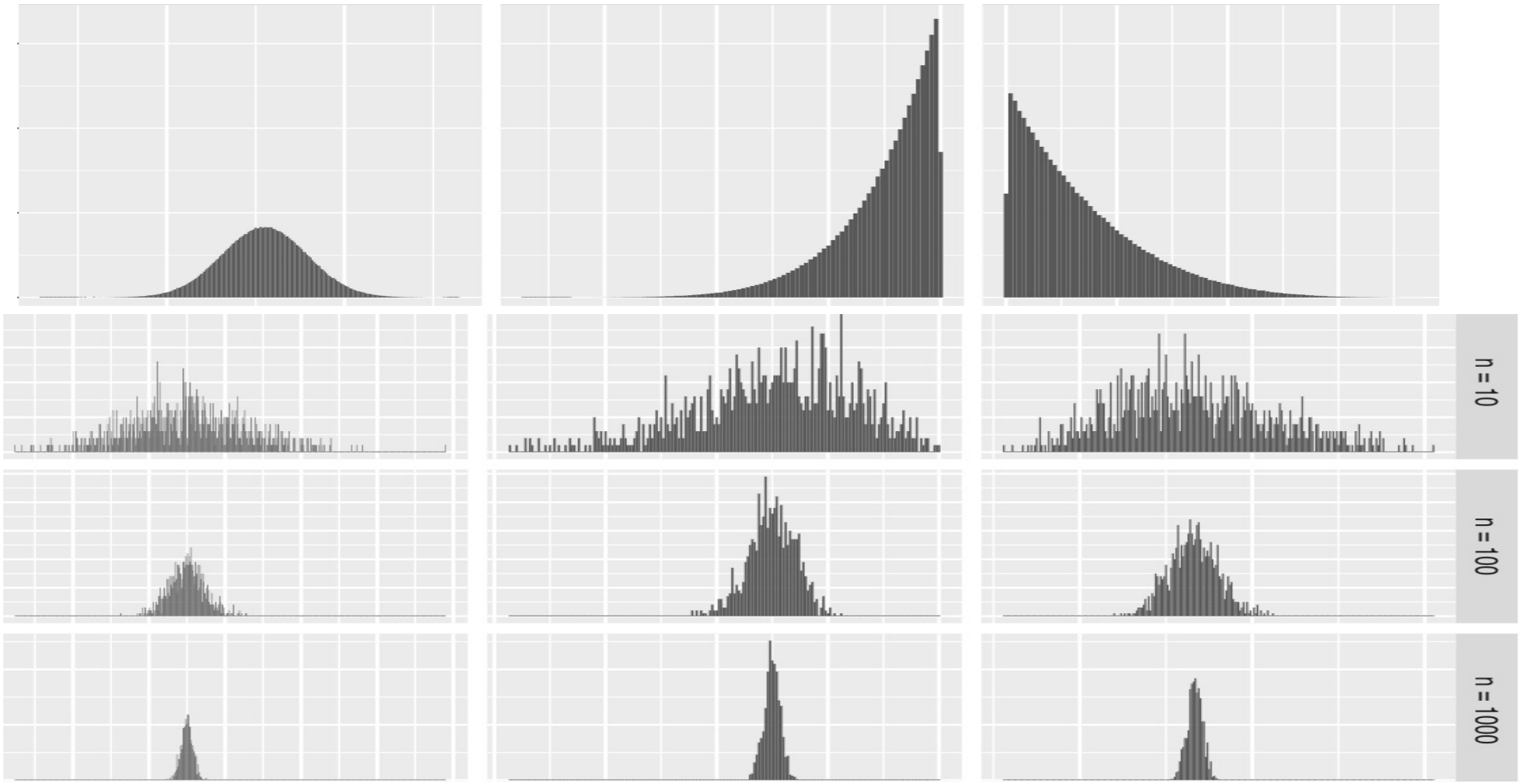
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 28 PART 1: 4/2/2021

Sections 8.3, 8.4

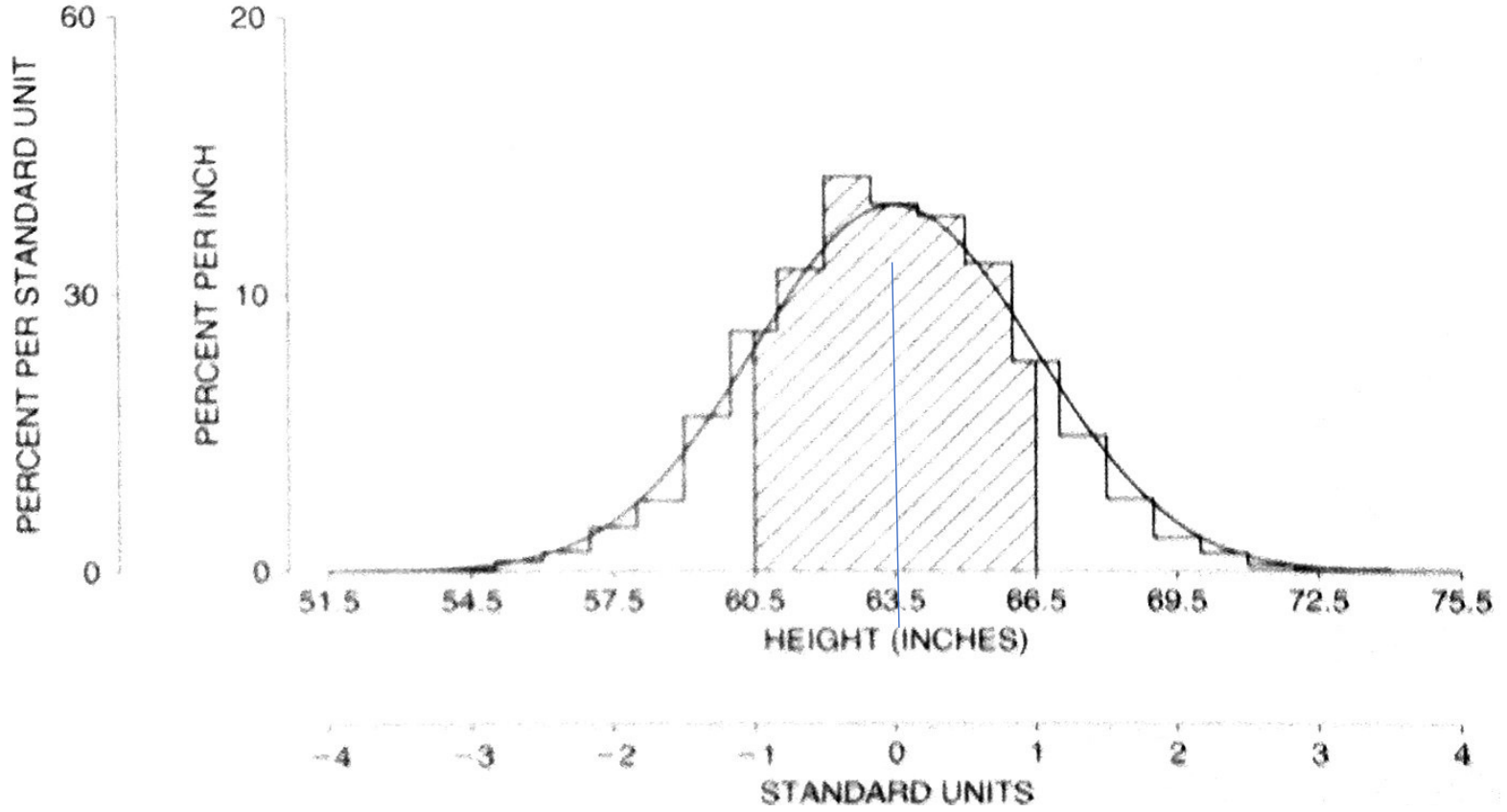
Using the Central Limit Theorem

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

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

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%.



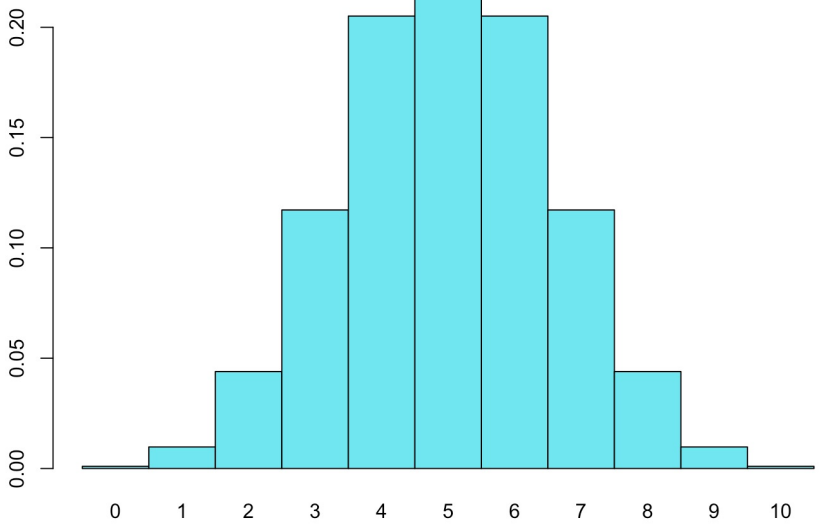
Picture from *Statistics,* by Freedman, Pisani, & Purves

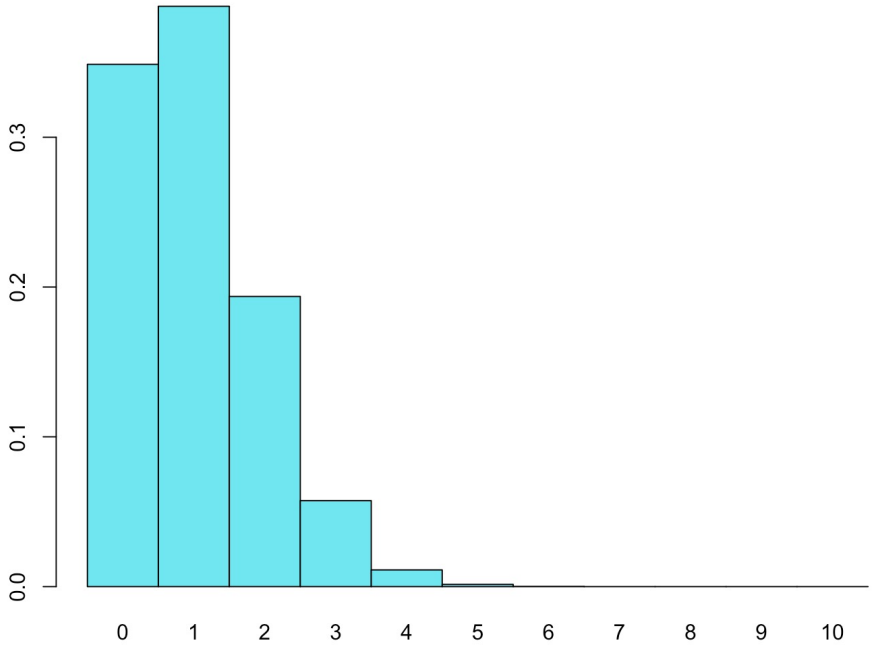
How large is "large"?

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"

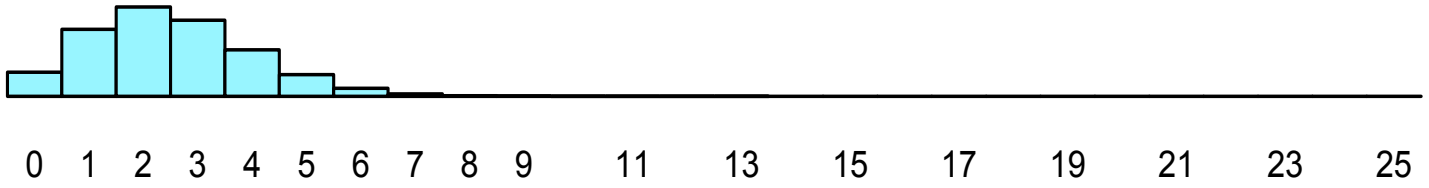
Answer: Well, it depends.



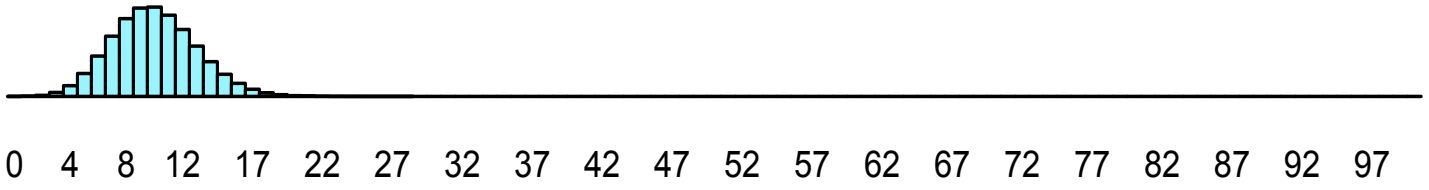


When p is small

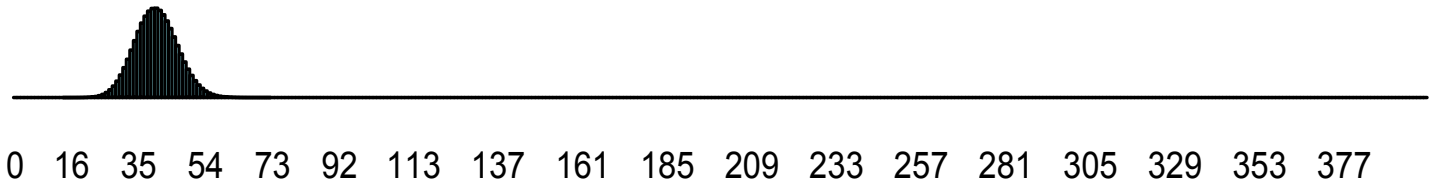
n=25



n=100



n=400



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.

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)?