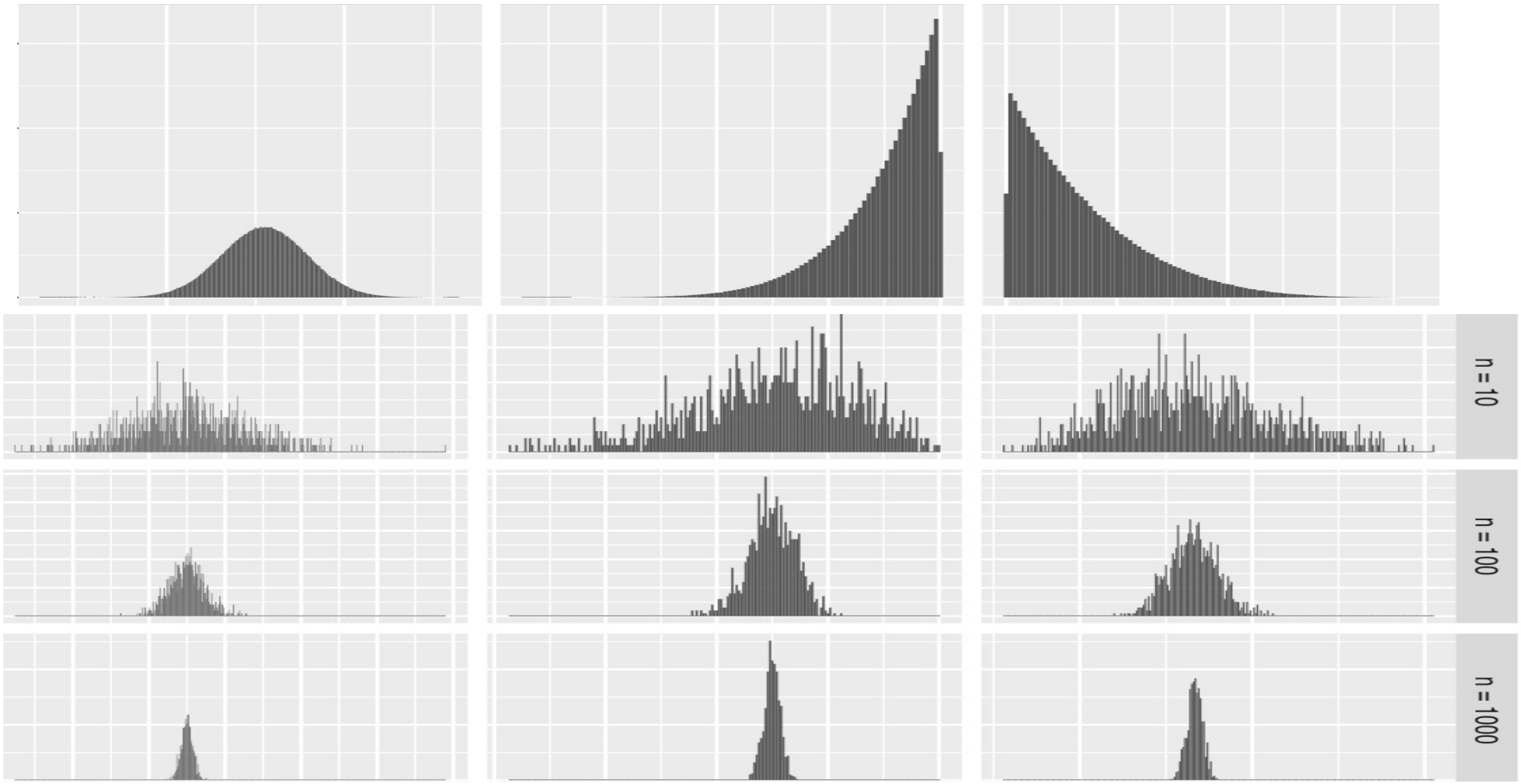
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 27: 3/31/2021

Sections 8.2, 8.3

The Normal distribution & using the Central Limit Theorem

The Central Limit Theorem

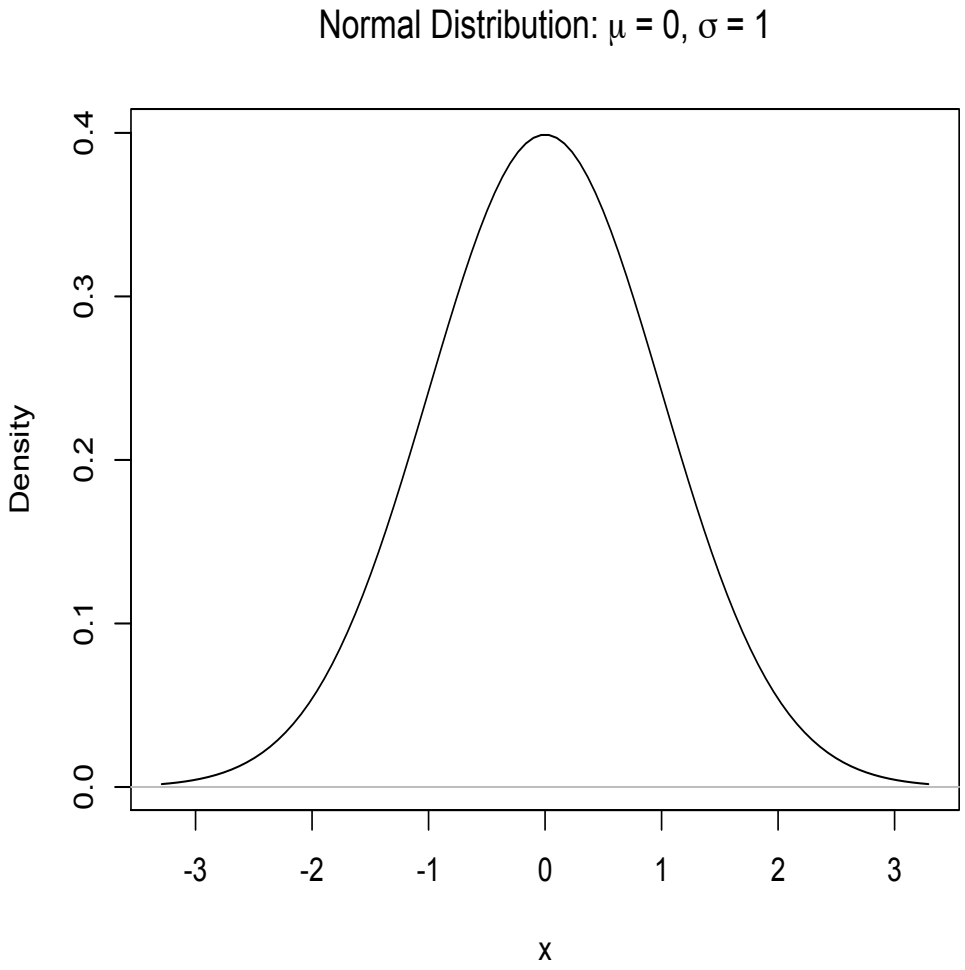
- Suppose that $X_1, X_2, \ldots, X_n$ are iid with mean μ and SD σ
- $S_n = X_1 + X_2 + \cdots + X_n$ is the sample sum
- then the distribution of S_n is *approximately normal* for large enough n.

- The distribution is approximately normal (bell-shaped) centered at $\boldsymbol{E(S_n) = n\mu}$ and the width of this curve is defined by $\boldsymbol{SD(S_n) = \sqrt{n}\,\sigma}$

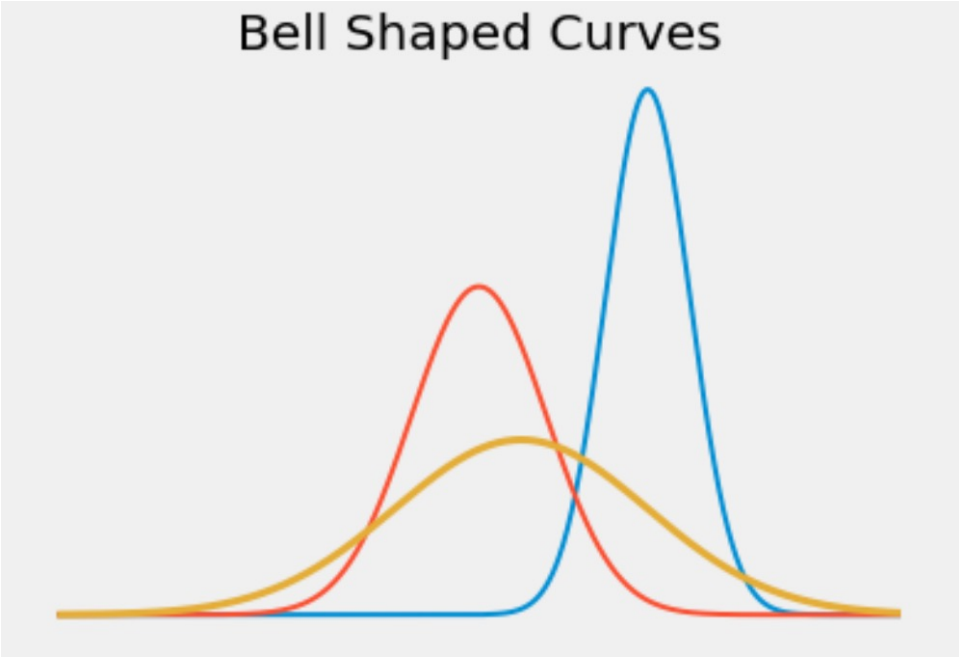
Bell curve: the Standard Normal Curve

- 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 $|z|>3$

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



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.

Standard normal cdf

- $\Phi(z) = \int_{-\infty}^{z} \phi(x)dx$

Standard normal cdf, symmetries, percentiles

Example: Find the area
(a) to the right of 1.25

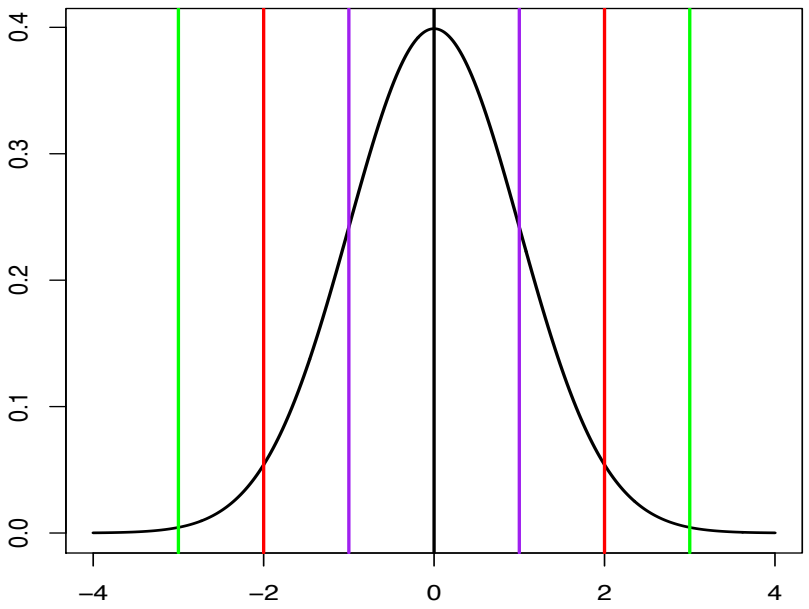
(b) between -0.3 and 0.9

(c) outside -1.5 and 1.5.

2. (a) Find z such that $\Phi(z) = 0.95$

(b) Find z so that the area in the **middle** is 0.95 ($\Phi(z) - \Phi(-z) = 0.95$)

How do we approximate the area of a range in a histogram?



- Many histograms bell-shaped, but not on the same scale, and not centered at 0
- Need to convert a value to ***standard units*** – see how many SDs it is above or below the average
- Then we can **approximate** the area of the histogram using the area under the standard normal curve (using for example, stats.norm.cdf for the actual numerical computation)

Total area under the curve = 100%
Curve is symmetric about 0
The areas between 1, 2 and 3 SDs away:
Between -1 and 1 the area is 68.27%
Between -2 and 2 the area is 95.45%
Between -3 and 3 the area is 99.73%

- 68%-95%-99.7% rule

  (**Empirical rule**)

Normal approximations : standard units

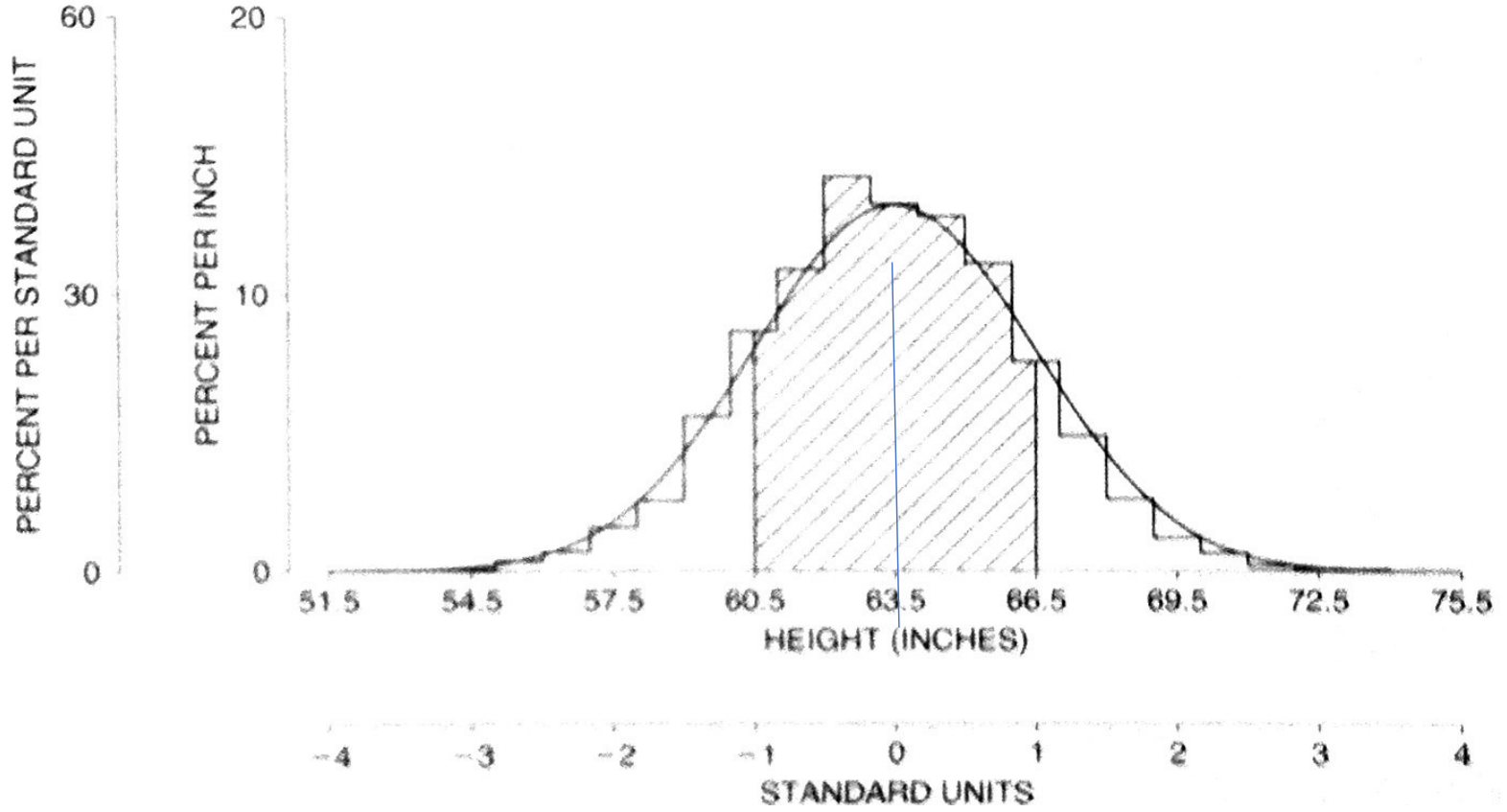
- Let X be any random variable, with expectation μ and SD σ, consider a new random variable that is a linear function of X, created by shifting X to be centered at 0, and dividing by the SD. If we call this new rv X^*, then X^* has expectation _____ and SD _____.
- $E(X^*) =$
- $SD(X^*) =$

- This new rv does not have units since it measures how far above or below the average a value is, in SD's. Now we can compare things that we may not have been able to compare.

- Because we can convert anything to standard units, ***every normal curve is the same.***

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%.



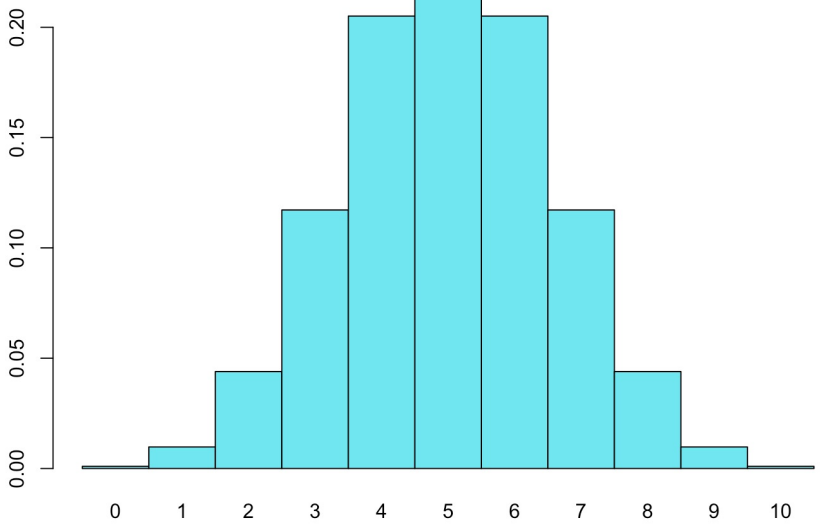
Picture from *Statistics,* by Freedman, Pisani, & Purves

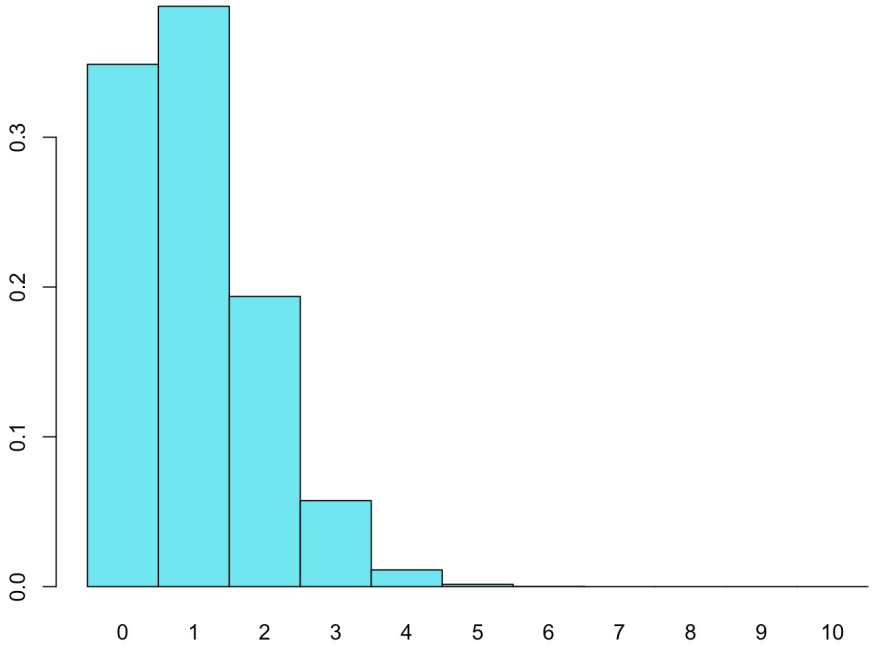
How large is "large"?

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"

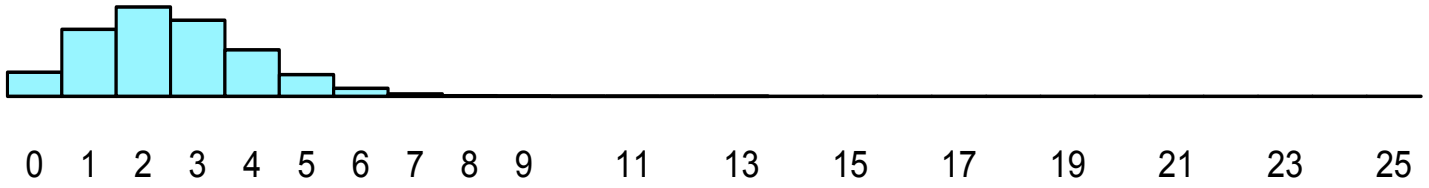
Answer: Well, it depends.



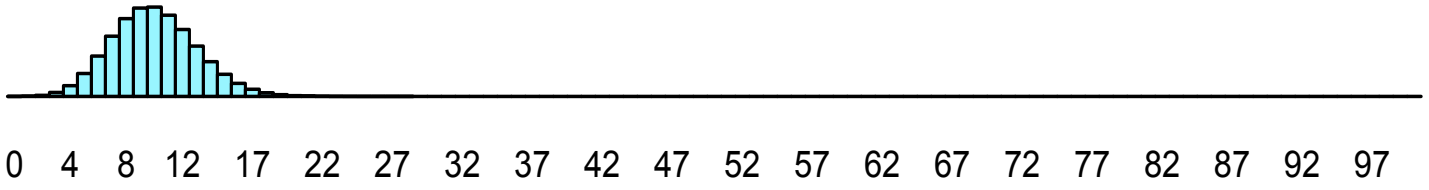


When p is small

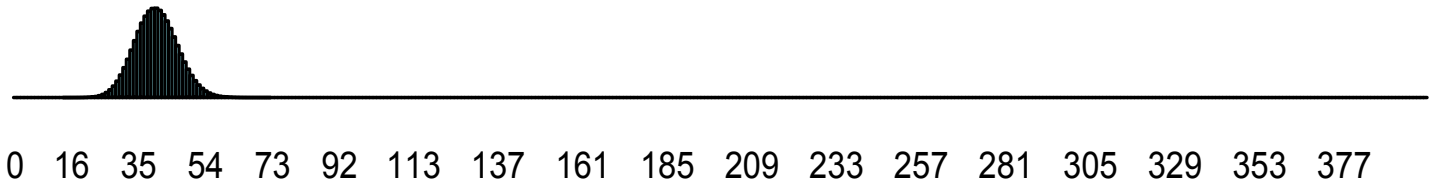
n=25



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)?