ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

DISTRIBUTION OF RANDOM VARIABLES:

THE NORMAL DISTRIBUTION

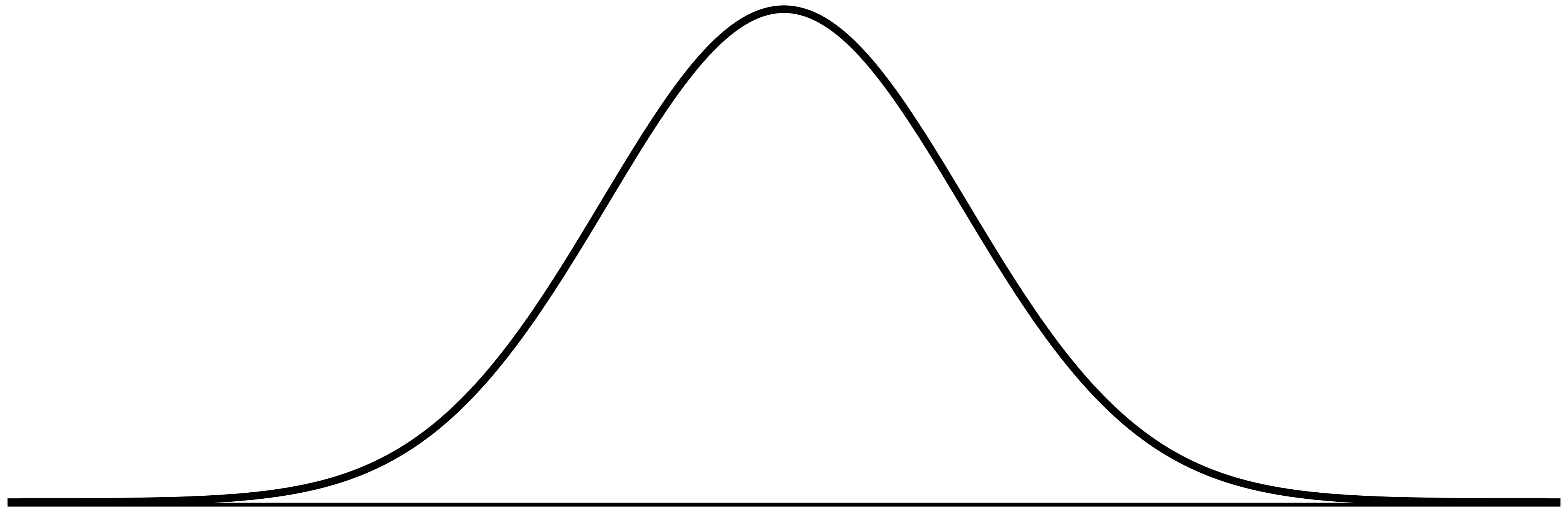
UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

The Normal Distribution

- The **normal distribution** is perhaps the most common, and most universal of all distributions one encounters in practice.
- Frequently used to model distributions of random variables with unknown distributions.



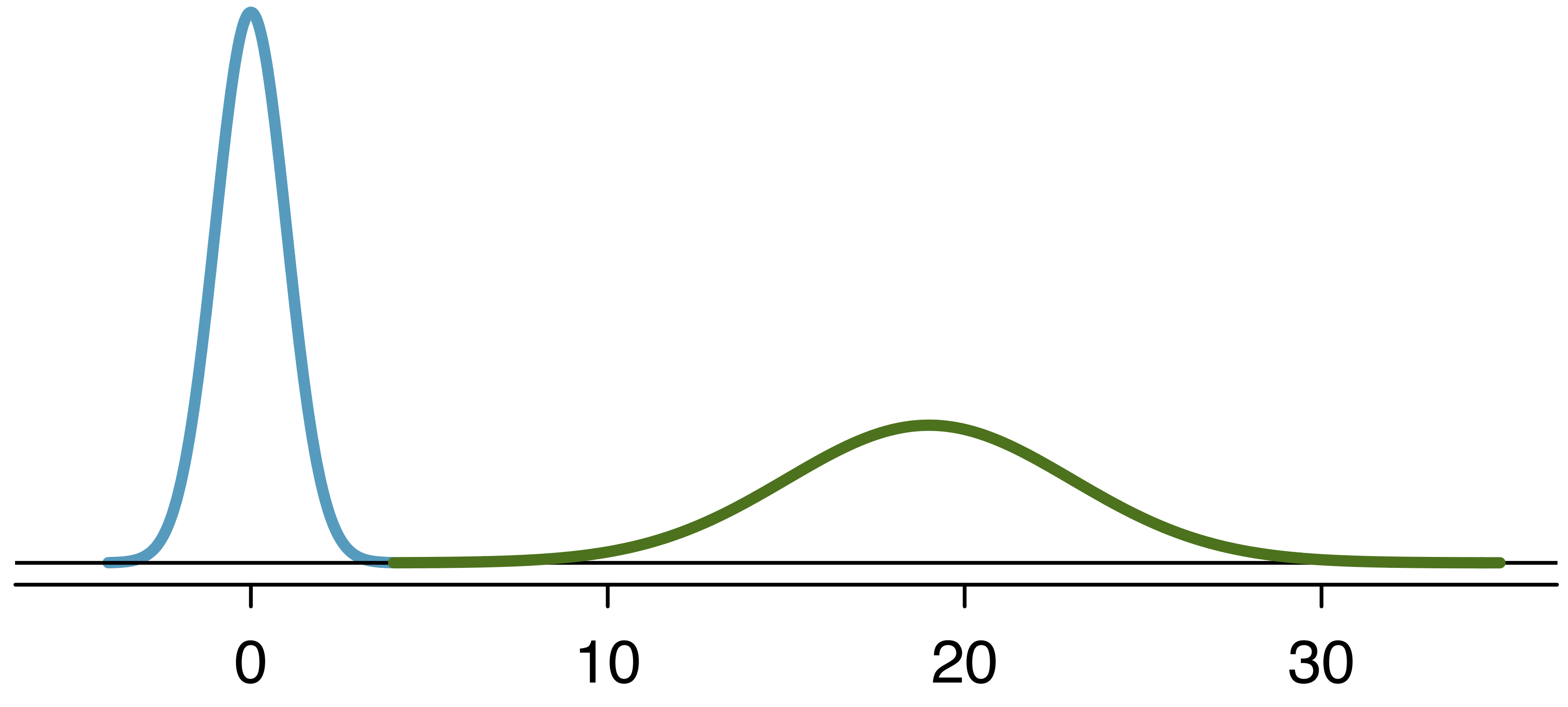
Normal Distribution. *Figure from Ch.4 of Openintro text.*

- This approximation is often justified by the **Central Limit Theorem:**

  The average of n samples of a random variable (under some general conditions) is a random variable whose distribution approaches a normal distribution as $n \to \infty$.

The Normal Distribution

- The **normal distribution** is parametrized by two parameters, the mean μ and standard deviation σ.
- We use the notation $X \sim \mathcal{N}(\mu, \sigma)$ to mean that the random variable X follows the normal



Two normal distributions: $\mathcal{N}(0,1)$ and $\mathcal{N}(19,4)$.
Figure from Ch. 4, Openintro text.

- The probability distribution function (pdf) associated with $\mathcal{N}(\mu, \sigma)$ is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\ \sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- $\mathcal{N}(0,1)$ is usually referred to as **the standard normal distribution.**

Standardizing and Z-scores

- If a random variable X follows the distribution $\mathcal{N}(\mu, \sigma)$, then the random variable $\dfrac{X-\mu}{\sigma}$ follows the **standard normal distribution $\mathcal{N}(0,1)$.**

- The **Z-score** of an observation x is the number of standard deviations that an observation falls above or below the mean. In other words:

$$Z = \frac{x-\mu}{\sigma}.$$

- Z-scores provide us with a way to put data on a standardized scale, to allow comparisons — and can be used with any distribution.

Example: Standardizing and Z-scores

- **Example:** If we want to compare the grades of two students on two different standardized tests (say from two different countries), where the total points are 20 and 100 respectively, we can instead compare their Z-scores.

- **Example:** Let $X \sim \mathcal{N}(1,3)$ and suppose we observe $x = 7.5$

Then, the associated Z-score is: $$Z = \frac{x - \mu}{\sigma} = \frac{7.5 - 1}{3} \approx 2.167$$

- This tells us that x is 2.167 standard deviations *above the mean.*

Example: Standardizing and Z-scores

- **Example:** Let $X \sim \mathcal{N}(1,3)$ and suppose we observe $x = -7.5$

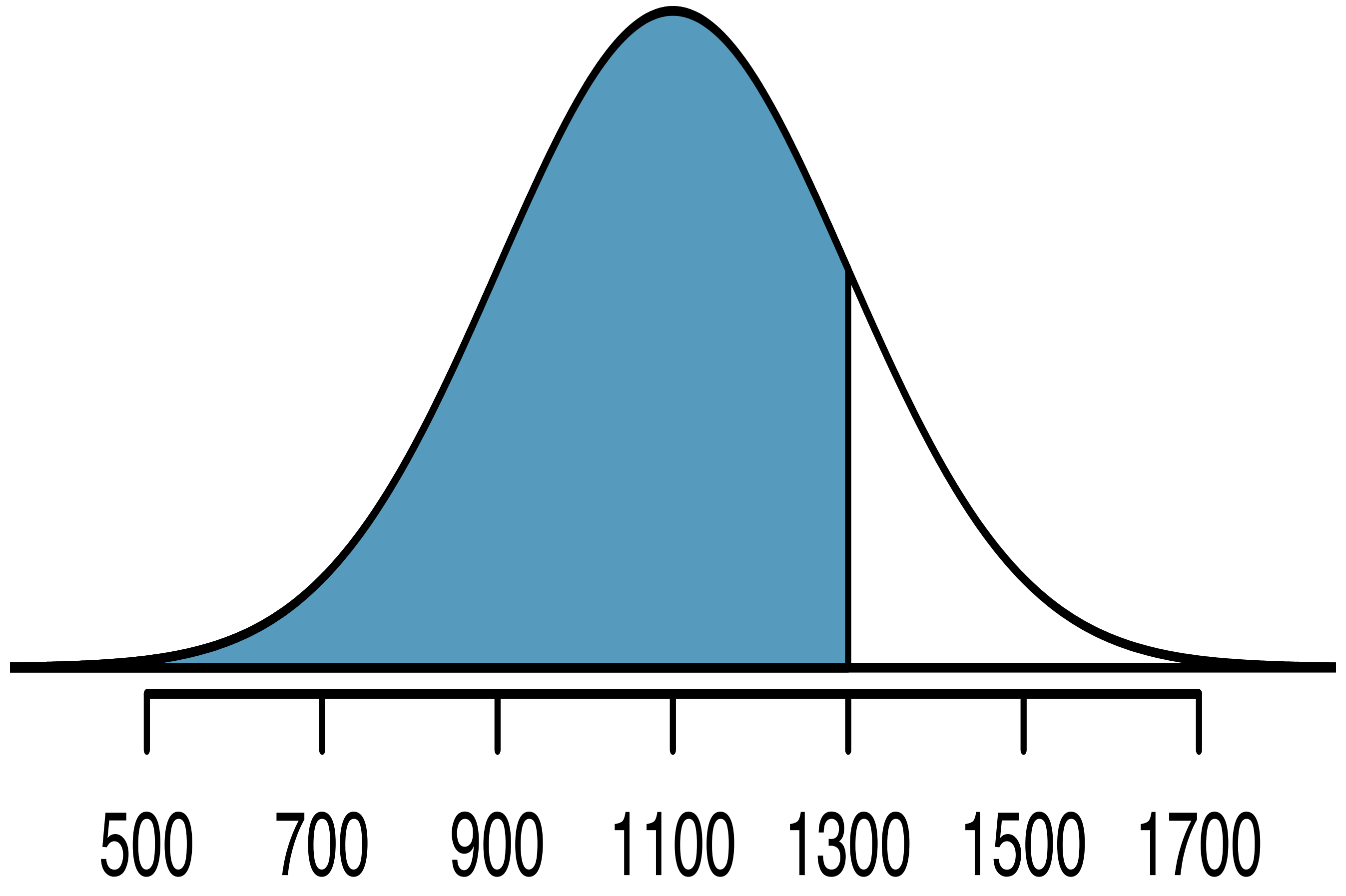
Then, the associated Z-score is: $$Z = \frac{x - \mu}{\sigma} = \frac{-7.5 - 1}{3} \approx -2.833$$

- This tells us that x is -2.833 standard deviations *below the mean.*

Finding Tail Areas

- In statistics, we will often need to calculate tail areas of distributions.
- **Example:** Ann scores 1300 on the SAT. Suppose that SAT scores have a normal distribution with $\mu = 1100$ and $\sigma = 200$. What fraction of people have an SAT score below 1300?
- The question is asking for a tail area like the shaded area on the pdf curve shown.
- Equivalently it is asking for:

$$\mathbb{P}(X \leq 1300) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{1300} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$



The shaded area represents the proportion of people who scores less than 1300.
Figure from Ch 4 of Openintro text

Finding Tail Areas

- Many ways to calculate this:
- **Statistical software:** Matlab, R, Python, SPSS. For example, in Matlab, the command

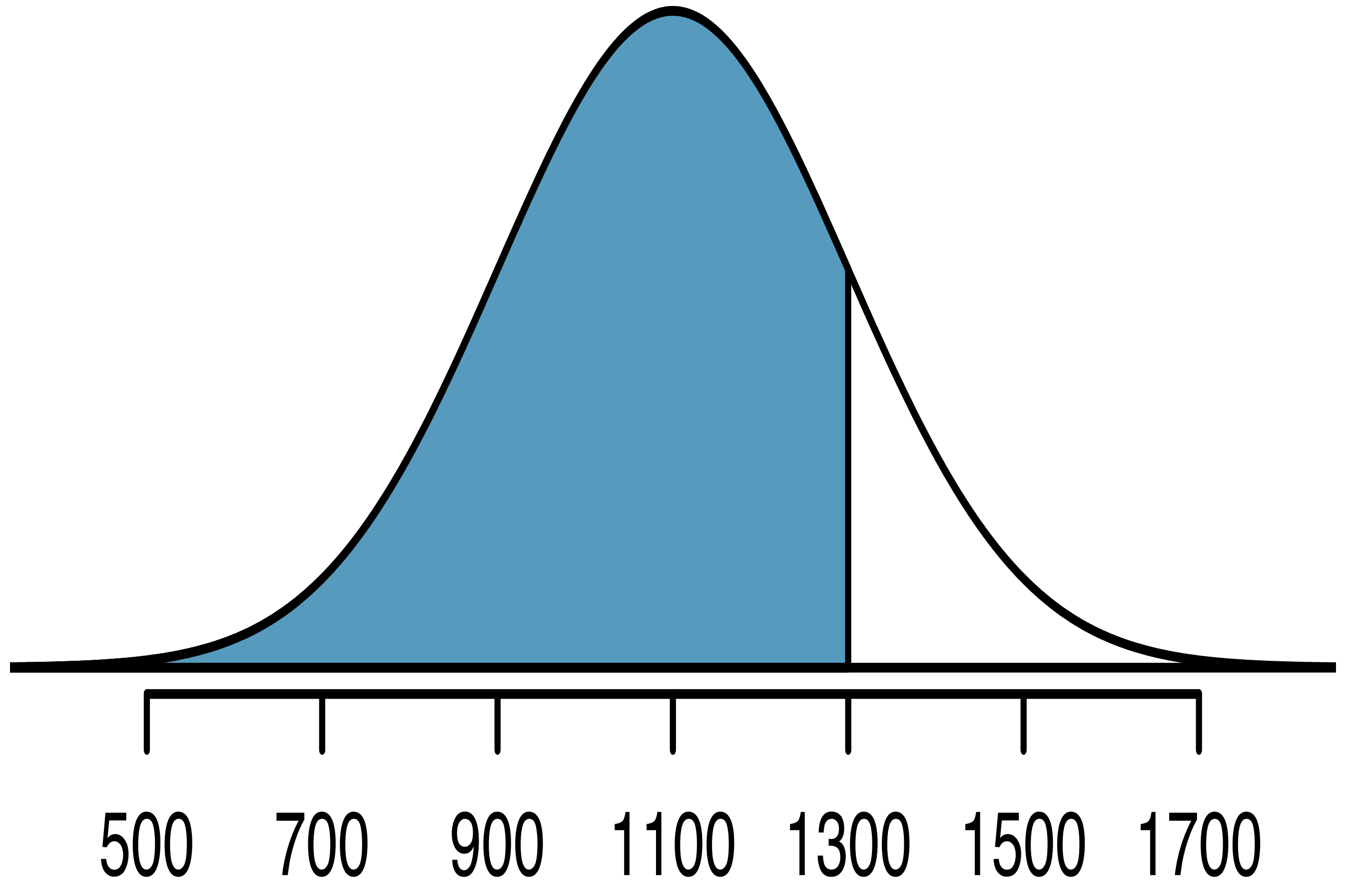
  normcdf(1300,1100, 200) gives
  $\mathbb{P}(X \leq 1300) \approx 0.8413.$

- **Using probability tables:** First calculate Z-score

$$Z = \frac{x-\mu}{\sigma} = \frac{1300-1100}{200} = 1.$$

  Then read the probability from the table for (Standard) Normal Probabilities using the associated Z-score.

- **Graphing calculators…**



The shaded area represents the proportion of people who scores less than 1300.
Figure from Ch 4 of Openintro text

Linking Probability and Statistics: Law of Large Numbers

- A Law of Large Numbers (LLN) is a theorem that describes the result of performing the same experiment many times.

- It states that the average of the results should be close to the expected value, and should get closer as more trials are performed.

- **The Strong Law of Large Numbers (SLLN)** states that if $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with expected value $\mu = \mathbb{E}(X_1) = \mathbb{E}(X_2) = \ldots$, then the **sample average**

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \quad \text{satisfies} \quad \mathbb{P}(\lim_{n \to +\infty} \bar{X}_n = \mu) = 1$$

- **Remark:** an intuitive reason that the statement is probabilistic is that $\bar{X}_n$ is itself a random variable!

Linking Probability and Statistics: Central Limit Theorem

- We saw that the SLLN tells us that the random variable $\bar{X}_n$ converges to $\mathbb{E}(X_i)$.

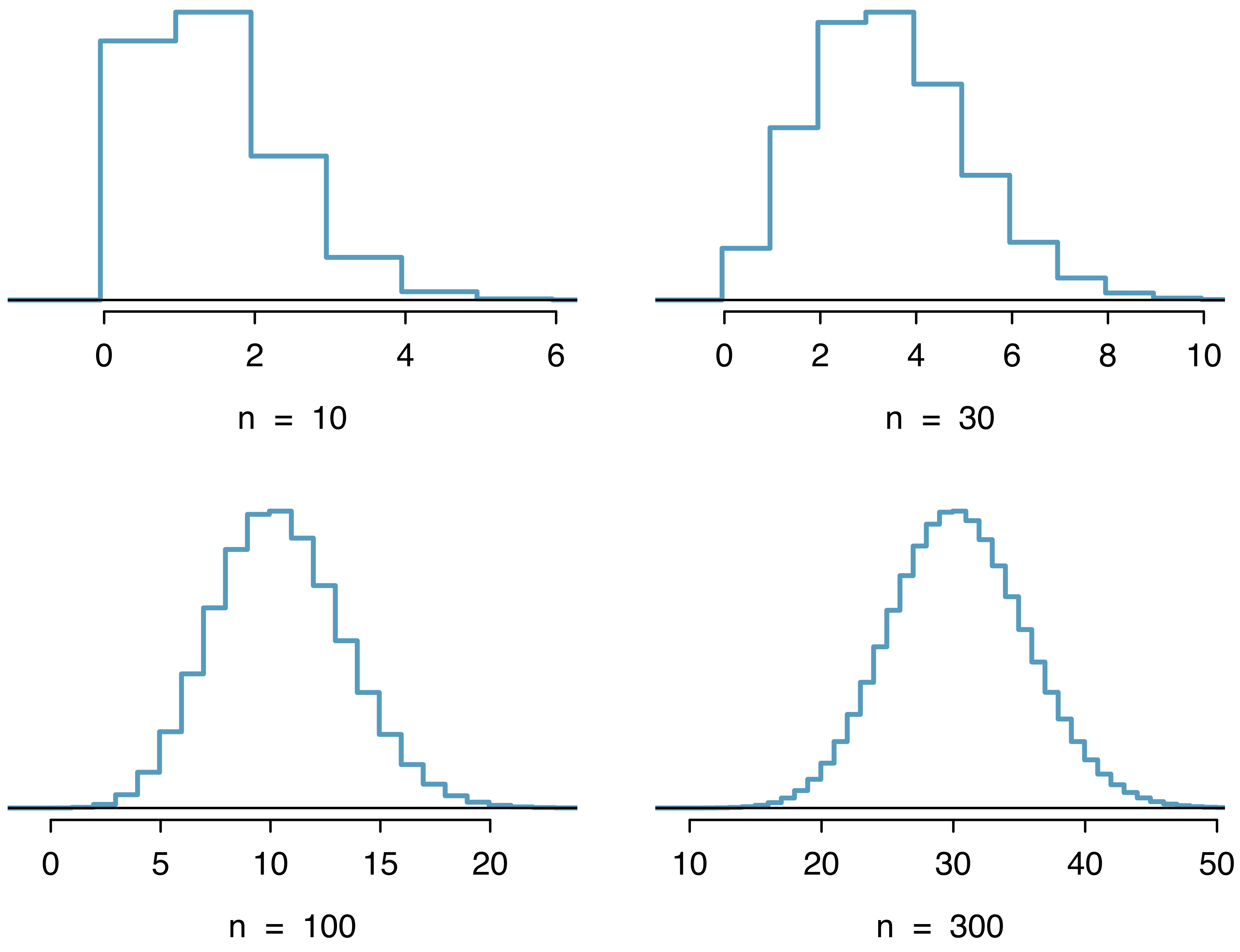
- But what is its distribution?

- **The Central Limit Theorem** states that if $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with expected value $\mu = \mathbb{E}(X_1) = \mathbb{E}(X_2) = \ldots$, and variance $\sigma^2 < \infty$, then the random variable

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \quad \text{converges to a normal distribution } \mathcal{N}(0,\sigma^2)$$

- Why is this important to statistics?

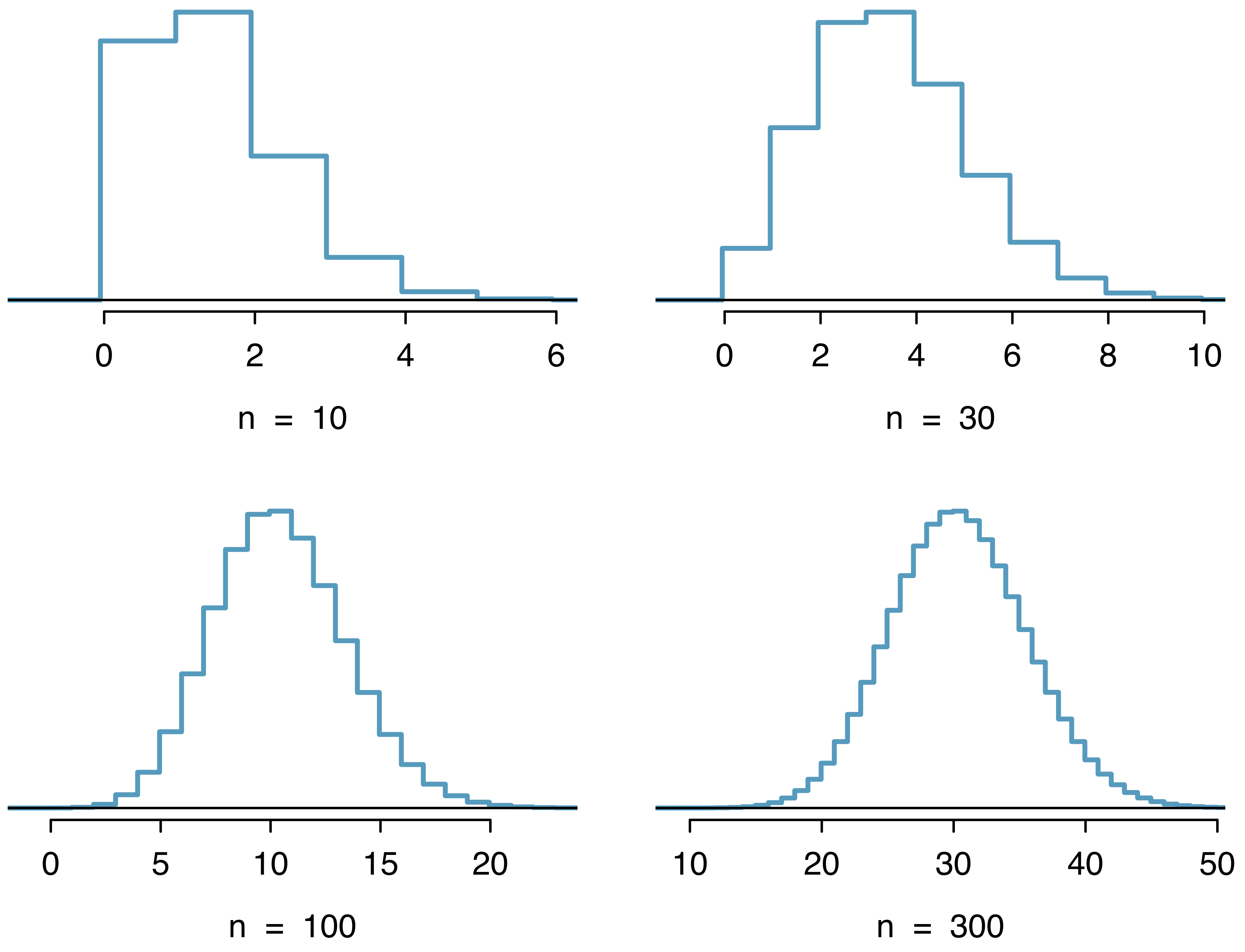
Approximating Binomials with Normal Distributions



Histograms of samples from $B(n, p)$ when $p = 0.10$ and $n = 10,\ 30,\ 100$, and 300, respectively. *Figure from open-intro text, Ch. 4.*

- Remember our question: What happens to a binomial distribution as the sample size increases?
- Because Binomials are sums of Bernoulli RV's, we can effectively use the CLT to justify approximating a binomial distribution with a normal distribution.

Approximating Binomials with Normal Distributions



Histograms of samples from $B(n, p)$ when $p = 0.10$ and $n = 10,\ 30,\ 100$, and 300, respectively. *Figure from open-intro text, Ch. 4.*

- The binomial distribution $B(n, p)$ is nearly normal when n is sufficiently large so that np and $n(1 - p)$ are both at least 10.
- The resulting normal distribution $\mathcal{N}(\mu, \sigma)$ has parameters $\mu = np$ and $\sigma = \sqrt{np(1 - p)}$ corresponding to the mean and standard deviation of the binomial distribution

Approximating Binomials with Normal Distributions

- Why is this useful? Recall that binomials satisfy $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.
- In statistical inference we will often need to calculate **tail probabilities** like $\mathbb{P}(X \geq k)$, which would be cumbersome with the binomial pmf.
- **Example:** Estimate the probability of observing 42 or fewer smokers in a sample of 400, if the true proportion of smokers is $p = 0.15$.

Solution:

- $np \geq 10$ and $n(1-p) \geq 10$, so we can approximate $B(n, p)$ with $\mathcal{N}(\mu, \sigma)$, where $\mu = np = 60$ and $\sigma = \sqrt{np(1-p)} \approx 7.14$.
- Next, either use statistical software to calculate $\mathbb{P}(X \leq 42)$ when $X \sim \mathcal{N}(\mu, \sigma)$, or use Z-scores: $Z = \dfrac{42 - 60}{7.14}$ to calculate the left-tail. Either way $\mathbb{P}(X \leq 42) \approx 0.0059$.