ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

DISTRIBUTION OF RANDOM VARIABLES:

THE BERNOULLI AND BINOMIAL DISTRIBUTIONS

UC San Diego
COMPUTER SCIENCE & ENGINEERING
HALICIOĞLU DATA SCIENCE INSTITUTE

Distributions of Random Variables

- In statistical inference (and data analysis, and many other areas) certain distributions arise quite frequently.
- Some important distributions for **discrete** random variables include the:
  - **Bernoulli distribution**
  - **Binomial distribution**
  - Geometric distribution, negative binomial distribution, poisson distribution …
- Some important distributions for **continuous** random variables include the:
  - **Normal distribution**
  - **Chi-squared distribution**
  - **t-distribution**
  - **F-distribution**
  - Logistic distribution …

Bernoulli Distribution

- Bernoulli random variables model processes with only two outcomes, say, "success" and "failure" (or heads and tails). A Bernoulli random variable typically assigns a value of 1 to "success" and 0 to "failure".

- Mathematically, a random variable X with a **Bernoulli** distribution takes the value 1 with probability p, and the value 0 with probability $1-p$.

- In other words $\mathbb{P}(X=1)=p$ and $\mathbb{P}(X=0)=(1-p)$.

- Consequently it has the **pmf** $p_X(x)=\begin{cases} p & \text{if} \quad x=1 \\ 1-p & \text{if} \quad x=0 \\ 0 & \text{otherwise.} \end{cases}$

Bernoulli Distribution

- It is easy to calculate that the expected value of a Bernoulli random variable is

$$\mu := \mathbb{E}(X) = p.$$

- While its variance is

$$\sigma^2 = \mathbb{E}(X^2) - \mu^2 = p(1-p)$$

- What can we reasonably model as a Bernoulli random variable?
  - Coin flips
  - Voting preference of an individual in a two-party system
  - Whether a particular product is defective…

Binomial Distribution

- The **binomial distribution** describes the probability of having exactly k successes in n *independent* Bernoulli trials, each with probability of success p

- We use the notation $X \sim B(n, p)$ to mean that the random variable X follows the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0,1]$.

- In that case, its pmf satisfies, for $k = 0,1,\ldots,n$

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Distribution

- Why does the Binomial pmf for $X \sim B(n, p)$ satisfy

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for.} \quad k = 0,1,\ldots,n?$$

- The probability of having any specified k *independent* Bernoulli trials be “successes" while the remaining $n - k$ are failures is $p^k(1 - p)^{n-k}$

- There are $\binom{n}{k} := \dfrac{n!}{k!(n-k)!}$ ways to specify k successes and $n - k$ failures.

Binomial Distribution

- It is easy to calculate that the expected value of a Binomial random variable X *once we recognize that* $X = X_1 + X_2 + \ldots + X_n$ *where each* X_i *is an independent Bernoulli random variable.* Consequently:

$$\mu := \mathbb{E}(X) = \mathbb{E}(X_1 + \ldots + X_n) = pn.$$

- While its variance is

$$\sigma^2 = np(1 - p)$$

- This also follows from the fact that the variance of a sum of independent random variables is the sum of the variances, and the variance of a Bernoulli random variable being $p(1 - p)$.

Example: Binomial Distributions

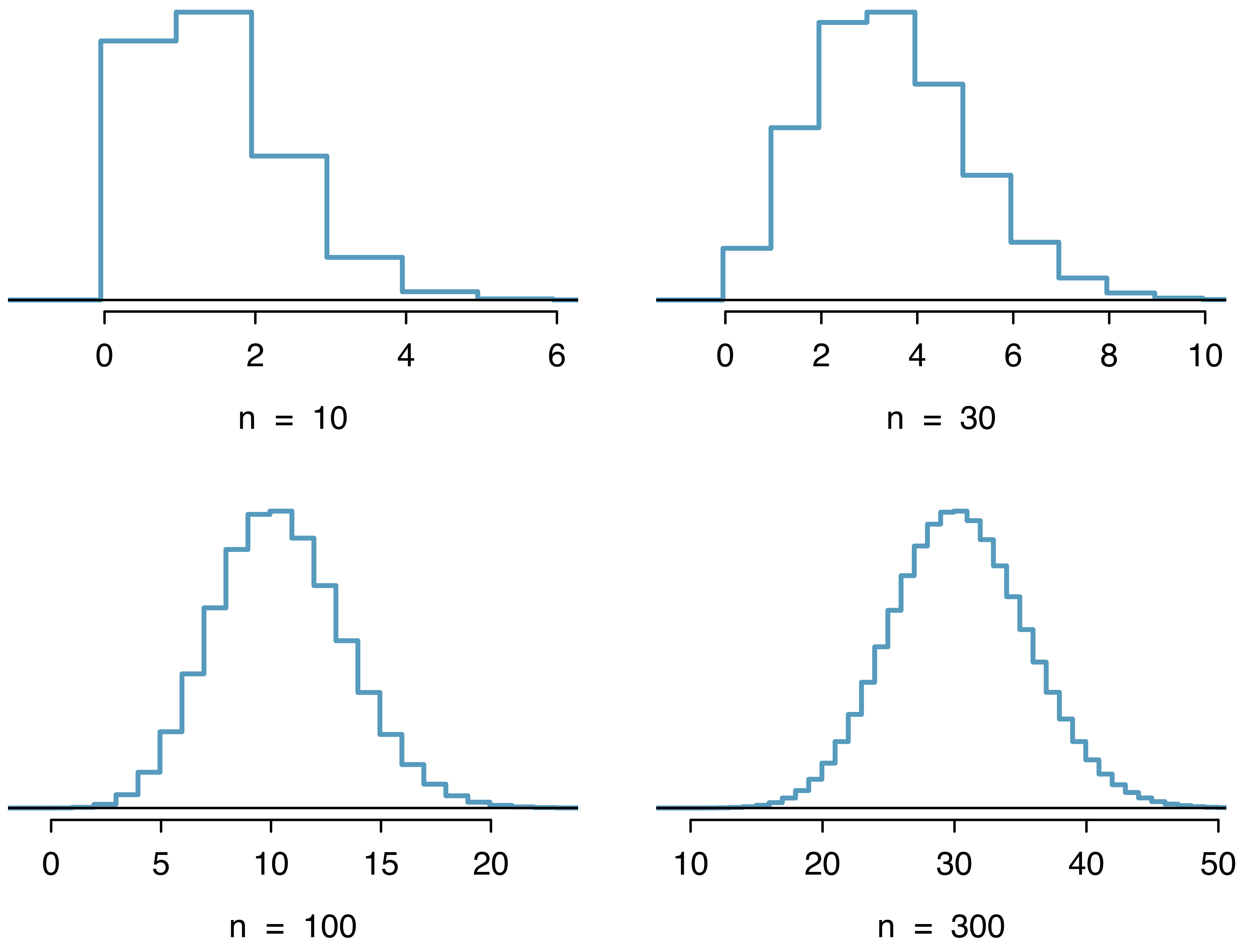
- **Example:** Suppose the probability of a child having a peanut allergy is $2\,\%$, and suppose that a classroom has 30 children (who are unrelated).

- Notice that we can use a Bernoulli distribution with parameter p, to model the probability that an individual child has a peanut allergy, and a binomial rv, X, with parameters $n = 30,\ p = 0.02$ to model the number of kids with the allergy

Example: Binomial Distributions

- **Q:** What is the probability that none of them has a peanut allergy?
- **A:** $$\mathbb{P}(X = 0) = 0.98^{30} \approx 0.5455$$
- **Q:** What is the probability that 3 of them have a peanut allergy?
- **A:** $$\mathbb{P}(X = 3) = \binom{30}{3} 0.02^{3} \times 0.98^{27} \approx 0.0188$$

Binomial Distributions: What does the PMF Look Like



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.*

- What happens to the shape of the distributions as the sample size n increases? What distribution does the hollow histogram resemble?