

Medical Statistics and Data Analysis

Lecture 6: Probability Distributions

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)

Assistant Professor of Biomedical Engineering

ShanghaiTech University

October 12, 2026

Learning Goals of This Lecture

- Discrete probability distributions
 - Discrete random variables
 - Probability-mass function (probability distribution)
 - CDF, expected value, and variance
 - Permutations and combinations
 - Binomial distribution
 - Poisson distribution
- Continuous probability distributions
 - Normal distribution
 - Covariance

Discrete Probability Distribution

Random Variables

- A random variable is a function that assigns numeric values to different events in a sample space.
- Two types of random variables: discrete and continuous.
 - A random variable for which there exists a discrete set of numeric values is a discrete random variable.
 - A random variable whose possible values cannot be enumerated is a continuous random variable.

Probability-Mass Function for a Discrete Random Variable

- The values taken by a discrete random variable and its associated probabilities can be expressed by a rule or relationship called a **probability-mass function** (PMF).
- A **probability-mass function**, sometimes also called a **probability distribution**, is a mathematical relationship, or rule, that assigns to any possible value r of a discrete random variable X the probability $Pr(X = r)$. This assignment is made for all values r that have positive probability.

- A probability-mass function (PMF) can be displayed in a tabular form, or it can be expressed as a mathematical formula giving the probabilities of all possible values.

TABLE 4.1 Probability-mass function for the hypertension-control example

$Pr(X = r)$	.008	.076	.265	.411	.240
r	0	1	2	3	4

- The probability of any particular value must be between 0 and 1, and the sum of the probabilities of all values must be exactly equal to 1.
- Thus, $0 < Pr(X = r) \leq 1$, $\sum_r Pr(X = r) = 1$, where the summation is taken over all possible values that have positive probability.

Relationship of Probability Distributions to Frequency Distributions

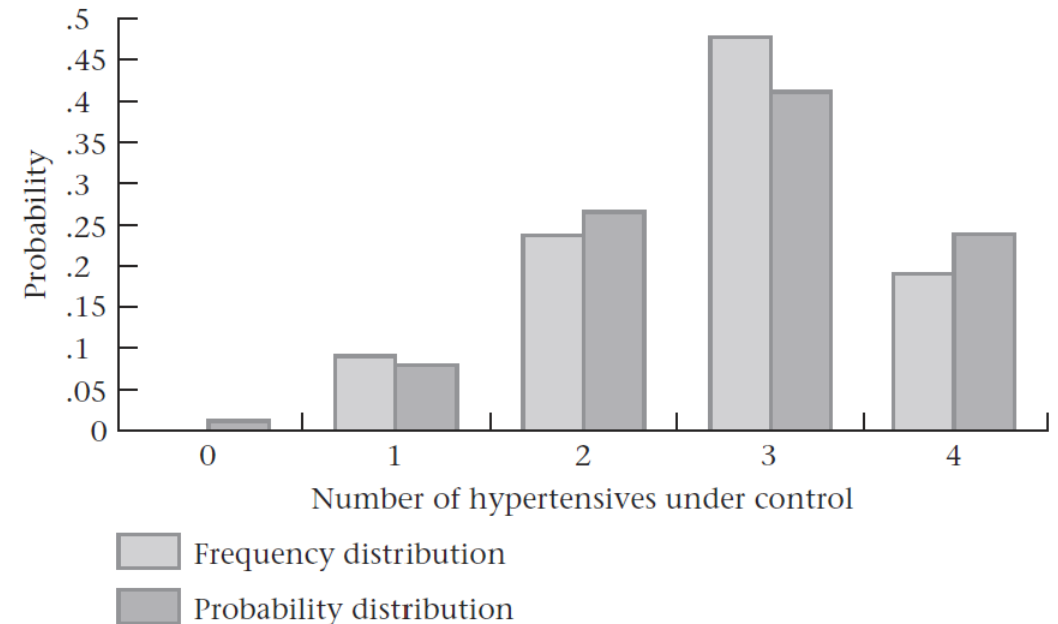
- Frequency distribution is described as a list of each value in the data set and a corresponding count of how frequently the value occurs.
- If each count is divided by the total number of points in the sample, then the frequency distribution can be considered as a sample analog to a probability distribution.
- Frequency distribution gives the actual proportion of points in a sample that correspond to specific values. Hence, the appropriateness of a model can be assessed by comparing the observed sample-frequency distribution with the probability distribution, also called as **goodness-of-fit**.

TABLE 4.2 Comparison of the sample-frequency distribution and the theoretical-probability distribution for the hypertension-control example

Number of hypertensives under control = r	Probability distribution $Pr(X = r)$	Frequency distribution
0	.008	.000 = 0/100
1	.076	.090 = 9/100
2	.265	.240 = 24/100
3	.411	.480 = 48/100
4	.240	.190 = 19/100

- The PMF derived from the binomial distribution is compared with the frequency distribution to determine whether the drug behaves with the same efficacy as predicted.

FIGURE 4.1 Comparison of the frequency and probability distribution for the hypertension-control example



Expected Value of a Discrete Random Variable

- If a random variable has a large number of values with positive probability then the pmf is not a useful measure.
- Measures of location and spread can be developed for a random variable in much the same way as for samples.
- The analog of the arithmetic mean $\bar{x}$ is called the **expected value** of a random variable, or **population mean**, and is denoted by $E(X)$ or μ and represents the “average” value of the random variable.
- The expected value of a discrete random variable is defined as $E(X) = \mu = \sum_{i=1}^k x_i P(X = x_i)$ where the x_i 's are the values the random variable assumes with positive probability.

TABLE 4.3 Probability-mass function for the number of episodes of otitis media in the first 2 years of life

r	0	1	2	3	4	5	6
$Pr(X=r)$	.129	.264	.271	.185	.095	.039	.017

$$E(X) = 2.038$$

Variance of a Discrete Random Variable

- The analog of the sample variance (s^2) for a random variable is called the **variance of a random variable**, or **population variance**, and is denoted by $Var(X)$ or σ^2 .

$$Var(X) = \sigma^2 = E[(X - \mu)^2] = \sum_{i=1}^k (x_i - \mu)^2 P(X = x_i)$$

- OR

$$Var(X) = \sigma^2 = E(X^2) - [E(X)]^2 = \sum_{i=1}^k x_i^2 P(X = x_i) - \mu^2$$

- The standard deviation of a random variable X , denoted by $sd(X)$ or σ , is defined by the square root of its variance.

$$\sigma = \sqrt{Var(X)}$$

- The variance represents the spread, relative to the expected value, of all values that have positive probability.
- Approximately 95% of the probability mass falls within two standard deviations (2σ) of the mean of a random variable of normal distribution.

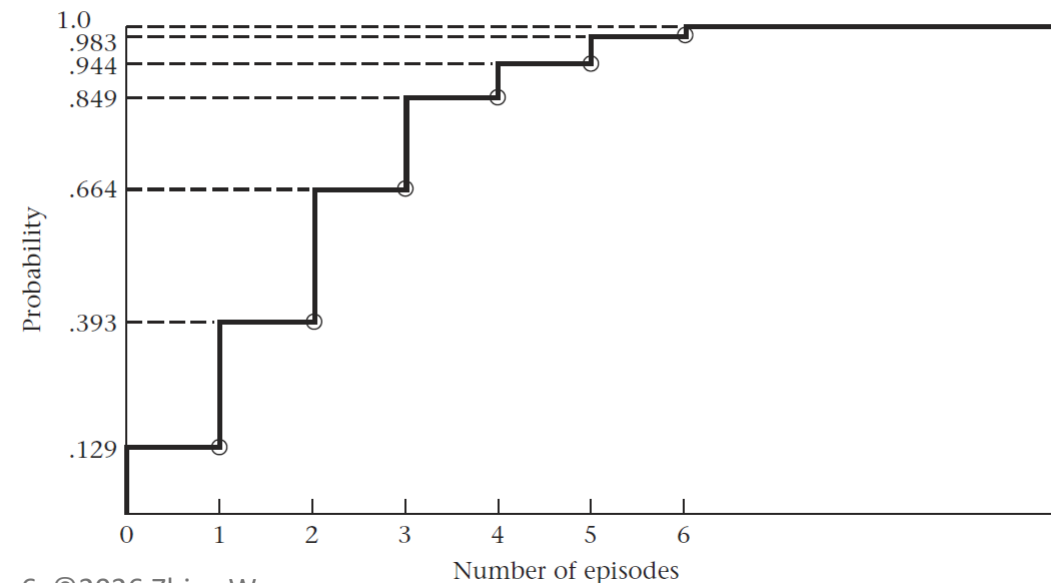
$$\sigma = \sqrt{1.967} \approx 1.403$$

Cumulative-Distribution Function of a Discrete Random Variable

- The **cumulative-distribution function (cdf)** of a random variable X is denoted by $F(X)$ and, for a specific value of x of X , is defined by $Pr(X \leq x)$ and denoted by $F(x)$.
- Discrete and continuous random variables can be distinguished based on each variable's cdf.

FIGURE 4.2 Cumulative-distribution function for the number of episodes of otitis media in the first 2 years of life

- For a discrete random variable, the cdf looks like a series of steps, called the **step function**.
- With the increase in number of values, the cdf approaches that of a smooth curve.
- For a continuous random variable, the cdf is a smooth curve.



Permutations

- In a **matched-pair design**, each sample/case is matched with a normal control of the same sex and age.
- Once the first control is chosen, the second control can be chosen in $(n - 1)$ ways. The third control can be chosen in $(n - 2)$ ways and so on. Thus, the first two controls can be chosen in any of $[n \times (n - 1)]$ ways.

- The number of permutations of n things taken k at a time is

$${}_nP_k = n(n - 1) \times \cdots \times (n - k + 1)$$

- It represents the number of ways of selecting k items of n , where the order of selection is important.
- For selecting n objects out of n , where order of selection matters

$${}_nP_n = n(n - 1) \times \cdots \times [(n - n + 1)] = n(n - 1) \times \cdots \times 2 \times 1$$

- The special symbol used for this quantity is $n!$, called **n factorial**.

$$n! = n(n - 1) \times \cdots \times 2 \times 1$$

$$0! = 1$$

Alternative Form for Permutations

- Expressing permutations in terms of factorials

$$\begin{aligned} {}_nP_k &= n(n-1) \times \cdots \times (n-k+1) \\ &= \frac{n(n-1) \times \cdots \times (n-k+1) \times (n-k) \times (n-k-1) \times \cdots \times 1}{(n-k) \times (n-k-1) \times \cdots \times 1} \\ &= \frac{n!}{(n-k)!} \end{aligned}$$

Combination

- In an **unmatched study design**, cases and controls are selected in no particular order.
- Thus, the method of selecting n things taken k at a time without respect to order is referred to as the number of **combinations**.

$${}_nC_k = \binom{n}{k} = \frac{n(n-1) \times \cdots \times (n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$$

- For any non-negative integers, n, k where $n \geq k$
$$\binom{n}{k} = \binom{n}{n-k}$$

- That is,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- If $k!$ is replaced with $(n-k)!$:

$$\binom{n}{n-k} = \frac{n!}{(n-k)! [n - (n-k)]!} = \frac{n!}{(n-k)! k!} = \binom{n}{k}$$

Binomial Distribution

- A sample of n independent trials, each of which can have only two possible outcomes, which are denoted as “success” and “failure.”
- The probability of a success at each trial is assumed to be some constant p , and hence the probability of failure at each trial is $1 - p = q$.
- If a neutrophil is denoted by an x and a non-neutrophil by an o .
- What is the probability of $oxo ox = Pr(oxo ox)$?

TABLE 4.5 Possible orderings for 2 neutrophils of 5 cells

<i>xxooo</i>	<i>oxxoo</i>	<i>ooxox</i>
<i>xoxoo</i>	<i>oxoxo</i>	<i>oooxx</i>
<i>xooxo</i>	<i>oxoox</i>	
<i>xooox</i>	<i>ooxxo</i>	

- That is, the probability of k successes within n trials:

$$\binom{n}{k} p^k (1 - p)^{n-k} = \binom{n}{k} p^k q^{n-k}$$

- The distribution of the number of successes in n statistically independent trials, where the probability of success on each trial is p , is known as the **binomial distribution** and has a probability-mass function given by

$$\Pr(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n$$

- Let X be a binomial random variable with parameters n and p .
- Let Y be a binomial random variable with parameters n and $q = 1 - p$.
- Then,

$$\Pr(X = k) = \binom{n}{k} p^k q^{n-k} = \binom{n}{n-k} q^{n-k} p^k = \Pr(Y = n - k)$$

- The probability of obtaining k successes for a binomial random variable X with parameters n and p is the same as the probability of obtaining $n - k$ successes for a binomial random variable Y with parameters n and q .

Evaluation of Binomial Distribution

- For sufficiently large n , the normal distribution can be used to approximate the binomial distribution and tables of the normal distribution can be used to evaluate binomial probabilities.
 - Conditions: $np \geq 5$ or 10, and p is not too far away from 0.5
- If the sample size is not large enough to use normal approximation, then a computer program (e.g., R) can be used to evaluate binomial probabilities.

TABLE 4.7 Calculation of the probability of at least 5 deaths among 100 women 60–64 years of age in 2004

```
This computes  $\Pr(X \leq 4)$ 
> pbinom(4,100, 0.009)
[1] 0.997809
This computes  $\Pr(X \geq 5)$ 
> 1-pbinom(4,100, 0.009)
[1] 0.00219104
```

Expected Value and Variance of the Binomial Distribution

- The expected value of a discrete random variable is

$$E(X) = \sum_{i=1}^R x_i \Pr(X = x_i)$$

- In the special case of a binomial distribution, the only values that take on positive probability are $0, 1, 2, \dots, n$ and these occur with probabilities

$$\binom{n}{0} p^0 q^n, \quad \binom{n}{1} p^1 q^{n-1}, \quad \dots$$

- Thus,

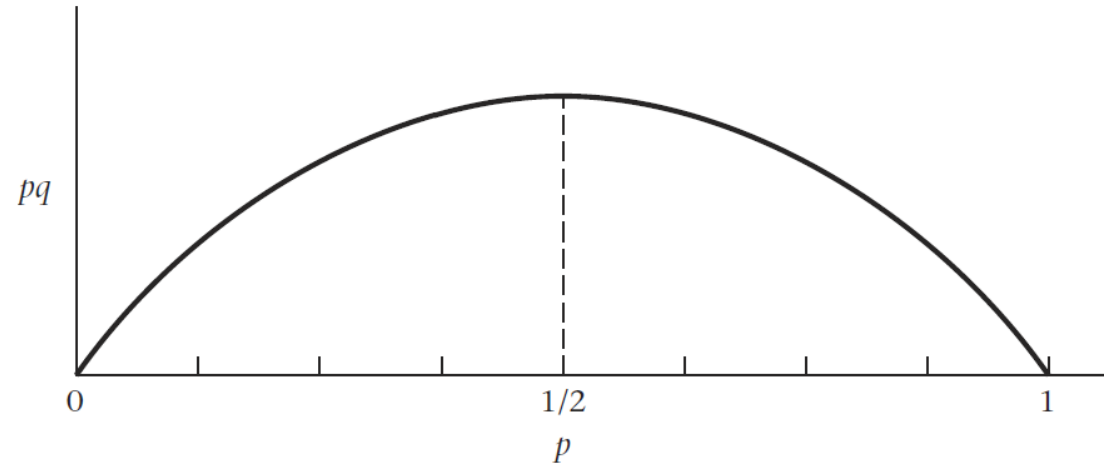
$$E(X) = \sum_{k=0}^n k \binom{n}{k} p^k q^{n-k}$$

that is, the expected value $= np$.

- Similarly, variance

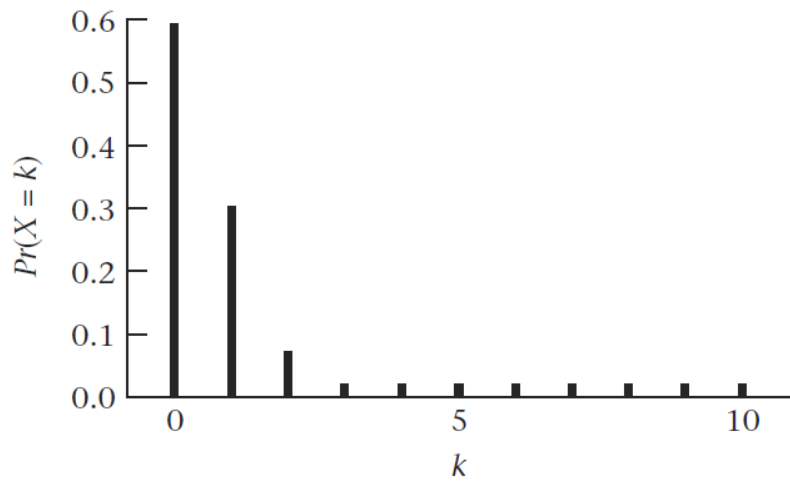
$$\text{Var}(X) = \sum_{k=0}^n (k - np)^2 \binom{n}{k} p^k q^{n-k} = npq = np(1 - p)$$

FIGURE 4.4 Plot of pq versus p

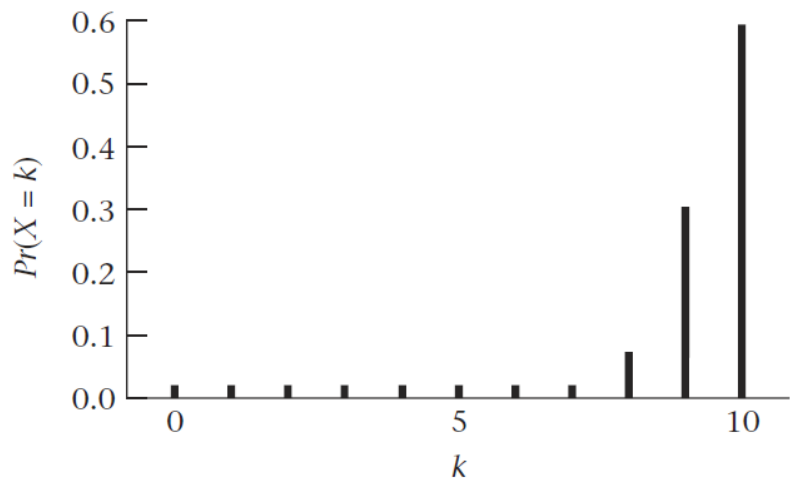


- The expected number of successes in n trials is the probability of success in one trial multiplied by n , which equals np .
- For a given number of trials n , the binomial distribution has the highest variance when $p = \frac{1}{2}$.
- Variance decreases as p moves away from $\frac{1}{2}$, becoming 0 when $p = 0$ or 1.

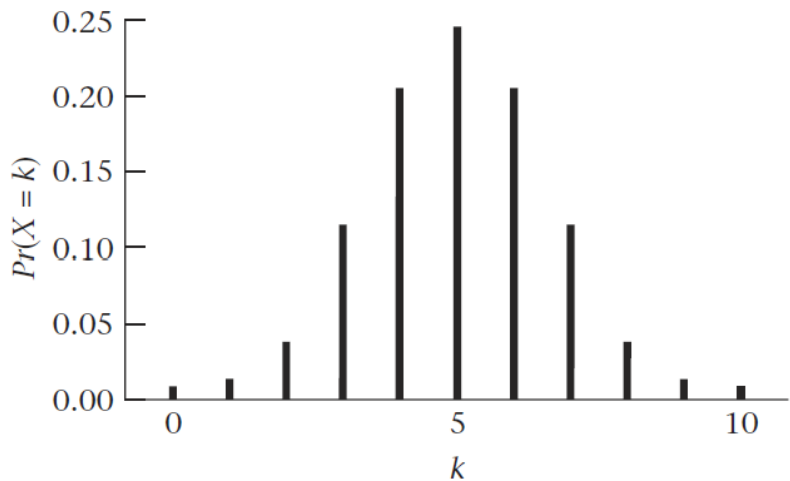
FIGURE 4.5 The binomial distribution for various values of p when $n = 10$



(a) $n = 10, p = .05$



(b) $n = 10, p = .95$



(c) $n = 10, p = .50$

Poisson Distribution

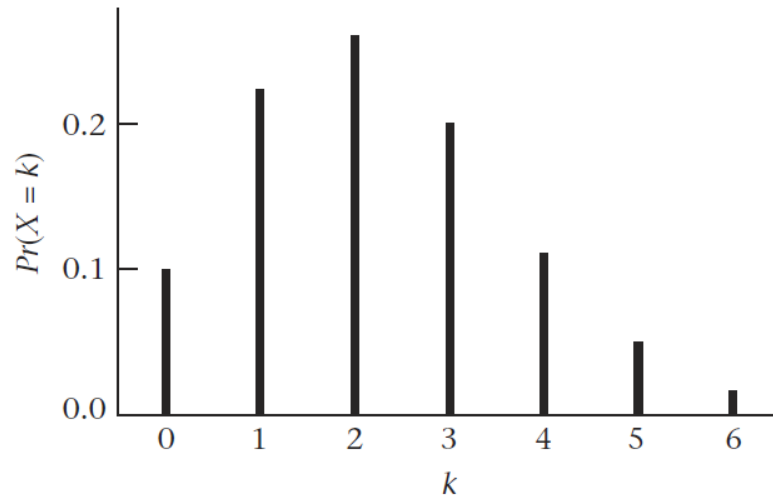
- The (second) most frequently used discrete distribution after binomial distribution. It is usually associated with rare events.
- The probability of k events occurring in a time period t for a Poisson random variable with parameter λ is

$$\Pr(X = k) = \frac{e^{-\mu} \mu^k}{k!}, \quad k = 0, 1, 2, \dots$$

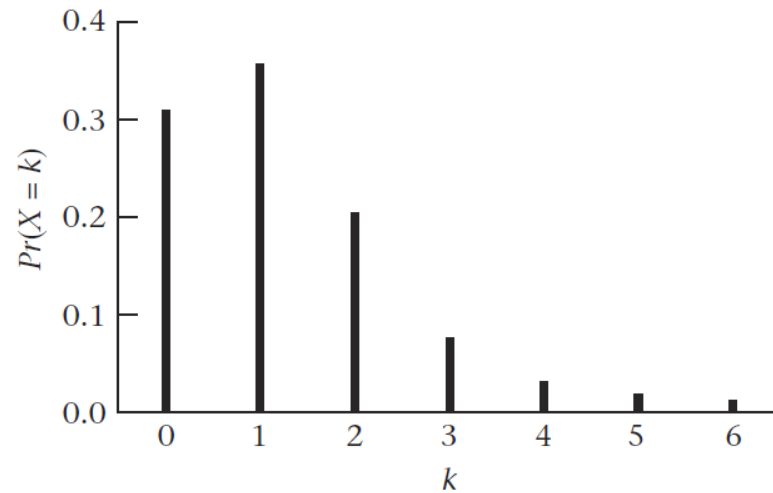
where $\mu = \lambda t$ and e is approximately 2.71828.

- λ = expected number of events / unit time; μ = expected number of events over time period t .
- For a Binomial distribution, number of trials n is finite, and the number of events can be no larger than n .
- For Poisson distribution, the number of trials is essentially infinite and the number of events can be indefinitely large. However, probability of k events becomes very small as k increases.

FIGURE 4.6 Distribution of the number of deaths attributable to typhoid fever over various time intervals



(a) 6 months



(b) 3 months

- Calculate the probability distribution of the number of deaths over a 6-month period and 3-month period.
- The plotted distribution tends to become more symmetric as the time interval increases or more specifically as μ increases.

Expected Value and Variance of the Poisson Distribution

- For a Poisson distribution with parameter μ , the mean and variance are both equal to μ .
- This guideline helps identify random variables that follow the Poisson distribution.
- If we have a data set from a discrete distribution where the mean and variance are about the same, then we can preliminarily identify it as a Poisson distribution.

TABLE 4.10 Number of deaths attributable to polio during the years 1968–1977

Year	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977
Number of deaths	15	10	19	23	15	17	23	17	26	15

Variance is approximately the same as the mean, therefore, the Poisson distribution would fit well here.

Poisson Approximation to the Binomial Distribution

- The binomial distribution with large n and small p can be accurately approximated by a Poisson distribution with parameter $\mu = np$.
 - Conditions: $n \geq 20$, $p \leq 0.05$
- The mean of this distribution is given by np and the variance by npq .
- Note that $q \cong 1$ (is approximately equal to 1) for small p , and thus $npq \approx np$. That is, the mean and variance are almost equal.
- The binomial distribution involves expressions such as $\binom{n}{k}$ and $(1 - p)^{n-k}$ which are cumbersome for large n .

TABLE 4.12 An example of the Poisson approximation to the binomial distribution for $n = 100$, $p = .01$, $k = 0, 1, \dots, 5$

k	Exact binomial probability	Poisson approximation	k	Exact binomial probability	Poisson approximation
0	.366	.368	3	.061	.061
1	.370	.368	4	.015	.015
2	.185	.184	5	.003	.003

Medical Application Example of Poisson Distribution

- Suppose the probability of a rare adverse drug reaction is 1%. A total of 100 patients receive the treatment, and the occurrence of adverse reactions is independent across patients.
- The number of patients experiencing an adverse drug reaction follows a binomial distribution:

$$X \sim \text{Binomial}(100, 0.01)$$

- This can be approximated by a Poisson distribution:

$$X \approx \text{Poisson}(1)$$

- Therefore, the probability that **none of the patients experiences an adverse drug reaction** is approximately:

$$P(X = 0) \approx e^{-1} = 0.3679$$

That is, approximately **36.79%**.

Summary

In this chapter, we discussed:

- **Random variables** and the distinction between discrete and continuous variables.
- **Specific attributes of random variables**, including notions of probability-mass function (probability distribution), cdf, expected value, and variance.
- **Sample frequency distribution** was described as a sample realization of a probability distribution, whereas sample mean ($\bar{x}$) and variance (s^2) are sample analogs of the expected value and variance, respectively, of a random variable.
- **Binomial distribution** was shown to be applicable to binary outcomes (“success” and “failure”).
- **Poisson distribution** as a classic model to describe the distribution of rare events.

Continuous Probability Distribution

Introduction to Continuous Probability Distribution

- The normal, or Gaussian or “bell-shaped,” distribution is the cornerstone of most methods of estimation and hypothesis testing.
- Many random variables, such as distribution of birth weights or blood pressures in the general population, tend to follow approximately a normal distribution.
- Those variables that are not themselves normal oftentimes are closely approximated by a normal distribution when summed.
- Using normal distribution is desirable since it is easy to use and tables for it are more widely available than are tables for other distributions.

- To develop an analog for a continuous random variable to the concept of a probability-mass function, we need to know which values are more probable than others and to what degree.

Figure 5.1 The pdf of DBP in 35- to 44-year-old men

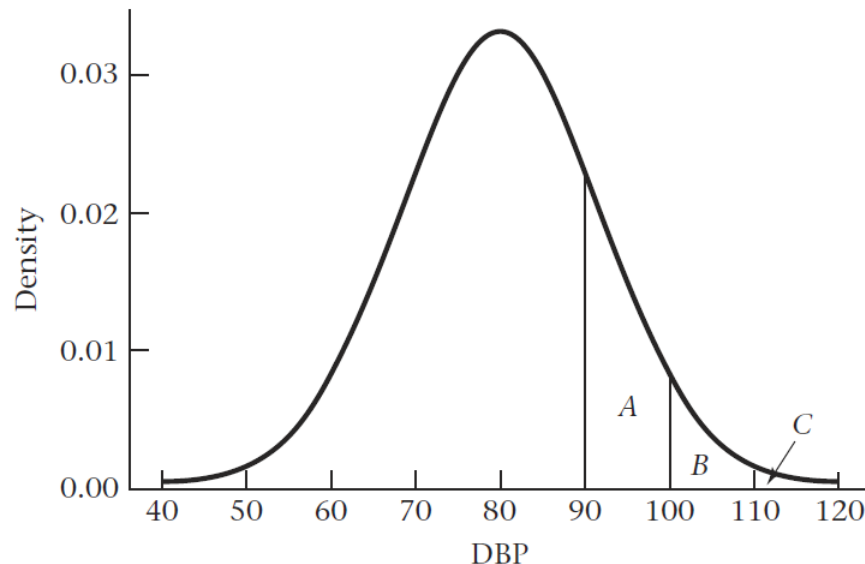
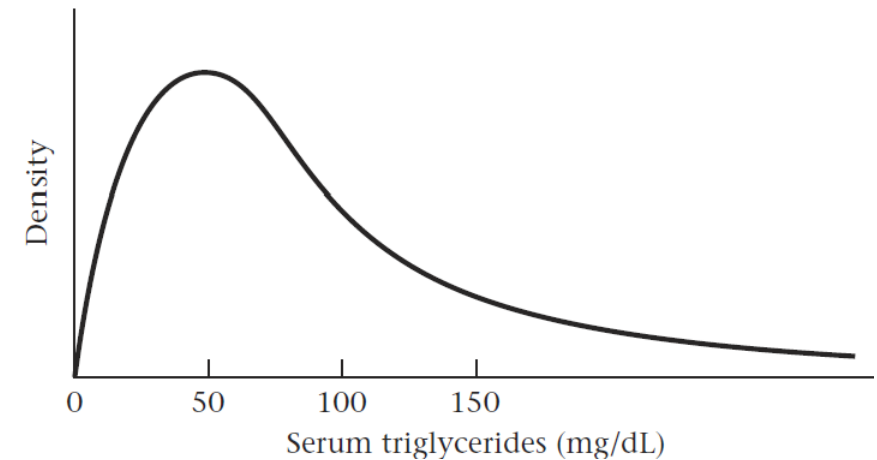


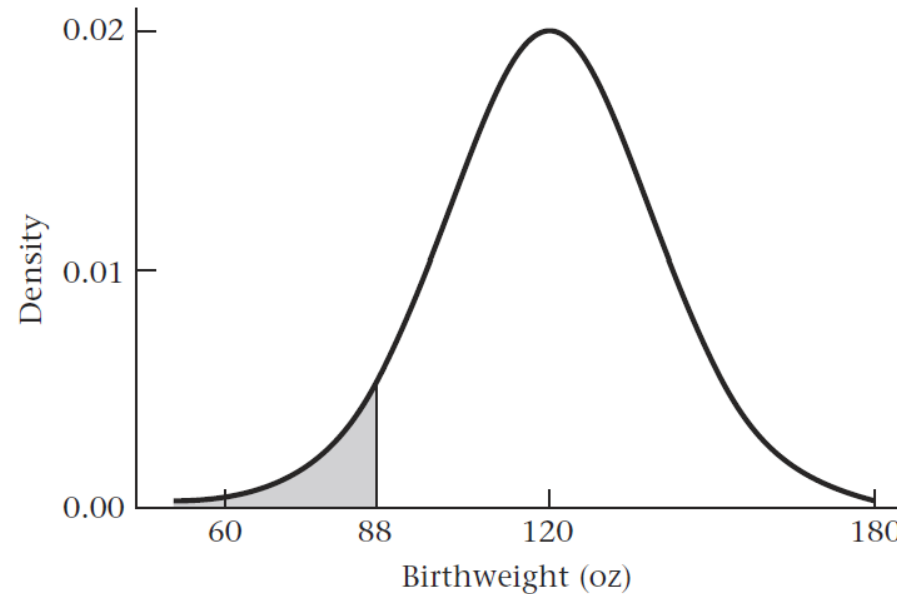
FIGURE 5.2 The pdf for serum triglycerides



- The cumulative-distribution function for the random variable X evaluated at the point a is defined as the probability that X will take on values $\leq a$. It is represented by the area under the pdf to the left of a .

- Generally, a distinction is not made between $Pr(X < x)$ and $Pr(X \leq x)$ when X is a continuous random variable because they represent the same quantity since the probability of individual values is 0; that is, $Pr(X = x) = 0$.
- Expected value and variance for continuous random variables have the same meaning as for discrete random variables.
- The expected value of a continuous random variable X , denoted by $E(X)$, or μ , is the average value taken on by the random variable.

FIGURE 5.3 The pdf for birthweight



- The **variance** of a continuous random variable X , denoted by $Var(X)$ or σ^2 is the average squared distance of each value of the random variable from its expected value, which is given by $E(X - \mu)^2$ and can be re-expressed in short form as $E(X^2) - \mu^2$. The **standard deviation**, or σ , is the squared root of the variance, that is, $\sigma = \sqrt{Var(X)}$.

Normal Distribution

- **Normal distribution** is the most widely used continuous distribution.
- It is vital to statistical work.
- It is also frequently called **Gaussian distribution**, after the well-known mathematician Karl Friedrich Gauss.
- Normal distribution is generally more convenient to work with than any other distribution.
- Body weights or DBPs approximately follow a normal distribution.
- Other distributions that are not themselves normal can be made approximately normal by transforming data onto a different scale, such as a logarithmic scale.



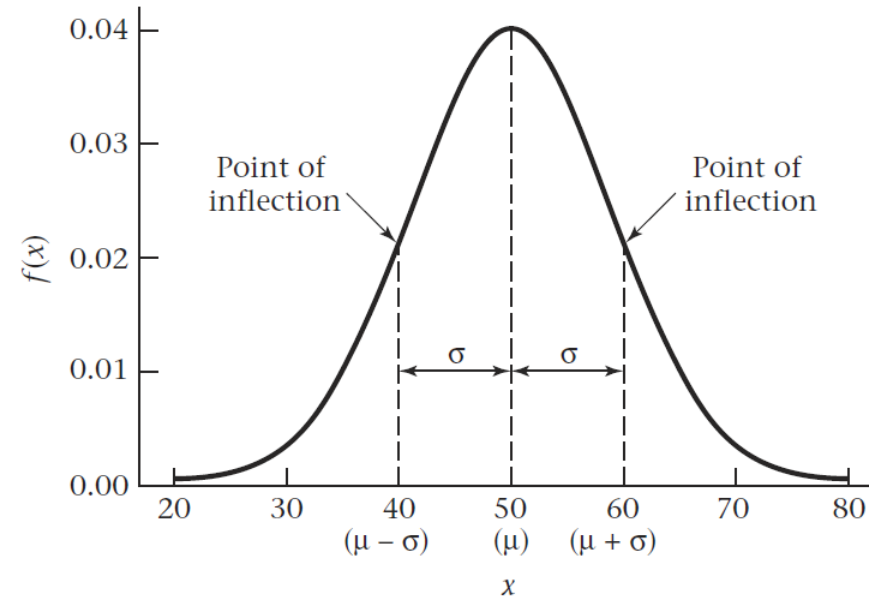
- Usually, any random variable that can be expressed as a sum of many other random variables can be well approximated by a normal distribution.
- Example., many physiologic measures are determined in part by a combination of several genetic and environmental risk factors can often be well approximated by a normal distribution.
- Most estimation procedures and hypothesis tests assume the random variable being considered has an underlying normal distribution.
- An important area of application of normal distribution is as an approximating distribution to other distributions. It is defined by its pdf, which is given as,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2} (x - \mu)^2 \right], \quad -\infty < x < \infty$$

for some parameters μ, σ , where $\sigma > 0$.

- exp is the power to which “e” (≈ 2.71828) is raised.

FIGURE 5.5 The pdf for a normal distribution with mean μ (50) and variance σ^2 (100)



- The density function follows a **bell-shaped curve**, with the mode at μ and the most frequently occurring values around μ . The curve is symmetric around μ , with points of inflection on either side of μ at $\mu - \sigma$ and $\mu + \sigma$.
- A **point of inflection** (拐点) is a point where its concavity (sign of second derivative) changes. Distance from μ to the points of inflection are an indicator of magnitude of σ .
- Using calculus methods, it can be shown that μ and σ^2 are the **expected value** and **variance**, respectively, of the distribution.

- A normal distribution with mean μ and variance σ^2 will generally be referred to as an **$N(\mu, \sigma^2)$ distribution**.
- The height of the normal distribution is always $1/(\sqrt{2\pi}\sigma)$.

FIGURE 5.6 Comparison of two normal distributions with the same variance and different means

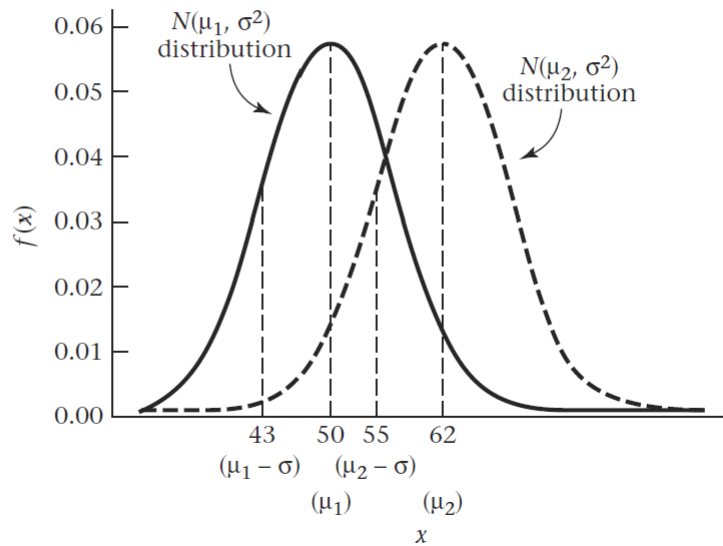
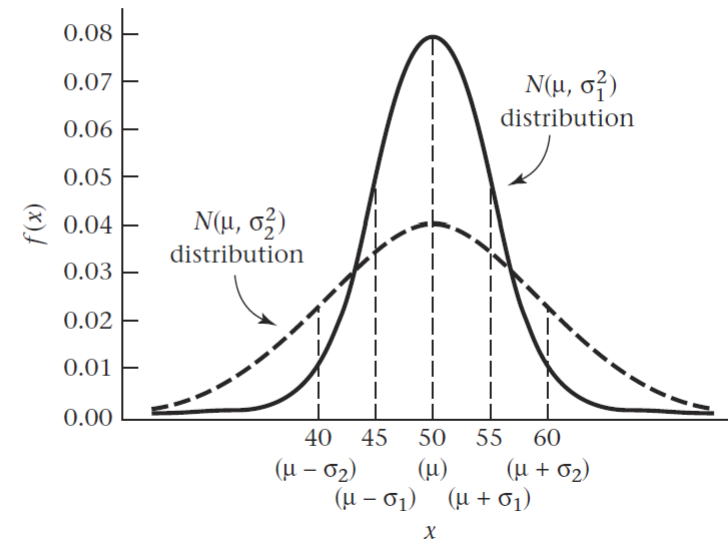


FIGURE 5.7 Comparison of two normal distributions with the same means and different variances



- A normal distribution with mean 0 and variance 1 is called a **standard normal distribution**, or unit normal distribution.
- This distribution is also called an **$N(0, 1)$ distribution**.

- $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < +\infty$
- This distribution is symmetric about 0, because $f(x) = f(-x)$.
- About 68% of the area under the standard normal density lies between +1 and -1, about 95% of the area lies between +2 and -2, and about 99% lies between +2.5 and -2.5.
- $\Pr(-1 < X < 1) = 0.6827$
- $\Pr(-1.96 < X < 1.96) = 0.95$
- $\Pr(-2.576 < X < 2.576) = 0.99$

FIGURE 5.8 The pdf for a standard normal distribution

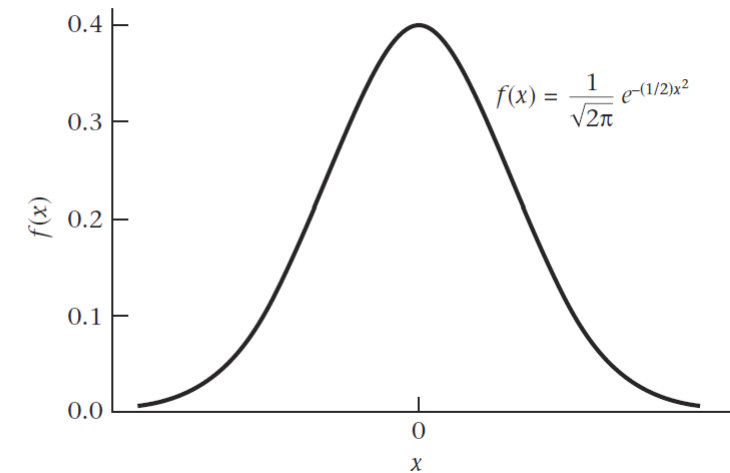
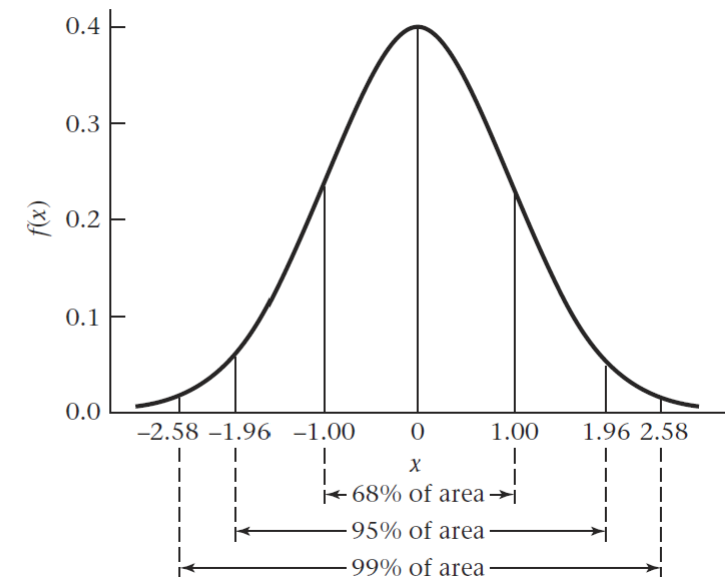


FIGURE 5.9 Empirical properties of the standard normal distribution



- The cumulative-distribution function (cdf) for a standard normal distribution is denoted by $\varphi(x) = Pr(X \leq x)$ where X follows an $N(0,1)$ distribution.
- The symbol $\sim$ is used as shorthand for the phrase “is distributed as.” Thus, $X \sim N(0,1)$ means that the random variable X is distributed as an $N(0,1)$ distribution.

FIGURE 5.10 The cdf $[\Phi(x)]$ for a standard normal distribution

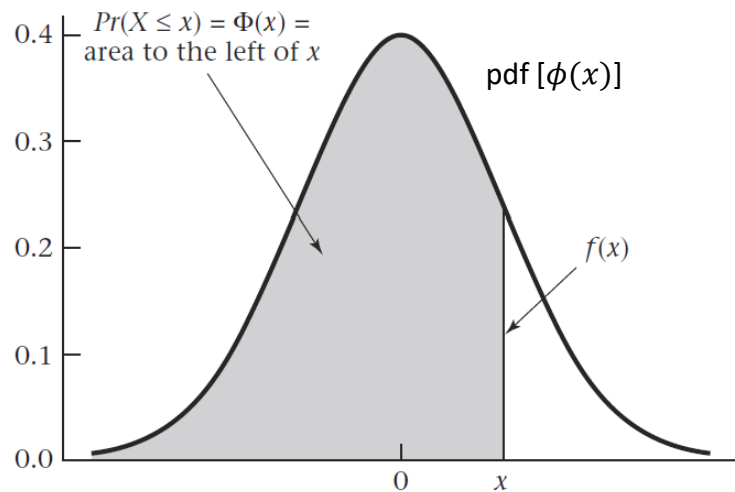
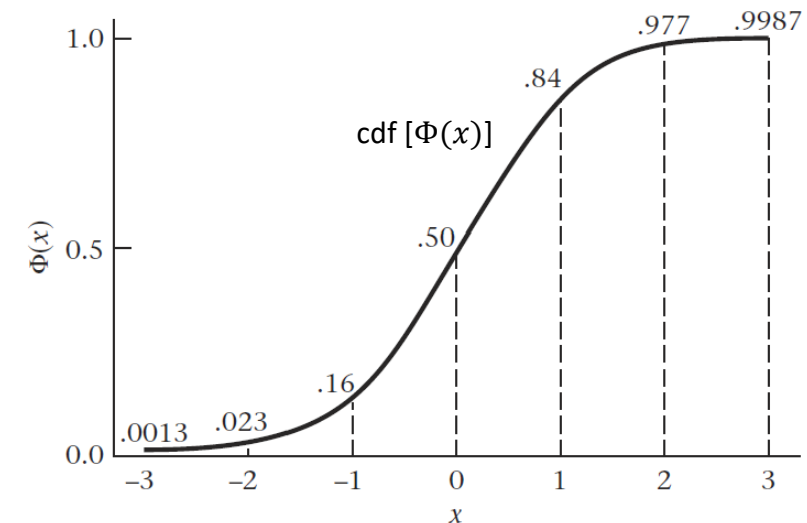


FIGURE 5.11 The cdf for a standard normal distribution $[\Phi(x)]$

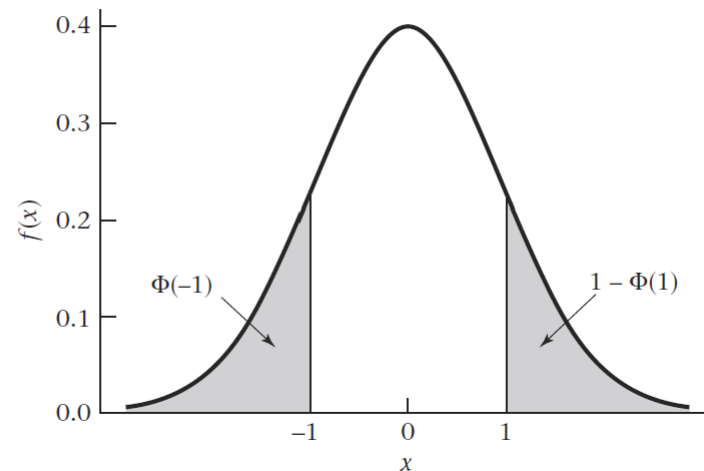


- The area to the left of x approaches 0 as x becomes small and approaches 1 as x becomes large.

Symmetry Properties of the Standard Normal Distribution

$$\Phi(-x) = \Pr(X \leq -x) = \Pr(X \geq x) = 1 - \Pr(X \leq x) = 1 - \Phi(x)$$

FIGURE 5.12 Illustration of the symmetry properties of the normal distribution

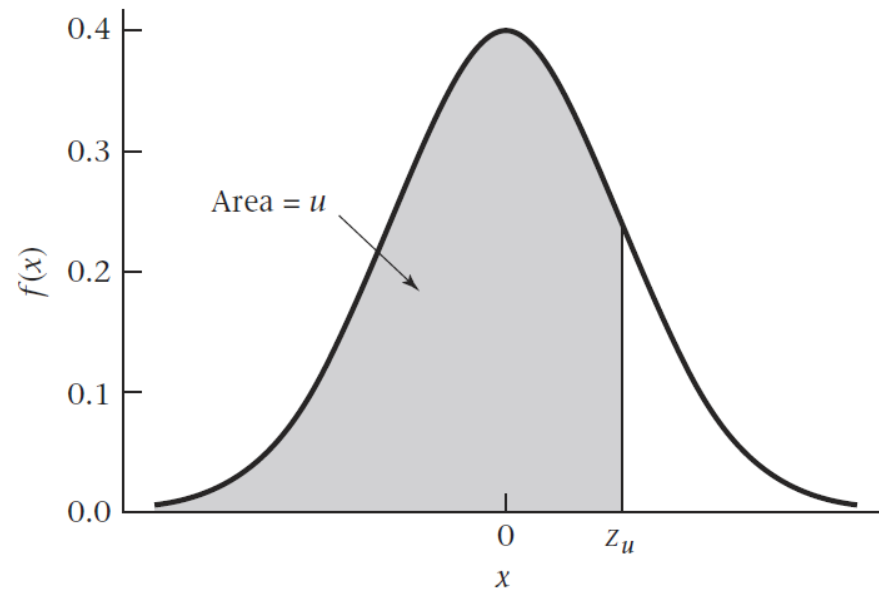


- A normal range for a biological quantity is often defined by a range within x standard deviations of the mean for some specified value of x .
- The probability of a value falling in this range is given by $\Pr(-x \leq X \leq x)$ for a standard normal distribution.

- The percentiles of a normal distribution are referred to in statistical inference. Example, the upper and lower fifth percentiles of the distribution may be referred to define a normal range of values.
- The $(100 \times u)$ th percentile of a standard normal distribution is denoted by z_u .

$$\Phi(z_u) = \Pr(X < z_u) = u \quad \text{where } X \sim N(0,1)$$

FIGURE 5.13 Graphic display of the $(100 \times u)$ th percentile of a standard normal distribution (z_u)



- The function z_u is sometimes referred to as the **inverse normal function**, or **quantile function**.

$$z_u = \Phi^{-1}(u) \quad \text{where } X \sim N(0,1)$$

- To evaluate z_u , we determine the area u in the normal tables and then find the value z_u that corresponds to this area.
- If $u < 0.5$, then we use the symmetry properties of the normal distribution to obtain

$$z_u = -z_{1-u}$$

where z_{1-u} can be obtained from the normal table.

- **Conversion from an $N(\mu, \sigma^2)$ Distribution to an $N(0, 1)$ Distribution**

- If $X \sim N(\mu, \sigma^2)$, then what is $\Pr(a < X < b)$ for any a, b ?
- Consider the random variable $Z = (X - \mu)/\sigma$.
- If $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$, then $Z \sim N(0,1)$.

- **Evaluation of Probabilities for Any Normal Distribution via Standardization**

- If $X \sim N(\mu, \sigma^2)$ and $Z = (X - \mu)/\sigma$
- This is known as **standardization of a normal variable**.

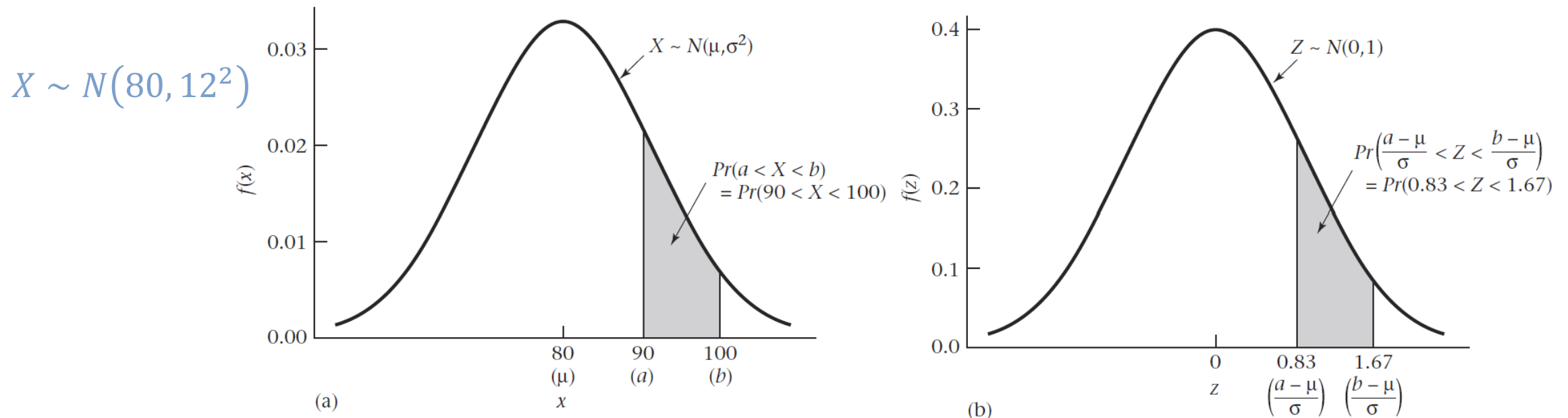
$$\begin{aligned}\Pr(a < X < b) &= \Pr\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)\end{aligned}$$

- For any probability expression concerning normal random variables of the form $\Pr(a < X < b)$, the population mean μ is subtracted from each boundary point and divided by the standard deviation σ to obtain an equivalent probability expression for the standard normal random variable Z ,

$$\Pr\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right)$$

- The standard normal tables are then used to evaluate this latter probability.

FIGURE 5.14 Evaluation of probabilities for any normal distribution using standardization



Linear Combinations of Random Variables

- Sums or difference or more complicated linear functions of random variables (either continuous or discrete) are often used.
- A linear combination L of the random variables $X_1, \dots, X_n$ is defined as any function of the form $L = c_1X_1 + \dots + c_nX_n$. A **linear combination** also called a **linear contrast** when $\sum c_i = 0$.
- To compute the expected value and variance of linear combinations of random variables, we use the principle that the expected value of the sum of n random variables is the sum of the n respective expected values.

$$E(L) = E(c_1X_1 + \dots + c_nX_n) = \sum_{i=1}^n c_i E(X_i), \quad L = \sum_{i=1}^n c_i X_i$$

- Variance of $L = \sum_{i=1}^n c_i X_i$, where $X_1, \dots, X_n$ are **independent** is

$$\text{Var}(L) = \sum_{i=1}^n c_i^2 \text{Var}(X_i)$$

- If $X_1, \dots, X_n$ are **independent normal random variables** with expected values $\mu_1, \dots, \mu_n$ and variances $\sigma_1^2, \dots, \sigma_n^2$, and L is any linear combination

$$L = \sum_{i=1}^n c_i X_i$$

then L is normally distributed with

- Expected value:

$$E(L) = \sum_{i=1}^n c_i \mu_i$$

- and

- Variance:

$$\text{Var}(L) = \sum_{i=1}^n c_i^2 \sigma_i^2$$

Dependent Random Variables

- The **covariance** between two random variables X and Y is denoted by $\text{Cov}(X, Y)$ and is defined by

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

- which can also be written as $E(XY) - \mu_x\mu_y$, where μ_x is the average value of X , μ_y is the average value of Y , and $E(XY)$ = average value of the product of X and Y .
- If X and Y are independent, then the covariance between them is 0.
- If large values of X and Y occur among the same subjects (as well as small values of X and Y), then the covariance is positive.
- If large values of X and small values of Y (or conversely, small values of X and large values of Y) tend to occur among the same subjects, then the covariance is negative.

- To obtain a measure of relatedness or association between two random variables X and Y , we consider the (population) **correlation coefficient**, denoted by $\text{Corr}(X, Y)$ or ρ and is defined by

$$\rho = \rho_{X,Y} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

- where σ_x and σ_y are the standard deviations of X and Y , resp.
- It is a **dimensionless** quantity that is independent of the units of X and Y and ranges between -1 and 1.
- A correlation coefficient of 0 indicates no linear correlation.
- If X and Y are jointly normally distributed, a correlation coefficient of 0 implies independence.
- Corr close to 1 implies nearly perfect positive dependence with large values of X corresponding to large values of Y and small values of X corresponding to small values of Y .
- Corr close to -1 implies $\approx$ perfect negative dependence, with large values of X corresponding to small values of Y and vice versa.

- To compute the variance of a linear contrast involving two dependent random variables X_1 and X_2 , we can use the general equation

$$\text{Var}(c_1X_1 + c_2X_2) = c_1^2 \text{Var}(X_1) + c_2^2 \text{Var}(X_2) + 2c_1c_2 \text{Cov}(X_1, X_2)$$

$$= c_1^2 \text{Var}(X_1) + c_2^2 \text{Var}(X_2) + 2c_1c_2\sigma_1\sigma_2 \text{Corr}(X_1, X_2)$$

- Variance of linear combination of random variables (general case)
- The variance of $L = \sum_{i=1}^n c_i X_i$ (the linear contrast)

$$\text{Var}(L) = \sum_{i=1}^n c_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=1, i < j}^n c_i c_j \text{Cov}(X_i, X_j)$$

$$= \sum_{i=1}^n c_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j=1, i < j}^n c_i c_j \sigma_i \sigma_j \text{Corr}(X_i, X_j)$$

- If $X_1, \dots, X_n$ are **normal random variables** with expected values $\mu_1, \dots, \mu_n$ and variances $\sigma_1^2, \dots, \sigma_n^2$ and L is any linear contrast

$$L = \sum_{i=1}^n c_i X_i$$

- then L is normally distributed with expected value

$$E(L) = \sum_{i=1}^n c_i \mu_i$$

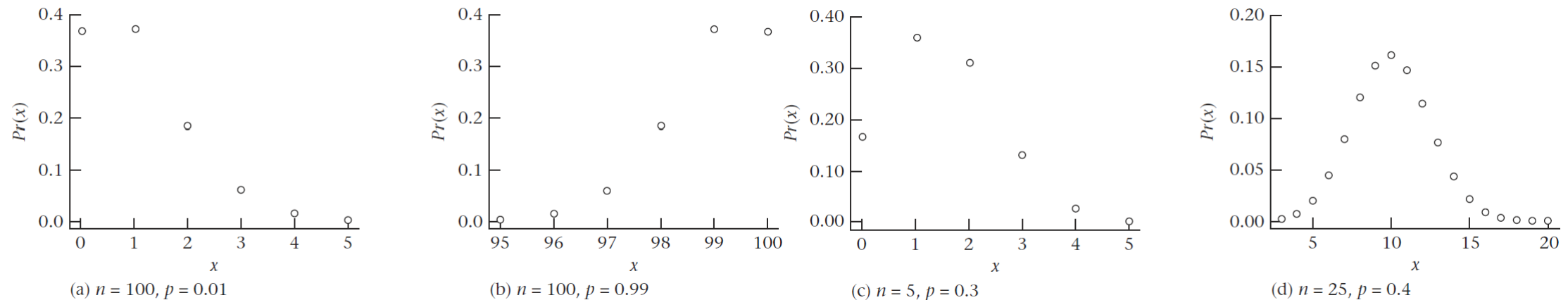
- and variance given by

$$\text{Var}(L) = \sum_{i=1}^n c_i^2 \sigma_i^2 + 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ i < j}}^n c_i c_j \sigma_i \sigma_j \rho_{ij}$$

Normal Approximation to the Binomial Distribution

- If n is large, the binomial distribution is difficult to work with and an approximation is easier to use rather than the exact binomial distribution.
- If n is moderately large and p is either near 0 or 1, then the binomial distribution will be very positively or negatively skewed, resp.
- If n is moderately large and p is not too extreme, then the binomial distribution tends to be symmetric and is well approximated by a normal distribution.

FIGURE 5.16 Symmetry properties of the binomial distribution



- If X is a binomial random variable with parameters n and p , then $\Pr(a \leq X_1 < b)$ is approximated by the area under an $N(np, npq)$ curve from $a - \frac{1}{2}$ to $b + \frac{1}{2}$.
- For the special case $a = b$, the binomial probability $\Pr(X = a)$ is approximated by the area under the normal curve from $a - \frac{1}{2}$ to $a + \frac{1}{2}$.
- Exception: $\Pr(X = 0)$ and $\Pr(X_1 = n)$ are approximated by the area under the normal curve to the left of $\frac{1}{2}$ and to the right of $n - \frac{1}{2}$, resp.
- If $c_1 = \dots = c_n = 1$, then a sum of normal random variables

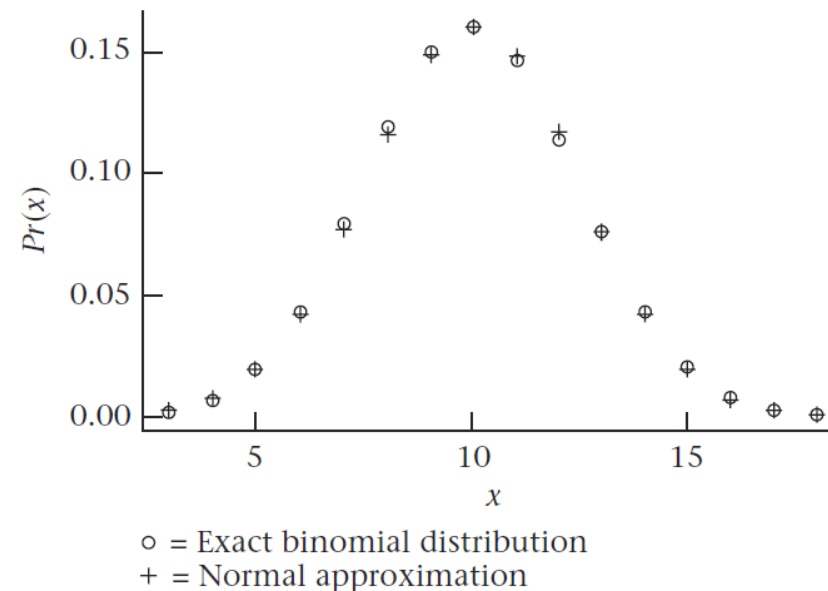
$$L = \sum_{i=1}^n X_i$$

is normally distributed.

- The normal approximation to the binomial distribution is a special case of the **central-limit theorem**. For large n , a sum of n random variables is approximately normally distributed even if the individual random variables being summed are not themselves normal.

- Let X_i be a random variable that takes on the value 1 with probability p and the value 0 with probability $q = 1 - p$. This random variable is referred to as **Bernoulli trial**. This is a special case of a random variable with $n = 1$.

FIGURE 5.17 The approximation of the binomial random variable X with parameters $n = 25$, $p = .4$ by the normal random variable Y with mean = 10 and variance = 6



- The normal distribution with mean np and variance npq can be used to approximate a binomial distribution with parameters n and p when $npq \geq 5$.

FIGURE 5.18 Normal approximation to the distribution of neutrophils

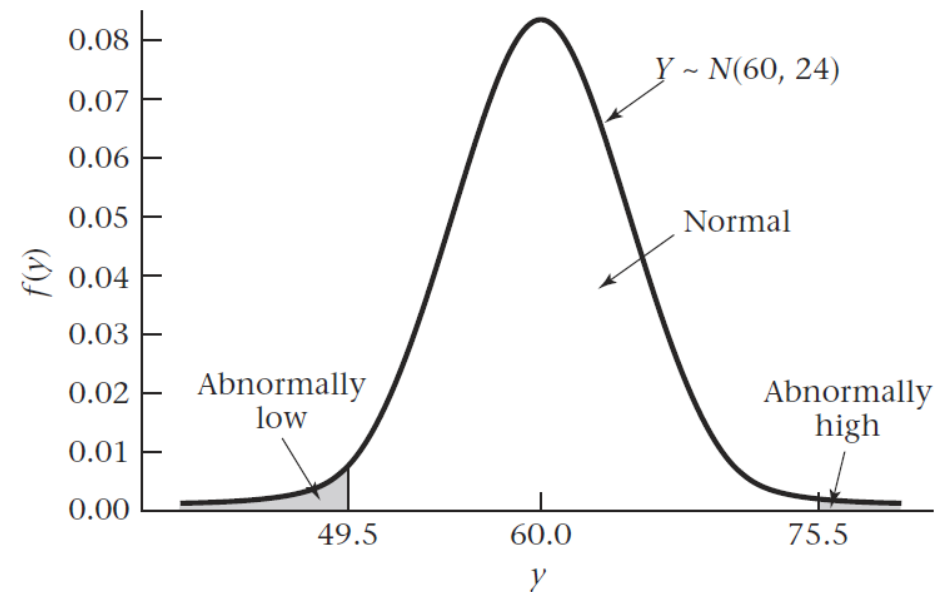
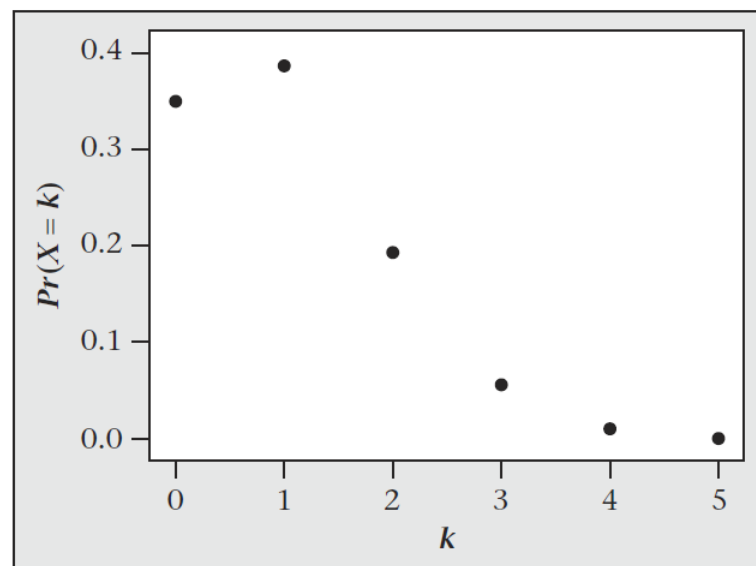
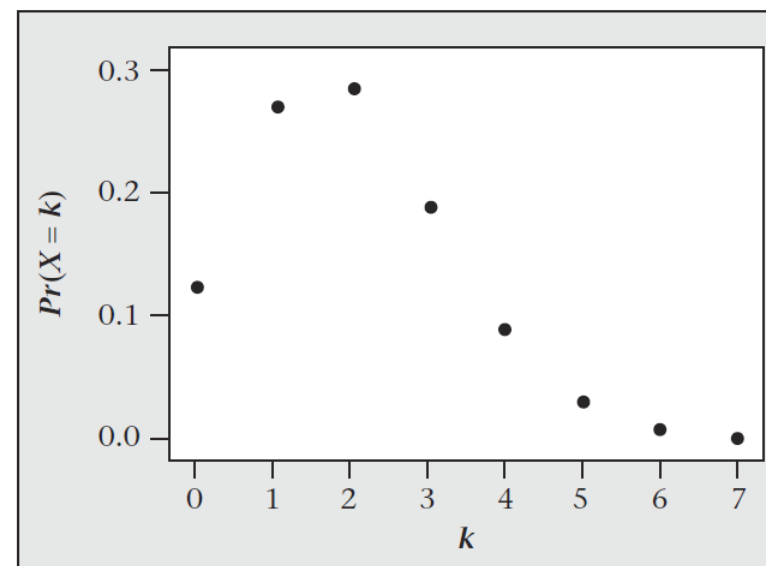


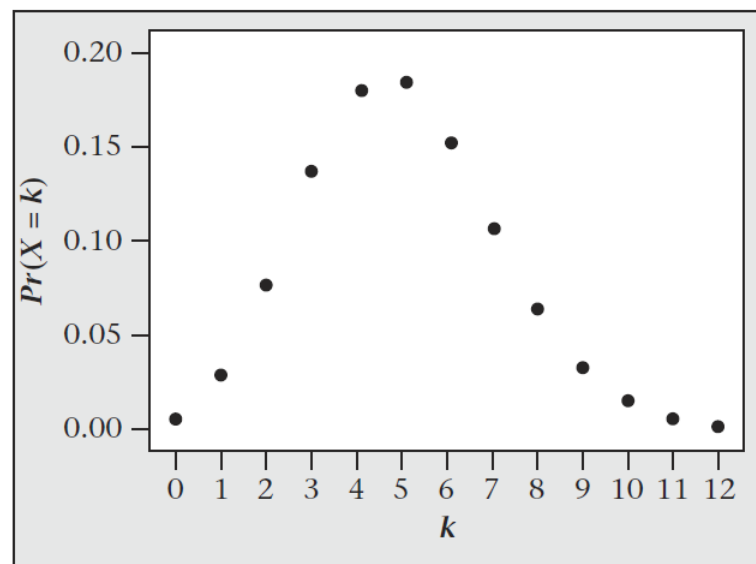
FIGURE 5.19 R plot of binomial distribution, $n = 10, 20, 50, 100, p = .1$



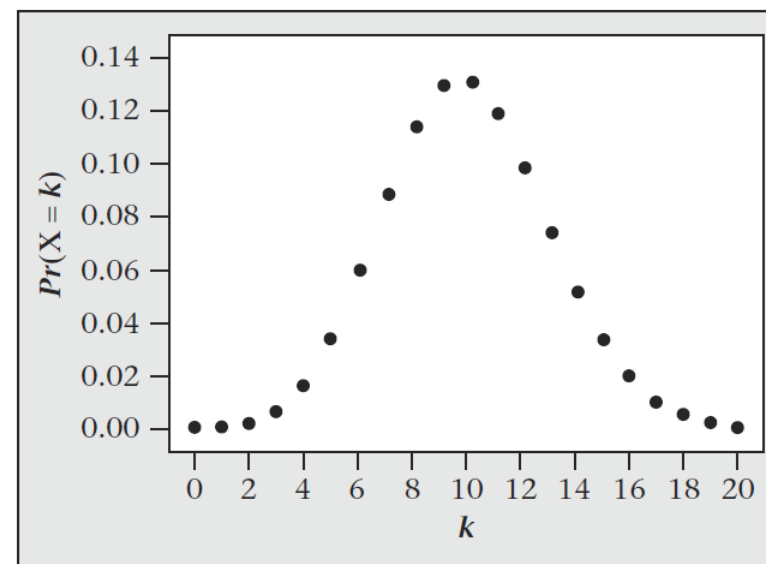
(a) $n = 10, p = 0.1$



(b) $n = 20, p = 0.1$



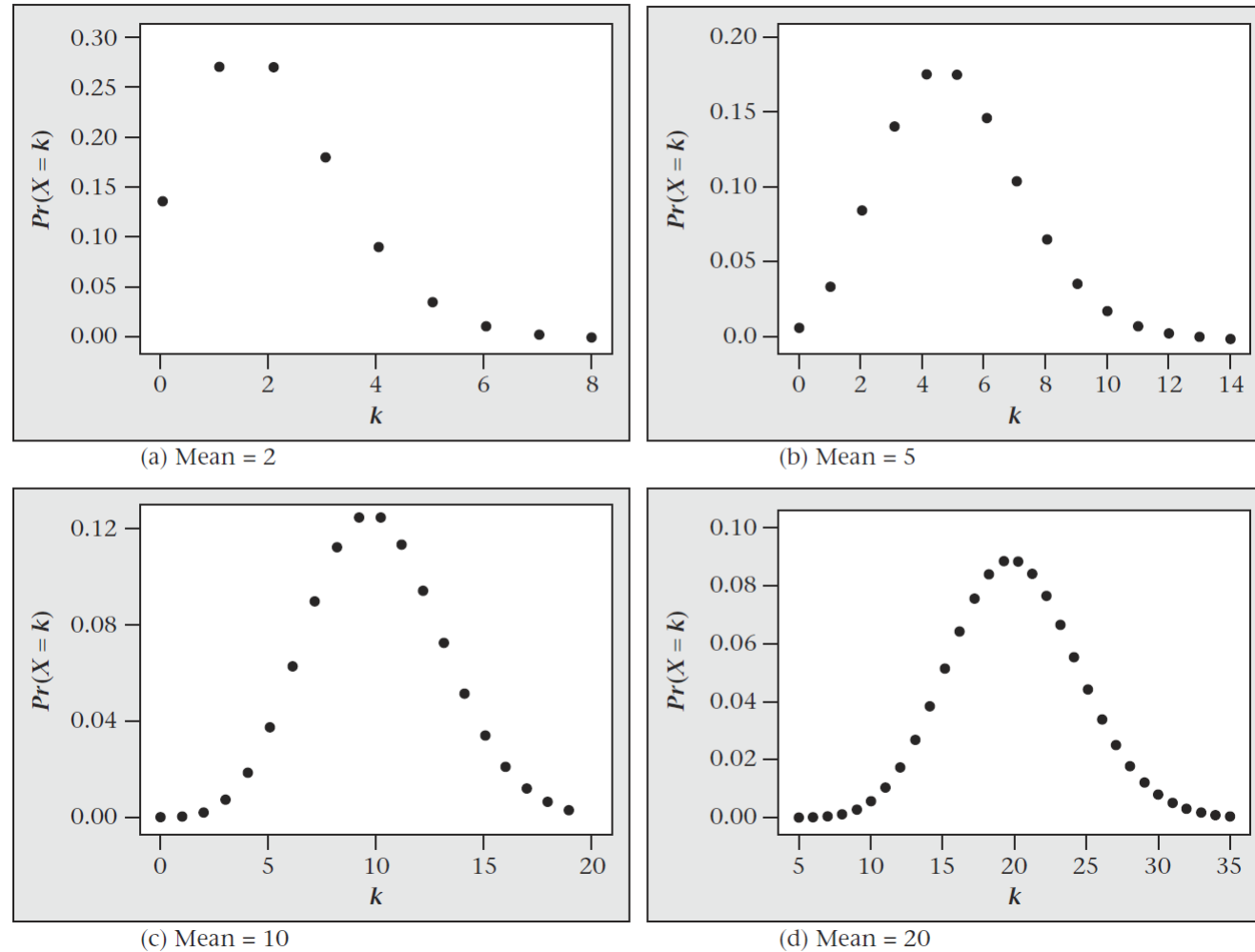
(c) $n = 50, p = 0.1$



(d) $n = 100, p = 0.1$

- The normal approximation to the binomial distribution works well (namely, $npq \geq 5$), which corresponds to n moderate and p not too large or too small, are generally not the same as the conditions for which the Poisson approximation to the binomial distribution works well.
- The normal distribution can also be used to approximate discrete distributions other than the binomial distribution, particularly the **Poisson distribution**, which is cumbersome to use for large values of μ .
- A Poisson distribution with parameter μ is approximated by a normal distribution with mean and variance both equal to μ .
- $Pr(X = x)$ is approximated by the area under an $N(\mu, \mu)$ density from $x - \frac{1}{2}$ to $x + \frac{1}{2}$ for $x > 0$ or by the area to the left of $\frac{1}{2}$ for $x = 0$. This approximation is used when $\mu \geq 10$.

FIGURE 5.21 R plot of Poisson distribution, $\mu = 2, 5, 10, 20$



- The normal approximation is clearly inadequate for $\mu = 2$, marginally inadequate for $\mu = 5$, and adequate for $\mu = 10$, and $\mu = 20$.