In this lecture, we will discuss the topics in Chapter 4 and 5:

- Discrete and continuous random variables
- Probability-mass function (probability distribution)
- CDF, expected value, and variance
- Permutations and combinations
- Binomial distribution
- Poisson distribution
- Normal distribution
- Linear combinations of random variables
- Normal approximation to binomial and poisson distributions
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

**Probability-mass function (PMF),** sometimes also called a **probability distribution,** is a mathematical relationship, or rule, such 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, namely, \( 0 < \Pr(X = r) \leq 1 \), \( \sum \Pr(X = r) = 1 \).

<table>
<thead>
<tr>
<th>( r )</th>
<th>( \Pr(X = r) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.008</td>
</tr>
<tr>
<td>1</td>
<td>.076</td>
</tr>
<tr>
<td>2</td>
<td>.265</td>
</tr>
<tr>
<td>3</td>
<td>.411</td>
</tr>
<tr>
<td>4</td>
<td>.240</td>
</tr>
</tbody>
</table>

**Table 4.1**  
*Probability-mass function for the hypertension-control example*

X = the number of patients of four who are brought under control on a trial of a new antihypertensive drug
Relationship of Probability Distributions to Frequency Distributions

Frequency distribution \((f(x))\): is described as a list of each value in the data set and a corresponding count of how frequently the value occurs.

\[
f(x)/\text{total # of sample points} = \text{a sample analog to the PMF}
\]

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**.

PMF derived from the binomial distribution is compared with the frequency distribution to determine whether the drug behaves with the same efficacy as predicted.
Expected Value and Variance of a Discrete Random Variable

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 is called the expected value of a random variable, or population mean, and is denoted by $E(X)$ or $\mu$ and represents the “average” value of the random variable.

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

- 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 $\text{Var}(X)$ or $\sigma^2$. Approximately 95% of the probability mass falls within two standard deviations ($2\sigma$) of the mean of a random variable.

$$\text{Var}(X) = \sigma^2 = \sum_{i=1}^{R} (x_i - \mu)^2 \Pr(X = x_i)$$

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

$$\sigma^2 = E(X - \mu)^2 = \sum_{i=1}^{R} x_i^2 \Pr(X = x_i) - \mu^2$$
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.

- 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

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

$$nP_k = n(n-1) \times \cdots \times (n-k+1)$$

representing the number of ways of selecting $k$ items of $n$, where the order of selection is important.

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.

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.

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

Expressing permutations in terms of factorials

$$nP_k = n(n-1) \times \cdots \times (n-k+1)$$

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

$$= n! / (n-k)!$$
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:

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

OR

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

For any non-negative integers, \( n \), \( k \) where \( n \geq k \),

\[
\binom{n}{k} = \binom{n}{n-k}
\]

OR

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

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 a failure at each trial is $1 - p = q$.

**Table 4.5**

<table>
<thead>
<tr>
<th>Possible orderings for 2 neutrophils of 5 cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>$xxoox$</td>
</tr>
<tr>
<td>$xoxoo$</td>
</tr>
<tr>
<td>$xooxo$</td>
</tr>
<tr>
<td>$xooox$</td>
</tr>
<tr>
<td>$ooxox$</td>
</tr>
</tbody>
</table>

If a neutrophil is denoted by an $x$ and a non-neutrophil by an $o$. What is the probability of $oxoox = \Pr(oxoox)$?

That is, the probability of $k$ successes within $n$ trials

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

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

$$Pr(X = k) = \binom{n}{k} p^k q^{n-k} = \binom{n}{n-k} q^{n-k} p^k = Pr(Y = n-k)$$
Some Examples of Binomial Distribution

Figure 4.4  The binomial distribution for various values of $p$ when $n = 10$

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

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

(c) $n = 10, p = .50$
For a binomial distribution, the only values that take on positive probability are 0, 1, 2,..., \( n \) and these occur with probabilities

\[
\binom{n}{0} p^0 q^n, \binom{n}{1} p^1 q^{n-1}, \ldots
\]

Thus, \( E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} \) the expected value = \( np \).

Similarly, variance is

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

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.
Poisson Distribution

The second most frequently used discrete distribution after binomial distribution. It is usually associated with rare events. With the following assumptions:

1) Pr(1 event) over \( \Delta t \) ≈ \( \lambda \Delta t \) for some constant \( \lambda \); Pr(0 event) over \( \Delta t \) ≈ 1 − \( \lambda \Delta t \); Pr(> 1 event) = 0.

2) Stationarity: the number of events per unit time is the same throughout the entire time interval \( t \).

3) Independence: If an event occurs within one time subinterval, then it has no bearing on the probability of event in the next time subinterval.

The probability of \( k \) events occurring in a time period \( t \) for a Poisson random variable with parameter \( \lambda \) is

\[
Pr(X = k) = e^{-\mu} \mu^k / k!, \quad k = 0, 1, 2, \ldots
\]

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

\( \lambda \) = expected no. of events /unit time;
\( \mu \) = expected no. 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.
Some Examples of Poisson Distribution

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 $\mu$ increases.

Figure 4.5  Distribution of the number of deaths attributable to typhoid fever over various time intervals

(a) 6 months

(b) 3 months
Expected Value and Variance of the Poisson Distribution

For a Poisson distribution with parameter \( \mu \), **the mean and variance are both equal to \( \mu \).**

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>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of deaths</td>
<td>15</td>
<td>10</td>
<td>19</td>
<td>23</td>
<td>15</td>
<td>17</td>
<td>23</td>
<td>17</td>
<td>26</td>
<td>15</td>
</tr>
</tbody>
</table>

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 $\mu = np$. The mean of this distribution is given by $np$ and the variance by $npq$.

Note that $q \approx 1$ 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>
<thead>
<tr>
<th>$k$</th>
<th>Exact binomial probability</th>
<th>Poisson approximation</th>
<th>$k$</th>
<th>Exact binomial probability</th>
<th>Poisson approximation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.366</td>
<td>.368</td>
<td>3</td>
<td>.061</td>
<td>.061</td>
</tr>
<tr>
<td>1</td>
<td>.370</td>
<td>.368</td>
<td>4</td>
<td>.015</td>
<td>.015</td>
</tr>
<tr>
<td>2</td>
<td>.185</td>
<td>.184</td>
<td>5</td>
<td>.003</td>
<td>.003</td>
</tr>
</tbody>
</table>
Summary of Discrete Probability Distribution

In chapter 4, 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 ($\overline{x}$) 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.
The probability-density function of the random variable $X$ is a function such that the area under the density-function curve between any two points $a$ and $b$ is equal to the probability that the random variable $X$ falls between, $a$ and $b$. Thus, the total area under the density-function curve over the entire range of possible values for the random variable is 1.

Areas $A$, $B$, $C$ ≈ probabilities of being mildly, moderately, and severely hypertensive. $Pr(X = x) = 0$.

The pdf has large values in regions of high probability and small values in regions of low probability A.

Expected value and variance for continuous random variables have the same meaning as for discrete random variables. **Expected value:** $E(X)$, or $\mu$

**Variance:** $\text{Var}(X)$ or $\sigma^2$, $E(X - \mu)^2$, $E(X^2) - \mu^2$.

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

- 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.

- 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.
The density function follows a bell-shaped curve, with the mode at $\mu$ and the most frequently occurring values around $\mu$. The curve is symmetric around $\mu$, with points of inflection on either side of $\mu$ at $\mu - \sigma$ and $\mu + \sigma$.

A point of inflection is a point at which the slope of the curve changes direction. Distance from $\mu$ to the points of inflection are an indicator of magnitude of $\sigma$.

Using calculus methods, it can be shown that $\mu$ and $\sigma^2$ are the expected value and variance, respectively, of the distribution.
Normal Distribution : Properties

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

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

\[
\begin{align*}
\Pr(-1 < X < 1) &= 0.6827 \\
\Pr(-1.96 < X < 1.96) &= 0.95 \\
\Pr(-2.576 < X < 2.576) &= 0.99
\end{align*}
\]
The cumulative-distribution function (cdf) for a standard normal distribution is denoted by $\Phi(x) = \Pr(X \leq x)$ 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.

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

\[ \phi(-x) = \Pr(X \leq -x) = \Pr(X \geq x) = 1 - \Pr(X \leq x) = 1 - \phi(x) \]

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.
Percentiles for the 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 $z_u$ satisfying

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

The function $z_u$ is sometimes referred to as the inverse normal function.

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.

\[
\Pr(a < X < b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)
\]

Figure 5.14 Evaluation of probabilities for any normal distribution using standardization

9/13/2016
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 $X_1, \ldots, X_n$ is defined as any function of the form $L = c_1X_1 + \ldots + c_nX_n$. A linear combination also called a linear contrast.

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.

$$L = \sum_{i=1}^{n} c_iX_i, \quad \mathbb{E}(L) = \sum_{i=1}^{n} c_i\mathbb{E}(X_i) = \sum_{i=1}^{n} c_i\mu_i$$

Variance of $L$ where $X_1, \ldots, X_n$ are independent is

$$L = \sum_{i=1}^{n} c_iX_i, \quad \text{Var}(L) = \sum_{i=1}^{n} c_i^2\text{Var}(X_i) = \sum_{i=1}^{n} c_i^2\sigma_i^2$$
The covariance between two random variables $X$ and $Y$ is denoted by $Cov(X,Y)$ and is defined by

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

which can also be written as $E(XY) - \mu_x \mu_y$, where $\mu_x$ is the average value of $X$, $\mu_y$ 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 correlation coefficient, denoted by $Corr(X,Y)$ or $\rho$ and is defined by

$$\rho = Corr(X,Y) = Cov(X,Y)/(\sigma_x \sigma_y)$$

where $\sigma_x$ and $\sigma_y$ 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.
If $X$ and $Y$ are approx. linearly related, 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:

$$Var(c_1 X_1 + c_2 X_2) = c_1^2 Var(X_1) + c_2^2 Var(X_2) + 2c_1c_2 Cov(X_1, X_2)$$

$$= c_1^2 Var(X_1) + c_2^2 Var(X_2) + 2c_1c_2 \sigma_x \sigma_y Corr(X_1, X_2)$$

Variance of linear combination of random variables (general case)
The variance of the linear contrast:

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

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

$$= \sum_{i=1}^{n} c_i^2 Var(X_i) + 2 \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \sigma_i \sigma_j Corr(X_i, X_j)$$
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. \( N(np, npq) \) can be used to approximate a binomial distribution with parameters \( n \) and \( p \) when \( npq \geq 5 \).
- If \( X \) is a binomial random variable with parameters \( n \) and \( p \), then \( \Pr(a \leq X < b) \) is approximated by the area under an \( N(np, npq) \) curve from \( a - \frac{1}{2} \) to \( b + \frac{1}{2} \).
- 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.
Normal Approximation to the Poisson Distribution

- 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 $\mu$.
- A Poisson distribution with parameter $\mu$ is approximated by a normal distribution with mean and variance both equal to $\mu$.

- $\Pr(X = x)$ 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 $\mu \geq 10$.

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

In chapter 5, we discussed

- Continuous random variables
- Probability-density function, which is the analog to a probability-mass function for discrete random variables
- Concepts of expected value, variance, and cumulative distribution for continuous random variables
- Normal distribution, which is the most important continuous distribution
- The two parameters: mean $\mu$ and variance $\sigma^2$
- Normal tables, which are used when working with standard normal distribution
- Electronic tables can be used to evaluate areas and/or percentiles for any normal distribution
- Properties of linear combinations of random variables for independent and dependent random variables
The End