4. CONTINUOUS RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

§ 4.1 Continuous Random Variables

Recall that for a discrete random variable, the set of all possible outcomes has to be discrete in the sense that the elements of the set can be listed sequentially beginning with the smallest number. On the other hand, if the set of all possible values of a random variable is an entire interval of numbers, then the random variable is said to be continuous. In this case, between any of the two possible values we can find infinitely many possible values. This implies that the values cannot be listed sequentially.

Example: For instance, let \( X = \) length of a randomly chosen rattle snake. Suppose the smallest rattle snake is of length \( a \) and the longest rattle snake is \( b \) units long. Then the set of all possible values lie between \( a \) and \( b \). We then write \( a \leq x \leq b \) as the range of possible values.

Example: Let \( X = \) reaction temperature (in °C) in a certain chemical process. Suppose the minimum temperature is \(-5^\circ C\) and maximum temperature for the process is \(5^\circ C\). Then the set of all possible values is \( \{ x : -5 \leq x \leq 5 \} \).

It is possible to argue that restrictions of our measuring instruments may sometimes limit us to a discrete world. For instance, our measurements of temperature may sometimes make us feel that temperature is discrete. However, continuous random variables and distributions often approximate real-world situations very well. Furthermore, problems based on continuous variables are often easier to solve than those based on discrete variables.

§ 4.2 Probability Distributions For Continuous Random Variables
In the discrete case, the probability distribution of a random variable called the probability mass function (pmf) shows us how the total probability of 1 is distributed among the possible values of the random variable. The probability distribution of a continuous random variable is called the probability density function.

Now, let $Y$ be a continuous random variable. Then a probability density function (pdf) of $Y$ is a function $f(y)$ such that for any two numbers $a$ and $b$, with $a \leq b$,

$$P(a \leq Y \leq b) = \int_a^b f(y)\,dy.$$ 

That is, if we plot the graph of $f(y)$ or the density curve, the probability that $Y$ takes a value between $a$ and $b$ is the area between these two numbers and under the graph of $f(y)$. That means, probabilities involving continuous random variables is the same as area under the curve. It therefore makes sense that for any constant, say $c$,

$$P(Y = c) = 0.$$ 

That is, the area under the graph at a single point is zero. This implies that if $Y$ is a continuous random variable, for any two numbers $a$ and $b$, with $a \leq b$,

$$P(a \leq Y \leq b) = P(a < Y \leq b) = P(a \leq Y < b) = P(a < Y < b).$$ 

We wish to emphasize the point that the equality of the four probabilities above holds for only continuous random variables.

For any function $f(y)$ to be a legitimate probability density function, it must satisfy

1. $f(y) \geq 0$, for all $y$.

2. $\int_{-\infty}^{\infty} f(y)\,dy = \text{total area under any density curve} = 1.$

**Example: Problem 5, Page 151** A college professor never finishes his lecture before the end of the hour and always finishes his lectures within 2 min after the hour. Let $X = $ the time that elapses between the end of the hour and the end of the lecture and suppose the pdf of $X$ is

$$f(x) = \begin{cases} kx^2 & 0 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the value of $k$. 

2
(b) What is the probability that the lecture ends within 1 min of the end of the hour?

Solution:

In our discussion on discrete random variables we studied some random variables with special mass functions such as binomial, hypergeometric and Poisson. The first random variable with a special density function we shall discuss is the uniformly distributed random variable. By definition, a continuous random variable $X$ is said to have a **uniform distribution** on the interval $[A, B]$ if the pdf of $X$ is

$$f(x; A, B) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \\ 0 & \text{otherwise.} \end{cases}$$

This is equivalent to the notion of equally likely outcomes of a sample space.

**Example: Problem 7, Page 151** The time $X$ (min) for a lab assistant to prepare the equipment for a certain experiment is believed to have a uniform distribution with parameters 25 and 35.

(a) Write the pdf of $X$ and sketch its graph.

(b) For any $a$ such that $25 < a < a + 2 < 35$, what is the probability that preparation time is between $a$ and $a + 2$ min?

Solution:

§4.3 Cumulative Distribution Functions and Expected Values

The cumulative distribution function (cdf) $F(x)$ of a continuous random variable, for every number $x$, is the entire area to the left of the point $x$ under the density curve $f(x)$ of $X$. That is,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy.$$ 

Notice that the variable in the integrand has been changed to $y$ in order not to confuse that with the upper limit of the integral. For the purpose of illustration, let us determine the cdf of the uniform random variable $X$ in Problem 7, Page 151. Now, for this problem,

$$f(x; 25, 35) = \begin{cases} \frac{1}{35-25} & 25 \leq x \leq 35 \\ 0 & \text{otherwise.} \end{cases}$$

It follows that for every $x$,

$$F(x) = \int_{-\infty}^{x} f(y)dy = \int_{25}^{x} \frac{1}{10}dy = \frac{x - 25}{10}.$$
Therefore the cdf of $X =$ time for a lab assistant to prepare equipment is

$$F(x) = \begin{cases} 
0 & x < 25 \\
\frac{x-25}{10} & 25 \leq x < 35 \\
1 & x \geq 35.
\end{cases}$$

The idea of complements can also be used to compute probabilities when dealing with continuous random variables. If $X$ is a continuous random variable and $a$ is any constant, then

$$P(X \geq a) = P(X > a) = 1 - P(X \leq a) = 1 - F(a).$$

Some students may have noticed that this is different from the definition for discrete random variables. Recall that if $Y$ is a discrete random variable and $a$ is any integer, then $P(Y \geq a) = 1 - P(Y < a) = 1 - F(a-1)$. Students should therefore take note of this difference between discrete and continuous random variable. Also, we note that if $X$ is a continuous random variable and $a$ and $b$ are constants, such that $a \leq b$, then

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

Again, this is slightly different from what we would have done if $X$ were to be discrete.

**Percentiles of a Continuous Distribution**

Now, we have seen that given a value, say $x = 4$ and the density function of the continuous random variable $X$, we can compute the cdf of $X$ at the point $x = 4$. For instance, let

$$f(x) = \begin{cases} 
0.5x & 0 \leq x \leq 2 \\
0 & \text{otherwise}.
\end{cases}$$

Then, it is easy to verify that

$$F(x) = \begin{cases} 
0 & x < 0 \\
0.25x^2 & 0 \leq x < 2 \\
1 & x \geq 2.
\end{cases}$$

Thus, we can see that at $x = 3$, $F(3) = 1$ and $F(-0.2) = 0$. Similarly, at $x = 1.5$, $F(1.5) = 0.25 \times (1.5)^2 = 0.375$. Now, the question is, suppose we know that $F(x) = 0.375$ and we are given the density function $f(x)$, is it possible to find the corresponding value of $x$, say $x = c$? In this case, we can see that the value of $x$ that gives $F(x) = 0.375$ is $x = 1.5$. In statistics, we call the point $x = 1.5$, the $(100 \times 0.375)$th **percentile** of the distribution of the continuous random variable $X$. Students may have observed that for both discrete and continuous random variables, the smallest value of the cdf is zero and the largest value is 1. A general definition of the percentile of a distribution is as follows.
**Definition:** Let $p$ be a number between 0 and 1. The $(100p)$th percentile of the distribution of a continuous random variable $X$, we shall denote by $c$, is that value for which

$$p = F(c) = \int_{-\infty}^{c} f(y)dy.$$  

Important cases of $p$ are as follows.

(a) $p = 0.5$ gives the 50th percentile which is also the median of the distribution of $X$. The median is the point on the $x$ axis which divides the area under the density curve of $X$ into two equal parts. In the course text, the median is denoted by $\bar{\mu}$.

(b) $p = 0.25$ gives the 25th percentile which is also the first quartile of the distribution of $X$. The first quartile is the point on the $x$ axis which divides the area under the density curve of $X$ into two parts such that the area to the left is 25% of the total area and the area to the right is 75% of the total area.

(c) $p = 0.75$ gives the 75th percentile which is also the third quartile of the distribution of $X$. The third quartile is the point on the $x$ axis which divides the area under the density curve of $X$ into two parts such that the area to the left is 75% of the total area and the area to the right is 25% of the total area.

As an example, suppose we wish to find the median of the distribution of the random variable $X$ with density function given by

$$f(x) = \begin{cases} 0.5x & 0 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$  

The first step is to compute $F(x)$. From previous results, we have

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.25x^2 & 0 \leq x < 2 \\ 1 & x \geq 2. \end{cases}$$  

Now, we are seeking the value of $x$ for which $F(x) = 50\% = 0.5$. Thus, we solve the equation

$$F(x) = 0.25x^2 = 0.5,$$

for $x$. We obtain $x = \pm\sqrt{\frac{0.5}{0.25}} = \pm\sqrt{2}$. Since the negative value falls outside the acceptable limits, the median of the distribution of $X$ is $x = \sqrt{2}$.

**Examples and Practice Problems** (see class notes for solutions)
Problem 1, Page 150

Let $X$ denote the amount of time for which a book on a 2-hour reserve at a college library is checked out by a randomly selected student and suppose that $X$ has density function

$$f(x) = \begin{cases} 
0.5x & 0 \leq x \leq 2 \\
0 & \text{otherwise}
\end{cases}$$

Calculate the following probabilities

(a) $P(X \leq 1)$
(b) $P(0.5 \leq X \leq 1.5)$
(c) $P(1.5 \leq X)$.

Problem 7, Page 151

The time $X$ (min) for a lab assistant to prepare the equipment for a certain experiment is believed to have a uniform distribution with $A = 25$ and $B = 35$.

(a) Write the pdf of $X$ and sketch its graph.
(b) What is the probability that preparation time exceeds 33 min?
(c) What is the probability that preparation time is within 2 min of the mean time?
(d) For any $a$ such that $25 < a < a+2 < 35$, what is the probability that preparation time is between $a$ and $a + 2$ min?

Problem

Find the cdf of $X =$ measurement error with pdf

$$f(x) = \begin{cases} 
\frac{3}{22}(4 - x^2) & -2 \leq x \leq 2 \\
0 & \text{otherwise}
\end{cases}$$

(a) Compute $P(X < 0)$.
(b) Compute $P(-1 < X < 1)$.
(c) Verify that the median of the distribution of $X$ is zero.

§4.4. Expected or Mean Values

Definition: The expected or mean value of a continuous random variable $X$ with pdf $f(x)$ is

$$\mu = \mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx.$$
If $h(X)$ is any function of $X$, then

$$
\mu_{h(X)} = E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f(x) \, dx.
$$

A very important special case which we have met in our study of discrete random variable is when $h(X) = (X - \mu)^2$. The expectation is then called the variance of $X$ and denoted by $\sigma^2$ or $\sigma^2_X$ or $V(X)$. That is,

$$
\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) \, dx.
$$

We repeat, once again, that for computational purposes, it is easier to use the expression

$$
V(X) = E(X^2) - [E(X)]^2.
$$

Examples and Practice Problems (see class notes for solutions)

**Problem 21, Page 160**

An ecologist wishes to mark off a circular sampling region having radius 10m. However, the radius of the resulting region is actually a random variable $R$ with pdf

$$
f(r) = \begin{cases} 
\frac{3}{4}[1 - (10 - r)^2] & 9 \leq r \leq 11 \\
0 & \text{otherwise}
\end{cases}
$$

(a) Compute $E(R)$ and $V(R)$.

(b) What is the expected area of the resulting circular region ?

**Problem 22, Page 160**

The weekly demand for propane gas (in 1000’s of gallons) from a particular facility is a continuous random variable $X$ with pdf

$$
f(x) = \begin{cases} 
2 \left(1 - \frac{1}{x^2}\right) & 1 \leq x \leq 2 \\
0 & \text{otherwise}.
\end{cases}
$$

(a) Compute the cdf of $X$.

(b) Obtain an expression for the $(100p)$th percentile. What is the median ?

(c) Compute the mean and variance of $X$.

(d) If 1.5 thousand gallons are in stock at the beginning of the week and no new supply is due in during the week, how much of the 1.5 thousand gallons is expected to be left at the end of the week ?
§4.5. The Normal Distribution

In statistics, the normal distribution is the most important distribution because several random variables that we encounter in real life have distributions that can be approximated by the normal distribution. Students who chose to pursue a career in statistics or take more statistics courses will find that the normal distribution is central to several statistical techniques that are used in data analysis.

**Definition:** A continuous random variable $X$ is said to have a normal distribution with parameters $\mu$ and $\sigma^2$ if the pdf of $X$ is

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

where $-\infty \leq \mu \leq \infty$ and $\sigma^2 > 0$.

It can be shown that the mean and variance of a normally distributed random variable $X$ are $E(X) = \mu$ and $V(X) = \sigma^2$. The function $g(x) = \exp(x)$ is called the exponential function with $f(1) = \exp(1) \approx 2.718282$. In statistics, the notation $X \sim N(\mu, \sigma^2)$ is interpreted to mean that the random variable $X$ has the normal distribution with mean $\mu$ and variance $\sigma^2$. Some properties of the normal distribution are as follows.

(a) The normal density curve is bell shaped.

(b) The normal density curve is symmetric about the mean $x = \mu$.

(c) The median of the normal density curve is equal to the mean of the normal density curve. That is, the curve is centered on $x = \mu$ and the point $x = \mu$ divides the area under the normal curve into two equal parts. Therefore, the area to the left of $x = \mu$ is 0.5 and the area to the right is also 0.5.

(d) If $X \sim N(\mu, \sigma_1^2)$ and $Y \sim N(\mu, \sigma_2^2)$ are two normally distributed random variables with same mean but different variance $\sigma_1^2 > \sigma_2^2$, then the density curve of the random variable $X$ with the larger variance $\sigma_1^2$ will be more spread out than the density curve for $Y$.

Now, let $X \sim N(\mu, \sigma^2)$ consider a change of variable from $x$ to $z$ by defining a new random variable

$$Z = \frac{X - \mu}{\sigma},$$

which comes from the argument of the exponential function in the expression for the normal density. This technique is called standardizing the random variable $X$. If $Q$ is a random variable with mean
\( \mu_Q = E(Q) \) and standard deviation \( \sigma_Q = \sqrt{V(Q)} \), then we can define a new random variable by standardizing \( Q \). In general, to standardize \( Q \), we subtract \( \mu_Q \) from \( Q \) and divide by \( \sigma_Q \) to obtain the expression, say \( U \), given by

\[
U = \frac{Q - \mu_Q}{\sigma_Q}.
\]

It is quite straightforward to verify that \( \mu_Z = E(Z) = 0 \) and \( \sigma^2_Z = V(Z) = 1 \). Now, using a common technique in statistic, which is beyond the scope of this course, we can see that the new density function for \( Z \), called the **standard normal density function** is

\[
f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad -\infty \leq z \leq \infty.
\]

The properties of the standard normal density curve are the same as before except that \( \mu = 0 \) and \( \sigma^2 = 1 \). We shall denote the cdf of a standard normal random variable \( Z \) by \( \Phi(Z) \)

\[
\Phi(Z) = P(Z \leq z) = \int_{-\infty}^{z} f(y; 0, 1) dy.
\]

Observe that computing probabilities involving normal random variables will require that we integrate the normal density function. This integral is not easy to evaluate. As a result, values of the cumulative distribution function \( \Phi(z) \) of the standard normal curve have been evaluated and tabulated for us to use.

**Examples (see notes for solutions)**

**Problem 26, Page 171**

Let \( Z \) be a standard normal random variable and calculate the following probabilities

(a) \( P(0 \leq Z \leq 2.17) \)

(b) \( P(-1.53 \leq Z \leq 2) \)

(c) \( P(-1.75 \leq Z) \)

**Problem 27, Page 171**

In each case, determine the value of the constant \( c \) that makes the probability statement correct.

(a) \( \Phi(c) = 0.9838 \)

(b) \( P(c \leq |Z|) = 0.016 \)
Percentiles of the Standard Normal Distribution

We recall that the point \( c \), such that \( \Phi(c) = P(Z \leq c) = p \) is called the \((100p)\)th percentile of the standard normal distribution. A notation that is commonly used in relation to the normal distribution is \( z_\alpha \), where \( z_\alpha \) is the \( 100(1 - \alpha)\)th percentile of the standard normal distribution. That is,

\[
\Phi(z_\alpha) = P(Z \leq z_\alpha) = 1 - \alpha.
\]

This means that the area under the standard normal curve to the right of \( z_\alpha \) is \( \alpha \) and the area to the left is \( 1 - \alpha \). As a result of symmetry of the standard normal curve about the point \( \mu_z = 0 \), we have that the area to the left of \( -z_\alpha \) is \( \alpha \) and the area to the right is \( 1 - \alpha \). The \( z_\alpha \) values are usually called critical values rather than percentiles for reasons that will be made clear later in our studies. This notation is particularly useful and convenient because later in this course we will need the critical values for some small values of \( \alpha \).

Example: From the standard normal table, we can see that

(a) \( z_{0.05} = 95\text{th percentile of the standard normal distribution} = 1.645 \). That is, for \( \Phi(z_{0.05}) = P(Z \leq z_{0.05}) = 1 - 0.05 = 0.95 \), we must have \( z_{0.05} = 1.645 \).

(b) Similarly, \( z_{0.025} = 97.5\text{th percentile of the standard normal distribution} = 1.96 \).

(c) \( z_{0.01} = 99\text{th percentile of the standard normal distribution} = 2.325 \).

Students will sometimes encounter problems that involve nonstandard normal distributions in the sense that although the distribution is normal, the mean is not zero and the variance may or may not be 1. In that case, we cannot use the standard normal tables directly to compute probabilities. We first have to transform or standardize the nonstandard normal random variable into a standard normal random variable as previously discussed before we can use the standard normal table. For instance, let \( X \sim N(\mu, \sigma^2) \) with \( \mu \neq 0 \) and \( \sigma \neq 1 \), and we wish to compute \( P(a \leq X \leq b) \). Since \( X \) has a nonstandard normal distribution, we first apply the transformation

\[
Z = \frac{X - \mu}{\sigma},
\]

to the variable \( X \) and the limits \( a \) and \( b \) and then use the standard normal table to obtain the required probability as follows.

\[
P(a \leq X \leq b) = P \left( \frac{a - \mu}{\sigma} \leq X - \mu \leq \frac{b - \mu}{\sigma} \right)
= P \left( \frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma} \right)
= \Phi \left( \frac{b - \mu}{\sigma} \right) - \Phi \left( \frac{a - \mu}{\sigma} \right).
\]
Example (see class notes for solution)

Problem 31, Page 171
Suppose the force acting on a column that helps to support a building is normally distributed with mean 15 kips and standard deviation 1.25 kips.

(a) What is the probability that the force is at most 18 kips?
(b) What is the probability that the force is between 10 and 12 kips?
(c) What is the probability that the force differs from 15 kips by at most 2 standard deviations?
(d) How would you characterize the largest 5% of all values of the force?

Problem 32, Page 171
The article “Reliability of Domestic-Waste Biofilm Reactors” (J. of Envir. Engr., 1995: 780 - 790) suggests that substrate concentration (mg/cm$^3$) of influent to a reactor is normally distributed with $\mu = 0.3$ and $\sigma = 0.06$.

(a) What is the probability that the concentration exceeds 0.25?
(b) How would you characterize the largest 5% of all concentration values?

The Normal Approximation to the Binomial Distribution

Students may have noticed that the normal distribution is a continuous distribution. However, it is sometimes used to approximate distributions that are not continuous. Surprisingly, these approximations are sometimes very close to the true value. In this section we shall discuss one of several possible approximations. In such cases a correction, called continuity correction is often introduced in the calculations to account for the fact that we are using a continuous distribution to approximate a discrete distribution.

Let $X \sim Bin(n, p)$. Then, for sufficiently large value of $n$, the distribution of $X$ can be approximated by a normal distribution with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$. The approximation appear to work well if both $np \geq 10$ and $n(1-p) \geq 10$. In particular, if $x$ is any one of the possible values of $X$, then

$$P(X \leq x) = B(x; n, p) \approx P(X \leq x + 0.5) = P \left( \frac{X - \mu}{\sigma} \leq \frac{x + 0.5 - \mu}{\sigma} \right)$$

$$= P \left( \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}} \right)$$

$$= \Phi \left( \frac{x + 0.5 - \mu}{\sqrt{np(1-p)}} \right).$$

11
Problem 51, Page 173
Suppose only 40% of all drivers in a certain state regularly wear a seat belt. A random sample of 500 drivers is selected. What is the probability that between 180 and 230 (inclusive) of the drivers in the sample regularly wear a seat belt?

5. JOINTLY DISTRIBUTED DISCRETE RANDOM VARIABLES AND RANDOM SAMPLES
In some experimental situations, the experimenter may define more than one random variable, say $X$ and $Y$, on the sample space of the experiment and may be interested in studying some properties of these random variables. The experimenter may wish to obtain the distribution of $X$ and the distribution of $Y$ separately. It is also possible to obtain the distribution of $X$ and $Y$ together (i.e. the joint distribution of $X$ and $Y$. The joint probability distribution of $X$ and $Y$ outlines how much probability mass is placed on each possible pair of values $(x, y)$. In this section, we shall restrict our discussion to joint distributions for discrete random variables.

**Definition:** Let $X$ and $Y$ be two discrete random variables defined on the sample space of an experiment. The joint probability mass function (pmf) $p(x, y)$ of $X$ and $Y$ is defined, for each pair of numbers $(x, y)$, by

$$p(x, y) = P(X = y \text{ and } Y = y).$$

Given the joint pmf of two discrete random variables, it is possible to obtain the pmf of each of the random variables separately. These separate pmfs are called marginal pmfs. The marginal pmf of, say $X$, is obtained by summing the joint pmf over all possible values of $Y$.

**Definition:** Let the joint probability mass function (pmf) of $X$ and $Y$ be $p(x, y)$. Then the marginal pmf of $X$, denoted by $p_X(x)$ is given by

$$p_X(x) = \sum_{all \ y} p(x, y).$$

Similarly, the marginal pmf of $Y$, denoted by $p_Y(y)$ is given by

$$p_Y(y) = \sum_{all \ x} p(x, y).$$
Given the marginal pmfs of $X$ and $Y$, we say that the two random variables are independent if the joint pmf is equal to the product of the marginal pmfs at all possible pair of values $(x, y)$. That is, if $X$ and $Y$ are independent then

$$p(x, y) = p_X(x) \cdot p_Y(y), \text{ for every } (x, y).$$

If this relation fails to hold, even for only one of all the possible values, then the random variables are not independent. They are dependent.

**Examples and Practice Problems (see class notes for solutions)**

**Problem 1, Page 215**

A service station has both self-service and full-service islands. On each island, there is a single regular unleaded pump with two hoses. Let $X$ denote the number of hoses being used on the self-service island, and let $Y$ denote the number of hoses on the full-service island in use. The joint pmf of $X$ and $Y$ are in the table below:

<table>
<thead>
<tr>
<th>$p(x, y)$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>0.1</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>1</td>
<td>0.08</td>
<td>0.20</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>0.06</td>
<td>0.14</td>
<td>0.30</td>
</tr>
</tbody>
</table>

(a) What is $P(X = 1 \text{ and } Y = 1)$?
(b) Compute $P(X \leq 1 \text{ and } Y \leq 1)$.
(c) Compute $P(X \neq 0 \text{ and } Y \neq 0)$.
(d) Compute $P(X + Y \leq 1)$.
(e) Compute the marginal pmf of $X$ and of $Y$.
(f) Are $X$ and $Y$ independent random variables?

**Problem 3, Page 216**

A certain market has both an express checkout line and a superexpress checkout line. Let $X_1$ denote the number of customers in line at the express checkout at a particular time of day and let $X_2$ denote the number of customers in line at the superexpress checkout at the same time. Suppose the joint pmf of $X_1$ and $X_2$ is as given in the accompanying table.

<table>
<thead>
<tr>
<th>$p(x, y)$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0.08</td>
<td>0.07</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>0.06</td>
<td>0.15</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>0.04</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.03</td>
<td>0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.01</td>
<td>0.05</td>
<td>0.06</td>
</tr>
</tbody>
</table>
(a) What is the probability that there is exactly one customer in each line?

(b) What is the probability that the numbers of customers in the two lines are identical?

(c) What is the probability that there are at least two more customers in one line than in the other line?

(d) What is the probability that the total number of customers in the two lines is exactly four? At least four?

Problem 4, Page 216

Return to the situation described in Problem 3, Page 212.

(a) Determine the marginal pmf of $X_1$ and then calculate the expected number of customers in line at the express checkout.

(b) Determine the marginal pmf of $X_2$.

(c) Are $X_1$ and $X_2$ independent?

5.1. Expected Values, Covariance and Correlation

Consider two jointly distributed discrete random variables $X$ and $Y$ with joint pmf $p(x, y)$. Let $h(X, Y)$ be any function of the random variables, then $h(X, Y)$ is itself a random variable. The expected or mean value of $h(X, Y)$ denoted by $\mu_{h(X,Y)}$ or $E[h(X,Y)]$ is given by

$$
\mu_{h(X,Y)} = E[h(X,Y)] = \sum_x \sum_y h(x, y) \cdot p(x, y).
$$

A special case is when $h(X, Y)$ takes the form $h(X, Y) = (X - \mu_X)(Y - \mu_Y)$. In that case, $E[h(X,Y)]$ is called the covariance of $X$ and $Y$ and denoted by $Cov(X,Y)$. That is, the covariance between $X$ and $Y$ is

$$
Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) \cdot p(x, y).
$$

For ease of computation it is best to use the formula

$$
Cov(X,Y) = E(XY) - \mu_X \cdot \mu_Y.
$$

If $X$ and $Y$ are independent, it can be shown that $E(XY) = E(X) \cdot E(Y) = \mu_X \cdot \mu_Y$. It is then clear that when $X$ and $Y$ are independent, $Cov(X,Y) = 0$. On the other hand $Cov(X,Y) = 0$ does not necessarily imply independence.
We have seen that the variance of a random variable $\sigma^2_X$ or $V(X)$ is a measure of the size of the error made by an experimenter. If the experimenter committed a lot of mistakes during experimentation, the variance will tend to be large. Alternatively, if the experimenter was very careful and made little or no mistakes during repeated trials of the experiment, variance will tend to be small. Now, the covariance between $X$ and $Y$ measures how the two variables are changing together. For instance if the values of $Y$ tend to increase as $X$ increases, then we say that there is a positive linear relationship between $X$ and $Y$. In that case, the value of $Cov(X,Y)$ will be positive and large if the linear relationship is very strong. On the other hand if the values of $Y$ tend to decrease as $X$ increases, then there is a negative linear relationship between $X$ and $Y$ and the $Cov(X,Y)$ will be negative. One problem with using covariance as a measure of the strength of linear relationships is the fact that it is usually affected by the unit in which the observations were measured. In order to remove the dependence of the covariance on the units, it is standardized to obtain the correlation coefficient denoted by $\rho_{X,Y}$ or simply $\rho$ or $Corr(X,Y)$ and given by

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y},$$

where $\sigma_X$ is the standard deviation of $X$ and $\sigma_Y$ is the standard deviation of $Y$.

Examples and Practice Problems (see class notes for solutions)

Refer to Problem 1, Page 215.

(a) Compute the expected total number of hoses in use $E(X + Y)$.

(b) Compute the covariance for $X$ and $Y$.

(c) Compute the correlation for $X$ and $Y$.

Problem 22, Page 224

An instructor has given a short quiz consisting of two parts. For a randomly selected student, let $X =$ the number of points earned on the first part and $Y =$ the number of points earned on the second part. Suppose that the joint pmf of $X$ and $Y$ is given in the accompanying table

<table>
<thead>
<tr>
<th>$p(x, y)$</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.02</td>
<td>0.06</td>
<td>0.02</td>
<td>0.10</td>
</tr>
<tr>
<td>5</td>
<td>0.04</td>
<td>0.15</td>
<td>0.20</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>0.01</td>
<td>0.15</td>
<td>0.14</td>
<td>0.01</td>
</tr>
</tbody>
</table>
(a) If the score recorded in the grade book is the total number of points earned on the two parts, what is the expected recorded score \( E(X + Y) \) ?

(b) If the maximum of the two scores is recorded, what is the expected recorded score.

(c) Compute the covariance for \( X \) and \( Y \).

(d) Compute the correlation for \( X \) and \( Y \).

(e) Instead of marking the first part out of 15 and the second part out of 10, suppose the instructor changed the points in the first part to \( y = 0, 1, 2, 3 \) and \( x = 0, 1, 2 \) for the second part. How would this affect the covariance and the correlation? Explain.

### 5.2. Statistics and Their Distribution

Suppose that the owner of a vending machine on campus is trying to determine how often the machine should be refilled in a week. The owner decides to observe the number of items left in the vending machine at the end of each day. Let \( X \) = number of items left in the machine on any given day. Now, at the end of Day 1 suppose the observed number of items is \( x_1 \). Similarly, let \( x_2, x_3, \ldots, x_{30} \) be the observed numbers from Day 2 to Day 30. We shall call these observed values a sample of size 30. From these observed numbers we can compute the average number of items the owner expects will be left over in any month of the year,

\[
\bar{x} = \frac{x_1 + x_2 + \cdots + x_{30}}{30} = \frac{1}{30} \sum_{i=1}^{30} x_i.
\]

We can also determine how the numbers vary from day to day,

\[
s^2 = \frac{1}{29} \sum_{i=1}^{30} (x_i - \bar{x})^2,
\]

called the sample variance. Suppose, we repeat the experiment for another 30 days, we expect the observed values \( x_1, x_2, x_3, \ldots, x_{30} \) to be slightly or very different from the observed values from the first experiment. This will in turn lead to different values of \( \bar{x} \) and \( s^2 \). This variation in observed values in turn implies that the value of any function of the sample observations, such as mean, variance, standard deviation, and so on, also varies from sample to sample.

**Definition:** A statistic is any quantity that is calculated from sample data.

Now, before we actually observe the daily number of left over items, the owner has no way of knowing what the value will be on any given day. The number could range from 0 to the total number of items the machine can contain, say \( n \). Therefore, before the experiment is actually performed, the
daily number of leftover items is itself viewed as a random variable. Denote the random variables by $X_1, X_2, X_3, \cdots, X_{30}$. Therefore, the mean and variances computed from these random variables will also be a random variable,

$$
\bar{X} = \frac{X_1 + X_2 + \cdots + X_{30}}{30} = \frac{1}{30} \sum_{i=1}^{30} X_i,
$$

and

$$
S^2 = \frac{1}{29} \sum_{i=1}^{30} (X_i - \bar{X})^2.
$$

Notice that we have used lower case or small letter $x$ for the actual observations and upper case or capital letter $X$ to represent the random variable. Thus, the sample mean regarded as a statistic is denoted by $\bar{X}$ with calculated or observed value $\bar{x}$. Similarly, the sample variance regarded as a statistic is represented as $S^2$ with $s^2$ as its computed value.

**Definition:** The random variables $X_1, X_2, \cdots, X_n$ are said to form a random sample of size $n$ if

1. the $X_i$’s are independent random variables.
2. every $X_i$ has the same distribution.

Another way of saying the same thing is to say that $X_1, X_2, \cdots, X_n$ are independent and identically distributed abbreviated as *iid*.

Recall that every random variable has a probability distribution. This implies that every statistic, being a random variable, has a probability distribution. The probability distribution of a statistic is sometimes called **sampling distribution** simply to emphasize how values of a statistic varies from sample to sample. The main issue is how the sampling distribution of a statistic can be derived. There are a number of ways in which this can be done. If the $X_i$’s is a random sample and their distribution is known, then we can use probability rules to determine the distribution of a statistic computed from the random sample. In other situations were it is too complicated to use probability rules we may obtain information about the sampling distribution of a statistic by using a computer to mimick the behaviour of the statistic through what is normally called a simulation experiment. We shall use an example to illustrate how the probability mass function of a statistic can be derived.

**Example (see class notes for solution)**

**Problem 38, Page 234**

There are two traffic lights on my way to work. Let $X_1$ be the number of lights at which I must stop and suppose that the distribution of $X_1$ is as follows:
Let $X_2$ be the number of lights at which I must stop on the way home; $X_2$ is independent of $X_1$. Assume that $X_2$ has the same distribution as $X_1$, so that $X_1$, $X_2$ is a random sample of size $n = 2$.

(a) Let $T_0 = X_1 + X_2$ and determine the probability distribution of $T_0$.

(b) Calculate $\mu_{T_0}$ and $\sigma_{T_0}^2$.

Problem 41, Page 236

Let $X$ be the number of packages being mailed by a randomly selected customer at a certain shipping facility. Suppose the distribution of $X$ is as follows.

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

(a) Consider a random sample of size $n = 2$ and let $\bar{X}$ be the sample mean number of packages shipped. Obtain the probability distribution of $\bar{X}$.

(b) Refer to Part (a) and calculate $P(\bar{X} \leq 2.5)$.

(c) Calculate $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}^2$.

(d) Again consider a random sample of size $n = 2$, but now focus on the statistic $R$ = the sample range (difference between the largest and the smallest values in the sample). Obtain the distribution of $R$.

5.3. The Distribution of the Sample Mean

In Problem 41, Page 236, we saw that the population mean of the random variable $X$, was $\mu = E(X) = 2$ and the population variance was $\sigma^2 = V(X) = E(X^2) - \mu^2 = 5 - 4 = 1$. The sample mean $\bar{X}$ of a random sample of size $n$ on the random variable $X$ provides an indication of what a typical value of $X$ will be and the sample variance $S^2$ tells us how spread out the values of $X$ will be from this typical value $\bar{X}$. Intuitively, we can say that since $\bar{X}$ is a typical value of $X$, the mean or expected value of $\bar{X}$ should be the same as that of $X$. That is, we expect that

$$\mu_{\bar{X}} = E(\bar{X}) = E(X) = \mu = 2.$$

In our example in class, Problem 41, Page 236, we saw that this was the case. However, we saw that the variance of $\bar{X}$ was slightly different from the variance of $X$, in the sense that

$$\sigma_{\bar{X}}^2 = V(\bar{X}) = \frac{V(X)}{n} = \frac{\sigma^2}{n} = \frac{1}{2} = 0.5.$$
This gives us an idea of the relationship between the mean of \( X, \mu \) and the mean of \( \bar{X}, \mu_{\bar{X}} \) and that between the variance of \( X, \sigma \) and the variance of \( \bar{X}, \sigma^2_{\bar{X}} \). This relationship can be extended to more general situations. We state this extension below.

**Proposition:** Let \( X_1, X_2, \ldots, X_n \) be a random sample from a distribution with mean value \( \mu \) and standard deviation \( \sigma \). Then

1. \( \mu_{\bar{X}} = E(\bar{X}) = E(X) = \mu. \)
2. \( \sigma^2_{\bar{X}} = V(\bar{X}) = \frac{V(X)}{n} = \frac{\sigma^2}{n}. \) and \( \sigma_{\bar{X}} = \sqrt{\frac{V(X)}{n}} = \frac{\sigma}{\sqrt{n}}. \)

In addition, if \( T_0 = X_1 + X_2 + X_3 + \cdots + X_n \) is the sample total, then \( E(T_0) = n\mu, \) \( \sigma^2_{T_0} = V(T_0) = n\sigma^2 \) and \( \sigma_{T_0} = \sigma\sqrt{n}. \)

We remark that the above statement refers to a random sample from any distribution. We state below the distribution of \( \bar{X} \) if the random sample is from a normal distribution.

**Proposition:** If in particular the random sample \( X_1, X_2, \ldots, X_n \) is from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), \( (X \sim N(\mu, \sigma^2)) \) then

1. \( \bar{X} \sim N\left( \mu, \frac{\sigma^2}{n} \right). \)
2. \( T_0 \sim N(n\mu, n\sigma^2). \)

**Remark:** Even when the random sample \( X_1, X_2, \ldots, X_n \) is not from a normal distribution but the sample size \( n \) is sufficiently large, it can be shown that both the sample mean \( \bar{X} \) and the sample total \( T_0 \) can be approximated by a normal distribution. That is, if the random sample is not from a normal distribution, then

1. \( \bar{X} \approx N\left( \mu, \frac{\sigma^2}{n} \right). \)
2. \( T_0 \approx N(n\mu, n\sigma^2). \)

The larger the value of \( n \), the better the approximation. This statement of approximation is called the **central limit theorem (CLT)**. In this class, by sufficiently large, we mean \( n > 30. \)

**Example (see class notes for solution)**

**Problems 46, 47, Page 242**

The inside diameter of a randomly selected piston ring is a random variable with mean value 12cm and standard deviation 0.04cm.
(a) If $\bar{X}$ is the sample mean diameter for a random sample of $n = 16$ rings, where is the sampling distribution of $\bar{X}$ centered, and what is the standard deviation of the $\bar{X}$ distribution?

(b) Suppose the distribution of diameter is normal, calculate $P(11.99 \leq \bar{X} \leq 12.01)$.

**Problem 48, Page 242**

Let $X_1, X_2, \ldots, X_{100}$ denote the actual net weights of 100 randomly selected 50-lb bags of fertilizer. If the expected weight of each bag is 50 and the variance is 1, calculate $P(49.75 \leq \bar{X} \leq 50.25)$.

**Problem 50, Page 243**

The breaking strength of a rivet has a mean value of 10,000 psi and a standard deviation of 500 psi. What is the probability that the sample mean breaking strength for a random sample of 40 rivets is between 9900 and 10,200?

### 5.4. The Distribution of a Linear Combination

Some students may have noticed that the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

can also be written as

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n.$$  

If we replace the coefficients of $X_1, X_2, \cdots, X_n$, in the expression for $\bar{X}$, by arbitrary constants $a_1, a_2, \cdots, a_n$, we obtain a more general expression

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n = \sum_{i=1}^{n} a_iX_i,$$

called a linear combination we can use to represent any statistic for different values of the constants.

We observe that the sample mean $\bar{X}$ is a special case of the linear combination $Y$ with $a_1 = \frac{1}{n}$, $a_2 = \frac{1}{n}, \cdots, a_n = \frac{1}{n}$. When $a_1 = 1, a_2 = a_3 = \cdots = a_n = 1$, the linear combination $Y$ becomes the sample total $T_0$. The next issue is how to compute the expectation and variance of a linear combination.

**Proposition:** Let $X_1, X_2, \cdots, X_n$ be random variables with mean values $E(X_1) = \mu_1$, $E(X_2) = \mu_2, \cdots, E(X_n) = \mu_n$ and variances $V(X_1) = \sigma_1^2$, $V(X_2) = \sigma_2^2, \cdots, V(X_n) = \sigma_n^2$. Then, the mean value of the linear combination $Y$ is

$$\mu_Y = E(a_1X_1 + a_2X_2 + \cdots + a_nX_n)$$

$$= a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n)$$

$$= a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n = \sum_{i=1}^{n} a_i\mu_i.$$
Furthermore, if the random variables $X_1, X_2, \ldots, X_n$ are independent, then the variance of the linear combination $Y$ is

\begin{align*}
\sigma_Y^2 &= V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\
&= a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n) \\
&= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2.
\end{align*}

Now, if the random variables $X_1, X_2, \ldots, X_n$ are not independent then the variance of the linear combination $Y$ is

\begin{align*}
\sigma_Y^2 &= V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j Cov(X_i, X_j).
\end{align*}

To fix ideas, we discuss the following examples.

**Example (see class notes for solution)**

**Problem 59, Page 247**

Let $X_1, X_2$ and $X_3$ represent the times necessary to perform three successive repair tasks at a certain service facility. Suppose they are independent normal random variables with expected values $\mu_1, \mu_2$ and $\mu_3$ and variances $\sigma_1^2, \sigma_2^2$ and $\sigma_3^2$ respectively. If $\mu_1 = \mu_2 = \mu_3 = 60$ and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 15$ calculate

(a) $P(X_1 + X_2 + X_3 \leq 200)$.

(b) $P(58 \leq \bar{X} \leq 62)$.

(c) $P(-10 \leq X_1 - 0.5X_2 - 0.5X_3 \leq 5)$.

(d) If $\mu_1 = 40$, $\mu_2 = 50$, $\mu_3 = 60$, $\sigma_1^2 = 10$, $\sigma_2^2 = 12$ and $\sigma_3^2 = 14$, calculate $P(X_1 + X_2 + X_3 \leq 160)$.

**Problem 70, Page 248**

Suppose we take a random sample of size $n$ from a continuous distribution having median 0 so that the probability of any one observation being positive is 0.5. We now disregard the signs of the observations, rank them from smallest to largest in absolute value, and then let $W =$ the sum of the ranks of the observations having positive signs. For example, if the observations are -0.3, 0.7, 2.1, and -2.5, then the
ranks of the positive observations are 2 and 3, so \( W = 5 \). In Chapter 15, \( W \) is called Wilcoxon’s signed rank statistic. \( W \) can be represented as follows:

\[
W = 1 \cdot Y_1 + 2 \cdot Y_2 + 3 \cdot Y_3 + \cdots + n \cdot Y_n = \sum_{i=1}^{n} i \cdot Y_i.
\]

where the \( Y_i \)'s are independent Bernoulli random variables, each with \( p = 0.5 \) \((Y_i = 1\) corresponds to the observation with rank \( i \) being positive). Compute the following:

(a) \( E(Y_i) \) and then \( E(W) \).

\[
\text{Hint} \quad \sum_{i=1}^{n} i = \frac{n(n + 1)}{2}.
\]

(b) \( V(Y_i) \) and then \( V(W) \).

\[
\text{Hint} \quad \sum_{i=1}^{n} i^2 = \frac{n(n + 1)(2n + 1)}{6}.
\]

6. POINT ESTIMATION

NOT PART OF SYLLABUS. SKIP TO CHAPTER 7.

7. INTERVAL ESTIMATION

Let us assume that we are trying to figure out the amount of money a student needs to survive in a typical university for a given semester. Now, based on our experience as students or other information available to us, in the form of data, we may estimate the amount in two ways. We can choose to use the data to specify a single amount, say $5000.00, as our estimate. Such an estimate which specifies a single value is called a point estimate. It is called a point estimate because it is a single value, a point in some sense. Such estimates are discussed in Chapter 6. Someone who is more careful and does not like to be wrong may suggest instead that we specify an interval as an estimate. So, we may say, the amount should lie between $3500 and $4500. Such an estimate is
called an interval estimate. We may decide not to be too precise by specifying a very wide range of values, say $3500 and $10000. In that case, the interval estimate will be unreliable and therefore not very helpful to students because it is too wide, the width being $10000 - $3500 = 6500. In this section, we will be focusing on how to use information in data to obtain interval estimates.

7.1. Basic Properties of Confidence Intervals

We shall introduce the basic concepts of interval estimation with an example that is practically unrealistic. We shall then consider situations that may arise in practice. Consider a random variable $X = \text{alcohol content of a certain brand of wine manufactured by a company}$. Suppose, the problem is to estimate the unknown population mean $\mu$ of the alcohol content of all bottles of this particular brand of wine manufactured by the company.

Let us assume that $X \sim N(\mu, \sigma^2)$, with $\sigma^2$ known and suppose that we take a random sample of size $n$ wine bottles of this brand with alcohol content $X_1, X_2, \cdots, X_n$. On the average, the alcohol content of the $n$ wine bottles will be $\bar{X}$. From results of §5, we know that $\bar{X} \sim N(\mu, \sigma^2/n)$. If we standardize $\bar{X}$, we obtain a standard normal random variable

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Using the cumulative standard normal probability table we find that

$$P(-1.96 < Z < 1.96) = 0.95.$$

If we replace $Z$ by the expression for standardizing the mean, we have

$$P \left( -1.96 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96 \right) = 0.95.$$

Then, by simplifying the expression within the brackets, we can rewrite the probability expression above as

$$P \left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) = 0.95.$$

This simply means that the probability that the random interval

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right),$$

will cover the “true” mean $\mu$ is 0.95. The interval is random because the statistic $\bar{X}$ is random. That is, before the experiment is performed and measurements are taken, we are 95% certain or confident
that the “true” mean $\mu$ will lie inside the interval above. Once the experiment is performed and measurements are observed, say $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, then the observed average or mean becomes $\bar{X} = \bar{x}$. The resulting fixed, non-random interval given by

$$\left( \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right),$$

or written as

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

or in a more compact form as

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}},$$

is called a 95% confidence interval for $\mu$. The situation we have used to obtain the expression for the confidence interval is practically unrealistic because of the assumptions we have made. We assumed that the random sample is from a normal population with unknown mean $\mu$ and known variance $\sigma^2$. In practice, the mean is needed in order to compute the variance. So, if the mean is unknown, it is highly unlikely that the variance will be known.

Some students may have noticed that there is a relationship between the value 1.96 in the expression for the confidence interval and the confidence level 0.95. We observe that 1.96 is the $(1 - 0.05/2)100^{th} = (1 - 0.025)100^{th} = 97.5^{th}$ percentile or critical value of the standard normal distribution. Using the notations from Chapter 4, we write

$$z_{0.05/2} = 1.96.$$ 

Thus, for a general confidence level, say 99% or 98% or 90%, a $(1 - \alpha)100^{th}$ confidence interval for the population mean $\mu$ of a distribution is obtained by replacing 1.96 with the critical value $z_{\alpha/2}$,

$$\left( \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right),$$

or $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$. The width of the $(1 - \alpha)100^{th}$ confidence interval is

$$\left( \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) - \left( \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

An interval width a large width shows that the error in our estimation is too large and therefore the interval will not be reliable or the estimate becomes unrealistic.

**Example (see class notes for solution)**

\[24\]
Problems 1, 5, Page 290

Assume that the helium porosity (in percentage) of coal samples taken from any particular seam is normally distributed with true standard deviation 0.75.

(a) What is the confidence level for the interval $\bar{x} \pm 2.81 \frac{\sigma}{\sqrt{n}}$?

(b) What value of $z_{\alpha/2}$ in the CI formula results in a confidence level of 99.7%?

(c) Compute a 98% CI for the true average porosity of a certain seam if the average porosity for 16 specimens from the seam was 4.56.

(d) How large a sample size is necessary if the width of the 95% interval is to be 0.4?

Problem 2, Page 290

Each of the following is a confidence interval for $\mu =$ true average (i.e., population mean) resonance frequency (Hz) for all tennis rackets of a certain type:

(114.4, 115.6) (114.1, 115.9)

(a) What is the value of the sample mean resonance frequency?

(b) Both intervals were calculated from the same sample data. The confidence level for one of these intervals is 90% and for the other is 99%. Which of the intervals has the 90% confidence level, and why?

7.2. Large-Sample Confidence Intervals for a Population Mean and Proportion

In §7.1 we made some assumptions which are practically unrealistic. We assumed that the variance of the population distribution $\sigma^2$ is known. The assumption that the random sample is from a normal population may be valid sometimes but may fail in other situations. We will now discuss large-sample confidence intervals whose validity does not depend on these assumptions.

Large-Sample Interval for $\mu$

Let $X_1, X_2, \cdots, X_n$ be a random sample from a population whose distribution is not known with mean $\mu$ and variance $\sigma^2$ that are also unknown. Recall that if the sample size $n$ is large, then by the Central Limit Theorem we discussed in Chapter 5, the distribution of the sample mean $\bar{X}$ is approximately normal with mean $E(\bar{X}) = \mu$ and variance $V(\bar{X}) = \sigma^2/n$. That is,

$$\bar{X} \approx N(\mu, \sigma^2/n).$$

It follows that the standardized random variable

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

25
has approximately the standard normal distribution. Therefore,

\[ P \left( -z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2} \right) \approx 1 - \alpha. \]

Following the technique used in §7.1 we can rearrange the expression in the parentheses to obtain an approximate \((1 - \alpha)100\%\) confidence interval for the population mean \(\mu\) as

\[ \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}. \]

The difficulty with this expression for confidence interval is that \(\sigma\) is not known. Therefore, since \(n\) is sufficiently large, we simply replace \(\sigma\) by the sample standard deviation

\[ s = \sqrt{\frac{1}{n-1} \left( \sum_{i=1}^{n} x_i^2 - n \bar{x}^2 \right)}, \]

to obtain the large-sample confidence interval for \(\mu\) with an approximate confidence level of \((1 - \alpha)100\%\) as

\[ \bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}. \]

Examples (see class notes for solution)

**Problem 13, Page 297**

The article “Extravisual Damage Detection ? Defining the Standard Normal Tree” *(Photogrammetric Engr. and Remote Sensing, 1981: 515-522)* discusses the use of color infrared photography in identification of normal trees in Douglas fir stands. Among data reported were summary statistics for green-filter analytic optical densitometric measurements on samples of both healthy and diseased trees. For a sample of 69 healthy trees, the sample mean dye-layer density was 1.028, and the sample standard deviation was 0.163.

(a) Calculate a 95\% (two-sided) CI for the true average dye-layer density for all such trees.

(b) Suppose the investigators had made a rough guess of 0.16 for the value of \(s\) before collecting data. What sample size would be necessary to obtain an interval width of 0.05 for a confidence level of 95\% ?
Large-Sample Interval for a Population Proportion \( p \)

In Chapter we discussed how the normal distribution can be used to approximate the binomial distribution. Now, suppose that \( X \) is a binomial random variable with parameters \( n \) and \( p \), where \( n \) is the number of trials and \( p \) is the probability or proportion of successes. We recall that if \( n \) is sufficiently large, then \( X \) has approximately a normal distribution with mean \( E(X) = np \) and variance \( V(X) = np(1 - p) \).

Now, if \( X = \text{number of successes in} \ n \ \text{trials} \), a reasonable estimator of the proportion of successes is \( \hat{p} = X/n \), the sample fraction of successes. Since \( \hat{p} \) is just a fraction of \( X \), the distribution of \( \hat{p} \) is also approximately normal. It can be shown that \( E(\hat{p}) = p \) and \( V(\hat{p}) = p(1 - p)/n \). Therefore, the standardized \( \hat{p} \) defined by

\[
Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}},
\]

is approximately standard normal. Thus, using the same approach as before, we find that

\[
P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha.
\]

If we rearrange the inequalities in the bracket, we obtain a **large-sample confidence interval for a population proportion** \( p \) with an approximate confidence level of \((1 - \alpha)100\%\) to be

\[
\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}.
\]

Suppose that the width of the confidence interval for \( p \) is \( w \). Then, the sample size required to obtain a confidence interval of width \( w \) is

\[
n = \frac{2z_{\alpha/2}^2\hat{p}(1 - \hat{p}) - z_{\alpha/2}^2w^2}{\sqrt{4z_{\alpha/2}^4\hat{p}(1 - \hat{p})[\hat{p}(1 - \hat{p}) - w^2] + w^2z_{\alpha/2}^4}}.
\]

Unfortunately, the formula for \( n \) involves the unknown \( \hat{p} \). It can be shown that the maximum value of \( \hat{p}(1 - \hat{p}) \) is attained when \( \hat{p} = 0.5 \). Therefore, using the value of \( \hat{p} = 0.5 \) will ensure that the width will be at most \( w \) regardless of what value of \( \hat{p} \) results from the sample. Alternatively, if the investigator believes strongly based on prior experience and information, that the true value of \( p \), say \( p_0 \), is less than 0.5 then the value \( p_0 \) can be used in place of \( \hat{p} \).

**Examples** (see class notes for solution)
Problem 21, Page 298
A random sample of 539 households from a certain city was selected, and it was determined that 133 of these households owned at least one firearm. Compute a 95% confidence interval for the proportion of all households in this city that own at least one firearm.

Problem 23, Page 298
The article “An Evaluation of Football Helmets Under Impact Conditions” (American Journal of Sports Medicine, 1984: 233-237) reports that when each football helmet in a random sample of 37 suspension-type helmets was subjected to a certain impact test, 24 showed a damage. Let $p$ denote the proportion of all helmets of this type that would show damage when tested in the prescribed manner.

(a) Calculate a 99% CI for $p$.

(b) What sample size would be required for the width of a 99% CI to be at most 0.10, irrespective of $\hat{p}$.

7.3. Intervals Based on a Normal Population Distribution

The interval estimates we discussed in §7.1 was not practical because of the assumptions underlying the estimates. In §7.2 we discussed intervals for large samples, e.g. that is when $n > 30$. Here, we discuss interval estimates for small samples and the assumptions under which the interval estimation method can be used. In §7.2 the intervals where obtained by invoking the central limit theorem because the sample size was assumed to be large. The key was that the standardized variable

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

has approximately the standard normal distribution when $n$ is large even when the population distribution of the random sample $X_1, X_2, \cdots, X_n$ is not the normal distribution.

Now, when the sample size $n$ is small, we cannot use the central limit theorem to obtain a random variable that has an approximate standard normal distribution. However, we can show that if $n$ is small and the random sample $X_1, X_2, \cdots, X_n$ of size $n$ is from a normal population, then the standardized variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

has a probability distribution that is called a student $t$-distribution with $n - 1$ degrees of freedom, where $S$ is the standard deviation. Notice that the standardized variable has not changed, we
have simply chosen to denote it by $T$ rather than $Z$ in order to emphasize the point that it is no longer normally distributed. Due to the complicated mathematical expression for the probability density function of $T$, we shall not provide the expression in this class. In addition, to assist us in computing probabilities without actually integrating the density function, a table of values of cumulative probability for the $t$-distribution is available in most elementary statistics text and will be made available to us in the class.

**Properties of $t$ Distributions**

The normal distribution is characterized by two parameters, the mean $\mu$ and the variance $\sigma^2$. By that we mean, once these two parameters are known, the normal distribution is completely determined. On the other hand, the $t$-distribution is characterized by only one parameter called the degrees of freedom we shall denote by the Greek letter $\nu$. Possible values of $\nu$ are $1, 2, 3, \cdots$. For a fixed value of $\nu$, some of the properties or features of the density curve of the $t$-distribution are as follows.

Let $t_\nu$ denote the density curve of the $t$-distribution with $\nu$ degrees of freedom. Then,

1. each $t_\nu$ curve is bell-shaped and centered at 0. *The standard normal $z$ curve also has this property.*
2. each $t_\nu$ curve is more spread out than the standard normal $z$ curve. In other words, the variance of data from a population with a $t$-distribution is more likely to be larger than the variance of data from a population with a standard normal distribution.
3. as $\nu$ increases, the spread of the corresponding $t_\nu$ curve decreases.
4. as $\nu$ becomes larger and larger or tends to infinity, the sequence of $t_\nu$ curves approaches the standard normal $z$ curve.

**Notation**

In Chapter 4, we introduced the $z_\alpha$ notation which we called the $z$ critical value. An equivalent notation for the $t$-distribution is $t_{\alpha, \nu} =$ the point on the $t$-axis of the $t_\nu$ curve for which the area to the right under the curve is $\alpha$. We shall call $t_{\alpha, \nu}$ a $t$ critical value.

**The One-Sample $t$ Confidence Interval for a Population mean**

Consider the random sample $X_1, X_2, \cdots, X_n$ of size $n$ from a normal population with $n < 30$. Let $\bar{X}$ be the sample mean. Then, the standardized random variable

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}},$$
has a student t-distribution with $n - 1$ degrees of freedom. Following the technique of §7.1, we can write that

$$P \left( -t_{\alpha/2,n-1} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_{\alpha/2,n-1} \right) \approx 1 - \alpha.$$ 

Therefore, if $\bar{x}$ is the observed sample mean and $s$ is the observed sample standard deviation computed from $n < 30$ observed data from a normal population with unknown mean $\mu$ and unknown variance $\sigma^2$, then a $(1 - \alpha)100\%$ confidence interval for the population $\mu$ is

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}.$$

Examples (see class notes for solution)

**Problem 32 Page 306**

A random sample of $n = 8$ E-glass fiber test specimens of a certain type yielded a sample mean interfacial shear yield stress of 30.2 and a sample standard deviation of 3.1. Assuming that interfacial shear yield stress is normally distributed, compute a 95% CI for true average stress.

**Problem 33 Page 306**

The article “Measuring and Understanding the Aging of Kraft Insulating Paper in Power Transformers” (IEEE Electrical Insul. Mag., 1996: 28-34) contained the following observations on degree of polymerization for paper specimens for which viscosity times concentration fell in a certain middle range

<table>
<thead>
<tr>
<th>418</th>
<th>421</th>
<th>421</th>
<th>422</th>
<th>425</th>
<th>427</th>
<th>431</th>
<th>434</th>
<th>437</th>
</tr>
</thead>
<tbody>
<tr>
<td>439</td>
<td>446</td>
<td>447</td>
<td>448</td>
<td>453</td>
<td>454</td>
<td>463</td>
<td>465</td>
<td></td>
</tr>
</tbody>
</table>

Calculate a two-sided 95% confidence interval for true average degree of polymerization. Does the interval suggest that 440 is a plausible value for true average degree of polymerization? What about 450?

**Problem 37, Page 307**

A study of the ability of individuals to walk in a straight line reported the accompanying data on cadence (strides per second) for a sample of $n = 20$ randomly selected healthy men:

.95, .85, .92, .95, .93, .86, 1.00, .92, .85, .81, .78, .93, .93, 1.05, .93, 1.06, 1.06, .96, .81, .96.

Assuming that the population distribution of cadence is approximately normal calculate and interpret a 95% confidence interval for population mean cadence.
A journal article reports that a sample of size 5 was used as a basis for calculating a 95% CI for the true average natural frequency (Hz) of delaminated beams of a certain type. The resulting interval was (229.764, 233.504). You decide that a confidence level of 99% is more appropriate than the 95% level used. What are the limits of the 99% interval?

8.1. Tests of Hypotheses

Companies, governments and businesses are always seeking better, more effective and efficient ways of doing things. This desire often lead to scientific research with the objective of helping managers decide whether current techniques or theory should be replaced by an alternative approach. Clearly, managers or government or any business will not replace their standard procedures unless there is very strong evidence that the new approach is better than the standard. For instance, suppose that a scientist claims that a new study has shown that on the average, a new technique for water treatment and purification developed by his team is better, cheaper and more efficient than current standard methods. Unless there is very strong evidence in support of this claim, no government will dismantle and replace its current water treatment plants, and the burden of proof is on the scientists and all those who believe in the alternative method. Let $\mu_1 =$ average bacteria count in water after treatment with the standard method and $\mu_2 =$ average bacteria count in water after treatment with new method.

Before the new treatment method was proposed, it was believed that any other method is at best as good as the standard method. That is, the standard method was initially favored. Therefore, the prior belief or initially favored claim can be written mathematically as $\mu_1 - \mu_2 \leq 0$. As a result of the new proposal, the scientist are claiming mathematically that there is an alternative given by $\mu_1 - \mu_2 > 0$. The initially favored claim or the current state of affairs will not be rejected in favor of the alternative unless there is strong evidence in favor of the alternative claim.

It is usual to refer to the prior belief as the null hypothesis denoted by $H_0$ and the claim that is to be proved or disproved or the claim that contradicts the prior belief as the alternative hypothesis and denoted by $H_a$. By definition, a statistical hypothesis or simply hypothesis is a claim or assertion about the value of a single parameter or about the values of several parameters or about the form of an entire probability distribution. For the water treatment example, we then write

$$H_0 : \mu_1 - \mu_2 \leq 0 \text{ versus } H_a : \mu_1 - \mu_2 > 0.$$
A test of hypotheses is a method for using sample data to decide whether to reject or not reject the null hypothesis $H_0$.

In order to decide whether to reject a claim or not, the investigator has to establish a reasonable rule based on available evidence or sample data that will form the basis of the decision. Such a rule is called a test procedure.

**Test Procedures**

Suppose that we claim that female students perform better than male students in mathematics courses at Memorial University. If $X =$ number of female students who pass mathematics courses, then $p = X/N =$ proportion of females with a pass in mathematics courses, where $N =$ total number of students in mathematics courses. Now, the prior belief or initially favored claim is that students performance in mathematics courses is gender neutral. Thus, the null hypothesis can be written mathematically as

$$H_0 : p = 0.5,$$

and the alternative hypothesis which we wish to prove or disprove is that

$$H_a : p > 0.5.$$

Since the number of students taking mathematics courses are so many, I decide to take a random sample of say $n$ students and observe whether the student is a male or female and whether the student passed or failed the course. We note that for each student, there is only two possible outcomes, pass (coded as 1) or fail (denoted by 0). At the end, we count the total number of 1’s which is the observed value, say $x$, of the random variable $X =$ number of female students among $n$ students who passed the mathematics course. Clearly, $X$ is a binomial random variable with parameters $n$ and unknown $p$. That is, $X \sim Bin(n, p)$.

Now, a reasonable rule for deciding whether to reject the null hypothesis $H_0$ or not will depend on $X$. Suppose, $n = 40$ and we found $x = 22$. Is this strong evidence in favour of rejecting $H_0$. The answer is, NO. The value $x = 22$ is probably due to the sample that we took. A different sample may turn up $x = 21$. However, we can decide that any value of $x \geq 28$ is strong enough for us to reject $H_0$. Now, in hypothesis testing, the set of all values for which we reject $H_0$ is called the critical region or rejection region. For this example, we reject $H_0$ for all values of $x$ greater than 28. Thus, the rejection or critical region we shall denote by $R$ can be written mathematically as

$$R = \{x : x > 28\} \text{ or } R = \{x : x = 28, 29, 30, \ldots, 40\}.$$
Clearly, our test procedure or rule for deciding on whether to reject or not reject $H_0$ is made up of two components,

(a) the value of the random variable $X$, which counts the number of 1’s or pass in the response or sample of students performance. In other words, $X$ is a function of the sample data. In this example, the function of the sample data $X$ is called a test statistic, and

(b) the critical or rejection region $R$ given by

$$ R = \{ x : x \geq 28 \} \text{ or } R = \{ x : x = 28, 29, 30, \ldots, 40 \}. $$

The value 28 shall be called the critical value (the value at which the rejection begins). That means, once the critical value is exceeded, the decision is to reject the null hypothesis. Now, we reject $H_0$ if the observed value of $X$, falls within the rejection region.

We generalize the ideas described above by stating that, in general, a test procedure is specified by the following;

1. A test statistic which is a function of the sample data on which the decision to reject $H_0$ or not to reject $H_0$ is to be based.

2. A critical or rejection region $R$ described by the set of all possible test statistic values for which $H_0$ will be rejected.

Again, we reject $H_0$ only if the observed or computed value of the test statistic falls in the critical or rejection region.

**Errors in Hypothesis Testing**

In every decision making process, there is always the risk of making the wrong decision. It is possible to reject $H_0$ when in fact $H_0$ is true. This error is called a type I error. It is also possible to not reject $H_0$ when it is false leading to a type II error. We can tabulate these errors as follows

<table>
<thead>
<tr>
<th>Decision</th>
<th>$H_0$ True</th>
<th>$H_0$ False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>Type I Error</td>
<td>Correct Decision</td>
</tr>
<tr>
<td>Do not Reject $H_0$</td>
<td>Correct Decision</td>
<td>Type II Error</td>
</tr>
</tbody>
</table>
In statistics, the chance of committing a type I error is called the level of significance and denoted by \( \alpha \). That is,

\[
\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ is true}).
\]

We denote the chance of committing a type II error by \( \beta \). That is

\[
\beta = P(\text{Type II error}) = P(\text{Do not Reject } H_0 | H_0 \text{ is false}).
\]

The desire of managers is always to make the chance of committing both Type I and Type II errors as small as possible. Unfortunately, it is impossible to minimize both simultaneously. Statisticians have found that as we decrease the size of \( \alpha \), the size of \( \beta \) becomes larger and larger. Thus, the approach that is commonly adopted is to fix a value for \( \alpha \) and construct a test procedure that will minimize the size of \( \beta \). The test procedure is then called a level \( \alpha \) test.

**Examples (see class notes for solution)**

**Problem 9, Page 325**

Two different companies have applied to provide cable television service in a certain region. Let \( p \) denote the proportion of all potential subscribers who favor the first company over the second. Consider testing \( H_0 : p = 0.5 \) versus \( H_a : p \neq 0.5 \) based on a random sample of 25 individuals. Let \( X \) denote the number in the sample who favor the first company and \( x \) represent the observed value of \( X \).

(a) What function of the random sample of 25 individuals will you use to decide whether to reject \( H_0 \) or not to reject \( H_0 \)?

(b) Which of the following rejection regions is most appropriate and why?

\[
R_1 = \{ x : x \leq 7 \text{ or } x \geq 18 \}, \quad R_2 = \{ x : x \leq 8 \}, \quad R_3 = \{ x : x \geq 17 \}.
\]

(c) In the context of this problem situation, describe what type I and type II errors are.

(d) What is the probability distribution of the test statistic \( X \) when \( H_0 \) is true? Use it to compute the probability of a type I error.

(e) Compute the probability of a type II error for the selected region when \( p = 0.3 \).

(f) Using the selected region, what would you conclude if 6 of the 25 queried favored company 1?
Problem 10, Page 325

A mixture of pulverized fuel ash and Portland cement to be used for grouting should have a compressive strength of more than 1300 KN/m$^2$. The mixture will not be used unless experimental evidence indicates conclusively that the strength specification has been met. Suppose compressive strength for specimens of this mixture is normally distributed with $\sigma = 60$. Let $\mu$ denote the true average compressive strength.

(a) What are the appropriate null and alternative hypothesis?

(b) Let $\bar{X} = \text{sample average compressive strength for } n = 20 \text{ randomly selected specimens}$. Consider a test procedure with test statistic $\bar{X}$ and rejection region $\bar{x} \geq 1331.26$. What is the probability distribution of the test statistic when $H_0$ is true? What is the probability that the mixture was judged to be satisfactory and used when it should not have been used?

8.2. Tests About a Population Mean

In our discussion on tests about a population mean, we shall begin with a case that is rarely met in practice. We begin by assuming that the population is normal and that $\sigma^2$, the population variance is known. Then we shall discuss the case of large-sample test where the distribution of the population is unknown and the population variance is also unknown. We will conclude this section with a discussion on the case where the sample size is small. In that case, we will require the population to be normally distributed. In all three cases, the null hypothesis of interest will be

$$H_0 : \mu = \mu_0,$$

and the alternative hypothesis will be one of the following:

(A) A one-sided upper-tailed test

$$H_a : \mu > \mu_0.$$

This is called an upper-tailed test because $H_0$ is rejected if the value of the test statistic is greater than a specified value. Thus, the rejection or critical region is often on the upper-tail of the distribution of the test statistic.

(B) A one-sided lower-tailed test

$$H_a : \mu < \mu_0.$$

This is said to be a lower-tailed test because $H_0$ is rejected if the value of the test statistic is less than a specified value. This implies that the rejection or critical region will lie on the lower-tail of the distribution of the test statistic.
(C) A two-tailed test

\[ H_a : \mu \neq \mu_0. \]

A test involving this alternative is said to be two-tailed because it is not specified whether \( \mu \) is negative or positive. Therefore, the critical region lies on both tails of the distribution of the test statistic.

**Case I: A Normal Population with Known \( \sigma \)**

Let \( X_1, X_2, \ldots, X_n \) represent a random sample of size \( n \) from a normal population. Then a reasonable function of the sample data that can be used for deciding whether to reject \( H_0 : \mu = \mu_0 \) or not is the sample mean. We saw in our previous example that to compute the probability of a type I error, we had to standardize \( \bar{X} \). Thus, rather than use \( \bar{X} \) we shall use the standardized \( \bar{X} \), when \( H_0 \) is true, as our test statistic. That is, the test statistic for testing \( H_0 : \mu = \mu_0 \) against any one of the three alternatives is given by

\[ Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}. \]

Now let \( z \) be a computed value of the test statistic \( Z \). For a fixed value of the significance level \( \alpha \), the rejection region for a level \( \alpha \) test is

(a) \( z \geq z_\alpha \) if the alternative hypothesis is \( H_a : \mu > \mu_0 \) (upper-tailed test).

(b) \( z \leq -z_\alpha \) if the alternative hypothesis is \( H_a : \mu < \mu_0 \) (lower-tailed test).

(c) \( |z| \geq z_{\alpha/2} \) if the alternative hypothesis is \( H_a : \mu \neq \mu_0 \) (two-tailed test).

When testing hypotheses about a parameter, it may be helpful to follow the steps below.

1. Identify and state the appropriate null and alternative hypotheses.

2. State the level of significance.

3. State the formula for the test statistic and compute its value.

4. Outline the rejection region and the decision rule.

5. Decide whether to reject \( H_0 \) and state your conclusion.
Determination of $\beta$ and Sample size $n$

We shall use the lower-tailed test with rejection region $z \leq -z_\alpha$ to illustrate how we can determine the probability of committing a type II error under Case I. This case is one of the few cases in which a closed form expression can be derived for both the sample size and $\beta$. From definition,

$$\beta = P(\text{Do not reject } H_0 \text{ when } H_0 \text{ is false}).$$

Now, rejecting $H_0$ simply means that the alternative $H_a$ is more plausible. However, under $H_a$ there are several possible values for $\mu$. Thus, the value of $\beta$ will depend on the value of $\mu$ that we use among all the eligible values under $H_a$. Now, suppose $\mu = \mu'$ is a value of $\mu$ that is less than $\mu_0$ under $H_a$. Then,

$$\beta(\mu') = P(Z > -z_\alpha \text{ when } \mu = \mu')$$

$$= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > -z_\alpha \text{ when } \mu = \mu'\right) \quad \text{(using the test statistic)}$$

$$= P\left(\bar{X} > -z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0 \text{ when } \mu = \mu'\right)$$

$$= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} > -z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu = \mu'\right) \quad \text{(since the true mean is } \mu')$$

$$= 1 - P\left(Z \leq -z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu = \mu'\right)$$

$$= 1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right).$$

Since $\mu'$ is less than $\mu_0$, the difference $\mu_0 - \mu'$ will be positive. Now, as $\mu'$ decreases, $\mu_0 - \mu'$ will be more positive and the value of $\beta(\mu')$ will become smaller. Similarly, we can show that

(a) For an upper-tailed test with $H_a : \mu > \mu_0$, the probability of committing a type two error is

$$\beta(\mu') = \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right),$$

where $\mu'$ is a value of $\mu$ that is greater than $\mu_0$ under $H_a$. For an upper or a lower-tailed one-sided test, it can be shown that the sample size $n$ for which a level $\alpha$ test also has $\beta(\mu') = \beta$ at the alternative value $\mu'$ is

$$\left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2.$$
(b) For a two-tailed test with $H_a : \mu \neq \mu_0$, the probability of committing a type two error is

$$\beta(\mu') = \Phi \left( z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma \sqrt{n}} \right) - \Phi \left( -z_{\alpha} + \frac{\mu_0 - \mu'}{\sigma \sqrt{n}} \right),$$

where $\mu'$ is a value of $\mu$ that is not equal to $\mu_0$ under $H_a$. For a two-tailed test, it can be shown that the approximate sample size $n$ for which a level $\alpha$ test also has $\beta(\mu') = \text{a specified } \beta$ at the alternative value $\mu'$ is

$$n = \left[ \frac{\sigma (z_{\alpha/2} + z_{\beta})}{\mu_0 - \mu'} \right]^2.$$

**Examples (see class notes for solution)**

**Problem 15, Page 331**

Let the test statistic $Z$ have a standard normal distribution when $H_0$ is true. Give the significance level for each of the following situations.

(a) $H_0 : \mu > \mu_0$, rejection region $z \geq 1.88$.

(b) $H_0 : \mu \neq \mu_0$, rejection region $z \geq 2.88$ or $z \leq -2.88$.

**Problem 19, Page 337**

The melting point of each of 16 samples of a certain brand of hydrogenated vegetable oil was determined, resulting in $\bar{x} = 94.32$. Assume that the distribution of melting point is normal with $\sigma = 1.2$.

(a) Test $H_0 : \mu = 95$ versus $H_a : \mu \neq 95$ using a two-tailed level 0.01 test.

(b) If a level 0.01 test is used, what is $\beta(94)$, the probability of a type II error when $\mu = 94$?

(c) What value of $n$ is necessary to ensure that $\beta(94) = 0.1$ when $\alpha = 0.01$?

**Case II: Large Sample Tests**

In this section we assume that the random sample $X_1, X_2, \ldots, X_n$ is from an unknown distribution with unknown mean $\mu$ and variance $\sigma^2$. We also assume that the sample size $n$ is large ($n > 30$). Then, from the central limit theorem a reasonable test statistic for testing $H_0 : \mu = \mu_0$ against one of the three possible alternatives is

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}},$$

which has approximately a standard normal distribution. The rejection regions remain the same as before.
Case III: Tests About $\mu$ When $n$ is small

When $n$ is small but the sample $X_1, X_2, \ldots, X_n$ is from a normal population, we have seen that the standardized random variable now has a t-distribution with $n - 1$ degrees of freedom. Thus, a reasonable test statistic for testing $H_0 : \mu = \mu_0$ against one of the three possible alternatives is

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

In this case, the rejection region for a level $\alpha$ test is

(a) $t \geq t_{\alpha,n-1}$ if the alternative hypothesis is $H_a : \mu > \mu_0$ (upper-tailed test).

(b) $t \leq -t_{\alpha,n-1}$ if the alternative hypothesis is $H_a : \mu < \mu_0$ (lower-tailed test).

(c) $|t| \geq t_{\alpha/2,n-1}$ if the alternative hypothesis is $H_a : \mu \neq \mu_0$ (two-tailed test).

Examples (see class notes for solution)

**Problem 27, Page 338**

The Charpy V-notch impact test is the basis for studying many material toughness criteria. This test was applied to 32 samples of a particular alloy at 110°F. The sample average amount of transverse lateral expansion was computed to be 73.1 mils, and the sample standard deviation was $s = 5.9$ mils. To be suitable for a particular application, the true average amount of expansion should be less than 75 mils. The alloy will not be used unless the sample provides strong evidence that this criterion has been met. Test the relevant hypotheses using $\alpha = 0.01$ to decide whether the alloy is suitable.

**Problem 29, Page 338**

The amount of shaft wear (.0001 in.) after a fixed mileage was determined for each of $n = 8$ internal combustion engines having copper lead as a bearing material, resulting in $\bar{x} = 3.72$ and $s = 1.25$. Assuming that the distribution of shaft wear is normal with mean $\mu$, test the hypotheses $H_0 : \mu = 3.5$ versus $H_a : \mu > 3.5$ at $\alpha = 0.05$.

**Problem 30, Page 338**

The recommended daily dietary allowance for zinc among males older than age 50 years is 15 mg/day. The article “Nutrient Intakes and Dietary Patterns of Older
Americans: A National Study” (J. Gerontology, 1992: M145-150) reports the following summary data on intake for a sample of males age 65-74 years: \( n = 115, \bar{x} = 11.3, \) and \( s = 6.43 \). Does this data indicate that average daily zinc intake in the population of all males age 65-74 falls below the recommended allowance?

**Problem 31, Page 338**

In an experiment designed to measure the time necessary for an inspector’s eyes to become used to the reduced amount of light necessary for penetrant inspection, the sample average time for \( n = 9 \) inspectors was 6.32 secs and the sample standard deviation was 1.65 sec. It has previously been assumed that the average adaptation time was at least 7 sec. Assuming adaptation time to be normally distributed, does the data contradict prior belief? Conduct your test at \( \alpha = 0.1 \).

### 8.3. Tests About a Population Proportion

Let \( p \) be the proportion of objects or individuals with a specified characteristic or property. Suppose we label those individuals or objects with the specified property, success (S) and those without the property, failure (F). Then \( p = \) population proportion of successes and \( 1 - p = \) population proportion of failures. When claims are made about \( p \), the problem is to test hypothesis about \( p \). In general, the null hypothesis is of the form

\[
H_0 : p = p_0,
\]

and the alternative hypothesis can be any one of the following three hypotheses.

(A) **Upper tailed test**: The alternative hypothesis is of the form

\[
H_a : p > p_0.
\]

(B) **Lower tailed test**: The alternative hypothesis is of the form

\[
H_a : p < p_0.
\]

(C) **Two-tailed test**: The alternative hypothesis is of the form

\[
H_a : p \neq p_0.
\]

It can be shown that, when the sample size \( n \) is large \((n > 30)\) and the null hypothesis is true, the test statistic

\[
Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},
\]

has approximately the standard normal distribution. The rejection regions for a level \( \alpha \) test are as follows.
(a) \( z \geq z_{\alpha} \) if the alternative hypothesis is \( H_a : \mu > \mu_0 \) (upper-tailed test).

(b) \( z \leq -z_{\alpha} \) if the alternative hypothesis is \( H_a : \mu < \mu_0 \) (lower-tailed test).

(c) \( |z| \geq z_{\alpha/2} \) if the alternative hypothesis is \( H_a : \mu \neq \mu_0 \) (two-tailed test).

Examples (see class notes for solution)

**Problem 35, Page 338**

State DMV records indicate that of all vehicles undergoing emissions testing during the previous year, 70\% passed on the first try. A random sample of 200 cars tested in a particular county during the current year yields 156 that passed on the initial test. Does this suggest that the true proportion for theis county during the current year differs from the previous statewide proportion? Test the relevant hypotheses using \( \alpha = 0.05 \).