Chapter 5: Joint Continuous Random Variables

5-2.1 Joint Probability Distribution

Definition

A joint probability density function for the continuous random variables $X$ and $Y$, denoted as $f_{XY}(x, y)$, satisfies the following properties:

1. $f_{XY}(x, y) \geq 0$ for all $x, y$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \, dx \, dy = 1$
3. For any region $R$ of two-dimensional space

$$P((X, Y) \in R) = \int_{R} \int f_{XY}(x, y) \, dx \, dy \quad (5-14)$$
Important Properties of Joint Probability Distributions, $f_{X,Y}(x,y)$

Marginal Distribution: $f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) \, dy$

Mean: $\mu_X = E(X) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x \, f_{XY}(x, y) \, dx \, dy$

Variance: $\sigma^2_X = V(X) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)^2 \, f_{XY}(x, y) \, dx \, dy$

Covariance: $\sigma_{XY} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)(y - \mu_Y) \, f_{XY}(x, y) \, dx \, dy$

Correlation (coefficient): $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
If $X$ and $Y$ are independent, then

$$f_{XY}(x,y) = f_X(x) f_Y(y)$$
5-5 Linear Combinations of Random Variables

Definition

Given random variables $X_1, X_2, \ldots, X_p$ and constants $c_1, c_2, \ldots, c_p$,

$$Y = c_1X_1 + c_2X_2 + \cdots + c_pX_p$$

(5-34)

is a linear combination of $X_1, X_2, \ldots, X_p$.

Mean of a Linear Combination

If $Y = c_1X_1 + c_2X_2 + \cdots + c_pX_p$,

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \cdots + c_pE(X_p)$$

(5-35)
5-5 Linear Combinations of Random Variables

Variance of a Linear Combination

If \( X_1, X_2, \ldots, X_p \) are random variables, and \( Y = c_1X_1 + c_2X_2 + \cdots + c_pX_p \), then in general

\[
V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \cdots + c_p^2V(X_p) + 2 \sum_{i<j} c_ic_j \text{cov}(X_i, X_j) \quad (5-36)
\]

If \( X_1, X_2, \ldots, X_p \) are independent,

\[
V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \cdots + c_p^2V(X_p) \quad (5-37)
\]
5-5 Linear Combinations of Random Variables

Reproductive Property of the Normal Distribution

If $X_1, X_2, \ldots, X_p$ are independent, normal random variables with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$, for $i = 1, 2, \ldots, p$, 

$$Y = c_1X_1 + c_2X_2 + \cdots + c_pX_p$$

is a normal random variable with 

$$E(Y) = c_1\mu_1 + c_2\mu_2 + \cdots + c_p\mu_p$$

and 

$$V(Y) = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \cdots + c_p^2\sigma_p^2 \quad (5-39)$$
Error Analysis for Nonlinear Equations

Consider a general nonlinear equation

\[ Y = f(X_1, X_2, ..., X_p) \]  

(1)

In order to derive analytical expressions for \( E(Y) \) and \( V(Y) \), Eqn. (1) is linearized by performing a Taylor Series Expansion at \( X_i = \bar{X}_i \), and truncating after the first-order terms:

\[ Y = f(X_1, X_2, ..., X_p) \approx f(\bar{X}_1, \bar{X}_2, ..., \bar{X}_p) + \sum_{i=1}^{p} \left( \frac{\partial f}{\partial X_i} \right) \frac{X_i - \bar{X}_i}{\bar{X}} \]  

(2)

Note that Eqn. (2) is a linear with respect to the \( \{X_i\} \) because function \( f \) and its derivatives become constants when they are evaluated at specific numerical values, the sample means.

The following important results can be derived\(^1\)\(^3\):

\[ E(Y) \approx f(\bar{X}_1, \bar{X}_2, ..., \bar{X}_p) \]  

(3)

\[ V(Y) \approx \sum_{i=1}^{p} \left[ \left( \frac{\partial f}{\partial X_i} \right)^2 \frac{s_i^2}{\bar{X}} \right] \]  

(4)

where \( \bar{X}_i \) denote the sample mean of \( X_i \) and \( s_i^2 \) denotes its sample variance.
Example: Error Analysis for a Chemical Reaction Rate Expression

Consider a reaction rate expression for a nonisothermal, irreversible reaction, \( A \rightarrow B \):

\[
r_A \approx f(C_A, T) = k_0 e^{-E/RT} C_A
\]

where \( k_0 \), \( E \) and \( R \) are constants. Experiments for typical conditions have determined values for the sample means, \( \bar{T} \) and \( \bar{C}_A \), and the sample standard deviations, \( s_T \) and \( s_{C_A} \). Derive expressions for the approximate mean and standard deviation of the reaction rate.

**Solution:** The mean value of \( r_A \) can be determined from Eqn. (3),

\[
E(r_A) \approx k_0 e^{-E/RT} \bar{C}_A
\]

The partial derivatives required for Eqn. (4) are:

\[
\left( \frac{\partial r_A}{\partial C_A} \right)_{\bar{T}} = k_0 e^{-E/RT}
\]

\[
\left( \frac{\partial r_A}{\partial T} \right)_{\bar{C}_A} = \frac{E}{RT^2} k_0 e^{-E/RT} \bar{C}_A
\]
Substituting into Eqn. (4) gives:

\[ V(r_A) \approx \left( \frac{\partial r_A}{\partial T} \right)^2 s_T^2 + \left( \frac{\partial r_A}{\partial \bar{C}_A} \right)^2 s_{\bar{C}_A}^2 \]

\[ V(r_A) \approx \left( \frac{E}{R \bar{T}^2} k_0 e^{-E/RT} \bar{C}_A \right)^2 s_T^2 + \left( k_0 e^{-E/RT} \right)^2 s_{\bar{C}_A}^2 \]

The standard deviation is \( \sigma_{r_A} \approx \sqrt{V(r_A)} \)

References


3. ChE 132C lectures (January, 2010).
6 Random Sampling and Data Description

CHAPTER OUTLINE

6-1 NUMERICAL SUMMARIES
6-2 STEM-AND-LEAF DIAGRAMS
6-3 FREQUENCY DISTRIBUTIONS AND HISTOGRAMS
6-4 BOX PLOTS
6-5 TIME SEQUENCE PLOTS
6-6 PROBABILITY PLOTS
Chapter 6: Random Sampling and Data Descriptions

- Data summaries and informative displays are essential to statistical analysis.
- In this chapter, we consider some basic concepts and standard graphical displays.
- **Reading assignment:** Chapter 6, but omit Sections 6.2 and 6.5.
6-1 Data Summary and Display

Definition

If the $n$ observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample mean** is

$$
\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}
$$

(6-1)
6-1 Numerical Summaries

Population Mean

For a finite population with $N$ measurements, the population mean is

$$\mu = \sum_{i=1}^{N} x_i f(x_i) = \frac{\sum_{i=1}^{N} x_i}{N}$$  \hspace{2cm} (6-2)

The sample mean is a reasonable estimate of the population mean.
Population Variance

When the population is finite and consists of $N$ values, we may define the population variance as

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$  \hspace{1cm} (6-5)

The sample variance is a reasonable estimate of the population variance.
If \( x_1, x_2, \ldots, x_n \) is a sample of \( n \) observations, the sample variance is

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

The sample standard deviation, \( s \), is the positive square root of the sample variance.
In the definition of \( s^2 \), why divide by \( n-1 \), instead of \( n \)?

- **Reason #1:**
  From the definition of \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \), it follows that:

\[
0 = \sum_{i=1}^{n} (x_i - \bar{x})
\]

Thus, only \( n-1 \) of the \( \{x_i - \bar{x}\} \) are independent.

- **Reason #2:**
  A dataset consisting of \( n \) data points can be considered to contain \( n \) degrees of freedom. One degree of freedom is used to calculate \( \bar{x} \). Thus, \( n-1 \) remain.

- **Reason #3:** \( s^2 \) is an unbiased estimator of \( \sigma^2 \).
6-1 Numerical Summaries

Definition

If the $n$ observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample range is

$$r = \max(x_i) - \min(x_i)$$

(6-6)
6-2 Random Sampling

Definitions

A population consists of the totality of the observations with which we are concerned.

A sample is a subset of observations selected from a population.
6-2 Random Sampling

Definitions

The random variables $X_1, X_2, \ldots, X_n$ are a random sample of size $n$ if (a) the $X_i$'s are independent random variables, and (b) every $X_i$ has the same probability distribution.

A statistic is any function of the observations in a random sample.
6-4 Box Plots

• The **box plot** is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of observations that lie unusually far from the bulk of the data.

• Whisker
• Outlier
• Extreme outlier
Figure 6-13 Description of a box plot.
7-1 Introduction

• The field of statistical inference consists of those methods used to make decisions or to draw conclusions about a population.

• These methods utilize the information contained in a random sample from the population in drawing conclusions.

• Statistical inference may be divided into two major areas:
  • Parameter estimation
  • Hypothesis testing
7-1 Introduction

Suppose that we want to obtain a point estimate of a population parameter. We know that before the data is collected, the observations are considered to be random variables, say $X_1, X_2, \ldots, X_n$. Therefore, any function of the observation, or any statistic, is also a random variable. For example, the sample mean $\overline{X}$ and the sample variance $S^2$ are statistics and they are also random variables.

Since a statistic is a random variable, it has a probability distribution. We call the probability distribution of a statistic a sampling distribution. The notion of a sampling distribution is very important and will be discussed and illustrated later in the chapter.

**Definition**

A point estimate of some population parameter $\theta$ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$. The statistic $\hat{\Theta}$ is called the point estimator.
Estimation problems occur frequently in engineering. We often need to estimate

- The mean $\mu$ of a single population
- The variance $\sigma^2$ (or standard deviation $\sigma$) of a single population
- The proportion $p$ of items in a population that belong to a class of interest
- The difference in means of two populations, $\mu_1 - \mu_2$
- The difference in two population proportions, $p_1 - p_2$
7-1 Introduction

Reasonable point estimates of these parameters are as follows:

- For $\mu$, the estimate is $\hat{\mu} = \bar{x}$, the sample mean.
- For $\sigma^2$, the estimate is $\hat{\sigma}^2 = s^2$, the sample variance.
- For $p$, the estimate is $\hat{p} = x/n$, the sample proportion, where $x$ is the number of items in a random sample of size $n$ that belong to the class of interest.
- For $\mu_1 - \mu_2$, the estimate is $\hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$, the difference between the sample means of two independent random samples.
- For $p_1 - p_2$, the estimate is $\hat{p}_1 - \hat{p}_2$, the difference between two sample proportions computed from two independent random samples.
7.2 Sampling Distributions and the Central Limit Theorem

Statistical inference is concerned with making decisions about a population based on the information contained in a random sample from that population.

Definitions:

The random variables $X_1, X_2, \ldots, X_n$ are a random sample of size $n$ if (a) the $X_i$’s are independent random variables, and (b) every $X_i$ has the same probability distribution.

A statistic is any function of the observations in a random sample.

The probability distribution of a statistic is called a sampling distribution.
7.2 Sampling Distributions and the Central Limit Theorem

If we are sampling from a population that has an unknown probability distribution, the sampling distribution of the sample mean will still be approximately normal with mean \( \mu \) and variance \( \sigma^2/n \), if the sample size \( n \) is large. This is one of the most useful theorems in statistics, called the central limit theorem. The statement is as follows:

If \( X_1, X_2, \ldots, X_n \) is a random sample of size \( n \) taken from a population (either finite or infinite) with mean \( \mu \) and finite variance \( \sigma^2 \), and if \( \bar{X} \) is the sample mean, the limiting form of the distribution of

\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}
\]

as \( n \to \infty \), is the standard normal distribution.
Sampling Distributions and the Central Limit Theorem

Figure 7-1 Distributions of average scores from throwing dice.
7.2 Sampling Distributions and the Central Limit Theorem

Approximate Sampling Distribution of a Difference in Sample Means

If we have two independent populations with means \( \mu_1 \) and \( \mu_2 \) and variances \( \sigma_1^2 \) and \( \sigma_2^2 \) and if \( \bar{X}_1 \) and \( \bar{X}_2 \) are the sample means of two independent random samples of sizes \( n_1 \) and \( n_2 \) from these populations, then the sampling distribution of

\[
Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}
\]  

(7-4)

is approximately standard normal, if the conditions of the central limit theorem apply. If the two populations are normal, the sampling distribution of \( Z \) is exactly standard normal.
7-3 General Concepts of Point Estimation

7-3.1 Unbiased Estimators

**Definition**

The point estimator $\hat{\Theta}$ is an *unbiased estimator* for the parameter $\theta$ if

$$E(\hat{\Theta}) = \theta$$  \hspace{1cm} (7-5)

If the estimator is not unbiased, then the difference

$$E(\hat{\Theta}) - \theta$$ \hspace{1cm} (7-6)

is called the **bias** of the estimator $\hat{\Theta}$. 
Unbiased Point Estimates

The following point estimates are unbiased:

Reasonable point estimates of these parameters are as follows:

- For \( \mu \), the estimate is \( \hat{\mu} = \bar{x} \), the sample mean.
- For \( \sigma^2 \), the estimate is \( \hat{\sigma}^2 = s^2 \), the sample variance.
- For \( p \), the estimate is \( \hat{p} = x/n \), the sample proportion, where \( x \) is the number of items in a random sample of size \( n \) that belong to the class of interest.
- For \( \mu_1 - \mu_2 \), the estimate is \( \hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2 \), the difference between the sample means of two independent random samples.
- For \( p_1 - p_2 \), the estimate is \( \hat{p}_1 - \hat{p}_2 \), the difference between two sample proportions computed from two independent random samples.
7-3 General Concepts of Point Estimation

7-3.2 Variance of a Point Estimator

**Definition**

If we consider all unbiased estimators of \( \theta \), the one with the smallest variance is called the **minimum variance unbiased estimator** (MVUE).

**Figure 7-5** The sampling distributions of two unbiased estimators \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \).
8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Known

8-2.1 Development of the Confidence Interval and its Basic Properties

• The endpoints or bounds \( l \) and \( u \) are called lower- and upper-confidence limits, respectively.

• Since \( Z \) follows a standard normal distribution, we can write:

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

Now manipulate the quantities inside the brackets by (1) multiplying through by \( \sigma/\sqrt{n} \), (2) subtracting \( \bar{X} \) from each term, and (3) multiplying through by \(-1\). This results in

\[
P \left\{ \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\} = 1 - \alpha
\]  

(8-6)
Interpreting a Confidence Interval

- The confidence interval is a **random interval**

- The appropriate interpretation of a confidence interval (for example on $\mu$) is: The observed interval $[l, u]$ brackets the true value of $\mu$, with confidence $100(1-\alpha)$. 

---

8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Known
8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Known

8-2.2 Choice of Sample Size

Definition

If \( \bar{x} \) is used as an estimate of \( \mu \), we can be \( 100(1 - \alpha)\% \) confident that the error \( |\bar{x} - \mu| \) will not exceed a specified amount \( E \) when the sample size is

\[
n = \left( \frac{z_{\alpha/2} \sigma}{E} \right)^2
\]  

(8-8)
8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Known

8-2.3 One-Sided Confidence Bounds

Definition

A 100\((1 - \alpha)\)% upper-confidence bound for \(\mu\) is

\[
\mu \leq u = \bar{x} + z_\alpha \sigma / \sqrt{n}
\]  

(8-9)

and a 100\((1 - \alpha)\)% lower-confidence bound for \(\mu\) is

\[
\bar{x} - z_\alpha \sigma / \sqrt{n} = l \leq \mu
\]  

(8-10)
8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

8-3.1 The \( t \) distribution

Let \( X_1, X_2, \ldots, X_n \) be a random sample from a normal distribution with unknown mean \( \mu \) and unknown variance \( \sigma^2 \). The random variable

\[
T = \frac{\bar{X} - \mu}{S/\sqrt{n}}
\]

has a \( t \) distribution with \( n - 1 \) degrees of freedom.
8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

8-3.1 The $t$ distribution

**Figure 8-4** Probability density functions of several $t$ distributions.
8-3 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

8-3.2 The $t$ Confidence Interval on $\mu$

If $\bar{x}$ and $s$ are the mean and standard deviation of a random sample from a normal distribution with unknown variance $\sigma^2$, a $100(1 - \alpha)$ percent confidence interval on $\mu$ is given by

$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n} \quad (8-18)$$

where $t_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point of the $t$ distribution with $n - 1$ degrees of freedom.

**One-sided confidence bounds** on the mean are found by replacing $t_{\alpha/2,n-1}$ in Equation 8-18 with $t_{\alpha,n-1}$. 
### t-distribution Table

\((\nu = n - 1)\)

<table>
<thead>
<tr>
<th>(\nu)</th>
<th>(\alpha)</th>
<th>.40</th>
<th>.25</th>
<th>.10</th>
<th>.05</th>
<th>.025</th>
<th>.01</th>
<th>.005</th>
<th>.0025</th>
<th>.001</th>
<th>.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>.325</td>
<td>1.000</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>127.32</td>
<td>318.31</td>
<td>636.62</td>
</tr>
</tbody>
</table>
8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

**Definition**

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$, and let $S^2$ be the sample variance. Then the random variable

$$X^2 = \frac{(n - 1) S^2}{\sigma^2}$$

(8-19)

has a chi-square ($\chi^2$) distribution with $n - 1$ degrees of freedom.
Figure 8-8 Probability density functions of several $\chi^2$ distributions.
8-4 Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

**Definition**

If $s^2$ is the sample variance from a random sample of $n$ observations from a normal distribution with unknown variance $\sigma^2$, then a $100(1 - \alpha)\%$ confidence interval on $\sigma^2$ is

$$\frac{(n - 1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n - 1)s^2}{\chi^2_{1-\alpha/2,n-1}} \quad (8-21)$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n - 1$ degrees of freedom, respectively. A confidence interval for $\sigma$ has lower and upper limits that are the square roots of the corresponding limits in Equation 8-21.
One-Sided Confidence Bounds

The 100(1 − α)% lower and upper confidence bounds on $\sigma^2$ are

$$\frac{(n - 1)s^2}{\chi^2_{\alpha,n-1}} \leq \sigma^2 \quad \text{and} \quad \sigma^2 \leq \frac{(n - 1)s^2}{\chi^2_{1-\alpha,n-1}}$$

(8-22)

respectively.
Chi-Squared Distribution Table

$(\nu = n - 1)$

![Chi-Squared Distribution](image)

### Table III

Percentage Points $\chi^2_{\alpha, \nu}$ of the Chi-Squared Distribution

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>.995</th>
<th>.990</th>
<th>.975</th>
<th>.950</th>
<th>.900</th>
<th>.500</th>
<th>.100</th>
<th>.050</th>
<th>.025</th>
<th>.010</th>
<th>.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.00+</td>
<td>.00+</td>
<td>.00+</td>
<td>.00+</td>
<td>.02</td>
<td>.45</td>
<td>2.71</td>
<td>3.84</td>
<td>5.02</td>
<td>6.63</td>
<td>7.88</td>
</tr>
<tr>
<td>2</td>
<td>.01</td>
<td>.02</td>
<td>.05</td>
<td>.10</td>
<td>.21</td>
<td>1.39</td>
<td>4.61</td>
<td>5.99</td>
<td>7.38</td>
<td>9.21</td>
<td>10.60</td>
</tr>
<tr>
<td>3</td>
<td>.07</td>
<td>.11</td>
<td>.22</td>
<td>.35</td>
<td>.58</td>
<td>2.37</td>
<td>6.25</td>
<td>7.81</td>
<td>9.35</td>
<td>11.34</td>
<td>12.84</td>
</tr>
<tr>
<td>5</td>
<td>.41</td>
<td>.55</td>
<td>.83</td>
<td>1.15</td>
<td>1.61</td>
<td>4.35</td>
<td>9.24</td>
<td>11.07</td>
<td>12.83</td>
<td>15.09</td>
<td>16.75</td>
</tr>
<tr>
<td>6</td>
<td>.68</td>
<td>.87</td>
<td>1.24</td>
<td>1.64</td>
<td>2.20</td>
<td>5.35</td>
<td>10.65</td>
<td>12.59</td>
<td>14.45</td>
<td>16.81</td>
<td>18.55</td>
</tr>
</tbody>
</table>
Normal Approximation for Binomial Proportion

If $n$ is large, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

is approximately standard normal.

The quantity $\sqrt{p(1-p)/n}$ is called the standard error of the point estimator $\hat{P}$. 
If \( \hat{p} \) is the proportion of observations in a random sample of size \( n \) that belongs to a class of interest, an approximate 100\((1 - \alpha)\)% confidence interval on the proportion \( p \) of the population that belongs to this class is

\[
\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \tag{8-25}
\]

where \( z_{\alpha/2} \) is the upper \( \alpha/2 \) percentage point of the standard normal distribution.
8-5 A Large-Sample Confidence Interval For a Population Proportion

Choice of Sample Size

The sample size for a specified value $E$ is given by

$$ n = \left( \frac{z_{\alpha/2}}{E} \right)^2 p(1 - p) \quad (8-26) $$

An upper bound on $n$ is given by

$$ n = \left( \frac{z_{\alpha/2}}{E} \right)^2 (0.25) \quad (8-27) $$
8-5 A Large-Sample Confidence Interval For a Population Proportion

One-Sided Confidence Bounds

The approximate 100(1 − α)% lower and upper confidence bounds are

\[ \hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \quad \text{and} \quad p \leq \hat{p} + z_\alpha \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \]  

(8-28)

respectively.
Good luck!