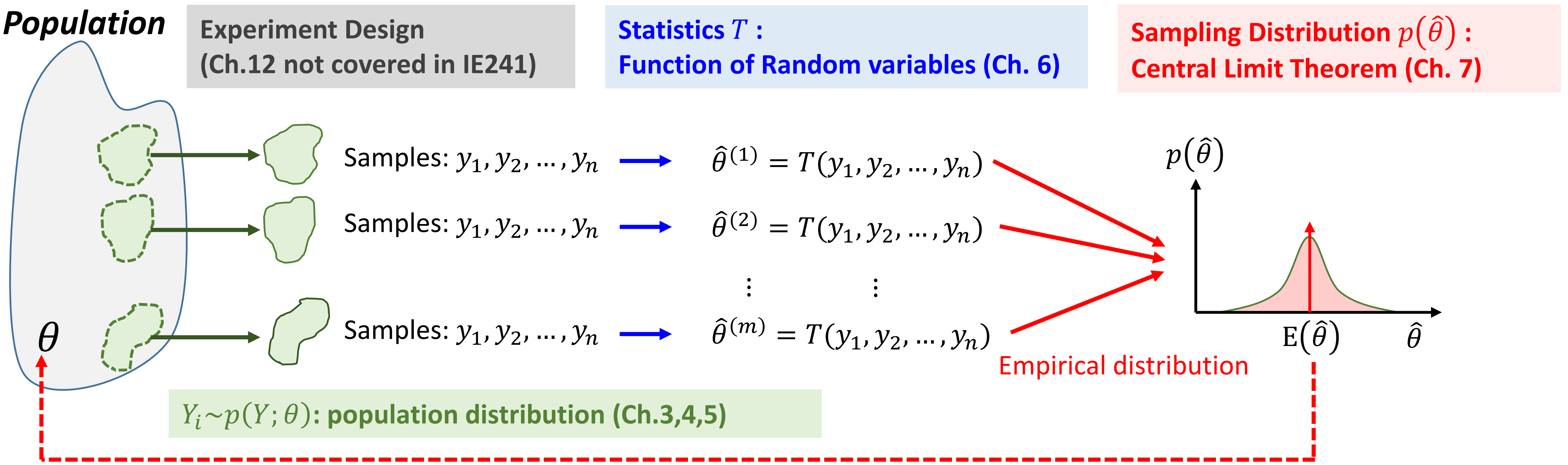
CHAPTER 7
Sampling Distributions and the Central Limit Theorem

Motivation

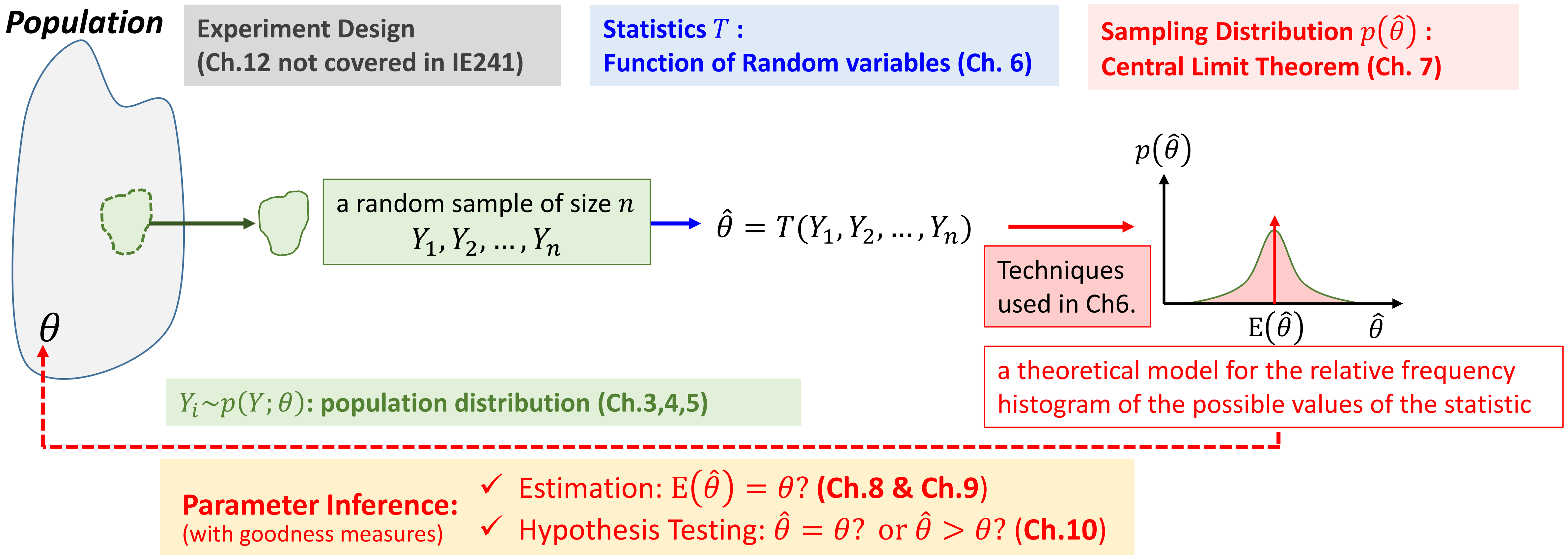


Parameter Inference:
(with goodness measures)

- ✓ Estimation: $\mathrm{E}(\hat{\theta}) = \theta$? **(Ch.8 & Ch.9)**
- ✓ Hypothesis Testing: $\hat{\theta} = \theta$? or $\hat{\theta} > \theta$? **(Ch.10)**

- **Probability Theory (Ch.2 $\sim$ Ch.6) plays an important role in inference** by computing the probability of the occurrence of the sample and connects the computed probability to the most probable target parameter.
- ***Estimator*** $\hat{\theta} = T(Y_1, Y_2, \dots, Y_n)$ for a target parameter θ is a function of the random variables observed in a sample and therefore itself is a random variable.
- **Sampling distribution** $p(\hat{\theta})$ can be used to evaluate the goodness of the ***estimator*** (confidence interval) and the errors (i.e., α and β errors) of ***hypothesis testing***.

Motivation



- **Probability Theory (Ch.2 ~ Ch.6) plays an important role in inference** by computing the probability of the occurrence of the sample and connects the computed probability to the most probable target parameter.
- ***Estimator*** $\hat{\theta} = T(Y_1, Y_2, \dots, Y_n)$ for a target parameter θ is a function of the random variables observed in a sample and therefore itself is a random variable.
- **Sampling distribution** $p(\hat{\theta})$ can be used to evaluate the goodness of the ***estimator*** (confidence interval) and the errors (i.e., α and β errors) of ***hypothesis testing***.

Definition

DEFINITION 7.1

A *statistic* is a function of the observable random variables in a sample and known constants.

We have encountered many statistics:

- the sample mean $\bar{Y}$,
- the sample variance S^2,
- $Y(n) = \max(Y_1, Y_2, \ldots, Y_n)$,
- $Y(1) = \min(Y_1, Y_2, \ldots, Y_n)$,
- the range $R = Y(n) - Y(1)$,
- The sample median,
- and so on

Example

EXAMPLE 7.1

A balanced die is tossed three times. Let Y_1, Y_2 , and Y_3denote the number of spots observed on the upper face for tosses 1, 2, and 3, respectively. Suppose we are interested in $\bar{Y} = (Y_1 + Y_2 + Y_3)/3$, the average number of spots observed in a sample of size 3. What are the mean, $\mu_{\bar{Y}}$,and standard deviation, $\sigma_{\bar{Y}}$, of Y? How can we find the sampling distribution of $\bar{Y}$?

Remind

EXAMPLE 5.27

Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$.
(These variables may denote the outcomes of n independent trials of an experiment.)
Define

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and show that $E(\bar{Y}) = \mu$ and $V(\bar{Y}) = \sigma^2/n$.

Example

SOLUTION 7.1

$\mu = E(Y_i) = 3.5$ and $\sigma^2 = V(Y_i) = 2.9167, i = 1,2,3.$

Since Y_1, Y_2 and Y_3 are independent random variables,

$E(Y) = \mu = 3.5, V(Y) = \frac{\sigma^2}{3} = \frac{2.9167}{3}\ .9722\ \ \sigma_Y = \sqrt{.9722} = .9860$

How can we derive the distribution of the random variable $\bar{Y}$?

The possible values of the random variable $W = Y_1 + Y_2 + Y_3$ are 3, 4, 5,...,18 and $\bar{Y} = W/3$. Because the die is balanced, each of the $6^3 = 216$ distinct values of the multivariate random variable (Y_1, Y_2, Y_3) are equally likely and $P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = p(y_1, y_2, y_3) = 1/216$

Therefore,

$P(\bar{Y} = 1) = P(W = 3) = p(1,1,1) = 1/216$

$P(\bar{Y} = 4/3) = P(W = 4) = p(1,1,2) + p(1,2,1) + p(2,1,1) = 3/216$

$P(\bar{Y} = 5/3) = P(W = 5) = p(1,1,3) + p(1,3,1) + p(3,1,1) + p(1,2,2) + p(2,1,2) + p(2,2,1) = 6/216 \ldots$

The probabilities $P(\bar{Y} = i/3), i = 7,8, \ldots, 18$ are obtained similarly.

Estimating a population mean μ

- The derivation of the sampling distribution of the random variable Y sketched in Example 7.1 utilizes ***the sample point approach*** that was introduced in Chapter 2.
  - ✓ Although it is not difficult to complete the calculations in Example 7.1 and give the exact sampling distribution for Y, the process is tedious.

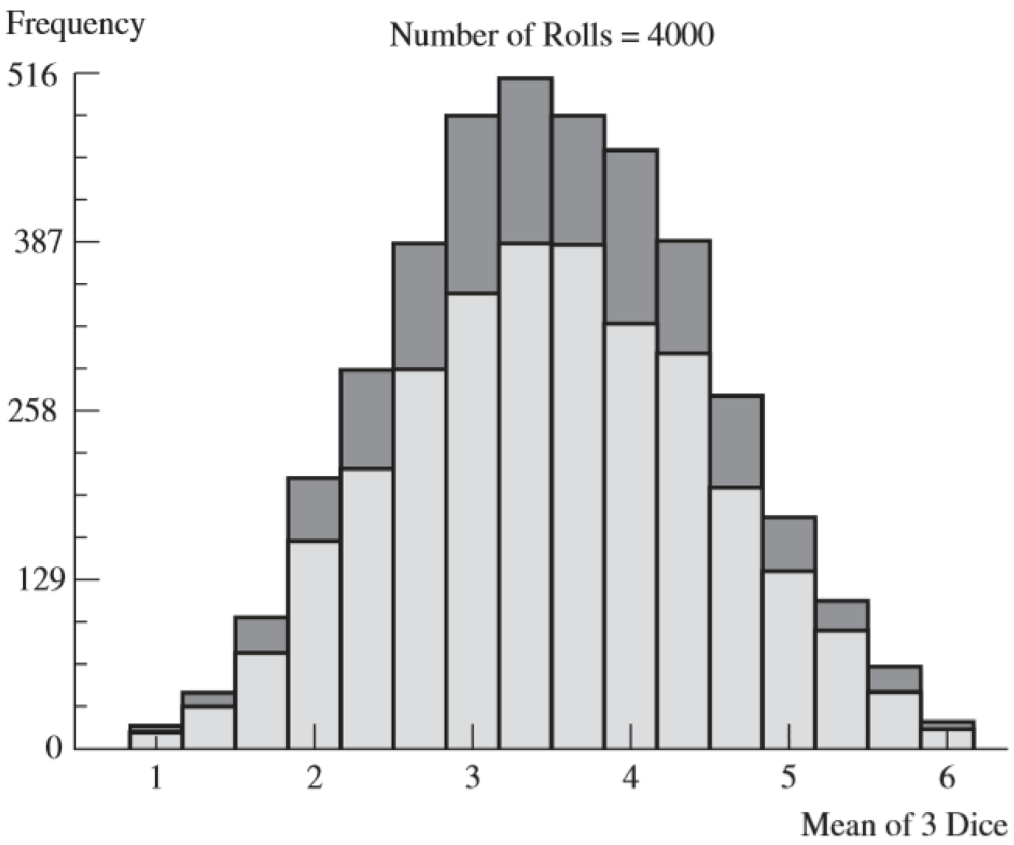
- How can we get an idea about the shape of this sampling distribution without going to the bother of completing these calculations?
  - One way is to simulate the sampling distribution by taking **repeated independent samples** each of size 3, computing the observed value y for each sample, and constructing a **histogram** of these observed values.

  - Another method is to employ techniques discussed in chapter 6 (**Functions of random variables**)

Estimating a population mean μ

Empirical distribution on the sample mean

- Simulate the sampling distribution by taking repeated independent samples each of size 3,
- Compute the observed value y for each sample,
- Construct a histogram of these observed values.



Pop Prob: (1) 0.167 (2) 0.167 (3) 0.167 (4) 0.167 (5) 0.167 (6) 0.167
Population: Mean = 3.500 StDev = 1.708
Samples = 4000 of size 3
Mean = 3.495
StDev = 0.981
+/– 1 StDev: 0.683
+/– 2 StDev: 0.962
+/– 3 StDev: 1.000

Motivation

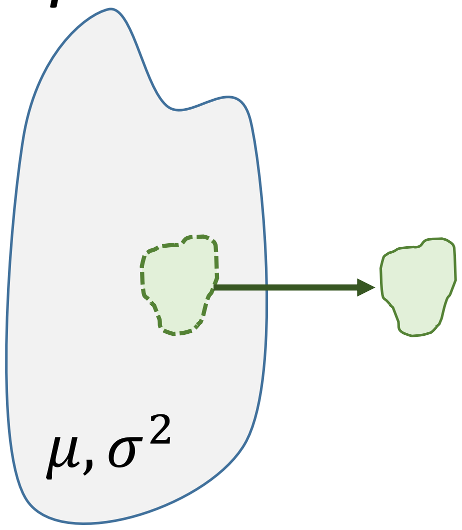
- We have already noted that many phenomena observed in the real world have relative frequency distributions that can be modeled adequately by a normal probability distribution.

  - Thus, in many applied problems, it is reasonable to assume that the observable random variables in a random sample, $Y_1, Y_2, \dots, Y_n$, are independent with the same normal density function.

  - In Exercise 6.43, you established that the statistic $\bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$ actually has a normal distribution

Motivation

Population



Statistics T :
Function of Random variables (Ch. 6)

Sampling Distribution $p(T)$:

$Y_i \sim N(\mu, \sigma^2)$

a random sample of size n
$Y_1, Y_2, \ldots, Y_n$

Sample mean

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Sample variance

(σ is assumed to be known)

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) = Z \sim N(0, 1^2)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2 \text{ with df} = n-1$$

$Z \sim N(0,1)$ $\quad W \sim \chi^2(v)$

(σ is assumed to be unknown)

$$\frac{Z}{\sqrt{W/v}} = \sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right) \sim t \qquad df = (n-1)$$

$$\frac{W_1/v_1}{W_2/v_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F \qquad \begin{array}{l} df_1 = (n_1 - 1) \\ df_2 = (n_2 - 1) \end{array}$$

Sample Mean

THEOREM 7.1

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from **a normal distribution** with mean μ and variance σ^2. Then

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is **normally distributed** with mean $\mu_{\bar{Y}} = \mu$ and variance $\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n}$.

Proof: Trivial by Theorem 6.3

μ will be retained for the mean of the random variables $Y_1, Y_2, \ldots, Y_n$

σ^2 will be retained for the variance of the random variables $Y_1, Y_2, \ldots, Y_n$

$\mu_{\bar{Y}}$ will be used to denote the mean of (the sampling distribution of) the random variable $\bar{Y}$

$\sigma^2_{\bar{Y}}$ will be used to denote the variance of (the sampling distribution of) the random variable $\bar{Y}$

Remind

THEOREM 6.3

Let $Y_1, Y_2, \ldots, Y_n$ be independent **normally distributed random variables** with $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma_i^2$, for all $i = 1,2, \ldots, n$, and let $a_1, a_2, \ldots, a_n$ be constants. If

$$U = \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n,$$

then ***U* is a normally distributed random variable** with

$$E(U) = \sum_{i=1}^{n} a_i \mu_i = a_1 \mu_1 + \cdots + a_n \mu_n$$

and

$$V(U) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 = a_1^2 \sigma_1^2 + \cdots + a_n^2 \sigma_n^2.$$

Sample Mean

THEOREM 7.1

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is normally distributed with mean $\mu_{\bar{Y}} = \mu$ and variance $\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}$.

Under the conditions of Theorem 7.1, Y is normally distributed with mean $\mu_Y = \mu$ and variance $\sigma_Y^2 = \frac{\sigma^2}{n}$.

It follows that

$$Z = \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right)$$

has a standard normal distribution. We will illustrate the use of Theorem 7.1 in the following example.

Example : Sample Mean

EXAMPLE 7.2

A bottling machine can be regulated so that it discharges an average of μ ounces per bottle. It has been observed that the amount of fill dispensed by the machine is normally distributed with $\sigma = 1.0$ ounce. A sample of $n = 9$ filled bottles is randomly selected from the output of the machine on a given day (all bottled with the same machine setting), and the ounces of fill are measured for each. Find the probability that the sample mean will be within .3 ounce of the true mean μ for the chosen machine setting.

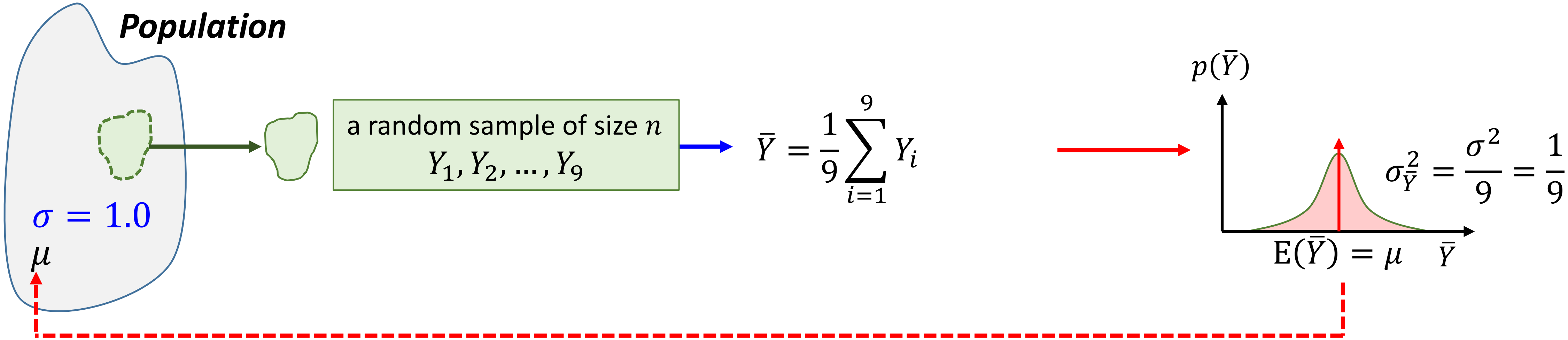
Example : Sample Mean

SOLUTION 7.2

If $Y_1, Y_2, \ldots, Y_9$ denote the ounces of fill to be observed, then we know that the Y_i 's are normally distributed with mean μ and variance $\sigma^2 = 1$ for $i = 1, 2, \ldots, 9$.

Therefore, by Theorem 7.1, Y possesses a normal sampling distribution with mean $\mu_{\bar{Y}} = \mu$ and variance $\sigma_{\bar{Y}}^2 = \sigma^2/n = 1/9$. We want to find

$$
\begin{aligned}
P(|\bar{Y} - \mu| \le 0.3) &= P(-0.3 \le \bar{Y} - \mu \le 0.3) \\
&= P\left(-\frac{0.3}{\sigma_{\bar{Y}}} \le \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} \le \frac{0.3}{\sigma_{\bar{Y}}}\right) \\
&= p(-0.9 \le Z \le 0.9) = 1 - 2(0.1841) = 0.6318
\end{aligned}
$$



Sum of Standardized Normal Random Variables

THEOREM 7.2

Let $Y_1, Y_2, \ldots, Y_n$ be defined as in Theorem 7.1. Then $Z_i = (Y_i - \mu)/\sigma$ are independent, standard normal random variables, $i = 1, 2, \ldots, n$, and

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2$$

has a χ^2 distribution with n degrees of freedom (df).

Proof:

Because $Y_1, Y_2, \ldots, Y_n$ is a random sample from a normal distribution with mean μ and variance σ^2, Example 6.10 implies that $Z_i = (Y_i - \mu)/\sigma$ has a standard normal distribution for $i = 1,2, \ldots, n$. Further, the random variables Z_i are independent. From **Theorem 6.4, we can have the result**.

Remind

THEOREM 6.4

Let $Y_1, Y_2, \ldots, Y_n$ be independent normally distributed random variables with $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma_i^2$, for all $i = 1,2, \ldots, n$, and define Z_i by

$$Z_i = \frac{Y_i - \mu_i}{\sigma_i}, \qquad i = 1,2, \ldots, n.$$

Then $\sum_{i=1}^{n} Z_i^2$ has a χ^2 distribution with n degrees of freedom.

Proof:

Note that Z_i is normally distributed with mean 0 and variance 1 by Example 6.10.

We have Z_i^2 is a χ^2-distributed random variable with 1 degree of freedom. Thus,

$$m_{Z_i^2}(t) = (1 - 2t)^{-1/2},$$

and from Theorem 6.2, with $V = \sum_{i=1}^{n} Z_i^2$,

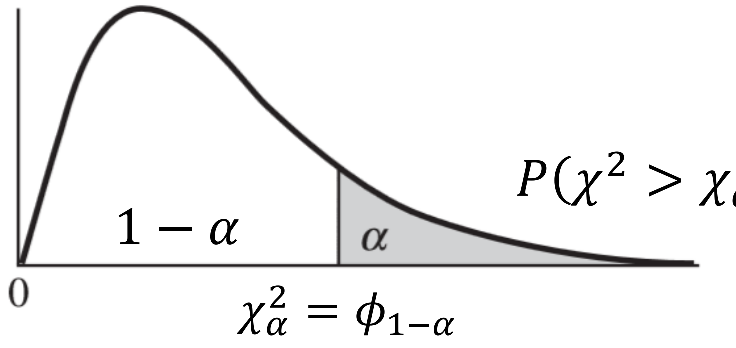
$$m_V(t) = \prod_{i=1}^{n} m_{Z_i^2}(t) = (1 - 2t)^{-n/2}.$$

Because moment-generating functions are unique, V has a χ^2 distribution with n degrees of freedom.

Sum of Standardized Normal Random Variables

- For random variables with χ^2 distribution, we can find

$P(\chi^2 \le \phi_{1-\alpha}) = 1 - \alpha$

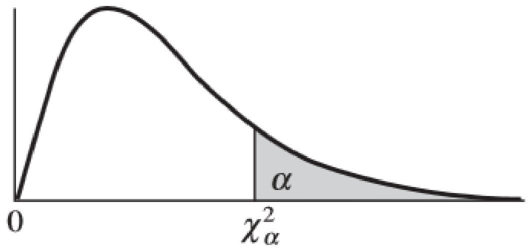


$P(\chi^2 > \chi^2_\alpha) = \alpha$

- For example, if the χ^2 random variable of interest has 10 df. Table 6, Appendix 3, can be used to find $\chi^2_{.90} = 4.96518$
  - ✓ That means, if Y has a χ^2 distribution with 10 df, $P(Y > 4.86518) = 0.9$
  - ✓ If follows that $P(Y \le 4.86518) = 1 - 0.9 = 0.10$
    - ➢ $\phi_{.10} = 4.86518$ (10 percentile)

Sum of Standardized Normal Random Variables

Table 6 Percentage Points of the χ^2 Distributions



df	$\chi^2_{0.995}$	$\chi^2_{0.990}$	$\chi^2_{0.975}$	$\chi^2_{0.950}$	$\chi^2_{0.900}$
1	0.0000393	0.0001571	0.0009821	0.0039321	0.0157908
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720
3	0.0717212	0.114832	0.215795	0.351846	0.584375
4	0.206990	0.297110	0.484419	0.710721	1.063623
5	0.411740	0.554300	0.831211	1.145476	1.61031
6	0.675727	0.872085	1.237347	1.63539	2.20413
7	0.989265	1.239043	1.68987	2.16735	2.83311
8	1.344419	1.646482	2.17973	2.73264	3.48954
9	1.734926	2.087912	2.70039	3.32511	4.16816
10	2.15585	2.55821	3.24697	3.94030	4.86518

Sum of Standardized Normal Random Variables

EXAMPLE 7.4

If $Z_1, Z_2, \ldots, Z_6$ denotes a random sample from the standard normal distribution, find a number b such that

$$P\left(\sum_{i=1}^{6} Z_i^2 \leq b\right) = 0.95$$

Find $\phi_{0.95}$ for $P(\chi^2 \leq \phi_{0.95}) = 0.95$

Find χ_α^2 for $P(\chi^2 > \chi_\alpha^2) = 1 - 0.95 = 0.05$

Sum of Standardized Normal Random Variables

SOLUTION 7.4

By Theorem 7.2,
$\sum_{i=1}^{6} Z_i^2$ has a χ^2 distribution with 6 df. Looking at Table 6, Appendix 3, in the row headed 6 df and the column headed $\chi^2_{0.05}$, we see the number 12.5916. Thus,

$$p\left(\sum_{i=1}^{6} Z_i^2 \geq 12.5916\right) = 0.05$$

or, equivalently,

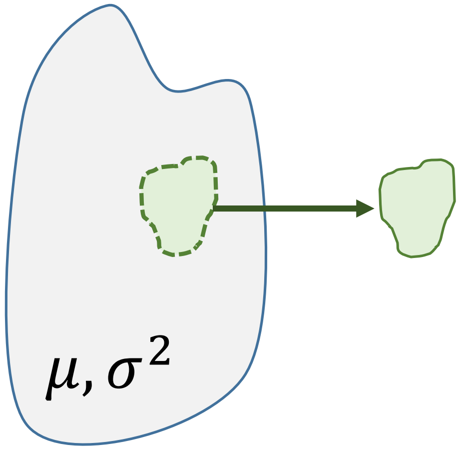
$$p\left(\sum_{i=1}^{6} Z_i^2 < 12.5916\right) = 0.95$$

and $b = 12.5916$ is the .95 quantile (95th percentile) of the sum of the squares of six independent standard normal random variables.

Where χ^2 distribution is used?

- The χ^2 distribution plays an important role in many inferential procedures.
- For example, suppose that we wish to make an inference about the population variance σ^2 based on a random sample $Y_1, Y_2, \dots, Y_n$ from a normal population.
- As we will show in Chapter 8, a good estimator of σ^2 is the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Population



$Y_i \sim N(\mu, \sigma^2)$

a random sample of size n
$Y_1, Y_2, \dots, Y_n$

Sample mean

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Sample variance

(σ is assumed to be known)

$$\sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right) = Z \sim N(0, 1^2)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2 \text{ with df} = n - 1$$

Sample Variance

THEOREM 7.3

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from **a normal distribution** with mean μ and variance σ^2. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

has a χ^2 distribution with $(n-1)$df. Also, $\bar{Y}$and S^2 are independent random variables.

Proof: omitted

Example : Sample Variance

EXAMPLE 7.5

In Example 7.2, the ounces of fill from the bottling machine are assumed to have a normal distribution with $\sigma^2 = 1$. Suppose that we plan to select a random sample of ten bottles and measure the amount of fill in each bottle. If these ten observations are used to calculate S^2, it might be useful to specify an interval of values that will include S^2 with a high probability. Find numbers b_1 and b_2 such that $P(b_1 \leq S_2 \leq b_2) = .90$.

Example : Sample Variance

SOLUTION 7.5

$$P(b_1 < S^2 < b_2) = P\left[\frac{(n-1)b_1}{\sigma^2} \le \frac{(n-1)S^2}{\sigma^2} \le \frac{(n-1)b_2}{\sigma^2}\right]$$

where $\sigma^2 = 1,\ \ (n-1)S^2 \sim \chi^2\ \ \mathrm{df} = n - 1 = 9$

Therefore, we can use Table 6, Appendix 3, to find two numbers a1 and a2 such that

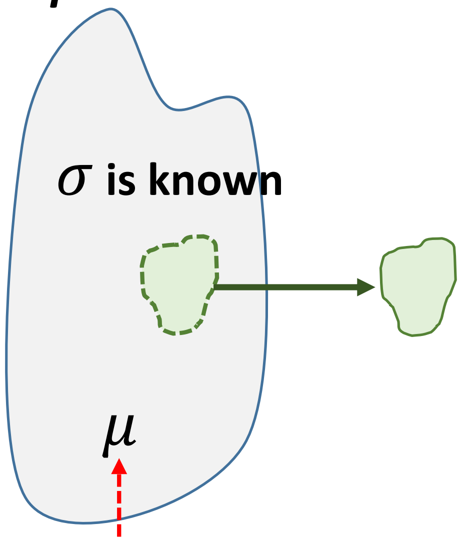
$$P[a_1 \le (n-1)S^2 \le a_2] = .90.$$

$$P(\chi^2 > \chi^2_\alpha) = 0.95 \rightarrow \chi^2_\alpha = a_1 = 3.325 \rightarrow b_1 = .369$$

$$P(\chi^2 > \chi^2_\alpha) = 0.05 \rightarrow \chi^2_\alpha = a_2 = 16.919 \rightarrow b_2 = 1.880$$

Motivation

Population



Statistics T :
Function of Random variables (Ch. 6)

$Y_i \sim N(\mu, \sigma^2)$

a random sample of size n
$Y_1, Y_2, \ldots, Y_n$

Sample mean

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

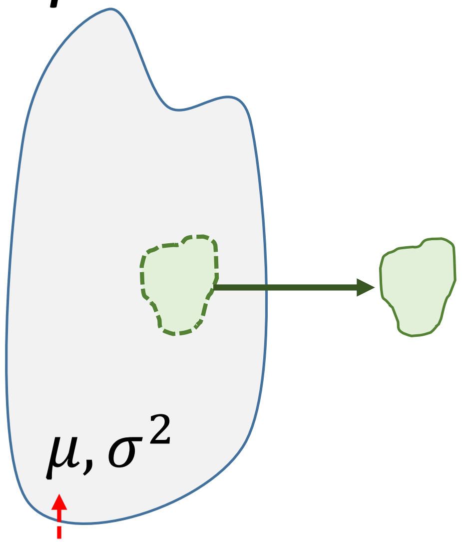
Sampling Distribution $p(T)$:

(σ is assumed to be known)

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) = Z \sim N(0, 1^2)$$

inference making procedures about the mean μ of a normal population with **known** variance σ^2.

Population



Statistics T :
Function of Random variables (Ch. 6)

$Y_i \sim N(\mu, \sigma^2)$

a random sample of size n
$Y_1, Y_2, \dots, Y_n$

Sample mean

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Sample variance

Sampling Distribution $p(T)$:

(σ is assumed to be known)

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) = Z \sim N(0, 1^2)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2 \text{ with df} = n-1$$

$Z \sim N(0,1)$ $\quad W \sim \chi^2(v)$

(σ is assumed to be unknown)

$$\frac{Z}{\sqrt{W/v}} = \sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right) \sim t \qquad df = (n-1)$$

$$\frac{W_1/v_1}{W_2/v_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F \qquad df_1 = (n_1 - 1),\ df_2 = (n_2 - 1)$$

inference making procedures about
the mean μ of a normal population with **unknown** variance σ^2

Definition : *t* distribution

DEFINITION 7.2

Let Z be a standard normal random variable and let W be a χ^2-distributed variable with ν df. Then, if Z and W are independent,

$$T = \frac{Z}{\sqrt{W/\nu}}$$

is said to have a t *distribution* with ν df.

- The general definition of a random variable that possesses a Student's t distribution (or simply a t distribution).

$$Z = \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right) \sim N(0, 1^2)$$

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2 \text{ df } \nu = n - 1$$

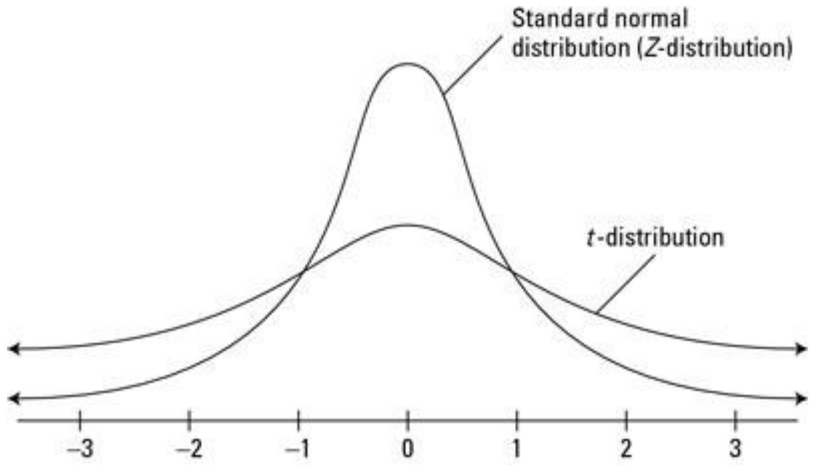
Z and W are independent (because Y and S2 are independent)

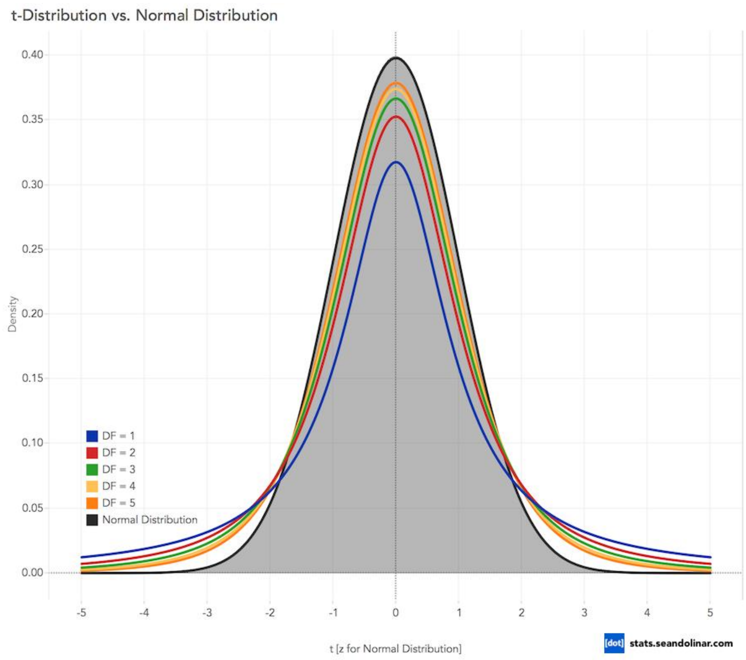
$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}(\bar{Y} - \mu)/\sigma}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}} = \sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right) \text{ has } t \text{ distribution with } (n-1)\text{df}$$

t distribution

- Like the standard normal density function, the t density function is symmetric about zero.

- Further, for $\nu > 1, E(T) = 0$; and for $\nu > 2, V(T) = \nu/(\nu - 2)$.
  - ✓ Thus, we see that, if $\nu > 1$, a t-distributed random variable has the same expected value as a standard normal random variable.
  - ✓ However, a standard normal random variable always has variance 1 whereas, if $\nu > 2$, the variance of a random variable with a t distribution always exceeds 1.

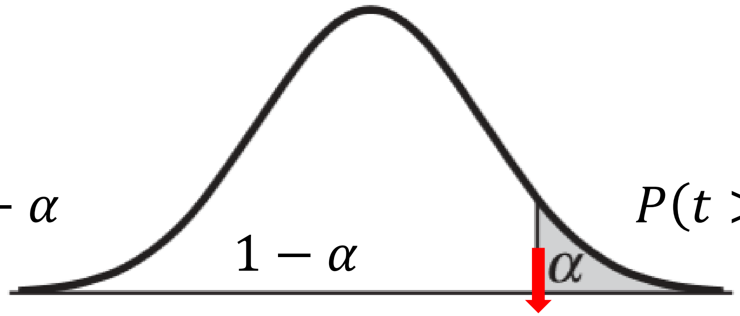




t distribution

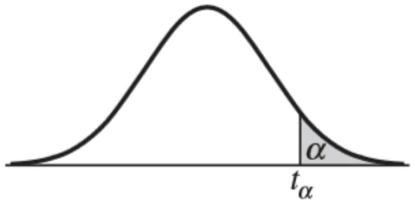
- For random variables with t distribution, we can find

$$P(t \leq \phi_{1-\alpha}) = 1 - \alpha \qquad P(t > t_\alpha) = \alpha$$



$$t_\alpha = \phi_{1-\alpha}$$

Table 5 Percentage Points of the *t* Distributions



$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
3.078	6.314	12.706	31.821	63.657	1
1.886	2.920	4.303	6.965	9.925	2
1.638	2.353	3.182	4.541	5.841	3
1.533	2.132	2.776	3.747	4.604	4

Example

EXAMPLE 7.6

The tensile strength for a type of wire is normally distributed with unknown mean μ and unknown variance σ^2. Six pieces of wire were randomly selected from a large roll; Y_i , the tensile strength for portion i, is measured for $i = 1, 2, \ldots, 6$. The population mean μ and variance σ^2 can be estimated by $\overline{Y}$ and S^2, respectively. Because $\sigma_{\overline{Y}}^2 = \sigma^2/n$, it follows that $\sigma_{\overline{Y}}^2$ can be estimated by S^2/n. Find the approximate probability that $\overline{Y}$ will be within $2S/\sqrt{n}$ of the true population mean μ.

Example

SOLUTION 7.6

$$P\left[-\frac{2S}{\sqrt{n}} \leq (\bar{Y} - \mu) \leq \frac{2S}{\sqrt{n}}\right] = P\left[-2 \leq \sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right) \leq 2\right]$$

$$= P[-2 \leq T \leq 2]$$

where $T \sim t(n - 1 = 5)$. Looking at Table 5, Appendix 3, we see that the upper-tail area to the right of 2.015 is .05. Hence,

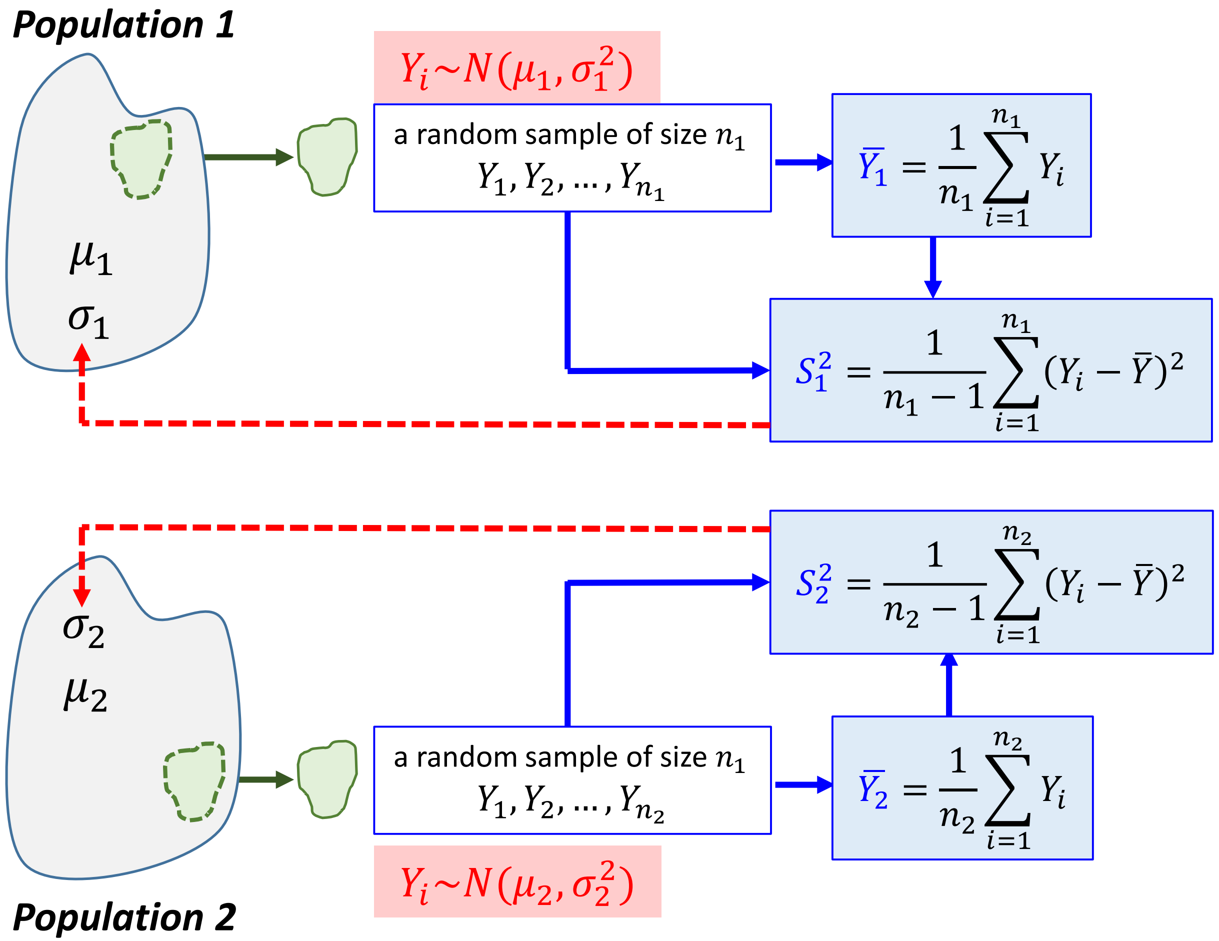
$$P(-2.015 \leq T \leq 2.015) = .90$$

Notice that, if σ^2 were known, the probability $\bar{Y}$ will fall within $2\sigma_{\bar{Y}}$ of μ would be given by

$$P\left[-2\left(\frac{\sigma}{\sqrt{n}}\right) \leq (\bar{Y} - \mu) \leq 2\left(\frac{\sigma}{\sqrt{n}}\right)\right] = P\left[-2 \leq \sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right) \leq 2\right]$$

$$= P(-2 \leq Z \leq 2) = 0.9544$$

Comparing Variance



$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \left(\frac{\sigma_2^2}{\sigma_1^2}\right)\left(\frac{S_1^2}{S_2^2}\right) \sim F$$

with $(n_1 - 1)$ df for numerator
$(n_2 - 1)$ df for denominator

- Thus, it seems intuitive that the ratio S_1^2/S_2^2 could be used to make inferences about the relative magnitudes of σ_1^2 and σ_2^2.

Definition : F distribution

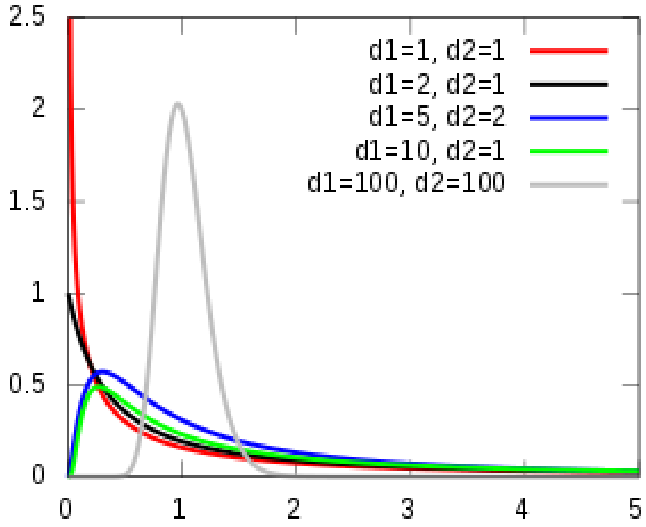
DEFINITION 7.3

Let W_1 and W_2 be *independent* χ^2-distributed random variables with ν_1 and ν_2 df, respectively. Then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

is said to have an F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom.

- F possesses an F distribution with ν_1 numerator and ν_2 denominator degrees of freedom, then
  - ✓ $E(F) = \nu_2/(\nu_2 - 2)$ if $\nu_2 > 2$.
  - ✓ Also, if $\nu_2 > 4$, then $V(F) = [2\nu_2^2(\nu_1 + \nu_2 - 2)]/[\nu_1(\nu_2 - 2)^2(\nu_2 - 4)]$.
  - ✓ Notice that the mean of an F distributed random variable depends only on the number of denominator degrees of freedom, ν_2.



Definition : F distribution

- Considering once again two independent random samples from normal distributions, we know that

  ✓ $W_1 = (n_1 - 1)S_1^2/\sigma_1^2$ have independent $\chi 2$ distributions with $\nu_1 \ = \ (n_1 - 1)$ df

  ✓ $W_2 = (n_2 - 1)S_2^2/\sigma_2^2$ have independent $\chi 2$ distributions with $\nu_2 \ = \ (n_2 - 1)$ df

- Thus, Definition 7.3 implies that

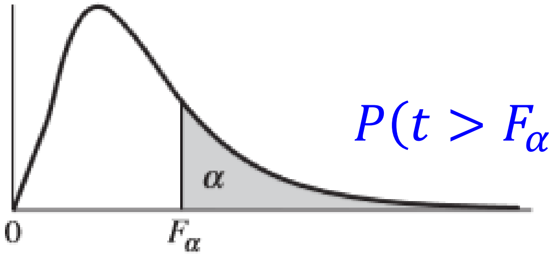
$$F = \frac{W_1/\nu_1}{W_2/\nu_2} = \frac{[(n_1 - 1)S_1^2/\sigma_1^2]/(n_1 - 1)}{[(n_2 - 1)S_2^2/\sigma_2^2]/(n_2 - 1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F distribution with $(n_1 - 1)$ numerator degrees of freedom and $(n_2 - 1)$ denominator degrees of freedom.

F distribution

Table 7 Percentage Points of the F Distributions

$P(F \leq \phi_{1-\alpha}) = 1 - \alpha$



$P(t > F_\alpha) = \alpha$

Denominator df	α	1	2	3	4	5	6	7	8	9
		Numerator df								
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
	.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3
	.010	4052	4999.5	5403	5625	5764	5859	5928	5982	6022
	.005	16211	20000	21615	22500	23056	23437	23715	23925	24091
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4

Example

EXAMPLE 7.7

If we take independent samples of size $n_1 = 6$ and $n_2 = 10$ from two normal populations with equal population variances, find the number b such that

$$P\left(\frac{S_1^2}{S_2^2} \leq b\right) = 0.95$$

Example

SOLUTION 7.7

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \sim F(\nu_1 = n_1 - 1 = 5, \nu_2 = n_2 - 1 = 9)$$

$$P\left(\frac{S_1^2}{S_2^2} \leq b\right) = 1 - P\left(\frac{S_1^2}{S_2^2} > b\right) = 0.95$$

$$\Rightarrow P\left(\frac{S_1^2}{S_2^2} > b\right) = 0.05$$

Therefore, we want to find the number b cutting off an upper-tail area of .05 under the F density function with 5 numerator degrees of freedom and 9 denominator degrees of freedom. Looking in column 5 and row 9 in Table 7, Appendix 3, we see that the appropriate value of b is 3.48.

$$\therefore b = 3.48$$

Motivation : Central limit theorem

THEOREM 7.1

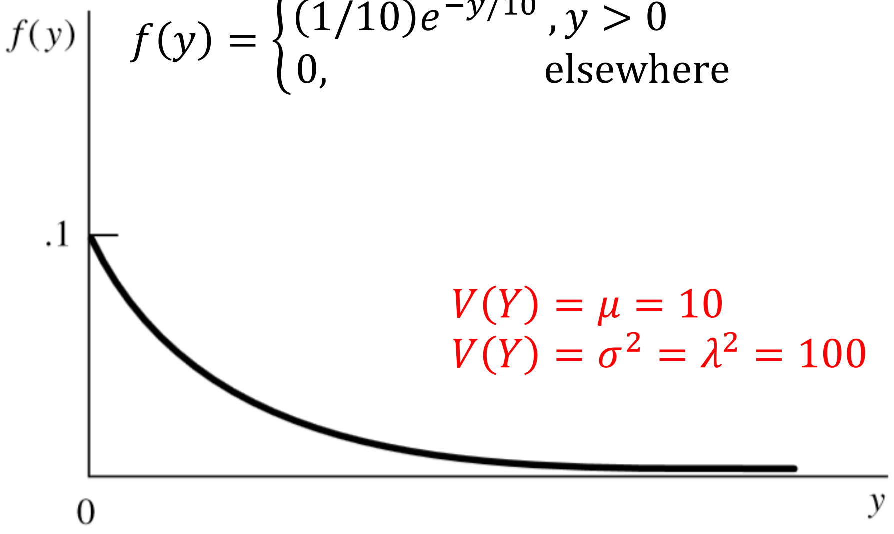
Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size n **from a normal distribution** with mean μ and variance σ^2. Then

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is **normally distributed** with mean $\mu_{\bar{Y}} = \mu$ and variance $\sigma_{\bar{Y}}^2 = \frac{\sigma^2}{n}$.

- But what can we say about the sampling distribution of $\bar{Y}$ if the variables Y_i are not normally distributed?
- Fortunately, $\bar{Y}$ will have a sampling distribution that is approximately normal if the sample size is large.
  - ➢ The formal statement of this result is **called the central limit theorem**

Motivation : Central limit theorem

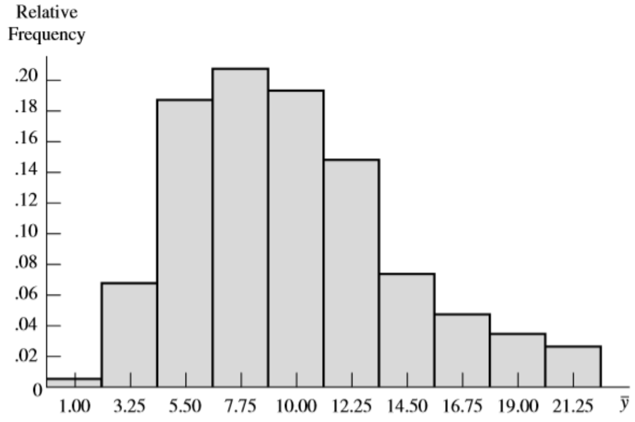


$$f(y) = \begin{cases} (1/10)e^{-y/10}, & y > 0 \\ 0, & \text{elsewhere} \end{cases}$$

$V(Y) = \mu = 10$

$V(Y) = \sigma^2 = \lambda^2 = 100$

a random sample of size n
$Y_1, Y_2, \dots, Y_n$ $\rightarrow$ $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$



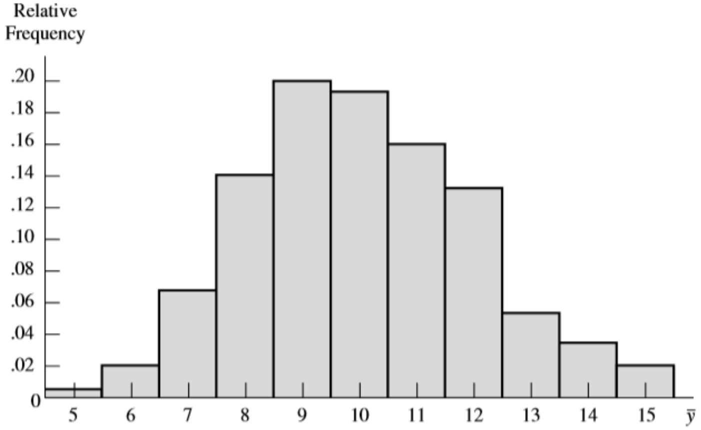


Table 7.1 Calculations for 1000 sample means

Sample Size	Average of 1000 Sample Means	$\mu_{\overline{Y}} = \mu$	Variance of 1000 Sample Means	$\sigma^2_{\overline{Y}} = \sigma^2/n$
$n = 5$	9.86	10	19.63	20
$n = 25$	9.95	10	3.93	4

Central Limit Theorem

THEOREM 7.4 (Central Limit Theorem)

Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed random variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2 < \infty$. Define

$$U_n = \frac{(\sum_{i=1}^{n} Y_i) - n\mu}{\sigma\sqrt{n}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then the distribution function of U_n converges to the standard normal distribution function as $n \to \infty$. That is,

$$\lim_{n\to\infty} P(U_n \leq u) = \int_{-\infty}^{\mathrm{u}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

for all u.

Example: Central Limit Theorem

EXAMPLE 7.8

Achievement test scores of all high school seniors in a state have mean 60 and variance 64. A random sample of $n = 100$ students from one large high school had a mean score of 58. Is there evidence to suggest that this high school is inferior? (Calculate the probability that the sample mean is at most 58 when $n = 100$.)

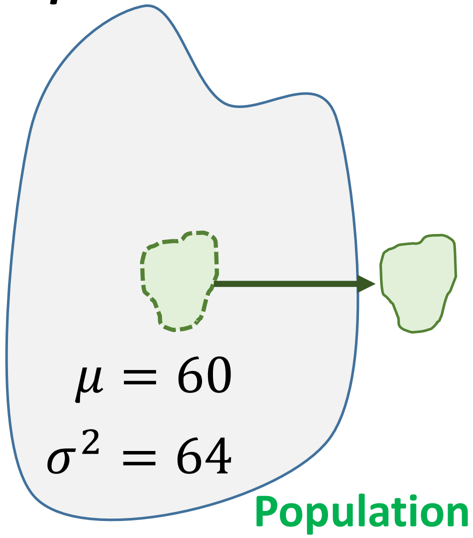
Example: Central Limit Theorem

SOLUTION 7.8

$$P(\bar{Y} \le 58) = P\left(\frac{\bar{Y} - 60}{8/\sqrt{100}} \le \frac{58 - 60}{8/\sqrt{100}}\right) \approx P(Z \le -2.5) = .0062$$

- Because this probability is so small, it is unlikely that the sample from the school of interest can be regarded as a random sample from a population with $\mu = 60$ and $\sigma^2 = 64$.
- The evidence suggests that the average score for this high school is lower than the overall average of $\mu = 60$.

Population



Population distribution

a random sample of size 100
$Y_1, Y_2, \ldots, Y_{100}$

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{Y} - 60}{8/\sqrt{10}} \sim N(0, 1^2)$$

Example: Central Limit Theorem

EXAMPLE 7.9

The service times for customers coming through a checkout counter in a retail store are independent random variables with mean 1.5 minutes and variance 1.0. Approximate the probability that 100 customers can be served in less than 2 hours of total service time.

Example: Central Limit Theorem

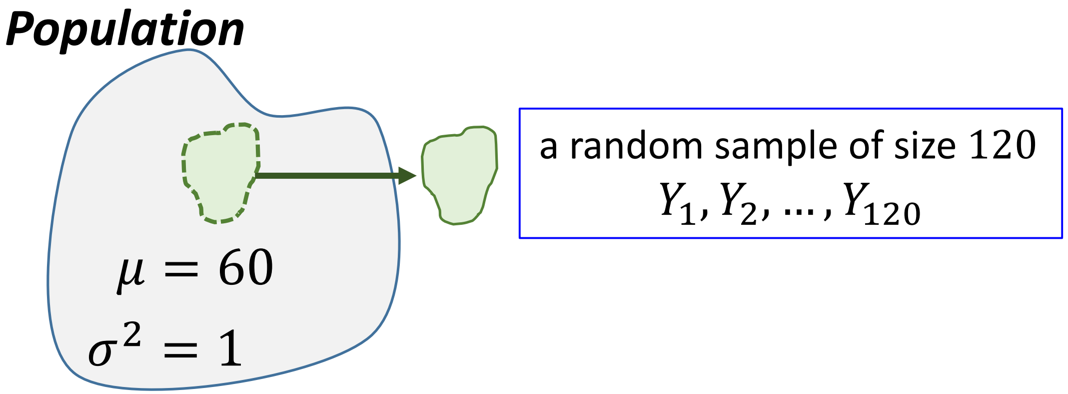
SOLUTION 7.9

Let Y_i the service time for the ith customer.

$$P\left(\sum_{i=1}^{100} Y_i \leq 120\right) = P\left(\bar{Y} \leq \frac{120}{100}\right) = P(\bar{Y} \leq 1.20)$$

Since the sample size is large, $\bar{Y}$ follows approximately normally distribution.

$$P(\bar{Y} < 1.20) = P\left(\frac{\bar{Y} - 1.50}{\frac{1}{\sqrt{100}}} \leq \frac{1.20 - 1.50}{\frac{1}{\sqrt{100}}}\right) \approx P(Z \leq -3) = 0.0013$$



$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{Y} - 1.5}{1/\sqrt{100}} \sim N(0,1^2)$$

Proof

THEOREM 7.5

Let Y and $Y_1, Y_2, Y_3, \ldots$ be random variables with moment-generating functions $m(t)$ and $m_1(t), m_2(t), m_3(t), \ldots$, respectively. If

$$\lim_{n\to\infty} m_n(t) = m(t) \text{ for all real } t,$$

then the distribution function of Y_n converges to the distribution function of Y as $n \to \infty$.

- Moment generation function converges→ PDF convergence

Proof

THEOREM 7.4 (Central Limit Theorem)

Let $Y_1, Y_2, \dots, Y_n$ be independent and identically distributed random variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2 < \infty$. Define

$$U_n = \frac{(\sum_{i=1}^{n} Y_i) - n\mu}{\sigma\sqrt{n}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then the distribution function of U_n converges to the standard normal distribution function as $n \to \infty$. That is,

$$\lim_{n\to\infty} P(U_n \le u) = \int_{-\infty}^{\mathrm{u}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

for all u.

We will sketch a proof of the central limit theorem for the case in which the moment generating functions exist for the random variables in the sample

Proof

Proof:

$$U_n = \sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) = \sqrt{n}\frac{1}{n}\left(\frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma}\right) = \sqrt{n}\frac{1}{n}\left(\frac{\sum_{i=1}^{n}(Y_i-\mu)}{\sigma}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \quad \text{where } Z_i = \frac{Y_i-\mu}{\sigma}$$

Because Y_i are i.i.d, Z_i are also i.i.d with $E(Z_i) = 0, V(Z_i) = 1$.

Since the moment-generating function of the sum of independent random variables is the product of their individual moment-generating functions,

$$m_{\Sigma Z_i}(t) = m_{\Sigma Z_1}(t) \times m_{\Sigma Z_2}(t) \times \cdots \times m_{\Sigma Z_n}(t) = \left[m_{\Sigma Z_i}(t)\right]^n$$

$$m_{U_n}(t) = m_{\frac{\Sigma Z_i}{\sqrt{n}}}(t) \left[m_{Z_i}\left(\frac{t}{\sqrt{n}}\right)\right]^n \qquad m_{X/n}(t) = \mathrm{E}\left(\exp\left(\frac{X}{n}t\right)\right) = \mathrm{E}\left(\exp\left(X\frac{t}{n}\right)\right) = m_X\left(\frac{t}{n}\right)$$

By Taylor's theorem, $m_{Z_i}(t) = m_{Z_i}(0) + \mathrm{m}'_{\mathrm{Z_i}}(0)t + m''_{Z_i}(\xi)\frac{t^2}{2}$ where $0 < \xi < t$,

$$\Rightarrow m_{Z_i}(t) = 1 + m''_{Z_i}(\xi)\frac{t^2}{2}, \quad \text{where } 0 < \xi < t,$$

$$m_{Z_i}(0) = E(e^{0Z_i}) = E(1) = 1, m'_{Z_i}(0) = E(Z_i) = 0$$

Proof

Proof:

$$m_{U_n}(t) = \left[1 + \frac{m''_{Z_1}(\xi_n)}{2}\left(\frac{t}{\sqrt{n}}\right)^2\right]^n = \left[1 + \frac{m''_{Z_1}(\xi_n)t^2/2}{n}\right]^n, \text{ where } 0 < \xi_n < \frac{t}{\sqrt{n}}$$

As $n \to \infty, \xi_n \to 0$ and $\dfrac{m''_{Z_1}(\xi_n)t^2}{2} \to \dfrac{m''_{Z_1}(0)t^2}{2} = \dfrac{E(Z_1^2)t^2}{2} = \dfrac{t^2}{2}$ because $\mathrm{E}(\mathrm{Z}_1^2) = \mathrm{V}(\mathrm{Z}_1) = 1$

Finally,

$$\lim_{n\to\infty} m_{U_n}(t) = \lim_{n\to\infty}\left[1 + \frac{m''_{Z_1}(\xi_n)t^2/2}{n}\right]^n = e^{t^2/2}$$

Because $\lim\limits_{n\to\infty} b_n = b$ then $\lim\limits_{n\to\infty}\left(1 + \frac{b_n}{n}\right)^n = e^b$

the moment-generating function for a standard normal random variable.

Applying Theorem 7.5, we conclude that U_n has a distribution function that converges to the distribution function of the standard normal random variable.

Motivation

- The central limit theorem also can be used to approximate probabilities for some discrete random variables when the exact probabilities are tedious to calculate.
- One useful example involves the binomial distribution for large values of the number of trials n.
- Suppose that Y has a binomial distribution with n trials and probability of success on any one trial denoted by p.
- If we want to find $P(Y \leq b)$, we can use the binomial probability function to compute $P(Y = y)$ for each nonnegative integer y less than or equal to b and then sum these probabilities.
  - ➢ Such direct calculation is cumbersome for large values of n for which tables may be unavailable.

Approximation

- we can view Y, the number of successes in n trials, as a sum of a sample consisting of 0s and 1s; that is,

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

where

$$X_i = \begin{cases} 1, & \text{if the ith trial results in success} \\ 0, & \text{otherwise} \end{cases}$$

- The random variables X_i for $i = 1, 2, \ldots, n$ are independent (because the trials are independent), and it is easy to show that $E(X_i) = p$ and $V(X_i) = p(1 - p)$ for $i = 1, 2, \ldots, n$. Consequently, when n is large, the sample fraction of successes,

$$\frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

possesses an approximately normal sampling distribution with mean $E(X_i) = p$ and variance $\frac{V(X_i)}{n} = \frac{p(1-p)}{n}$

Example

EXAMPLE 7.10

Candidate A believes that she can win a city election if she can earn at least 55% of the votes in precinct 1. She also believes that about 50% of the city's voters favor her. If $n = 100$ voters show up to vote at precinct1, what is the probability that candidate A will receive at least 55% of their votes?

Example

SOLUTION 7.10

Because it is reasonable to assume that $X_i, i = 1, 2, \ldots, n$ are independent, the central limit theorem implies that $X = Y/n$ is approximately normally distributed with mean $p = .5$ and variance $pq/n = (.5)(.5)/100 = .0025$. Therefore,

$$P\left(\frac{Y}{n} \geq .55\right) = P\left(\frac{\frac{Y}{n} - 0.5}{\sqrt{0.0025}} \geq \frac{.55 - .50}{0.05}\right) \approx P(Z \geq 1) = .1587$$