Chapter 5. Joint Probability Distributions and Random Sample

Chapter 5: Joint Probability Distributions and Random Sample

- **5.1. Jointly Distributed Random Variables**
- **5.2. Expected Values, Covariance, and Correlation**
- **5. 3. Statistics and Their Distributions**
- **5.4. The Distribution of the Sample Mean**
- **5.5. The Distribution of a Linear Combination**

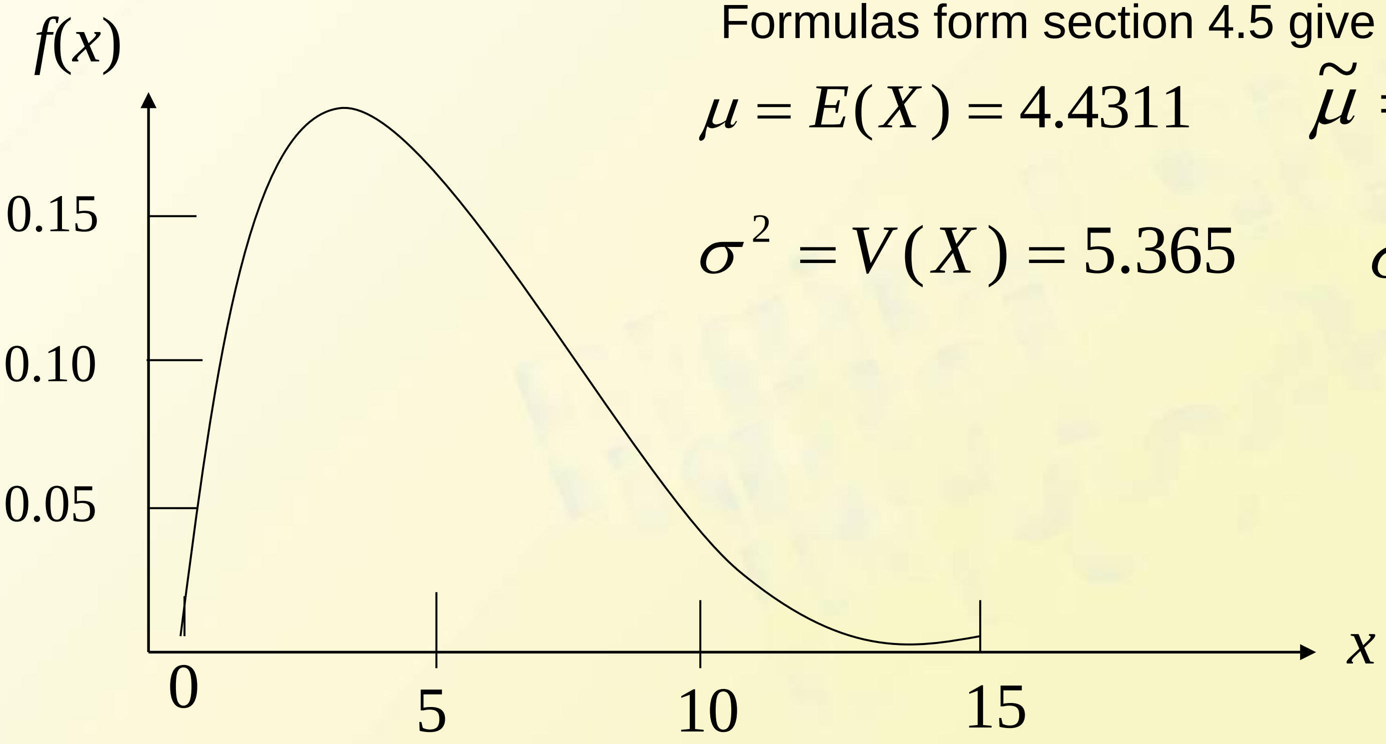
5.3 Statistics and Their Distributions

From this section, we consider function of *n* random variables $X_1, X_2, \ldots, X_n$ focusing especially on their average (X1,X2, …, Xn)/n. we call any such function, itself a random variable, a **statistic.**

As before, we studied some statistics: sample mean, sample standard deviation or sample fourth spread----also varies from sample to sample.

Example 5.19

Given a Weibull Population with α=2, β=5. The corresponding density curve is shown in Fig.5.6.



Formulas form section 4.5 give

$$\mu = E(X) = 4.4311 \qquad \tilde{\mu} = 4.1628$$

$$\sigma^2 = V(X) = 5.365 \qquad \sigma = 2.316$$

In section 4.5 we studied the Weibull Distribution

- **The Weibull Distribution**

  A random variable X is said to have a Weibull distribution with parameters α and β ($\alpha > 0, \beta > 0$) if the cdf of X is

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-(x/\beta)^{\alpha}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

When α =1, the pdf reduces to the exponential distribution (with λ =1/ β), so the exponential Distribution is a special case of both the gamma and Wellbull distributions.

5.3 Statistics and Their Distributions

▪The Weibull Distribution

- Mean and Variance

$$\mu = \beta\Gamma\left(1+\frac{1}{\alpha}\right); \quad \sigma^2 = \beta^2\left\{\Gamma\left(1+\frac{2}{\alpha}\right)-\left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^2\right\}$$

- The cdf of a Weibull Distribution

$$F(x;\alpha,\beta)=\begin{cases} 0 & x<0 \\ 1-e^{-(x/\beta)^{\alpha}} & x\geq 0 \end{cases}$$

5.3 Statistics and Their Distributions

- **Example 5.19 (Cont')**

We used MINITAB to generate six different samples, each with *n=10*.

Sample	1	2	3	4	5	6
1	6.1171	5.07611	3.46710	1.55601	3.12372	8.93795
2	4.1600	6.79279	2.71938	4.56941	6.09685	3.92487
3	3.1950	4.43259	5.88129	4.79870	3.41181	8.76202
4	0.6694	8.55752	5.14915	2.49795	1.65409	7.05569
5	1.8552	6.82487	4.99635	2.33267	2.29512	2.30932
6	5.2316	7.39958	5.86887	4.01295	2.12583	5.94195
7	2.7609	2.14755	6.05918	9.08845	3.20938	6.74166
8	10.2185	8.50628	1.80119	3.25728	3.23209	1.75486
9	5.2438	5.49510	4.21994	3.70132	6.84426	4.91827
10	4.5590	4.04525	2.12934	5.50134	4.20694	7.26081

Example 5.19 (Cont')

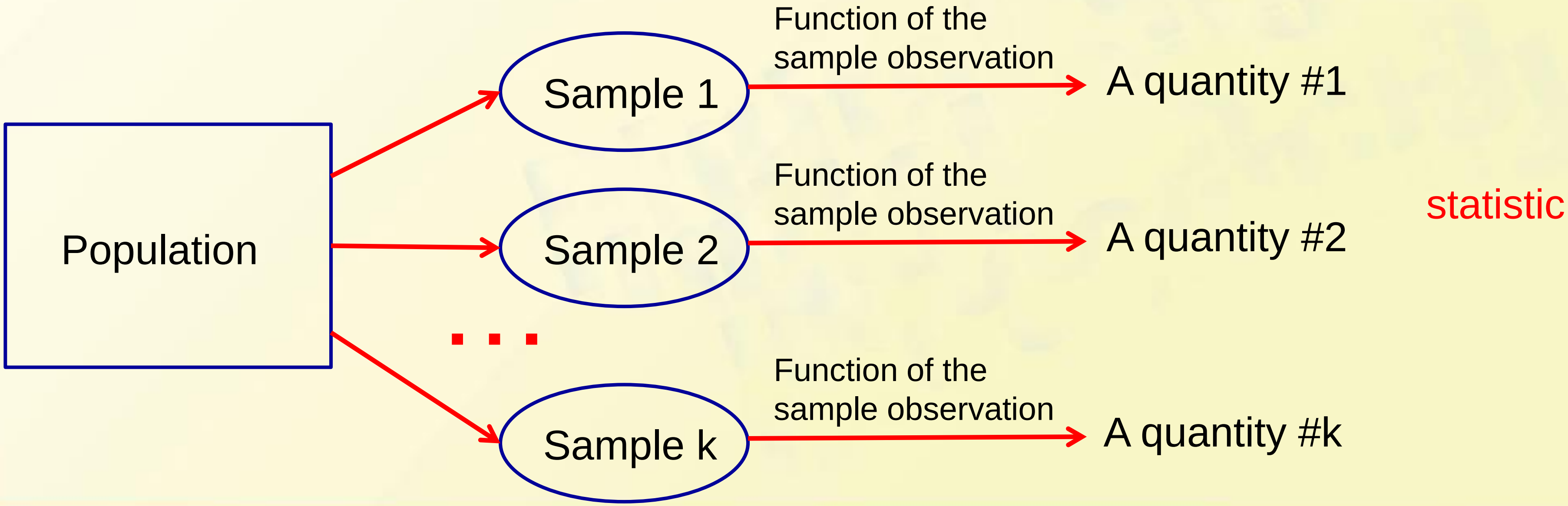
Sample	1	2	3	4	5	6
Mean	4.401	5.928	4.229	4.132	3.620	5.761
Median	4.360	6.144	4.608	3.857	3.221	6.342
Standard Deviation	2.642	2.062	1.611	2.124	1.678	2.496

For sample mean, none of the estimates from these six samples is identical to what is being estimated($\mu = 4.4311$) .

The estimates from the second and sixth samples are much too large, whereas the fifth sample gives a substantial underestimate. All six of the resulting estimates are in error by at least a small amount

5.3 Statistics and Their Distributions

In summary, the value of the individual sample observations vary form sample to sample, so in general the value of any quantity computer from sample data, and the value of a sample characteristic used as an estimate of the corresponding population characteristic, will virtually never coincide with what is being estimated.



5.3 Statistics and Their Distributions

- **Statistic**

A statistic is any quantity whose value can be calculated from **sample data** (with a function).

- Prior to obtaining data, there is **uncertainty** as to what value of any particular statistic will result. Therefore, **a statistic is a random variable.** A **statistic** will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

- The probability distribution of a statistic is sometimes referred to as its **sampling distribution**. It describes how the statistic varies in value across all samples that might be selected.

5.3 Statistics and Their Distributions

- The probability distribution of any particular statistic depends on

1. The population distribution, *e.g.* the normal, uniform, etc. , and the corresponding parameters
2. The sample size n (refer to Ex. 5.20 & 5.30)
3. The method of sampling, *e.g.* sampling with **replacement or without replacement**

- **Example**

Consider selecting a sample of size $n = 2$ from a population consisting of just the three values 1, 5, and 10, and suppose that the statistic of interest is the sample variance.

- If sampling is done "**with replacement**", then $S^2 = 0$ will result if $X_1 = X_2$.
- If sampling is done "**without replacement**", then S^2 can not equal 0.

5.3 Statistics and Their Distributions

- Random Sample

The rv's $X_1, X_2, \ldots, X_n$ are said to form **a (simple) random sample** of size n if

1. The *Xi's* are independent rv's.
2. Every *Xi* has the same probability distribution.

When conditions 1 and 2 are satisfied, we say that the X_i's are independent and identically distributed (**i.i.d**)

Note: Random sample is one of commonly used sampling methods in practice.

5.3 Statistics and Their Distributions

- Random Sample

- Sampling with **replacement** or from an **infinite population** is random sampling

- Sampling **without replacement** from a finite population is generally considered **not random sampling**. However, if the sample **size *n* is much smaller than the population size *N* ($n/N \le 0.05$), it is approximately random sampling.**

Note: The virtue of random sampling method is that the probability distribution of any statistic can be more easily obtained than for any other sampling method.

5.3 Statistics and Their Distributions

- Deriving the Sampling Distribution of a Statistic

➢ Method #1: Calculations based on probability rules

e.g. Example 5.20 & 5.21

➢ Method #2:

Carrying out a simulation experiments

e.g. Example 5.22 & 5.23

- **Example 5.20**

A large automobile service center charges \$40, \$45, and \$50 for a tune-up of four-, six-, and eight-cylinder cars, respectively. If 20% of its tune-ups are done on four-cylinder cars, 30% on six-cylinder cars, and 50% on eight-cylinder cars, then the probability distribution of revenue from a single randomly selected tune-up is given by

x	40	45	50
$p(x)$	0.2	0.3	0.5

$\mu = 46.5$

$\sigma^2 = 15.25$

Suppose on a particular day only two servicing jobs involve tune-ups.

Let X_1 = the revenue from the first tune-up &

X_2 = the revenue from the second,

which constitutes a random sample with the above probability distribution.

5.3 Statistics and Their Distributions

- **Example 5.20 (Cont')**

x_1	x_2	$p(x_1,x_2)$	$\bar{x}$	$\bar{s}^2$
40	40	0.04	40	0
40	45	0.06	42.5	12.5
40	50	0.10	45	50
45	40	0.06	42.5	12.5
45	45	0.09	45	0
45	50	0.15	47.5	12.5
50	40	0.10	45	50
50	45	0.15	47.5	12.5
50	50	0.25	50	0

Known the Population Distribution

x	40	42.5	45	47.5	50
$p_x(x)$	0.04	0.12	0.29	0.30	0.25

$$\mu_{\bar{X}} = E(\bar{X}) = 46.5 = \mu$$

$$\sigma_{\bar{X}}^2 = V(\bar{X}) = \sum \bar{x}^2 \cdot P_{\bar{X}}(\bar{x}) - \mu_{\bar{X}}^2 = 7.625 = \frac{15.25}{2} = \frac{\sigma^2}{2}$$

The variance of $\bar{X}$ is precisely half that of the original variance (because n=2)

Similarly,

$$P_{S^2}(50) = P(S^2 = 50) = P(X_1 = 40, X_2 = 50 \ \ or \ \ X_1 = 50, \ X_2 = 40) = 0.10 + 0.10 = 0.20$$

s^2	0	12.5	50
$p_{s^2}(s^2)$	0.38	0.42	0.20

$$\mu_{S^2} = E(S^2) = \sum s^2 \cdot P_{\overline{X}}(\overline{x}) - \mu_{\overline{x}}^2 = (0)(0.38) + (12.5)(0.42) + (50)(0.20) = 15.25 = \sigma^2$$

That is, the $\overline{X}$ Sampling distribution is centered at the population mean μ

And the S^2 Sampling distribution is centered at the population variance σ^2

5.3 Statistics and Their Distributions

- **Example 5.20 (Cont')**

$\bar{x}$	40	42.5	45	47.5	50
$p_{\bar{x}}(x)$	0.04	0.12	0.29	0.30	0.25

n=2

$\bar{x}$	40	41.25	42.5	43.75	45	43.26	47.5	48.75	50
$p_{\bar{x}}(x)$	0.0016	0.0096	0.0376	0.0936	0.1761	0.2340	0.2350	0.1500	0.0625

n=4

. . .

5.3 Statistics and Their Distributions

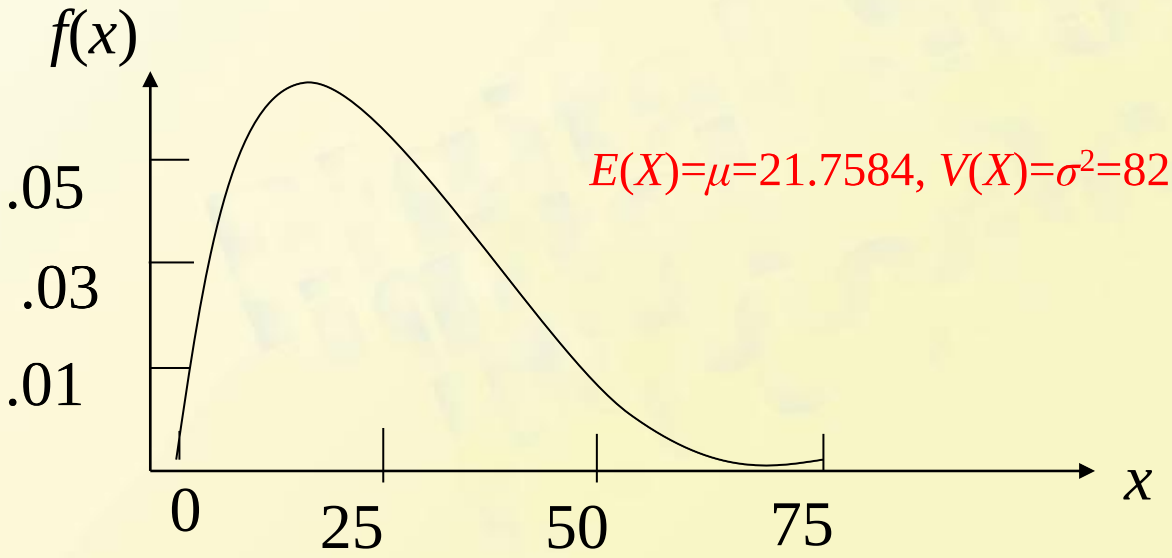
- **Simulation Experiments**

  This method is usually used when a derivation via probability rules is too difficult or complicated to be carried out. Such an experiment is virtually always done with the aid of a computer. And the following characteristics of an experiment must be specified:

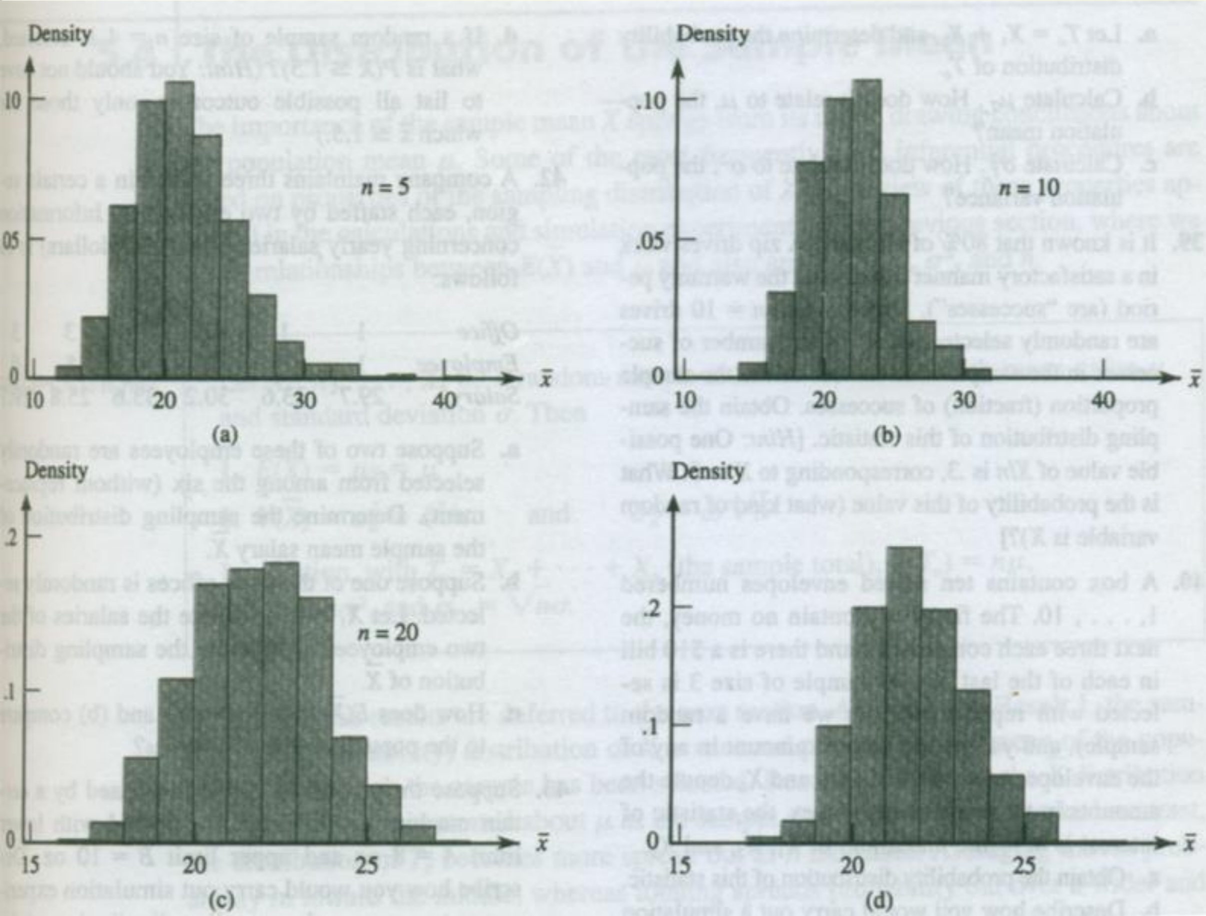
- The statistic of interest (*e.g.* sample mean, S, etc.)
- The population distribution (normal with $\mu = 100$ and $\sigma = 15$, uniform with lower limit $A = 5$ and upper limit $B = 10$, etc.)
- The sample size n (*e.g.*, $n = 10$ or $n = 50$)
- The number of replications k (*e.g.*, $k = 500$ or 1000) (the actual sampling distribution emerges as $k \rightarrow \infty$)

- **Example 5.23**

Consider a simulation experiment in which the population distribution is quite skewed. Figure shows the density curve of a certain type of electronic control (actually a **lognormal distribution with $E(\ln(X)) = 3$ and $V(\ln(X))=.4$**).



Example 5.23 (Cont')



1. **Center of the sampling distribution remains at the population mean.**
2. **As *n* increases:**
   - ✓ **Less skewed ("more normal")**
   - ✓ **More concentrated ("smaller variance")**

Sample histogram for $\bar{X}$ based on 500 samples ,each consisting of n observations:

(a) n=5; (b) n=10; (C) n=20; (d) n=30

5.4 The Distribution of the Sample Mean

- **Proposition**

Let $X_1, X_2, \ldots, X_n$ be a **random sample** (i.i.d. rv's) from a distribution with **mean value μ and standard deviation σ**. Then

$$E(\overline{X}) = \mu_{\overline{X}} = \mu$$

$$V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2 / n \quad and \quad \sigma_{\overline{X}} = \sigma / \sqrt{n}$$

In addition, with $T_o = X_1 + \ldots + X_n$ (the sample total),

$$V(T_0) = n\sigma^2, \; and \quad \sigma_{T_0} = \sqrt{n} / \sigma$$

Refer to 5.5 for the proof!

5.4 The Distribution of the Sample Mean

- **Example 5.24**

In a notched tensile fatigue test on a titanium specimen, the expected number of cycles to first acoustic emission (used to indicate crack initiation) is μ = 28,000, and the standard deviation of the number of cycles is σ = 5000.

Let $X_1, X_2, \ldots, X_{25}$ be a random sample of size 25, where each X_i is the number of cycles on a different randomly selected specimen. Then

$$E(\bar{X}) = \mu = 28,000, E(T_0) = n\mu = 25(28000) = 700,000$$

The standard deviations of $\bar{X}$ and T_o are

$$\sigma_{\bar{X}} = \sigma / \sqrt{n} = \frac{5000}{\sqrt{25}} = 1000$$

$$\sigma_{T_0} = \sqrt{n}\sigma = \sqrt{25}(5000) = 25,000$$

5.4 The Distribution of the Sample Mean

- Proposition

  Let X_1, X_2, …, X_n be a random sample from a normal distribution with mean μ and standard deviation σ.

  Then for **any *n***, $\bar{X}$ is normally distributed (with mean μ and standard deviation $\sigma/\sqrt{n}$), as is T_o (with mean $n\mu$ and standard deviation $\sqrt{n}\sigma$).

- **Example 5.25**

  The time that it takes a randomly selected rat of a certain subspecies to find its way through a maze is a normally distributed rv with μ = 1.5 min and σ = .35 min. Suppose five rats are selected. Let X_1, X_2, …, X_5 denote their times in the maze. Assuming the X_i's to be a random sample from this normal distribution.

- **Q #1:** What is the probability that the total time $T_o = X_1+X_2+\ldots+X_5$ for the five is between 6 and 8 min?
- **Q #2:** Determine the probability that the sample average time $\bar{X}$ is at most 2.0 min.

5.4 The Distribution of the Sample Mean

- **Example 5.25 (Cont')**

A #1: T_o has a normal distribution with $\mu_{To} = n\mu = 5(1.5) = 7.5$ min and variance $\sigma_{To}^2 = n\sigma^2 = 5(0.1225) = 0.6125$, so $\sigma_{To} = 0.783$ min. To standardize T_o, subtract μ_{To} and divide by σ_{To}:

$$P(6 \le T_o \le 8) = P\left(\frac{6-7.5}{0.783} \le Z \le \frac{8-7.5}{0.783}\right)$$
$$= P(-1.92 \le Z \le 0.64) = \Phi(0.64) - \Phi(-1.92) = 0.7115$$

A #2:

$$E(\bar{X}) = \mu = 1.5 \qquad \sigma_{\bar{X}} = \sigma/\sqrt{n} = 0.35/\sqrt{5} = 0.1565$$

$$P(\bar{X} \le 2.0) = P\left(Z \le \frac{2.0-1.5}{0.1565}\right)$$
$$= P(Z \le 3.19) = \Phi(3.19) = 0.9993$$

5.4 The Distribution of the Sample Mean

- The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a **random sample** from a distribution (may or may not be normal) with mean μ and variance σ^2.

Then if n is sufficiently large, $\bar{X}$ has approximately a normal distribution with

$$\mu_{\bar{X}} = \mu, \sigma_{\bar{X}}^2 = \sigma^2 / n$$

T_o also has approximately a normal distribution with

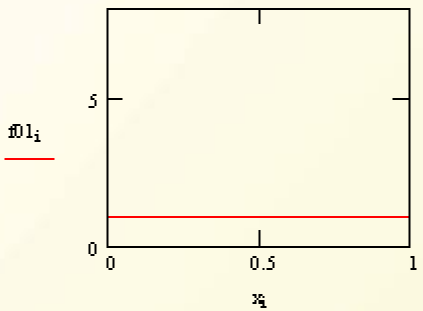
$$\mu_{T_0} = n\mu, \sigma_{T_0}{}^2 = n\sigma^2$$

The larger the value of n, the better the approximation

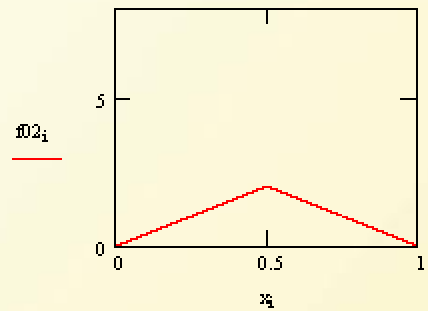
Usually, If n > 30, the Central Limit Theorem can be used.

5.4 The Distribution of the Sample Mean

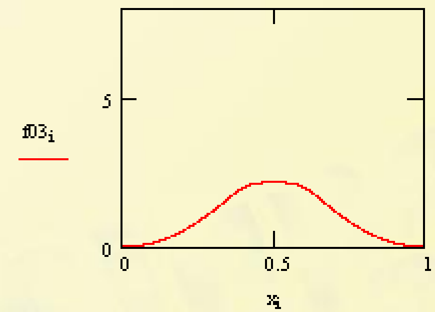
- An Example for Uniform Distribution



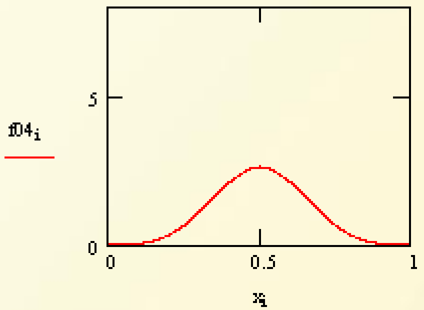
NonNormal Distribution of X



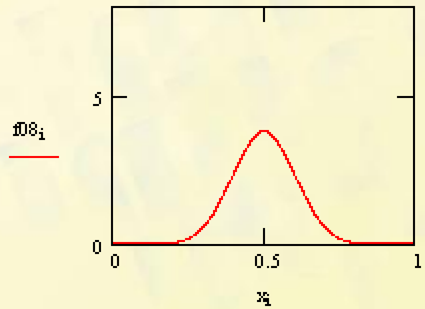
Distribution of Xbar when sample size is 2



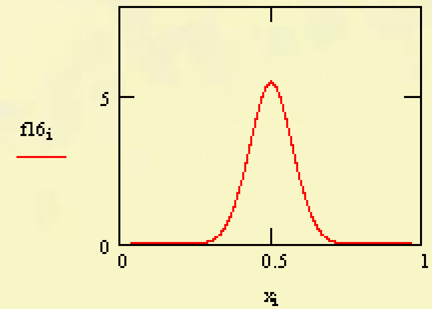
Distribution of Xbar when sample size is 3



Distribution of Xbar when sample size is 4



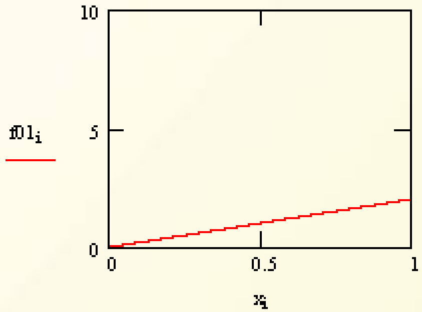
Distribution of Xbar when sample size is 8



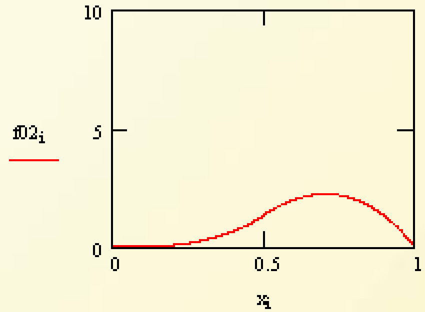
Distribution of Xbar when sample size is 16

5.4 The Distribution of the Sample Mean

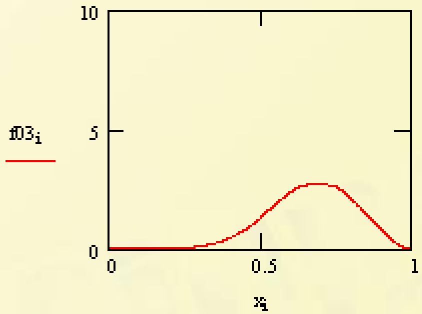
- An Example for Triangular Distribution



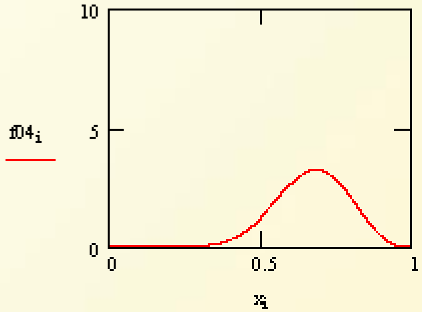
NonNormal Distribution of X



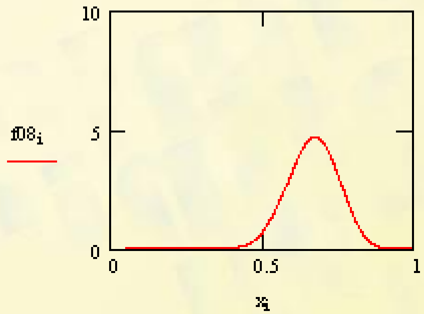
Distribution of Xbar when sample size is 2



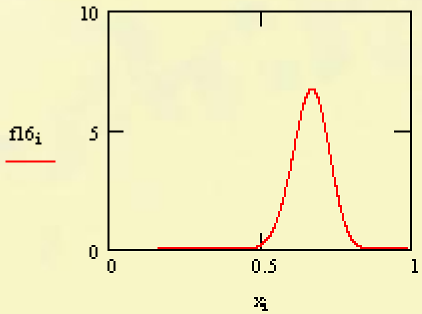
Distribution of Xbar when sample size is 3



Distribution of Xbar when sample size is 4



Distribution of Xbar when sample size is 8



Distribution of Xbar when sample size is 16

5.4 The Distribution of the Sample Mean

- **Example**

When a batch of a certain chemical product is prepared, the amount of a particular impurity in the batch is a random variable with mean value 4.0g and standard deviation 1.5g. If 50 batches are independently prepared, what is the (approximate) probability that the sample average amount of impurity $\bar{X}$ is between 3.5 and 3.8g?

Here $n = 50$ is large enough for the CLT to be applicable. $\bar{X}$ then has approximately a normal distribution with mean value $\mu_{\bar{X}} = 4.0$ and $\sigma_{\bar{X}} = 1.5/\sqrt{50} = 0.2121$, so

$$P(3.5 \le \bar{X} \le 3.8) \approx P\left(\frac{3.5-4.0}{0.2121} \le Z \le \frac{3.8-4.0}{0.2121}\right) = \Phi(-0.94) - \Phi(-2.36) = 0.1645$$

5.5 The Distribution of a Linear Combination

- Linear Combination

Given a collection of n random variables $X_1, \ldots, X_n$ and n numerical constants $a_1, \ldots, a_n$, the rv

$$Y = a_1 X_1 + \cdots + a_n X_n = \sum_{i=1}^{n} a_i X_i$$

is called a linear combination of the X_i's.

5.5 The Distribution of a Linear Combination

Proposition

Let X_1, X_2, …, X_n have mean values μ_1, …, μ_n respectively, and variances of σ_1^2, …., σ_n^2, respectively.

1. Whether or not the X_i's are independent,

$$E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i E(X_i) = \sum_{i=1}^{n} a_i \mu_i$$

2. If X_1, X_2, …, X_n are independent,

$$V(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 V(X_i) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

$$\sigma_{a_1 X_1 + \cdots + a_n X_n} = \sqrt{a_1^2 \sigma_1^2 + \cdots a_n^2 \sigma_n^2}$$

2. For any X_1, X_2, …, X_n,

$$V(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j)$$

5.5 The Distribution of a Linear Combination

- Proof: $E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i E(X_i) = \sum_{i=1}^{n} a_i \mu_i$

For the result concerning expected values, suppose that X_i's are continuous with joint pdf $f(x_1,\ldots,x_n)$. Then

$$E(\sum_{i=1}^{n} a_i X_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\sum_{i=1}^{n} a_i x_i) f(x_1,\ldots,x_n) dx_1 \ldots dx_n$$

$$= \sum_{i=1}^{n} a_i \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i f(x_1,\ldots,x_n) dx_1 \ldots dx_n$$

$$= \sum_{i=1}^{n} a_i \int_{-\infty}^{\infty} x_i f_{X_i}(x_i) dx_i$$

$$= \sum_{i=1}^{n} a_i E(X_i)$$

5.5 The Distribution of a Linear Combination

- Proof: $V(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(X_i, X_j)$

$$V\left(\sum_{i=1}^{n} a_i X_i\right) = E\left[\left(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} a_i \mu_i\right)^2\right]$$

$$= E\left\{\left[\sum_{i=1}^{n} a_i \left(X_i - \mu_i\right)\right]^2\right\} \quad = E\left\{\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j (X_i - \mu_i)(X_j - \mu_j)\right\}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j E[(X_i - \mu_i)(X_j - \mu_j)]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(X_i, X_j)$$

When the X_i's are independent, $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$, and

$$V\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(X_i, X_j) \quad = \sum_{i=1}^{n} a_i^2 V(X_i)$$

- **Example 5.29**

A gas station sells three grades of gasoline: regular unleaded, extra unleaded, and super unleaded. These are priced at \$1.20, \$1.35, and \$1.50 per gallon, respectively. Let X_1, X_2 and X_3 denote the amounts of these grades purchased (gallon) on a particular day.

Suppose the X_i's are **independent** with $\mu_1 = 1000$, $\mu_2 = 500$, $\mu_3 = 300$, $\sigma_1 = 100$, $\sigma_2 = 80$, and $\sigma_3 = 50$. The revenue from sales is $\boldsymbol{Y = 1.2X_1 + 1.35X_2 + 1.5X_3}$. Compute $E(Y)$, $V(Y)$, σ_Y.

Solution:

$$E(Y) = 1.2\mu_1 + 1.35\mu_2 + 1.5\mu_3 = \$2325$$

$$V(Y) = (1.2)^2\sigma_1^2 + (1.35)^2\sigma_2^2 + (1.5)^2\sigma_3^2 = 31,689$$

$$\sigma_Y = \sqrt{31,689} = \$178.01$$

5.5 The Distribution of a Linear Combination

- Corollary (the different between two rv's)

  $E(X_1 - X_2) = E(X_1) - E(X_2)$ **and, if X_1 and X_2 are *independent*,** $V(X_1 - X_2) = V(X_1) + V(X_2)$.

The expected value of a difference is the difference of the two expected values, but the variance of a difference between two independent variables is the **sum**, not the difference, of the two variances

Example 5.30

A certain automobile manufacturer equips a particular model with either a six-cylinder engine or a four-cylinder engine. Let X_1 and X_2 be fuel efficiencies for independently and randomly selected six-cylinder and four-cylinder cars, respectively. With $\mu_1 = 22$, $\mu_2 = 26$, $\sigma_1 = 1.2$, and $\sigma_2 = 1.5$,

Find $E(X_1 - X_2)$, $V(X_1 - X_2)$, $\sigma_{X_1 - X_2}$.

Solution:

$$E(X_1 - X_2) = \mu_1 - \mu_2 = 22 - 26 = -4$$

$$V(X_1 - X_2) = \sigma^2{}_1 + \sigma^2{}_2 = (1.2)^2 + (1.5)^2 = 3.69$$

$$\sigma_{X_1 - X_2} = \sqrt{3.69} = 1.92$$

5.5 The Distribution of a Linear Combination

- **Proposition**

If X_1, X_2, …, X_n are independent, normally distributed rv's (with possibly different means and/or variances), then **any linear combination** of the X_i's also has a normal distribution.

Example 5.31 (Ex. 5.29 Cont')

The total revenue from the sale of the three grades of gasoline on a particular day was $Y = 1.2X_1+1.35X_2+1.5X_3$, and we calculated $\mu_Y = 2325$ and $\sigma_Y = 178.01$). If the X_i's are normally distributed, the probability that the revenue exceeds 2500 is ?

Solution:

$$P(Y \geq 2500) = P(Z > \frac{2500-2325}{178.01})$$

$$= P(Z > 0.98) = 1 - \Phi(0.98) = 0.1635$$