

## Lecture 1

# 1 Review of Some Statistical Concepts

The notation  $\sum$  (sigma), in mathematical term, denotes the **summation**

$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n \quad (\text{Eq.1})$$

The noteworthy properties of summation include:

1.  $\sum_{i=1}^n k = nk$
2.  $\sum_{i=1}^n kx_i = k \sum_{i=1}^n x_i$ , where  $k$  is a constant term.
3.  $\sum_{i=1}^n (a + bx_i) = na + b \sum_{i=1}^n x_i$ , where  $a$  and  $b$  are constants.
4.  $\sum_{i=1}^n (X_i + Y_i) = \sum_{i=1}^n X_i + \sum_{i=1}^n Y_i$ , where  $a$ ,  $b$  and  $k$  are constant.

**Multiple summation** is the summation of variable that is in the form of matrix, shown as,

$$\sum_{i=1}^n \sum_{j=1}^m x_{ij} = \sum_{i=1}^n (x_{i1} + x_{i2} + \cdots + x_{im}) = (x_{11} + x_{21} + \cdots + x_{n1}) + (x_{12} + x_{22} + \cdots + x_{n2}) + \cdots + (x_{1m} + x_{2m} + \cdots + x_{nm}) \quad (\text{Eq.2})$$

where

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}_{n \times m}$$

The significant properties of multiple summations are:

1.  $\sum_{i=1}^n \sum_{j=1}^m X_{ij} = \sum_{j=1}^m \sum_{i=1}^n X_{ij}$
2.  $\sum_{i=1}^n \sum_{j=1}^m X_i Y_j = \sum_{i=1}^n X_i \times \sum_{j=1}^m Y_j$
3.  $\sum_{i=1}^n \sum_{j=1}^m (X_{ij} + Y_{ij}) = \sum_{i=1}^n \sum_{j=1}^m X_{ij} + \sum_{i=1}^n \sum_{j=1}^m Y_{ij}$

$$4. (\sum_{i=1}^n X_i)^2 = \sum_{i=1}^n X_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n X_i X_j$$

The product operator  $\prod$  is defined as:

$$\prod_{i=1}^n x_i = x_1 * x_2 * \dots * x_n \quad (\text{Eq.3})$$

## 2 Experiment

**Sample space** is the set of all possible results of an experiment. For example, if you toss the coin twice, all feasible outcomes are composed of head twice, head followed by tail, tail followed by head, and tail twice. Let H and T denotes head and tail, respectively. The sample space can be written as,

$$SS = \{HH, HT, TH, TT\}$$

**Sample Point** is the member of sample space, eg. the event that head occurs twice from tossing a coin twice. Specifically, sample point is,

$$SP = HH \text{ or } HT \text{ or } TH \text{ or } TT$$

**Events** are the set of specific consequences of the experiment such as the events that head occurs twice. Events are the subset of sample space.

$$A = \text{the event that head occurs twice} = \{HH\}$$

Events are **mutually exclusive**, if the occurrence of one event makes no other events in sample space possible. As an illustration, for the experiment of tossing two coins once, let C be the event that both turn head and D be the event that both turn tail. Since C and D cannot happen at the same time, these two events are said to be mutually exclusive. Another example is the experiment of drawing one card from the standard 52-card deck, let E be the event that the rank of card is King and F be the event that suit of card is Clubs. As the event E and F can occur simultaneously, namely the King of Clubs, the two events are not mutually exclusive.

Events are **collectively exhaustive** if they cover all possible outcomes in the sample space. With the experiment of tossing the coin twice, let A be the event that head appears twice, B be the event that tail appears twice, and C be the event that head and tail each appear once. In this case, A, B and C are collectively exhaustive since all events cover all possible results from sample space; that is, HH, HT, TH and TT.

### 3 Probability and Random Variable

**Probability** is the possibility that any event will occur, given some specific sample space.

Let  $A$  be the event occurring in the given sample space and  $P(A)$  be the probability that  $A$  will happen. Then,  $P(A)$  is defined as;

$$P(A) = \frac{\text{the number of times the event } A \text{ will occur}}{\text{the number of all possible outcomes in sample space}} \quad (\text{Eq.4})$$

For instance, to draw one card from the standard 52-card deck, let  $A$  be the event that the rank of card is 2. Times the event will occur is 4 and the amount of all possible outcomes is 52; hence, the probability of  $A$  is  $\frac{4}{52}$  or  $\frac{1}{13}$ .

Some properties of probability are;

1.  $0 \leq P(A) \leq 1$
2. If  $A$ ,  $B$  and  $C$  are exhaustive set, then,

$$P(A) + P(B) + P(C) = 1$$

3. If  $A$ ,  $B$  and  $C$  are mutually exclusive, then,

$$P(A + B + C) = P(A) + P(B) + P(C)$$

Suppose that the results of an experiment are in the form of value, the variable, whose value is determined by one of those results, is known as **Random Variable**. Random variable can be either **discrete** or **continuous value**.

For discrete random variable, the example is the sum of the values on the face of two dice, when rolling two dice once. In other word, the obtained sum will range from 2 to 12, and it is impossible to get 2.5 or 3.5.

For continuous random variable, the example is the height of the high-school student, constricted to the range from 160 to 180 centimetres. It can be seen that the value of the height need not be the integers and can take the value of 160.5 or 160.52 centimetres.

These two distinct characteristics of random variable enable us to classify them into different probability density functions, which would be stated in Section 2.4.

## 4 Probability Density Function

As the value of random variable depends on an experiment, the **probability density function** would portray the overall image of possible random results. The type of the probability density function relies on the characteristics of the random variable. In this section, many important types are discussed.

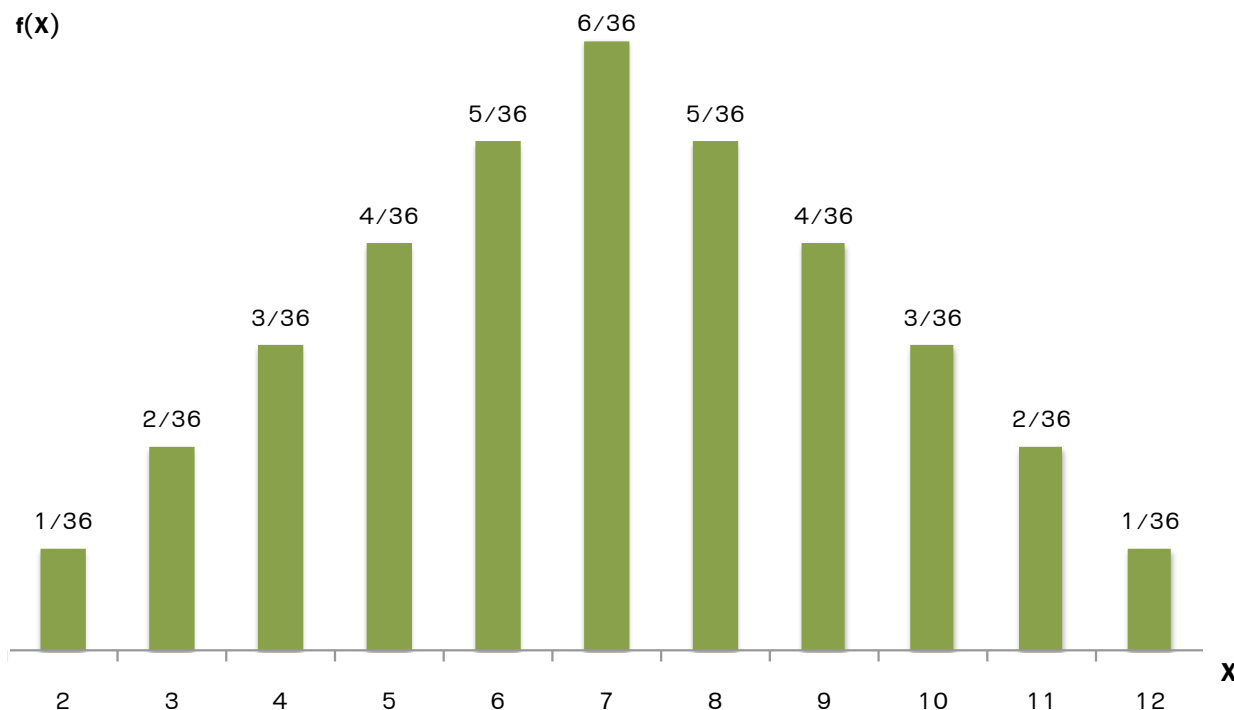
### 4.1 Probability Density Function for Discrete Random Variable

Let  $X$  be the discrete random variable with the value  $x_1, x_2, \dots, x_n$  and we get,

$$\begin{aligned} f(x) &= P(X = x_i) & \text{for } i &= 1, 2, \dots, n \\ f(x) &= 0 & \text{for } x &\neq x_i \end{aligned}$$

**Example:** Let  $X$  be random variable of the sum of values on the face of two dices. The value might be 2 or 12, that is the value from both rolling round is 1 or 6, respectively. The Figure 2-1 summarizes all possible results#

Figure 2-1: Probability Density function of the Sum of Values on the Side of the Dice, Obtained from Rolling the Dice Twice



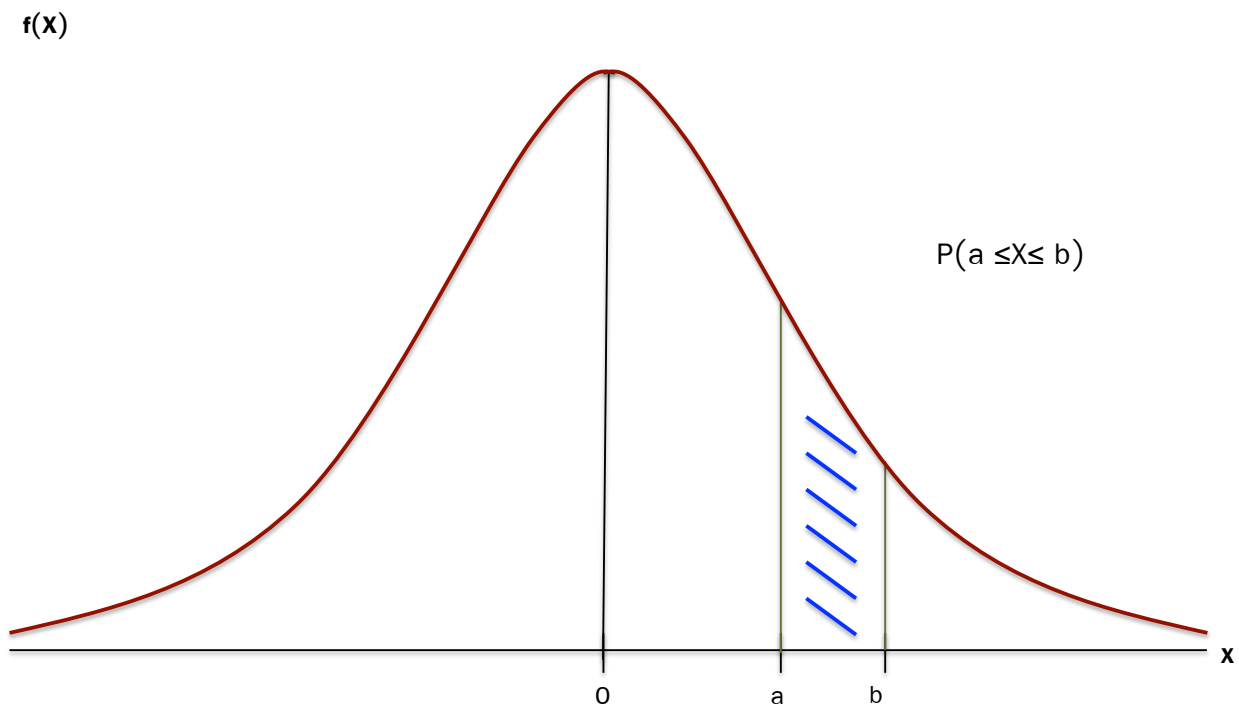
## 4.2 Probability Density Function for Continuous Random Variable

Let  $X$  be the continuous random variable. The probability density function of  $X$  satisfies the three following conditions.

1.  $f(x) \geq 0$
2.  $\int_{-\infty}^{\infty} f(x)dx = 1$
3.  $\int_a^b f(x)dx = P(a \leq x \leq b)$

Figure 2-2 exhibits the probability density function for the continuous random variable, where the area under the curve represents the probability that the variable will lay on that range. Specifically,  $P(a \leq X \leq b)$  means the probability that  $X$  will take the value between  $a$  and  $b$ .

Figure 2-2: Probability Density Function for Continuous Random Variable



**Example-**

### 4.3 Joint Probability Density Function

In this subsection, only **joint probability density function** for discrete variable is discussed. Let  $X$  and  $Y$  be discrete random variables. The joint probability density function, identifying the probability that  $X$  and  $Y$  happen simultaneously, is written as,

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

**Example:** The following table explains the joint probability density function.

Table 2-1: The table illustrating the joint probability density function of  $X$  and  $Y$

|   |   | X    |      |      |
|---|---|------|------|------|
|   |   | -1   | 0    | 1    |
| Y | 1 | 0.11 | 0.08 | 0.05 |
|   | 2 | 0.09 | 0.05 | 0.03 |
|   | 3 | 0.35 | 0.07 | 0.17 |

According to the table, the probability that random variable  $X$  will be 0 and random variable  $Y$  will be 3 is 0.07 or 7 percent. In mathematical term, it can be written as  $f(X = 0, Y = 3) = 0.07$ .

## 4.4 Marginal Probability Density Function

The above joint probability density function  $f(X, Y)$  shows the joint distribution of two variables. On the other hand, **marginal probability density function** with respect to joint probability function, displays the probability density function of single variable like  $f(X)$ ,  $f(Y)$ , which can be derived from;

$$\begin{aligned} f(X) &= \sum_Y f(X, Y) && \text{called} && \text{marginal PDF of X} \\ f(Y) &= \sum_X f(X, Y) && \text{called} && \text{marginal PDF of Y} \end{aligned}$$

where  $\sum_Y$  or  $\sum_X$  means the summation of probability over all values of X and Y respectively.

**Example:** According to Table 2-1 above, marginal PDF of X is obtained from

$$\begin{aligned} f(X = -1) &= \\ &= \\ &= \\ &= \\ f(X = 0) &= \sum_Y f(X = 0, Y) \\ &= f(X = 0, Y = 1) + f(X = 0, Y = 2) + f(X = 0, Y = 3) \\ &= 0.08 + 0.05 + 0.07 \\ &= 0.20 \\ f(X = 1) &= \sum_Y f(X = 1, Y) \\ &= f(X = 1, Y = 1) + f(X = 1, Y = 2) + f(X = 1, Y = 3) \\ &= 0.05 + 0.03 + 0.17 \\ &= 0.25 \end{aligned}$$

marginal PDF of Y is obtained from

$$\begin{aligned} f(Y = 1) &= \\ &= \\ &= \\ &= \\ f(Y = 2) &= \sum_X f(X, Y = 2) \\ &= f(X = -1, Y = 2) + f(X = 0, Y = 2) + f(X = 1, Y = 2) \\ &= 0.09 + 0.05 + 0.03 \\ &= 0.17 \\ f(Y = 3) &= \sum_X f(X, Y = 3) \\ &= f(X = -1, Y = 3) + f(X = 0, Y = 3) + f(X = 1, Y = 3) \\ &= 0.35 + 0.07 + 0.17 \\ &= 0.59 \end{aligned}$$

According to the calculation above, the result can be summarized into Table 2-2.

Table 2-2: Table demonstrating joint probability of random variable  $X$  and  $Y$

|   |   | X                |                 |                 |                      |
|---|---|------------------|-----------------|-----------------|----------------------|
|   |   | -1               | 0               | 1               |                      |
| Y | 1 | 0.11             | 0.08            | 0.05            | $f(Y = 1)$<br>=      |
|   | 2 | 0.09             | 0.05            | 0.03            | $f(Y = 2)$<br>=      |
|   | 3 | 0.35             | 0.07            | 0.17            | $f(Y = 3)$<br>=      |
|   |   | $f(X = -1)$<br>= | $f(X = 0)$<br>= | $f(X = 1)$<br>= | $f(X) =$<br>$f(Y) =$ |

## 4.5 Conditional Probability Density Function

**Conditional probability density function** is the probability of one event given that some events have already occurred. The function is written as,

$$f(X|Y) = P(X = x|Y = y)$$

This function can be obtained from the joint probability density function through,

$$f(X|Y) = \frac{f(X, Y)}{f(Y)}$$

**Example:** According to Table 2.1, find  $f(X = 1|Y = 2)$  and  $f(Y = 2|X = 0)$

$$\begin{aligned}
 f(X = 0|Y = 1) &= \\
 &= \\
 &= \\
 &= \\
 f(Y = 2|X = 0) &= \\
 &= \\
 &=
 \end{aligned}$$

**Example:** Let event  $A$  be tossing the dice once and the point is odd number and  $B$  be the tossing the dice once and the point is at least 5. Find the probability that the point coming up is odd given that the point has to be at least 5.

Answer  $A$  and  $B$  will occur simultaneously if the point from tossing the dice is 5; so, the joint probability of  $A$  and  $B$  is  $\frac{1}{6}$ . The probability that  $B$  occurs is  $\frac{2}{6}$ . Hence, the conditional probability of  $A$  given  $B$  is

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\frac{1}{6}}{\frac{2}{6}} = \frac{1}{2}$$

## 4.6 Statistical Independence

Two random variables are **independent** if the resulting value of one variable does not affect the resulting value of the other; namely,

$$f(X, Y) = f(X)f(Y)$$

**Example:** Consider Mr. Ake's expenditure for a meal and the Miss Somsri's expenditure for a dessert. Given that they do not know each other, the realization of Mr. Ake's expenditure does not imply the realization of Miss Somsri's expenditure. We can, thus, conclude that the expenditures of these two people are independent#

**Example:** Consider drawing cards sequentially from the standard 52-card deck without putting it back into the deck. Once the first card is drawn, the probability of drawing the second card will be influenced because the amount of cards in the deck is reduced. In this case, it can be concluded that drawing the first and second card are not independent#

## 5 Expectation, Variance, Covariance and Correlation

### 5.1 Mean or Expected Value

Because the value of random variable hinges on the value of random results of experiment which cannot be determined certainly, statisticians have invented the measures of central tendency of the random variable. One of them is **expected value**, indicating the mean of the random variable.

For discrete random variable, the expected value is calculated by;

$$E(X) = \sum_{i=1}^n x_i f(x_i) = x_1 f(x_1) + x_2 f(x_2) + \dots + x_n f(x_n)$$

For continuous random variable, the expected value is calculated by,

$$E(X) = \int_a^b x f(x) dx$$

where;

$E(X)$  is the measure of central tendency of random variable, resulting from repeated trial of experiment.

$\sum_{i=1}^n x_i f(x_i)$  is the average of random variable weighted by the probability corresponding to each value.

$a$  and  $b$  are the lowest and highest constant possible respectively.

**Example:** Find the expected value of rolling two dice once (Figure 2-1)

**Example:**

Crucial properties of expected value include:

1.  $E(b) = b$
2.  $E(aX + b) = aE(X) + b$
3.  $E(XY) = E(X)E(Y)$ ; given that  $X$  and  $Y$  are independent
4.  $E(g(X)) = \sum_x g(X)f(X)$

where  $a$  and  $b$  are constant.

**Conditional expectation value** is the expectation value of random variable under some conditions such as expected value of  $X$  conditional on  $Y$  or  $E(X|Y = 5)$

Let  $f(X, Y)$  be the joint probability function of  $X$  and  $Y$ . The expectation of  $X$  conditional on some value of  $Y$  is defined as,

For discrete random variable  $E(X|Y = y) = \sum_X X_i f(X|Y = y)$

For continuous random variable  $E(X|Y = y) = \int_{-\infty}^{\infty} X_i f(X|Y = y)$

### Example

## 5.2 Variance

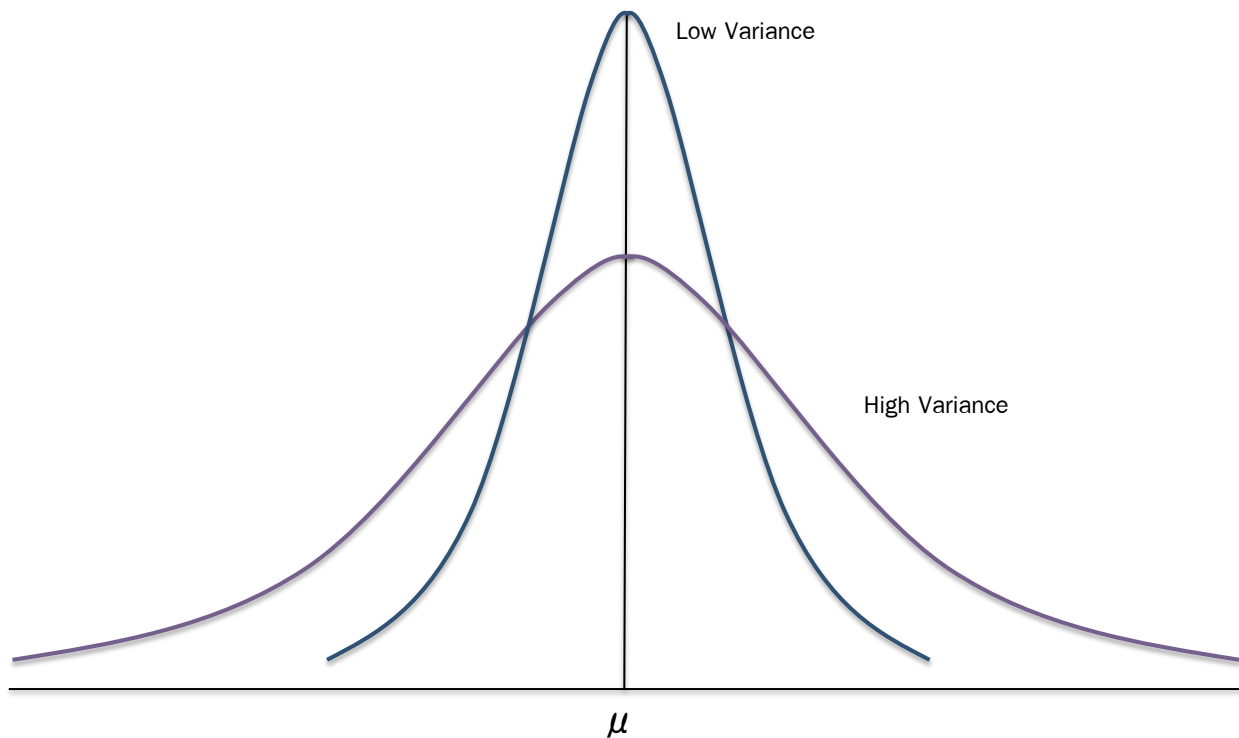
**Variance** is the measure of dispersion of the value of variable around the expected value. The higher the variance, the more dispersing the random variable (Figure 2-3). If  $X$  is the random variable with expected value  $\mu$ , we get;

$$Var(X) = \sigma_X^2 = E[X - E(X)]^2 = E(X)^2 - \mu^2 \quad (\text{Eq.5})$$

From,

$$\begin{aligned} \text{Var}(X) &= \sigma_X^2 \\ &= E[X - E(X)]^2 \\ &= E[X^2 - 2XE(X) + (E(X))^2] \\ &= E(X^2) - 2(E(X))^2 + (E(X))^2 \\ &= E(X)^2 - \mu^2 \end{aligned}$$

Figure 2-3: Distribution of Random Variables with Different Variance



Important properties of expected value include;

1.  $\text{Var}(b) = 0$
2.  $\text{Var}(aX + b) = a^2\text{Var}(X)$
3.  $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ ; given that X and Y are independent
4.  $\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$

where  $a$  and  $b$  are constant.

## Lecture 2

### 5.3 Conditional Variance

The conditional variance of  $X$  is given  $Y = y$  is defined as following:

$$\begin{aligned} \text{var}(X|Y = y) &= E \{ [X - E(X|Y = y)]^2 | Y = y \} \\ &= \sum_x [X - E(X|Y = y)]^2 f(x|Y = y) \\ &= \int_{-\infty}^{\infty} [X - E(X|Y = y)]^2 f(x|Y = y) dx \end{aligned} \quad (\text{Eq.6})$$

#### Example

## Properties of conditional expectation and conditional variance

## 5.4 Covariance

**Theorem.** Let  $X$  and  $Y$  be two random variables with means  $\mu_x$  and  $\mu_y$ , respectively. Then, we can define the covariance between these two variables as following:

$$\text{cov}(X, Y) = E \{ (X - \mu_x) (Y - \mu_y) \} = E(XY) - \mu_x \mu_y \quad (\text{Eq.7})$$

If  $X$  and  $Y$  are continuous random variables we can calculate their  $\text{cov}(X, Y)$ :

$$\begin{aligned} \text{cov}(X, Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X - \mu_x) (Y - \mu_y) f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} XY f(x, y) dx dy - \mu_x \mu_y \end{aligned} \quad (\text{Eq.8})$$

### Properties of Covariance

1. If  $X$  and  $Y$  are independent, the covariance between  $X$  and  $Y$  is zero.

Proof:

2.  $\text{cov}(a + bX, c + dY) = bd * \text{cov}(X, Y)$ , where  $a, b, c$ , and  $d$  are constants.

**Example** Suppose the joint PDF of random variables  $X$  and  $Y$  can be represented as in the below table. What is the covariance between  $X$  and  $Y$ ?

|   |   | X               |                 |                 |                      |
|---|---|-----------------|-----------------|-----------------|----------------------|
|   |   | 1               | 2               | 3               |                      |
| Y | 1 | 0.25            | 0.25            | 0               | $f(Y = 1)$<br>=      |
|   | 2 | 0               | 0.25            | 0.25            | $f(Y = 2)$<br>=      |
|   |   | $f(X = 1)$<br>= | $f(X = 2)$<br>= | $f(X = 3)$<br>= | $f(X) =$<br>$f(Y) =$ |

Next, we will turn our attention to seeing how we can apply the covariance to calculate the correlation between the random variables X and Y

## 5.5 Correlation

When we calculate the covariance of X and Y, it reflects the units of both random variables. However, it is useful to have a **dimensionless measure of dependency** by calculating the correlation instead.

**Definition** Let X and Y be any two random variables (discrete or continuous) with standard deviation  $\sigma_X$  and  $\sigma_Y$ , respectively. The **correlation coefficient** of X and Y, denoted **corr(X,Y)** or  $\rho_{XY}$  ( the greek letter "rho") is defined as:

$$\rho_{XY} = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}} = \frac{cov(x, y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

**Example** Suppose the join PDF of random variables X and Y can be represented as in the below table. What is the correlation between X and Y?

|   |   | X                    |                     |                     |                          |
|---|---|----------------------|---------------------|---------------------|--------------------------|
|   |   | 1                    | 2                   | 3                   |                          |
| Y | 1 | 0.25                 | 0.25                | 0                   | $f(Y = 1)$<br>=0.5       |
|   | 2 | 0                    | 0.25                | 0.25                | $f(Y = 2)$<br>=0.5       |
|   |   | $f(X = 1)$<br>= 0.25 | $f(X = 2)$<br>= 0.5 | $f(X = 3)$<br>=0.25 | $f(X) = 1$<br>$f(Y) = 1$ |

From the definition,  $\rho_{XY}$  is measure of linear association between two random variables. The value of  $\rho$  lies between -1 and +1,  $-1 \leq \rho_{XY} \leq +1$ . We can interpret the value of correlation as:

- ▶ If  $\rho_{XY} = 1$ , then X and Y are perfectly, positively, linearly correlated.
- ▶ If  $\rho_{XY} = -1$ , then X and Y are perfectly, negatively, linearly correlated.
- ▶ If  $\rho_{XY} = 0$ , then X and Y are completely, un-linearly correlated. This means that X and Y may correlated in some other manner i.e. a parabolic manner., but NOT in a linear manner
- ▶ If  $\rho_{XY} \leq 0$ , then X and Y are positively, linearly correlated, but NOT perfectly.
- ▶ If  $\rho_{XY} \geq 0$ , then X and Y are negatively, linearly correlated, but NOT perfectly.

**Theorem.** If X and Y are independent random variables, then:

$$\text{corr}(X, Y) = \text{cov}(X, Y) = 0$$

**Example:** Let X = the outcome of a fair, black, 6-sided die.

Let Y = outcome of a fair, red, 4-sided die.

What is the covariance of X and Y? What is the correlation of X and Y?

**NOTE:** The converse of the theorem is NOT NECESSARILY CORRECT!

**Example:** Let  $X$  and  $Y$  be two discrete random variables with the following joint PDF:

|   |   | X               |                 |                 |                 |
|---|---|-----------------|-----------------|-----------------|-----------------|
|   |   | 0               | 1               | 2               |                 |
| Y | 0 | 0               | 0.20            | 0.10            | $f(Y = 0)$<br>= |
|   | 1 | 0.20            | 0.40            | 0               | $f(Y = 1)$<br>= |
|   | 2 | 0.10            | 0               | 0               | $f(Y = 2)$<br>= |
|   |   | $f(X = 0)$<br>= | $f(X = 1)$<br>= | $f(X = 2)$<br>= |                 |

What is the correlation between  $X$  and  $Y$ ? And, are  $X$  and  $Y$  independent?

## 5.6 Variances of Correlated Variables

Let  $X$  and  $Y$  be two random variables, then

$$\begin{aligned} \text{var}(X + Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y) \\ &= \text{var}(X) + \text{var}(Y) + 2\rho\sigma_x\sigma_y \\ \text{var}(X - Y) &= \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y) \\ &= \text{var}(X) + \text{var}(Y) - 2\rho\sigma_x\sigma_y \end{aligned} \tag{Eq.9}$$

The generalized result:

Let  $\sum_{i=1}^n X_i = X_1 + X_2 + \cdots + X_n$ , then the variance of the linear combination  $\sum X_i$  is:

$$\begin{aligned} \text{var}\left(\sum_{i=1}^n X_i\right) &= \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i < j} \text{cov}(X_i, X_j) \\ &= \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i < j} \rho_{ij} \sigma_i \sigma_j \end{aligned} \tag{Eq.10}$$

**Example:**

what is the  $\text{var}(X_1 + X_2 + X_3)$ ?

## Lecture 3

### 5.7 Higher Moments of Probability Distributions

In the previous subsection, we have already discussed about mean, variance, and covariance as the measures of the first and second moments of univariate and multivariate PDFs. Besides the first two moments, we are occasionally interested in the higher moments such as the third and fourth moments which are normally applied in studying the “Shape” of the distribution. In general, the  $r^{th}$  moments about the mean is defined as

$$r^{th} \text{moment} : E(X - \mu)^r$$

By the definition of  $r^{th}$  moments, we can easily define the third and fourth moments as:

**Third moment:**

$$E(X - \mu)^3$$

**Fourth moment:**

$$E(X - \mu)^4$$

We can study the shape of the distribution by calculating **skewness** and **kurtosis**.

**SKEWNESS** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

One measure of skewness is defined as:

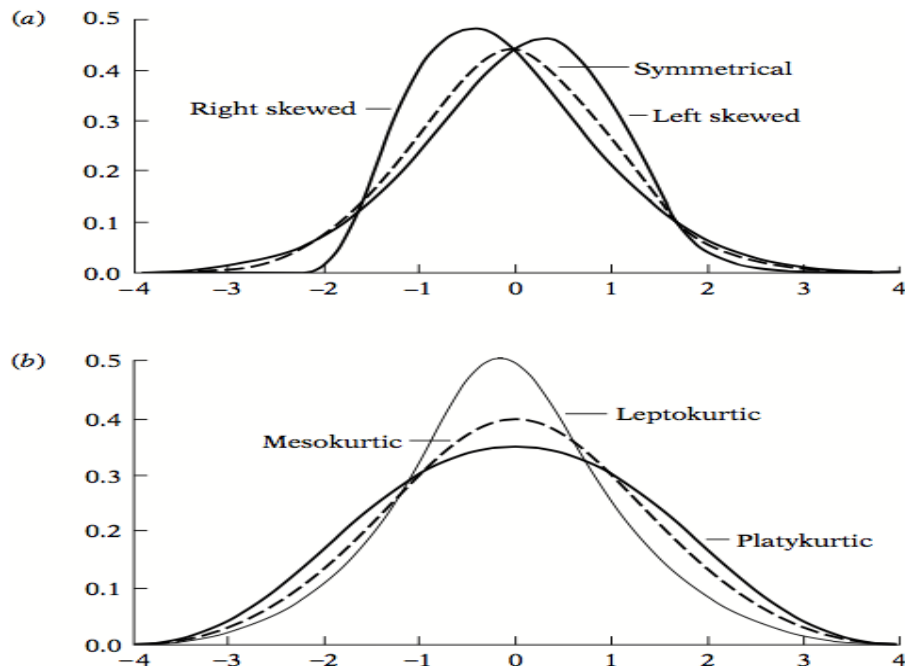
|                                                                                                                                                                       |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\begin{aligned} S &= \frac{E(X - \mu)^3}{\sigma^3} \\ &= \frac{\text{third moment about the mean}}{\text{cube of the standard deviation}} \end{aligned} \tag{Eq.11}$ |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|

**KURTOSIS** is a measure of the peakedness of the probability distribution of a real-valued random variable

We can also measure kurtosis as:

$$\begin{aligned}
 S &= \frac{E(X - \mu)^4}{\sigma^4} \\
 &= \frac{\text{fourth moment about the mean}}{\text{square of the second moment}}
 \end{aligned}
 \tag{Eq.12}$$

- ♣ **Platykurtic (fat or short-tailed)**  $\implies$  PDFs with Kurtosis  $< 3$
- ♣ **Leptokurtic (slim or long-tailed)**  $\implies$  PDFs with Kurtosis  $> 3$
- ♣ **Mesokurtic (which is the normal distribution)**  $\implies$  PDFs with Kurtosis  $= 3$



**Figure 1.** (a) Skewness; (b)Kurtosis

## 6 Some important probability distribution

### 6.1 Normal Distribution

A continuous random variable  $X$  has a normal distribution with mean  $\mu$  and variance  $\sigma^2$  if its probability density function (pdf) is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right) \quad \text{where} \quad -\infty < x < \infty$$

**NOTE:** The normal distribution can be described by two parameters

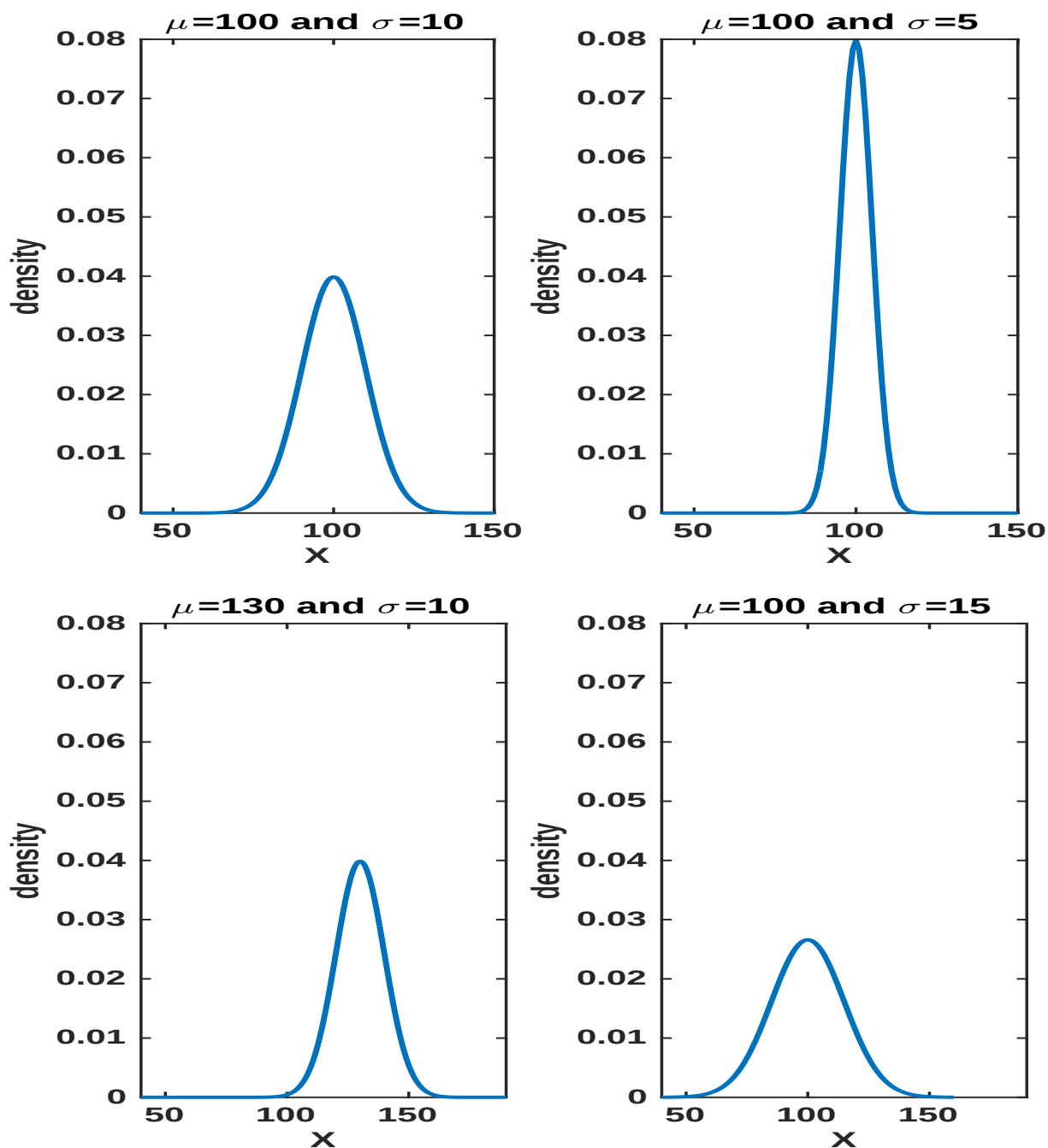
■  $\mu$  = The mean of the distribution.

■  $\sigma$  = The standard deviation of the distribution.

Therefore, changing the values of  $\mu$  and  $\sigma$  alter the positions and shapes of the distributions.

If  $X$  is Normally distributed with mean  $\mu$  and standard deviation  $\sigma$ , we can write it as:

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



**Figure 2.** Compare the mean and standard deviation of the normal distribution

### The properties of the normal distribution.

- ★ It is symmetrical around its mean value.
- ★ About 68 percent of the area under the normal distribution lies between the value  $\mu \pm \sigma$
- About 95 percent of the area under the normal distribution lies between the value  $\mu \pm 2\sigma$
- About 99.7 percent of the area under the normal distribution lies between the value  $\mu \pm 3\sigma$  (as shown in figure 2)

★ We can convert the given normally distributed variable  $X$  with mean  $\mu$  and  $\sigma^2$  into the standardized normal variable  $Z$  by calculating  $Z$  where  $Z$  can be defined as:

$$Z =$$

With the standardized normal variable  $Z$ , we can rewrite the normal pdf as:

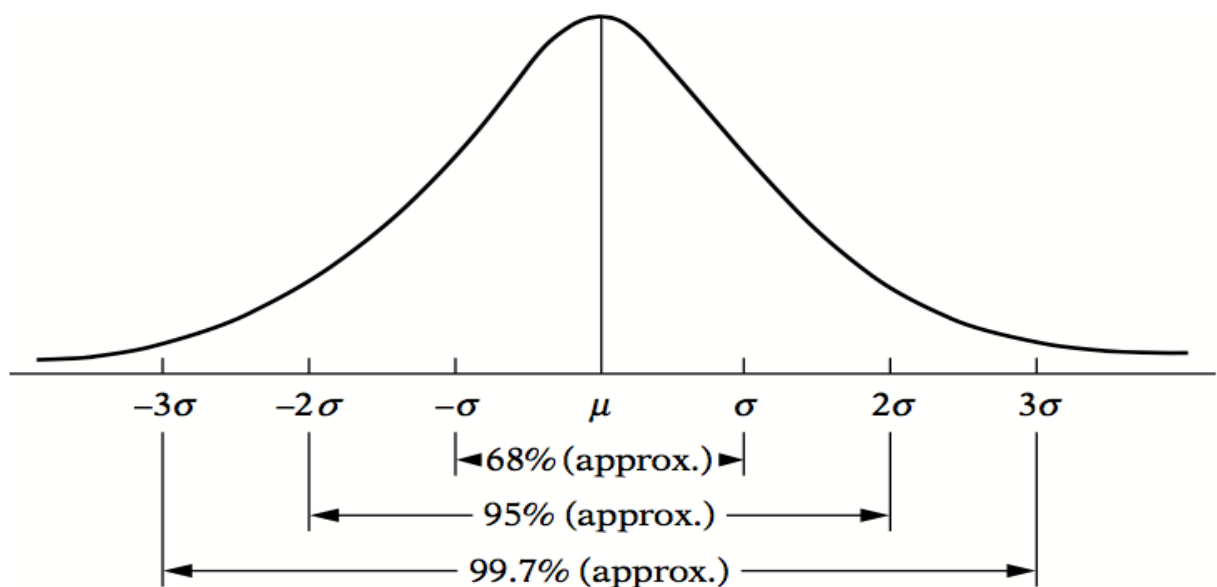
$$f(Z) =$$

In sum, you can see that we convert the given normally distributed variable  $X$  into the standardized normal variable by:

- (i) Subtracting the mean  $\mu$
- (ii) Dividing by the standard deviation  $\sigma$

♡ Subtracting the mean re-centers the distribution on zero.

♡ Dividing by the standard deviation re-scales the distribution so it has standard deviation 1.

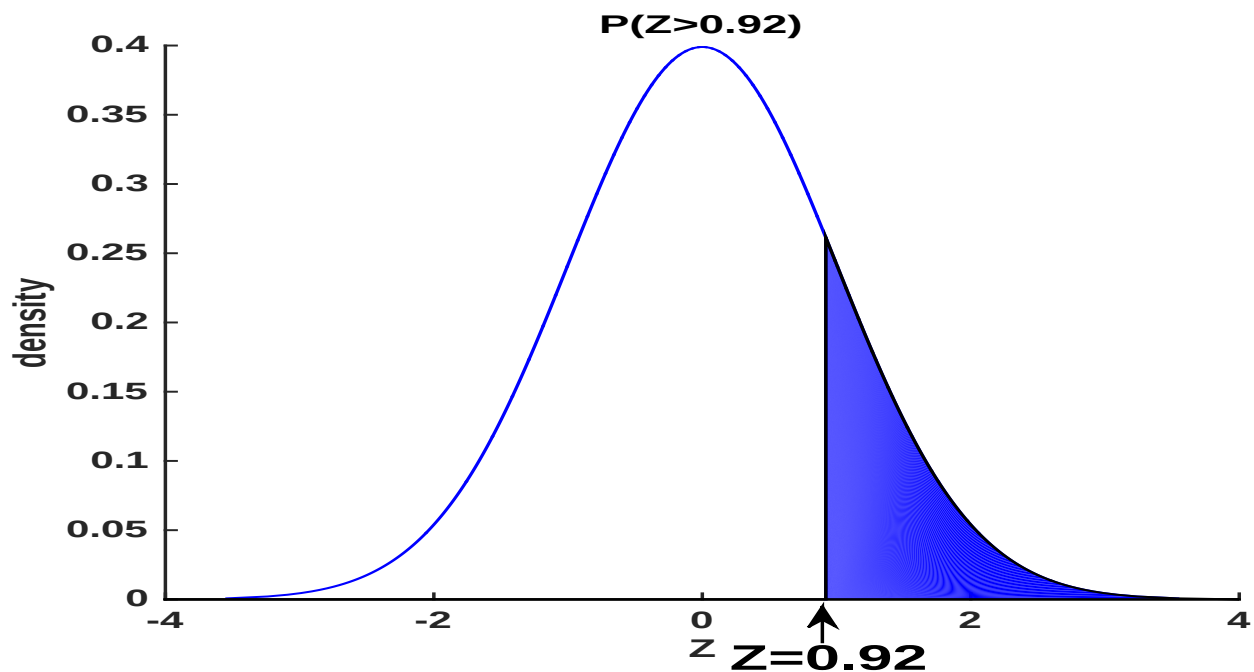


**Figure 3.** Areas under the normal distribution

It should be remarked that its mean value is zero and its variance is unity for any standardized variable.

By convention, we can denote a normally distributed variable  $X$  with zero mean and unit variance as

$$X \sim N(0, 1)$$



**Figure 4.** If  $Z \sim N(0,1)$ , the probability that  $P(Z > 0.92)$

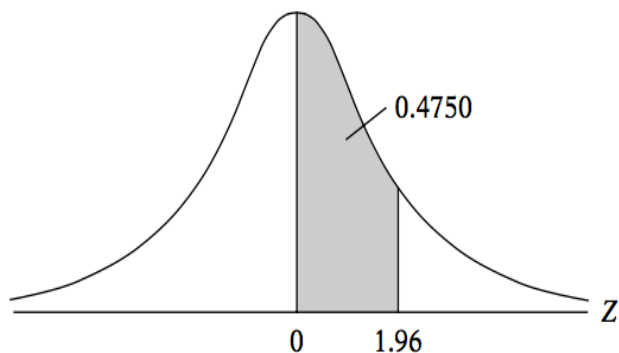
**Example** If  $Z \sim N(0,1)$  what is  $P(Z > 0.92)$ ?

## AREAS UNDER THE STANDARDIZED NORMAL DISTRIBUTION

### Example

$$\Pr(0 \leq Z \leq 1.96) = 0.4750$$

$$\Pr(Z \geq 1.96) = 0.5 - 0.4750 = 0.025$$



| Z   | .00   | .01   | .02   | .03   | .04   | .05   | .06   | .07   | .08   | .09   |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.0 | .0000 | .0040 | .0080 | .0120 | .0160 | .0199 | .0239 | .0279 | .0319 | .0359 |
| 0.1 | .0398 | .0438 | .0478 | .0517 | .0557 | .0596 | .0636 | .0675 | .0714 | .0753 |
| 0.2 | .0793 | .0832 | .0871 | .0910 | .0948 | .0987 | .1026 | .1064 | .1103 | .1141 |
| 0.3 | .1179 | .1217 | .1255 | .1293 | .1331 | .1368 | .1406 | .1443 | .1480 | .1517 |
| 0.4 | .1554 | .1591 | .1628 | .1664 | .1700 | .1736 | .1772 | .1808 | .1844 | .1879 |
| 0.5 | .1915 | .1950 | .1985 | .2019 | .2054 | .2088 | .2123 | .2157 | .2190 | .2224 |
| 0.6 | .2257 | .2291 | .2324 | .2357 | .2389 | .2422 | .2454 | .2486 | .2517 | .2549 |
| 0.7 | .2580 | .2611 | .2642 | .2673 | .2704 | .2734 | .2764 | .2794 | .2823 | .2852 |
| 0.8 | .2881 | .2910 | .2939 | .2967 | .2995 | .3023 | .3051 | .3078 | .3106 | .3133 |
| 0.9 | .3159 | .3186 | .3212 | .3238 | .3264 | .3289 | .3315 | .3340 | .3365 | .3389 |
| 1.0 | .3413 | .3438 | .3461 | .3485 | .3508 | .3531 | .3554 | .3577 | .3599 | .3621 |
| 1.1 | .3643 | .3665 | .3686 | .3708 | .3729 | .3749 | .3770 | .3790 | .3810 | .3830 |
| 1.2 | .3849 | .3869 | .3888 | .3907 | .3925 | .3944 | .3962 | .3980 | .3997 | .4015 |
| 1.3 | .4032 | .4049 | .4066 | .4082 | .4099 | .4115 | .4131 | .4147 | .4162 | .4177 |
| 1.4 | .4192 | .4207 | .4222 | .4236 | .4251 | .4265 | .4279 | .4292 | .4306 | .4319 |
| 1.5 | .4332 | .4345 | .4357 | .4370 | .4382 | .4394 | .4406 | .4418 | .4429 | .4441 |
| 1.6 | .4452 | .4463 | .4474 | .4484 | .4495 | .4505 | .4515 | .4525 | .4535 | .4545 |
| 1.7 | .4454 | .4564 | .4573 | .4582 | .4591 | .4599 | .4608 | .4616 | .4625 | .4633 |
| 1.8 | .4641 | .4649 | .4656 | .4664 | .4671 | .4678 | .4686 | .4693 | .4699 | .4706 |
| 1.9 | .4713 | .4719 | .4726 | .4732 | .4738 | .4744 | .4750 | .4756 | .4761 | .4767 |
| 2.0 | .4772 | .4778 | .4783 | .4788 | .4793 | .4798 | .4803 | .4808 | .4812 | .4817 |
| 2.1 | .4821 | .4826 | .4830 | .4834 | .4838 | .4842 | .4846 | .4850 | .4854 | .4857 |
| 2.2 | .4861 | .4864 | .4868 | .4871 | .4875 | .4878 | .4881 | .4884 | .4887 | .4890 |
| 2.3 | .4893 | .4896 | .4898 | .4901 | .4904 | .4906 | .4909 | .4911 | .4913 | .4916 |
| 2.4 | .4918 | .4920 | .4922 | .4925 | .4927 | .4929 | .4931 | .4932 | .4934 | .4936 |
| 2.5 | .4938 | .4940 | .4941 | .4943 | .4945 | .4946 | .4948 | .4949 | .4951 | .4952 |
| 2.6 | .4953 | .4955 | .4956 | .4957 | .4959 | .4960 | .4961 | .4962 | .4963 | .4964 |
| 2.7 | .4965 | .4966 | .4967 | .4968 | .4969 | .4970 | .4971 | .4972 | .4973 | .4974 |
| 2.8 | .4974 | .4975 | .4976 | .4977 | .4977 | .4978 | .4979 | .4979 | .4980 | .4981 |
| 2.9 | .4981 | .4982 | .4982 | .4983 | .4984 | .4984 | .4985 | .4985 | .4986 | .4986 |
| 3.0 | .4987 | .4987 | .4987 | .4988 | .4988 | .4989 | .4989 | .4989 | .4990 | .4990 |

**Example** If  $Z \sim N(0,1)$  what is  $P(-0.64 < Z < 0.43)$ ?

**Example** If  $X \sim N(3500, 500^2)$  what is  $P(X < 3100)$ ?

Let  $X_1 \sim N(\mu_1, \sigma_1^2)$  and  $X_2 \sim N(\mu_2, \sigma_2^2)$  and assume that  $X_1$  and  $X_2$  are independent. If we have the linear combination between  $X_1$  and  $X_2$  where we can write it as:

$$Y = aX_1 + bX_2,$$

where  $a$  and  $b$  are the constant terms. Then

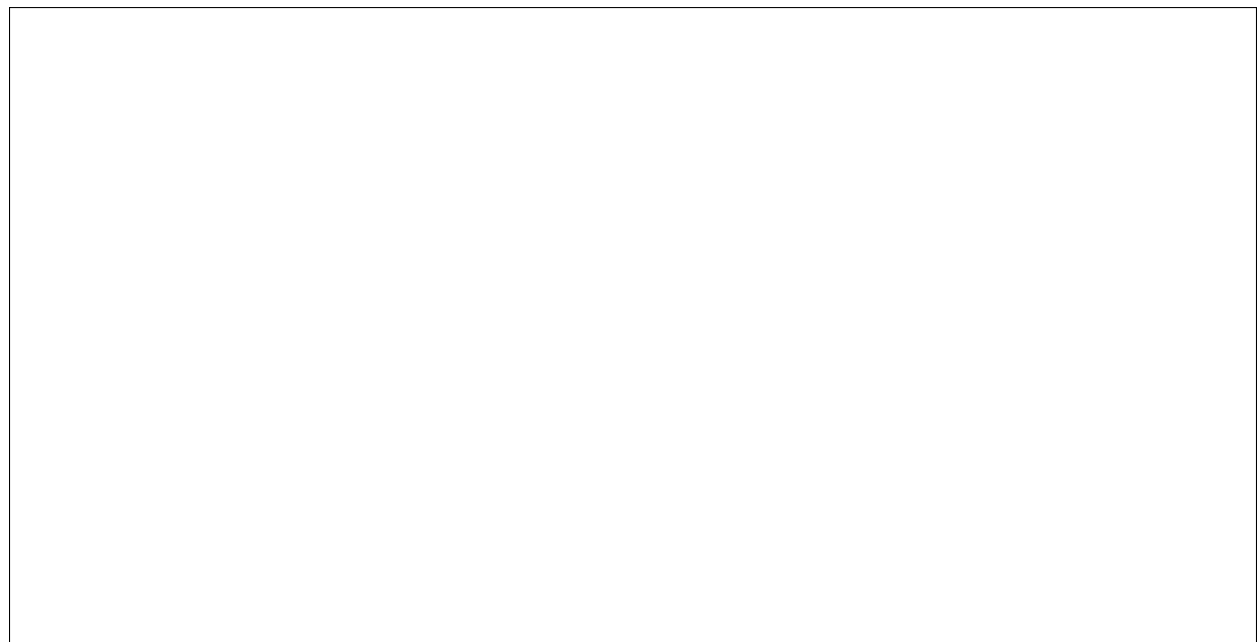
$$Y \sim N[(a\mu_1 + b\mu_2), (a^2\sigma_1^2 + b^2\sigma_2^2)]$$

In other words, **a linear combination of normally distributed variables is itself normally distributed.**

**Central limit theorem** Let  $X_1, X_2, \dots, X_n$  denote  $n$  independent random variables and

$$X_i \sim N(\mu, \sigma)$$

Let  $\bar{X} = \sum \frac{X_i}{n}$ , then as  $n$  increases indefinitely (i.e,  $n \rightarrow \infty$ ),



**The third and fourth moments of the normal distribution:**

**Third moment:**  $E(X - \mu)^3 = 0$

**Fourth moment:**  $E(X - \mu)^4 = 3\sigma^4$

## Lecture 4

### 6.2 The $\chi^2$ (Chi-Square) Distribution

Let  $Z_1, Z_2, \dots, Z_k$  be **independent standardized normal variables**. Then the quantity

$$Z = \sum_{i=1}^k Z_i^2$$

is said to possess the  $\chi^2$  with  $k$  degree of freedom (df)

**Properties of the  $\chi^2$  distribution are as follows:**

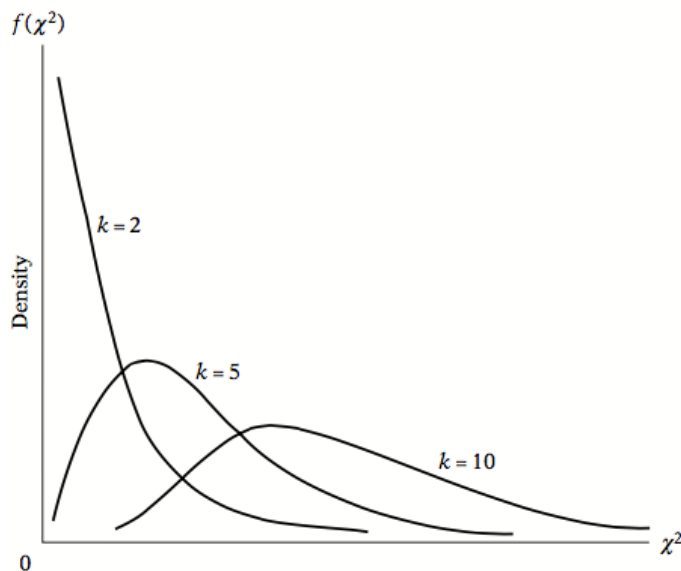
1. The  $\chi^2$  distribution is a skewed distribution where the degree of the skewness depending on the df. As the number of df increases, the distribution becomes more symmetrical. For the df excess of 100, the variable

$$\frac{\sqrt{2\chi^2} - \sqrt{(2k-1)}}{\sqrt{2}}$$

can be converted to a standardized normal variable, where  $k$  is the df.

2. The mean of the chi-square distribution is  $k$ , and its variance is  $2k$ , where  $k$  is the df.

3. If  $Z_1$  and  $Z_2$  are two independent chi-square variables with  $k_1$  and  $k_2$  df, then the sum of  $Z_1 + Z_2$  is also a chi-square with  $df = k_1 + k_2$



**Figure 5.** Density function of the  $\chi^2$  variable

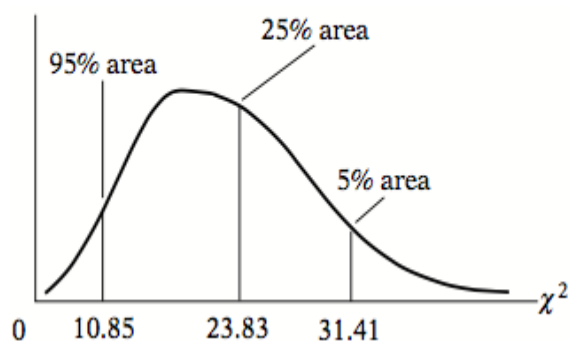
## UPPER PERCENTAGE POINTS OF THE $\chi^2$ DISTRIBUTION

### Example

$$\Pr(\chi^2 > 10.85) = 0.95$$

$$\Pr(\chi^2 > 23.83) = 0.25 \quad \text{for } df = 20$$

$$\Pr(\chi^2 > 31.41) = 0.05$$



| Degrees of freedom \ Pr | .995                     | .990                    | .975                    | .950                    | .900     |
|-------------------------|--------------------------|-------------------------|-------------------------|-------------------------|----------|
| 1                       | $392704 \times 10^{-10}$ | $157088 \times 10^{-9}$ | $982069 \times 10^{-9}$ | $393214 \times 10^{-8}$ | .0157908 |
| 2                       | .0100251                 | .0201007                | .0506356                | .102587                 | .210720  |
| 3                       | .0717212                 | .114832                 | .215795                 | .351846                 | .584375  |
| 4                       | .206990                  | .297110                 | .484419                 | .710721                 | 1.063623 |
| 5                       | .411740                  | .554300                 | .831211                 | 1.145476                | 1.61031  |
| 6                       | .675727                  | .872085                 | 1.237347                | 1.63539                 | 2.20413  |
| 7                       | .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  |
| 11                      | 2.60321                  | 3.05347                 | 3.81575                 | 4.57481                 | 5.57779  |
| 12                      | 3.07382                  | 3.57056                 | 4.40379                 | 5.22603                 | 6.30380  |
| 13                      | 3.56503                  | 4.10691                 | 5.00874                 | 5.89186                 | 7.04150  |
| 14                      | 4.07468                  | 4.66043                 | 5.62872                 | 6.57063                 | 7.78953  |
| 15                      | 4.60094                  | 5.22935                 | 6.26214                 | 7.26094                 | 8.54675  |
| 16                      | 5.14224                  | 5.81221                 | 6.90766                 | 7.96164                 | 9.31223  |
| 17                      | 5.69724                  | 6.40776                 | 7.56418                 | 8.67176                 | 10.0852  |
| 18                      | 6.26481                  | 7.01491                 | 8.23075                 | 9.39046                 | 10.8649  |
| 19                      | 6.84398                  | 7.63273                 | 8.90655                 | 10.1170                 | 11.6509  |
| 20                      | 7.43386                  | 8.26040                 | 9.59083                 | 10.8508                 | 12.4426  |
| 21                      | 8.03366                  | 8.89720                 | 10.28293                | 11.5913                 | 13.2396  |
| 22                      | 8.64272                  | 9.54249                 | 10.9823                 | 12.3380                 | 14.0415  |
| 23                      | 9.26042                  | 10.19567                | 11.6885                 | 13.0905                 | 14.8479  |
| 24                      | 9.88623                  | 10.8564                 | 12.4011                 | 13.8484                 | 15.6587  |
| 25                      | 10.5197                  | 11.5240                 | 13.1197                 | 14.6114                 | 16.4734  |
| 26                      | 11.1603                  | 12.1981                 | 13.8439                 | 15.3791                 | 17.2919  |
| 27                      | 11.8076                  | 12.8786                 | 14.5733                 | 16.1513                 | 18.1138  |
| 28                      | 12.4613                  | 13.5648                 | 15.3079                 | 16.9279                 | 18.9392  |
| 29                      | 13.1211                  | 14.2565                 | 16.0471                 | 17.7083                 | 19.7677  |
| 30                      | 13.7867                  | 14.9535                 | 16.7908                 | 18.4926                 | 20.5992  |
| 40                      | 20.7065                  | 22.1643                 | 24.4331                 | 26.5093                 | 29.0505  |
| 50                      | 27.9907                  | 29.7067                 | 32.3574                 | 34.7642                 | 37.6886  |
| 60                      | 35.5346                  | 37.4848                 | 40.4817                 | 43.1879                 | 46.4589  |
| 70                      | 43.2752                  | 45.4418                 | 48.7576                 | 51.7393                 | 55.3290  |
| 80                      | 51.1720                  | 53.5400                 | 57.1532                 | 60.3915                 | 64.2778  |
| 90                      | 59.1963                  | 61.7541                 | 65.6466                 | 69.1260                 | 73.2912  |
| 100*                    | 67.3276                  | 70.0648                 | 74.2219                 | 77.9295                 | 82.3581  |

\*For df greater than 100 the expression  $\sqrt{2\chi^2} - \sqrt{(2k-1)} = Z$  follows the standardized normal distribution, where  $k$  represents the degrees of freedom.

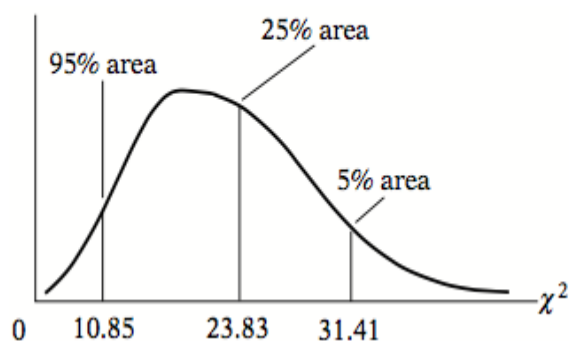
## UPPER PERCENTAGE POINTS OF THE $\chi^2$ DISTRIBUTION

### Example

$$\Pr(\chi^2 > 10.85) = 0.95$$

$$\Pr(\chi^2 > 23.83) = 0.25 \quad \text{for } df = 20$$

$$\Pr(\chi^2 > 31.41) = 0.05$$



| .750     | .500    | .250    | .100    | .050    | .025    | .010    | .005    |
|----------|---------|---------|---------|---------|---------|---------|---------|
| .1015308 | .454937 | 1.32330 | 2.70554 | 3.84146 | 5.02389 | 6.63490 | 7.87944 |
| .575364  | 1.38629 | 2.77259 | 4.60517 | 5.99147 | 7.37776 | 9.21034 | 10.5966 |
| 1.212534 | 2.36597 | 4.10835 | 6.25139 | 7.81473 | 9.34840 | 11.3449 | 12.8381 |
| 1.92255  | 3.35670 | 5.38527 | 7.77944 | 9.48773 | 11.1433 | 13.2767 | 14.8602 |
| 2.67460  | 4.35146 | 6.62568 | 9.23635 | 11.0705 | 12.8325 | 15.0863 | 16.7496 |
| 3.45460  | 5.34812 | 7.84080 | 10.6446 | 12.5916 | 14.4494 | 16.8119 | 18.5476 |
| 4.25485  | 6.34581 | 9.03715 | 12.0170 | 14.0671 | 16.0128 | 18.4753 | 20.2777 |
| 5.07064  | 7.34412 | 10.2188 | 13.3616 | 15.5073 | 17.5346 | 20.0902 | 21.9550 |
| 5.89883  | 8.34283 | 11.3887 | 14.6837 | 16.9190 | 19.0228 | 21.6660 | 23.5893 |
| 6.73720  | 9.34182 | 12.5489 | 15.9871 | 18.3070 | 20.4831 | 23.2093 | 25.1882 |
| 7.58412  | 10.3410 | 13.7007 | 17.2750 | 19.6751 | 21.9200 | 24.7250 | 26.7569 |
| 8.43842  | 11.3403 | 14.8454 | 18.5494 | 21.0261 | 23.3367 | 26.2170 | 28.2995 |
| 9.29906  | 12.3398 | 15.9839 | 19.8119 | 22.3621 | 24.7356 | 27.6883 | 29.8194 |
| 10.1653  | 13.3393 | 17.1170 | 21.0642 | 23.6848 | 26.1190 | 29.1413 | 31.3193 |
| 11.0365  | 14.3389 | 18.2451 | 22.3072 | 24.9958 | 27.4884 | 30.5779 | 32.8013 |
| 11.9122  | 15.3385 | 19.3688 | 23.5418 | 26.2962 | 28.8454 | 31.9999 | 34.2672 |
| 12.7919  | 16.3381 | 20.4887 | 24.7690 | 27.5871 | 30.1910 | 33.4087 | 35.7185 |
| 13.6753  | 17.3379 | 21.6049 | 25.9894 | 28.8693 | 31.5264 | 34.8053 | 37.1564 |
| 14.5620  | 18.3376 | 22.7178 | 27.2036 | 30.1435 | 32.8523 | 36.1908 | 38.5822 |
| 15.4518  | 19.3374 | 23.8277 | 28.4120 | 31.4104 | 34.1696 | 37.5662 | 39.9968 |
| 16.3444  | 20.3372 | 24.9348 | 29.6151 | 32.6705 | 35.4789 | 38.9321 | 41.4010 |
| 17.2396  | 21.3370 | 26.0393 | 30.8133 | 33.9244 | 36.7807 | 40.2894 | 42.7956 |
| 18.1373  | 22.3369 | 27.1413 | 32.0069 | 35.1725 | 38.0757 | 41.6384 | 44.1813 |
| 19.0372  | 23.3367 | 28.2412 | 33.1963 | 36.4151 | 39.3641 | 42.9798 | 45.5585 |
| 19.9393  | 24.3366 | 29.3389 | 34.3816 | 37.6525 | 40.6465 | 44.3141 | 46.9278 |
| 20.8434  | 25.3364 | 30.4345 | 35.5631 | 38.8852 | 41.9232 | 45.6417 | 48.2899 |
| 21.7494  | 26.3363 | 31.5284 | 36.7412 | 40.1133 | 43.1944 | 46.9630 | 49.6449 |
| 22.6572  | 27.3363 | 32.6205 | 37.9159 | 41.3372 | 44.4607 | 48.2782 | 50.9933 |
| 23.5666  | 28.3362 | 33.7109 | 39.0875 | 42.5569 | 45.7222 | 49.5879 | 52.3356 |
| 24.4776  | 29.3360 | 34.7998 | 40.2560 | 43.7729 | 46.9792 | 50.8922 | 53.6720 |
| 33.6603  | 39.3354 | 45.6160 | 51.8050 | 55.7585 | 59.3417 | 63.6907 | 66.7659 |
| 42.9421  | 49.3349 | 56.3336 | 63.1671 | 67.5048 | 71.4202 | 76.1539 | 79.4900 |
| 52.2938  | 59.3347 | 66.9814 | 74.3970 | 79.0819 | 83.2976 | 88.3794 | 91.9517 |
| 61.6983  | 69.3344 | 77.5766 | 85.5271 | 90.5312 | 95.0231 | 100.425 | 104.215 |
| 71.1445  | 79.3343 | 88.1303 | 96.5782 | 101.879 | 106.629 | 112.329 | 116.321 |
| 80.6247  | 89.3342 | 98.6499 | 107.565 | 113.145 | 118.136 | 124.116 | 128.299 |
| 90.1332  | 99.3341 | 109.141 | 118.498 | 124.342 | 129.561 | 135.807 | 140.169 |

Source: Abridged from E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, vol. 1, 3d ed., table 8, Cambridge University Press, New York, 1966. Reproduced by permission of the editors and trustees of *Biometrika*.

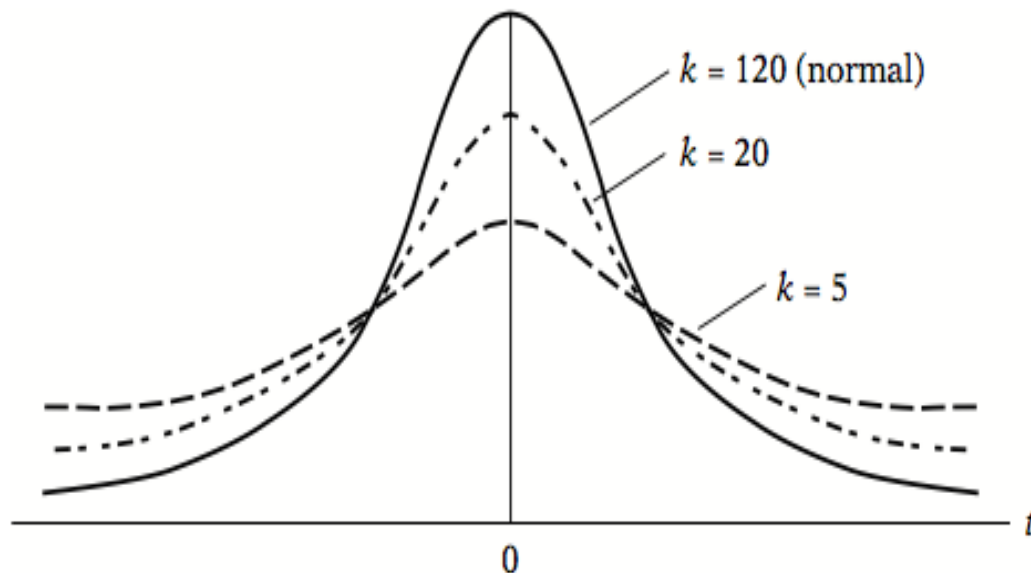
### 6.3 Student's t Distribution

If  $Z_1$  is a standardized normal variable and  $Z_2$  is the chi-square distribution with  $k$  degree of freedom and is distributed independently of  $Z_1$ , then the Student's t distribution ( $t_k$ ) with  $k$  degree of freedom can be represented as

$$\begin{aligned} t &= \frac{Z_1}{\sqrt{(Z_2/k)}} \\ &= \frac{Z_1 \sqrt{k}}{\sqrt{Z_2}} \end{aligned} \quad (\text{Eq.13})$$

**Properties of the Student's t distribution are as follows:**

1. The t distribution is symmetrical, BUT it is flatter than the normal distribution. However, as the df increase, the t distribution is converted to the normal distribution.
2. The mean of the t distribution is zero, and the variance is  $\frac{k}{k-2}$



**Figure 6.** Density function of the student's t distribution

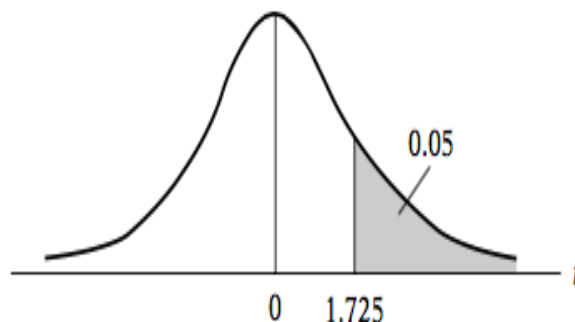
## PERCENTAGE POINTS OF THE $t$ DISTRIBUTION

### Example

$$\Pr(t > 2.086) = 0.025$$

$$\Pr(t > 1.725) = 0.05 \quad \text{for } df = 20$$

$$\Pr(|t| > 1.725) = 0.10$$



| Pr \ df | 0.25<br>0.50 | 0.10<br>0.20 | 0.05<br>0.10 | 0.025<br>0.05 | 0.01<br>0.02 | 0.005<br>0.010 | 0.001<br>0.002 |
|---------|--------------|--------------|--------------|---------------|--------------|----------------|----------------|
| 1       | 1.000        | 3.078        | 6.314        | 12.706        | 31.821       | 63.657         | 318.31         |
| 2       | 0.816        | 1.886        | 2.920        | 4.303         | 6.965        | 9.925          | 22.327         |
| 3       | 0.765        | 1.638        | 2.353        | 3.182         | 4.541        | 5.841          | 10.214         |
| 4       | 0.741        | 1.533        | 2.132        | 2.776         | 3.747        | 4.604          | 7.173          |
| 5       | 0.727        | 1.476        | 2.015        | 2.571         | 3.365        | 4.032          | 5.893          |
| 6       | 0.718        | 1.440        | 1.943        | 2.447         | 3.143        | 3.707          | 5.208          |
| 7       | 0.711        | 1.415        | 1.895        | 2.365         | 2.998        | 3.499          | 4.785          |
| 8       | 0.706        | 1.397        | 1.860        | 2.306         | 2.896        | 3.355          | 4.501          |
| 9       | 0.703        | 1.383        | 1.833        | 2.262         | 2.821        | 3.250          | 4.297          |
| 10      | 0.700        | 1.372        | 1.812        | 2.228         | 2.764        | 3.169          | 4.144          |
| 11      | 0.697        | 1.363        | 1.796        | 2.201         | 2.718        | 3.106          | 4.025          |
| 12      | 0.695        | 1.356        | 1.782        | 2.179         | 2.681        | 3.055          | 3.930          |
| 13      | 0.694        | 1.350        | 1.771        | 2.160         | 2.650        | 3.012          | 3.852          |
| 14      | 0.692        | 1.345        | 1.761        | 2.145         | 2.624        | 2.977          | 3.787          |
| 15      | 0.691        | 1.341        | 1.753        | 2.131         | 2.602        | 2.947          | 3.733          |
| 16      | 0.690        | 1.337        | 1.746        | 2.120         | 2.583        | 2.921          | 3.686          |
| 17      | 0.689        | 1.333        | 1.740        | 2.110         | 2.567        | 2.898          | 3.646          |
| 18      | 0.688        | 1.330        | 1.734        | 2.101         | 2.552        | 2.878          | 3.610          |
| 19      | 0.688        | 1.328        | 1.729        | 2.093         | 2.539        | 2.861          | 3.579          |
| 20      | 0.687        | 1.325        | 1.725        | 2.086         | 2.528        | 2.845          | 3.552          |
| 21      | 0.686        | 1.323        | 1.721        | 2.080         | 2.518        | 2.831          | 3.527          |
| 22      | 0.686        | 1.321        | 1.717        | 2.074         | 2.508        | 2.819          | 3.505          |
| 23      | 0.685        | 1.319        | 1.714        | 2.069         | 2.500        | 2.807          | 3.485          |
| 24      | 0.685        | 1.318        | 1.711        | 2.064         | 2.492        | 2.797          | 3.467          |
| 25      | 0.684        | 1.316        | 1.708        | 2.060         | 2.485        | 2.787          | 3.450          |
| 26      | 0.684        | 1.315        | 1.706        | 2.056         | 2.479        | 2.779          | 3.435          |
| 27      | 0.684        | 1.314        | 1.703        | 2.052         | 2.473        | 2.771          | 3.421          |
| 28      | 0.683        | 1.313        | 1.701        | 2.048         | 2.467        | 2.763          | 3.408          |
| 29      | 0.683        | 1.311        | 1.699        | 2.045         | 2.462        | 2.756          | 3.396          |
| 30      | 0.683        | 1.310        | 1.697        | 2.042         | 2.457        | 2.750          | 3.385          |
| 40      | 0.681        | 1.303        | 1.684        | 2.021         | 2.423        | 2.704          | 3.307          |
| 60      | 0.679        | 1.296        | 1.671        | 2.000         | 2.390        | 2.660          | 3.232          |
| 120     | 0.677        | 1.289        | 1.658        | 1.980         | 2.358        | 2.617          | 3.160          |
| ∞       | 0.674        | 1.282        | 1.645        | 1.960         | 2.326        | 2.576          | 3.090          |

## 6.4 The F Distribution

If  $Z_1$  and  $Z_2$  are independently distributed chi-square variables with  $k_1$  and  $k_2$  df, respectively, the (Fisher's) F distribution with  $k_1$  and  $k_2$  df can be written as

$$F = \frac{Z_1/k_1}{Z_2/k_2}$$

**The F distribution has the following properties:**

1. The F distribution is skewed to the right, but as  $k_1$  and  $k_2$  become large, the F distribution is converted to normal distribution.
2. The mean value of an F-distributed variable is  $\frac{k_2}{(k_2-2)}$ , and its variance is

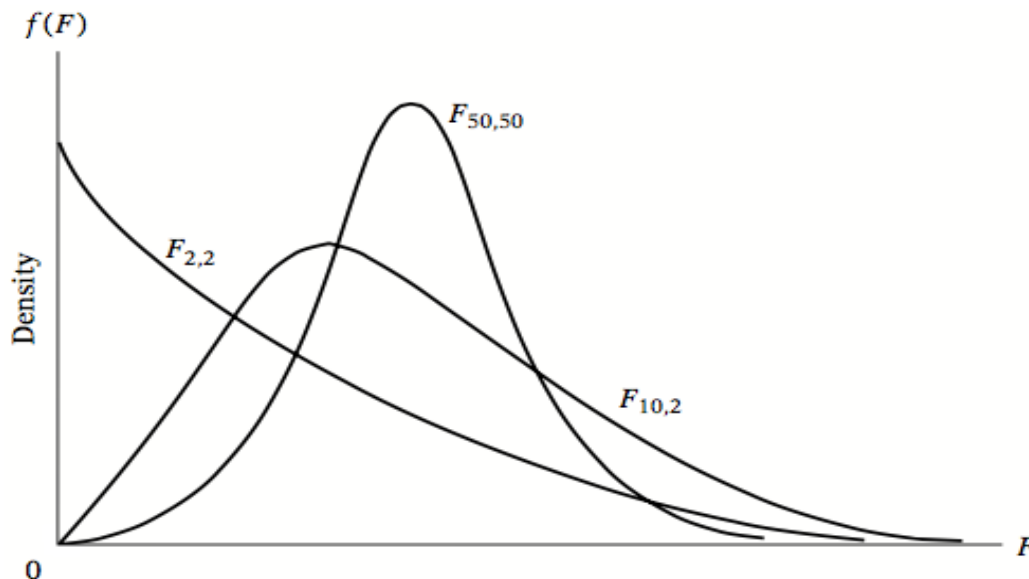
$$\frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$$

3. The square of a t-distributed random variable with  $k$  df is equivalent to an F distribution with 1 and  $k$  df.

$$t_k^2 = F_{1,k}$$

4. If the denominator df,  $k_2$ , is fairly large, we can get the following relationship

$$k_1 F \sim \chi_{k_1}^2$$



**Figure 7.** Density function of F distribution

## Lecture 5

### CHAPTER 2: TWO-VARIABLE REGRESSION ANALYSIS

#### 2.1 Example

In order to understand two-variable regression, consider the data given in Table 1.

The data in the below table refer to a total **Population** of 42 families with their weekly income (X) and weekly consumption expenditure (Y).

**Table 1.** Weekly family Expenditure (Y), Baht and Income (X), Baht

|                                        | X=Weekly family Income, Baht |      |      |      |      |      |
|----------------------------------------|------------------------------|------|------|------|------|------|
|                                        | 500                          | 600  | 700  | 800  | 900  | 1000 |
| Y= Weekly<br>Family Expenditure        | 360                          | 376  | 458  | 610  | 600  | 700  |
|                                        | 313                          | 475  | 422  | 468  | 531  | 679  |
|                                        | 322                          | 380  | 498  | 575  | 670  | 730  |
|                                        | 310                          | 382  | 560  | 542  | 630  | 591  |
|                                        | 390                          | 390  | 442  | 588  | 544  | 550  |
|                                        | 315                          | 425  | 440  | 466  | 565  | 620  |
|                                        | 390                          | 442  | -    | 461  | -    | 695  |
|                                        | 400                          | -    | -    | -    | -    | 635  |
| Total                                  | 2800                         | 2870 | 2820 | 3710 | 3540 | 5200 |
| Conditional<br>means of Y,<br>$E(Y X)$ | 350                          | 410  | 470  | 530  | 590  | 650  |
| Notes -                                |                              |      |      |      |      |      |

**Conditional expected value of weekly consumption expenditure given the income level =X ,  $E(Y|X)$**

**Unconditional expected value , $E(Y)$**

**Table 2.** Conditional Probabilities  $p(Y|X_i)$  for the Weekly Family Income (X) and Expenditure (Y)

|                                  |  | X=Weekly family Income, Baht |     |     |     |     |      |
|----------------------------------|--|------------------------------|-----|-----|-----|-----|------|
|                                  |  | 500                          | 600 | 700 | 800 | 900 | 1000 |
| Y= Weekly<br>Family Expenditure  |  | 1/8                          | 1/7 | 1/6 | 1/7 | 1/6 | 1/8  |
|                                  |  | 1/8                          | 1/7 | 1/6 | 1/7 | 1/6 | 1/8  |
|                                  |  | 1/8                          | 1/7 | 1/6 | 1/7 | 1/6 | 1/8  |
|                                  |  | 1/8                          | 1/7 | 1/6 | 1/7 | 1/6 | 1/8  |
|                                  |  | 1/8                          | 1/7 | 1/6 | 1/7 | 1/6 | 1/8  |
|                                  |  | 1/8                          | 1/7 | 1/6 | 1/7 | 1/6 | 1/8  |
|                                  |  | 1/8                          | 1/7 | -   | 1/7 | -   | 1/8  |
|                                  |  | 1/8                          | -   | -   | -   | -   | 1/8  |
| Conditional means of Y, $E(Y X)$ |  | 350                          | 410 | 470 | 530 | 590 | 650  |

Notes -

**Figure 2.1: Conditional Distribution of Expenditure for Various Levels of Income**

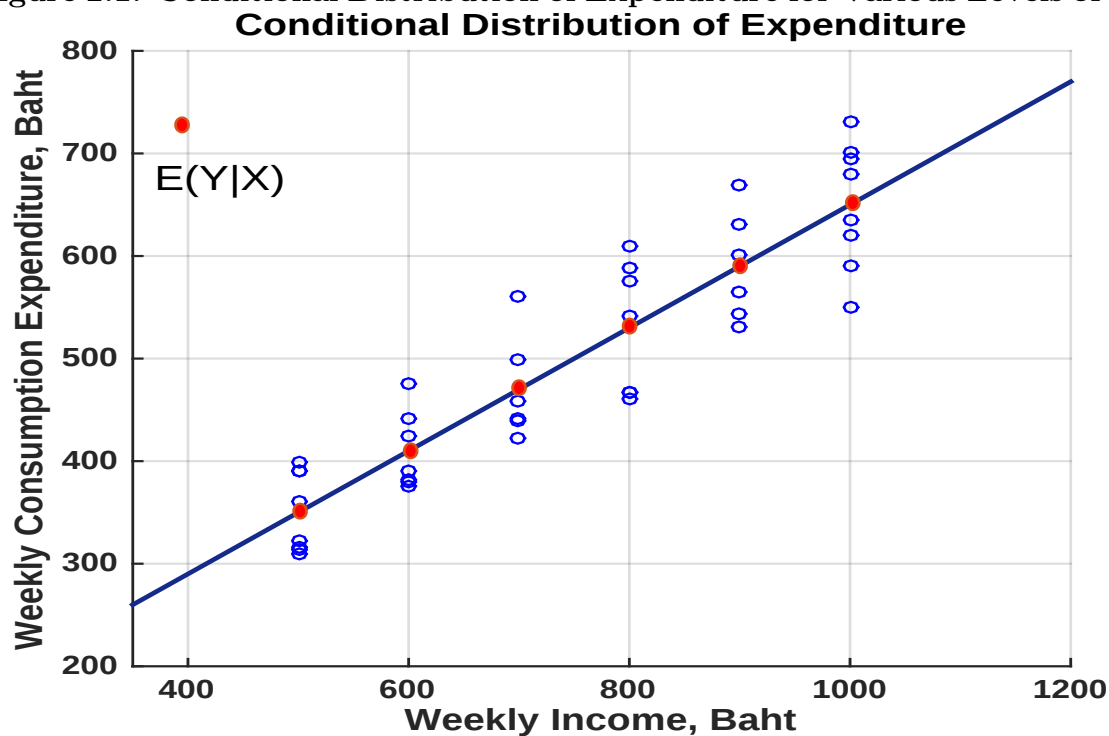
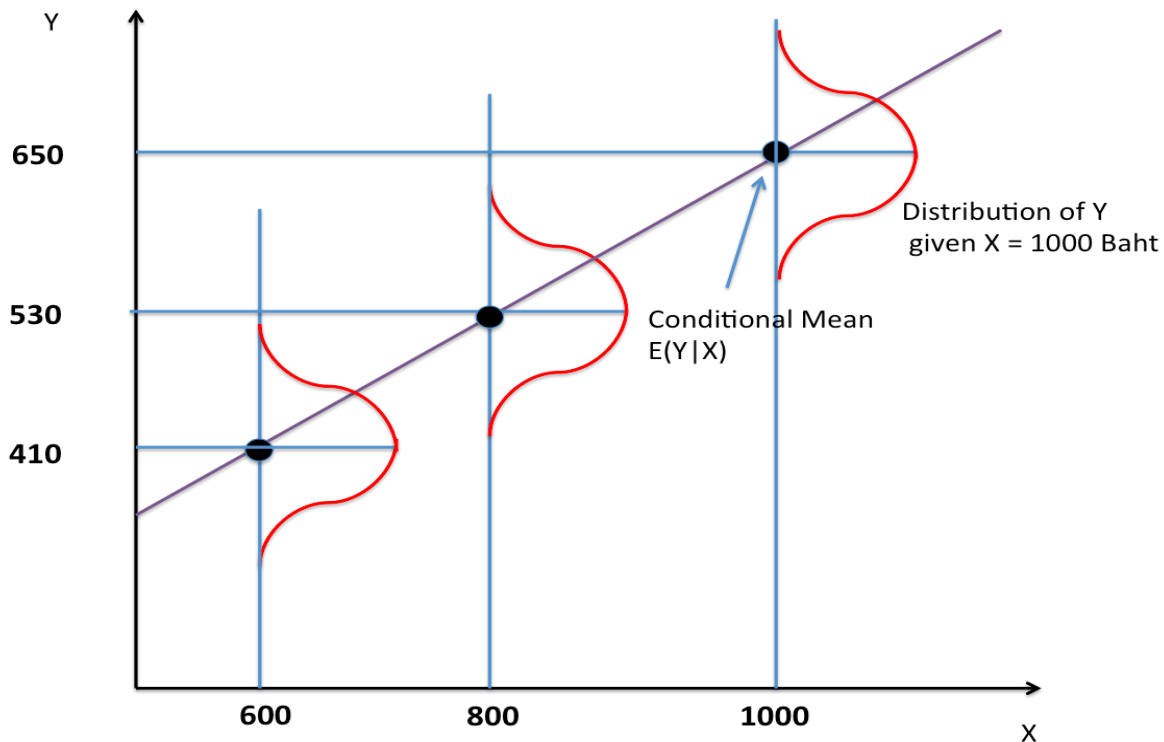


Figure 2.2: Population Regression Line (PRL)



## 2.2 The Concept of Population Regression Function (PRF)

The population regression function (PRF) can be written as the function of  $X_i$ :

What form does the function  $f(X_i)$  assume?

If we assume the PRF  $E(Y|X_i)$  is a linear function of  $X_i$ , we get

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

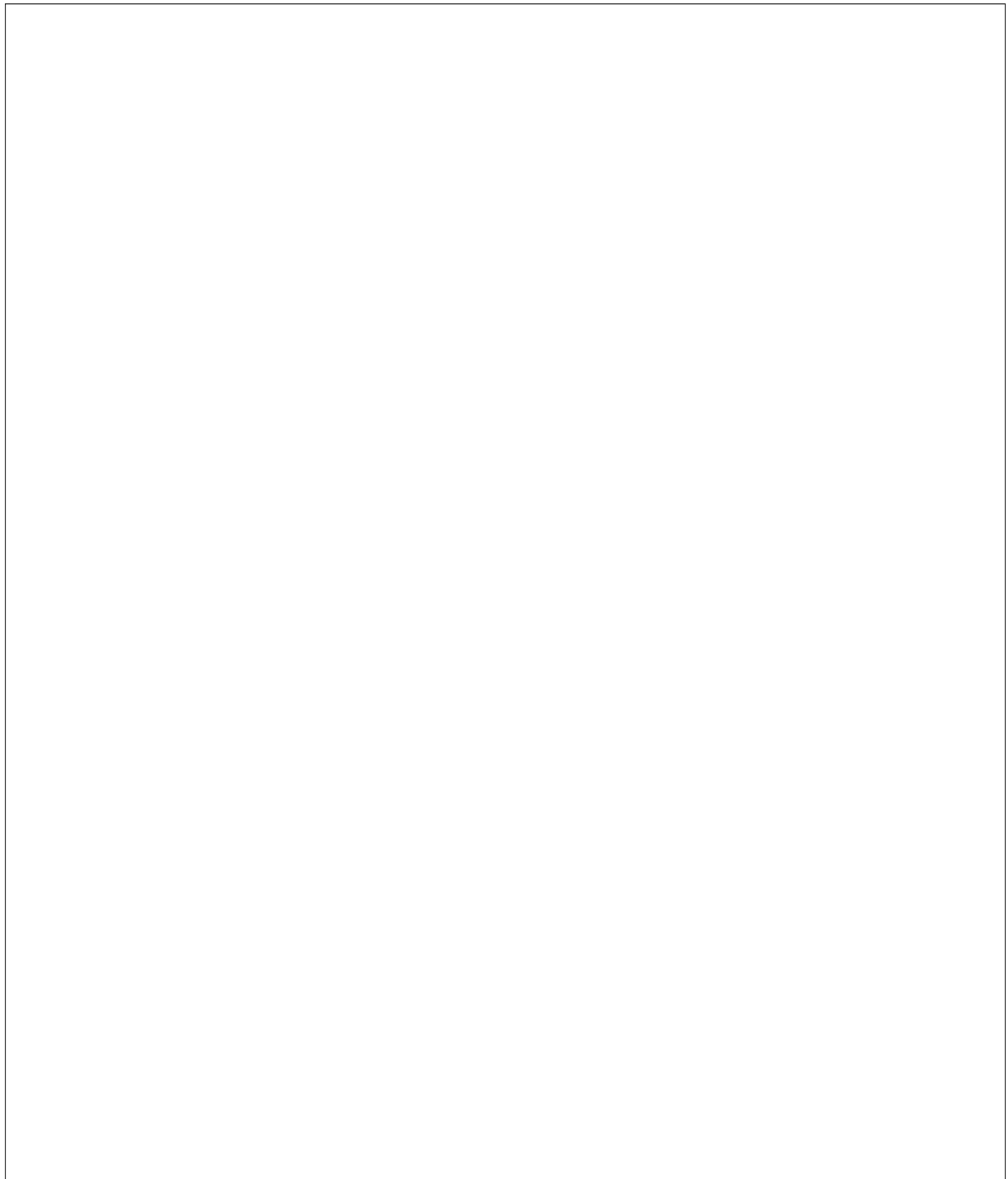
## 2.3 What is the meaning of the term LINEAR?

**LINEARITY in the variables**

**LINEARITY in the parameters**

## 2.4 Stochastic Specification of PRF

We can write the **deviation** of an individual  $Y_i$  around its expected value as follows:



## **2.5 The roles of the stochastic disturbance term**

1. Vagueness of theory
2. Unavailability of data
3. Core variables versus peripheral variables
4. Intrinsic randomness in human behavior
5. Poor proxy variable
6. Principle of parsimony
7. Wrong functional form

## Lecture 6

### 2.6 The Sample Regression Function (SRF)

As mentioned, in the real situation, we cannot find out all the population of Y values corresponding to the fixed X's. We only have a sample of Y values corresponding to some fixed X's.

Therefore, our goal in this section is to estimate the population regression line (PRF) on the basis of the **SAMPLE INFORMATION**.

As a result, for the fixed X's as given in table 1, we only have a randomly selected sample of Y values. For example, table 5 and table 4 show a random sample from the population of table 1

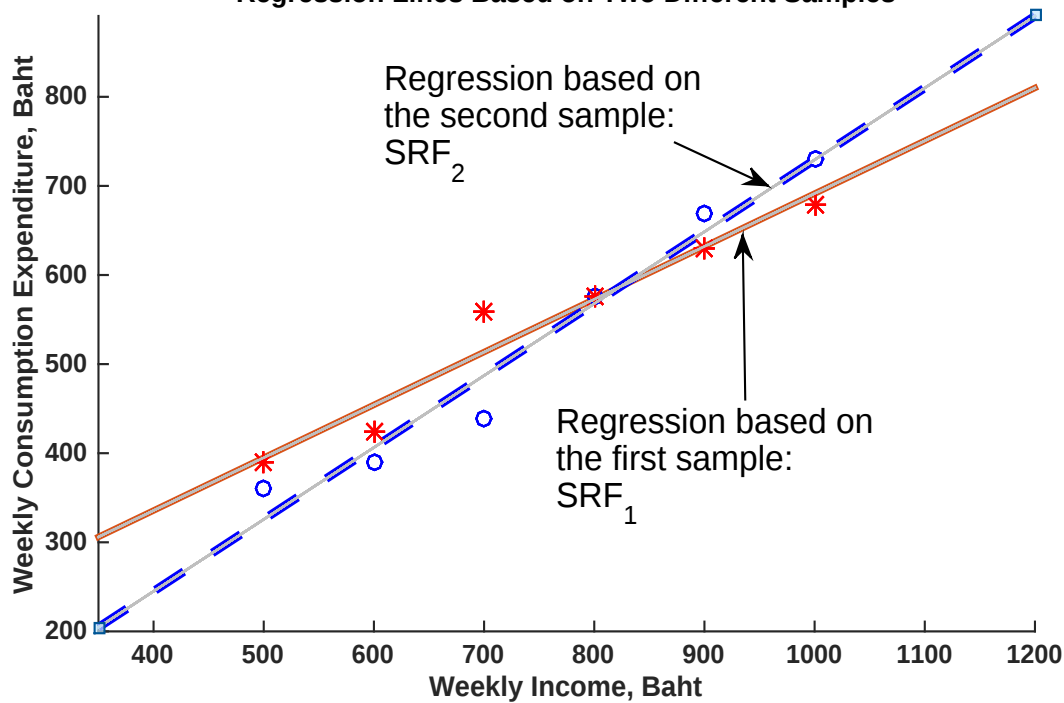
**Table 3.** A Random Sample From the Population

| <b>X</b> | <b>Y</b> |
|----------|----------|
| 500      | 390      |
| 600      | 425      |
| 700      | 560      |
| 800      | 575      |
| 900      | 630      |
| 1000     | 679      |

**Table 4.** Another Random Sample From the Population

| <b>X</b> | <b>Y</b> |
|----------|----------|
| 500      | 360      |
| 600      | 390      |
| 700      | 440      |
| 800      | 575      |
| 900      | 670      |
| 1000     | 730      |

Figure 2.3: Regression lines based on two different samples  
**Regression Lines Based on Two Different Samples**



The sample regression function (SRF) can be written as:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

where  $\hat{Y}$  is read as “Y-hat”

$\hat{Y}_i$  = estimator of  $E(Y|X_i)$

$\hat{\beta}_1$ =estimator of  $\beta_1$

$\hat{\beta}_2$ =estimator of  $\beta_2$

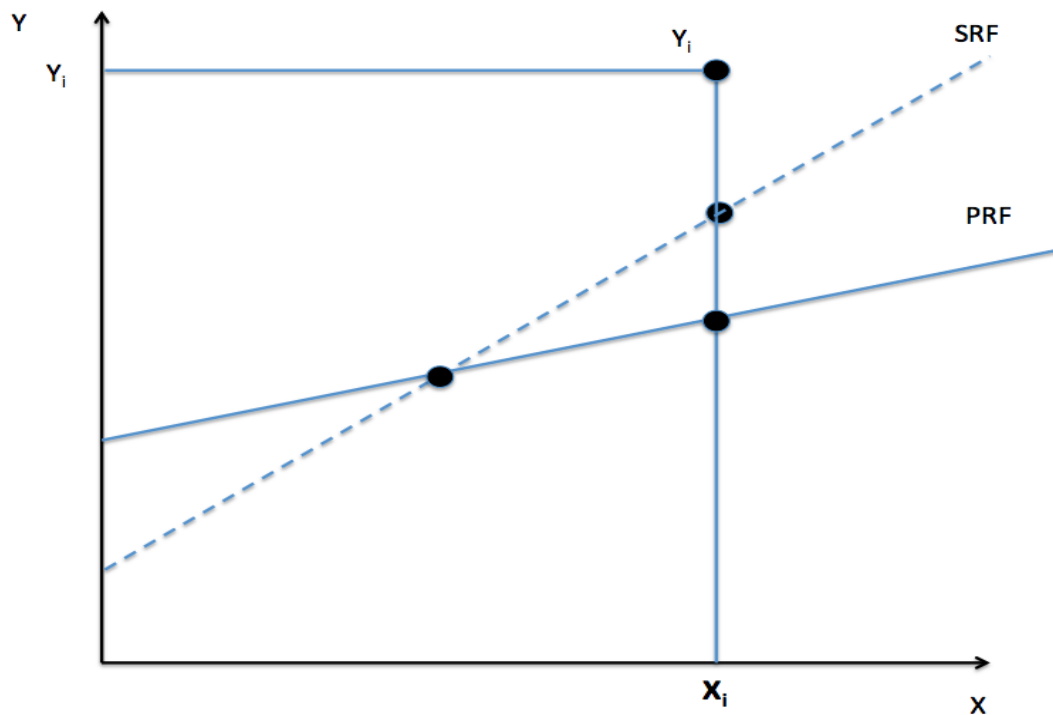
We can express the SRF in its stochastic form as follows:

$$Y_i = \beta_1 + \beta_2 X_i + \mu_i$$

In sum, our ultimate goal is to estimate  
**the PRF**

on the basis of  
**the SRF**

Figure 2.4: Sample and Population Regression Lines



## CHAPTER 3: TWO-VARIABLE REGRESSION MODEL: THE PROBLEM OF ESTIMATION

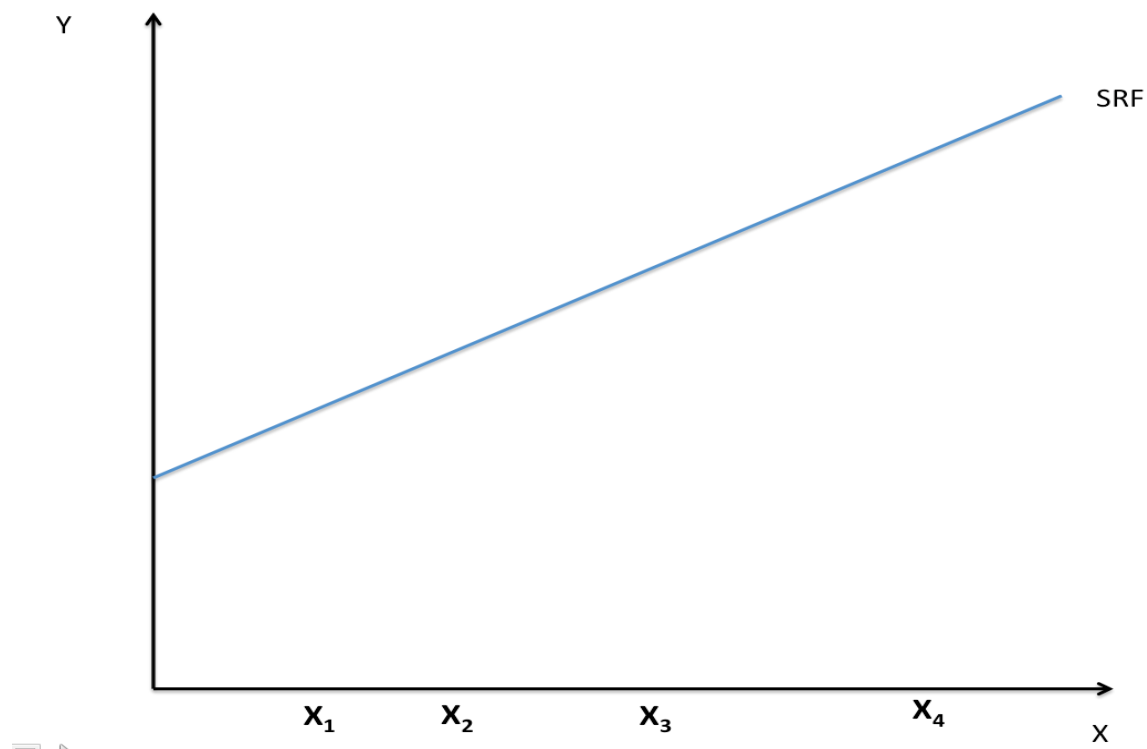
As mentioned in the previous chapter, our main objective is to estimate the population regression function (PRF) based on the basis of the sample regression function (SRF) as accurately as possible.

In this chapter, we are going to discuss two methods of estimation:

- (1) Ordinary Least Squares (OLS) and
- (2) Maximum Likelihood (ML).

### 3.1 The Method of Ordinary Least Squares (OLS)

Figure 3.1: Least-Squares Criterion



## The Method to Find Out the Least-Squares Estimators: $\hat{\beta}_1$ and $\hat{\beta}_2$

## Lecture 7

From the SRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

Now, we obtain the **least-squares estimators**:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} \\ &= \bar{Y} - \hat{\beta}_2 \bar{X}\end{aligned}\tag{Eq.1}$$

$$\hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}\tag{Eq.2}$$

If we define  $\bar{X}$  and  $\bar{Y}$  to be the sample means of X and Y. Then:

$$\begin{aligned}x_i &= (X_i - \bar{X}) \\ y_i &= (Y_i - \bar{Y})\end{aligned}\tag{Eq.3}$$

We can have the alternative expressions for  $\hat{\beta}_2$ :

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} \\ &= \frac{\sum x_i Y_i}{\sum X_i^2 - n \bar{X}^2} \\ &= \frac{\sum X_i y_i}{\sum X_i^2 - n \bar{X}^2}\end{aligned}\tag{Eq.4}$$

Show that

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

**EXAMPLE**

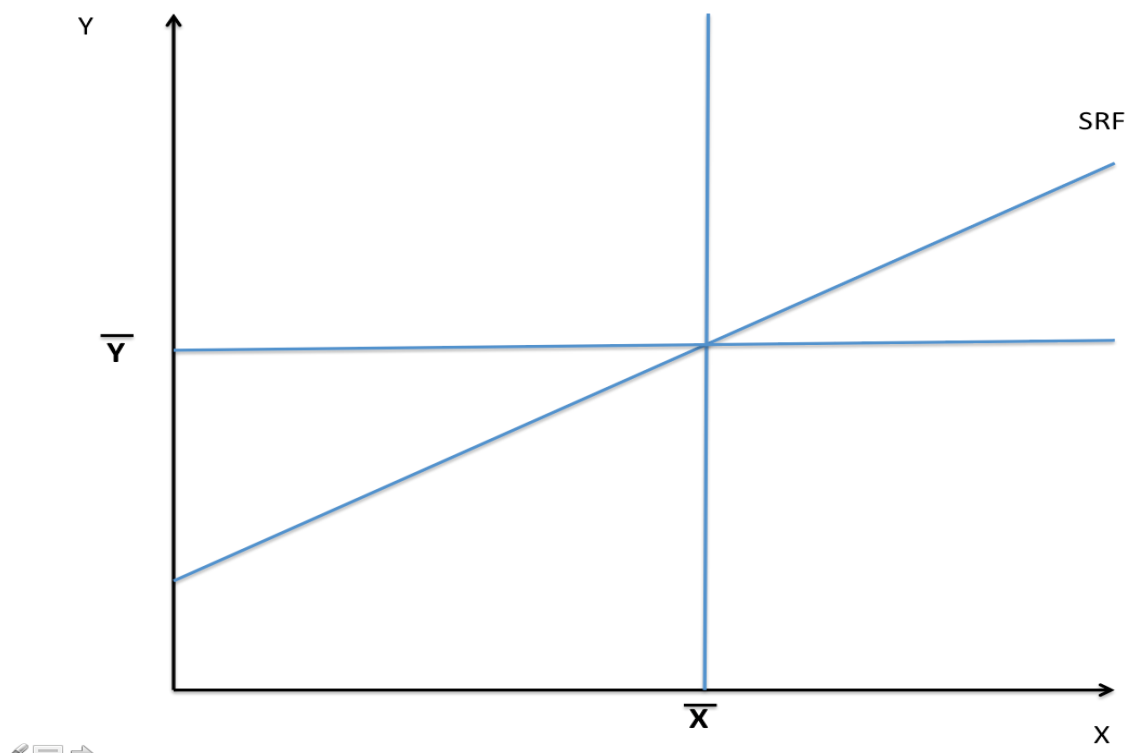
**Table 5.** A Random Sample From the Population

| <b>X</b> | <b>Y</b> |
|----------|----------|
| 500      | 390      |
| 600      | 425      |
| 700      | 560      |
| 800      | 575      |
| 900      | 630      |
| 1000     | 679      |

**Table 6.** Raw Data Based on the Sample Data on Table 5

| (1)<br>$Y_i$ | (2)<br>$X_i$ | (3)<br>$Y_i X_i$ | (4)<br>$X_i^2$ | (5)<br>$x_i = X_i - \bar{X}$ | (6)<br>$y_i = Y_i - \bar{Y}$ | (7)<br>$x_i^2$ | (8)<br>$x_i y_i$ | (9)<br>$\hat{Y}_i$ | (10)<br>$\hat{u}_i = Y_i - \hat{Y}_i$ | (11)<br>$\hat{Y}_i \hat{u}_i$ |
|--------------|--------------|------------------|----------------|------------------------------|------------------------------|----------------|------------------|--------------------|---------------------------------------|-------------------------------|
| 390          | 500          |                  |                |                              |                              |                |                  |                    |                                       |                               |
| 425          | 600          |                  |                |                              |                              |                |                  |                    |                                       |                               |
| 560          | 700          |                  |                |                              |                              |                |                  |                    |                                       |                               |
| 575          | 800          |                  |                |                              |                              |                |                  |                    |                                       |                               |
| 630          | 900          |                  |                |                              |                              |                |                  |                    |                                       |                               |
| 679          | 1000         |                  |                |                              |                              |                |                  |                    |                                       |                               |
| Sum          |              |                  |                |                              |                              |                |                  |                    |                                       |                               |
| Mean         |              |                  |                |                              |                              |                |                  |                    |                                       |                               |
| 750          |              |                  |                |                              |                              |                |                  |                    |                                       |                               |

Figure 3.2: Sample Regression Line Based on the Data of Table 5



## Lecture 8

### The numerical and statistical properties of OLS estimators

1. The OLS estimators  $\hat{\beta}_1$  and  $\hat{\beta}_2$  are expressed solely in terms of the observable (Sample size) and quantities (i.e X and Y).

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} \\ &= \bar{Y} - \hat{\beta}_2 \bar{X}\end{aligned}\tag{Eq.5}$$

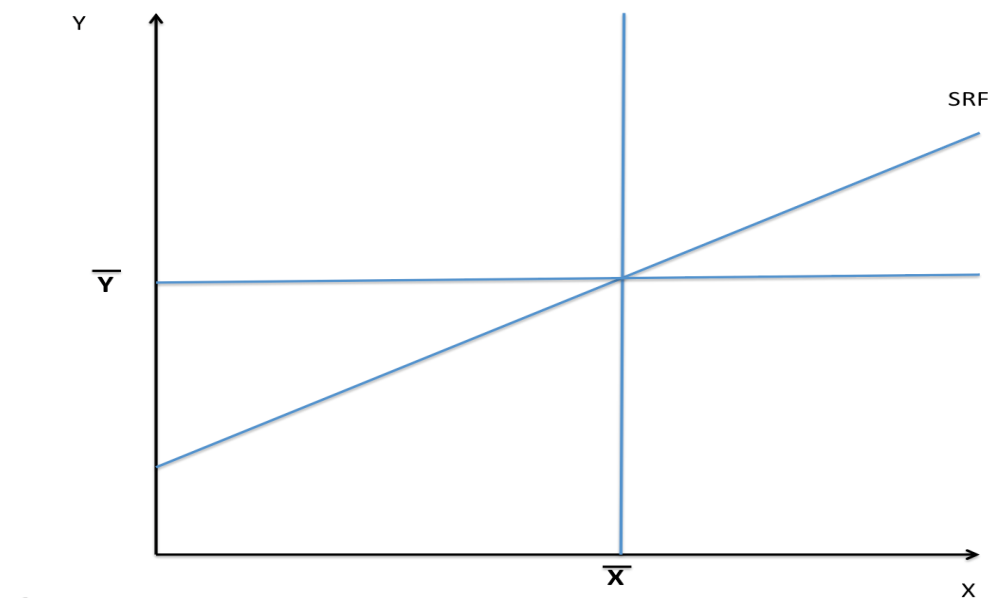
$$\hat{\beta}_2 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}\tag{Eq.6}$$

2. They are **point estimators**.

3. The regression line has the following properties.

3.1 The sample regression function (SRF) passes through the sample means of Y and X ( $\bar{Y}$  and  $\bar{X}$ ).

**Figure 3.3: The Sample regression Line Passes through the Sample Mean Values of Y and X**



3.2 The mean value of the estimated  $Y = \hat{Y}_i$  is equal to the mean value of the actual  $Y$ .

3.3. The mean value of the residuals  $\hat{u}_i$  is zero.

3.4 The residuals  $\hat{u}_i$  are uncorrelated with the predicted  $\hat{Y}_i$ .

3.5 The residuals  $\hat{u}_i$  are uncorrelated with  $X_i$ .

## 3.2 The Assumptions Underlying the Method of Least Squares

**Assumption 1: Linear regression model**

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

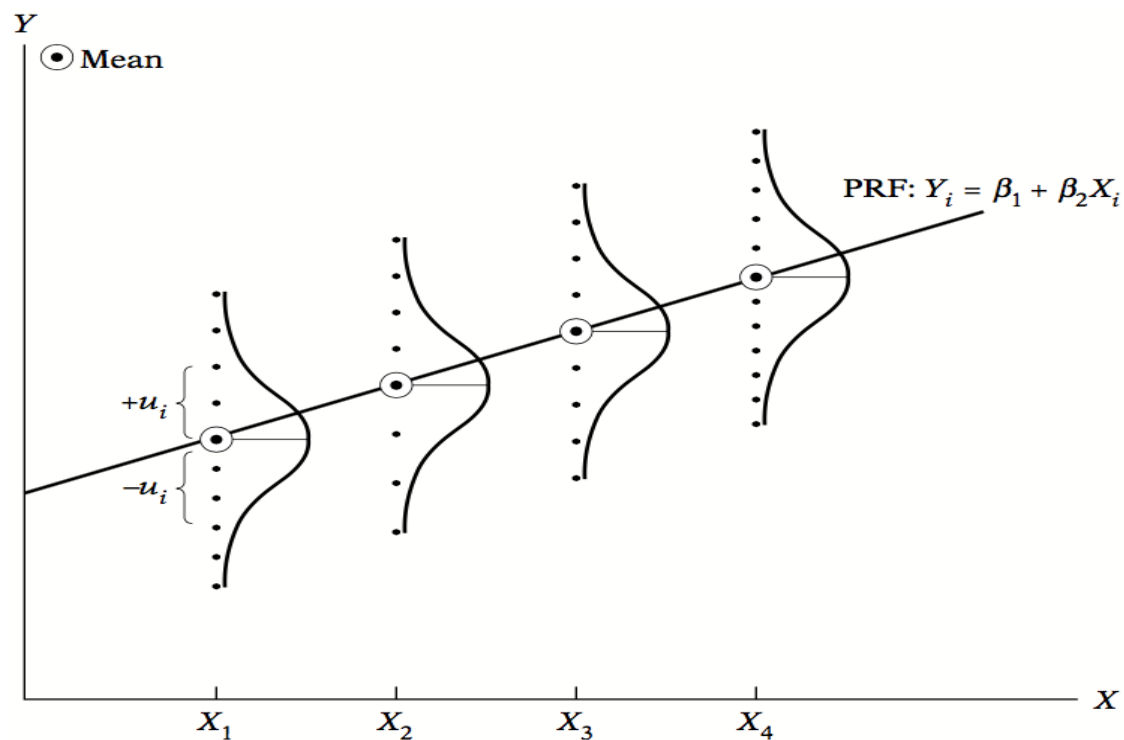
**Assumption 2: X values are fixed in repeated sampling**

X is assumed to be nonstochastic.

**Assumption 3: Zero mean value of disturbance  $u_i$**

$$E(u_i | X_i) = 0$$

Figure 3.4: Conditional Distribution of the Disturbances  $u_i$



Assumption 4: Homoscedasticity or Equal Variance of  $u_i$

Figure 3.5: Homoscedasticity

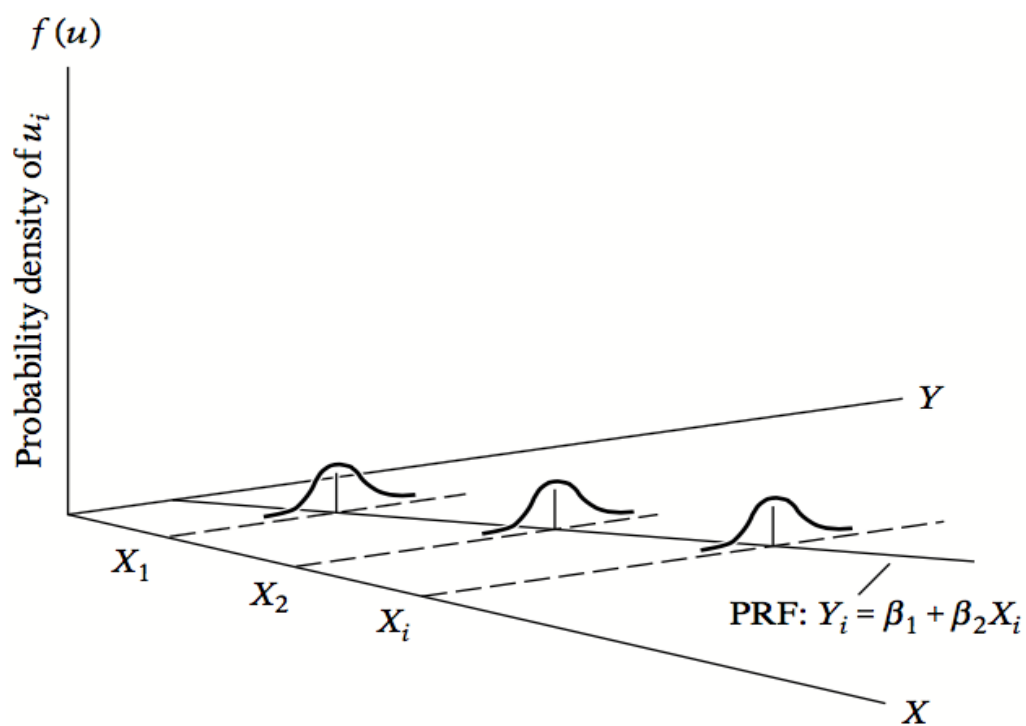
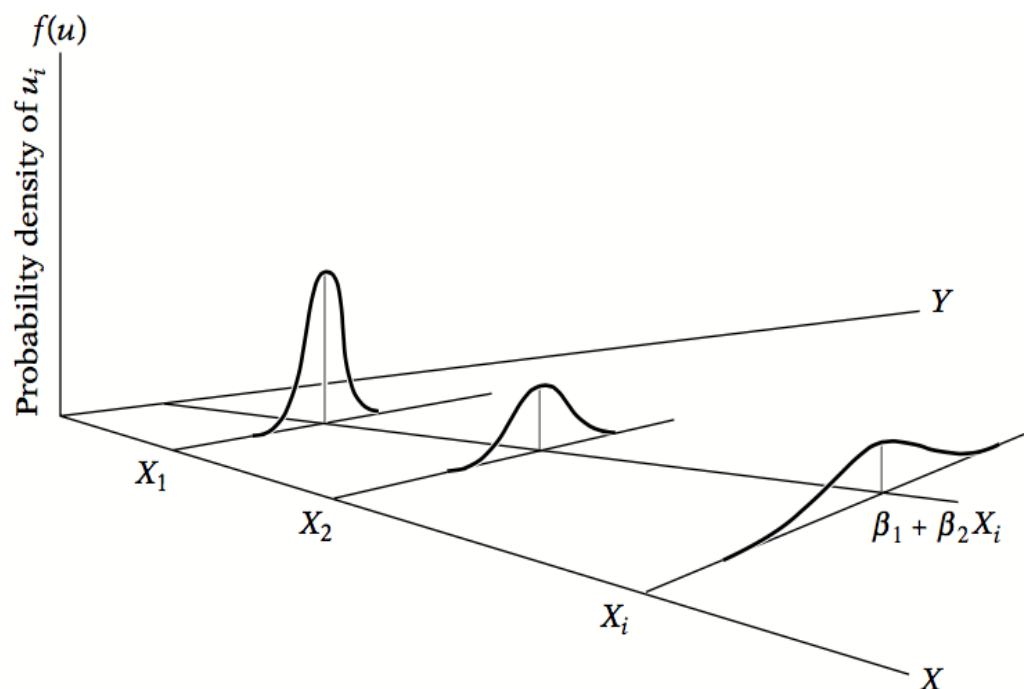


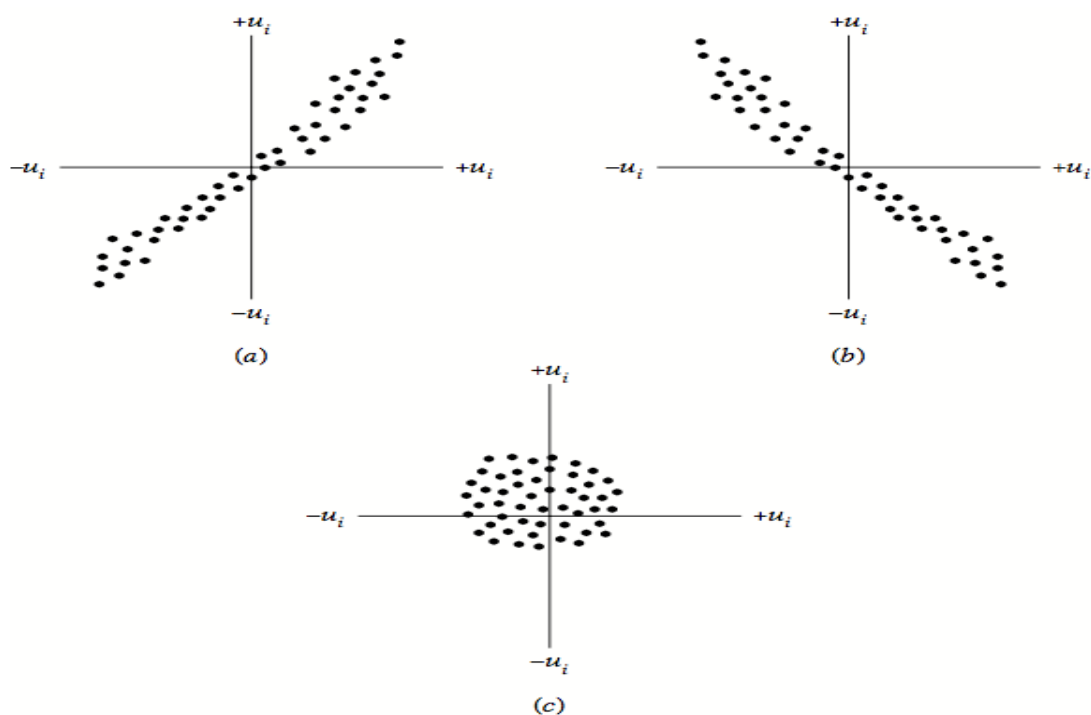
Figure 3.6: Heteroscedasticity



**Assumption 5: No Autocorrelation Between the Disturbances**

**Assumption 6: Zero Covariance Between  $u_i$  and  $X_i$**

Figure 3.7: Patterns of Correlation Among the disturbances



**Assumption 7:** The number of observations  $n$  must be greater than the number of parameters to be estimated.

**Assumption 8:** Variability in  $X$  values.

**Assumption 9:** The regression model is correctly specified.

**Assumption 10:** There is no perfect multicollinearity.

### 3.3 Standard Errors of Least-Squares Estimates

The standard errors of the OLS estimates can be obtained as follows:  
We know that

$$\hat{\beta}_2 = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum k_i Y_i$$

where

$$k_i = \frac{x_i}{\sum x_i^2}$$

**The properties of the weights  $k_i$**

1. The  $k_i$  are nonstochastic.
2.  $\sum k_i = 0$
3.  $\sum k_i^2 = \frac{1}{\sum x_i^2}$
4.  $\sum k_i x_i = \sum k_i X_i = 1$

Since

$$var(\hat{\beta}_2) = E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2$$

**First Step**

Find the  $E(\hat{\beta}_2)$

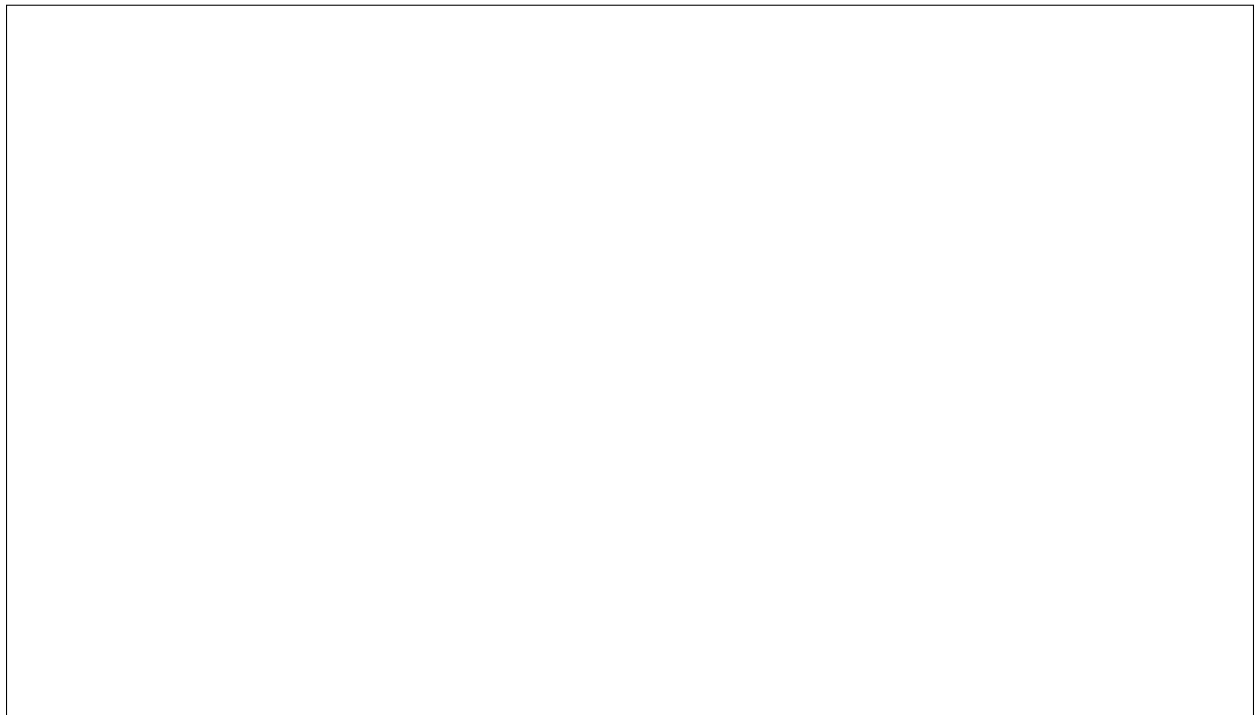
## Second Step

Using the definition of variance

$$\text{var}(\hat{\beta}_2) = E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2$$



The covariance between  $\hat{\beta}_1$  and  $\hat{\beta}_2$



## The Least-Square Estimator of $\sigma^2$



## Lecture 9

In sum, the standard errors of the OLS estimators can be obtained as follow:

$$\begin{aligned}\text{var}(\hat{\beta}_2) &= \frac{\sigma^2}{\sum x_i^2} \\ \text{se}(\hat{\beta}_2) &= \frac{\sigma}{\sqrt{\sum x_i^2}}\end{aligned}\tag{Eq.7}$$

$$\begin{aligned}\text{var}(\hat{\beta}_1) &= \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2 \\ \text{se}(\hat{\beta}_1) &= \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}} \sigma\end{aligned}\tag{Eq.8}$$

We can estimate the  $\sigma^2$  from the data where the formula for the estimated  $\sigma^2$  is following :

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n - 2}$$

where

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2$$

The alternative expression for computing  $\sum \hat{u}_i^2$  is

$$\sum \hat{u}_i^2 = \sum y_i^2 - \frac{(\sum x_i y_i)^2}{\sum x_i^2}$$

The covariance between  $\hat{\beta}_1$  and  $\hat{\beta}_2$  is:

$$\begin{aligned}\text{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\bar{X} \text{var}(\hat{\beta}_2) \\ &= -\bar{X} \left( \frac{\sigma^2}{\sum x_i^2} \right)\end{aligned}\tag{Eq.9}$$

### 3.4 Properties of Least-Squares Estimators: The Gauss-Markov Theorem

Given the assumptions of the classical linear regression model, the least-square estimators are satisfied the optimum properties which is known as “**The Gauss- Markov Theorem.**” To understand this theorem, we need to know the small-sample properties of an estimator first.

#### The Small-Sample Properties of An Estimator

##### 1. Unbiasedness

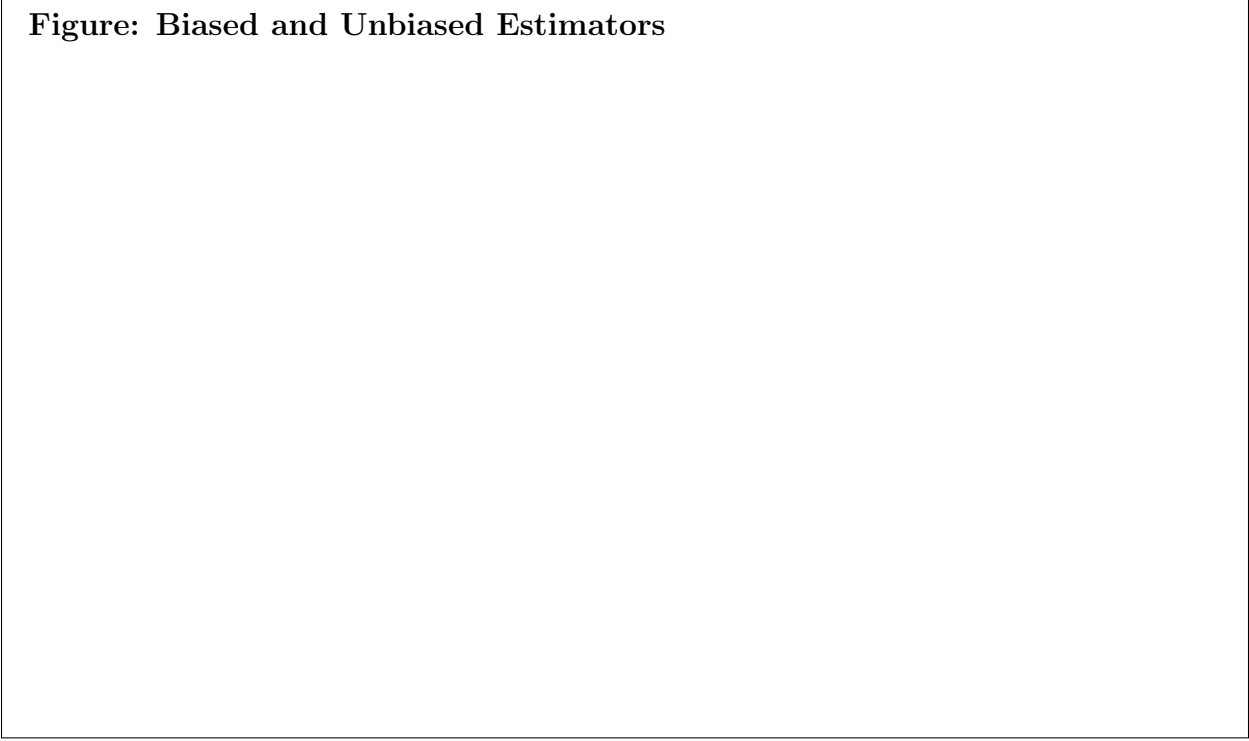
An estimator  $\hat{\theta}$  is said to be an unbiased estimator of  $\theta$  if the expected value of  $\hat{\theta}$  is equal to the true  $\theta$

$$E(\hat{\theta}) = \theta$$

Therefore, if the expected value of  $\hat{\theta}$  is not equal to the true  $\theta$ , then the estimator is said to be biased. We can calculate the biased as:

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

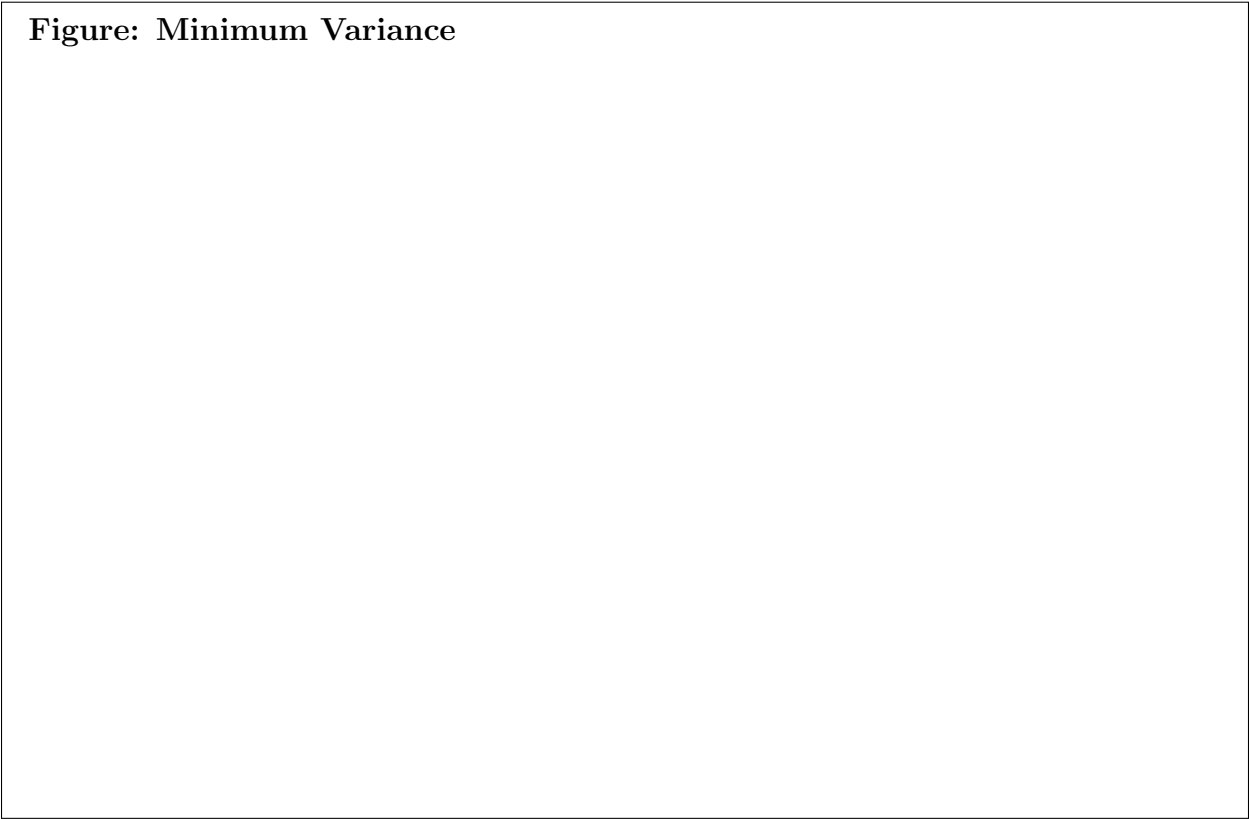
**Figure: Biased and Unbiased Estimators**



## 2. Minimum Variance

$\hat{\theta}_1$  is said to be a minimum variance estimator of  $\theta$  if the variance of  $\hat{\theta}_1$  is smaller than or at most equal to the variance of  $\hat{\theta}_2$ , which is any other estimator of  $\theta$

**Figure: Minimum Variance**



## 3. Best Unbiased or Efficient Estimator = property 1 + property 2

If  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are two unbiased estimators of  $\theta$  and the variance of  $\hat{\theta}_1$  is smaller than or at most equal to the variance of  $\hat{\theta}_2$ , then  $\hat{\theta}_1$  is a **minimum-variance unbiased estimator or best unbiased estimator**.

## 4. Linearity

An estimator  $\hat{\theta}$  is said to be a linear estimator of  $\theta$  if it is a linear function of the sample observations. For example:

$$\bar{X} = \frac{1}{n} \sum X_i = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

Thus,  $\bar{X}$  is a linear estimator because it is a linear function of the  $X$  values.

## Best Linear Unbiased Estimators : BLUE

The estimator  $\hat{\theta}$  is called as the Best Linear Unbiased Estimator **BLUE** if it is satisfied the properties 1,2,4 that is  $\hat{\theta}$  is linear, is unbiased, and has the minimum variance in the class of all linear unbiased estimators of  $\theta$ .

## Minimum Mean-Square-Error (MSE) Estimator

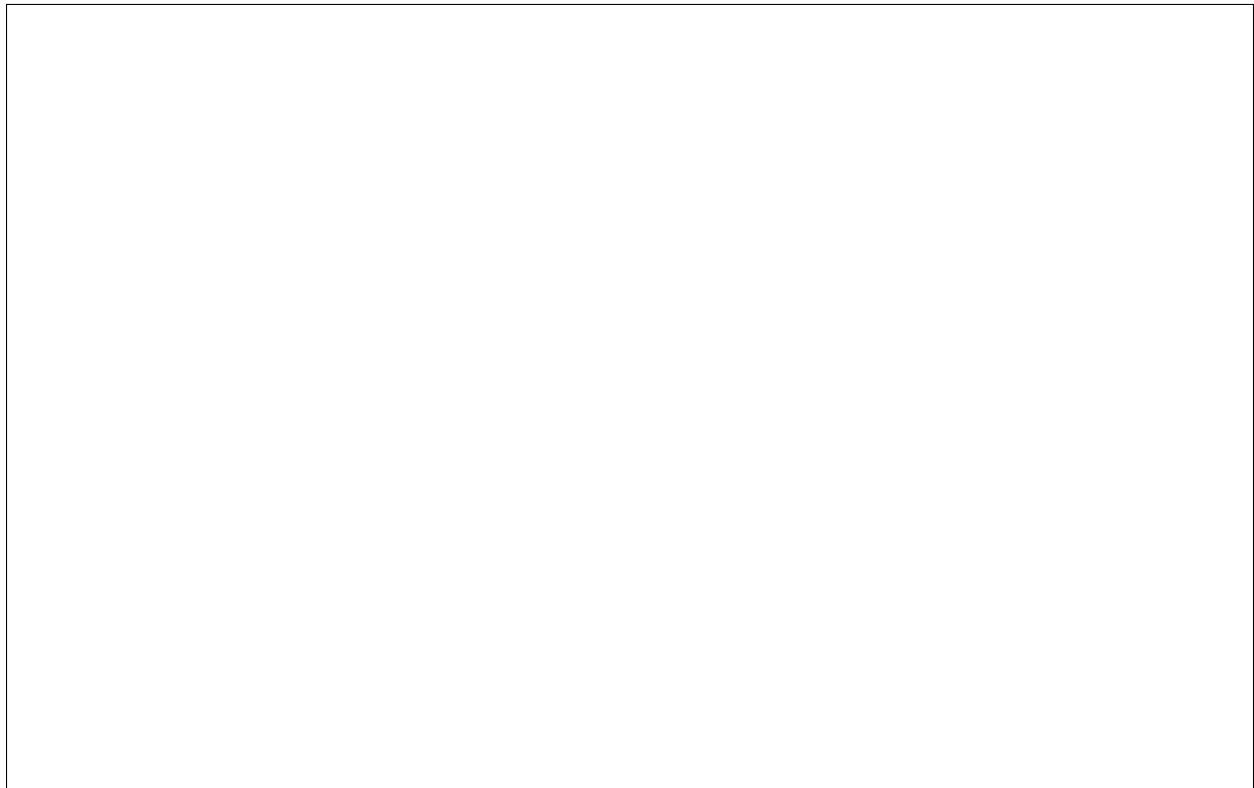
The MSE measures dispersion around the true value of the parameter. It is defined as:

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

However, the variance of  $\hat{\theta}$  measures the dispersion of the distribution of the distribution of  $\hat{\theta}$  around its mean or expected value.

$$\text{var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$$

The relationship between the  $\text{MSE}(\hat{\theta})$  and the  $\text{var}(\hat{\theta})$  is as follows:



## Lecture 10

An estimator  $\hat{\beta}_2$  is said to be a best linear unbiased estimator (BLUE) of  $\beta_2$  if the following hold:

♣ **It is linear.** It is the linear function of a random variable.

♣ **It is unbiased.** That is  $E(\hat{\beta}_2)$  is equal to the true value,  $\beta_2$

♣ **It has the minimum variance in the class of all such linear unbiased estimators.**

**Gauss-Markov Theorem:** Given the assumptions of the classical linear regression model, the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE.

### 3.5 A measure of goodness of fit: $r^2$

In this section, we are going to study the goodness of fit of the fitted regression line to a set of data. Let us consider the following example:

Suppose we were to estimate the family expenditure (Y) based on our information from a random sample (as in Table 5 on page 48).

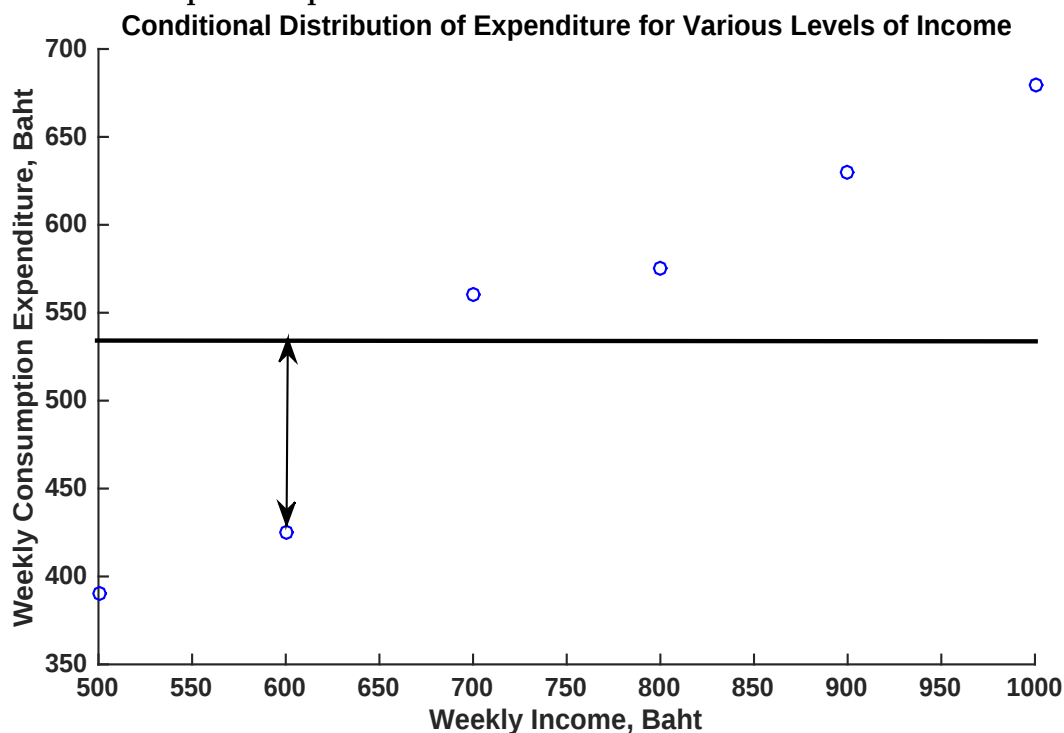
What will happen if we set the estimated Y to be  $\bar{Y}$ ?

**Table 7.** Estimating the expenditure of the household

| Family Number (i) | Actual<br>$Y_i$ | Estimate<br>$\hat{Y}_i = \bar{Y}$ | Error in Estimation<br>$Y_i - \bar{Y}$ | Errors Squared<br>$(Y_i - \bar{Y})^2$ |
|-------------------|-----------------|-----------------------------------|----------------------------------------|---------------------------------------|
| 1                 | 390             | 543                               | -153                                   | 23460.03                              |
| 2                 | 425             | 543                               | -118                                   | 13963.36                              |
| 3                 | 560             | 543                               | 17                                     | 283.36                                |
| 4                 | 575             | 543                               | 32                                     | 1013.36                               |
| 5                 | 630             | 543                               | 87                                     | 7540.03                               |
| 6                 | 679             | 543                               | 136                                    | 18450.69                              |
| Sum               | 3259            | 3259                              | 0                                      | 64710.83                              |

We can see all this graphically:

Figure 3.8: Graphic Representation



**Question:** Can we determine the total estimation error for this sample data?

**Answer:** Yes, we can calculate the total (combined) amount of estimation error for all observations in the sample when **using the mean as the estimate** as following:

$$TSS = \sum (Y_i - \bar{Y})^2$$

It is called the total sum of squares (TSS) which is the total variation of the actual Y values about their sample mean.

Since our objective in estimation is to minimize error (maximize precision), we need to cut down the amount of the estimation error (TSS).

We can achieve this by using information about other variables suspected to be strong predictors (strongly related to) the expenditure of the families.

We now can attempt to estimate the expenditure from the information on the income level of the family, rather than from its own mean.

**Table 8.** Estimating the expenditure of the household with income

| <b>Family (i)</b> | Actual | Income |               |               |                              |                   |
|-------------------|--------|--------|---------------|---------------|------------------------------|-------------------|
|                   | $Y_i$  | $X_i$  | $X - \bar{X}$ | $Y - \bar{Y}$ | $(X - \bar{X})(Y - \bar{Y})$ | $(X - \bar{X})^2$ |
| 1                 | 390    | 500    | -250          | -153.17       | 38291.67                     | 62500             |
| 2                 | 425    | 600    | -150          | -118.17       | 17725.00                     | 22500             |
| 3                 | 560    | 700    | -50           | 16.83         | -841.67                      | 2500              |
| 4                 | 575    | 800    | 50            | 31.83         | 1591.67                      | 2500              |
| 5                 | 630    | 900    | 150           | 86.83         | 13025.00                     | 22500             |
| 6                 | 679    | 1000   | 250           | 135.83        | 33958.33                     | 62500             |
| Sum               | 3259   | 4500   | 0             | 0             | 103750                       | 175000            |

From the table 8, we can calculate the simple regression as following:

Figure 3.9: Breakdown of the variation of  $Y_i$  into two components  
Conditional Distribution of Expenditure for Various Levels of Income

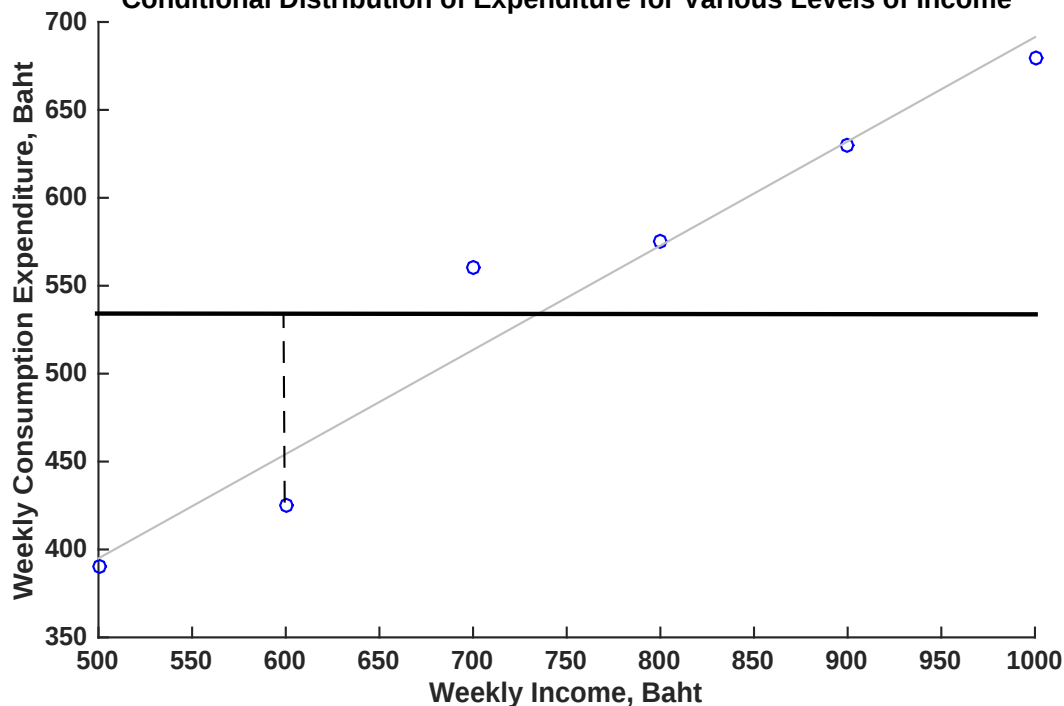


Table 9. Estimating the expenditure of the household with income

| Family (i) | Actual<br>$Y_i$ | Income<br>$X_i$ | Regression Estimate<br>$\hat{Y}$ | Residual<br>$Y - \hat{Y}$ | Residual squared<br>$(Y - \hat{Y})^2$ |
|------------|-----------------|-----------------|----------------------------------|---------------------------|---------------------------------------|
| 1          | 390             | 500             | 394.95                           | -4.95                     | 24.53                                 |
| 2          | 425             | 600             | 454.24                           | -29.24                    | 854.87                                |
| 3          | 560             | 700             | 513.52                           | 46.48                     | 2160.04                               |
| 4          | 575             | 800             | 572.81                           | 2.19                      | 4.80                                  |
| 5          | 630             | 900             | 632.10                           | -2.10                     | 4.39                                  |
| 6          | 679             | 1000            | 691.38                           | -12.38                    | 153.29                                |
| Sum        | 3259            | 4500            | 0                                | 0                         | 3201.90                               |

From the table 9, we can calculate the estimation error we have committed by using the regression line as:

$$RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum \hat{u}_i^2$$

where RSS stands for the residual sum of squares. which is the unexplained variation of the Y values about the regression line.

Total Baseline Error using the mean (SS Total) =

New or Remaining Error (SS Error or SS Residual) =

**QUESTION:** How much of the original estimation error have we explained away (eliminated) by using the regression model (instead of the mean)?

**ANS**

**QUESTION:** What % of estimation error have we explained (eliminated by using the regression model)?

**ANS**

**QUESTION:** What does the remaining% represent?

**ANS**

Percent of variation (differences) in expenditures that can be accounted for by: (a) all other potential predictors not included in the model, beyond income levels, and (b) unexplainable random/chance variations.

$$r^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

♣  $r^2$  is a measure of our success regarding accuracy of our estimation effort.

♣  $r^2$  = % of estimation error that we have been able to explain away by using the regression model, instead of using the mean.

♣  $r^2$  indicates how much better we can predict Y from information about Xs, rather than from using its own mean.

♣  $r^2$  = % of differences (variations) in Y values that is explained by (attributable to) differences in X values.

## Lecture 11

### CHAPTER 4: Classical Normal Regression Model (CNLRM)

We know that the classical theory of statistical inference consists of:

#### 1. Estimation

We have covered this topic since we were able to estimate the parameters  $\beta_1, \beta_2$ , and  $\sigma^2$  by using the method of OLS.

We also proved that these estimators  $\hat{\beta}_1$ ,  $\hat{\beta}_2$  and  $\hat{\sigma}$  satisfy several desirable statistical properties, such as unbiasedness, minimum variance, and linearity (BLUE property).

However,  $\hat{\beta}_1$ ,  $\hat{\beta}_2$  and  $\hat{\sigma}$  change their values from sample to sample. The following tables show the two different sets of  $\hat{\beta}_1$ ,  $\hat{\beta}_2$  and  $\hat{\sigma}$  depending on the two different sample data.

**Table 10.** Estimating the expenditure of the household with income

| Family (i) | Actual<br>$Y_i$ | Income<br>$X_i$ | Regression Estimate<br>$\hat{Y}$ | Residual<br>$Y - \hat{Y}$ | Residual squared<br>$(Y - \hat{Y})^2$ |
|------------|-----------------|-----------------|----------------------------------|---------------------------|---------------------------------------|
| 1          | 390             | 500             | 394.95                           | -4.95                     | 24.53                                 |
| 2          | 425             | 600             | 454.24                           | -29.24                    | 854.87                                |
| 3          | 560             | 700             | 513.52                           | 46.48                     | 2160.04                               |
| 4          | 575             | 800             | 572.81                           | 2.19                      | 4.80                                  |
| 5          | 630             | 900             | 632.10                           | -2.10                     | 4.39                                  |
| 6          | 679             | 1000            | 691.38                           | -12.38                    | 153.29                                |
| Sum        | 3259            | 4500            | 0                                | 0                         | 3201.90                               |

If we use this sample data. We can estimate:

$$\hat{\beta}_1 = 98.524$$

$$\hat{\beta}_2 = 0.593$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n - 2} = \frac{3201.90}{6 - 2} = 800.476$$

**Table 11.** Estimating the expenditure of the household with income with another sample data

| Family (i) | Actual<br>$Y_i$ | Income<br>$X_i$ | Regression Estimate<br>$\hat{Y}$ | Residual<br>$Y - \hat{Y}$ | Residual squared<br>$(Y - \hat{Y})^2$ |
|------------|-----------------|-----------------|----------------------------------|---------------------------|---------------------------------------|
| 1          | 390             | 500             | 325.71                           | 64.29                     | 4132.65                               |
| 2          | 425             | 600             | 406.43                           | 18.57                     | 344.90                                |
| 3          | 560             | 700             | 487.14                           | 72.86                     | 5308.16                               |
| 4          | 575             | 800             | 567.86                           | 7.14                      | 51.02                                 |
| 5          | 630             | 900             | 648.57                           | -18.57                    | 344.90                                |
| 6          | 679             | 1000            | 729.29                           | -50.29                    | 2528.65                               |
| Sum        | 3259            | 4500            | 0                                | 0                         | 12710.29                              |

If we use this sample data. We can estimate:

$$\hat{\beta}_1 = -77.857$$

$$\hat{\beta}_2 = 0.807$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n - 2} = \frac{12710.29}{6 - 2} = 3177.571$$

From the example, you can easily see that these estimators are **RANDOM VARIABLES**. Therefore, we need to learn another part of statistical inference which is call **Hypothesis Testing**.

## 2. Hypothesis Testing

The main objective is to find out how close of  $\hat{\beta}_1$  and  $\hat{\beta}_2$  to the true  $\beta_1$  and the true  $\beta_2$ , respectively. Also, we would like to see how close of  $\hat{\sigma}^2$  compared to the true  $\sigma^2$ .

To achieve this goal, we need to know the probability distributions of  $\hat{\beta}_1$ ,  $\hat{\beta}_2$ , and  $\hat{\sigma}^2$ .

Consider the estimator of  $\beta_2$ :

$$\hat{\beta}_2 = \sum k_i Y_i$$

We can write the above equation as:

$$\hat{\beta}_2 = \sum k_i (\beta_1 + \beta_2 X_i + u_i)$$

From this equation, the probability distribution of  $\hat{\beta}_2$  will depend on the assumption made about the probability distribution of  $u_i$

#### 4.1 The Normality Assumption for $u_i$

In the classical normal linear regression model (CNLRM), we assume that each  $u_i$  is distributed normally :

$$u_i \sim N(0, \sigma^2)$$

where

Mean:

$$E(u_i) = 0$$

Variance:

$$\begin{aligned} E[u_i - E(u_i)]^2 &= E(u_i^2) = \sigma^2 \\ cov(u_i, u_j) &= E\{[u_i - E(u_i)][u_j - E(u_j)]\} = E(u_i u_j) = 0 \end{aligned}$$

Therefore,

$$u_i \sim N(0, \sigma^2)$$

Also,  $u_i$  and  $u_j$  are not only uncorrelated but also independently distributed.

we can then write the above equation as:

$$u_i \sim NID(0, \sigma^2)$$

where NID stands for normally and independently distributed.

## 4.2 Properties of OLS estimators under the normality assumption

1. They are unbiased.
2. They have minimum variance.
3. By 1+2 properties, they are minimum-variance unbiased, or efficient estimators.
4.  $\hat{\beta}_1$  is normally distributed with:

$$\text{Mean: } E(\hat{\beta}_1) = \beta_1$$

$$\text{var}(\hat{\beta}_1) = \sigma_{\beta_1}^2 = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2$$

Therefore,

$$\hat{\beta}_1 \sim N(\beta_1, \sigma_{\beta_1}^2)$$

By the properties of the normal distribution, we can:

5.  $\hat{\beta}_2$  is normally distributed with

$$\text{Mean: } E(\hat{\beta}_2) = \beta_2$$

$$\text{var}(\hat{\beta}_2) = \sigma_{\beta_2}^2 = \frac{\sigma^2}{\sum x_i^2}$$

or more compactly

$$\hat{\beta}_2 \sim N(\beta_2, \sigma_{\beta_2}^2)$$

then we can define the standard normal distribution as

6.  $(n-2)(\hat{\sigma}^2/\sigma^2)$  is distributed as the  $\chi^2$  (chi-square) distribution with  $(n-2)$  df.
7.  $(\hat{\beta}_1, \hat{\beta}_2)$  are distributed independently of  $\hat{\sigma}^2$
8.  $\hat{\beta}_1$  and  $\hat{\beta}_2$  have the minimum variance in the entire class of unbiased estimators, whether linear or not.
9. we can find out the probability distribution of  $Y_i$  as following:

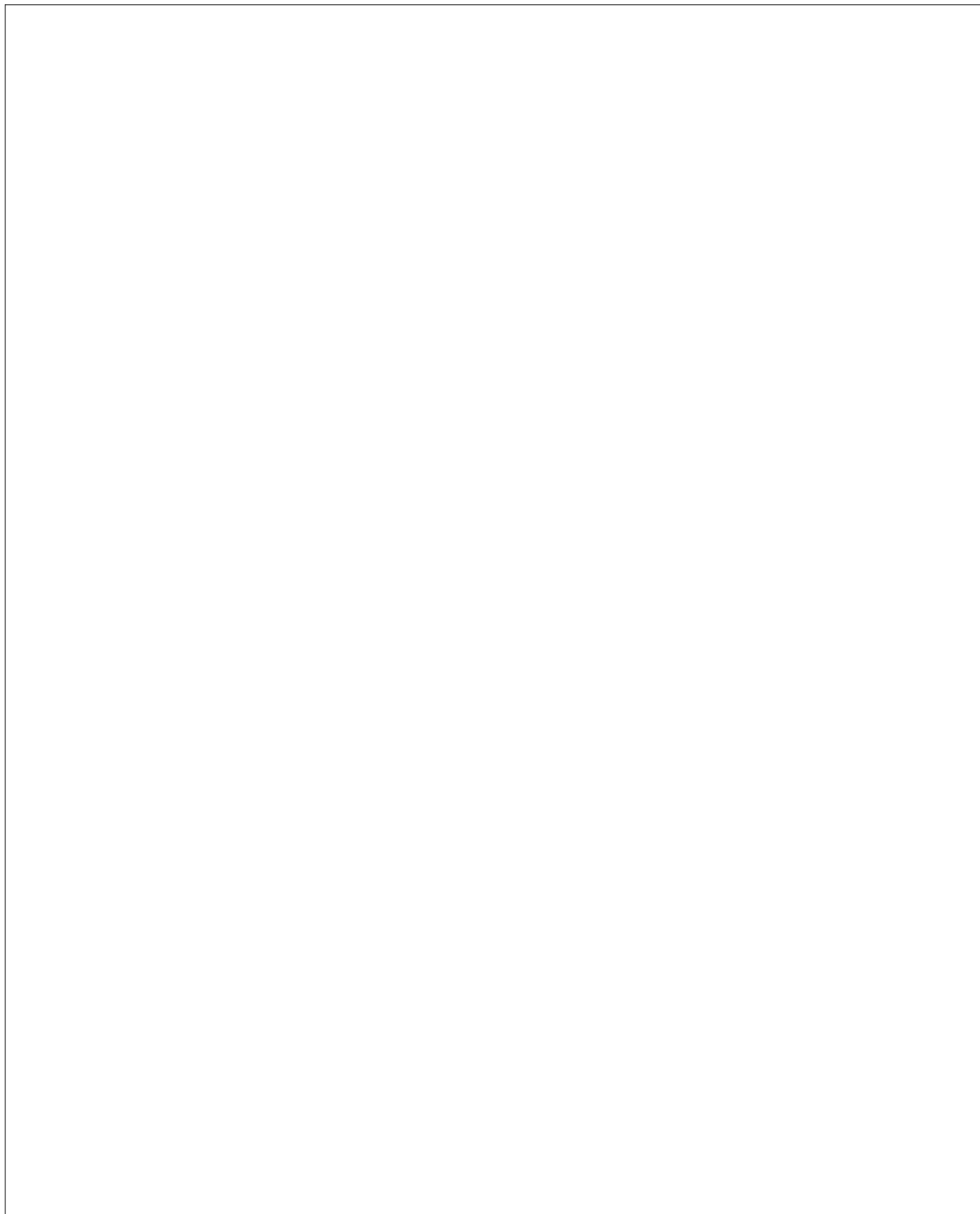
## Lecture 12

# CHAPTER 5: Interval Estimation and Hypothesis Testing

### Interval Estimation



## 5.1 Confidence Intervals for Regression Coefficients $\beta_1$ and $\beta_2$



### In Sum

A  $100(1 - \alpha)$  percent **confidence interval** for  $\beta_2$  can be defined as:

$$\hat{\beta}_2 \pm t_{\alpha/2} \text{se}(\hat{\beta}_2)$$

or

$$\Pr[\hat{\beta}_2 - t_{\alpha/2} \text{se}(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} \text{se}(\hat{\beta}_2)] = 1 - \alpha$$

Analogously, we can define  $100(1 - \alpha)$  percent **confidence interval** for  $\beta_1$  as:

$$\hat{\beta}_1 \pm t_{\alpha/2} \text{se}(\hat{\beta}_1)$$

or

$$\Pr[\hat{\beta}_1 - t_{\alpha/2} \text{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2} \text{se}(\hat{\beta}_1)] = 1 - \alpha$$

## Example

**Table 12.** Estimating the expenditure of the household with income

| Family (i) | Actual<br>$Y_i$ | Income<br>$X_i$ | $X - \bar{X}$ | $Y - \bar{Y}$ | $(X - \bar{X})(Y - \bar{Y})$ | $(X - \bar{X})^2$ |
|------------|-----------------|-----------------|---------------|---------------|------------------------------|-------------------|
| 1          | 390             | 500             | -250          | -153.17       | 38291.67                     | 62500             |
| 2          | 425             | 600             | -150          | -118.17       | 17725.00                     | 22500             |
| 3          | 560             | 700             | -50           | 16.83         | -841.67                      | 2500              |
| 4          | 575             | 800             | 50            | 31.83         | 1591.67                      | 2500              |
| 5          | 630             | 900             | 150           | 86.83         | 13025.00                     | 22500             |
| 6          | 679             | 1000            | 250           | 135.83        | 33958.33                     | 62500             |
| Sum        | 3259            | 4500            | 0             | 0             | 103750                       | 175000            |

**Table 13.** Estimating the expenditure of the household with income

| Family (i) | Actual<br>$Y_i$ | Income<br>$X_i$ | Regression Estimate<br>$\hat{Y}$ | Residual<br>$Y - \hat{Y}$ | Residual squared<br>$(Y - \hat{Y})^2$ |
|------------|-----------------|-----------------|----------------------------------|---------------------------|---------------------------------------|
| 1          | 390             | 500             | 394.95                           | -4.95                     | 24.53                                 |
| 2          | 425             | 600             | 454.24                           | -29.24                    | 854.87                                |
| 3          | 560             | 700             | 513.52                           | 46.48                     | 2160.04                               |
| 4          | 575             | 800             | 572.81                           | 2.19                      | 4.80                                  |
| 5          | 630             | 900             | 632.10                           | -2.10                     | 4.39                                  |
| 6          | 679             | 1000            | 691.38                           | -12.38                    | 153.29                                |
| Sum        | 3259            | 4500            | 0                                | 0                         | 3201.90                               |

**Confidence Interval for  $\beta_2$**

**Confidence Interval for  $\beta_1$**

## 5.2 Confidence Interval for $\sigma^2$

