Probability & Statistics

COURSE CODE: MT2005

Chapter 1 Introduction to Statistics

Statistics

Statistics is the art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.

Data Collection & Descriptive Statistics

Sometimes a statistical analysis begins with a given set of data: For instance, the government regularly collects and publicizes data concerning earthquake occurrences, the unemployment rate and the rate of inflation. Statistics can be used to describe, summarize, and analyze these data.

In some situations, data are not yet available; in such cases statistical theory can be used to design an appropriate experiment to generate data. For instance, suppose that an instructor is interested in determining which of two different methods for teaching computer programming to beginners is most effective. To study this question, the instructor might divide the students into two groups, and use a

different teaching method for each group. At the end of the class the students can be tested and the scores of the members of the different groups compared. If the data, consisting of the test scores of members of each group, are significantly higher in one of the groups, then it might seem reasonable to suppose that the teaching method used for that group is superior.

It is important to note, however, that in order to be able to draw a valid conclusion from the data, it is essential that the students were divided into groups in such a manner that neither group was more likely to have the students with greater natural aptitude for programming. For instance, the instructor should not have let the male class members be one group and the females the other. For if so, then even if the women scored significantly higher than the men, it would not be clear whether this was due to the them,

or to the fact that women may be inherently better than men at learning programming skills. The accepted way of avoiding this pitfall is to divide the class members into the two groups "at random." This term means that the division is done in such a manner that all possible choices of the members of a group are equally likely.

At the end of the experiment, the data should be described. For instance, the scores of the two groups should be presented. In addition, summary measures such as the average score of members of each of the groups should be presented. This part of statistics, concerned with the description and summarization of data, is called descriptive statistics.

Inferential Statistics & Probability Models

After the preceding experiment is completed and the data are described and summarized, we hope to be able to draw a conclusion about which teaching method is superior. This part of statistics, concerned with the drawing of conclusions, is called inferential statistics.

To be able to draw a conclusion from the data, we must take into account the possibility of chance. For instance, suppose that the average score of members of the first group is quite a bit higher than that of the second. Can we conclude that this increase is due to the teaching method used? Or is it possible that the teaching method was not responsible for the increased scores but rather that the higher scores of the first group were just a chance occurrence? For instance, the fact that a coin comes up heads 7 times in 10 flips does not necessarily mean that the coin is more likely to come up heads than tails in future flips.

Indeed, it could be a perfectly ordinary coin that, by chance, just happened to land heads 7 times out of the total of 10 flips. (On the other hand, if the coin had landed heads 47 times out of 50 flips, then we would be quite certain that it was not an ordinary coin.)

To be able to draw logical conclusions from data, we usually make some assumptions about the chances (or probabilities) of obtaining the different data values. The totality of these assumptions is referred to as a probability model for the data.

Population & Samples

In statistics, we are interested in obtaining information about a total collection of elements, which we will refer to as the population. The population is often too large for us to examine each of its members. For instance, we might have all the residents of a given state, or all the television sets produced in the last year by a particular manufacturer, or all the households in a given community. In such cases, we try to learn about the population by choosing and then examining a subgroup of its elements. This subgroup of a population is called a sample.

The sample is to be informative because it is representative of the population.

Describing Data Set

The numerical findings of a study should be presented clearly, concisely, and in such a manner that an observer can quickly obtain a feel for the essential characteristics of the data. Over the years it has been found that tables and graphs are particularly useful ways of presenting data, often revealing important features such as the range, the degree of concentration, and the symmetry of the data.

Frequency Tables & Graphs

A data set having a relatively small number of distinct values can be conveniently presented in a frequency table. For instance, Table 1 is a frequency table for a data set consisting of the starting yearly salaries (to the nearest thousand dollars) of 42 recently graduated students with B.S. degrees in computer science. Table 1 tells us, among other things, that the lowest starting salary of $47,000 was received by four of the graduates, whereas the highest salary of $60,000 was received by a single student. The most common starting salary was $52,000, and was received by 10 of the students.

Table 1 *Starting Yearly Salaries*

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

Data from a frequency table can be graphically represented by a line graph that plots the distinct data values on the horizontal axis and indicates their frequencies by the heights of vertical lines. A line graph of the data presented in Table 1 is shown in Figure 1.

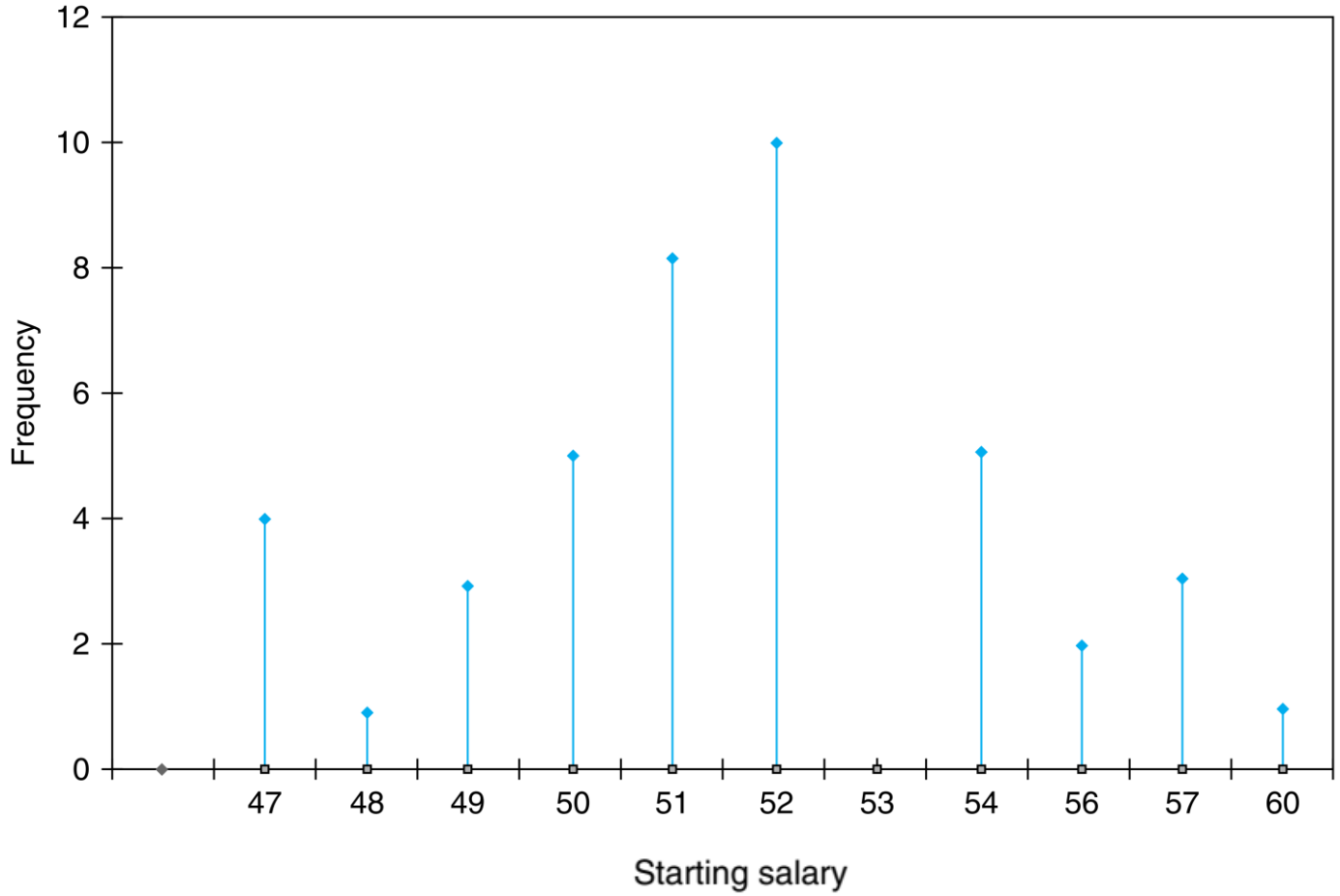


Figure 1

When the lines in a line graph are given added thickness, the graph is called a bar graph.

Figure 2 presents a bar graph.

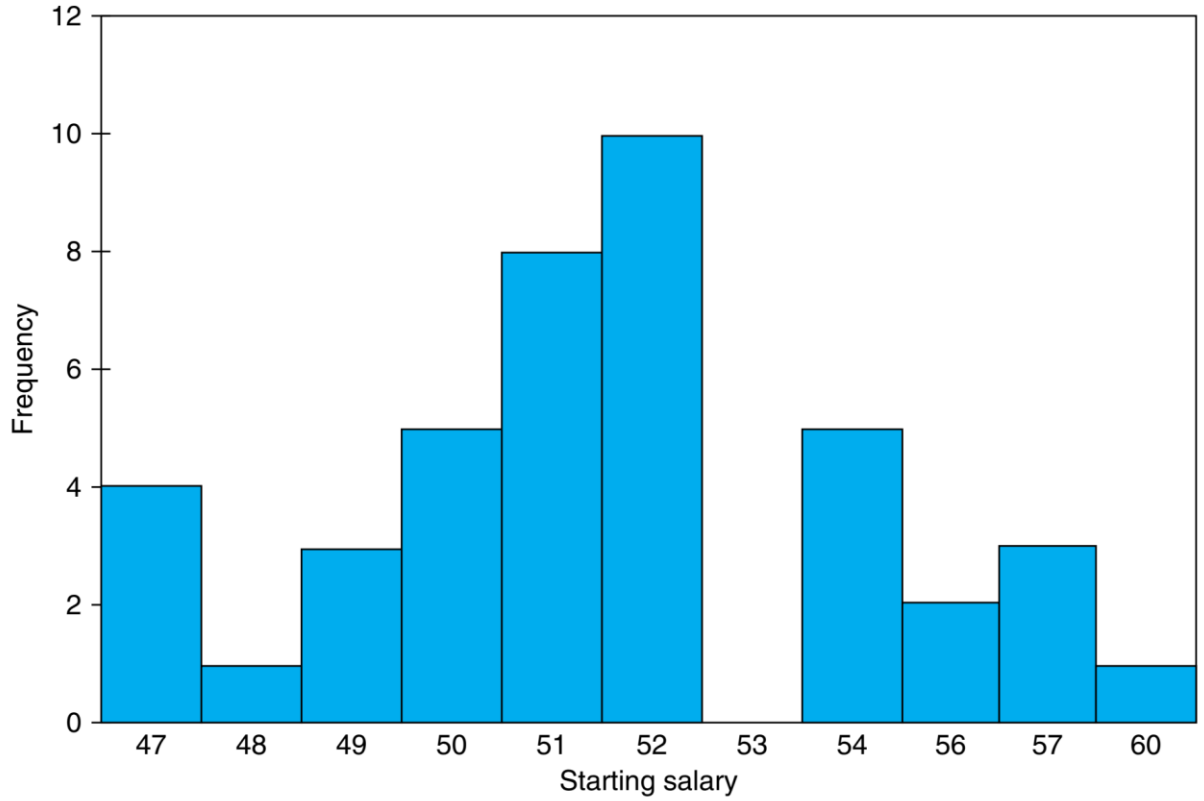


Figure 2 *Bar graph for starting salary data.*

Another type of graph used to represent a frequency table is the frequency polygon, which plots the frequencies of the different data values on the vertical axis, and then connects the plotted points with straight lines. Figure 3 presents a frequency polygon for the data of Table 1.

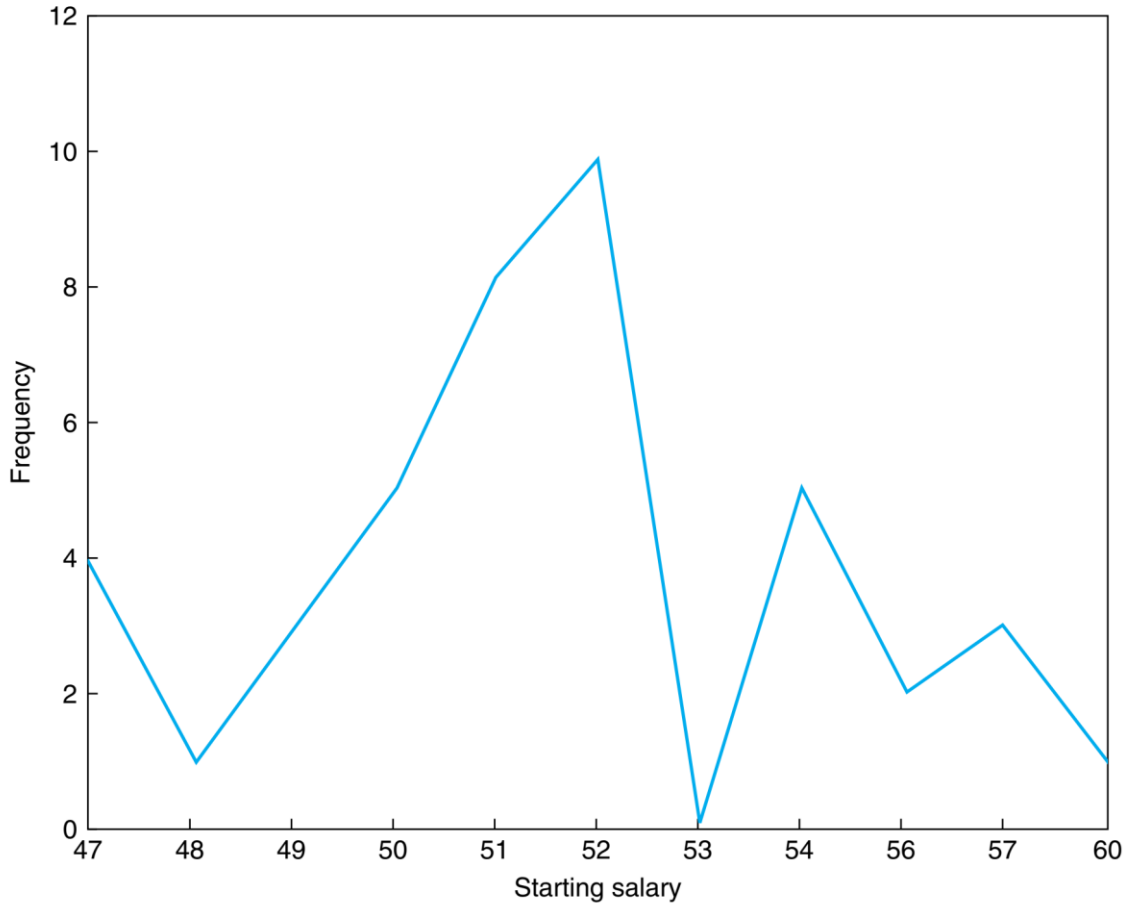


Figure 3

Frequency polygon for starting salary data.

Relative Frequency Tables & Graphs

Consider a data set consisting of n values. If f is the frequency of a particular value, then the ratio f/n is called its relative frequency. That is, the relative frequency of a data value is the proportion of the data that have that value. The relative frequencies can be represented graphically by a relative frequency line or bar graph or by a relative frequency polygon. Indeed, these relative frequency graphs will look like the corresponding graphs of the absolute frequencies except that the labels on the vertical axis are now the old labels (that gave the frequencies) divided by the total number of data points.

Table 4 is a relative frequency table for the data of Table 1. The relative frequencies are obtained by dividing the corresponding frequencies of Table 1 by 42, the size of the data set.

Starting Salary	Frequency
47	4/42 = .0952
48	1/42 = .0238
49	3/42
50	5/42
51	8/42
52	10/42
53	0
54	5/42
56	2/42
57	3/42
60	1/42

Figure 4

Pie Chart

We can construct pie chart by dividing a circle into various sections or slices. It should be used when we want to compare individual categories with the whole. If you want to compare the values of categories with each other, a bar chart may be more useful.

Problem

The following table shows the yearly budget of a family

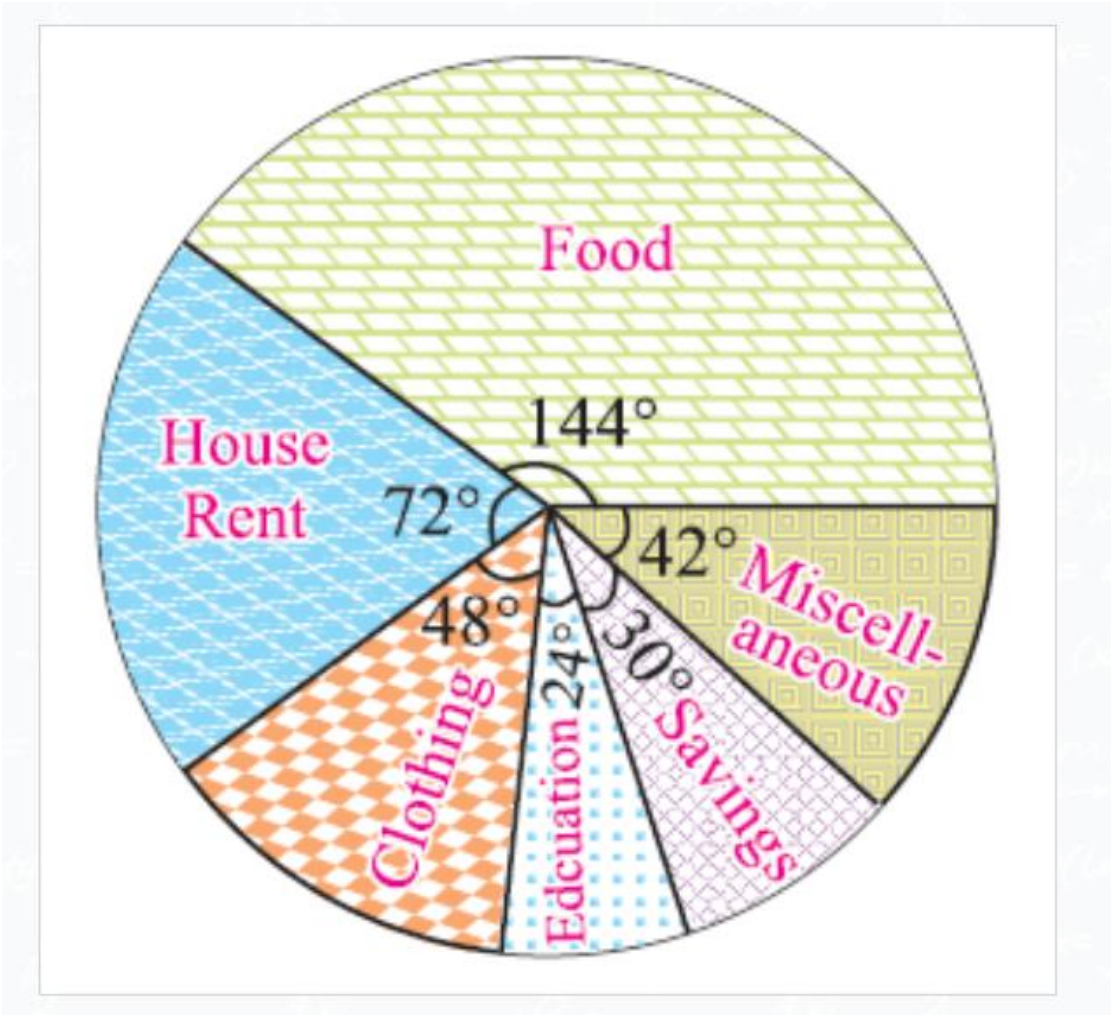
Particulars	Food	House Rent	Clothing	Education	Savings	Miscella-neous
Expenses (in $)	4800	2400	1600	800	1000	1400

Draw a pie chart to represent the above information.

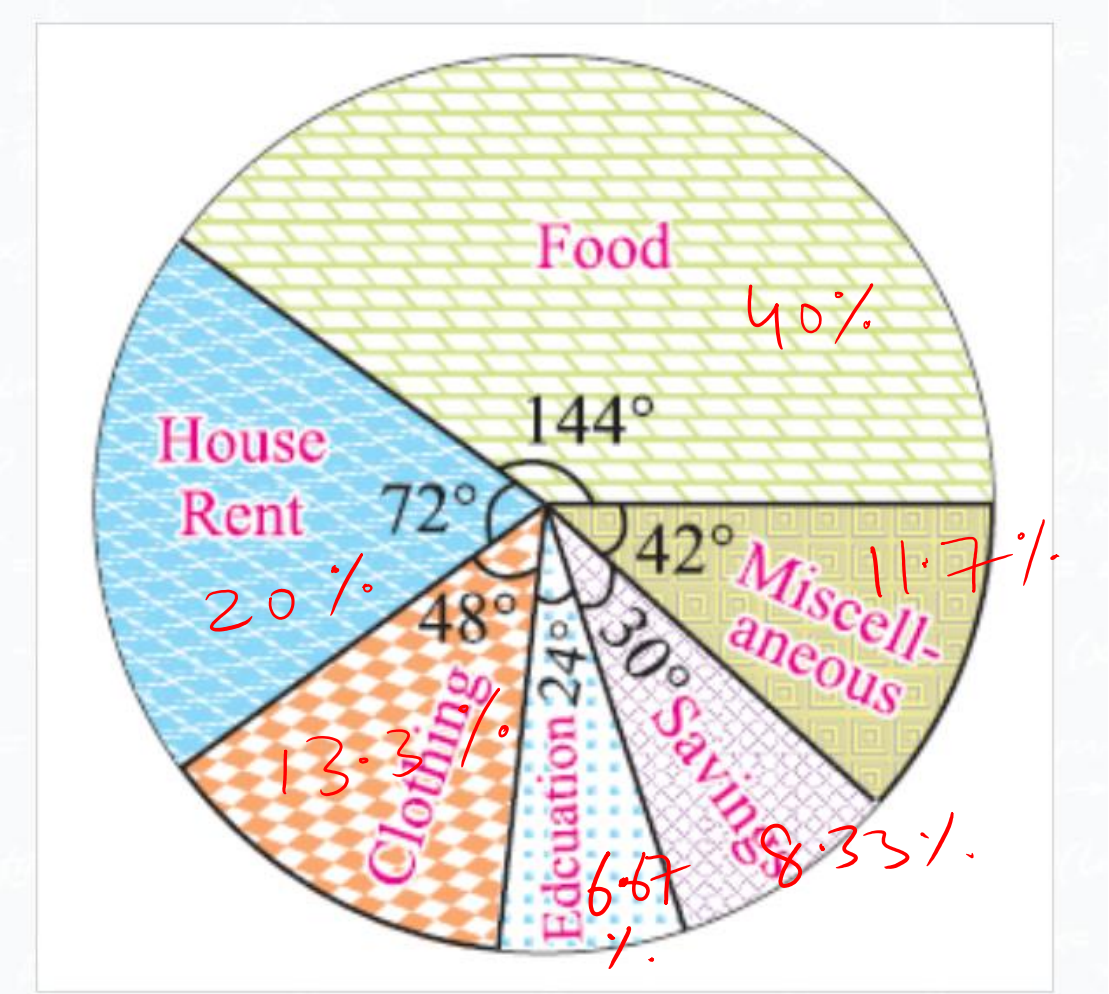
Solution

Particulars	Expenses ($)	Central angle
Food	4800	$\frac{4800}{12000} \times 360° = 144°$
House rent	2400	$\frac{2400}{12000} \times 360° = 72°$
Clothing	1600	$\frac{1600}{12000} \times 360° = 48°$
Education	800	$\frac{800}{12000} \times 360° = 24°$
Savings	1000	$\frac{1000}{12000} \times 360° = 30°$
Miscellaneous	1400	$\frac{1400}{12000} \times 360° = 42°$
Total	**12000**	**360°**

From the table, we obtain the required pie chart as shown below.



Particulars	Expenses ($)	%
Food	4800	$\frac{4800}{12000} \times 100 = 40$
House rent	2400	$\frac{2400}{12000} \times 100 = 20$
Clothing	1600	$\frac{1600}{12000} \times 100 = 13.3$
Education	800	6.67
Savings	1000	8.33
Miscellaneous	1400	11.7
Total	**12000**	100


Food
40%
144°
House Rent
72°
20%
42°
Miscell-aneous
11.7%
48°
Clothing
13.3%
24°
Education
6.67%
30°
Savings
8.33%