3 Describing data by tables and graphs

[Johnson & Bhattacharyya (1992), Weiss (1999) and Freund (2001)]

3.1 Qualitative variable

The number of observations that fall into particular class (or category) of the qualitative variable is called the **frequency** (or **count**) of that class. A table listing all classes and their frequencies is called a **frequency distribution**.

In addition of the frequencies, we are often interested in the **percentage** of a class. We find the percentage by dividing the frequency of the class by the total number of observations and multiplying the result by 100. The percentage of the class, expressed as a decimal, is usually referred to as the **relative frequency** of the class.

$$\text{Relative frequency of the class} = \frac{\text{Frequency in the class}}{\text{Total number of observation}}$$

A table listing all classes and their relative frequencies is called a **relative frequency distribution**. The relative frequencies provide the most relevant information as to the pattern of the data. One should also state the sample size, which serves as an indicator of the creditability of the relative frequencies. Relative frequencies sum to 1 (100%).

A **cumulative frequency** (**cumulative relative frequency**) is obtained by summing the frequencies (relative frequencies) of all classes up to the specific class. In a case of qualitative variables, cumulative frequencies makes sense only for ordinal variables, not for nominal variables.

The qualitative data are presented graphically either as a **pie chart** or as a horizontal or vertical **bar graph**.

A pie chart is a disk divided into pie-shaped pieces proportional to the relative frequencies of the classes. To obtain angle for any class, we multiply the relative frequencies by 360 degrees, which corresponds to the complete circle.

A horizontal bar graph displays the classes on the horizontal axis and the frequencies (or relative frequencies) of the classes on the vertical axis. The frequency (or relative frequency) of each class is represented by vertical bar

whose height is equal to the frequency (or relative frequency) of the class. In a bar graph, its bars do *not* touch each other. At vertical bar graph, the classes are displayed on the vertical axis and the frequencies of the classes on the horizontal axis.

Nominal data is best displayed by pie chart and ordinal data by horizontal or vertical bar graph.

EXAMPLE 3.1. Let the blood types of 40 persons are as follows:

O O A B A O A A A O B O B O O A O O A A A A AB A B A A O O A O O A A A O A O O AB

Summarizing data in a frequency table by using SPSS:

Analyze -> Descriptive Statistics -> Frequencies,
Analyze -> Custom Tables -> Tables of Frequencies

Table 1: Frequency distribution of blood types

BLOOD

		Statistics	
BLOOD		Frequency	Percent
Valid	O	16	40.0
	A	18	45.0
	B	4	10.0
	AB	2	5.0
	Total	40	100.0

Graphical presentation of data in SPSS:

Graphs -> Interactive -> Pie -> Simple,
Graphs -> Interactive -> Bar

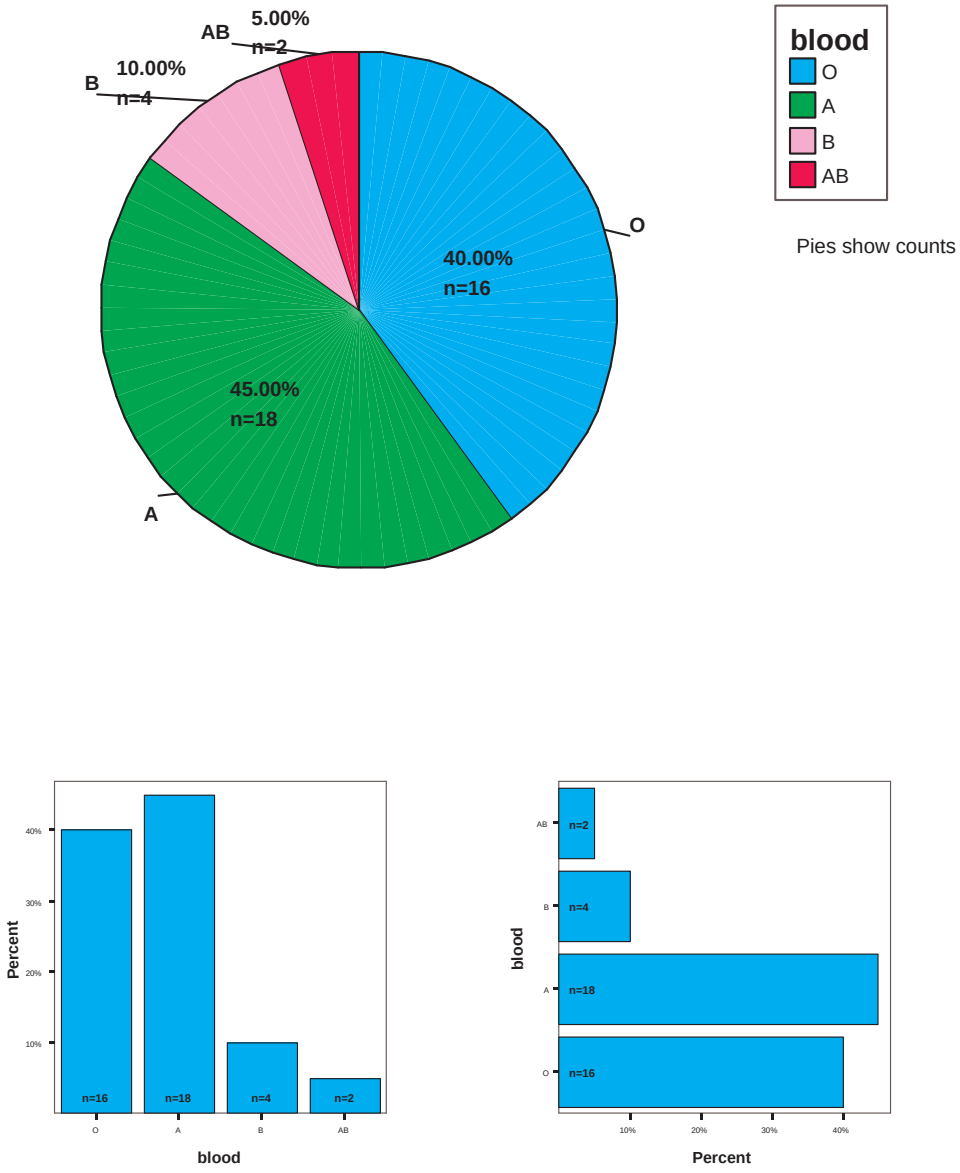

Figure 2: Charts for blood types

3.2 Quantitative variable

The data of the quantitative variable can also presented by a frequency distribution. If the discrete variable can obtain only few different values, then the data of the discrete variable can be summarized in a same way as qualitative variables in a frequency table. In a place of the qualitative categories, we now list in a frequency table the distinct numerical measurements that appear in the discrete data set and then count their frequencies.

If the discrete variable can have a lot of different values or the quantitative variable is the continuous variable, then the data must be **grouped** into classes (categories) before the table of frequencies can be formed. The main steps in a process of grouping quantitative variable into classes are:

(a) Find the minimum and the maximum values variable have in the data set

(b) Choose intervals of equal length that cover the range between the minimum and the maximum *without* overlapping. These are called **class intervals**, and their end points are called **class limits**.

(c) Count the number of observations in the data that belongs to each class interval. The count in each class is the class frequency.

(c) Calculate the relative frequencies of each class by dividing the class frequency by the total number of observations in the data.

The number in the middle of the class is called **class mark** of the class. The number in the middle of the upper class limit of one class and the lower class limit of the other class is called the **real class limit**. As a rule of thumb, it is generally satisfactory to group observed values of numerical variable in a data into 5 to 15 class intervals. A smaller number of intervals is used if number of observations is relatively small; if the number of observations is large, the number on intervals may be greater than 15.

The quantitative data are usually presented graphically either as a **histogram** or as a horizontal or vertical bar graph. The histogram is like a horizontal bar graph except that its bars *do* touch each other. The histogram is formed from grouped data, displaying either frequencies or relative frequencies (percentages) of each class interval.

If quantitative data is discrete with only few possible values, then the variable should graphically be presented by a bar graph. Also if some reason it is more reasonable to obtain frequency table for quantitative variable with unequal class intervals, then variable should graphically also be presented by a bar graph!

EXAMPLE 3.2. Age (in years) of 102 people:

34,67,40,72,37,33,42,62,49,32,52,40,31,19,68,55,57,54,37,32,
54,38,20,50,56,48,35,52,29,56,68,65,45,44,54,39,29,56,43,42,
22,30,26,20,48,29,34,27,40,28,45,21,42,38,29,26,62,35,28,24,
44,46,39,29,27,40,22,38,42,39,26,48,39,25,34,56,31,60,32,24,
51,69,28,27,38,56,36,25,46,50,36,58,39,57,55,42,49,38,49,36,
48,44

Summarizing data in a frequency table by using SPSS:

Analyze -> Descriptive Statistics -> Frequencies,
Analyze -> Custom Tables -> Tables of Frequencies

Table 2: Frequency distribution of people's age

Frequency distribution of people's age

		Frequency	Percent	Cumulative Percent
Valid	18 - 22	6	5.9	5.9
	23 - 27	10	9.8	15.7
	28 - 32	14	13.7	29.4
	33 - 37	11	10.8	40.2
	38 - 42	19	18.6	58.8
	43 - 47	8	7.8	66.7
	48 - 52	12	11.8	78.4
	53 - 57	12	11.8	90.2
	58 - 62	4	3.9	94.1
	63 - 67	2	2.0	96.1
	68 - 72	4	3.9	100.0
	Total	102	100.0	

Graphical presentation of data in SPSS:

Graphs -> Interactive -> Histogram,
Graphs -> Histogram

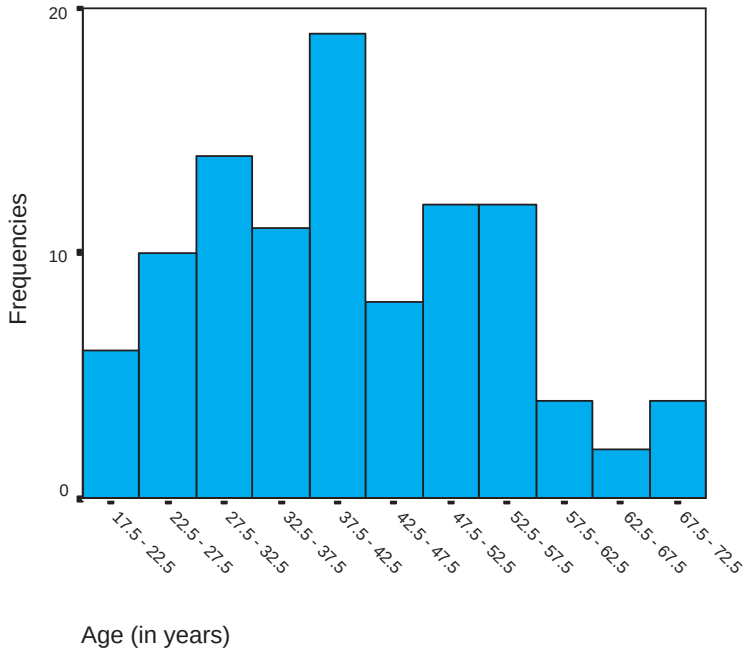

Figure 3: Histogram for people's age

EXAMPLE 3.3. Prices of hotdogs ($/oz.):

0.11,0.17,0.11,0.15,0.10,0.11,0.21,0.20,0.14,0.14,0.23,0.25,0.07,
0.09,0.10,0.10,0.19,0.11,0.19,0.17,0.12,0.12,0.12,0.10,0.11,0.13,
0.10,0.09,0.11,0.15,0.13,0.10,0.18,0.09,0.07,0.08,0.06,0.08,0.05,
0.07,0.08,0.08,0.07,0.09,0.06,0.07,0.08,0.07,0.07,0.07,0.08,0.06,
0.07,0.06

Frequency table:

Table 3: Frequency distribution of prices of hotdogs

Frequencies of prices of hotdogs ($/oz.)

		Frequency	Percent	Cumulative Percent
Valid	0.031-0.06	5	9.3	9.3
	0.061-0.09	19	35.2	44.4
	0.091-0.12	15	27.8	72.2
	0.121-0.15	6	11.1	83.3
	0.151-0.18	3	5.6	88.9
	0.181-0.21	4	7.4	96.3
	0.211-0.24	1	1.9	98.1
	0.241-0.27	1	1.9	100.0
	Total	54	100.0	

or alternatively

Table 4: Frequency distribution of prices of hotdogs (Left Endpoints Excluded, but Right Endpoints Included)

Frequencies of prices of hotdogs ($/oz.)

		Frequency	Percent	Cumulative Percent
Valid	0.03-0.06	5	9.3	9.3
	0.06-0.09	19	35.2	44.4
	0.09-0.12	15	27.8	72.2
	0.12-0.15	6	11.1	83.3
	0.15-0.18	3	5.6	88.9
	0.18-0.21	4	7.4	96.3
	0.21-0.24	1	1.9	98.1
	0.24-0.27	1	1.9	100.0
	Total	54	100.0	

Graphical presentation of the data:

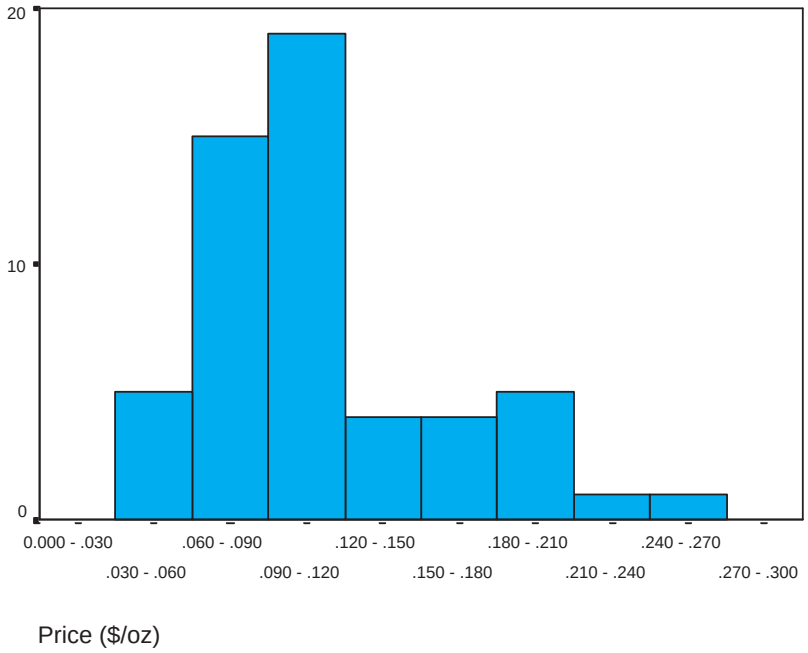

Figure 4: Histogram for prices

Let us look at another way of summarizing hotdogs' prices in a frequency table. First we notice that minimum price of hotdogs is 0.05. Then we make decision of putting the observed values 0.05 and 0.06 to the same class interval and the observed values 0.07 and 0.08 to the same class interval and so on. Then the class limits are choosen in way that they are middle values of 0.06 and 0.07 and so on. The following frequency table is then formed:

Table 5: Frequency distribution of prices of hotdogs

Frequencies of prices of hotdogs ($/oz.)

		Frequency	Percent	Cumulative Percent
Valid	0.045-0.065	5	9.3	9.3
	0.065-0.085	15	27.8	37.0
	0.085-0.105	10	18.5	55.6
	0.105-0.125	9	16.7	72.2
	0.125-0.145	4	7.4	79.6
	0.145-0.165	2	3.7	83.3
	0.165-0.185	3	5.6	88.9
	0.185-0.205	3	5.6	94.4
	0.205-0.225	1	1.9	96.3
	0.225-0.245	1	1.9	98.1
	0.245-0.265	1	1.9	100.0
	Total	54	100.0	

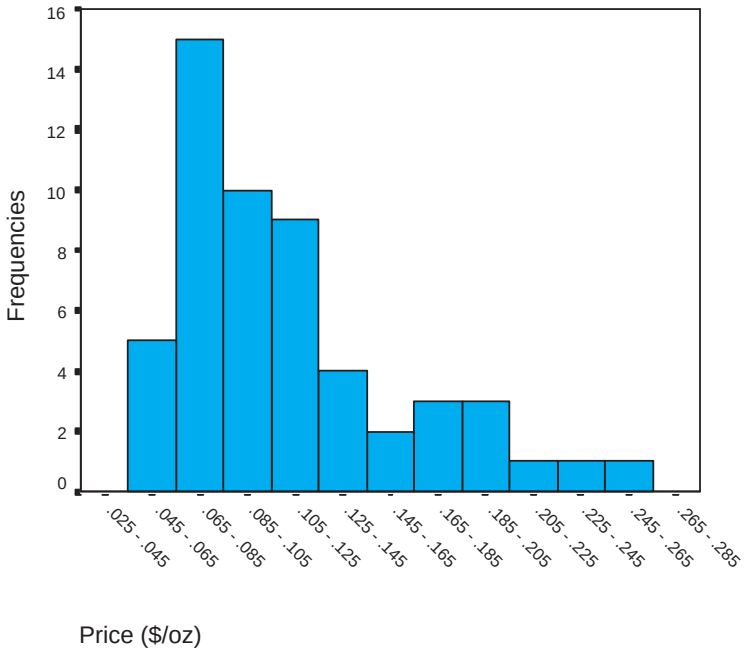

Figure 5: Histogram for prices

Another types of graphical displays for quantitative data are

(a) **dotplot**
Graphs -> Interactive -> Dot

(b) **stem-and-leaf diagram** of just **stemplot**
Analyze -> Descriptive Statistics -> Explore

(c) **frequency** and **relative-frequency polygon** for frequencies and for relative frequencies (**Graphs -> Interactive -> Line**)

(d) **ogives** for cumulative frequencies and for cumulative relative frequencies (**Graphs -> Interactive -> Line**)

3.3 Sample and Population Distributions

Frequency distributions for a variable apply both to a population and to samples from that population. The first type is called the **population distribution** of the variable, and the second type is called a **sample distribution**. In a sense, the sample distribution is a blurry photograph of the population distribution. As the sample size increases, the sample relative frequency in any class interval gets closer to the true population relative frequency. Thus, the photograph gets clearer, and the sample distribution looks more like the population distribution.

When a variable is continous, one can choose class intervals in the frequency distribution and for the histogram as narrow as desired. Now, as the sample size increases indefinitely and the number of class intervals simultaneously increases, with their width narrowing, the shape of the sample histogram gradually approaches a smooth curve. We use such curves to represent population distributions. Figure 6. shows two samples histograms, one based on a sample of size 100 and the second based on a sample of size 2000, and also a smooth curve representing the population distribution.

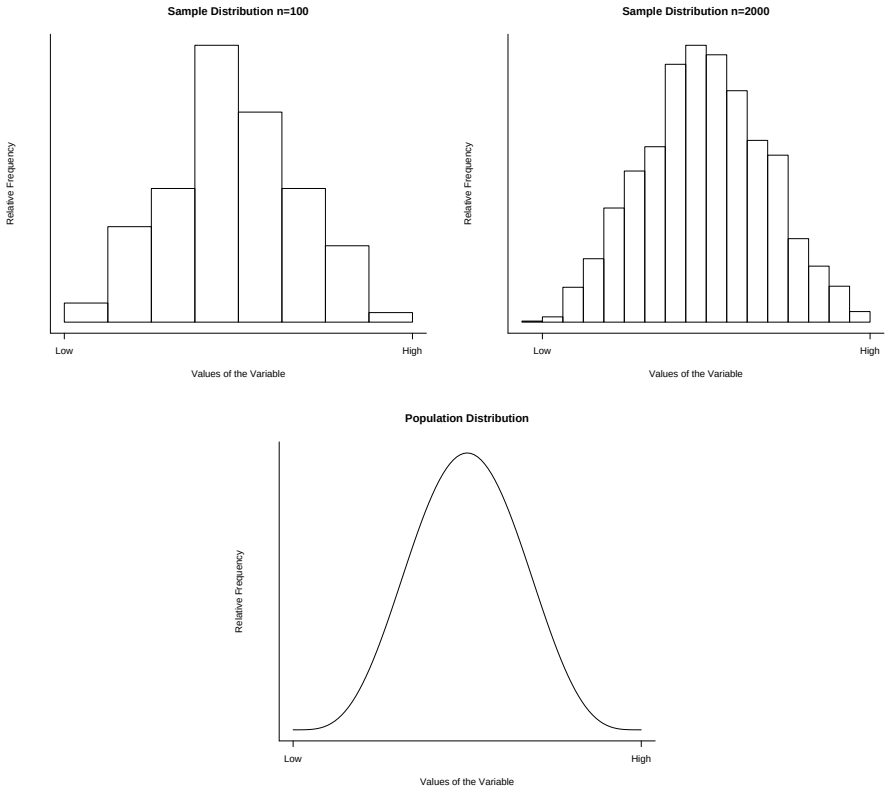

Figure 6: Sample and Population Distributions

One way to summarize a sample of population distribution is to describe its shape. A group for which the distribution is bell-shaped is fundamentally different from a group for which the distribution is U-shaped, for example.

The bell-shaped and U-shaped distributions in Figure 7. are **symmetric**. On the other hand, a nonsymmetric distribution is said to be **skewed to the right** or **skewed to the left**, according to which tail is longer.

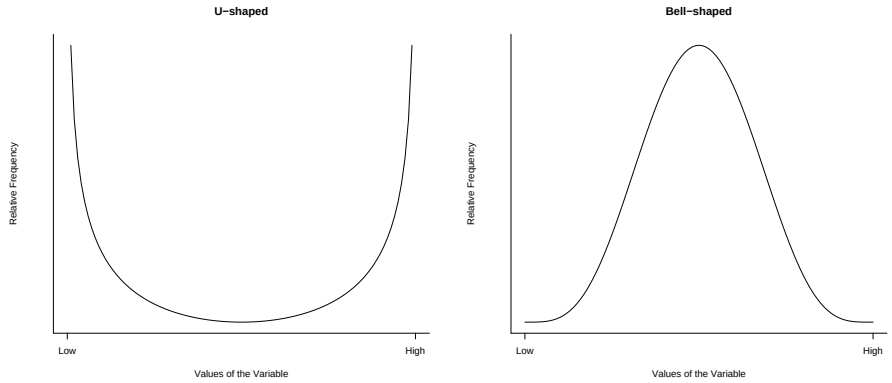

Figure 7: U-shaped and Bell-shaped Frequency Distributions

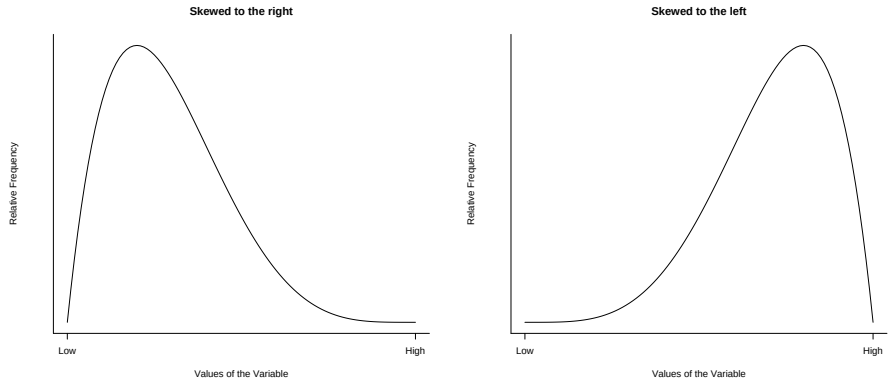

Figure 8: Skewed Frequency Distributions