3 Descriptive Analysis and Presentation of Bivariate Data

3.1 Bivariate Data

Weighing Your Fish with a Ruler

Weighing Your Fish with a Ruler

Bivariate data The values of two different variables that are obtained from the same population element.

Each of the two variables may be either *qualitative* or *quantitative*. As a result, three combinations of variable types can form bivariate data:

1. Both variables are qualitative (both attribute).
2. One variable is qualitative (attribute), and the other is quantitative (numerical).
3. Both variables are quantitative (both numerical).

Two Qualitative Variables

Two Qualitative Variables

When bivariate data result from two qualitative (attribute or categorical) variables, the data are often arranged on a **cross-tabulation** or **contingency table**. Let’s look at an example.

Example 1 – *Constructing Cross-Tabulation Tables*

Thirty students from our college were randomly identified and classified according to two variables: gender (M/F) and major (liberal arts, business administration, technology), as shown in Table 3.1.

Name	Gender	Major	Name	Gender	Major	Name	Gender	Major
Adams	M	LA	Feeney	M	T	McGowan	M	BA
Argento	F	BA	Flanigan	M	LA	Mowers	F	BA
Baker	M	LA	Hodge	F	LA	Ornt	M	T
Bennett	F	LA	Holmes	M	T	Palmer	F	LA
Brand	M	T	Jopson	F	T	Pullen	M	T
Brock	M	BA	Kee	M	BA	Rattan	M	BA
Chun	F	LA	Kleeberg	M	LA	Sherman	F	LA
Crain	M	T	Light	M	BA	Small	F	T
Cross	F	BA	Linton	F	LA	Tate	M	BA
Ellis	F	BA	Lopez	M	T	Yamamoto	M	LA

Genders and Majors of 30 College Students **[TA03-01]**

Table 3.1

Example 1 – *Constructing Cross-Tabulation Tables*

cont'd

These 30 bivariate data can be summarized on a 2×3 cross-tabulation table, where the two rows represent the two genders, male and female, and the three columns represent the three major categories of liberal arts (LA), business administration (BA), and technology (T).

The entry in each cell is found by determining how many students fit into each category. Adams is male (M) and liberal arts (LA) and is classified in the cell in the first row, first column.

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

See the red tally mark in Table 3.2.

Gender	Major		
	LA	BA	T
M	||||| (5)	|||||	(6)	||||| || (7)
F	|||||	(6)	|||| (4)	|| (2)

Cross-Tabulation of Gender and Major (tallied)

Table 3.2

The other 29 students are classified (tallied, shown in black) in a similar fashion.

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

The resulting 2 × 3 cross-tabulation (contingency) table, Table 3.3, shows the frequency for each cross-category of the two variables along with the row and column totals, called *marginal totals* (or *marginals*). The total of the marginal totals is the grand total and is equal to *n*, the sample size.

	Major			
Gender	LA	BA	T	Row Total
M	5	6	7	18
F	6	4	2	12
Col. Total	11	10	9	30

Cross-Tabulation of Gender and Major (frequencies)

Table 3.3

Example 1 – *Constructing Cross-Tabulation Tables* cont’d

Contingency tables often show percentages (relative frequencies). These percentages can be based on the entire sample or on the subsample (row or column) classifications.

Percentages Based on the Grand Total (Entire Sample)

The frequencies in the contingency table shown in Table 3.3 can easily be converted to percentages of the grand total by dividing each frequency by the grand total and multiplying the result by 100.

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

For example, 6 becomes 20% $\left[\left(\frac{6}{30}\right) \times 100 = 20\right]$.
See Table 3.4.

Gender	Major			
	LA	BA	T	Row Total
M	17%	20%	23%	60%
F	20%	13%	7%	40%
Col. Total	37%	33%	30%	100%

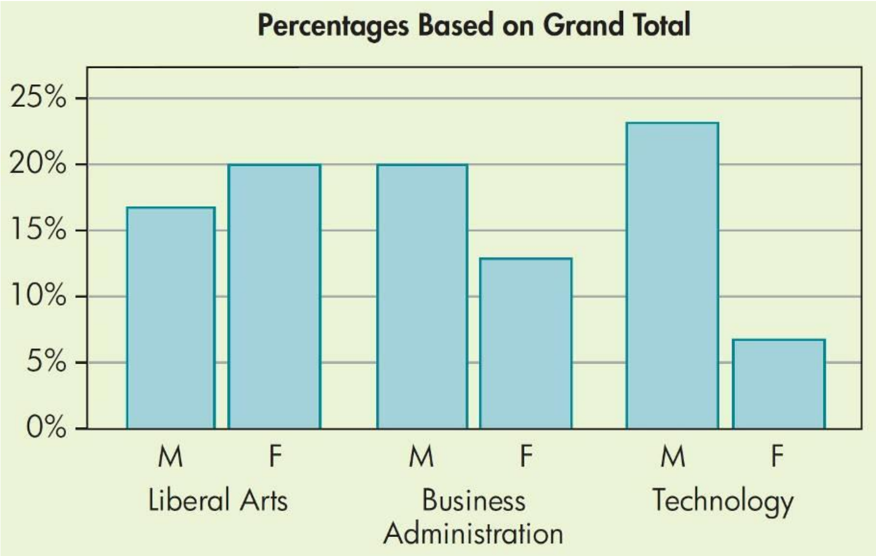
Cross-Tabulation of Gender and Major (relative frequencies; % of grand total)

Table 3.4

From the table of percentages of the grand total, we can easily see that 60% of the sample are male, 40% are female, 30% are technology majors, and so on.

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

These same statistics (numerical values describing sample results) can be shown in a bar graph (see Figure 3.1).



Bar Graph

Figure 3.1

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

Table 3.4 and Figure 3.1 show the distribution of male liberal arts students, female liberal arts students, male business administration students, and so on, relative to the entire sample.

	Major			
Gender	LA	BA	T	Row Total
M	17%	20%	23%	60%
F	20%	13%	7%	40%
Col. Total	37%	33%	30%	100%

Cross-Tabulation of Gender and Major (relative frequencies; % of grand total)

Table 3.4

Example 1 – *Constructing Cross-Tabulation Tables*

cont'd

Percentages Based on Row Totals

The frequencies in the same contingency table, Table 3.3, can be expressed as percentages of the row totals (or gender) by dividing each row entry by that row's total and multiplying the results by 100.

	Major			
Gender	**LA**	**BA**	**T**	**Row Total**
M	5	6	7	18
F	6	4	2	12
Col. Total	11	10	9	30

Cross-Tabulation of Gender and Major (frequencies)

Table 3.3

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

Table 3.5 is based on row totals.

Gender	Major LA	BA	T	Row Total
M	28%	33%	39%	100%
F	50%	33%	17%	100%
Col. Total	37%	33%	30%	100%

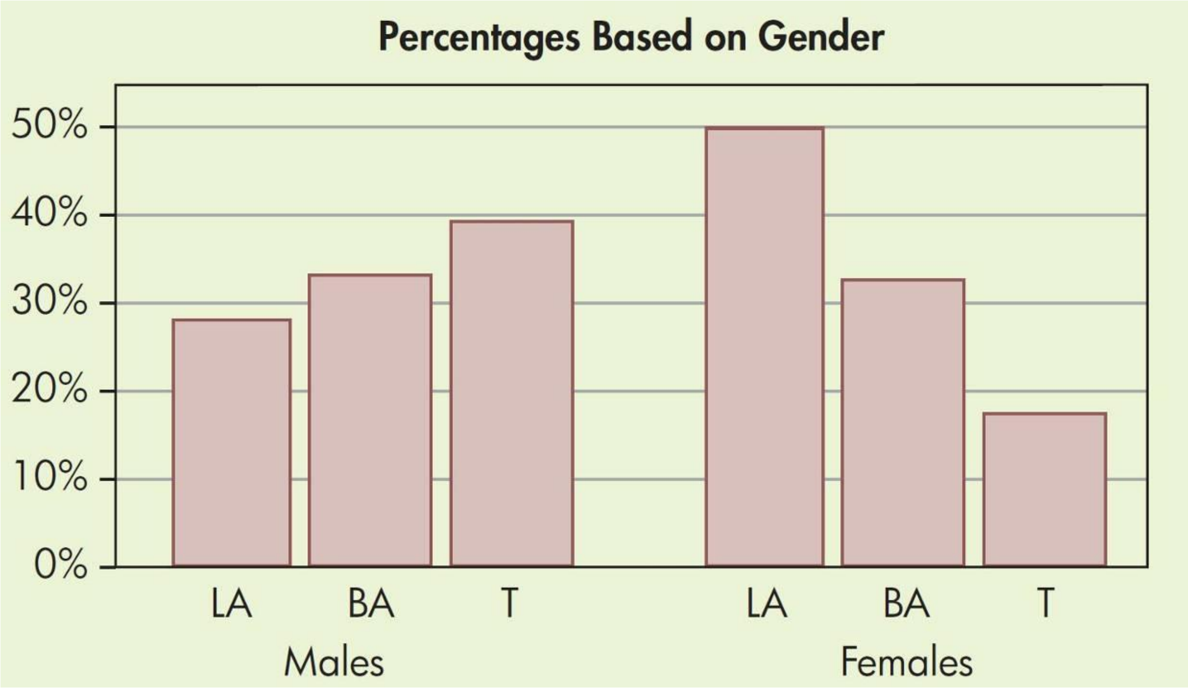
Cross-Tabulation of Gender and Major (% of row totals)

Table 3.5

From Table 3.5 we see that 28% of the male students are majoring in liberal arts, whereas 50% of the female students are majoring in liberal arts.

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

These same statistics are shown in the bar graph in Figure 3.2.



Bar Graph

Figure 3.2

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

Table 3.5 and Figure 3.2 show the distribution of the three majors for male and female students separately.

	Major			
Gender	**LA**	**BA**	**T**	**Row Total**
M	28%	33%	39%	100%
F	50%	33%	17%	100%
Col. Total	37%	33%	30%	100%

Cross-Tabulation of Gender and Major (% of row totals)

Table 3.5

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

Percentages Based on Column Totals

The frequencies in the contingency table, Table 3.3, can be expressed as percentages of the column totals (or major) by dividing each column entry by that column's total and multiplying the result by 100.

	Major			
Gender	LA	BA	T	Row Total
M	5	6	7	18
F	6	4	2	12
Col. Total	11	10	9	30

Cross-Tabulation of Gender and Major (frequencies)

Table 3.3

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

Table 3.6 is based on column totals.

	Major			
Gender	LA	BA	T	Row Total
M	45%	60%	78%	60%
F	55%	40%	22%	40%
Col. Total	100%	100%	100%	100%

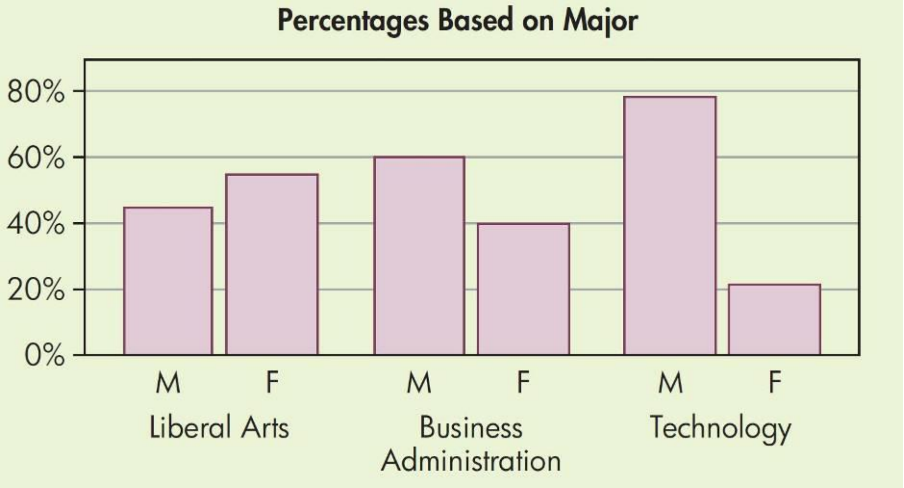
Cross-Tabulation of Gender and Major (% of column totals)

Table 3.6

From Table 3.6 we see that 45% of the liberal arts students are male, whereas 55% of the liberal arts students are female.

Example 1 – *Constructing Cross-Tabulation Tables* cont'd

These same statistics are shown in the bar graph in Figure 3.3.



Bar Graph

Figure 3.3

Table 3.6 and Figure 3.3 show the distribution of male and female students for each major separately.

One Qualitative and One Quantitative Variable

One Qualitative and One Quantitative Variable

When bivariate data result from one qualitative and one quantitative variable, the quantitative values are viewed as separate samples, each set identified by levels of the qualitative variable.

Example 2 – *Constructing Side-by-side Comparisons*

The distance required to stop a 3000-pound automobile on wet pavement was measured to compare the stopping capabilities of three tire tread designs (see Table 3.7).

Design A (n = 6)			Design B (n = 6)			Design C (n = 6)		
37	36	38	33	35	38	40	39	40
34	40	32	34	42	34	41	41	43

Stopping Distances (in feet) for Three Tread Designs **[TA03-07]**

Table 3.7

Tires of each design were tested repeatedly on the same automobile on a controlled wet pavement.

Example 2 – *Constructing Side-by-side Comparisons* cont'd

The design of the tread is a qualitative variable with three levels of response, and the stopping distance is a quantitative variable.

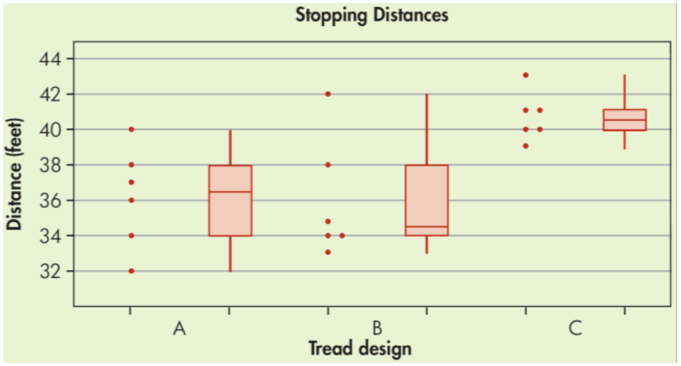
The distribution of the stopping distances for tread design A is to be compared with the distribution of stopping distances for each of the other tread designs.

This comparison may be made with both numerical and graphic techniques.

Example 2 – *Constructing Side-by-side Comparisons*

cont'd

Some of the available options are shown in Figure 3.4, Table 3.8, and Table 3.9.



Dotplot and Box-and-Whiskers Display Using a Common Scale

Figure 3.4

	Design A	Design B	Design C
High	40	42	43
Q_3	38	38	41
Median	36.5	34.5	40.5
Q_1	34	34	40
Low	32	33	39

5-Number Summary for Each Design

Table 3.8

	Design A	Design B	Design C
Mean	36.2	36.0	40.7
Standard deviation	2.9	3.4	1.4

Mean and Standard Deviation for Each Design

Table 3.9

Two Quantitative Variables

Two Quantitative Variables

When the bivariate data are the result of two quantitative variables, it is customary to express the data mathematically as **ordered pairs** (x, y), where x is the **input variable** (sometimes called the **independent variable**) and y is the **output variable** (sometimes called the **dependent variable**).

The data are said to be *ordered* because one value, x, is always written first.

They are called *paired* because for each x value, there is a corresponding y value from the same source.

Two Quantitative Variables

For example, if x is height and y is weight, then a height value and a corresponding weight value are recorded for each person.

The input variable, x, is measured or controlled in order to predict the output variable, y.

Suppose some research doctors are testing a new drug by prescribing different dosages and observing the lengths of the recovery times of their patients.

Two Quantitative Variables

The researcher can control the amount of drug prescribed, so the amount of drug is referred to as x.

In the case of height and weight, either variable could be treated as input and the other as output, depending on the question being asked. However, different results will be obtained from the regression analysis, depending on the choice made.

In problems that deal with two quantitative variables, we present the sample data pictorially on a *scatter diagram*.

Two Quantitative Variables

Scatter diagram A plot of all the ordered pairs of bivariate data on a coordinate axis system. The input variable, *x*, is plotted on the horizontal axis, and the output variable, *y*, is plotted on the vertical axis.

Note

When you construct a scatter diagram, it is convenient to construct scales so that the range of the *y* values along the vertical axis is equal to or slightly shorter than the range of the *x* values along the horizontal axis.

This creates a “window of data” that is approximately square.