Lecture 1

1.1. Data and Statistics

Summary

Statistics is the art and science of collecting, analyzing, presenting, and interpreting data. Nearly every college student majoring in business or economics is required to take a course in statistics.

Data consist of the facts and figures that are collected and analyzed. Four **scales** of measurement used to obtain data on a particular variable include **nominal**, **ordinal**, **interval**, and **ratio**. The scale of measurement for a variable is **nominal** when the data use labels or names to identify an attribute of an element. The scale is **ordinal** if the data demonstrate the properties of nominal data and the order or rank of the data is meaningful. The scale is **interval** if the data demonstrate the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Finally, the scale of measurement is **ratio** if the data show all the properties of interval data and the ratio of two values is meaningful.

For purposes of statistical analysis, data can be classified as **qualitative** or **quantitative**. **Qualitative** data use labels or names to identify an attribute of each element. Qualitative data use either the nominal or ordinal scale of measurement and may be nonnumeric or numeric. **Quantitative** data are numeric values that indicate how much or how many. Quantitative data use either the interval or ratio scale of measurement. Ordinary arithmetic operations are meaningful only if the data are quantitative. Therefore, statistical computations used for quantitative data are not always appropriate for qualitative data.

In Sections 1.4 and 1.5 we introduced the topics of descriptive statistics and statistical inference. **Descriptive statistics** are the tabular, graphical, and numerical methods used to summarize data. The process of statistical inference uses data obtained from a **sample** to make estimates or test hypotheses about the characteristics of a **population**. In the last section of the chapter we noted that computers facilitate statistical analysis. The larger data sets contained in Excel files.

Glossary

Statistics The art and science of collecting, analyzing, presenting, and interpreting data.

Data The facts and figures collected, analyzed, and summarized for presentation and interpretation.

Data set All the data collected in a particular study.

Elements The entities on which data are collected.

Variable A characteristic of interest for the elements.

Observation The set of measurements obtained for a particular element.

Nominal scale The scale of measurement for a variable when the data use labels or names to identify an attribute of an element. Nominal data may be nonnumeric or numeric.

Ordinal scale The scale of measurement for a variable if the data exhibit the properties of nominal data and the order or rank of the data is meaningful. Ordinal data may be nonnumeric or numeric.

Interval scale The scale of measurement for a variable if the data demonstrate the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure. Interval data are always numeric.

Ratio scale The scale of measurement for a variable if the data demonstrate all the properties of interval data and the ratio of two values is meaningful. Ratio data are always numeric.

Qualitative data Labels or names used to identify an attribute of each element. Qualitative data use either the nominal or ordinal scale of measurement and may be nonnumeric or numeric.

Quantitative data Numeric values that indicate how much or how many of something. Quantitative data are obtained using either the interval or ratio scale of measurement.

Qualitative variable A variable with qualitative data.

Quantitative variable A variable with quantitative data.

Cross-sectional data Data collected at the same or approximately the same point in time.

Time series data Data collected over several time periods.

Descriptive statistics Tabular, graphical, and numerical summaries of data.

Population The set of all elements of interest in a particular study.

Sample A subset of the population.

Census A survey to collect data on the entire population.

Sample survey A survey to collect data on a sample.

Statistical inference The process of using data obtained from a sample to make estimates or test hypotheses about the characteristics of a population.

1.2. Descriptive Statistics: Tabular and Graphical Presentation

Summary

A set of data, even if modest in size, is often difficult to interpret directly in the form in which it is gathered. Tabular and graphical methods provide procedures for organizing and summarizing data so that patterns are revealed and the data are more easily interpreted.

Frequency distributions, **relative frequency distributions**, **percent frequency distributions**, **bar graphs**, and **pie charts** were presented as tabular and graphical procedures for summarizing qualitative data.

Frequency distributions, **relative frequency distributions**, **percent frequency distributions**, **histograms**, **cumulative frequency distributions**, **cumulative relative frequency distributions**, **cumulative percent frequency distributions**, and **ogives** were presented as ways of summarizing quantitative data.

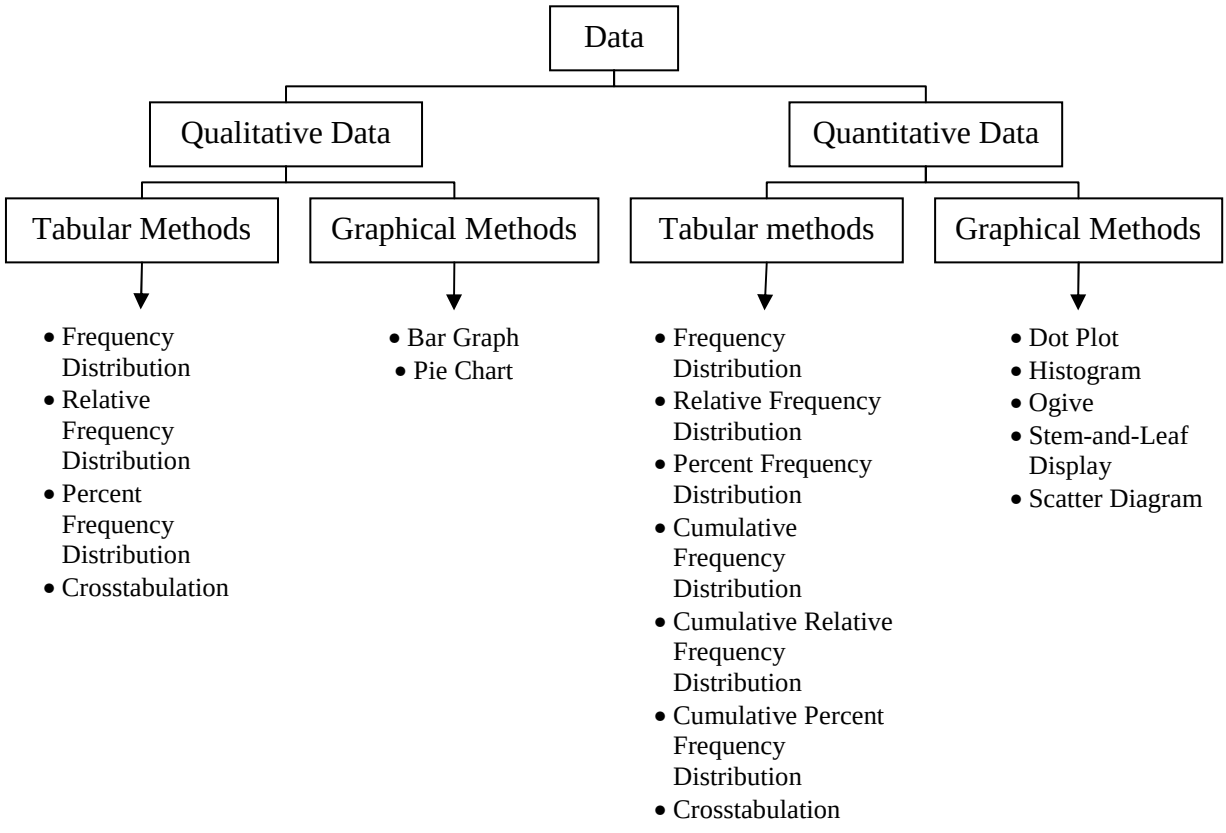
A **stem-and-leaf display** provides an exploratory data analysis technique that can be used to summarize quantitative data.

Crosstabulation was presented as a tabular method for summarizing data for two variables.

The **scatter diagram** was introduced as a graphical method for showing the relationship between two quantitative variables. Figure 2.9 shows the tabular and graphical methods presented in this chapter.

With large data sets, computer software packages are essential in constructing tabular and graphical summaries of data. **Excel** can be used for this purpose.

FIGURE.1 TABULAR AND GRAPHICAL METHODS FOR SUMMARIZING DATA



Glossary

Qualitative data Labels or names used to identify categories of like items.

Quantitative data Numerical values that indicate how much or how many.

Frequency distribution A tabular summary of data showing the number (frequency) of! items in each of several nonoverlapping classes.

Relative frequency distribution A tabular summary of data showing the fraction or proportion of data items in each of several nonoverlapping classes.

Percent frequency distribution A tabular summary of data showing the percentage of items in each of several nonoverlapping classes.

Bar graph A graphical device for depicting qualitative data that have been summarized in a frequency, relative frequency, or percent frequency distribution.

Pie chart A graphical device for presenting data summaries based on subdivision of a circle into sectors that correspond to the relative frequency for each class.

Class midpoint The value halfway between the lower and upper class limits.

Histogram A graphical presentation of a frequency distribution, relative frequency distribution, or percent frequency distribution of quantitative data constructed by placing the class intervals on the horizontal axis and the frequencies, relative frequencies, or percent frequencies on the vertical axis.

Cumulative frequency distribution A tabular summary of quantitative data showing the number of items with values less than or equal to the upper class limit of each class.

Cumulative relative frequency distribution A tabular summary of quantitative data showing the fraction or proportion of items with values less than or equal to the upper class limit of each class.

Cumulative percent frequency distribution A tabular summary of quantitative data showing the percentage of items with values less than or equal to the upper class limit of each class.

Ogive A graph of a cumulative distribution.

Exploratory data analysis Methods that use simple arithmetic and easy-to-draw graphs to summarize data quickly.

Dot plot A graphical device that summarizes data by the number of dots above each data value on the horizontal axis.

Stem-and-leaf display An exploratory data analysis technique that simultaneously rank orders quantitative data and provides insight about the shape of the distribution.

Crosstabulation A tabular summary of data for two variables. The classes for one variable are represented by the rows; the classes for the other variable are represented by the columns.

Simpson's paradox Conclusions drawn from two or more separate crosstabulations that can be reversed when the data are aggregated into a single crosstabulation.

Scatter diagram A graphical presentation of the relationship between two quantitative variables. One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.

Trendline A line that provides an approximation of the relationship between two variables.

Key Formulas

Relative Frequency

$$\frac{Frequency\ of\ the\ class}{n} \tag{1.1}$$

Approximate Class Width

$$\frac{Largest\ data\ value - Smallest\ data\ value}{Number\ of\ classes} \tag{1.2}$$

1.3. Descriptive Statistics: Numerical Measures

Summary

In this lecture we introduced several **descriptive statistics** that can be used to summarize the location, variability, and shape of a data distribution. Unlike the tabular and graphical procedures introduced in Chapter 2, the measures introduced in this chapter summarize the data in terms of numerical values. When the numerical values obtained are for a sample, they are called sample statistics. When the numerical values obtained are for a population, they are called population parameters. Some of the notation used for sample statistics and population parameters follow (In statistical inference, the sample statistic is referred to as the point estimator of the population parameter).

	Sample Statistic	Population Parameter
Mean	$\bar{x}$ or m	μ
Proportion	p	π
Variance	s^2	σ^2
Standard deviation	s	σ
Covariance	s_{xy}	σ_{xy}
Correlation	r_{xy}	ρ_{xy}

As measures of central location, we defined the **mean**, **median**, and **mode**. Then the concept of percentiles was used to describe other locations in the data set. Next, we presented the **range**, **interquartile range**, **variance**, **standard deviation**, and **coefficient of variation** as measures of variability or dispersion. Our primary measure of the shape of a data distribution was the **skewness**. Negative values indicate a data distribution skewed to the left. Positive values indicate a data distribution skewed to the right. We then described how the mean and standard deviation could be used, applying Chebyshev's theorem and the empirical rule, to provide more information about the distribution of data and to identify outliers.

In Section 3.4 we showed how to develop a five-number summary and **a box plot** to provide simultaneous information about the location, variability, and shape of the distribution. In Section 3.5 we introduced **covariance** and the **correlation coefficient** as measures of association between two variables. In the final section, we showed how to compute a **weighted mean** and how to calculate a mean, variance, and standard deviation for **grouped data**.

Glossary

Sample statistic A numerical value used as a summary measure for a sample (e.g., the sample mean, $\bar{x}$, the sample variance, s^2, and the sample standard deviation, s).

Population parameter A numerical value used as a summary measure for a population (e.g., the population mean, μ, the population variance, σ^2, and the population standard deviation, σ).

Point estimator The sample statistic, such as $\bar{x}$, s^2, and s, when used to estimate the corresponding population parameter.

Mean A measure of central location computed by summing the data values and dividing by the number of observations.

Median A measure of central location provided by the value in the middle when the data are arranged in ascending order.

Mode A measure of location, defined as the value that occurs with greatest frequency.

Percentile A value such that at least p percent of the observations are less than or equal to this value and at least $(100 - p)$ percent of the observations are greater than or equal to this value. The 50th percentile is the median.

Quartiles The 25th, 50th, and 75th percentiles, referred to as the first quartile, the second quartile (median), and third quartile, respectively. The quartiles can be used to divide a data set into four parts, with each part containing approximately 25% of the data.

Range A measure of variability, defined to be the largest value minus the smallest value.

Interquartile range (IQR) A measure of variability, defined to be the difference between the third and first quartiles.

Variance A measure of variability based on the squared deviations of the data values about the mean.

Standard deviation A measure of variability computed by taking the positive square root of the variance.

Coefficient of variation A measure of relative variability computed by dividing the standard deviation by the mean and multiplying by 100.

Skewness A measure of the shape of a data distribution. Data skewed to the left result in negative skewness; a

symmetric data distribution results in zero skewness; and data skewed to the right result in positive skewness.

z-score A value computed by dividing the deviation about the mean (x_i — $\bar{x}$) by the standard deviation s. A z-score is referred to as a standardized value and denotes the number of standard deviations x_i is from the mean.

Chebyshev's theorem A theorem that can be used to make statements about the proportion of data values that must be within a specified number of standard deviations of the mean. **At least (1– 1/z^2) of the data must be within *z* standard deviations from the mean, where *z* – any value > 1.**

Empirical rule A rule that can be used to compute the percentage of data values that must be within one, two, and three standard deviations of the mean for data that exhibit a bell-shaped distribution.

Outlier An unusually small or unusually large data value.

Five-number summary An exploratory data analysis technique that uses five numbers to summarize the data: smallest value, first quartile, median, third quartile, and largest value.

Box plot A graphical summary of data based on a five-number summary.

Covariance A measure of linear association between two variables. Positive values indicate a positive relationship; negative values indicate a negative relationship.

Correlation coefficient A measure of linear association between two variables that takes on values between -1 and +1. Values near +1 indicate a strong positive linear relationship, values near -1 indicate a strong negative linear relationship; and values near zero indicate the lack of a linear relationship.

Weighted mean The mean obtained by assigning each observation a weight that reflects its importance.

Grouped data Data available in class intervals as summarized by a frequency distribution. Individual values of the original data are not available..

Key formulas

Sample Mean

$$\bar{x} = \frac{\sum x_i}{n} \tag{1.3}$$

Population Mean

$$\mu = \frac{\sum x_i}{N} \tag{1.4}$$

Sample proportion

$$p = \frac{\sum (x_i = true)}{n} \tag{1.5}$$

Interquartile Range

$$IQR = Q_3 - Q_1 \tag{1.6}$$

Population Variance

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{1.7}$$

Sample Variance

$$s^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n-1} \tag{1.8}$$

Standard Deviation

$$Sample\ standard\ deviation = s = \sqrt{s^2} \tag{1.9}$$

$$Population\ standard\ deviation = \sigma = \sqrt{\sigma^2} \tag{1.10}$$

Coefficient of Variation

$$\left(\frac{Standard\ deviation}{Mean} \times 100\right)\% \tag{1.11}$$

Skewness of the sample data

$$Skewness = \frac{n}{(n-1)(n-2)} \sum_i \left(\frac{x_i - \bar{x}}{s}\right)^3 \tag{1.12}$$

z-Score

$$z_i = \frac{x_i - \bar{x}}{s} \tag{1.13}$$

Sample Covariance

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{1.14}$$

Population Covariance

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N} \tag{1.15}$$

Pearson Product Moment Correlation Coefficient: Sample Data

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{1.16}$$

Pearson Product Moment Correlation Coefficient: Population Data

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{1.17}$$

Weighted Mean

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \tag{1.18}$$

Sample Mean for Grouped Data

$$\bar{x} = \frac{\sum f_i M_i}{n} \tag{1.19}$$

Sample Variance for Grouped Data

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n-1} \tag{1.20}$$

Population Mean for Grouped Data

$$\mu = \frac{\sum f_i M_i}{N} \tag{1.21}$$

Population Variance for Grouped Data

$$\sigma^2 = \frac{\sum f_i (M_i - \mu)^2}{N} \tag{1.22}$$

Lecture supplementary material

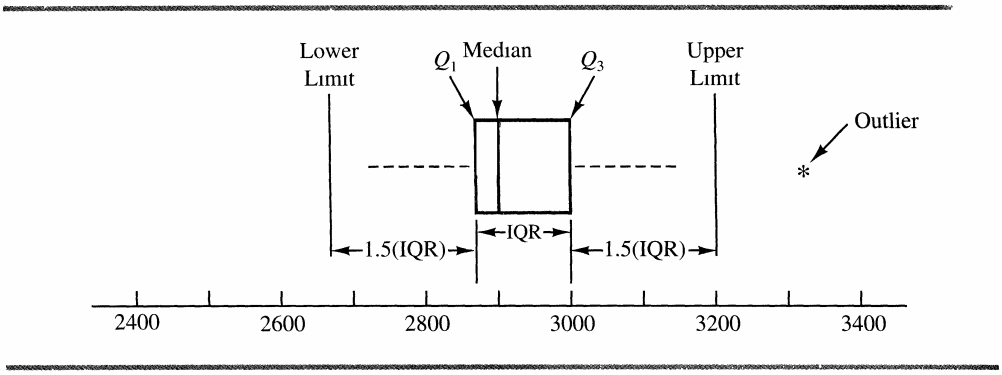
Suppose that a college placement office sent a questionnaire to a sample of graduates requesting information on monthly starting salaries. Table 1.1 shows the

TABLE 1.1 MONTHLY STARTING SALARIES FOR A SAMPLE OF 12 GRADUATES

Graduate	1	2	3	4	5	6	7	8	9	10	11	12
Starting Salary, $	2850	2950	3050	2880	2755	2710	2890	3130	2940	3325	2920	2880

2710	2755	2850	2880	2880	2890	2920	2940	2950	3050	3130	3325
		Q_1=2865			Q_2=2905 (Median)			Q_3=3000			

FIGURE 3.5 BOX PLOT OF THE STARTING SALARY DATA WITH LINES SHOWING THE LOWER AND UPPER LIMITS



NOTE: other variants of calculation of "whiskers" exist!