Chapter 1

Overview and Descriptive Statistics

What is statistics and why it is important?

The discipline of statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.

Without the presence of uncertainty and variation, there would be little need for statistical methods or statisticians.

Its ultimate goal is translating data into knowledge and understanding of the world around us.

Statistics is the art (and science) of learning from data. It is used to do inference or prediction based on the collected data.

Statistics is very important for the decision making in industry and all other areas.

Examples:

1. Weather forecasts: Computer models are built using statistics that compare prior weather conditions with current weather to predict future weather.

2. Political forecasting: Election forecasting models usually combine hundreds of opinion polls with historical and demographic information to calculate the odds.

3. Auto and home insurance: The premium rate that an insurance company charges you is based upon statistics from all drivers or homeowners in you area.

4. Biology and public health: Statisticians are helping find the important genes, called markers, which influence the whole network of genes. These markers can be used to predict disease risk.

5. Stock market: Stock analysts also use statistical and computational models to forecast what is happening in the economy and to predict the stock trend in the future.

6. Quality Testing: Companies make thousands of products every day and they want to make sure that good quality items are sold. But a company can't test each and every time that they sell a product to you, the consumer. So the company uses statistics to test just a few, called a sample, of what they make. If the sample passes quality tests, then the company assumes that all the items made in the group are good.

1.1 Populations, Samples, and Processes

Table 1: Student information collected from STAT 155 class.

	Gender	Height(in)	IQ
Student 1	M	70	110
Student 2	M	73	121
Student 3	F	65	108
Student 4	M	78	135
Student 5	F	63	115
Student 6	F	68	138
Student 7	F	72	113
Student 8	M	76	123
Student 9	M	68	117
Student 10	F	66	140

Introduction of Basic Terms

- **Population** – the entire collection of objects whose properties are to be analyzed in a particular study.

- **Sample** – a subset of the population.

- **Variable** – any characteristic of interest for each object in a population or a sample.
  alphabet.
  x = Gender
  y = Height
  z = IQ
- **Observation** – the set of measurements obtained for a particular object.

- **Parameter** – a numerical value summarizing the population data.

- **Statistic** – a numerical value summarizing the sample data.

Types of Variables

- **Qualitative/Categorical variables** use labels or names to identify an attribute of an element. Each data value belongs to one of a set of categories. *Arithmetic operations, such as addition and averaging, are NOT meaningful for data resulting from a categorical variable.*

- **Quantitative/Numerical variables** use numeric values that represent different magnitudes of the variable. *Arithmetic operations, such as addition and averaging, are meaningful for data resulting from a numerical variable.*

- **Discrete variable** – a quantitative variable that can take on only a finite or at most a countably infinite number of values. Intuitively, a discrete variable can assume values corresponding to isolated points along a line interval. That is, there is a gap between any two values.

- **Continuous variable** – a quantitative variable that can assume an uncountable number of values. Intuitively, a continuous variable can assume any value along a line interval, including every possible value between any two values.

Univariate, Bivariate and Multivariate data

- **Univariate data** – observations on a single variable

- **Bivariate data** – observations on two variables

- **Multivariate data** – observations on more than one variable

Branches of Statistics

- **Descriptive Statistics** – summarize and present important features of the data in a form that is easy for a reader to understand. These summaries may be tabular, graphical, or numerical.
- **Inferential statistics** – use techniques for generalizing from a sample to a population.

The measurements we make of a variable vary from object to object.

Likewise, results of descriptive and inferential statistics vary, depending on the sample chosen.

The study of variability is a key part of statistics.

Collecting Data

Statistics deals not only with the organization and analysis of data once it has been collected but also with the development of techniques for collecting data.

Sample survey is the process of collecting data on a sample. **Census** is the process of collecting data on the entire population.

It is important to obtain good, representative data. Inferences are made based on statistics obtained from the data. If data are not properly collected, an investigator may not be able to answer the questions under consideration with a reasonable degree of confidence.

With simple random sampling, each subset of objects of the specified size in the population has the same chance of being the sample. This is desirable, because then the sample tends to be a good reflection of the population.

Sometimes alternative sampling methods can be used to make the selection process easier, to obtain extra information, or to increase the degree of confidence in conclusions.

One such method, stratified sampling, entails separating the population units into nonoverlapping groups and taking a sample from each one using simple random sampling.

Frequently, a convenience sample is obtained by selecting objects without systematic randomization.

1.2 Pictorial and Tabular Methods in Descriptive Statistics

Descriptive statistics can be divided into two general subject areas. In section 1.2, we consider representing a data set using visual techniques.

- stem-and-leaf displays
- dotplots
- histograms

In sections 1.3 and 1.4, we will develop some numerical summary measures for data sets.

Notation

n = sample size (number of observations in the sample)

Data values occurring in a sample are symbolically represented by $x_1, x_2, x_3, \cdots, x_n$.

Stem-and-Leaf Display

It's a quick way to obtain an informative visual representation of a numerical data set.

Steps for Constructing a Stem-and-Leaf Display

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
2. List possible stem values in a vertical column.

3. Record the leaf for every observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

A stem-and-leaf display conveys information about the following aspects of the data:

- identification of a typical or representative value
- extent of spread about the typical value
- presence of any gaps in the data
- extent of symmetry in the distribution of values
- number and location of peaks
- presence of any outlying values

Dotplots

A dotplot is an attractive summary of numerical data when the data set is reasonable small or there are relatively few distinct data values.

Each observation is represented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically.

As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

Shapes of Distributions

Uniform

Unimodal Bimodal Multimodal

Mound-shaped, Bell-shaped, Symmetrical

Positively Skewed, Right Skewed

Negatively Skewed, Left Skewed

Exercise 1.12 The accompanying specific gravity values for various wood types used in construction appeared in the article Bolted Connection Design Values Based on European Yield Model (*J. of Structural Engr.*, 1993: 2169-2186):

.31 .35 .36 .36 .37 .38 .40 .40 .40
.41 .41 .42 .42 .42 .42 .42 .43 .44
.45 .46 .46 .47 .48 .48 .48 .51 .54
.54 .55 .58 .62 .66 .66 .67 .68 .75

Construct a stem-and-leaf display using repeated stems and comment on any interesting features of the display.

Histograms

Constructing a Histogram for **Discrete** Data

1. Determine the frequency and relative frequency of each value of x.
2. Mark possible x values on a horizontal scale.
3. Above each value, draw a rectangle whose height is the relative frequency (or frequency) of that value.

*Note: The same method can be applied to **qualitative/categorical** data too.*

Consider data consisting of observations on a **discrete variable** x:

- The **frequency** of any particular x value is the number of times that value occurs in the data set.
- The **relative frequency** of a value is the fraction or proportion of times the value occurs.

  $$\text{relative frequency of a value} = \frac{\text{\# of times the value occurs}}{\text{\# of observations in the data set}}$$

- A **frequency distribution** is a tabulation of the frequencies and/or relative frequencies.

Constructing a Histogram for **Continuous** Data: **Equal** Class Widths

1. Subdivide the measurement axis into a suitable number of equal-width **class intervals** or **classes**. Each observation is contained in exactly one class. An observation on a boundary is placed in the interval to the right of the boundary.
2. Determine the frequency and relative frequency for each class.
3. Mark the class boundaries on a horizontal measurement axis.
4. Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

*Note: The same method can be applied to **discrete** data too.*

There are no hard-and-fast rules concerning either the number of classes or choice of classes themselves. Between 5 and 20 classes will be satisfactory for most data sets. Generally, the larger the number of observations in a data set, the more classes should be used. A reasonable rule of thumb is

$$\#\text{ of classes} \approx \sqrt{\#\text{ of observations}}$$

Equal-width classes may not be a sensible choice if a data set "stretches out" to one side or the other – there are some regions of the measurement scale that have a high concentration of data values and other parts where data is quite sparse.

In that case, using a small number of equal-width classes results in almost all observations falling in just one or two of the classes. If a large number of equal-width classes are used, many classes will have zero frequency.

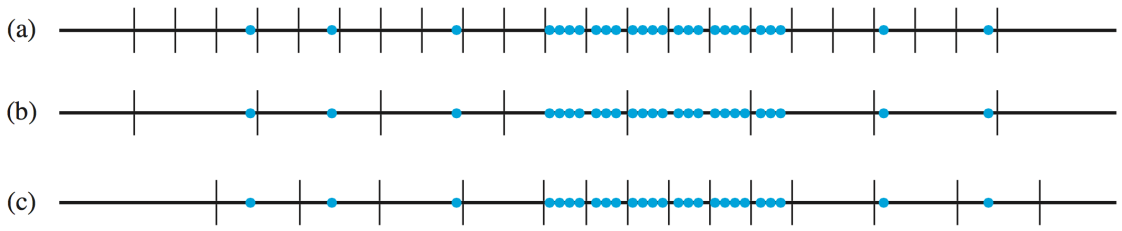

Figure 1.9 Selecting class intervals for "varying density" data: (a) many short equal-width intervals; (b) a few wide equal-width intervals; (c) unequal-width intervals

A sound choice is to use a few wider intervals near extreme observations and narrower intervals in the region of high concentration.

Constructing a Histogram for **Continuous** Data: **Unequal** Class Widths

1. Subdividing the measurement axis to a reasonable number of unequal-width class intervals or classes.
2. Determining frequencies and relative frequencies.
3. Calculate the height of each rectangle

$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

Note: The same method can be applied to ***discrete*** *data too.*

The resulting rectangle heights are usually called **densities**, and the vertical scale is the **density scale**.

This prescription will also work when class widths are equal.

A density histogram has one interesting property. Multiplying both sides of the formula for density by the class width gives

$$\begin{aligned} \text{relative frequency} &= (\text{class width})(\text{density}) \\ &= (\text{rectangle width})(\text{rectangle height}) \\ &= \text{rectangle area} \end{aligned}$$

That is, the area of each rectangle is the relative frequency of the corresponding class.

Furthermore, since the sum of relative frequencies should be 1, the total area of all rectangles in a density histogram is 1.

Exercise 1.27 The paper "Study on the Life Distribution of Microdrills" (*J. of Engr. Manufacture*, 2002: 301305) reported the following observations, listed in increasing order, on drill lifetime (number of holes that a drill machines before it breaks) when holes were drilled in a certain brass alloy.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

(a) Construct a relative frequency histogram based on the equal-width class intervals $0-<100$, $100-<200$, $200-<300$, $\cdots$, and comment on features of the histogram.

(b) Construct a density histogram based on the following unequal-width class intervals $0-<50$, $50-<100$, $100-<150$, $150-<200$, $200-<300$, $300-<600$, and comment on features of the histogram.

(c) What proportion of the lifetime observations in this sample are less than 100? What proportion of the observations are at least 200?

1.3 Measures of Location

- mean
- median
- quartiles
- trimmed mean

Mean – the arithmetic average of the data. Mean is calculated as the sum of all observations divided by the number of observations.

Sample mean $\overline{x}$ of observations $x_1, x_2, x_3, \cdots, x_n$ is:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For reporting $\overline{x}$, it is recommended to use a decimal accuracy of one digit more than the accuracy of the x_i's.

The arithmetic mean is the most widely used measure of central location. However, it is oversensitive to extreme/outlying values and must be used with caution.

Median – the middle value.

Sample median $\widetilde{x}$ is the middle sorted observation. That is, we want a value such that half of the data is smaller than it and half is greater than it.

Steps to find the sample median:

1. Rank the n observations from smallest to largest.

2. If n is odd, median equals to the middle value,
   $\widetilde{x} = (\frac{n+1}{2})^{\text{th}}$ ordered value;
   If n is even, there are two middle values whose average equals the median,
   $\widetilde{x}$ = average of $(\frac{n}{2})^{\text{th}}$ and $(\frac{n+1}{2})^{\text{th}}$ ordered values.

Unlike the mean, the median is insensitive to extreme/outlying values.

In many samples, the relationship between the mean and the median can be used to assess the shape of a distribution.

Positively/Right Skewed

Symmetrical

Negatively/Left Skewed

Example: In 1994, the Major League Baseball Players association claimed that the median salary for a baseball player was $450,000. The owners reported that the mean salary was $1,168,263. These numbers tell different stories about baseball salaries.

Quartiles – values of the variable that separate a ranked data set into 4 equal parts.

Order n observations from smallest to largest and separate the smallest half from the largest half; the median is included in both halves if n is odd.

- Q_1 **lower quartile** or **lower fourth** – the median of the smallest half of the data
- Q_2 – the median of the data
- Q_3 **upper quartile** or **upper fourth** – the median of the largest half of the data

Quartiles will be used to construct boxplot in Section 1.4.

The mean is quite sensitive to a single outlier, whereas the median is impervious to many outliers.

The mean and the median are at opposite extremes of the same family of measures. The mean is the average of all the data, whereas the median results from eliminating all but the middle one or two values and then averaging.

That is, the mean involves trimming 0% from each end of the sample, whereas for the median the maximum possible amount is trimmed from each end.

Trimmed mean $\overline{x}_{tr(100\alpha)}$ – a compromise between the mean $\overline{x}$ and the median $\widetilde{x}$ in which the smallest $100\alpha\%$ and the largest $100\alpha\%$ of the data are eliminated and the average is computed from what is left over.

A trimmed mean with a moderate trimming percentage – someplace between 5% and 25% – will yield a measure of center that is neither as sensitive to outliers as is the mean nor as insensitive as the median.

If the desired trimming percentage is $100\alpha\%$ and $n\alpha$ is not an integer, the trimmed mean must be calculated by interpolation using the appropriate weighted average.

Exercise 1.40 Compute the sample mean, the sample median, the quartiles, and the 10% trimmed mean for the lifetime data given in Exercise 1.27, and compare these measures.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

Categorical Data and Sample Proportions

Consider sampling a population that consists of only two categories. If we let x denote the number in the sample falling in category 1, then the number in category 2 is $n - x$. The relative frequency or **sample proportion** in category 1 is x/n and the sample proportion in category 2 is $1 - x/n$.

Focus attention on a particular category and code the sample results so that a 1 is recorded for an observation in the category and a 0 for an observation not in the category. Then the sample proportion of observations in the category is the sample mean of the sequence of 1s and 0s.

Thus, a sample mean can be used to summarize the results of a categorical sample.

These remarks also apply to situations in which categories are defined by grouping values in a numerical sample or population (e.g., we might be interested in knowing whether individuals have owned their present automobile for at least 5 years, rather than studying the exact length of ownership).

---

	sample statistic	population parameter
mean		
median		
proportion		

Exercise 1.41 A sample of n = 10 automobiles was selected, and each was subjected to a 5-mph crash test. Denoting a car with no visible damage by S (for success) and a car with such damage by F, results were as follows:

S S F S S S F F S S

(a) What is the value of the sample proportion of successes x/n?

(b) Replace each S with a 1 and each F with a 0. Then calculate $\overline{x}$ for this numerically coded sample. How does $\overline{x}$ compare to x/n?

(c) Suppose it is decided to include 15 more cars in the experiment. How many of these would have to be S's to give x/n = .80 for the entire sample of 25 cars?

1.4 Measures of Variability

- range
- variance
- standard deviation

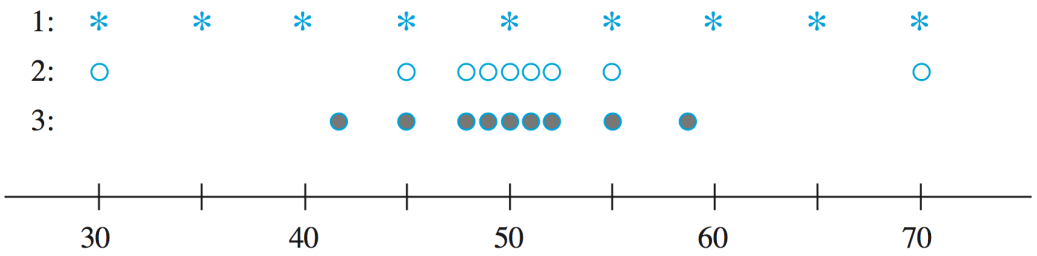

Figure 1.19 Samples with identical measures of center but different amounts of variability

Range – The difference between the largest and smallest sample values.

A defect of the range is that it depends on only the two most extreme observations and disregards the positions of the remaining $n-2$ values.

Our primary measures of variability involve the **deviations** from the mean, $x_i - \overline{x}$, for $i = 1, 2, \cdots, n$.

A deviation will be positive if the observation is larger than the mean and negative if the observation is smaller than the mean.

Note that sum of deviations $= \sum(x_i - \overline{x}) = 0$.

Sample variance – the average of the squared deviations from the mean.

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

Why $n-1$?

It is customary to refer to s^2 as being based on $n-1$ **degrees of freedom** (df). This terminology results from the fact that although s^2 is based on n deviations, these sum to zero. Therefore, only $n-1$ of the deviations are freely determined.

An alternative computational formula for the numerator of s^2 is

$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - n\overline{x}^2$$

Sample standard deviation – the (positive) square root of the sample variance.

$$s = \sqrt{s^2}$$

The unit of s is the same as the unit of each of the x_is, and is more easily interpreted than the variance. The bigger s is, the more spread out the data is.

---

	sample statistic	population parameter
variance		
standard deviation		

Sample statistics are **point estimates** of corresponding population parameters. We'll learn point estimation in Chapter 6.

Some Properties of the Mean, Variance and Standard Deviation

- Change the origin:
  If $y_i = x_i + c$, for $i = 1, 2, \cdots, n$, then

$$\begin{aligned} \overline{y} &= \overline{x} + c \\ s_y^2 &= s_x^2 \\ s_y &= s_x \end{aligned}$$

- Change the scale:
  If $y_i = cx_i$, for $i = 1, 2, \cdots, n$, then

$$\begin{aligned} \overline{y} &= c\overline{x} \\ s_y^2 &= c^2 s_x^2 \\ s_y &= cs_x \end{aligned}$$

Therefore, adding a constant to each data value does not change the sample variance; whereas, multiplying each data value by a constant results in a new sample variance that is equal to the old one multiplied by the square of the constant.

Exercise 1.45 The value of Young's modulus (GPa) was determined for cast plates consisting of certain intermetallic substrates resulting in the following sample observations ("Strength of Modulus of a Molybdenum-Coated Ti-25Al-10Nb-3U-1Mo Intermetallic," *J. of Materials Engr. And Performance*, 1997: 46-50):

116.4 115.9 114.6 115.2 115.8

(a) Calculate $\overline{x}$ and the deviations from the mean.

(b) Use the deviations calculated in part (a) to obtain s^2 and s.

(c) Calculate s^2 by using the computational (shortcut) formula for S_{xx}.

(d) Subtract 100 from each observation to obtain a sample of transformed values. Now calculate the sample variance of these transformed values can compare it to s^2 for the original data.

Boxplot – a pictorial summary that describes the following most prominent features of the data

- center
- spread
- the extent and nature of any departure from symmetry
- identification of "outliers"

Steps for Constructing a Boxplot that Shows Outliers

1. Draw a horizontal measurement scale.
2. Draw vertical lines at the lower fourth/quartile Q_1, the median $\widetilde{x}$, and the upper fourth/quartile Q_3. Enclose these vertical lines in a box.
3. Compute the **fourth spread** f_s or the **interquartile range (IQR)**. IQR $= f_s =$ upper fourth $-$ lower fourth $= Q_3 - Q_1$
4. Detect outliers.
   Any observation farther than $1.5f_s$ beyond the closest fourth is an **outlier**. That is, an outlying value is less than $Q_1 - 1.5f_s$ or greater than $Q_3 + 1.5f_s$.
   An outlier is **extreme** if it is more than $3f_s$ beyond the nearest fourth, and it is **mild** otherwise.
5. Draw a whisker from Q_1 to the smallest data value that is larger than the lower fence and a whisker from Q_3 to the largest data value that is smaller than the upper fence.
6. Represent each mild outlier by a closed circle and each extreme outlier by an open circle.

Outliers distort both the mean and the standard deviation, since neither is resistant. Statistical inference based a set of data that contains outliers could be flawed.

Example Use the lifetime data given in Exercise 1.27, construct a boxplot that shows outliers. Describe/summarize the data.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

Distribution Shape Based upon Boxplot

If the median is near the center of the box and each of the horizontal lines is of approximately equal length, then the distribution is roughly symmetric.

If the median is to the left of the center of the box or the right line is substantially longer than the left line, the distribution is positively/right skewed.

If the median is to the right of the center of the box or the left line is substantially longer than the right line, the distribution is negatively/left skewed.

Comparative Boxplots

A comparative or side-by-side boxplot is a very effective way of revealing similarities and differences between two or more data sets consisting of observations on the same variable.