Introduction

Describing Data Sets

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.

Frequency Tables and Graphs

A data set having a relatively small number of distinct values can be conveniently presented in a *frequency table*.

Starting Salary	Frequency
47	4
48	1
49	3

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.

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

Relative Frequency Tables and Graphs

Consider a data set consisting of n values. If f is the frequency of a particular value, then the ratio

$$\frac{f}{n}$$

is called its *relative frequency*.

A *pie chart* is often used to indicate relative frequencies when the data are not numerical in nature.

A bar graph plot of class data, with the bars placed adjacent to each other, is called a *histogram*.

The vertical axis of a histogram can represent either the class frequency or the relative class frequency; in the former case the graph is called a *frequency histogram* and in the latter a *relative frequency histogram*.

A cumulative frequency plot is called an *ogive*.

An efficient way of organizing a small- to moderate-sized data set is to utilize a *stem and leaf plot*. Such a plot is obtained by first dividing each data value into two parts – its stem and its leaf.

Summarizing Data Sets

Sample Mean, Sample Median, and Sample Mode

Sample Mean

To begin, suppose that we have a data set consisting of the n numerical values $x_1, x_2, \ldots, x_n$.
The sample mean is the arithmetic average of these values.

Definition The *sample mean*, designated by $\bar{x}$, is defined by

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

The computation of the sample mean can often be simplified by noting that if for constants a and b

$$y_i = ax_i + b, i = 1, \ldots, n$$

then the sample mean of the data set $y_1, y_2, \ldots, y_n$ is

$$\bar{y} = \sum_{i=1}^{n} \frac{(ax_i + b)}{n} = \sum_{i=1}^{n} \frac{ax_i}{n} + \sum_{i=1}^{n} \frac{b}{n} = a\bar{x} + b$$

Example The winning scores in the U.S. Masters golf tournament in the years from 1982 to 1991 were as follows:

$$284, 280, 277, 282, 279, 285, 281, 283, 278, 277$$

Find the sample mean of these scores.

Sometimes we want to determine the sample mean of a data set that is presented in a frequency table listing the k distinct values $v_1, v_2, \ldots, v_k$ having corresponding frequencies $f_1, f_2, \ldots, f_k$.
Since such a data set consists of $n = \sum_{i=1}^{k} f_i$ observations, which the value v_i appearing f_i times, for

each $i = 1, 2, \ldots, k$, it follows that the sample mean of these n data values is

$$\bar{x} = \sum_{i=1}^{k} \frac{v_i f_i}{n}$$

By writing the preceding as

$$\bar{x} = \frac{f_1}{n} v_1 + \frac{f_2}{n} v_2 + \cdots + \frac{f_k}{n} v_k$$

we see that the sample mean is a *weighted average* of the distinct values, where the weight given to the value v_i is equal to the proportion of the n data values that are equal to v_i, $i = 1, 2, \ldots, k$.

Sample Median

Another statistic used to indicate the center of a data set is the *sample median*; loosely speaking, it is the middle value when the data set is arranged in increasing order.

Definition Order the values of a data set of size n from smallest to largest.
If n is odd, the *sample median* is the value in position $\frac{n+1}{2}$.
If n is even, the *sample median* is the average of the values in positions $\frac{n}{2}$ and $\frac{n}{2} + 1$.

The sample mean and sample median are both useful statistics for describing the central tendency of a data set.
The sample mean makes use of all the data values and is affected by extreme values that are much larger or smaller than the others; the sample median makes use of only one or two of the middle values and is thus not affected by extreme values.
Which of them is more useful depends on what one is trying to learn from the data.

Sample Mode

Another statistic that has been used to indicate the central tendency of a data set is the *sample mode*, defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values*.

Sample Variance and Sample Standard Deviation

Sample Variance

Whereas we have presented statistics that describe the central tendencies of a data set, we are also interested in ones that describe the spread or variability of the data values.
A statistic that could be used for this purpose would be one that measures the average value of the squares of the distances between the data values and the sample mean.
This is accomplished by the *sample variance*, which for technical reasons divides the sum of the squares of the differences by $n-1$ rather than n, where n is the size of the data set.

Definition The *sample variance*, call it s^2, of the data set $x_1, x_2, \ldots, x_n$ is defined by

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

An Algebraic Identity The following algebraic identity is often useful for computing the sample variance:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x^2 - n\bar{x}^2$$

The computation of the sample variance can also be eased by noting that if

$$y_i = a + bx_i, i = 1, 2, \ldots, n$$

then$\bar{y} = a + b\bar{x}$, and so

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

That is, if s_y^2 and s_x^2 are the respective sample variance, then

$$s_y^2 = b^2 s_x^2$$

In other words, 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.

Sample Standard Deviation

The positive square root of the sample variance is called the *sample standard deviation*.
Definition The quantity s, is defined by

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)}}$$

is called the *sample standard deviation*.

The sample standard deviation is measured in the same units as the data.

Sample Percentiles and Box Plots

Sample Percentiles

Loosely speaking, the sample 100p percentile of a data set is that value such that $100p$ percent of the data values are less than or equal to it, $0 \leq p \geq 1$.
More formally, we have the following definition.

Definition The *sample* $100p$ *percentile* is that data value such that $100p$ percent of the data are less than or equal to it and $100(1-p)$ percent are greater than or equal to it.
If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these two values.

To determine the sample $100p$ percentile of a data set of size n, we need to determine the data values such that

1. At least np of the values are less than or equal to it
2. At least $n(1-p)$ of the values are greater than or equal to it

To accomplish this, first arrange the data in increasing order.
Then, note that if np is not an integer, then the only data value that satisfies the preceding conditions is the one whose position when the data are ordered from smallest to largest is the smallest integer exceeding np.
On the other hand, if np is an integer, then it is easy to check that both the values in positions np and $np+1$ satisfy the preceding conditions, and so the sample $100p$ percentile is the average of these values.

Sample Quantile

The sample 25 percentile is called the *first quartile*; the sample 50 percentile is called the *sample median* or the *second quartile*; the sample 75 percentile is called the *third quartile*.

Box Plot

A *box plot* is often used to plot some of the summarizing statistics of a data set.
A straight line segment stretching from the smallest to the largest data value is drawn on a horizontal axis; imposed on the line is a "box," which starts at the first and continues to the third quartile, with the value of the second quartile indicated by a vertical line.

The length of the line segment on the box plot, equal to the largest minus the smallest data value, is called the *range* of the data.
Also, the length of the box itself, equal to the third quartile minus the first quartile, is called the *interquartile range*.

Chebyshev's Inequality

Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set.
Assuming that $s > 0$, Chebyshev's inequality states that for any value of $k \geq 1$, greater than $100(1 - frac1k^2)$ percent of the data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$.

Definition

Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, x_2, \ldots, x_n$, where $s > 0$.
Let

$$S_k = \{i, 1 \leq i \leq n : |x_i - \bar{x}| < ks\}$$

and let $N(S_k)$ be the number of elements in the set S_k.
Then, for any $k \geq 1$,

$$\frac{N(S_k)}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2}$$

Because Chebyshev's inequality holds universally, it might be expected for given data that the actual percentage of the data values that lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$ might be quite a bit larger than the bound given by the inequality.

Suppose now that we are interested in the fraction of data values that exceed the sample mean by at least k sample standard deviations, where k is positive.
That is, suppose that $\bar{x}$ and s are the sample mean and the sample standard deviation of the data set $x_1, x_2, \ldots, x_n$.

Then, with

$$N(K) = \text{number of i} : x_i - \bar{x} \geq ks$$

And clearly,

$$\frac{N(k)}{n} \leq \frac{\text{number of i} : |x_i - \bar{x}| \geq ks}{n} \leq \frac{1}{k^2}$$

this is by Chebyshev's Inequality

The One-Sided Chebyshev Inequality

For $k > 0$,

$$\frac{N(k)}{n} \leq \frac{1}{1 + k^2}$$

Normal Data Sets

Many of the large data sets observed in practice have histograms that are similar in shape.
These histograms often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion.
Such data sets are said to be *normal* and their histograms are called *normal histograms*.

If the histogram of a data set is close to being a normal histogram, then we say that the data set is *approximately normal*.

The Empirical Rule

If a data set is approximately normal with sample mean $\bar{x}$ and sample standarad deviation s, then the following statements are ture.

1. Approximately 68 percent of the observations lie within

$$\bar{x} \pm s$$

2. Apporximatedly 95 percent of the observations lie within

$$\bar{x} \pm 2s$$

3. Apporximatedly 99.7 percent of the observations lie within

$$\bar{x} \pm 3s$$

A data set that is obtained by sampling from a population that is itself made up of subpopulations of different types is usually not normal.
Rather, the histogram form such a data set often appears to resemble a combining, or superposition, of normal histograms and thus will often have more than one local peak or hump.
Because the histogram will be higher at these local peaks than at their neighboring values, these peaks are similar to modes.
A data set whose histogram has two local peaks is said to be *bimodal*.

Paired Data Sets and The Sample Correlation Coefficient

We are often concerned with data sets that consist of pairs of values that have some relationship to each other.
If each element in such a data set has an x value and a y value, then we represent the ith data point by the pair (x_i, y_i).

A useful way of portraying a data set of paired values is to plot the data on a two-dimensional graph, with the x-axis representing the x value of the data and the y-axis representing the y value.
Such a plot is called a *scatter diagram*.

A question of interest concerning paired data sets is whether large x values tend to be paired with large y values, and small x values with small y values; if this is not the case, then we might question whether large values of one of the variables tend to be paired with small values of the other.

Suppose that the data set consists of the paired values $(x_i, y_i), i = 1, 2, \ldots, n$.
To obtain a statistic that can be used to measure the association between the individual values of a set of paired data, let $\bar{x}$ and $\bar{y}$ denote the sample means of the x values and the y values, respectively.
For data pair i, consider $x_i - \bar{x}$ the deviation of its x value from the sample mean, and $y_i - \bar{y}$ the deviation of its y value from the sample mean.

When large values of the x variable tend to be associated with large values of the y variable and small values of the x variable tend to be associated with small values of the y variable, then the signs, either positive or negative, of $x_i - \bar{x}$ and $y_i - \bar{y}$ will tend to be the same.

Now, if $x_i - \bar{x}$ and $y_i - \bar{y}$ both have the same sign (either positive or negative), then their product $(x_i - \bar{x})(y_i - \bar{y})$ will be positive.
Thus, it follows that when large x values tend to be associated with large y values and small x values are associated with small y values, then $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ will tend to be a large positive number.

To determine what it means for $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ to be "large," we standardize this sum first by dividing by $n - 1$ and then by dividing by the product of the two sample standard deviations.
The resulting statistic is called the *sample correlation coefficient*.

Definition

Let s_x and s_y denote, respectively, the sample standard deviations of the x values and the y values.
The *sample correlation coefficient*, call it r, of the data pairs $(x_i, y_i), i = 1, 2, \ldots, n$ is defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

When $r > 0$ we say that the sample data pairs are positively correlated, and when $r < 0$ we say that they are negatively correlated.

Properties of r

1. $-1 \leq r \leq 1$
2. If for constants a and b, with $b > 0$,

   $$y_i = a + bx_i, i = 1, 2, \ldots, n$$

   then $r = 1$
3. If for constants a and b, with $b < 0$,

   $$y_i = a + bx_i, i = 1, 2, \ldots, n$$

   then $r = 1$
4. If r is the sample correlation coefficient for the data pairs $x_i, y_i, i = 1, 2, \ldots, n$ then it is also the sample correlation coefficient for the data pairs

   $$a + bx_i, c + dy_i, i = 1, 2, \ldots, n$$

   provided that b and d are both positive or both negative.

We will now prove the first three properties of the sample correlation coefficient r.
That is, we will prove that $|r| \leq 1$ with equality when the data lie on a straight line.

Correlation Measures Association, Not Causation

The explanation for such an association lies with an unexpressed factor that is related to both variables under consideration.