Summary Statistics

Introduction

In the preceding chapter, we studied the frequency distribution of a data set with stem-and-leaf plots and frequency tables. Although those techniques were very useful, they did not allow us to make concise statements about the distribution as a whole. To do this, we need *numerical summary measures* of the data ("summary statistics"). Taken together, such measures provide a great deal of information about a data set.

To illustrate these type of summary measures, let us consider as a simple data set consisting of the following ten age values:

21 42 5 11 30 50 28 27 24 52

In discussing these data, let:

- n represent the sample size (e.g., $n = 10$)
- X represent the *variable* (e.g., AGE)
- x_i represent the *value* of the i^{th} observation (e.g., $x_1 = 21$)

The symbol Σ (capital "sigma") is the summation sign, indicating all values should be added. For the illustrative data set, $\Sigma x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} = 21 + 42 + 5 + 11 + 30 + 50 + 28 + 27 + 24 + 52 = 290$.

Measures of Central Location

Mean

When mentioned without specification, the term *mean* refers to the *arithmetic average* of a data set. Statisticians refer to two different types of means (arithmetic averages). The population mean and the sample mean.

The **population mean** (μ; pronounced "mu") is:

$$m = \frac{\sum x_i}{N} = \frac{1}{N}\sum x_i \qquad \textbf{(3.1)}$$

where Σx represents the sum of all values in the population and N represents the population size. For example, the sum of all values (Σx) for ages listed in Appendix 1 is 17,703 and the population size (N) is 600. Therefore, the population mean age (μ_{age}) = 17,703 / 600 = 29.505.

Although knowledge of the population mean is often valuable, it is often difficult (or impossible) to get information on the entire population. This forces us to study the population mean indirectly, through the sample mean. The **sample mean** ($\bar{x}$; pronounced "x bar") is:

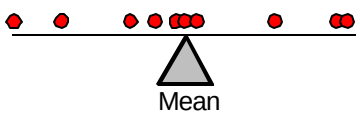
$$\bar{x} = \frac{\sum x_i}{n} = \frac{1}{n}\sum x_i \qquad \textbf{(3.2)}$$

where Σx represents the sum of all values in the sample and n represents the sample size. For the illustrative data set, $\Sigma x_i = 290$ and $n = 10$. Therefore, $\bar{x} = 290 / 10 = 29.0$.

Notice that the operations specified in Formula 3.1 and Formula 3.2 are identical; they both tell you to add all the values and divide by the number of observations. Therefore, whether you are addressing a population mean or sample mean is based on whether the data are *thought* to represent all possible values you are interested in (in which case you are dealing with μ) or only a subset of all possible values of interest (in which case you are dealing with $\bar{x}$). Since we rarely have data on all possible values, we are usually calculating $\bar{x}$, i.e., we *rarely* calculate μ directly.

Interpretation of the mean: Most people intuitively know how to interpret an mean (arithmetic averages). However, there are additional insights you may wish to keep in mind.

First, the mean of a distribution represents its **gravitational center**. That is, the mean is where the distribution would *balance* if placed on a "numerical scale" (figure, right).



Second, the **population mean** is often called the expected value, because if you were to select one observation at random from the population, the population mean would provide a reasonable expectation for that value.

Third, **the sample mean** a good reflection of several different things that you might want to know. It is a good reflection of individual value drawn at random from the sample. It is also a good reflection of an individual value drawn at random from the population. Finally, it is a good estimate of the population mean.

Reporting statistical results: Statistical results should be rounded before they are reported. In general, your final results should be reported to one decimal beyond the initial precision of the data. For example, if age is measured to the nearest year, the mean age should be reported to the nearest tenth of a year (e.g., 29.0 years). To attain one decimal place accuracy for a mean, intermediate calculations should carry at least three decimal places. Also, always indicate the units of measures when reporting statistics. For example, the mean age of the sample is 29.0 *years* (Not just "29.0").

Median

An other type of measure of central location ("average") is the median.

> *The median is the value that is greater than or equal to half of the values in the data set.*

To determine the median, data are ordered from low to high, forming an **ordered array.** The ordered array for the illustrative data is:

```
5   11   21   24   27   28   30   42   50   52
```

The distance from the lowest value in the ordered array to any point in the array is referred to as *depth*.

$$\text{The median has a } \textit{depth} \text{ of } \frac{n+1}{2} \qquad \textbf{(3.3)}$$

For the illustrative example (n = 10), the median has a depth of (10 + 1) / 2 = 5.5. Since this is a non-integer, the median falls between two values. In such instances, the median is the average of the adjacent values. For the illustrative data set, the median is the average of 27 and 28, or 27.5.

```
                                  27.5
5   11   21   24   27  | 28   30   42   50   52
```

When n is odd, the depth of the median will be an integer. For this data set:

```
4    7    8    11    12
```

n = 5 and the median has a depth of (5 + 1) / 2 = 3. Therefore, the median of this second data set is 8.

Mode

The mode, the last type of average we will consider, is *the most frequently occurring value in a data set.* For example, in the data set {4, 7, 7, 7, 8, 8, 9}, the mode is 7, since 7 appears three times in the data set.

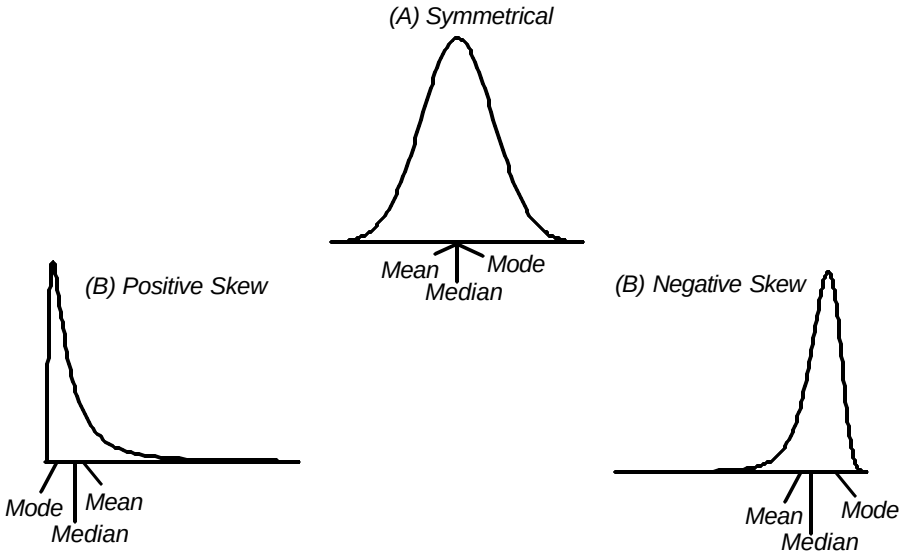
When each value of a data set occurs only once, the data set has no mode. For example, the data set {5, 11, 21, 24, 27, 28, 30, 42, 50, 52} has no mode.

When data sets are small to moderate in size, the mode is rarely used.

> **SPSS:** Means, medians, and modes are computed with the `Analyze | Descriptive Statistics | Explore` command.

Comparison of the Mean, Median, and Mode

The mean, median, and mode are equivalent when the distribution is unimodal and symmetrical. However, with asymmetry, the median is approximately one-third the distance between the mean and mode:



One might then ask which of these statistics is best when asymmetry exists. Although there is no prescriptive formula to answer this question, there are some advantages and disadvantages to each measure of central location.

First, the mean offers the advantages of familiarity and efficiency. It also has a theoretical advantage when making inferences about a population. (This theoretical advantage will be covered in Chapter 5.) However, the mean is is *markedly* influenced by asymmetry and outliers. In such circumstances, it is prone to *misinterpretation*. An often cited example of this is the typical salary of employees, where the salary of highly paid executives skews the average income toward a misleadingly high value. Another example is the average price of homes (in which case high priced homes skew the data in a positive direction). In such circumstances, the median is less likely to be misinterpreted, and is therefore the preferred measure of central location.

A procedure used to diagnose asymmetry is to compare the mean and median of a distribution. When the mean is greater than the median, we have evidence of a positive skew. When the mean is about equal to the median, the distribution is symmetrical. When the mean is less than the median, the distribution has a negative skew:

mean > median — positive skew
mean ≅ median — symmetry
mean < median — negative skew

Quartiles and Other Markers on the Distribution

Quartiles

A **quantile** is any of several ways of dividing the total number of observations into equally sized groups (i.e., each group having the same number of observations). For example, dividing the data up into four equally sized groups results in a type of quantile called **quartiles:** the first quartile marks the bottom quarter of the data, the second quartile marks the middle of the data set (the "second quartile" and median are synonymous), and the third quartile cuts off the top quarter of the data sets.

A general recipe for finding quartiles (according to Tukey's method) is:

(A) Put the data in rank order (i.e., create an ordered array).
(B) Divide the data into two groups by finding its median.
(C) Find the median of the low group. This is the first quartile (Q1)
(D) Find the median of the high group. This is the third quartile (Q3)

For the illustrative data:

```
  5     11     21     24     27    28     30     42     50     52
               |                 |                 |
               Q1                m                 Q3
```

The median is 27.5. The "low group" consisting of {5, 11, 21, 24, 27} has a middle value of 21. This is the **first quartile** (**Q1**). The "high group" is {28, 30, 42, 50, 52}. The middle value of the high group is 42. This is the **third quartile** (**Q3**) of the data set.

Consider this second illustrative example:

```
1.47    2.06    2.36    3.43    3.74    3.78   3.94
              |           |           |
             Q1         median       Q3
```

Here, *n* is odd. The median is 3.43, as marked. When the median represents an actual value (i.e., where *n* is odd), include it in both the low group and high groups when splitting the data. Therefore, the "low group" is {1.47, 2.06, 2.36, 3.43}. The middle of this low group (Q1) is the average of 2.06 and 2.36, or 2.21. The "high group" is: {3.43, 3.74, 3.78, 3.94}. The middle of this high group (Q3) is the average of 3.74 and 3.78, or 3.76.

Percentiles

Another important type of quantile is the percentile. Percentiles divide a data set into 100 equally-sized groups. Therefore, percentiles indicate the percentage of values located below a given value. A good definition for a percentile is, the p^{th} percentile is the value that is greater than or equal to *p* percent of other data points.

Notice that the 25*th* percentile is greater than or equal to 25 percent of the data points. This is synonymous with Q1. The 75*th* percentile is greater than or equal to 75 percent of the data points. This is synonymous with Q3.

Notice that in a large data set, the cumulative relative frequency of a data set is its percentile. For example, a cumulative relative frequency of 95% is greater than or equal to 95% of the data points. It is therefore the 95*th* percentile of the data set. We will <u>not</u> learn how to extrapolate percentiles in small data sets.

Five-Point Summaries

A good picture of a distribution can be achieved by listing its:

- Minimum (Q0)
- First quartile (Q1)
- Median (Q2)
- Second quartile (Q3)
- Maximum (Q4)

This divides the data set up into four equally sized segments and is called a **five-point summary**. The five-point summary for the main illustrative data set in this chapter (data on page 1, calculations throughout the chapter) is 5, 21, 27.5, 42, 52.

Boxplot

The **box-and-whiskers plot** (**"boxplot"**) is a graphical technique that displays a five-point summary and potential outliers. The **box** of a boxplot shows the location of Q1 and Q3. A **line in the box** locates the median. **Whiskers** extend from the top of the box and from the bottom of the box showing high and low values. A procedure for constructing a boxplot is:

(A) Draw a linear axis that covers the range of values. (Use graph paper!)
(B) Next to this axis, draw a box extending from Q1 to Q3.
(C) Inside the box, draw a line that locates the median.
(D) Then, calculate the interquartile range (IQR) as:

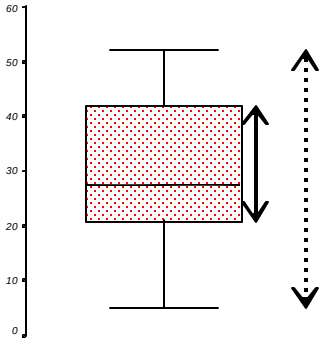
$$IQR = Q3 - Q1 \tag{3.4}$$

From this, calculate the **lower fence and upper fence** as follows:

$$\begin{aligned} \text{Fence}_{\text{lower}} &= Q1 - (1.5)(IQR) \\ \text{Fence}_{\text{upper}} &= Q3 + (1.5)(IQR) \end{aligned} \tag{3.5}$$

(E) Determine if there are any values above the upper fence or below the lower fence. If there are any such values, plot these as separate points on the graph. These are called **outside values**.

(F) If there are no outside values, whiskers are drawn from the upper extent of the box ("upper hinge") to the maximum, and from the lower extent of the box ("bottom hinge") to the minimum. If there are outside values, the lower whisker extends from Q1 to the smallest value still within the lower fence (lower inside value). The upper whisker extends from Q3 to the largest value still within the upper fence (upper inside value).



For the illustrative data, the box extends from Q1 (21) to Q3 (42). A line in the box locates the median at 27.5. The *IQR* = 42 – 21 = 21, so $\text{Fence}_{\text{Upper}}$ = 42 + (1.5)(21) = 73.5. No value is more than 73.5, so there are no outside

values on the top. Therefore, the upper whisker is drawn from Q3 (42) to the maximum (52). $Fence_{Lower} = 21 - (1.5)(21) = -10.5$. No value is less than 10.5, so there are no outside values on the bottom. The lower whisker is drawn from (Q1) 21 to the minimum (5). The final boxplot is shown in the figure to the right.

***Interpretation*:**

- The box contains the middle 50 percent of the data.
- The position of the median and the box itself identify the **center** of the distribution.
- The length of the box is called the "hinge spread" (solid vertical line in the previous figure). This provides a visual representation of the **spread** of the distribution. A less reliable measure of spread is the "whisker spread" (dotted vertical line in the previous figure).
- The **symmetry** of distribution can be judge by the position of the median within the box and box within the whiskers. Also, the presence of outside values toward one side of the box suggests asymmetry.

Judgements about a distribution work best when the sample is large to moderate in size.

Boxplot Illustrative Example 2:

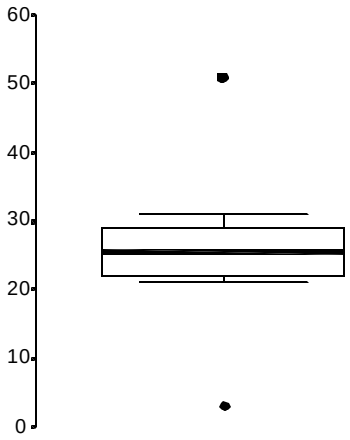
Let us look at an additional data set to illustrate boxplots. Data are:

```
 3    21     22     24     25     26     28     29     31     51
 |           |             |             |             |
Q0          Q1          median          Q3            Q4
```

The five-point summary is: 3, 22, 25.5, 29, 51. The IQR = 29 – 22 = 7.

The $Fence_{Upper} = 29 + (1.5)(7) = 39.5$. Therefore, there is one outside value on the top (51). This outside value is plotted separately. The next highest value, 31, is inside the fence, thus delineating the upper inside value. The upper whisker is drawn from Q3 (29) to this value (31).

The $Fence_{Lower} = 22 - (1.5)(7) = 11.5$. There is one value outside of this fence (3). This point is plotted separately. The next lowest value, 21, is the inside value on the bottom, thus demarcating the lower whisker. The lower whisker extends from Q1 (22) to the lower-inside-value (21). The boxplot is shown in the figure to the right.



SPSS: Boxplots are produced with the `Analyze | Descriptive Statistics | Explore` command.

Measures of Spread

Variance

"Spread" refers to the dispersion of data points around the data set's center. There are several ways to quantify spread, the most common being the variance and standard deviation.

To understand these statistics, it helps to understand what is meant by a deviation. The **deviation** of a data point is its difference from the mean:

$$\text{deviation}_i = x_i - \bar{x} \quad \textbf{(3.6)}$$

For example, the very small data set {1, 3} has a mean of 2 and deviations of (1 – 2) = – 1 and (3 – 2) = +1, respectively.

Although the sum of these deviations may seem like a good basis for a measure of spread, sums of deviations will always equal zero (e.g., for the illustrative example above, the sum of the deviations = – 1 + 1 = 0). Therefore, the sum of the deviations can not be used to measure spread. To get around this problem, statisticians *square* the deviations before summing them. This statistic, known as a **sum of squares (*SS*)**, is:

$$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \textbf{(3.7)}$$

For example, the sum of squares for the data set {1, 3} is $(1 - 2)^2 + (3 - 2)^2 = 1 + 1 = 2$.

The variance of the data can now be calculated. We have two formulas for variances. The **population variance** (σ^2; pronounced "sigma squared") is:

$$s^2 = \frac{SS}{N} \quad \textbf{(3.8)}$$

Notice that his is the mean sum of squares. However, since we rarely have data on the entire population, we usually must calculate the **sample variance,** which is:

$$s^2 = \frac{SS}{(n-1)}$$

For the main illustrative data set in this chapter (i.e., the data set in which $n = 10$ and $\bar{x} = 29.0$) , $SS = (21 - 29)^2 + (42 - 29)^2 + (5 - 29)^2 + (11 - 29)^2 + (30 - 29)^2 + (50 - 29)^2 + (28 - 29)^2 + (27 - 29)^2 + (24 - 29)^2 + (52 - 29)^2 = 2134$, and $s^2 = 2134 / (10 - 1) = 237.1111$.

One problem with the variance is that it carries units *squared*. For example, the variance of the illustrative data is 237.111 **years²**. This makes it difficult to interpret (since we don't think in squared units). To get around this problem, we take the square root of the variance. This is called the standard deviation.

Standard Deviation

The **standard deviation** (*syn*: "root mean square") is simply the square root of the variance. The **population standard deviation** (σ) is the square root of the population variance:

$$= \sqrt{s^2} = \sqrt{\frac{SS}{N}} \qquad \textbf{(3.9)}$$

The **sample standard deviation** (s) is the square root of the sample variance:

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}} \qquad \textbf{(3.10)}$$

The standard deviation of the illustrative example = sqrt(237.1111 years^2) = 15.4 years. By square-rooting the variance, units of "years squared" convert to "years".

Interpretation: Interpreting a standard deviation is not as easy as, say, interpreting a mean. One thing to keep in mind is that big standard deviations are associated with big "spreads" and small standard deviations are associated with small data spreads. For example, if the standard deviation of the age of two population are 15 years and 2 years, respectively, it can be safely said that the first population has much more age variability than the second population.

But how do we interpret a single standard deviation? One way is to indicate the percent of data that is within a specified number of standard deviations of the mean. We have two rules for applying this approach. The first rule applies **when the distribution is normal**.[1] When this is the case, we can say:

- about 68% of the all values will lie within 1 standard deviation from the mean. These boundaries are $\mu \pm \sigma$.
- about 95% of all values will lie within 2 standard deviations from the mean. These boundaries are $\mu \pm 2\sigma$.
- nearly all values will lie within 3 standard deviations from the mean. These boundaries are $\mu \pm 3\sigma$.

For example, *if* we assume that ages of a population are normal distributed (possibly, a bad assumption) with a mean (μ) of 30 and standard deviation (σ) of 10, then 68% of the population will be in the age range 30 ± 10 (20 to 40), 95% will be in the age range 30 ± 20 (10 to 50), and nearly all will be in the age range 30 ± 30 (0 to 60).

For distributions that are not normal, **Chebyshev's rule** applies. Chebyshev's rule states:

- *at least* 75% of the values lie within 2 standard deviations from the mean
- *at least* seven-eighths lie within 3 standard deviations from the mean

For a population with a mean age of 30 years and standard deviation of 10 years, for instance, we know with that *at least* 75% of the values lie in the range 30 ± 20 (10 to 50) and at least seven-eighths lie in the range 30 ± 30 (0 to 60)
.

[1] The normal distribution will be introduced in the next chapter. For now let us note that a normal distribution is bell-shaped and is neither too flat nor too peaked ("mesokurtotic").

The Interquartile Range

The interquartile range (*IQR*) is a measure of spread based on a distribution's quartiles. Recall that quartile one (Q1) is the 25th percentile of a data set. Quartile 3 (Q3) is the 75th percentile of the data set. (Page 3.5 discussed how these statistics are calculated.) The interquartile range is simply the difference between these quartiles:

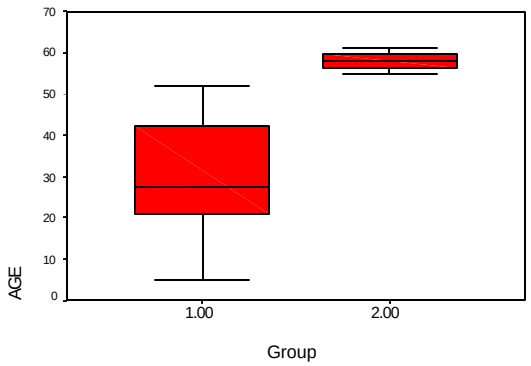
$$IQR = Q3 - Q1 \qquad \textbf{(3.11)}$$

For the illustrative data, Q1 = 21 and Q3 = 42. Therefore, *IQR* = 42 – 21 = 21.

The interquartile range is related to the median, and is relatively unaffected by outliers. It is, therefore, a "robust"[2] measure of spread.

Interpretation: The *IQR* contains the middle 50 percent of data. Groups with large *IQR*s have greater variability than groups with smaller *IQR*s.

A good way to compare *IQR*s is with side-by-side boxplots of groups. For example, in the figure to the right, it is clear that Group 1 has greater variability than Group 2.



SPSS: Although interquartile ranges are reported in output from SPSS's `Analyze | Descriptive Statistics | Explore` command, these IQRs *should be ignored*. (They will differ from your hand calculations.)

[2] The term robust implies that it is relatively resistant to the influence of outliers and distributional asymmetry.