3

Data Description: Measures of Central Tendency and Variability

The previous chapter highlighted the subjective limitation of interpreting graphical presentations of data in general. In addition to this limitation, another important shortcoming stems from the fact that graphs are difficult to use for inferential purposes. In other words, they are not very helpful when one is interested in drawing conclusions about large sets of observations (populations) using data obtained from samples. Unlike these graphs, numerical descriptors in terms of indexes are very useful summaries of data. We primarily employ them with quantitative variables. Two commonly used types of measures in descriptive statistics are the so-called measures of central tendency or location and measures of variability or dispersion, which we discuss in turn throughout this chapter.

3.1. MEASURES OF CENTRAL TENDENCY

3.1.1. The mode

For any qualitative or quantitative variable, the *mode* is the score that occurs the most often in a given data set. We define the *mode* as the score(s) in a given data set (variable) that occurs with the highest frequency. For example, let us consider the following twelve intelligence (IQ) test scores:

95, 97, 100, 101, 103, 101, 102, 105, 101, 95, 97, 101.

The mode of this data set is 101—as can be easily ascertained—since it occurs four times in this set of scores. In this simple example, there is only one mode, but in general their number can be larger, as seen in the next examples.

The mode can be readily obtained with R by examining via the stem-and-leaf plot the frequencies with which scores in a given sample or group are taken (see Chapter 2 as to how to obtain the steam-and-leaf plot). The

Copyright 2013. Rowman & Littlefield Publishers.
All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

EBSCO Publishing : eBook Collection (EBSCOhost) - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD
AN: 507188 ; Tenko Raykov, George A. Marcoulides.; Basic Statistics : An Introduction with R
Account: ns145102.main.eds

score(s) with the highest frequency(-ies) is then the mode of the examined variable in the studied group. To illustrate, let us take a look at the stem-and-leaf plot (Figure 2.7) of the mathematics test–taking anxiety (MTA) scores in Example 2.2 considered in the last chapter. For ease of presentation we repeat the stem-and-leaf plot below as Figure 3.1. As can be readily seen from the graph in Figure 3.1, there are three modes in the MTA example data set. These are the scores 19, 23, and 25. The reason is that each of them occurs five times among the 36 anxiety scores, and all remaining scores occur less frequently.

We note in passing that in a multiple-variable data set, each variable has its own mode that need not be the score of the same person across the variables, nor for that matter the same score(s). For example, suppose that in a study of mathematics ability three tests are administered to high school seniors: (i) an algebra test, (ii) a geometry test, and (iii) a trigonometry test. Then the mode on the algebra test could be 25, the mode on the geometry test 27, and the mode on the trigonometry test 22. The modes can also be scores obtained by different subjects across these three tests.

We emphasize that the mode need not be uniquely defined. This is because there may be more than just a single number that occurs with the highest observed frequency. In fact, it all depends on the data of the variable under consideration, and some data sets may have more than one mode. A way in which this can happen is when we consider several groups of subjects but disregard group membership (i.e., consider them as a single data set), e.g., males and females on a variable of interest. Data sets (variables) with two modes are often referred to as *bi-modal*, and similarly defined in terms of the number of modes are *tri-modal* or *multi-modal* data sets or variables in general. For instance, the MTA data set presented in Figure 3.1 (Example 2.2 from Chapter 2) is tri-modal as we saw above. Also, in a data set where each separate score appears the same number of times, either score can be viewed as a mode.

An especially useful feature of the mode is that it is in general not affected by extreme observations (e.g., even in the problematic cases when these are the results of entry errors). This feature is often referred to as ''resistance'' or ''robustness'' of the mode to extreme values. To illustrate, consider the following example. Suppose that when recording or entering the data in the first

```
1 | 577899999
2 | 1111223333344
2 | 555557778999
3 | 12
```

FIGURE 3.1.
Stem-and-leaf plot of the anxiety scores (Example 2.2).

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

considered example in this chapter with the IQ test scores, a researcher mistakenly entered 995 rather than 95. That is, in his/her version of the above data set, the following were the actually entered (recorded) scores

995, 97, 100, 101, 103, 101, 102, 105, 101, 95, 97, 101.

However, despite this flagrant error, the mode is still 101 as it occurs the most frequently. That is, in this case the mode is unaffected by the data entry/data recording mistake.

At the same time, it is worthwhile noting that it is also possible that an entry or recording error might be made when entering the score that most frequently occurs in the data set, in which case the mode may well be affected. For example, when the second most frequent score appears only one less times than the actual mode in the original data set, then due to such a recording error both these scores may appear equally often. In such a case, two modes will be proclaimed for a data set that actually has only a single mode. Obviously a variety of other examples of similar errors can be easily constructed as well, where the actual mode of a data set is misrepresented due to a data entry or recording error.

A useful feature of the mode is that while it is easily defined for quantitative variables, it is the only measure of "central tendency" that is meaningful for a qualitative variable. (The concept of "central tendency" is not well defined for such variables, and we use it here in a loose sense, viz., in the sense of most "fashionable'" category—i.e., one with the highest observed frequency.) Finally, we note that the mode has the unique feature of being a score that is actually contained in the data set under consideration. That is, the mode is in fact a score taken by/measured on a studied subject (unit of analysis). This need not be the case with the two other measures of central tendency that we turn to next.

3.1.2. The median

The above-mentioned resistance or robustness feature of the mode is also shared by another measure of central tendency, called the median. The *median* can be defined as the middle value of a set of scores under consideration. When the data set has an uneven number of scores, it is the middle value when they are arranged from lowest to highest. In contrast, when the data set contains an even number of scores, then it is the average of the middle two values. We note in passing that this arrangement from lowest to highest is frequently referred to as "rank ordering" of the observed data. (We usually imply an increasing, or ascending, rank ordering when using this term in the remainder of the book, although a reverse ordering can also be meaningful in some cases.)

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

From the above-provided definition a simple, ''manual'' procedure to obtain the median—especially with small data sets—would consist of using the R command 'sort' to rank order a given set of scores, and then work out the median by counting from left or right half as many scores as there are in the data set (or averaging the middle two scores if there is an even number of scores in the data set). For instance, for the earlier mathematics test–taking anxiety example (Chapter 2; see Table 2.3), we can ''manually'' obtain the median of the MTA score—denoted 'y' there—as follows. First we sort the data:

```
> sort(y)
```

This command produces the following output, whereby the prefixes [1] and [23] signal that the first and 23rd consecutive MTA scores follow immediately after them (which prefixes we just ignore most of the time):

```
 [1] 15 17 17 18 19 19 19 19 19 21 21 21 21 22 22 23 23 23 23 23 24 24
[23] 25 25 25 25 25 27 27 27 28 29 29 29 31 32
```

To find the median of this variable, we need to take the middle score, if sample size is an uneven number, or the average of the middle two scores, if sample size is even. In this particular example, we happen to know that the sample size is 36, but at times the sample size may not be known beforehand. In such cases, in order to work out the sample size for a given study or variable, we can use the command 'length'. This command determines the length of the array (row, set) of scores that comprise the available observations on a variable in question. For our MTA example, the command produces the following result (given immediately beneath the command):

```
> length(y)
[1] 36
```

That is, the sample size is 36 here—something we knew beforehand for this MTA example but may not know in another data set of interest.

Since the number of scores on the variable 'y' of interest in this data set is even, in order to work out its median we take the average of the 18th and 19th scores from left (or from right) in their above-sorted sequence. As it happens, both these middle scores are 23, and thus they are each equal to their average, which is declared to be the median value of the MTA variable, viz., 23.

We would like to note here that all of these same computational activities

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

can be accomplished alternatively in an automated fashion with the specific R command 'median':

```
> median(y)
```

This command yields for the mathematics test–taking anxiety (MTA) example the median value of

```
[1] 23
```

which as expected, is identical to the answer we manually obtained above.

Like the mode, the median is ''resistant'' or ''robust'' with regard to (a relatively limited number of) extreme values, such as abnormal values on a given variable. For example, if we have entered the value 117 in lieu of the second number 17 in the MTA data set, the median would still be 23. (As an aside, this can be readily checked out, by first manipulating the data in this way, saving it under a new name, reading it into R as usual, and then evaluating the median on the ''new'' variable 'y'.)

3.1.3. The mean

A measure of central tendency that is very popular, but does not share the above resistance or robustness property with regard to extreme values, is the mean. The mean is defined for a quantitative variable only, as the arithmetic average of the scores under consideration (see further below for a more formal definition). To obtain the mean with R, we use the command 'mean'. To illustrate with the data from the earlier MTA example, the command

```
> mean(y)
```

yields

```
[1] 23.25
```

Since the mean is so widely used, we will spend next some time on its more formal discussion (e.g., Raykov & Marcoulides, 2011, ch. 2). Let us first denote a given variable of interest by the letter y—a notation that will be used quite often in the rest of this book. The mean of y in a population of N subjects—N being typically large and finite, as is usually the case in current behavioral and social research and assumed throughout the rest of this book—is defined as:

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

$$\mu_y = \frac{1}{N}(y_1 + y_2 + \ldots + y_N) = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{3.1}$$

where y_1 through y_N denote the values of the variable y for the members of the population, beginning with the first and up to the Nth member, and μ_y designates the mean of y in the population of interest. (We may eventually dispense with using the sub-index $(_y)$ attached to the symbol μ for population mean values in later discussions, when no confusion may arise.) We note also the presence of the summation symbol, Σ, in Equation (3.1). This symbol is utilized to denote the process of adding together all y scores with sub-indexes ranging from $i = 1$ to $i = N$, i.e., the sum $y_1 + y_2 + \ldots + y_N$. We will also frequently use this short summation index, Σ (with appropriate ranges of associated sub-indexes), in the rest of the book.

We rarely have access, however, to an entire population of interest. Rather, we typically only have available a sample from the population of interest. How can we then use this sample to extract information about the population mean, i.e., obtain a good "guess" of the population mean? To this end, we wish to combine in an appropriate way the studied variable values obtained from the sample, so as to render such a good "guess" of the population mean. This process of combining appropriately the sample values to furnish information about an unknown quantity, like the mean, is called in statistical terminology *estimation*. Unknown population quantities, like the mean (or median), which characterize a population distribution on a variable of interest, are called *parameters*. We wish to estimate these parameters using data obtained from the sample.

For any given parameter, we accomplish this estimation process utilizing an appropriate combination, or function, of the sample values or scores on the studied variable. This function is typically referred to as *statistic*. That is, we estimate unknown parameters using statistics. A major property of "good" statistics is that they appropriately combine sample values (observations) in order to extract as much as possible information about the values of unknown parameters in a population(s) under investigation. When using statistics to estimate parameters, the statistics are often referred to as *estimators* of these parameters. That is, in a sense an estimator is a statistic, or a formula, that is generally applicable for the purpose of estimating a given parameter. In a given sample, the value that the statistic takes represents the *estimate* of this specific parameter.

Returning to the mean considered earlier in this section, as we indicated before it is estimated by the arithmetic average of the scores on y obtained in the studied sample, i.e., by

$$\hat{\mu}_y = \bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{3.2}$$

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

where a caret (^) symbolizes estimator of the parameter underneath—a practice that is also followed throughout the rest of this book. In Equation (3.2), y_1 through y_n now symbolize the scores in the available sample, and n denotes as usual its size (i.e., the number of subjects (units of analysis) in the sample). We stress that Equation (3.2) represents the mean estimator, often denoted $\bar{y}$. The value obtained when the sample scores on the variable y are entered into the right-hand of this equation, is the mean estimate for that sample, $\bar{y}$. That is, the value obtained in $\bar{y}$ is an estimate furnished using a statistic (estimator) to estimate an unknown population parameter in a given sample, the mean. It is these estimates of parameters that are of key interest and are usually reported in most empirical research.

An important limitation of the mean as indicated above is the fact that it is sensitive (i.e., not resistant or not robust) to abnormal values, such as excessively high or low values on a studied variable. In particular, even a single data entry error can change substantially the mean. For example, if we incorrectly entered the value 117 instead of the accurate value of 17 in the MTA data set in Table 2.2, the resulting mean of 25.85 would be quite different from the mean of the original data/variable that was found above to be 23.25.

For this reason, a variant of the mean has also been developed, the so-called trimmed mean. One may be interested in the 5%-trimmed mean, which is the mean of the "middle" 90% of the scores in a given data set. To obtain this trimmed mean, after rank ordering the scores, the top and bottom 5% of the scores are first deleted. The average of the remaining scores in the now reduced data set is then the so-called 5%-trimmed mean. (We note that a higher percentage could also be dropped from either end if needed, e.g. 10%, if there are reasons to consider it; this would lead to the so-called 10%-trimmed mean.) The 5%-trimmed mean is readily obtained with R using the subcommand 'trim' of the command 'mean' as follows (assuming the variable denoted 'y' is of interest):

```
> mean(y, trim=5)
```

Specifically for the data set in Example 2.2 (see the MTA example in Chapter 2), the 5%-trimmed mean is obtained with this R command as

```
[1] 23
```

which is quite similar to the untrimmed mean of 23.25 we found earlier in this subsection. This similarity would actually be expected, given that none of the 36 scores on MTA in this example appears "abnormal" (i.e., "sticks out" from the rest in terms of its magnitude—see, e.g., Raykov & Marcoulides, 2008, ch. 3, for a more detailed and nontechnical discussion of the concept

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

of an ''outlier,'' often used synonymously to represent an unusual, aberrant, abnormal, or extreme value, score, or observation that could be the result of either gross data entry errors or perhaps originating from a person or unit that is not a member of the population of interest).

While the measures of central tendency discussed in this section are very useful for providing summary information pertaining to the central location of a variable under study (e.g., the MTA score in the considered example), none of them contains information about any potential individual differences that might be present within the data. We address this issue next.

3.2. MEASURES OF VARIABILITY

Individual differences on studied variables are frequently of special interest in empirical research. In fact, in many areas scientists are particularly interested in explaining individual differences on certain variables—usually referred to as dependent variables, response variables, or outcome variables—in terms of differences on other variables, typically called independent variables, ''predictors,'' explanatory variables, or covariates. For instance, do individual differences in parental motivational practices or upbringing style account for possible observed differences in mathematics achievement in elementary and secondary school? Similarly, do urban/suburban differences account for disparities in mathematics achievement? Or perhaps it is gender, SES, or school sector differences that can explain potential achievement differences?

In order to evaluate individual differences, however, some new measures called *measures of variability or dispersion* are needed. These are necessitated by the observation mentioned above that the measures of central tendency do not address the stated concerns. Rather, the needed new measures of variability should respond to the following main question: ''To what extent is the mean on a variable informative, in the sense of being 'representative' of the scores that this variable takes in a group of studied individuals (or units of analysis)?'' For distributions with a wide spread around the mean, the latter is obviously far less informative than in distributions for which the majority of scores are tightly clustered around the mean. We stress that none of the central tendency measures contain information that bears upon the answer to this question about individual differences, since all these measures are primarily concerned with the location of the scores rather than with their differences.

As it turns out, the last question is commonly answered using the important concepts of variance or standard deviation of a given variable. These two measures, like the mean, are typically defined for quantitative variables. To introduce them, we look first at individual scores and how close they are to the mean. Specifically, the degree to which the mean is representative of most

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

scores in a population on a variable under consideration (i.e., the extent to which the values on this variable are dispersed around its mean there) is captured at the individual level by the *deviation scores*. For a given population of size N, let us denote the scores on the studied variable y as y_1, y_2,..., y_N. Then the deviation scores are defined as follows:

$$u_i = y_i - \mu_y \ (i = 1, 2, \ldots, N). \tag{3.3}$$

In a given sample, these individual deviations are obtained by subtracting the average of the scores in the sample from each score (since the average is the estimate of the mean in the available sample). Denoting the n sample values as $y_1, y_2, \ldots y_n$, the deviation scores in the sample are $u_i = y_i - \bar{y}$ $(i = 1, \ldots, n)$; as mentioned before, usually n is much smaller than N). Hence, to furnish the individual deviation scores with R, after rendering the mean estimate we simply subtract it from each observed score. For our earlier utilized MTA example (Example 2.2 in Chapter 2), we obtain them as follows:

```
> u = y - mean(y)
```

To see the result of this action, we ask R to print to the screen the elements of the newly obtained vector u, by simply stating its symbol at the R prompt:

```
> u
```

This prints to the computer screen the individual deviation scores as follows (the prefixes "[13]" and "[25]" signal that the 13th and 25th individual deviation score follows immediately after them, in this listing, and are ignored as usual):

```
 [1] -8.25 -6.25 -4.25 -2.25 -0.25  8.75 -5.25 -4.25 -1.25 -0.25  1.75  -2.25
[13] -1.25  1.75  1.75  0.75 -0.25  3.75 -4.25 -6.25 -2.25  5.75  3.75   7.75
[25] -4.25 -4.25 -0.25  1.75  1.75  3.75  5.75  5.75 -2.25  0.75 -0.25   4.75
```

While the individual deviation scores represent the degree to which individuals (individual units of analysis) deviate from the mean, they have the property that they always sum up to zero, no matter how large any one of them is. In other words, the sum of the deviations of individual scores around the mean will always be equal to zero. Indeed, if we sum them up using R (for which we can use the command 'sum' as indicated in the preceding chapter), we readily observe that the result is zero:

```
> sum(u)
```

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

which returns as expected

```
[1] 0
```

Thus, any data set—no matter how many scores it consists of or how different the individual scores are from their mean—has the same overall sum of the individual deviation scores, viz., zero. Hence, it would seem that this information does not appear to help much in differentiating between different sets of scores. In addition, the individual mean deviations are scores that are characteristic for each person studied. Therefore, when one simply examines these deviations, no data reduction (summarization) is actually being achieved. Yet this type of reduction or summarization is what is frequently sought when using statistics in most empirical research.

For these reasons, we need a summary measure of individual differences, which measure does not share the limitations just mentioned. At the population level, such a measure is the *variance* of the studied variable, which is defined as follows (later in the book, for ease of presentation we may dispense with the subindex 'y' to σ^2, when no confusion may arise):

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N} u_i^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2. \tag{3.4}$$

Based on this equation, it should be evident that the variance is the average squared mean deviation. (We assume throughout the rest of the book that the variance, as well as the mean, of any variable considered is finite, which is a fairly mild if at all restrictive assumption in empirical research.) We also observe from this definition in Equation (3.4) that the variance has as units of measurement the *squared* units underlying the individual scores (measurements). Because of these squared units, the variance is somewhat difficult to directly interpret in an empirical setting. To avoid this problem of interpreting squared units, the standard deviation is also considered, which is defined as the square root of the variance:

$$\sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N} u_i^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{3.5}$$

where a positive square root is taken.

Equations (3.4) and (3.5) allow us to determine the variance and standard deviation of a studied random variable y, if an entire (finite) population of concern were available. As already mentioned, this will rarely be the case in empirical social and behavioral research that typically works, due to a number of reasons, with samples from populations of interest. As a consequence, the variance and standard deviation for a variable of concern are estimated in an available sample correspondingly by using the following equations:

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

$$\hat{\sigma}_y^2 = s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.6}$$

and

$$\hat{\sigma}_y = s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.7}$$

with a positive square root taken in the last equation. We emphasize that we divide in Equation (3.6) the sum of squared mean deviations by $(n - 1)$ rather than by n. We do this in order to obtain an ''unbiased'' estimate/estimator of variance. This means that the resulting estimate on average equals the population variance, across possible samples taken from the population (all of them being with the same size, n). This unbiasedness feature is a desirable property for any estimator of any parameter. Alternatively, if we divide just by n in Equation (3.6), then on average—that is, across repeated sampling from the same population—the variance estimate will underestimate the true variance (i.e., will be associated with a negative bias).

The variance and standard deviation sample estimates are also readily obtained with R using the commands

```
> var(y)
```

and

```
> sd(y)
```

respectively. We note in passing that the variance and standard deviation of the original data in y are the same as those of the individual deviation scores u. (While this is a generally valid result (e.g., Agresti & Finlay, 2009), one can readily see it on the above example data using R, by applying instead the last two 'var' and 'sd' commands on the set of scores u representing these individual deviations.)

Although the variance and standard deviation of a given variable are quite useful indexes, as we will see later in this book, they have the limitation that they are not immediately interpretable in a substantive domain in terms of individual variation. For instance, what does it actually mean to say that an IQ score measured on second graders has a variance of nine or a standard deviation of three? We will see below that under some additional assumptions these numbers do actually attain important meaning. However, if the assumptions are not correct, these numbers tell us little about the degree to which individual scores are indeed scattered around the mean on a studied variable.

A measure of variability that can be immediately interpreted is the *range*. It

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

is defined as the difference between the largest and smallest scores in a given data set (on a studied variable). Denoting by y this variable, the range r is defined as

$$r = \max(y) - \min(y), \tag{3.8}$$

where max(.) and min(.) are the largest and smallest score on y, respectively.

With R, we obtain the range by simply using its definitional equation:

```
> r = max(y) - min(y)
```

The result of this activity is that the range of the variable y is stored or created as the object r. As before, in order to see its contents—e.g., to see what the range of the MTA score is—we need to state next the symbol of the range at the R prompt:

```
> r
```

which yields

```
[1] 17
```

That is, the distance between the highest and lowest mathematics test–taking anxiety score in the example studied sample of 36 students is equal to 17. Alternatively, we can also use the command 'range', to obtain the smallest and largest number in the data set:

```
> range(y)
```

In the above MTA example (Example 2.2 in Chapter 2), this command returns

```
[1] 15 32
```

From this result, the range is obviously determined to be $r = 32 - 15 = 17$, although we would need to manually obtain it in this case.

An interesting relation holds between the standard deviation s_y and the range on a given variable (y) (e.g., Ott & Longnecker, 2010). Accordingly, its range is generally expected to be about four times larger than its standard deviation:

$$r \approx 4s_y. \tag{3.9}$$

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

We stress, however, that this is only an approximate relationship, and it is presented here merely to provide a rough connection between the two indexes involved.

An important limitation of the range is the fact that it is obviously affected by "outliers," i.e., unusually large/small observations—e.g., data entry errors (e.g., Raykov & Marcoulides, 2008, ch. 3). As such, the range may give misleading indications of what may be a very large dispersion of scores around their mean, which is however spurious due to the presence of extreme scores in a given data set.

A measure of variability that does not suffer from this drawback (at least not to the same degree in general) is the *inter-quartile range* (IQR). The IQR is defined as the interval that contains the middle 50% of the scores on a given variable. That is, the IQR represents the distance between the median of all scores that are positioned below the median of the variable being studied, on the one hand, and the median of all scores positioned above the median on that variable. With R, we obtain the IQR using the same-named command, 'IQR' (note the capitals). For our earlier MTA example, we thus obtain (with the result again given beneath the R command used):

```
> IQR(y)
[1] 5
```

That is, the middle half (in the rank ordering of) the scores of the $n = 36$ students examined in this anxiety study differ from each other by up to five points.

Unfortunately, any of the summary indexes discussed in this section has the limitation that sets of scores with very different mean values could still have the same variability measures. In order to relate variability to the position of the actual range within which scores vary from one another, the *coefficient of variation* (CV) can be used. The CV index provides by definition the variability per unit mean and is determined as

$$c_y = \sigma_y \,/\, \mu_y \tag{3.10}$$

for a given variable y in a population of interest. In an available considered sample, the CV can obviously be estimated as

$$\hat{c}_y = \hat{\sigma}_y \,/\, \hat{\mu}_y = s_y \,/\, \bar{y}\ . \tag{3.11}$$

For our earlier MTA example, this coefficient is obtained with R as follows:

```
> c = sd(y)/mean(y)
```

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

which yields

```
[1] 1.789
```

The CV becomes quite relevant when one considers different populations that may have similar variability but different mean values. Under such circumstances, their CV indexes will also differ, as a reflection of the population differences.

3.3. THE BOXPLOT

We have discussed so far in this chapter a number of measures for determining both central tendency and variability in studied data sets. With this multitude of indexes, a natural question that arises is whether there may be a way of integrating them into a single "picture" of data under consideration. As it turns out, this goal is readily accomplished with the so-called boxplot.

The *boxplot* is a very popular graphical device in applications of descriptive statistics, which integrates a number of measures of central tendency and variability. Indeed, with just a single glance at a boxplot, one can obtain a fairly large amount of information about the set of scores considered. To define the boxplot, we need first to attend to the concept of quartile.

3.3.1. Quartiles

The term *quartile* is commonly used to describe the division of observations into defined intervals based on the values of the data. (We note in passing that a quartile can also be thought of as a particular quantile; see below.) There are two main quartiles for a given variable, a lower and an upper quartile. The *lower quartile* cuts out at the left (i.e., the lowest) 25% of the scores on the distribution, i.e., the lowest quarter of the distribution. Conversely, the *upper quartile* cuts out at the right (i.e., the highest) 25% of the scores, i.e., the upper quarter of the distribution. That is, the upper quartile can be thought of as being the median of all scores that are positioned above the median of the original data set; similarly, the lower quartile would be the median of the scores below the median of the original data. In other words, the earlier discussed inter-quartile range (IQR) is just the difference between these two quartiles, which thus enclose the middle 50% of the scores on a given variable.

The lower and upper quartiles are readily obtained with R using the command 'quantile'. To obtain the lower quartile, we use

```
> quantile(y, 1/4)
```

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

while we can determine the upper quartile with

```
> quantile(y, 3/4).
```

We stress the use of the numbers 1/4 and 3/4 at the end of these commands (and after a comma separating them from the name of the variable in question), since 1/4 and 3/4 of all scores correspond to the left of the lower and to the left of the upper quartile, respectively.

To illustrate, for our earlier MTA example (Example 2.2 of Chapter 2), we obtain the following values of these quartiles (with the results again given beneath the command used):

```
> quantile(y, 1/4)
 25%
20.5

> quantile(y, 3/4)
 75%
25.5
```

That is, the smallest 25% of the anxiety scores in this example are no higher than 20 (since they are all whole numbers, as seen from Table 2.2 in Chapter 2), while the highest 25% of the anxiety scores are at least 26.

These two quartiles can be readily obtained for a quantitative variable in a given data set, along with its mean, median, maximal, and minimal value, using alternatively the command 'summary'. This R command produces the *six-point summary of the variable*. For instance, considering again the MTA example, we obtain the following results (given beneath the used R command):

```
> summary(y)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
15.00   20.50   23.00   23.25   25.50   32.00
```

As before, we observe from this output the minimum and maximum values being 15 and 32 respectively (see earlier discussion of the 'range' command in this chapter). We also notice the values of the mean and median being correspondingly 23.25 and 23, and finally the lower and upper quartiles as 20.5 and 25.5, respectively. With these features, the command 'summary' provides a quick and convenient summarization of the data on a quantitative variable.

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

3.3.2. Definition of the boxplot and its empirical construction

Returning to our discussion of the graphical device of a boxplot for a studied variable, the IQR is represented by a "box" in it. The two horizontal ends (lower and upper) of this box, also called *hinges*, represent the lower and upper quartiles. To these hinges, two *whiskers* are attached. The whiskers extend from the lower and upper hinges to the most extreme observation that is still within 1.5 × IQR units away from the nearest hinge ('×' denoting multiplication). The observations further away from the median are presented by separate points and can be viewed informally as extreme scores (possible outliers; cf. Raykov & Marcoulides, 2008, ch. 3).

We obtain the boxplot with R using the command 'boxplot':

```
> boxplot(y, main = "Boxplot of Anxiety Scores", ylab = "Anxiety  Score")
```

where we now add in the subcommand 'ylab' a title for the vertical axis—the one of anxiety scores. This command yields Figure 3.2 for the earlier MTA example.

In this displayed boxplot, the thick horizontal line in its middle represents the median of the anxiety scores of interest, which is equal to 23. As mentioned before, the two thinner lines (hinges) enclose the box at the upper and lower quartiles of 20.5 and 25.5, respectively. That is, the distance of the latter two statistics is the IQR, viz., 5 in this example. In other words, this box encloses the "middle" half of the scores, as one moves along the vertical axis

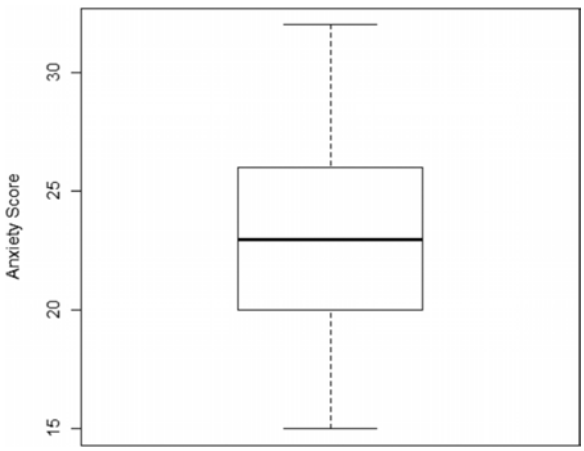

FIGURE 3.2.
Boxplot of anxiety scores.

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

representing the range of the MTA scores under consideration. In the boxplot, the whiskers are the dashed lines that stretch away from the hinges and until the smallest and largest value (15 and 32, respectively) in the sample, which fall within 1.5 IQR units from the closest box end. As mentioned before, when a data set contains possible "outliers" (i.e., extremely small or large values), they are visually located outside of the whiskers. We do not have such "outliers" in the present data set, since none of its scores extends further than 1.5 $\times$ IQR $=$ 7.5 points from the lower and upper quartiles (20.5 and 25.5, respectively).

There are a number of important pieces of information that can be extracted by examining a boxplot for a variable under consideration. First, the median can be easily located by examining where the thick horizontal line is located within the box—the height of this line is the median. (See the vertical axis in Figure 3.2 and the score on it pertaining to the line; that score is the median, 23, in the presently considered example.) Second, the length of the box (i.e., the IQR) is a measure of variability, as discussed earlier in this chapter. This measure is visually represented by the distance between the lower and upper hinges of the box. In case the median is positioned in the center of the box and the two whiskers are of the same length, with no points further away from the lower and upper ends of the whiskers, the distribution of the variable is fairly symmetric. If the median is not positioned in the center of the box, however, there is evidence of some asymmetry in this distribution. In that case, the longer tail of the distribution is to be found in the direction of the longer whisker (and further extreme observations if any). This asymmetry usually is opposite to the direction in which one finds the hinge that is closer to the median. Points outside of the ends of the whiskers are also indicative of possible outliers as mentioned before.

When a distribution is not symmetric, it is called *skewed*. In a skewed distribution, the scores above (below) the median are spread more and further away from it, than the scores below (above) the median. More specifically, there are two types of skewed distributions—positively and negatively skewed distributions. A positively skewed distribution has the median usually positioned lower than the center of the box and closer to the lower whisker. In such a distribution, the scores above the median are spread more and further away from it, than the scores below the median. That is, the right tail of the distribution is longer. Conversely, a negatively skewed distribution has the median usually positioned closer to the upper hinge, and the lower whisker being longer than the upper whisker. In such a distribution, the scores below the median are spread more and further away from it, than the scores above the median; that is, the left tail of the distribution is longer. We will discuss further the notion of asymmetry and quantify it in a later chapter of the book.

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

3.3.3. Boxplots and comparison of groups of scores

Boxplots are also very useful when comparing informally several groups of scores. For instance, an educational researcher may be interested in comparing college aspiration scores for male and female applicants. We illustrate the details in the following example.

Example 3.1 (college aspiration in high school students): Let us consider a study in which 40 high school sophomores (boys and girls) were administered a scale evaluating their levels of college aspiration. Using a boxplot graphical device, we are interested in comparing the two sets of obtained scores for boys and for girls. The data set CH3_EX31.dat contains their data (whereby the following designations are used in the file: id = identifier, y = college aspiration score, g = gender—0 for boys, 1 for girls) and is presented in Table 3.1.

To examine simultaneously the boxplots of both genders, after reading in the entire data set (see Chapter 2), we use with R the following command:

```
> boxplot(y~g, main = "Boxplot of College Aspiration Scores by  Gender",
ylab = "College Aspiration Score", xlab = "Gender: Boys = 0, Girls  = 1")
```

We note the use of the symbol '∼', which effectively requests comparison of the variable 'y' within the values of the variable 'g', i.e., for each of the genders here. Furthermore, we have also added the 'xlab' subcommand to provide a title for the horizontal axis. The last stated R command yields the graphical representation provided in Figure 3.3.

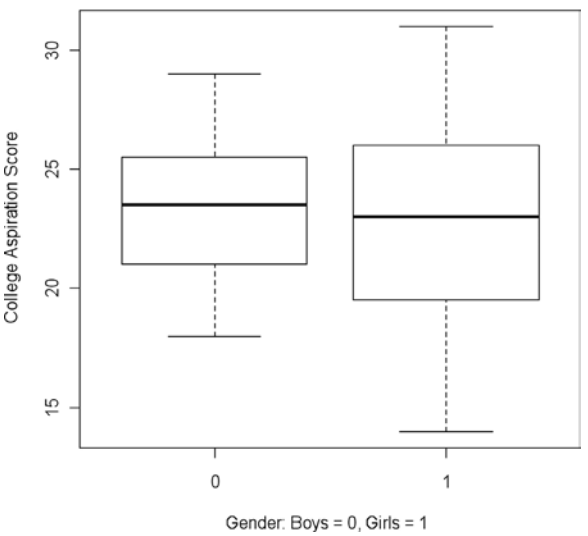

FIGURE 3.3.
Simultaneous boxplots for boys and girls on a college aspiration measure.

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

Table 3.1 Data from a college aspiration study ($n = 40$).

id	y	g
1	14	1
2	18	1
3	17	1
4	22	0
5	23	1
6	31	1
7	18	0
8	19	0
9	22	1
10	23	1
11	25	0
12	21	0
13	22	0
14	25	0
15	25	0
16	24	0
17	23	1
18	26	0
19	19	1
20	17	1
21	21	1
22	29	0
23	27	0
24	31	1
25	20	1
26	19	0
27	23	0
28	25	1
29	25	0
30	27	1
31	28	0
32	29	1
33	21	0
34	24	1
35	23	1
36	28	0
37	21	0
38	24	1
39	23	0
40	28	1

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use

As can be seen by examining Figure 3.3, boys have a slightly higher median than girls on the college aspiration scale; they also have less pronounced inter-individual differences. Additional group comparisons can also be readily made using this data set. For example, in this group comparison context, a question of actual interest could be whether these observed group differences are "real," i.e., if they exist also in the two examined populations (of boys and of girls) from which these two samples came. After all, the research question was concerned with population differences to begin with—whether such existed—rather than with differences in randomly drawn samples from them. Answering this important question will be the concern of later chapters of the book. Before we turn our attention to them, however, we need to discuss in the next chapter a concept of fundamental relevance for statistics and its applications, that of probability and its related notions.

EBSCOhost - printed on 9/5/2024 7:36 PM via UNIVERSIDAD NACIONAL ABIERTA Y A DISTANCIA - UNAD. All use subject to https://www.ebsco.com/terms-of-use