APPENDIX N

SUMMATION NOTATION AND ARITHMETIC OPERATIONS

Much knowledge of statistics is best communicated by means of equations or formula. These equations represent computations that need to be performed to produce the statistics and, in many cases, the conceptual meaning of the statistics as well. Appendix A presents the basic concepts needed to understand and follow the basic statistical formula used in this book.

Summation Notation

Summation notation uses letters to represent variables. The specific letters are arbitrary, although we will use y in general to represent the criterion variable and x to represent the predictor variable. For the examples in this section, let y represent the number of errors on some experimental task for each of four participants who obtained the following scores: 2, 3, 5, 6.

The letter y represents the variable in general. Sometimes we wish to identify specific observations. A subscript (i.e., a number) can be assigned to y to indicate specific observations. For example, y_1 = first observation, y_2 = second observation, and so on. For our four observations, we would have: $y_1 = 2$, $y_2 = 3$, $y_3 = 5$, $y_4 = 6$. One analogy that may help to understand this notation is a street address. There is a street (the letter representing the variable, y), a street number (the subscripts for the letter: 1, 2, 3, 4), and the inhabitants of those addresses (the observed values: 2, 3, 5, and 6). Students with some familiarity with computers or other areas of mathematics will recognize that y names an array, which consists of locations and values.

For reasons that will be clear shortly, we also want to have a general letter to represent the subscript values (i.e., the street numbers). We will use i to represent the subscript for the ith observation; therefore, y_i stands for the ith observation. Our general notation for any score is y_i, where $y_i = y_1, y_2, y_3, \ldots$ (the ... are called ellipses and indicate continuation of a series). The subscript i assumes sequential values to indicate each observation (i.e., i = 1, 2, 3, ...).

It is also desirable to have a general letter to represent the number of observations. We will use *n* to represent the number of observations; hence, y_n = the last observation. Our general notation to this point can be summarized as $y_i = y_1, y_2, \ldots, y_n$, for i = 1, 2, ..., n.

Statistical analysis involves some basic calculations or operations on the set of numbers

represented by y_i. For example, some statistics require the calculation of the sum of the numbers. We could represent the sum as: $y_1 + y_2 + ... + y_n$, but this would quickly get very awkward as most formula involve several different sums combined together in various ways.

Instead we will use the Greek uppercase sigma, Σ, as a symbol for summation or addition. In this notation, $\Sigma y_i = y_1 + y_2 + ... + y_n$. Note that the full summation notation is slightly more elaborate than is presented here. The full notation would include values below and above sigma to indicate the range of subscripts to sum over. We will almost invariably be summing all of the observations (i.e., from $i = 1$ to n), so this additional information is unnecessary at the present time.

To apply this notation to our small set of numbers: $\Sigma y_i = y_1 + y_2 + y_3 + y_4 = 2 + 3 + 5 + 6 = 16$. The elegant thing about summation notation is that no matter how many numbers there are to add, the notation remains simply Σy_i.

When working with more than one variable, there will be additional letters or subscripts to represent the other variables. To expand our example, let $x_i = 3, 4, 4, 5$, and let letters that share the same subscript indicate paired observations (i.e., the first participant obtained a score of 3 on x and 2 on y, the second participant obtained a score of 4 on x and 3 on y, and so on).

Arithmetic Operations

Because summation is often paired with other arithmetic operations, it is important to perform the different operations in the proper order. The standard order of operations is: (1) operations within brackets, (2) powers, (3) multiplication and division (x, / or ÷), and (4) addition and subtraction (+, -). This is an important issue because sometimes the order of operations does not matter and at other times it is critical for obtaining the correct result.

The first example involves multiple additions, one specified within brackets and the other indicated by summation. Because summation and addition are at the same level (indeed they are the same operation), it does not make any difference whether one first sums the pairs of observations and then adds the sums, or sums the variables and adds the resulting sums.

$$\Sigma(x_i + y_i) = (3 + 2) + (4 + 3) + (4 + 5) + (5 + 6) = 5 + 7 + 9 + 11 = 32$$

$$\Sigma(x_i + y_i) = \Sigma x_i + \Sigma y_i = (3 + 4 + 4 + 5) + (2 + 3 + 5 + 6) = 16 + 16 = 32$$

We can show algebraically that these are equivalent by simply rearranging the terms in

the expanded summation to put the xs and ys together:

$$\Sigma(x_i + y_i) = x_1 + y_1 + x_2 + y_2 + ... + x_n + y_n$$
$$= (x_1 + x_2 + ... + x_n) + (y_1 + y_2 + ... + y_n) = \Sigma x_i + \Sigma y_i$$

Such equalities do not necessarily occur when operations of different orders are involved. Consider the following calculation:

$$\Sigma x_i y_i = x_1 y_1 + x_2 y_2 + ... + x_n y_n$$

This calculation involves multiplication and addition. The multiplication must be done first; that is, $\Sigma x_i y_i = \Sigma(x_i y_i)$. For our sample data:

$$\Sigma x_i y_i = 3x2 + 4x3 + 4x5 + 5x6 = 6 + 12 + 20 + 30 = 68$$

We should not confuse $\Sigma x_i y_i$ with $\Sigma x_i \Sigma y_i$, which instructs us to multiply together the results of the separate summations of x and y. That is, $\Sigma x_i \Sigma y_i = (x_1 + x_2 + ... + x_n) \times (y_1 + y_2 + ... + y_n) = (\Sigma x_i)(\Sigma y_i)$. This gives a very different result than $\Sigma x_i y_i$.

$$\Sigma x_i \Sigma y_i = (3 + 4 + 4 + 5) \times (2 + 3 + 5 + 6) = 16 \times 16 = 256$$

One thing that you should pay special attention to as we review basic statistical operations is that operations that appear (wrongly) to indicate the same quantity may not be equivalent, as above. A second is that operations that appear (again wrongly) to be different may actually represent the same quantity, as shown shortly.

Statistics often involve the calculation of some quantity, such as the mean *M*, and then the use of that quantity in other formula. Quantities that do not change are called constants, which we can represent in general by the letter k. Constants demonstrate some interesting properties when included in summations. In particular, their sum can be calculated by or represented as multiplication of the constant by the number of items being summed. That is, $\Sigma k = k + k + ... + k = nk$. Consider the constant 4.0 for our hypothetical sample of four observations. (Note that here is one example where it would be helpful to have the full summation notation indicating that the sum is from i = 1 to 4).

$$\Sigma 4.0 = 4.0 + 4.0 + 4.0 + 4.0 = 16 = 4 \times 4.0$$

This allows us to determine certain quantities in alternative ways. Consider the sum of observed variable values minus a constant k:

$$\Sigma(y_i - k) = y_1 - k + y_2 - k + ... + y_n - k$$

$$= (y_1 + y_2 + \ldots y_n) + (k + k + \ldots + k) = \Sigma y_i - nk$$

Using the sample x data and a constant of 4.0, we obtain:

$$\Sigma(x_i - 4.0) = (3 - 4.0) + (4 - 4.0) + (4 - 4.0) + (5 - 4.0) = -1 + 0 + 0 + 1 = 0$$

or

$$\Sigma x_i - nk = (3 + 4 + 4 + 5) - 4 \times 4.0 = 16 - 16 = 0$$

Let us modify this formula slightly to illustrate the importance of the order of operations principles. If we square the differences (or deviations) between x_i and 4.0, then the sum of the squared deviations is not 0 because squaring the deviations before summing them eliminates the negative value(s) and makes the sum of the squared deviations positive:

$$\Sigma(x_i - 4.0)^2 = -1^2 + 0^2 + 0^2 + 1^2 = 1 + 0 + 0 + 1 = 2$$

The operation of summing and then squaring would require additional brackets to give the summation priority over the squaring operation. That is, $[\Sigma(x_i - 4.0)]^2 = 0 \times 0 = 0$. When multiple brackets are used in formula, it is important to attend carefully to the order of operations. To simplify interpretation of such formula, it is common to use different styles of brackets (e.g., the parentheses and square brackets above).

Definitional and Calculational Formula

Another point to note here briefly is that there may be several different variants of formula for calculating the same statistical quantity. Often one of the equations communicates more clearly the meaning of the statistic; such formula are called definitional or conceptual formula. But such equations may not be the simplest to use when doing calculations. In this case it may be better to use computational or calculation version of the formula, although now the meaning of the statistic will be less apparent. Consider the following example discussed more fully in the text.

The sum of the squared deviations of the ys about the mean of y (SS_y for short) can be calculated in the following manner, where $M_y = \Sigma y_i/n = (2+3+5+6)/4 = 4.0$:

$$SS_y = \Sigma(y_i - M_y)^2 = -2^2 + -1^2 + 1^2 + 2^2 = 4 + 1 + 1 + 4 = 10$$

or, it can be calculated by the following computational formula:

$$SS_y = \Sigma y_i^2 - (\Sigma y_i)^2/n = 74 - 16^2/4 = 74 - 256/4 = 74 - 64 = 10$$

The same value results in both cases (except for differences due to rounding error),

although the formula appear dissimilar in their superficial form.

Working with Statistical Formula

Most statistical formula involve the rather simple calculations that we have reviewed here. Because the formula combine together operations in various ways, however, the calculations might appear more complex and confusing than they actually are. A few suggestions might help you to appreciate the underlying simplicity of the formula.

One fact that can simplify understanding of more complex formula, especially alternative forms of such formula, is that the same quantity can often be calculated in different ways, as we just showed for SS_y. Let us illustrate with another example, $\Sigma(y_i - 4.0)^2 = 2$, that seemingly different operations may represent the same quantity. We have already shown that:

$$\Sigma(y_i - 4.0)^2 = -1^2 + 0^2 + 0^2 + 1^2 = 1 + 0 + 0 + 1 = 2.0$$

Now consider the following seemingly different operation:

$$\Sigma y_i^2 - (\Sigma y_i)^2/n = (3^2 + 4^2 + 4^2 + 5^2) - (3 + 4 + 4 + 5)^2/4$$
$$= 66 - 16^2/4 = 66 - 64 = 2.0$$

That is, these two seemingly different operations lead to the same value. This is because the value of 4.0 in the first version is actually the mean of the set of scores, represented by *M* or $\bar{y}$. That is, $M = \Sigma y_i/n = 16/4 = 4.0$. What we have demonstrated here through example is that:

$$\Sigma(y_i - M)^2 = \Sigma y_i^2 - (\Sigma y_i)^2/n$$

We will elaborate on this equality in the section on descriptive statistics and will learn to call the computed quantity the Sum of the Squared deviations about the mean, or SS for short. For now, let us simply use it to illustrate how to work effectively with more complex formula. Consider two alternative formula for the standard deviation, which is discussed conceptually in the section on descriptive statistics (recall that $\bar{y}$ represents the mean of the y scores):

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1}}$$

Note that the equality of these formula is obvious once we recognize that the numerator in both cases is the sum of squared deviations about the mean (SS). This also means that the formula could be simplified by substituting SS for the numerator giving:

$$s = \sqrt{\frac{SS}{n-1}}$$

Examine the following two formula for a few minutes to see how they might be conceptualized or simplified. The quantity computed here is the correlation coefficient, *r*, which is discussed in a later chapter.

$$\frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{[\sum x^2 - \frac{(\sum x)^2}{n}][\sum y^2 - \frac{(\sum y)^2}{n}]}}$$

You should first have noticed that the two quantities in the denominator are the SSs for x and y, respectively. We will learn later that the numerator quantity is the sum of cross-products (SCP for short), with the left equation using the definitional form and the right the computational form. Once these basic elements have been identified, then the formula can be written much more simply as:

$$\frac{SCP}{\sqrt{SS_x SS_y}}$$

The important lessons to learn from these examples are: (1) that statistical formula are hierarchical in nature with later formula incorporating earlier (more basic) formula, (2) to learn well the basic statistical quantities as you come across them in the book and class, and (3) to look

for the basic elements in later formula so as to not be intimidated by their more complex form. An analogy can be made to reading. Sentences are composed of strings of letters that make up words. Learning the words allows one to understand sentences, which would otherwise be long strings of meaningless characters. Statistical formula similarly involve meaningful units analogous to words (e.g., SS) and learning the basic vocabulary simplifies the understanding of statistical formula. The present book involves learning the meaning of some of these basic terms.