Understanding Equations in Basic Statistics

Keith A. Markus

July 8, 2019

Part of learning statistics is learning about a number of different equations used in statistics. Learning about equations is not necessarily something that comes naturally. It is a skill that you develop with practice. The purpose of this document is to provide some concepts for thinking about equations and strategies for learning to understand them. These will be helpful in developing your skill at understanding equations.

Understanding the Anatomy of Equations

Consider the equation $a = c + d - b$. The hallmark of an equation is the equal sign that divides the left-hand-side (LHS) from the right-hand side (RHS) of the equation. If there is no equal sign, then the expression is probably a formula rather than an equation. The same equation can be written different ways. Mathematicians might be perfectly happy to write the above equation as $a + b = c + d$ or perhaps as $0 = (a + b) - (c + d)$. In statistics, we typically like to write equations so that nothing appears on the LHS except for the quantity that we want to solve for. In the above example, only a appears on the LHS and this is the quantity that we want to solve for. We might therefore refer to the equation as the equation for a. We refer to the value that we obtain as the solution to the equation.

The elements separated by arithmetic operators on the RHS of the equation are called parameters in mathematics. When we substitute in specific values for the symbolic parameters these values are called arguments. This mathematical terminology differs from the way we use the term 'parameter' in statistics. So, we do not use this terminology very often when we talk about statistics. However, it can be useful for conceptualizing the parts of an equation. In the above example, b, c, and d are parameters. If we substitute 1 for all the parameters, we get arguments and the equation becomes $a = 1 + 1 - 1$. We can thus solve for $a = 1$. In statistics, we generally like to write equations in a form that makes the RHS either easy to calculate or conceptually informative.

Sometimes we use variants of the same equation to solve for different quantities. For example, when designing a study and doing statistical power analysis we might choose a sample size and solve for statistical power or choose a value for statistical power and solve for the sample size. We could rearrange the above example as $b = c + d - a$ to make b the quantity that we solve for.

Whenever you encounter a new equation in statistics, it is a good idea to pause from reading and examine the equation. Identify the parts of the equation. Make sure that you understand the mathematical notation well enough that you can calculate the solution to the equation from a set of arguments to the equation.

1. Identify the parameters in the rearranged version of the equation that solves for b.
2. Substitute arguments for the parameters in both the version that solves for a and the version that solves for b. Confirm that the same set of values solve both forms of the equation.
3. Can you rewrite the equation to make c the quantity that is solved for? (If not, search the web for "solving linear equations" or dig out your old algebra book.[1])

1 For complex problems, there is software to solve equations symbolically (e.g., see http://www.sagemath.org/) but this would be overkill for rearranging simple equations such as the one given above.

Understanding Equations with Examples

Whenever you encounter a new equation, once you understand its parts, make up some examples of arguments and calculate the solution. Jeremy Kun (2018)[2] recommends this because the examples will serve as a basis for you to build up intuitions about how the equation works and what it means. Let's take a new example that is more concrete but still very simple.

$$P_C = \frac{P_B}{C_B} \quad (1)$$

In *Equation 1*, P_C stands for the number of pages per chapter in a textbook, P_B stands for the number of pages in the entire textbook, and C_B stands for the number of chapters in the textbook. So, once we understand the structure of the equation, the next thing we should do is substitute some arguments for P_B and C_B and solve for P_C. It is okay to substitute any values you like but sometimes it is helpful to think about interesting cases. Let's look at some examples for this equation.

Suppose that a book has 500 pages and 10 chapters. This example makes the arithmetic easy and we can easily calculate that this book has 50 pages per chapter (on average, the chapters do not have to be all the same length) because 500 / 10 = 50.

Suppose that a book has 0 pages. We can substitute in any positive number of chapters, and we will always get zero pages per chapter. For example, 0 / 10 = 0. This makes sense because the book has no pages to divide among the chapters.

Suppose that $P_B = C_B$, for example, consider a book with 10 pages and 10 chapters. In this case, the book has one page per chapter.

Finally, suppose that a book has no chapters. In this case, no matter what value we substitute for the length of the book, the number of pages per chapter is not defined. (P_C is not defined in such cases because division by zero is not defined in mathematics.) For example, 100 / 0 is not defined. This makes sense too, because a book with no chapters cannot divide up its pages among the chapters that it does not have.

With a simple equation like *Equation 1*, we can do the arithmetic in our head, by hand, or with a hand calculator. However, sometimes, to avoid being distracted by the arithmetic so that we can focus on the equation, it may be helpful to use a spreadsheet to do the calculations for us. We could set up the above examples in a spreadsheet as follows.

	A	B	C
1	P[B]	N[C]	P[C]

2 Kun, J. (2018). *A programmer's introduction to mathematics*. Oakland, CA: Author. Kun was writing about mathematical definitions. However, it is not much of a leap from definitions to equations because many equations in basic statistics are used as definitions for the quantity on the LHS of the equation.

2	500	10	50
3	0	10	0
4	10	10	1
5	100	0	#DIV/0!

The first row contains column labels typed as text. I have used brackets to indicate the subscript. You could spell out the full words if you wanted to.

The numbers in the first two columns are typed in verbatim from the above examples. These are our arguments to *Equation 1*. The numbers in the third column are not typed in. Instead, we type in *Equation 1*. In cell C2, for example, we would type this as "= A1 / B1" with no quotation marks. If we then copy and paste (or drag with the cursor) this down the column, most spreadsheets will automatically update the rows to 2 to 5. In cell C5, this formula returns an error message warning us that we are trying to divide by zero.

1. We can rearrange *Equation 1* to solve for the length of the book as $P_B = C_B \times P_C$. Identify the parameters and the quantity solved for in this new equation.
2. Make up some examples of arguments and solve for the solution to better understand this equation.
3. Try setting up your examples in a simple spreadsheet just for practice.

Understanding Equations with Graphs

Sometimes it is helpful to see how the solution to an equation changes over a range of values for one or more of the arguments. You could do that with a large table. However, when tables become too large, it can be hard to take in the patterns in the numbers. So, graphs offer a useful alternative for getting to know an equation in these circumstances. Because it is written for beginners, this document will focus on drawing graphs in a spreadsheet. However, as you progress, you may find it more convenient to draw graphs in statistical software such as *R*.[3]

Drawing graphs in a spreadsheet is usually a matter of setting up the values that you want to graph in adjacent columns and then choosing the type of graph that you want. We will use line graphs. For simplicity, we can continue working with *Equation 1*. This time, instead of a few choice examples, we want to compute P_C for a sequence of values within a range. Suppose that we focus for the moment on C_B. How does the number of pages per chapter vary across values of the number of chapters?

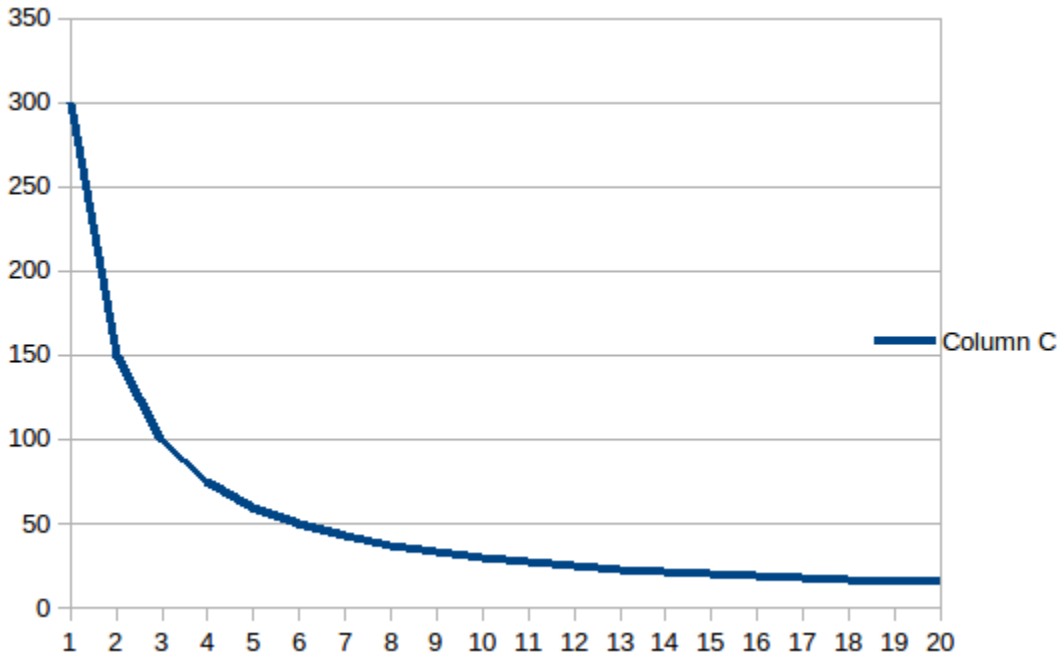
To calculate P_C, we need to choose a fixed value of P_B. Suppose we work with 300. This seems like a plausible number of pages for a textbook and we can try other values later.

We can set up the spreadsheet similarly as we did for our earlier examples. However, this time we will use a range of C_B values going in small steps. Suppose we are interested in the range from 1 to 20 chapters. It does not make sense to have fractional numbers of chapters, so we can go in increments of one chapter from 1 to 20. To save space, I am only going to show you the first five rows of the table, but we would continue the same way for 5 to 20 chapters. If you set it up correctly, you should see 15 pages per chapter for a book with 20 chapters and 300 pages.

3 See https://www.r-project.org/

	A	B	C
1	P[B]	N[C]	P[C]
2	300	1	300
3	300	2	150
4	300	3	100
5	300	4	75

To create the chart, I selected columns B and C within the range of my table of values using my mouse. Then I inserted a chart and selected a line chart as my type. I then had to check a box to tell the spreadsheet that column B was a label. This produced the below graph.



We can see that the number of chapters is listed on the horizontal axis and the pages per chapter is listed on the vertical axis. If you took the time, you could tidy up the graph with labels. However, detailed instructions for using a particular spreadsheet are beyond the scope of this document. Use whatever software you are most comfortable with and consult the help for that software or search online for specific instructions. (I used LibreOffice Calc for these graphs and used LibreOffice Writer for this document.)

Looking at the graph, we can easily see that as the number of chapters goes up, the number of pages per chapter goes down (assuming, as we did, a fixed number of pages for the book). If we look further, we can see that for books with a small number of chapters, the number of pages per chapter drops very quickly with each additional chapter. For example, moving from two chapters to three

chapters reduces the pages per chapter from 150 to 100. However, when there are more chapters, then the pages per chapter reduces more slowly with each additional chapter. For example, there is almost no visible difference from a book with 19 chapters (rounds to 15.79 pages per chapter) to a book with 20 chapters (15 pages per chapter). It is impossible to have less than zero pages per chapter. So, the graph is reaching a lower asymptote of zero. Even if we extended the number of chapters to positive infinity, the line would never cross zero. As a result, it goes down by less and less with each added chapter to keep it from crossing zero.

Sometimes we have to use common sense (or theory) to put limits on the range for which we plot a graph. For example, we have already seen that if we extend one more value to the left, so that $C_B = 0$, then P_C becomes undefined. This will appear as a break in the line because there is no value to plot. However, you have to be careful because if you do not include zero as a value on the horizontal axis to be plotted, but include values on both sides of zero, the line might run continuously through zero by connecting dots on both sides. Likewise, we can calculate $P_C = -300$ for $C_B = -1$. This is mathematically defined. However, it does not make sense for a book to have a negative number of chapters. As a result, it does not make sense to plot P_C for negative values of C_B. We could plot values greater than 20, but it seems unlikely that a book intended as a textbook would have so many chapters.

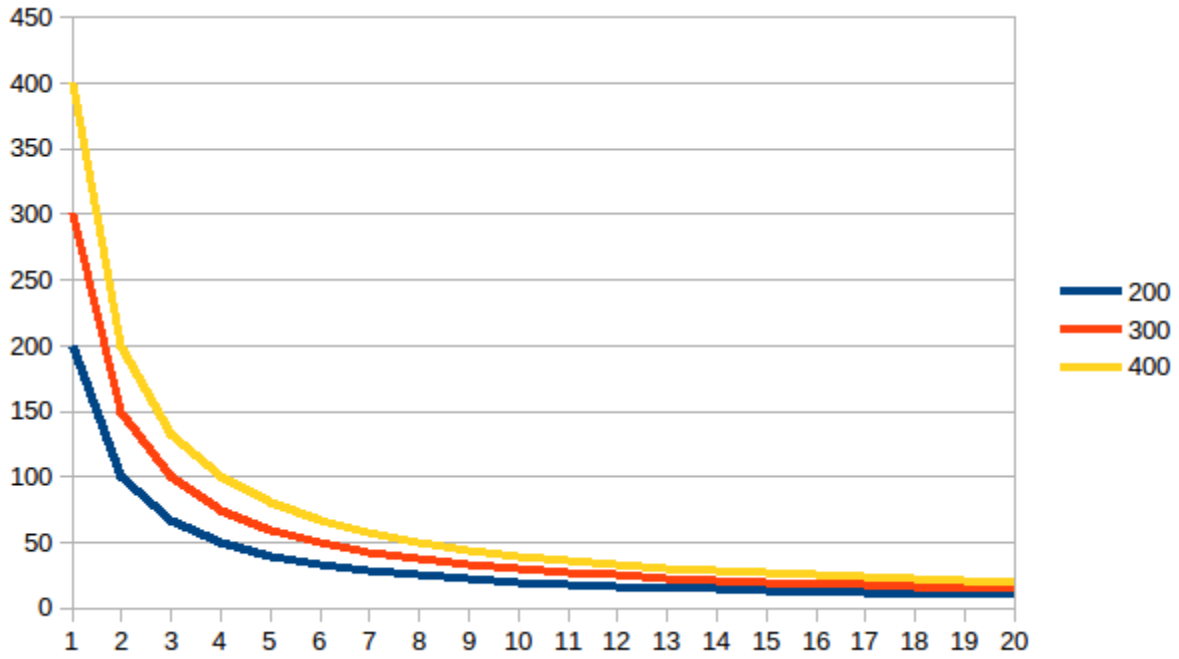
How far can we generalize the above conclusions about *Equation 1* regarding books of different lengths? To explore that, we can add more columns for different values of P_B. It is easier to draw the graphs in a spreadsheet if we keep all the columns of data in the same rows. So, a good strategy is to remove P_B from the first column and instead place it in a row that spans the columns. Then each column of P_C values reflects the P_B values above it. In this case, for example, the value of 100 pages per chapter in cell B3 reflects a book length of 200 pages given in cell B1.

	A	B	C	D
1	N[C]	200	300	400
2	1	200	300	400
3	2	100	150	200
4	3	66.67	100	133.33
5	4	50	75	100

In cell B2, you would enter the equation as "= B$1 / $A2" again with no quotation marks. The purpose of the dollar signs is to make it easier to drag the formula to the other cells. When you put a dollar sign in front of an element of a cell label, that element of the label does not change when you drag across cells. In this case, when I dragged across to cells C1 and D1, the reference to cell A2 remains A2 instead of updating to B2 and C2 as I go across the table. Likewise, when I then dragged the three cells down 20 rows, the reference to row 1 remained row 1 instead of updating to 2 through 19.

I rounded the values in cells B4 and D4 to two decimal places. By default, your spreadsheet might show more decimal places. Again, I am only showing you the first four rows of data but there are actually 20 rows in the full table.

Now I can highlight the table to draw the graph. In this case, I told the spreadsheet that both the first column and the first row were labels. As a result, not only do the C_B values appear on the horizontal axis but the P_B values appear in the legend on the right margin.



Looking at the graph, we can see that the lines are very similar. The yellow line is highest because this represents the book with 400 pages and thus has the most pages per chapter. Likewise, the blue line represents the shortest book and has the fewest pages per chapter. However, in every case we see the same basic features that we saw above. As the number of chapters goes up, the pages per chapter consistently goes down. Moreover, all three lines approach an asymptote of zero, and thus the amount by which the pages per chapter goes down gets less and less as the number of chapters increases.

From the graph, we can also see that the pages per chapter differs most across book lengths for books with relatively few chapters. As the number of chapters increases, the pages per chapter in books of different lengths become more and more similar.

1. Extend the number of lines in the graph to cover book lengths of 100 and 500.
2. Try plotting the version of *Equation 1* that solves for the length of the book in terms of the number of chapters and the pages per chapter.
3. In statistics we sometimes make use of mathematical functions such as square root, natural log, and natural exponent. You can calculate these functions in your spreadsheet, perhaps by using *sqrt(x)*, *log(x)*, and *exp(x)* where you substitute a cell reference for *x*. Graph these functions with *x* on the horizontal axis to get a feel for how they look.

Conclusion

That concludes this tutorial on understanding equations in basic statistics. Use these concepts and strategies to build your skill at understanding equations. However, remember that it is a skill. Skills do not develop all at once in a flash of insight, they come gradually with practice. So, be patient. Like other skill based activities, understanding equations gets easier with practice.

None of the equations discussed as examples were actually statistical equations. As you encounter statistical equations in the textbook, apply your skills to understand those equations. Remember that in statistical equations, the quantity solved for often does not represent anything as concrete as the number of pages per chapter in a textbook. Instead, it often represents a more abstract statistical quantity that is useful because it follows a certain statistical distribution or has some other useful property. Often, understanding the equation is key to understanding such quantities.