Notes by David Groisser, Copyright ©1998, 2018, 2019

What is a solution?

Colloquially, a *solution to a problem* usually means some method by which you find an answer. Mathematicians, however, distinguish the *method* from the *answer*, especially when the problem concerns one or more equations. When a mathematician uses the phrase *solution to* [or *of*] *an equation* (s)he almost always is referring to the answer (a value or set of values for a variable or variables), NOT to any method by which the answer is found.

Let's elaborate. A (non-definitional)[1] equation makes an assertion that two particular things are equal, a statement that is either TRUE or FALSE. For example, the equation $3^2 = 9$ is a true statement, while $1 = 0$ is a false statement. When an equation contains one or more variable objects (perhaps numbers, perhaps functions, perhaps something else), the equation is meant to stand for a *family* of statements, one statement for each possible value of the variable(s).

Example 1. Consider the equation

$$x^2 = 2 \tag{1}$$

for a variable real number (or *unknown*) x. This equation is a true statement when x is $\sqrt{2}$ or $-\sqrt{2}$, but is false otherwise.

Example 2. Consider the differential equation

$$f' = f \tag{2}$$

for an unknown real-valued function f whose domain is the whole real line. In this case, "=" means "equal as functions"; the equation "$f' = f$" is short-hand for the statement

$$f'(x) = f(x) \text{ for every value of } x \text{ in the domain of } f.$$

The equation $f' = f$ is a true statement if $f(x) = e^x$ for all x, or more generally if there is some real number c such that f is given by $f(x) = ce^x$ (for all x); it is a false statement if f is any other function. The "true" variable in equation (2) is the *function* f, not the number x.[2]

[1] An example of a *definitional* is the equation in "Define a function f by $f(x) = x^2$."

[2] Often we write the differential equation (2) using notation such as "$\frac{dy}{dx} = y$", in which case it is common to calll x the *independent variable* and y the *dependent variable*. There is nothing wrong with this terminology, but in it the word "variable" has a differently nuanced meaning from what we have used above, coming from the concept of *function* rather than from *family of true/false statements*. When we say that x is the independent variable in "$\frac{dy}{dx} = y$", we mean simply that x is the letter we are choosing for elements of the domain of the functions we're considering. When we say that y is the dependent variable in this equation, we may mean one or both of the following: y is the letter we're choosing for the *functions* we're considering—exactly what the letter f was used for in equation (2)—or that y is the letter we're using for the *outputs* of those functions (so that we can use the xy plane when we graph those functions).

More generally, consider the equation

$$\text{some stuff done to some variables} = \text{something else} \tag{3}$$

(the "something else" may or may not involve variables). It must be specified what sort of object the variable(s) is (are) allowed to be. In Example 1, we allowed x to be any real number, but we could have considered the same equation just for positive real numbers, or just for integers. There are exactly two real numbers x for which $x^2 = 2$ is a true statement; there is exactly one positive x for which $x^2 = 2$ is true; and there are no integers x at all for which $x^2 = 2$ is true. In Example 2, we allowed the variable (f, not x!) to be any real-valued function with domain the whole real line. So in general, one considers a set of the *allowed values* of the variable(s).

Below I'll phrase things as if there's only one variable in a general equation, but the same considerations apply no matter how many variables there are.

Definition of "a solution". Consider the equation (3) for a specified set of allowed values of the variable. An allowed value v is a *solution* of this equation if, when the value v is plugged into (3), the resulting equation is a true statement.

Thus, in Example 1, $\sqrt{2}$ is a solution of (1), and so is $-\sqrt{2}$; there are exactly two solutions. In Example 2, the function f defined by $f(x) = 17e^x$ is a solution, and so is the function f defined by $f(x) = ce^x$ for any other constant c; there are infinitely many solutions.

Note that we have just defined the phrase "a solution"; I was careful not to use the phrase "the solution". The definite article "the" should only be used when the object it refers to is *unique* (i.e. there is only one such object). We *can* attach meaning to "the solution of an equation" when the equation has more than one solution, *provided that we use the following definition* (which not everybody does):

Definition of the solution. THE solution of (3) (for a specified set of allowed values of the variable) is the *solution set*, i.e. the set consisting of ALL allowed values which are solutions of (3). (Sometimes this is called the "general solution".)

Thus, in Example 1, the solution is the set $\{\sqrt{2}, -\sqrt{2}\}$, usually abbreviated $\pm\sqrt{2}$. Had we restricted the set of allowed values to be positive real numbers, the solution would have been the single number $\sqrt{2}$; had we restricted the set of allowed values to be integers, the solution would have been the empty set.

Important: Thus, *a* solution of an equation is an allowed value of the variable that makes the equation a true statement; *the* solution is the set of all allowed values that make the equation a true statement. The phrase *solution of an equation* DOES NOT refer to any method for producing values of the variable that make the equation true. For example, 3 is a solution of the equation $x^5 + 10x - 23 = 250$ for a real variable x; there are no if's and's, or but's about it. The fact that I don't have a formula for finding the general solution (or even one solution) of a general fifth-degree equation is irrelevant to deciding whether 3 is a solution of the particular equation above.

Here is another way of looking at the same thing. In high school, you may have been given problems that stated "solve the following equation and check your answer". If I tell you "Show that 3 is a solution of $x^5 + 10x - 23 = 250$" I'm basically saying, in high-school terms: I've done the solving, you do the checking.

Of course, one of the central themes in mathematics is how to find solutions of equations. The definitions above are not made to minimize the importance of having methods for finding solutions; they're just made so that when we talk to each other we can both be sure what we're talking about. When we try to solve equations, what's of the utmost importance is that our method produce a solution, not that it be some cookbook recipe that produces an answer whose significance we don't understand.

Even mathematicians have other uses for the word "solution", depending on context. There is generally a distinction between a *solution to* [or *of*] *a problem* and a *solution to* [or *of*] *an equation*; the former sometimes refers to a method for finding the latter.

Many of the same considerations apply to the verb "to solve" as to the noun "solution". The phrase "[this value of the variable] *solves* a certain equation" is synonymous with "[this value of the variable] is a solution of the equation". The instruction "solve [this equation]" usually means "find THE solution of [this equation]". As an exercise, figure out the different meanings of "solves" in the following two sentences.

(1) 5 solves the equation $x^2 = 25$.

(2) Fred solves the equation $x^2 = 25$.