

Beyond Facts

In some of the sciences there is an unstated understanding that facts and theory are separate: According to this idea, numbers are “done” by technicians grappling with the messy stuff of data and reality. Meanwhile, theory is done by theorists, preferably pure theorists, applying their mind to pure ideas and general principles. The idea that such a division works is destructive and dead wrong. It probably grew up as an attempt to make the best of incomplete and divided training among (my generation) of scientists, converting the unhappy fact of division into a virtue: People who work with words are presumed to be theoreticians; people who work with data are presumed to be methodologists.

But verbiage does not make the speaker a theoretician any more than numbers make their user a methodologist. There simply is no such division among active scientists: We construct and develop our ideas by working with the facts; we shape our analysis of the facts in order to shape and test our ideas — there is no division. You will still find a few grand theorists who disdain data. You may ignore them. You will still find a few data crunchers who think that science is a pile of facts, and that greater science is a greater pile. You may ignore them.

The second of the two concepts is a strategy of reverse logic. Non scientists tend to think that scientists “prove” their theories, that we “prove” the laws of physics and “prove” the theory of evolution. We don’t. The trouble is that we can never “prove” anything about the world: The very best we can say is that a theory is consistent with the data — that we have no counter-evidence, not yet. In contrast, disproof can be clear and definitive — just once instance of an event, of something that can’t happen, according to the theory and the theory has been disproved. (or is in need of modification). That’s all it takes. So since disproof can be relatively easy to spot (and proof is impossible) we tend to work by a reverse logic that makes the most of our errors.

And then, bringing the two concepts to together, one of the things we learn the most from is *badly* behaved data — because surprises force us to rethink what’s going on.

The Potato Hypothesis

The place where the ideas and the data merge is in the equations of the hypothesis. Recapitulating, I referred to the pseudo equation

$$Data = Signal + Noise$$

which allowed me to isolate noise as noise = data - signal

$$Data - Signal = Noise$$

Then I suggested that the analyst construct a hypotheses about the signal and evaluate the hypothesis by looking at the residuals: if the residuals look like the noise, then the hypotheses captured the signal

$$Data - Hypothesis = Residual$$

There is the hypothesis. In the example, the hypothesis was a rather dry statement using a linear equation for y and x as well as estimates of the values for the intercept and the slope. That was about all I could say about a hypothesis when I was using only letters and numbers, x, y, 10, 20, and so forth. But when the hypothesis is about a phenomenon — about some process for which you have gathered the data, then things get much more interesting.

Here are some data. The data are meant to describe the response of crops of potatoes to the application of fertilizer.

The data come from a controlled experiment (Rothamsted Experimental Station Report, 1933) on the effects of increasing amounts of a mixture of the standard crop fertilizers on the yields of potatoes that was carried out in 1933 at the Midland Agricultural College in England. The mixture contained 1 part of sulfate of ammonia, 3 parts of superphosphate, and 1 part of sulfate of potash. The amounts were 0, 4, 8, and 12 cwt per acre, the cwt unit, called hundredweight, being actually 112 lb. Owing to natural variability, the yields of potatoes under a given amount of the mixture vary from plot to plot. The yield figures shown for each amount in table 9.2.1 are the means of random samples of four plots.

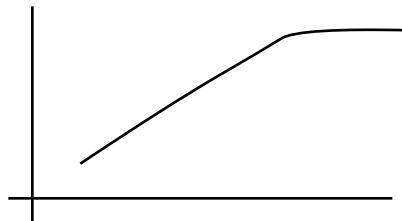
From *Statistical Methods, Seventh Edition*, Snedecor and Cochran, Iowa State University Press, 1980, pp 149-150:

YIELD OF POTATOES IN RESPONSE TO FERTILIZER	
Fertilizer in cwt per acre X_i	Potatoes in tonnes per acre Y_i
0	8.34
4	8.89
8	9.16
12	9.50

Now — think! What do I expect and why? That is the half of hypothesis construction that was missing when I used only numbers. What do I expect? Crudely, I expect a positive response to fertilizer, more fertilizer, more potatoes.

Less crudely, I don't know enough about crops to offer a very clever hypothesis. I think of plants as grabbing up nutrients from the soil. The more the nutrient available, the more that will be assimilated by the plant, and the greater the yield — up to a point when the growth approaches the limits of the organism itself.

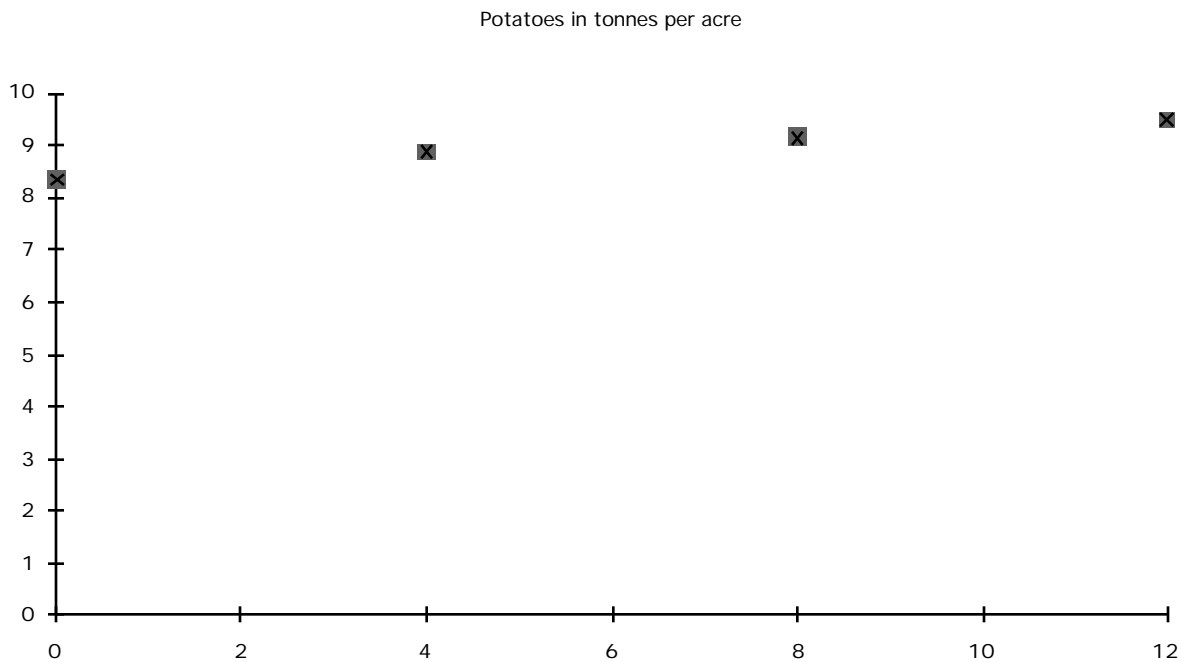
Those are my initial thoughts. Now I crystallize these thoughts into a picture of my expectation, and, if I can, into an equation that will join the thought to the fact. My words suggest a sketch of a linear response, for a part of the range, followed by a bend approaching an asymptote.



What specific equation would I attach to my sketch? I'll have to do it in two parts: Direct proportionality to the fertilizer initially. That would be a line. Later? I would expect something like each additional unit of fertilizer having an equal effect on the distance between the yield and the asymptote. Likely, however, I will not have enough data to track this down very precisely.

Now, I'm ready to look at the data.

			intercept is:	0
			slope is:	0
Unit	fertilizer in cwt per acre	Potatoes in tonnes per acre	expected= intercept + slope Variable x	Residuals= y-expected *
# 1	0	8.34	0	8.34
# 2	4	8.89	0	8.89
# 3	8	9.16	0	9.16
# 4	12	9.50	0	9.50



Let me look at this. Before the beginning: Who, what where...? I probably have enough in that descriptive paragraph. Now, are these variables well-behaved? If I evaluate them in terms of their one variable distributions, then I can't tell: The data are from a controlled experiment. That means that I can't really think about the distribution of the applications of fertilizer. The experimenter chose the distribution that, so there is nothing "natural" to be discovered from looking at this distribution to see how it behaved — although it is likely that the experimenter considered the zero, (no fertilizer), and the three applications to be equally spaced. So, I'm unable to consider either the first or the second criterion for a well behaved variable. But the third criterion, linearity, looks useful: The graph suggests that the relation is approximately linear, suggesting that each variable is approximately well behaved. There is more data on the variation of the yield per acre, these are average yields and that average must refer to some distribution. But I don't have it. The fifth criterion is sense. And that is a problem. Whether the data come in pounds, or ounces, or cwt, I can do the arithmetic. But the purpose of the first steps of a data analysis is to feed the intuition of the analyst. So let me tell you about the analyst. I'm an American. My intuition knows pounds and tons, not cwt. What does a cwt of fertilizer look like? Is it a shovelful, or a truckload? I can figure it out. But if I use these units of measure then I am going to slow down my work because every stage will need to be translated until it makes sense. That means I'm not really able to *think* about this. I don't know what 4 cwt of fertilizer is, nor for that matter do I understand a tonne of potatoes. In this primitive sense, I need a change of the unit of measure —so that I can "think" about these without burdening my intuition with translations.

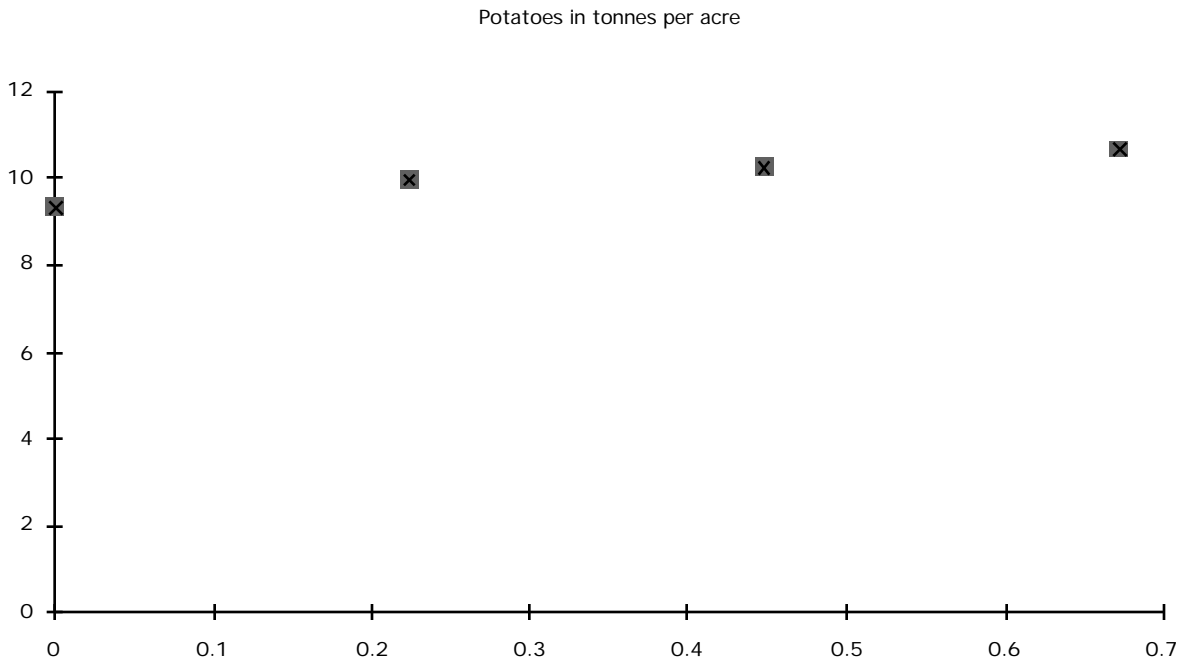
So — to the dictionary. A cwt is a hundredweight. So much for word origins: *cent* (hundred) *weight*, cwt. And what is a hundred weight? It says, "a unit of weight, equal to 100 pounds in the United States and 112 pounds in England". The data are British, and I should take note of that spelling "tonne", also British. So I have 112 pounds of fertilizer for each hundredweight; .056 tons (American) of fertilizer for each hundredweight. Changing the unit measure, the three applications are .22 tons, .45 tons and .67 tons per acre.

My dictionary does not include "tonne", but it has two definitions of "ton", one British. That should be it. A ton, it tells me is "a unit of weight equal to 2,240 pounds avoirdupois (1,016.06 kilograms) commonly used in Great Britain: in full **long ton**, **shipping ton**." Also, "a unit of weight equal to 2,000 pounds avoirdupois (or 907.20 kilograms), com-

monly used in the United States, Canada, South Africa, etc.: in full **short ton.**” So multiplying 8.34 by 2,240 pounds and dividing by 2,000 pounds, the yield of an unfertilized acre is 9.3 tons. And successive yields are 10.0 tons, 10.3 tons, and 10.6 tons per acre.

Converting to these units of measure, I'll start again

			intercept is:	0
			slope is:	0
			intercept is:	0.0
			slope is:	0.0
Unit	fertilizer in tons per acre	Potatoes in tons per acre	expected= intercept + slope Variable x	Residuals= y-expected *
# 1	0	9.3408	0.0	
# 2	0.224	9.9568	0.0	
# 3	0.448	10.2592	0.0	
# 4	0.672	10.6400	0.0	



Now, using these units of measure, I'm willing to look. And what do I get from this integration of ideas and number? I get the message that my first ideas were an exhibit of sloppy thinking. I said that yield would be proportional to fertilizer. Note that I drew my sketch pointed toward zero. I was thinking *proportionality* and built it in to my sketch (although my words were more ambiguous). Drawing the line through zero is what "proportionality" means: y is proportional to x , i.e.,

$$y = mx$$

(not $y = mx + b$)

But that is not the relation shown in the picture. The picture clearly shows that soil is perfectly capable of producing potatoes without the assistance of fertilizer. My thinking was too narrow. I was thinking about fertilizer and yield and human intervention, as if fertilizer were necessary to induce nature to grow crops. And so I simply failed to step back far enough to take into account the very hefty yield that nature can deliver when there is no fertilizer at all.

So, chastised by the data, let me revise my hypothesis by stating it more precisely. I expect the *increase* in yield to be proportional to the increase in *fertilizer*. That idea is matched by a linear equation with an intercept.

I know this revised hypothesis will be in the "ball park" of the data, because I've looked at the data. But I can still test the idea by a straight forward application of reverse logic: Is this hypothesis a fair description of the signal? If it is, then the residuals will look like noise. So, with reverse logic, I direct my attention to the residuals — to evaluate the acceptability of the hypothesis. As for the specific values of slope and intercept, I don't have enough experience to have an expectation or a hypothesis. I'll take these from the data.

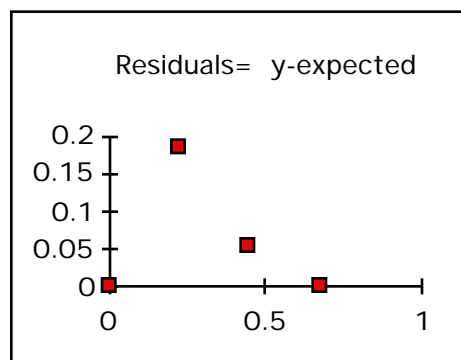
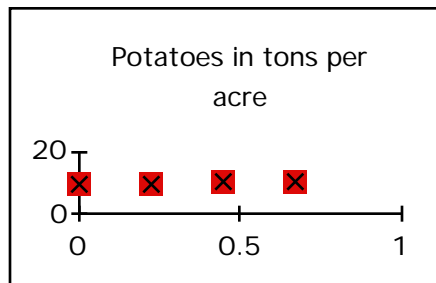
From the data, a first guess for the intercept is obvious, 9.34 tons per acre because that is the observed value of the yield when the fertilizer is 0. A first guess for the slope is also obvious, the vertical rise on the graph is from 9.3 to 10.6. The horizontal run on the graph is from 0 to .67. And the slope is the rise divided by run. That gives me a first guess of approximately 1.93 tons of potatoes per ton of fertilizer.

Even without a graph, that is very interesting: *1.9 tons of potatoes per ton of fertilizer*. That surprises me: To increase the yield of potatoes by

1.9tons I have to apply 1 ton of fertilizer to the soil!. (And the 1.9 tons of potatoes are 95% water.) Non agriculturist that I am, non-biologist that I am, I'm astonished, one ton of fertilizer spread on the field to get 2 additional tons of potatoes.

Husbanding my three pieces of information, intercept, slope, and residuals, I'll try placing these estimates for the intercept and the slope into the hypothesis. Then I will look at the pattern of the residuals, asking do these residuals look like noise?

			intercept is:	9.34
			slope is:	1.93
Unit	fertilizer in tons per acre	Potatoes in tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
# 1	0	9.3408	9.340	0.0008
# 2	0.224	9.9568	9.772	0.18448
# 3	0.448	10.2592	10.205	0.05456
# 4	0.672	10.6400	10.637	0.00304



Residuals with respect to the line with intercept = 9.34, slope = 1.93.

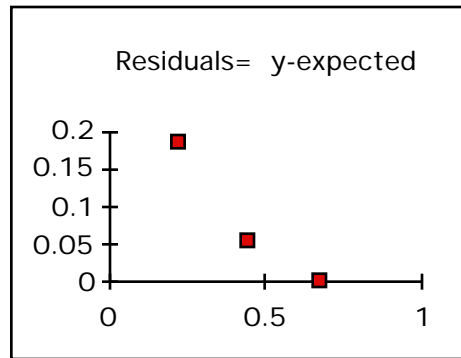
No, these residuals certainly suggest a pattern among the residuals, not noise. So, by reverse logic, my hypothesis doesn't represent the signal. My equation was wrong. And what is important is not just that my equation was wrong. What is important is that my thinking was wrong. Writing my thoughts into the equation allowed me to test the ideas with data. My ideas were wrong again.

I'll keep going with these data but from here forward I am not going to be able to test my subsequent . I'll call the next idea a hypothesis. But by now I've seen so much of the data that I am really just writing a hypothesis to fit the facts — to test it I would need other data.

With that proviso, I *think* that what I see in these data is that the soil alone (without fertilizer) was capable of a yield of approximately 9.3 tonnes per acre. The first addition of 0.2 tons of fertilizer to the untreated soil added about 0.6 tons to the yield, .2 tons in, .6 tons out. And each additional .2 tons produced an increase of .33 tons, .2 tons in .3 tons out. In words, the untreated soil, without fertilizer, comes within 88% of the maximum yield achieved with the heaviest application of fertilizer (calculating $9.34/10.64 = 88\%$). The first application of fertilizer to the untreated soil has a disproportionately large gain as compared to subsequent applications. Perhaps the effect of the fertilizer is catalytic as well as directly nutritive so that the initial application enables the organism to use nutrients that were already present. This is a one-shot effect. As a result further increments of fertilizer will add only the nutritional effect.

For my graph, the only thing to be "tested" after all this handling of the data is that the last two increments are approximately equal. Limiting my graph to the last two increments (the last three data points):

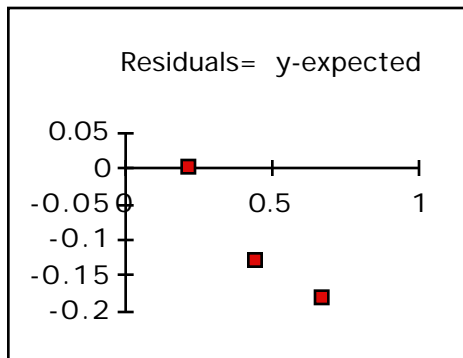
			intercept is:	9.34
			slope is:	1.93
Unit	fertilizer in tons per acre	Potatoes in tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
# 1				
# 2	0.224	9.9568	9.772	0.18448
# 3	0.448	10.2592	10.205	0.05456
# 4	0.672	10.64	10.637	0.00304



Residuals with respect to the line with
intercept = 9.34, slope = 0.097.

From that, I can bring the first residual down to zero by treating its value, .184, as a positive signal. I remove this positive signal from the residuals by adding it to the hypothesis.

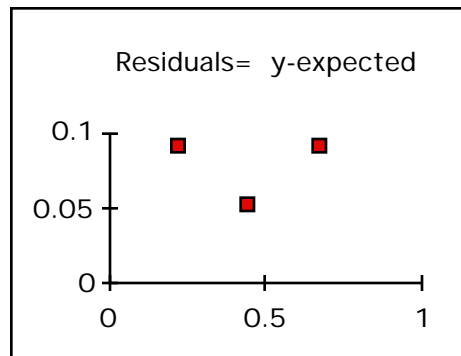
			intercept is:	9.524
			slope is:	1.93
Unit	fertilizer in tons per acre	Potatoes in tons per acre	expected= intercept + slope Variable x	Residuals= y-expected *
# 1				
# 2	0.224	9.9568	9.956	0.00048
# 3	0.448	10.2592	10.389	-0.12944
# 4	0.672	10.64	10.821	-0.18096



Residuals with respect to the line with
intercept = 9.524, slope = 1.93.

Then I can work on the slope suggested by the residuals: The residuals show a vertical descent of $-.181$ over a horizontal run of $.448$. Calculating $-.181/.448 = -.405$, that is a signal of $-.405$ in the residuals. I take it out of the residuals by adding negative $-.405$ to the hypothesis.

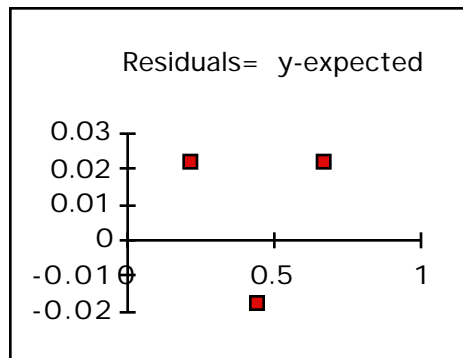
			intercept is:	9.524
			slope is:	1.525
Unit	fertilizer in tons per acre	Potatoes in tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
# 1				
# 2	0.224	9.9568	9.866	0.0912
# 3	0.448	10.2592	10.207	0.052
# 4	0.672	10.64	10.549	0.0912



Residuals with respect to the line with
intercept = 9.524, slope = 1.525.

In turn, I can go back to the intercept for fine tuning. Note that it *is* fine tuning, because the size of these residuals is small. It now shows an intercept of about .07 in the residuals. I take this .07 out of the residuals by adding it to the hypothesis, getting:

			intercept is:	9.594
			slope is:	1.525
Unit	fertilizer in tons per acre	Potatoes in tons per acre	expected= intercept + slope * Variable x	Residuals= y-expected
# 1				
# 2	0.224	9.9568	9.936	0.0212
# 3	0.448	10.2592	10.277	-0.018
# 4	0.672	10.64	10.619	0.0212



Residuals with respect to the line with
intercept = 9.594, slope = 1.525.

Those are about as harmless a set of residuals as I can imagine but, again, these residuals aren't really a test of anything because, now that I am down to three data points, if there is any error at all (which there always is), this is what it has to look like, down/up or up/down. (Anything else, for example, up/up would have been interpreted as a slope and then removed from the residuals.)

So, what do I know? Let's put it in order, noting that the most obvious things were obvious, once there was a picture.

I know that the untreated fields averaged ____ tons of potatoes per acre. (I know that by converting the data to common units and reading the data.)

I know that the lightest application of fertilizer received the greatest return in terms of additional tons of potatoes per ton of fertilizer, about 2.75 tons of potatoes per ton of fertilizer. (I know the effect from looking at the residuals. I know the number by simple calculation directly from the data.)

I know that additional fertilizer had a lower marginal return of potatoes about 1.525 tons of potatoes per ton of fertilizer. (I know this by examination of secondary residuals and by fitting a line to the last three of the four data points.)

I also know that my initial ideas had very little to do with what was found. Certainly it is not true that growth is proportional to fertilizer. It is not even true that additional growth is proportional to additional fertilizer. Instead additional growth is realized at one rate for the first application of fertilizer and at a lower rate for greater applications of fertilizer. And, finally, these data provide no evidence of a diminishing rate of return.

In case you hadn't noticed, there is an almost inverse relation between the utility of each piece of information gained from various aspects of these data and the amount of work that was necessary to extract that information. Most of what was learned: By learning that my hypothesis was sloppy (and wrong), and then by getting a description of the actual behavior of the data — these things can be read directly from the data and the second graph. The less obvious things, uncovered with page after page of technical virtuosity, added detail: Before I went through this detailed procedure I knew that the slope for the last three data points was about 1.5 (in the phrase “.2 tons in, .3 tons out”). Now I know it is more like 1.525 than 1.5 — which is to say the largest technical display in this analysis was attached to the smallest gain of real information, increasing the precision of the estimate from approximately 1.5 to approximately 1.525.

That is a hard blow to the ego of the data analyst. By the end of an analysis you are focused on your most technically sophisticated efforts. But very often, the technological sophistication added little to what was known when you graphed two well-behaved variables, one against the other. That was your real act of sophistication: The real act of sophistication is getting the initial display right so that the most

important results become the most visible features of the graph. Once the initial display is under control, the obvious results became obvious.

The next blow to the analyst comes in the write up. Truth is, nobody cares how hard you worked. “Heh! I’ve got 14 pages of work here (in manuscript).” Tough. That’s how much work it takes, but no one cares. They want to know, in this case, about fertilizer and potatoes. Ontogony may recapitulate phylogeny (in biology), but in data analysis, the report does not recapitulate the research — except to leave some evidence for the cognoscenti: The cognocenti (other analysts) need assurance that you actually did the research and they need sufficient information to allow them to reproduce the work for themselves if they care to.

Data from the Rothamstead research station present a picture of the increases in crop yield that may be obtained through the application of fertilizer. The data show that the lightest application of fertilizer brought a return of 3 tons of potatoes per ton of fertilizer, approximately a seven percent increase compared to the untreated fields. Additional applications of fertilizer at rates that were double and triple the initial application achieved a smaller increase of approximately 1.5 additional tons of potatoes per additional ton of fertilizer

The data come from a controlled experiment (*Rothamsted Experimental Station Report, 1933*) on the effects of increasing amounts of a mixture of the standard crop fertilizers on the yields of potatoes that was carried out in

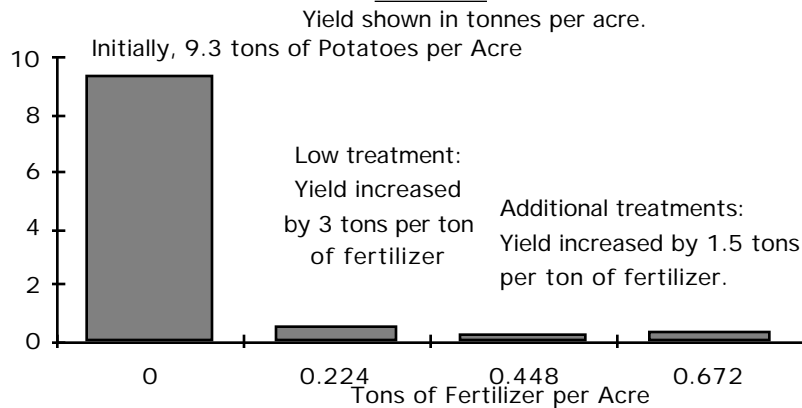
1933 at the Midland Agricultural College in England. The fertilizer contained 1 part of sulfate of ammonia, 3 parts of superphosphate, and 1 part of sulfate of potash, with applications and average crop yields as shown in the table below.

Yields of Potatoes Corresponding to Different Amounts of Fertilizer

Fertilizer	Average Yield (cwt per acre)(tonnes per acre)
0	8.34
4	8.89
8	9.16
12	9.50

Secondary source: *Statistical Methods*, Seventh Edition, Snedecor and Cochran, Iowa State University Press, 1980, p 152.

Initial and Incremental yield of Potatoes in response to Fertilizer



Exercise: From *Statistical Methods*, Snedecor and Cochran, Iowa State University Press, 1980, p. 153:

Two problems:

In a controlled experiment on the effects of increasing amounts of mixed fertilizers on sugar beets conducted at Redbourne, Lincs, England, in 1933, the mean yields of sugar beet roots and tops ($n=5$) for each amount X are as follows:

X (cwt/acre)	0	4	8	12	16
Roots (T/acre)	14.42	15.31	15.62	15.94	15.76
Tops (T/acre)	7.48	8.65	9.74	11.00	11.65

Problem 1: Analyze the data for roots.

Problem 2: Analyze the data for tops.