Don't Drive Into Smoke: Evaluating Data

Normand R. St-Pierre[1]
Department of Animal Sciences
The Ohio State University

Abstract

All observations (data) contain errors. Understanding the sources of these errors is important to reach the correct decision from the data, or else you risk driving into smoke. Some sources of errors are linked to the physical limitations of the measuring devices. This is the type of errors that people working in the physical sciences are accustomed to. Reporting data with more digits than what is legitimate from the precision of the instrument is frequent, but very misleading. People working with live things, such as cows, must understand that data also contain errors because living entities vary. For example, the milk production and body weight of a given cow continuously vary. The sizes of the daily variation of many traits within a cow are such that little can be inferred from one single datum. In addition, there is variation amongst animals treated alike, which is the basis of replication in research. Because cows within a pen are not independent, any factors common to a pen will affect all animals within it. Looking at feed analyses, data contain errors (variation) that are intrinsic to the feed (i.e., true), and errors that are due to the observer. In most instances, the sampling variation in forages is such that little can be inferred from a single sample. Much progress would be made if 2 independent samples were taken and assayed each time a nutritionist need data on feed composition.

Introduction

Little did we know that the first implementation of the transistor by 3 American physicists in 1947 would lead to the mountains of electronic data now inundating the scientific disciplines. Data are now so much embedded into the scientific process that some have expressed doubts whether Einstein would be a successful scientist had his career been delayed by one century. Massive amount of data are invading not only the scientific world but also the management of various processes of which agriculture, in general and dairying in particular, have vastly benefited from. We have reached a time when observations must be quantified in the form of data if they are to carry credibility and be acted upon. Unfortunately, too often people forget that data contain inherent errors, that these errors are of many types, and which in the end, substantially affects their interpretation. In this paper, we review the different types of errors using different examples. This leads us to many cautions regarding possible misuse and abuse of data.

Data Contain Errors

Imagine for a second that you are part of a photo safari in the Australian Outback with the specific goal of taking pictures of *Macrofus rufus*, better known as red kangaroos. Soon after you disrupted a large mob of kangaroos

[1]Contact at: 2029 Fyffe Court, 221A Animal Science Building, Columbus, OH-43210-1095, (614) 292-6507, FAX: (614) 292-1515, Email: st-pierre.8@osu.edu.

from their afternoon nap, you find yourself hanging for dear life in the passenger seat of an Australian ute (better known in North America as a pick-up truck), as the driver speeds through the arid landscape in reckless pursuit of the bouncing animals. Before long, a large male is isolated from the rest of the group, bouncing at full speed in a direction parallel to that of your vehicle. The animal is perfectly positioned, perpendicular to your vehicle. You take your video-camera out and point it toward the animal through the side window. For sure, the animal is bouncing up and down: that's how kangaroos "run". But as you put the camera's viewfinder to your eye, the animal appears to be bouncing even more. That's because the bounces of your camera are being added to the bounces of the kangaroos. The total (apparent) bounces are made of two parts: the intrinsic or true bounces of the kangaroo, and the extrinsic (or virtual) bounces of the camera. If you try to measure the true height of a bounce of a given kangaroo, you must factor out the bounces of the camera/instrument doing the measurement. Likewise, all measurements contain errors. As in our kangaroo example, some error components represent true variation, whereas other components are linked to the observer and add to the noise.

Errors in Measuring Milk Yield and Body Weight

Let's take the task of measuring milk yield and body weight in a herd as examples. What are the different types of errors?

Physical measurement errors

Even if the total milk production from a given cow at last milking is put in a milk can, which is weighed, we still don't know the exact value of her milk production. The precision of our measurement is determined by the precision of the measuring instrument. We might be able to say that her production was above 41 lb and less than 42 lb, or that it was "around" 41.2 lb, but we likely will not have a scale with a precision down to ±1 μg. For management purposes, we know that this degree of precision is not needed, but that doesn't negate the fact there is, and always will be, some measurement errors. This type of errors is what people working in the physical sciences generally have to deal with. In measuring the diameter of (say) a bolt, ever more precise measuring devices can be used, but an error always remains. Repeated measurement of the same bolt produces slightly different measurements (data) of the diameter of the bolt, but this is not due to the one bolt changing its diameter. This type of error is entirely due to the observer (the camera in our kangaroo example). The physical error involved in measuring milk yield may or may not be of practical consequence; this depends on the precision of the measuring device. But for other types of data acquired on a dairy farm, the measurement error may be consequential. Think of the scales on mixer wagons; typically, what is their precision? Most of the scales I have worked with have a precision of ±10 lb. Hence if the scale indicates that 320 lb of supplement were added to the mixer, the correct interpretation would be that somewhere between 310 and 330 lb of the supplement were added. This error may no longer be insignificant. Likewise, most livestock scales also have an error of ±10 lb. Say that you weigh your animals once per day (automatic scale on the return alley). The weight on a given animal will have a physical measurement error of ±10 lb each time the animal is weighed. If Bertha weighed 1,350 (±10 lb) yesterday and 1,340 (±10 lb) today, I cannot conclude that she has lost 10 lb in 1 day! To gain precision would require a more precise scale.

Unfortunately, data are often being reported with considerably more digits than what is warranted by the precision of the instrument

(or method) generating the data. There is no reason to report in vitro digestibility of a single forage sample with 2 decimal digits when the error of the method is somewhere around ±5%.

Variation in the unit itself

In the physical sciences, what is being measured doesn't change: the diameter of the one bolt being measured doesn't change (we ignore the effect of changing temperature, etc., for the sake of simplicity here). But biological entities keep changing through time. We know that we have a physical measurement error when we measured Bertha's milk production yesterday morning and that we also had physical measurement errors when we measured her milk production this morning. But these errors generally pale in comparison to the size of the variation (errors) due to the cow (the biological unit). On a well-managed farm, the standard deviation (**SD**) of daily milk yield on the same animal over a period of one week is generally in the 7 lb/day range. So if Bertha produced 100 lb yesterday and 95 lb today, we really cannot say that she is down 5 lb/day in production. Our data say that the amount measured today was 5 lb less than the amount measured yesterday, but we really cannot say much about the production status of Bertha.

The total weight of a cow is the sum of her true physical weight (generally expressed as empty body weight) plus all of her gut and bladder contents (plus milk in her udder, which technically is no longer part of her body – at least for some of the milk). Graduate students doing their first digestion trial with full collection of feces and urine are always amazed at the amount of feces and urine that an average cow excretes in a day (roughly 150 lb). Whenever Bertha drinks, she easily gains 10 lb (she easily drinks 200 to 250 lb/day of water). Whenever she defecates, she looses 10 lb. All of a sudden, the error with a one point in time measurement of her body weight is no longer just the precision of the scale (±10 lb), but also the variation of Bertha's apparent weight (±10 lb), which is not really Bertha's true weight to begin with.

Variation among biological units

Whenever someone says that the cows in the first pen are producing (say) 90 lb/day, they really mean that they are *averaging* about 90 lb/day. Of course, by now we understand that the 90 lb is an approximation because of the measurement error and the variation within a cow. But there is more error than that when we look at milk production for a given pen. In that pen, there are some Berthas producing over 110 lb/day, while other Berthas are below 70 lb/day. Hence, the pen contains at a minimum the sum of the errors of all the animals it contains. Fortunately, some of these errors cancel each other. *If* all the cows in a 100-cow pen were *independent* and *if* the daily SD on each animal is 7 lb/day, then we would expect the SD of milk production for the whole pen to be $7/\sqrt{100} = 0.7$ lb/day. In practice, the SD of milk production from a 100-cow pen is always greater than that; sometimes much greater than that. The reason being that cows within a pen are not independent of one another. A pen of cows is not milked at exactly the same time every day of the week: "things" happen… If the waterline to the high pen froze during the night, pretty much all Berthas in that one pen will be down in milk the next day, and they will all have lost apparent weight. A frozen pipe is an easily identifiable factor, but there are 50,000 things that affect cow production – some of which are known, whereas others are not. As nutritionists, we tend to see things through the glasses of nutrition. When we investigate the cause for an apparent drop in milk production, we tend to focus on nutrition because that's what we do and sometimes sell. I am afraid, however, that all too

often, we fix non-existent problems by changing the nutrition (we cure a non existent "disease"), or we fix a self-curable, non-nutritional problem by changing the nutrition. The analogy that I have used is that of a kid coming back from school not feeling well. You give him a teaspoon of a magic elixir and send him to bed. He feels great the next morning. Maybe the kid was just dead tired!

The issue with the pen variation is that too many people deny that it even exists. This is exactly what is being implied by anybody who conducts "field research" with one pen fed a control diet and one pen fed the "treatment" diet, and uses the individual cows as experimental units, as if they were independent of one another. How often have you heard "the 2 pens were identical"? Well, if they were, the variation between replicated pens would be very, very small and an experiment with 2 pens on a control diet and 2 pens on a treatment diets should detect statistically significant differences that would be incredibly small. Four pens of 100 cows each would detect differences in milk yield of less than 0.1 lb/day in a 60 day production trial. To my knowledge, every time that replicated pens have been correctly used in a field trial, the power of the experiment was considerably less than one would expect if the pen had a small effect. The pen effect is real and considerably greater than what most people think. Hence, the so-called experiments with one pen per treatment lead to pure statistical fantasy, vastly incorrect P-values, amounting to a momentous drive through a big cloud of smoke. In fact, one would suspect that the experimenters may have inhaled too much of the smoke themselves. Reporting wrong probabilities is far worse than not reporting any probability at all.

Errors in Feed Composition

Suppose that you receive 5 loads of distillers dried grains with solubles (**DDGS**). You sample each one and send the 5 samples to a feed laboratory and request a neutral detergent fiber (**NDF**) assay. Results for the 5 samples are: 27, 29, 30, 31, and 33% NDF. The mean (arithmetic average, which represents the expectation) of the 5 samples is 30%. But how would you express the variation between the 5 samples? One could use the range, which is the difference between the maximum and the minimum values. In our example, the range is 33 – 27 = 5%. This gives an idea of the variation, but it is sensitive to only the 2 extreme values: the NDF content of the 3 other samples are not part of the variation assessment. This makes the range very sensitive to outliers, which is not a good thing. Statisticians have long used a measure of variation that expresses the spread of observations in a manner that is less sensitive to outliers than the range and also uses all the measurements in its determination. This is the variance. Expressed in words, it is the sum of the squares of the difference between each measurement and the mean, with said sum divided by the total number of observations minus one. For our simple example:

$$Var(NDF) = \frac{(27-30)^2 + (29-30)^2 + (30-30)^2 + (31-30)^2 + (33-30)^2}{5-1} = 5$$

The immediate issue one has with the variance is with its units of expression: the result (i.e., 5) is not in the same units as the measurements. The variance is not 5% because the expression is a sum of squared percentages: it is 5%-squared, a highly inconvenient, if not completely mentally, intractable unit. The solution is simply to take the square root of the variance (i.e., the SD). Hence, in our example:

$$SD(NDF) = \sqrt{5} = 2.23$$

The standard deviation is expressed in the same units as the measurements themselves. Therefore, we can summarize the results from our 5 samples as: mean = 30%, SD = 2.23%.

If the sample is moderately large and follows what is known as the normal distribution (the infamous bell-shaped curve), then approximately 2/3 of the observations will be within +/- 1 SD of the mean, and approximately 95% will be within +/- 2 SD of the mean. The sample in our example was not very large (n = 5), but we would expect that approximately 67% of all loads of DDGS to have an NDF between 30 – 2.23 = 27.8% and 30 + 2.23 = 32.2%.

The reason that we examined variance as an expression of variation is because we want to decompose the total variation in measurements into separate components. However, these components are additive only in the variance scale and not in the standard deviation scale. As long as the components (factors) are independent, we have:

$$\text{Var(whole)} = \text{Var(factor A)} + \text{Var (factor B)} + \ldots$$

but

$$\text{SD(whole)} \neq \text{SD(factor A)} + \text{SD(factor B)} + \ldots$$

This will be important in our understanding of the factors contributing to the variances of various feedstuffs and their partitioning into components.

Sources of variation in forages: short-term

Over a total of 14 consecutives days, we had nutritionists and trained farm personnel take multiple samples of corn and haycrop silages on 14 Ohio and Vermont farms (St-Pierre and Weiss, 2015). To be more precise, the sampler took 2 independent samples of each forage on each farm and on each day of the 14-day sampling period. At the lab, each assay was run in duplicate on each of the 2 samples from each silage, from a given farm on a given day. This elaborate sampling scheme allowed us to separate the total variance of a given nutrient into 4 distinct components. First is the variation due to farm. This component is easy to understand: it represents how much the true values of a given nutrient (e.g., NDF) in corn and haycrop silages vary from farm to farm. The second component is the variation due to day. This component quantifies how much the true values of a given nutrient in corn and haycrop silages vary from day to day on a given farm. The third component quantifies the variation due to sampling, or how much the true value of a given nutrient in corn or haycrop silages varies from sample to sample taken on the same farm and on the same day. The fourth and last component is labeled 'analytical' and identifies the variation within a single lab of a given assay on a set sample. One should note that all assays were conducted in our research lab. Hence, the variation that would exist between assays conducted on the same sample but by different labs was not present in our study. Depending on the assay, the lab-to-lab variation can be substantial. Some components of this variance decomposition are real (intrinsic): the variation due to farm and day represent true variation in the composition of the forage. Other components are extrinsic (virtual): the variation due to sampling and analytical variance are part of the noise inherent to the measurements. The variation due to laboratory would be added to the virtual components had it been measured. As in the Australian safari analogy, the virtual components do not contribute to the true feed variation, but they add to the total, overall perception of variation. Conceptually, all the variation could be in the virtual components, indicating a perfectly uniform feed that appears to vary just because of the errors in the multiple steps of the measurements: the kangaroo would be completely still, but the binoculars are bouncing all over the place.

Table 1 summarizes the measurements for 4 nutrients/chemical groups for corn and

haycrop silages. The means are close to what one would find in standard tables of feed composition, such as those of NRC (2001). The table also shows that there is considerable variation around the means. The coefficient of variation (**CV**) is simply the ratio of the SD divided by the mean, multiplied by 100 to have a metric expressed as a percentage. For corn silage, the CV of DM (14.1%) is about the same as the CV of ash (14.0%). But the variance components are very different between the two. Table 2 shows the decomposition of the variance for the 2 feeds and the 4 nutrients/chemicals. The only components relevant to a given farm are the variation due to day, sampling, and analytical. How much a forage varies from farm to farm doesn't affect how much it varies on my farm. Although the CV of DM and ash in corn silage are nearly identical, the sources of the variation are entirely different. For a given farm, nearly half (46.8%) of the DM variance is from day-to-day (true variation). Hence, DM of corn silage should be measured frequently. The situation is entirely different for ash where only 17.8% of the variance is due from day-to-day variation and over 80% is due to sampling (i.e., noise). Therefore, frequent measurements of ash in corn silage would be mostly a waste of time and money.

For all nutrients and both silages, analytical variation was the smallest contributor to variation within a given farm. In general, the SD for assay were substantially greater for corn silage than haylage. This can be explained by the greater degree of difficulty in sub-sampling corn silage in the lab due to the large differences in nutrient composition among particles of corn silage. On-farm sampling was the greatest source of variation for nutrients other than DM. The substantial amount of observer variation (sampling + analytical) relative to the amount of true day-to-day variation indicates the followings:

1. A change in nutrient composition between 2 samples of silages taken over a short period of time is often just noise (i.e., not true). Modifying the diet on an apparent change in composition from one forage sample is generally not wise. Most of the time you will be driving through smoke!

2. Much progress would be made in controlling variation in forage composition if a minimum of 2 *independent* samples were taken and sent to the lab any time that the forage is sampled. Here, the word *independent* is critical. The whole sampling process has to be repeated.

Sources of variation in forages: long-term

In a parallel study to the one we have already briefly described, 14 nutritionists took monthly samples of every major feed (forage and concentrates) and high group TMR on 47 farms throughout the United States (St-Pierre and Weiss, 2015). Results for the mean composition and variance components of the forages that were sampled and for the major nutrients are reported in Table 3. As expected, monthly variation was greater than daily variation. Although sampling variation generally made up a smaller proportion of within-farm variation when silages were sampled over a 12-mo period than when sampled over a 2-wk period, sampling variation was still substantial. This suggests that duplicate, independent samples and averaging is beneficial for many different sampling schedules.

Conclusions

A number is meaningless unless you have quantified its error.

Acknowledgement

I want to acknowledge the considerable contribution that my colleague Dr. Bill Weiss made in stimulating some of the thoughts and generating most of the data used in this paper. In addition, portions of this paper were extracted from a chapter of an electronic book soon to be published by the American Dairy Science Association (Champaign, IL).

References

National Research Council. 2001. Nutrient requirements of dairy cattle. 7th rev. ed. Natl. Acad. Press, Washington, DC.

St-Pierre, N.R., and W.P. Weiss. 2015. Partitioning variation in nutrient composition data of common feeds and mixed diets on commercial dairy farms. J. Dairy Sci. 98:5004-5015.

Table 1. Descriptive statistics for corn silage and haycrop silages sampled over 14 consecutive days on 11 Ohio and Vermont farms (% of DM) (St-Pierre and Weiss, 2015).

Item	Mean	SD	CV	Range	10th to 90th Percentile
Corn silage (n = 504)					
DM	37.0	5.23	14.2	26.2-49.0	30.3-44.0
NDF	39.1	4.03	10.3	30.8-50.9	34.3-45.0
Starch	32.8	4.33	13.2	14.3-44.0	27.1-38.6
Ash	3.57	0.50	14.0	2.4-8.3	3.1-4.2
Haycrop silage (n = 504)					
DM	41.7	8.00	19.2	28.3-70.5	31.9-52.1
NDF	49.9	6.61	13.2	32.7-65.2	43.0-59.5
CP	16.3	2.72	16.7	10.8-23.2	12.7-19.5
Ash	9.30	1.86	20.0	6.3-15.3	6.9-11.6

Table 2. Farm, sampling, analytical, and true day-to-day variation in nutrient composition (% of DM) of corn silage and haycrop silage sampled over 14 consecutive days on 11 Ohio and Vermont farms (St-Pierre and Weiss, 2015).

	SD				% of within-farm variance		
Item	Farm	Day	Sampling	Analytical	Day	Sampling	Analytical
Corn silage							
DM	5.00	1.21	0.96	0.86	46.8	29.5	23.7
NDF	3.68	1.31	1.61	0.89	33.6	50.9	15.5
Starch	4.22	1.29	2.10	0.82	24.7	65.3	10.0
Ash	0.34	0.16	0.34	0.05	17.8	80.6	1.6
Haycrop silage							
DM	7.37	2.71	1.89	0.74	64.0	31.2	4.8
NDF	6.97	1.67	1.61	0.75	46.9	43.6	9.5
CP	2.47	0.59	0.89	0.44	26.1	59.4	14.5
Ash	1.86	0.33	0.59	0.10	23.4	74.7	1.9

Table 3. Mean composition and estimates of total, farm-to-farm, and within-farm variation (i.e., residual) for various forages on 47 farms over a period of 12 months (St-Pierre and Weiss, 2015).

			Standard Deviations		
	Mean	10-90%tile	Total	Farm	Residual
Corn silage (n = 627)					
DM, %	34.1	29.8 – 38.7	3.70	2.54	2.83
CP, %	8.00	7.0 – 8.9	1.03	0.84	0.60
NDF, %	40.8	36.3 – 46.4	4.23	3.81	2.66
Ash, %	4.3	3.1 – 6.6	1.46	1.29	0.62
Legume hay (n = 263)					
DM, %	88.4	85.3 – 91.6	3.37	2.04	2.80
CP, %	21.4	18.5 – 24.4	2.41	1.51	2.00
NDF, %	36.7	31.0 – 43.3	5.03	3.45	3.92
Ash, %	10.9	9.1 – 12.8	1.95	0.95	1.74
Legume silage (n = 453)					
DM, %	44.2	32.8 – 55.6	9.12	7.74	6.22
CP, %	21.7	19.0 – 24.2	1.99	1.10	1.68
NDF, %	40.0	34.7 – 45.8	4.55	3.02	3.51
Ash, %	11.1	9.3 – 13.4	1.80	1.38	1.22
Mixed hay (n = 41)					
DM, %	86.1	82.7 – 88.4	2.95	2.26	2.37
CP, %	15.2	10.2 – 19.6	3.39	2.61	2.69
NDF, %	54.8	48.2 – 61.9	5.83	0	5.83
Ash, %	8.7	7.2 – 10.4	1.72	1.13	1.43
Mixed silage (n = 101)					
DM, %	43.5	31.2 – 58.5	10.45	8.63	7.80
CP, %	18.1	15.6 – 20.3	1.92	0.97	1.70
NDF, %	48.3	43.2 – 53.8	4.67	2.81	4.00
Ash, %	9.7	8.5 – 11.2	1.27	0.77	1.04
Small grain silage (n = 94)					
DM, %	35.2	29.8 – 40.3	7.08	9.20	3.47
CP, %	12.8	9.1 – 17.2	3.17	3.32	1.63
NDF, %	53.9	46.1 – 63.1	6.56	5.82	3.73
Ash, %	12.8	10.1 – 15.0	3.73	3.13	2.91
Straw (n = 127)					
DM, %	88.0	84.0 – 92.3	5.14	1.44	4.95
CP, %	4.8	3.4 – 6.6	1.48	0.43	1.42
NDF, %	78.7	72.1 – 83.1	5.06	3.83	3.75
Ash, %	7.6	4.8 – 11.6	2.84	2.57	1.47