Harvard Data Science Review • Issue 1.2, Fall 2019

Data Have a Limited Shelf Life

Stephen M. Stigler[1]

[1]**Department of Statistics, Physical Sciences Division, University of Chicago, Chicago, Illinois, United States of America**

The MIT Press

Published on: Nov 30, 2020

DOI: https://doi.org/10.1162/99608f92.f9a1e510

ABSTRACT

Data, unlike some wines, do not improve with age. The contrary view, that data are immortal, a view that may underlie the often-observed tendency to recycle old examples in texts and presentations, is illustrated with three classical examples and rebutted by further examination. Some general lessons for data science are noted, as well as some history of statistical worries about the effect of data selection on induction and related themes in recent histories of science.

Keywords: dead data, zombie data, post-selection inference, history

1. Introduction

We commonly encounter repeated use of the same data sets in statistical exposition; that is, in textbooks, in lectures, and as examples in theoretical papers. These data sets may be termed classical, even though they may be of recent origin. They can supply a link to reality without the need of a lengthy explanation. But how strong is that link? I will argue that data have a limited shelf life. To see what data look like after too long a time on the shelf, see Figure 1.

Figure 1. The Hyrtl Skull Collection. Reproduced with permission from The Mütter Museum of the College of Physicians of Philadelphia.

Put bluntly, old data may be dead data, in both statistical exposition and in science more generally. I hasten to add that I do not mean that all old dead data are without value. Often these museum pieces are extremely valuable, just like the great art of our other museums. But the value may be quite different from that of a new and unanalyzed set of scientific data produced to help cope with a problem that is often not yet well-posed. Often old data are no more than decoration; sometimes they may be misleading in ways that cannot easily be discovered.

I will present three examples of classical data sets used for statistical exposition, to show both the problems and the benefits, and I will end with some implications going beyond classical data sets, highlighting general concerns for the era of big data in science.

2. Three Examples From Classical Data Sets

I will define a classical data set as a set of data that has been collected for some scientific or commercial purpose, and has been employed for instruction or exposition by several people. Generally, the data are ‘full’ in some sense (there is no immediate prospect for more data—questions that arise may not be dealt with by simply getting more data).

Mesures de la circonférence des poitrines des soldats écossais.

MESURES de la POITRINE.	NOMBRE d'hommes.	NOMBRE PROPORTIONNEL.	PROBABILITÉ d'après L'OBSERVATION.	RANG dans LA TABLE.	RANG d'après le CALCUL.	PROBABILITÉ d'après LA TABLE.	NOMBRE D'OBSERVATIONS calculé.
Pouces.							
33	3	5	0,5000			0,5000	7
34	18	31	0,4995	52	50	0,4993	29
35	81	141	0,4964	42,5	42,5	0,4964	110
36	185	322	0,4823	33,5	34,5	0,4854	323
37	420	732	0,4501	26,0	26,5	0,4531	732
38	749	1305	0,3769	18,0	18,5	0,3799	1333
39	1073	1867	0,2464	10,5	10,5	0,2466	1358
			0,0597	2,5	2,5	0,0628	
40	1079	1882	0,1285	5,5	5,5	0,1359	1987
41	954	1628	0,2913	13	13,5	0,3034	1675
42	658	1148	0,4061	21	21,5	0,4130	1096
43	370	645	0,4706	30	29,5	0,4690	560
44	92	160	0,4866	35	37,5	0,4911	221
45	50	87	0,4953	41	45,5	0,4980	69
46	21	38	0,4991	49,5	53,5	0,4996	16
47	4	7	0,4998	56	61,8	0,4999	3
48	1	2	0,5000			0,5000	1
	5738	1,0000					1,0000

Figure 2. From Quetelet (1846). Columns 1 and 2 give chest circumference (in inches) and frequency count for 5,738 Scottish soldiers.

My first example may even be the first such data set in the history of statistics. It comes well after a number of earlier data sets, such as John Graunt's 1662 bills of mortality and Gauss's 1803 data on the orbit of the asteroid Ceres, which do not meet my conditions. The example that does is Quetelet's data on the chest circumference of Scottish soldiers (Figure 2). These were first published in this form by the Belgian statistician Adolphe Quetelet in an 1846 textbook to illustrate his method for fitting a normal distribution to grouped data, and they appeared in many later publications by him and others (Quetelet, 1846). The data are the frequency counts in the second column, representing the distribution of 5,738 Scottish soldiers' chest measurements, from the

smallest at 33 inches to the largest at 48 inches. Since the Scots of that era were generally short, that largest measure may be of someone approaching spherical shape.

The second is better known; it is von Bortkiewicz's data on deaths by horse kick in the Prussian Cavalry (Figure 3). It was first published in 1898, and has appeared in many subsequent publications (Andrews & Herzberg, 1985; von Bortkiewicz, 1898; Winsor, 1947). It shows the number of deaths in this dramatic fashion, classified by year (from 1875 to 1894) and by Cavalry Corps. They are most commonly viewed as an example of how well the Poisson distribution fits the occurrence of what must be considered a rare event.

	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
G	—	2	2	1	—	—	1	1	—	3	—	2	1	—	—	1	—	1	—	1
I	—	—	—	2	—	3	—	2	—	—	—	1	1	1	—	2	—	3	1	—
II	—	—	—	2	—	2	—	—	1	1	—	—	2	1	1	—	—	2	—	—
III	—	—	—	1	1	1	2	—	2	—	—	—	1	—	1	2	1	—	—	—
IV	—	1	—	1	1	1	1	—	—	—	—	1	—	—	—	—	1	1	—	—
V	—	—	—	—	2	1	—	—	1	—	—	1	—	1	1	1	1	1	1	—
VI	—	—	1	—	2	—	—	1	2	—	1	1	3	1	1	1	—	3	—	—
VII	1	—	1	—	—	—	1	—	1	1	—	—	2	—	—	2	1	—	2	—
VIII	1	—	—	—	1	—	—	1	—	—	—	—	1	—	—	—	1	1	—	1
IX	—	—	—	—	—	2	1	1	1	—	2	1	1	—	1	2	—	1	—	—
X	—	—	1	1	—	1	—	2	—	2	—	—	—	—	2	1	3	—	1	1
XI	—	—	—	—	2	4	—	1	3	—	1	1	1	1	2	1	3	1	3	1
XIV	1	1	2	1	1	3	—	4	—	1	—	3	2	1	—	2	1	1	—	—
XV	—	1	—	—	—	—	—	1	—	1	1	—	—	—	2	2	—	—	—	—

a) Man kann im gegebenen Fall zunächst einmal genau in derselben Weise verfahren wie in den beiden vorangehenden. Man findet:

Jahres-ergebnis	Zahl der Fälle, in denen das nebenstehende Jahresergebnis eingetreten ist	zu erwarten war
0	144	143,1
1	91	92,1
2	32	33,3
3	11	8,9
4	2	2,0
5 u. mehr	—	0,6

$$\{\varepsilon_0'(x)\}^2 = 0{,}70\ (0{,}05); \qquad \{\varepsilon_0''(x)\}^2 = 0{,}73\ (0{,}09);$$

$$\varepsilon_0'(x) = 0{,}84\ (0{,}03); \qquad \varepsilon_0''(x) = 0{,}85\ (0{,}05).$$

Figure 3. Data on deaths by horse kick in 14 Prussian Cavalry Corps over 1875–1894, from Bortkiewicz (1898). The tabulation at the bottom compares the frequency of deaths per Corps with a fitted Poisson frequency distribution (von Bortkiewicz calculated the expected frequencies for each of the Corps separately and added them up).

The third example is more recent. It is what is commonly referred to as the Cushny-Peebles data, as published by William Sealy Gosset in the 1908 *Biometrika* article that introduced Student's *t* test, and it was the first and most visible application of that test (Figure 4). The data purport to give for 10 patients the additional time they slept under each of two different sleeping potions, and Gosset applies the *t* test to the 10 paired differences to find, as we would write today, $t = 4.06$ with nine degrees of freedom: a statistically significant difference. Gosset would be criticized today for misinterpreting the *p* value as a posterior probability, saying "the odds are

about 666 to 1 that [the second drug] is the better soporific" (Gosset, 1908, p.21). The data were reprinted by R. A. Fisher in his 1925 book that brought the *t* test to wide attention (quietly avoiding Gosset's misstatement about *p* values), and they continued to appear there through 14 editions over 40 years, and in many other places.

Illustration I. As an instance of the kind of use which may be made of the tables, I take the following figures from a table by A. R. Cushny and A. R. Peebles in the *Journal of Physiology* for 1904, showing the different effects of the optical isomers of hyoscyamine hydrobromide in producing sleep. The sleep of 10 patients was measured without hypnotic and after treatment (1) with D. hyoscyamine hydrobromide, (2) with L. hyoscyamine hydrobromide. The average number of hours' sleep gained by the use of the drug is tabulated below.

The conclusion arrived at was that in the usual dose 2 was, but 1 was not, of value as a soporific.

Additional hours' sleep gained by the use of hyoscyamine hydrobromide.

Patient	1 (Dextro-)	2 (Laevo-)	Difference (2–1)
1.	+ ·7	+1·9	+1·2
2.	−1·6	+ ·8	+2·4
3.	− ·2	+1·1	+1·3
4.	−1·2	+ ·1	+1·3
5.	−1	− ·1	0
6.	+3·4	+4·4	+1·0
7.	+3·7	+5·5	+1·8
8.	+ ·8	+1·6	+ ·8
9.	0	+4·6	+4·6
10.	+2·0	+3·4	+1·4
	Mean + ·75	Mean +2·33	Mean +1·58
	S. D. 1·70	S. D. 1·90	S. D. 1·17

First let us see what is the probability that 1 will on the average give increase of sleep; i.e. what is the chance that the mean of the population of which these experiments are a sample is positive. $\frac{+\cdot 75}{1\cdot 70} = \cdot 44$ and looking out $z = \cdot 44$ in the

Figure 4: From Gosset (1908)

The three examples have much in common. They all help tell a story, and the story seems clear to any listener. We all can imagine the collection of Scotsmen of different shapes and sizes and understand that demonstrating the measurements are normal can stand as at least weak confirmation that they can be considered as a statistical group, conceivably to be compared to French or Belgian soldiers.

We all can vividly imagine a cavalry officer carelessly walking behind a horse at exactly the wrong moment, perhaps due to drink or simply inexperience, and we can marvel how such a purely random occurrence can, in the aggregate, so closely fit such an exact distribution as the Poisson. Neither of these examples requires scientific expertise or more context than a simple description. The third example is a bit more technical with

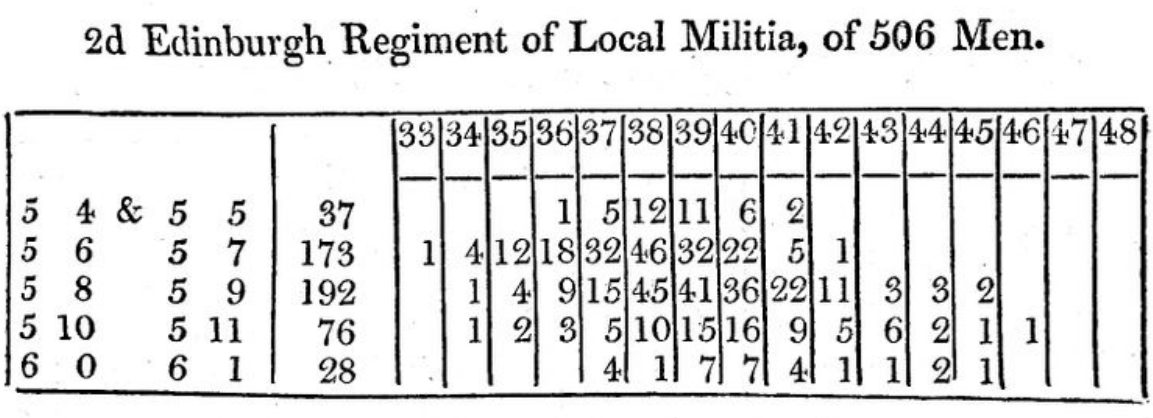
the fancy names for the drugs, but the idea of comparing two sleeping potions is clear and it gives life and motive to what would otherwise be a dull numerical exercise.

3. The Three Examples Examined

All three seem ideal ways to engage and explain statistical ideas as classical data sets are supposed to. But are they really ideal? Let us ask the questions that should always be asked when presented with data, namely, where did these data come from? Why and how were they collected? What decisions led to them being presented in this way? Why the interest in Scottish chests? In Prussian horses? In these two drugs?

It happens that in these cases, all of these questions can be answered, and the answers can greatly change the way we view the examples.

Quetelet tells us in his textbook where he got the data, from an 1817 Edinburgh medical journal (Quetelet, 1817). And there we find a surprise. The data there were gathered by an army contractor charged with making uniforms, and they are much more interesting than what Quetelet presented. The contractor gave cross-classified data on both chests and height separately for 11 different militias (Figure 5). To a later eye the pattern of association of chests and heights, the elliptical shape showing the correlation that so excited Francis Galton in the 1880s when he discovered that concept, jumps out to the viewer.



2d Edinburgh Regiment of Local Militia, of 506 Men.

						33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
5	4	&	5	5	37				1	5	12	11	6	2							
5	6		5	7	173	1	4	12	18	32	46	32	22	5	1						
5	8		5	9	192		1	4	9	15	45	41	36	22	11	3	3	2			
5	10		5	11	76		1	2	3	5	10	15	16	9	5	6	2	1	1		
6	0		6	1	28					4	1	7	7	4	1	1	2	1			

Figure 5. Frequency data for one of the 11 militias, from Edinburgh (1817). Height (left margin) in feet and inches, grouped; chest (top margin) in inches.

To be sure, the data are truncated by height, since the militias did not accept the very short or very tall. Quetelet showed no interest in the height. He wanted untruncated data and for that he needed the marginal totals by chest circumference. The contractor gave marginal totals for height by militia and overall, but not for chests. Quetelet had to work to get his data. He computed the needed combined marginal totals for the 11 militias, and, all in all, not too badly. Most of his totals were wrong, but not so much as to upset his conclusion via an eyeball test of fit that the Scots were normal. The grand total of his totals, 5,738, was closer to the total computed from the contractor's data, 5,732, than were many of the individual numbers (see Table 1). Quetelet

gave a wrong version of the data, and he missed the chance to investigate if there was a difference between militias, not to mention discover correlation, but neither upset his basic purpose of illustrating his method. For more on Quetelet and these data, see Stigler (1986, chapter 5, esp. pp. 203–220.)

Table 1. Comparison of Quetelet's marginal totals with those derived from the original publication.

Inches	1817 data	Quetelet
33	3	3
34	19	18
35	81	81
36	189	185
37	409	420
38	753	749
39	1062	1073
40	1082	1079
41	935	934
42	646	658
43	313	370
44	168	92
45	50	50
46	18	21
47	3	4
48	1	1
Total	5732	5738

von Bortkiewicz's data turn out to have come from about 100 large volumes of Prussian state statistics, published over 20 years. Each year's data comprised at least three volumes for a total of well over 1,000 pages. Buried in one of each year's later volumes were the military statistics, and buried in those are the data on cavalry deaths for the year, in a large two-page spread. One line of that table, number 17, gave deaths by horse kick separately for 14 Corps, both for members of cavalry and "over all," presumably including other personnel (Figure 6).

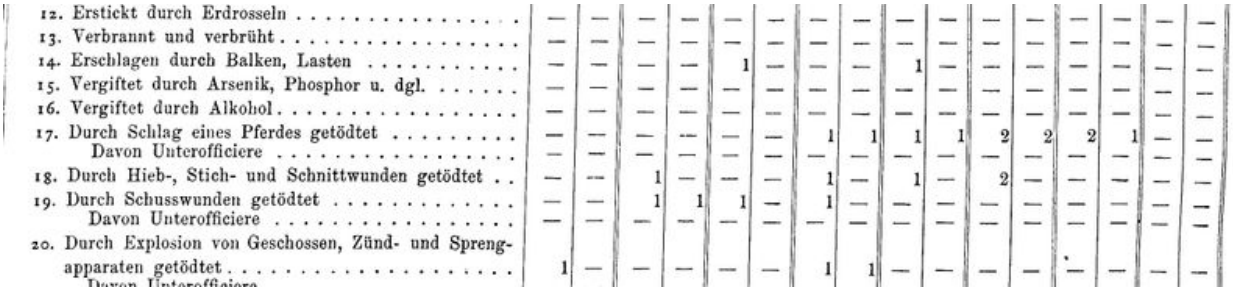


Figure 6. Part of Line 17 from the 1879 Prussian State Statistics (*Prüssische Statistik, Vol. 60*).

von Bortkiewicz had gone through these 100 volumes and selected the single statistic for death by horse kick, using the "over all" figure. He had avoided several other attractive categories, including deaths by falling from a window, from arsenic, falling from a horse, and being struck by lightning. Or did he also try those categories and reject them in favor of horse kicks? He didn't say. The end result was a highly selected data set, selected with a particular purpose in mind. He wanted to demonstrate what he called "the law of small numbers," to illustrate that any real differences (such as between years and between Corps of different sizes) would, for rare occurrences, be masked by the large Poisson variation. Ironically, the data he presented can in modern hands be used to demonstrate a converse of this. An analysis by generalized linear models will detect both Corps and year differences.

The story regarding the Cushny-Peebles data is much more complicated. Gosset gave the source as a 1904 article in the *Journal of Physiology*, and there was indeed an article by the two authors that year, but the actual source was a follow-up article in the 1905 volume. In 1905 Arthur Cushny, a Scotsman, moved to University College London after 12 years at the University of Michigan in Ann Arbor, where he had worked with a medical student, A. R. Peebles, and where the work was done. The data they presented came from a study performed at the Michigan Asylum for the Insane in Kalamazoo. That asylum changed its name in 1911 and is now the Kalamazoo Regional Psychiatric Hospital, currently the largest mental health institution in Michigan. (For an excellent discussion of these data, see Senn & Richardson, 1994; also Preece, 1982, and Zabell, 2008).

According to Cushny and Peebles, the study was conducted at the asylum by three doctors Cushny had worked with, and the drugs in question were being administered as "hypnotics" in that ward to calm the patients. The principal drug was supplied for the study by the company Merck, which was apparently supporting the work in

some way. Worries about potential conflicts of interest had not yet arisen in 1905. The experiment, such as it was, included 10 patients who received all four treatments: a control and three different drugs and one patient (number 11) for whom there was no control. There were several replications, varying from three to six for the drugs. After a drug administration there would be a night with no drug; those nights served as the "controls," apparently assuming no carryover effect. And, of course, in that era there was no randomization, blinding, or informed consent.

	Controls (no hypnotic)		0·6 mg. L-Hyoscyamine HBr.			0·6 mg. L-Hyoscine HBr.			0·6 mg. R-Hyoscine HBr.		
Patient	No. of observations	Average hours of sleep	No. of observations	Average hours of sleep	Increase over controls	No. of observations	Average hours of sleep	Increase over controls	No. of observations	Average hours of sleep	Increase over controls
1	9	0·6	6	1·3	0·7	6	2·5	1·9	6	2·1	1·5
2	9	3·0	6	1·4	−1·6	6	3·8	0·8	6	4·4	1·4
3	8	4·7	6	4·5	−0·2	6	5·8	1·1	6	4·7	0·0
4	9	5·5	3	4·3	−1·2	3	5·6	0·1	3	4·8	−0·7
5	9	6·2	3	6·1	−0·1	3	6·1	−0·1	3	6·7	0·5
6	8	3·2	4	6·6	3·4	3	7·6	4·4	3	8·3	5·1
7	8	2·5	3	6·2	3·7	3	8·0	5·5	3	8·2	5·7
8	7	2·8	6	3·6	0·8	6	4·4	1·6	5	4·3	1·5
9	8	1·1	5	1·1	0·0	6	5·7	4·6	5	5·8	4·7
10	9	2·9	5	4·9	2·0	5	6·3	3·4	6	6·4	3·5
11	—	—	2	6·3	—	2	6·8	—	2	7·3	—

Figure 7. From Cushny and Peebles (1905).

The table (Figure 7) gives the average hours of sleep in each case and the "increase over controls" for the drugs. Without benefit of a *t* test, Cushny and Peebles judged one of the drugs ineffective and the other two (including the Merck drug) equally effective. Gosset copied the "increase over control" column for only the ineffective drug and the Merck drug (with one typo for patient number 5 that did not affect his calculations). He ignored patient number 11. He mislabeled both columns, both now listed as drugs that were in fact not involved in this trial. His two columns were then correlated, not only by patient but also by both being differences from the same control, although when he analyzed only the difference between the two columns that would not have had much effect. But the data used were means based upon different sample sizes (3 to 6) and this would have led to different patient variances.

In all editions up to 1935 Fisher reproduced the data with Gosset's errors, but in January of that year an NYU researcher named Isidor Greenwald wrote to Fisher calling attention to the labeling errors. When Fisher sent the letter to Gosset, Gosset replied to Fisher, "That blighter is of course perfectly right and of course it doesn't really matter two straws." Fisher altered later editions, calling the drugs simply "A" and "B," and at that point the example became strictly a numerical, not a scientific example.

4. The Vitality of Data

In all three cases the example was grounded in real data. In all three cases, the closer you examine the situation, the more complicated it becomes. There was in each case a large degree of selection of data that was unreported. There are errors of transcription and commission. There are doubts about important aspects that cannot be answered at this point in time. And there is the unescapable conclusion that each data set had become only ornamental at an early stage of its history. There are gains from such reexamination of old data—the story can become much richer statistically (as with von Bortkiewicz). And there are losses—we find the greatest statistician of the 20th century was content to sweep all problems away under the labels A and B. But after this investigation it is not possible to view them in the same way as before. Scientifically, they are dead data.

These are examples; they are themselves a selected sample. But in many years of following up on sources, I have found such cases of dead data to be more the rule than the exception. Some cases have been more egregious, such as a popular text that in one edition presented some numbers (integers from 1 to 6) as the results of throwing dice; in another edition for business students the same numbers became the sales of men's suits. Some were more benign, as the passage of time had simply rendered the example unconvincing: a 1958 argument by Joseph Berkson based on observational studies undermined the case against cigarettes as a cause of lung cancer by showing that the elevated risk of lung cancer was matched by an elevated risk of many other diseases where smoking was not a suspected cause. Berkson may have legitimately pointed to other environmental and social causes of lung cancer at the time of that study, but the hypothesized effect of smoking has by now broadened well beyond lung cancer (Berkson, 1958). Sometimes the same fallacy based upon dead data keeps returning, as when ‘a study shows’ some effect of left-handedness, such as shorter life spans, when demographers have known for decades that such results are vitiated by strong cohort effects: the frequency of being born lefthanded may be roughly constant over time, but the practice of forcing lefties to retrain as righties has changed greatly over time and varies in different cultures. Since handedness recorded at death need not then agree with handedness at birth, younger dead lefthanders are greatly oversampled in many studies (e.g., Peto, Burgess, & Beaton, 1994).

Data may die of old age, but there are other related risks. Data may die simultaneously with the person who collected them, who is then not around to answer questions that arise. In one well-known case, psychologist Cyril Burt gathered data over many years with psychological and physical measures on pairs of twins, publishing several studies using a data set that increased in size over the years. He was posthumously charged with fraud for seeming anomalies. In one case the accusation was based on a discovery that some (but not all) correlation coefficients did not change when the sample size increased in the later articles. In another case the accuser claimed some data fit a normal distribution "too well." Burt was not there to explain, and the charges were generally accepted as true. Later reexaminations of the cases showed the anomalies can reasonably be innocently explained. When the sequential nature of the data is taken into account —in a multivariate setting increased data may be only available for some coordinates—then only some correlations will be changed

(Joynson, 1989). And in the second case, when data are rescaled and grouped (as Burt stated in the article he had done) the appearance of fit to the normal distribution will be dramatically increased, and the evidence for that charge evaporated when examined this way (Stigler & Rubin, 1979). Nonetheless, Burt's data are now properly regarded as dead, for lack of what we would now consider adequate documentation.

5. Some More General Implications

These lessons from classical data sets are equally important more generally. In the euphoria of the current age there seems to be a growing sense that data are immortal; that they are collected and then more are added to the collection, and more, and that the value of the collection can only grow with size and time. I question that idea.

The life span of data usually has an ending, and it is much earlier than generally realized. That end comes when people stop using the data in favor of later, different data for the same or similar goals, or when they become irrelevant. We may dig out old data for various good reasons, but they are dead when they cease to play a legitimate active role in the scientific discussion. They die when they do not record what later researchers want, or when the trust in them has diminished, or when the questions they can answer are no longer asked, or when new, better data supersede them. Successful resurrection is only slightly more common than in Christian theology.

A real worry today is that with big data it becomes harder to ask the questions we should ask as ‘proof of life.’ In my examples there was sufficient information to uncover major data selection issues and sampling issues. Should we really think a poorly planned study drugging bipolar inmates in an asylum can yield much information on the efficacy of sleeping potions for a more general population? Would we be likely to uncover such things in a part of a huge modern study?

There are modern cases where massive data sets are accumulated over time as part of a grand project, for example, in genomic sciences. A potential problem in such cases is that conditions change with time; new and better measuring devices can be more sensitive in a way that creates a bias (see Jordan, 2019). Many other seemingly benign examples of what was once called ‘interlaboratory differences’ can pass unnoticed, yet have a significant affect in the aggregate.

Another potential problem arises when well-meaning people may try to breathe new life into dead data by recasting them in modern form. For example, they may add small amounts of noise (‘jitter’) to grouped data to make them more realistic, or expand summarized data by simulation in an attempt to recreate the original data as it may have been before it was reduced to summary statistics, or even add noise to render the data less vulnerable to violations of privacy. The intentions may be good, but the result may produce unfortunate side effects. Such data might be called ‘zombie data.’

Unless we can ask of big data the same questions we should ask of small data, and get satisfactory answers, we cannot have confidence in the conclusions. Different people with different questions to answer could ask

different questions. In that sense data could be dead to one user but not another. But biases and sampling problems do not go away simply by repetition. It is clear the problem is too large and the data of today too diverse to hope for a general solution. Awareness of the problem is at least a first step.

6. Some Historical Questions

The intended message in this article is statistical, not historical. The examples, while of historical interest, were chosen to illustrate modern statistical pitfalls, such as the ubiquity of unnoticed data selection issues. Why have the data been selected and presented in this form? Might it have been in service of some strongly held hypothesis? Or maybe the hypothesis has been in fact suggested by the same data presented to support it, a natural practice, but one that can approach circular reasoning. There is an interesting history to these problems as well, one that has caught the attention of historians in recent times.

The statistical issues are not new. Data selection can raise problems with induction, and the statistical side of this was noted as long ago as the 1760s. Richard Price, in an application he added to Thomas Bayes's famous article introducing Bayesian reasoning, called specific attention to the error of letting the data suggest a hypothesis and then using the same data to confirm it without making any allowance for this selection. A century later the Cambridge logician John Venn identified the same difficulty more generally in his book *The Logic of Chance*, noting that the choice of which variables to include was extremely influential, and was at the heart of inductive logic. (For Price and Venn, see Stigler, 2018). In 1885 the Cambridge economist Alfred Marshall issued an even stronger warning on data selection: “The most reckless and treacherous of all theorists is he who professes to let facts and figures speak for themselves, who keeps in the background the part he has played, perhaps unconsciously, in selecting and grouping them” (as quoted in Stigler, 2016, p. 202).

Thomas S. Kuhn, in a provocative article that presaged his vastly influential 1962 book *The Structure of Scientific Revolutions*, argued that the examples in scientific textbooks comparing, say, theoretical predictions based on Boyle's Law with real values derived from a laboratory experiment, were not in fact tests of the theory (as many would suppose), but rather functioned as a description—a definition of—what would be considered the degree of reasonable agreement to be expected from the theory in the practice of that time (Kuhn, 1961). The point was that the measurements of physical science were to be interpreted in the context of the time—of the theory and experimental practice of the time—in which they were made. In later years a number of historians and philosophers of science have carried such ideas to data more generally. For example, Ted Porter's book *Trust in Numbers* looked at the development of expertise in the 19th century and argued that the choice of what to count, what and how to measure, depended on political contexts (Porter, 1995). In recent work he went deeply into the generation of data in 18th- and 19th-century asylums that was then used to examine issues in mental health, illuminating the challenges of ‘big data’ from a past era when the plasticity of categorization resulted in data being deduced from conclusions (Porter, 2018). In these and others' such studies (e.g., Gitelman, 2013; Radin, 2017), the role of context in shaping data selection and form—context in

temporal, political, and social as well as scientific terms—has been shown to be a powerful and interesting phenomenon.

Disclosure Statement

Stephen M. Stigler has no financial or non-financial disclosures to share for this article.

References

Andrews, D. F., & Herzberg, A. M. (1985). *Data: A collection of problems from many fields for the student and research worker*. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-5098-2

Berkson, J. (1958). Smoking and lung cancer: Some observations on two recent reports. *JASA, 53,* 28–38. https://doi.org/10.1080/01621459.1958.10501421

Cushny, A., & Peebles, A. R. (1905). The action of optical isomers II: Hyoscines. *Journal of Physiology, 32*(5–6), 501–510. https://doi.org/10.1113/jphysiol.1905.sp001097

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Gitelman, L. (Ed.). (2013). *Raw data is an oxymoron*. Cambridge, MA: MIT Press.

Gosset, W. S. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–24. https://doi.org/10.2307/2331554

Jordan, M. (2019). Artificial intelligence—The revolution hasn't happened yet. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.f06c6e61

Joynson, R. B. (1989). *The Burt Affair*. New York, NY: Routledge.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis*, *52*(2), 161–193. https://doi.org/10.1086/349468

Peto, R., Burgess, A., & Beaton, A. A. (1994). Left handedness and mortality: Causal inferences cannot be trusted. *BMJ*, *308*(6925), 408–408. https://doi.org/10.1136/bmj.308.6925.408

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Porter, T. M. (2018). *Genetics in the madhouse: The unknown history of human heredity*. Princeton, NJ: Princeton University Press.

Preece, D. A. (1982). t is for trouble (and textbooks): A critique of some examples of paired-samples t-test. *The Statistician, 31*(2), 169–195. https://doi.org/10.2307/2987888

Prüssische Statistik (1875–1894). Vols. 38, 46, 50, 55, 60, 63, 67, 80, 84, 87, 91, 95, 99, 108, 114, 118, 124, 132, 135, 139.

Quetelet, A. (1817). Statement of the sizes of men in different counties of Scotland, taken from the local militia. *Edinburgh Medical and Surgical Journal, 13*, 260–264.

Quetelet, A. (1846). *Lettres à S.A.R. Le Duc Régnant de Saxe-Cobourg et Gotha, sur la Theorie des Probabilités, appliquée aux Sciences Morales et Politiques*. Brussels, Belgium: Hayez. (Translated as Letters Addressed to H. R. H. the Grand Duke of Saxe Coburg and Gotha, on the Theory of Probabilities as Applied to the Moral and Political Sciences, 1849. London, UK: Layton.)

Radin, J. (2017). "Digital natives": How medical and indigenous histories matter for big data. *Osiris, 32*(1), 43–64. https://doi.org/10.1086/693853

Senn, S., & Richardson, W. (1994). The first t-test. *Statistics in Medicine, 13*(8), 785–803. https://doi.org/10.1002/sim.4780130802

Stigler, S. M. (1986). *The history of statistics*. Cambridge, MA: Harvard University Press.

Stigler, S. M. (2016). *The seven pillars of statistical wisdom*. Cambridge, MA: Harvard University Press.

Stigler, S. M. (2018). Richard Price, the first Bayesian. *Statistical Science, 33*(1), 117–125.

Stigler, S. M., & Rubin, D. B. (1979). "Burt's Tables" and "Dorfman's Data Analysis" (the latter with D. B. Rubin). *Science, 204*, 242–245, and *205*, 1204–1206.

von Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen*. Leipzig, Germany: Teubner.

Winsor, C. P. (1947). Das Gesetz der kleinen Zahlen [The law of small numbers] *Human Biology, 19*, 154–161.

Zabell, S. (2008). On student's 1908 article "The Probable Error of a Mean." *JASA, 103*(481), 1–7.
