Shining a Light on Dark Data



Professor David Hand discusses the perils of "dark data" – where the data that you *don't* have can be just as important, if not *more* important than the data you do have. He provides examples of where dark data has let to catastrophic outcomes and provides insights into managing the risks of dark data, and even, how to start using dark data to your advantage.

When you first start to learn statistics, you learn from small complete data sets. You calculate the mean of ten numbers, their median, their standard deviation. Later you learn to use computers, carrying out more sophisticated statistical analyses like regression or analysis of variance. Again, initially, when you learn how to use these tools – the statistical methods and the software which implements them – you apply them to perfect data, to clean and complete data sets. This all makes perfect sense. If you are trying to learn about averages or regression analysis, you don't want to be sidetracked by having to deal with peripheral matters.

Except that, when you graduate from learning how to use the tools and begin to apply them to real data sets, you realise that you have been misled. You learn that real data are not as clean as those early ten-point data sets suggested, that the data for your regression aren't as balanced as in the text books. At first, you ignore the problems. You discount the fact that quite a few people you approached for your survey refused to take part so your data set does not include them, that you were unable to take accurate measurements of some of the chemical concentrations, that some of the numbers were so outlandish that you felt obliged to drop them from the analysis, that the data collected when the patients were admitted to the hospital isn't *exactly* what you need to answer your research question, that the instrument providing the telemetry feed never gives a value greater than 2.3.

But then you begin to think more closely about the issues. What if those who refused to respond to the survey were different from those who agreed? Would this mean my statements about the overall population were wrong? What if the telemetry instrument had a hard ceiling at 2.3, so that it could never read above this, even if the actual values were way larger? Would this mean I was being very misled about how the system was working?

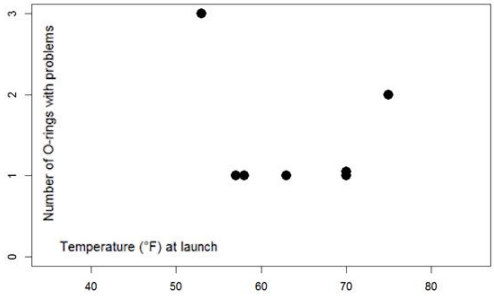
These problems, these data inadequacies, are examples of *dark data*. Dark data are, in short, data you don't have. They are the data you believe you have or hoped to have or would have had *if* something else had happened but you don't actually have. And the painful truth is that real life is riddled with *dark data* problems. Dark data are ubiquitous. It means that, as well as looking at the numbers we see in front of us, we need to ask *what am I not seeing? What is missing? What are these numbers concealing?* What is worse, these dark data can have very serious consequences. More than just meaning your analysis is wrong, it can result in lost fortunes and even deaths. Here are some examples.

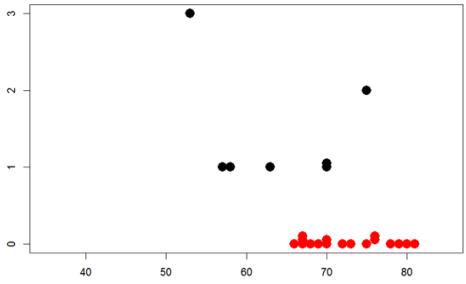
How it all went horribly wrong

You may well be familiar with the 1986 Challenger Space Shuttle disaster, in which the teleconference the night before studied the relationship between O-ring distress and air temperature at launch for those launches which had had O-ring problems. As the Rogers Commission report on the disaster says ([1], p146-7):

"there is nothing irregular in the distribution of O-ring 'distress' over the spectrum of joint temperatures at launch between 53 degrees Fahrenheit and 75 degrees Fahrenheit."

But what was revealing about the plot was not so much as what it showed, but as what it didn't show. In particular, no data was included for those launches which had had *no* O-ring problems.

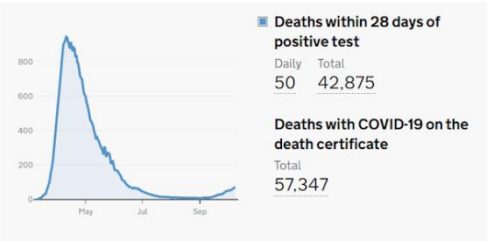




Why include data which appeared to show nothing about the problem being studied? But including those extra (dark data) points in the graph (red points in graph opposite) gives a completely different picture – so different that it is impossible to believe the launch would have gone ahead if the conference participants had seen it. The original data plot had just seven points. Seven astronauts died in the disaster.

A completely different kind of dark data arises from ambiguities or differences in definitions. The police recorded crime (PRC) statistics show that crime in the UK went up between 1997 and 2003 (from 4.6 million to 5.5 million), but the Crime Survey for England and Wales (CSE&W) show that crime *dropped* over the same period (from 16.5 to 12.4 million). If you had studied just one of these data sets you might be very misled – especially if you were really interested in the question answered by the other. The results are different, of course, because the two sources used different definitions of what a crime was. The CSE&W is a

victim survey, asking people what crimes they have experienced. That means it omits some – like possession of drugs. On the other hand it will include crimes that people didn't report to the police (perhaps because they felt no action would be taken). In contrast, the PRC omits those not reported. And there is a complicated potential feedback effect: the number of people arrested for drug possession will depend on the level of police activity, and that will depend on the perceived need for police activity.



And for a third, and again very different type of dark data, consider what summary statistics conceal. You may learn that a particular country has 20 cases of Covid-19 per day per 100,000 in the population. But what this doesn't tell you is that Covid-19 outbreaks tend to be clustered. A look at a map of where the cases occur will show you hotspots, where the rate is much higher than 20 per 100,000, and the rest of the country where it is much lower. In fact, the coronavirus pandemic is a rich source of examples of dark data. Early reported infection rates were wildly misleading because they reported only people with symptoms, and only those that came forward at that. Death rates can be misleading, according to whether they report people who died *of* Covid or *with* Covid, and how effective the strategy is for identifying such people. Tests are not perfect, and always have a non-zero rate of false positives and false negatives. If critical risk factors like age and deprivation are not measured we will not learn how important they are: what else have we missed? All of these, and other causes, serve to conceal what is really going on and have the potential to mislead any analysis.

"The coronavirus pandemic is a rich source of examples of dark data"

A taxonomy of dark data

These examples show that dark data can occur (or perhaps that should be 'not occur') in all sorts of ways. In my book *Dark Data: Why What You Don't Know Matters* [2] (and see https://darkdata.website/) I give a taxonomy of 15 types of dark data. These are not mutually exclusive – just because your data suffers from one of them does not mean it will not also suffer from others. These 15 types are:

1. *Data we know are missing*
2. *Data we don't know are missing*
3. *Choosing just some cases*
4. *Self-selection*
5. *Missing What Matters*
6. *Data Which Might Have Been*
7. *Changes with Time*
8. *Definitions of Data*
9. *Summaries of Data*
10. *Measurement Error and Uncertainty*
11. *Feedback and Gaming*
12. *Information Asymmetry*
13. *Intentionally Darkened Data*
14. *Fabricated and Synthetic Data*
15. *Extrapolating beyond Your Data*

The first two will be familiar to you, because they are Donald Rumsfeld's famous 'known unknowns' and 'unknown unknowns'. A familiar example of the first would be a missing value in a table giving ages of people. If a value is missing, we still know that that person had an age; we just don't know what it is. An example of the second is given by the Challenger disaster: the participants in the teleconference did not recognise that they were missing (crucial) data. Another example of this second type arose as a consequence of a problem with a NASA satellite, which suggested that the extent of fires in August 2020 in Brazil's section of the Amazon rainforest was 5% lower than the previous August [3]. It was only later that extra data showed this was not true, and in fact the number of fires had increased over the year [4].

The third and fourth type of dark data above, choosing just some cases and self-selection, are particularly common and pernicious types of dark data. If people are allowed to decide whether or not to be included in a database (to share their medical records, for example) then there is huge potential for the database to be non-representative of the population, with clear risk of subsequent analysis being misleading. If all the people who begin to feel better drop out of a clinical trial, so that you don't know that they are recovering, the implications for the conclusions are obvious. The Gartner organisation focuses on type 3 dark data (also called 'cold data' by some), ignoring data which have been collected and stored [5] but not analysed – perhaps they were collected for reasons of regulatory compliance.

And so it goes on. Dark data are ubiquitous, cropping up in all sorts of unexpected ways, possibly misleading you with potentially serious consequences.

"I have to chuckle whenever I read yet another description of American frontier log cabins as having been well crafted or sturdily or beautifully built. The much more likely truth is that 99% of frontier log cabins were horribly built—it's just that all of those fell down. The few that have survived intact were the ones that were well made. That doesn't mean all of them were." Mike Johnston at The Online Photographer [7]

What can we do about dark data?

So, that's the problem. But what can we do about it?

An obvious strategy which can sometimes be applied (at least if you know you have dark data) is to get more data. Larger samples reduce measurement inaccuracy, a sample of applicants you would normally have rejected gives you more information on overall population risk characteristics (so, for example, you can build better credit scoring models), follow up those who don't respond to surveys, and so on.

These are all sensible things to do. But there's an old saying that all complex problems have simple, straightforward, and easy-to-understand wrong answers. The same applies to strategies for coping with dark data. While basic statistics textbooks don't deal with the problem at all, implicitly encouraging people to ignore it, some statistical software packages include such simple strategies as default solutions, encouraging people to think they are reasonable approaches. One such strategy is *complete case analysis*, where you drop incomplete records. That's fine, but on the one hand you might discover it leaves you with very little data left to work with, or even none at all, while on the other hand how do you know that the complete cases are properly representative of the population? You could be being seriously misled.

Anthropologist Stephen Jay Gould wrote: "In July 1982, I learned that I was suffering from abdominal mesothelioma … mesothelioma is incurable, with a median mortality of only eight months after discovery." But what Gould realised was that that average mixed together all sorts of different people. He went on to say "half the people will live longer; now what are my chances of being in that half. I read for a furious and nervous hour and concluded, with relief: damned good. I possessed every one of the characteristics conferring a probability of longer life: I was young; my disease had been recognized in a relatively early stage; I would receive the nation's best medical treatment; I had the world to live for; I knew how to read the data properly and not despair." Gould lived for 20 years after his diagnosis, eventually dying of an unrelated cancer. [8]

Another simple strategy is to use all the data you have. Calculate the average height from those who gave their height, and the average weight from those who gave their weight. That's also fine – except what if the shorter people and the fatter people were more likely not to answer. Your two averages could give the impression of a population of very tall very thin people. Worse, correlations calculated from pairs of variables in this way can lead to contradictions: the sample correlation between x and y, and between y and z, contradicts that between x and z.

A third simple strategy is imputation: replace all the missing age values by the average of the ages recorded, for example. Except that this will underestimate the variance of age in the data. Having said that, modern sophisticated methods for coping with dark data – like the so-called *EM-algorithm* – are based on elaborations of this idea.

"And that, perhaps, is the most important message … be suspicious, recognise that you might be missing something"

To cope with dark data effectively, it is necessary to have an understanding of why it is dark. That means, in the first place that you have to know it is lurking out there. And that, perhaps, is the most important message of this article: *be suspicious, recognise that you might be missing something*. Having done that, you can try to detect it. While detection is very context dependent, varying from situation to situation and data set to data set, there are general strategies, like using expected properties of data – like the Benford distribution [6], for example.

Data can be dark for a huge variety of reasons, but a useful categorisation is into data which are unobserved for reasons unrelated to the study (e.g. sickness preventing students from sitting an exam), data which are unobserved for reasons related to data you do have (e.g. exam results preferentially missing for students who previously did badly on tests), and data which are unobserved because of the values they would have had (e.g. a student realising he had revised for the wrong exam, and so skipping it). I call these *Not Data Dependent*, *Seen Data Dependent*, and *Unseen Data Dependent* respectively in my book.

Using dark data to your advantage

My message is that dark data are ubiquitous and dangerous. But there is more. In fact, you can take advantage of dark data, in what I call the strategic application of ignorance. I do not mean hiding information from people, as in fraud, but something much more ethical. In fact, you already do this in other contexts, such as when you use a computer password (itself dark data) to keep data dark from others. And statisticians do it all the time, in all sorts of other ways. To spot this, it is necessary to look at statistical analyses from the other perspective, the dark data perspective.

So, for example, sample surveys study just a (carefully randomly selected) portion of a population – they keep the remainder dark – while randomised controlled trials conceal the true treatment each person receives.

But there's even more. Some of the most sophisticated tools use unobserved *data which might have been*. For example:

- in simulation, where data *which could have arisen from a model* are generated
- in regularisation, equivalent to slightly randomly perturbing the observed data
- in bootstrap methods, which take multiple subsamples from the observed data
- in boosting methods which replicate misclassified cases and force models to focus attention on them
- in Bayesian methods, which combine the data you have got with imprecisely specified extra data in the form of priors.

While it is possible to take advantage of dark data ideas, my main aim in this article has been to draw attention to the risks. These risks have been with us since we began collecting data, and they have not gone away with computers and big data. Computers, while enabling us to do amazing things, necessarily stand between us and the data. They enable us to see projections – summaries, plots, and so on – of the data. And big data do not erase the dark data challenges. Rather, they amplify the risks: you cannot examine your billion point data set point-by-point (there are only 32 million seconds in a year) but have to rely on computer summaries. Big data have all the problems of small data, and extra problems of their own.

It is always necessary to ask: *what might we be missing*?

References

[1] Rogers Commission: Report to the President by the Presidential Commission on the Space Shuttle Challenger Accident, June 6th 1986, NASA, Washington DC

[2] Hand, D.J. (2020) Dark Data: Why What You Don't Know Matters. Princeton University Press.

[3] www.reuters.com/article/us-brazil-environment-fires-exclusive/exclusive-brazil-amazon-fires-likely-worst-in-10-years-august-data-incomplete-government-researcher-says-idUSKBN25T349

[4] unearthed.greenpeace.org/2020/08/13/amazon-rainforest-fires-brazil-protected-areas

[5] www.gartner.com/en/information-technology/glossary/dark-data

[6] Berger A and Hill T (2015) An introduction to Benford's law. Princeton University Press, Princeton

[7] theonlinephotographer.typepad.com/the_online_photographer/2017/02/i-find-this-a-particularly-poignant-picture-its-preserved-in-the-george-grantham-bain-collection-at-the-library-of-congres.html

[8] csn.cancer.org/node/213889

Reference websites accessed Oct 2020.
Image Attributions
Lead article: 191433897 © Angkana Kittayachaweng | Dreamstime.com
Covid-19 graph (from coronavirus.data.gov.uk/deaths) licensed under Open Government Licence v3.0

Professor David Hand, Imperial College London

David Hand is Emeritus Professor of Mathematics and a Senior Research Investigator at Imperial College, London. He is a Fellow of the British Academy and a former President of the Royal Statistical Society. His 31 books include *Principles of Data Mining*, *Measurement Theory and Practice*, *The Improbability Principle*, *The Wellbeing of Nations*, and *Dark Data*.