3rd International Conference
on Public Policy (ICPP3)
June 28-30, 2017 – Singapore

Panel T14P03 Session 1

The Data/Sensor Revolution and Public Policy

Title of the paper

From Dots to Distributions: Why a Statistician's Approach to Big Data Matters

Author(s)

Jason Kok, Autoriti Monetari Brunei Darussalam,
Brunei Darussalam (jason.kok@ambd.gov.bn)

Date of presentation

30 June 2017

Abstract

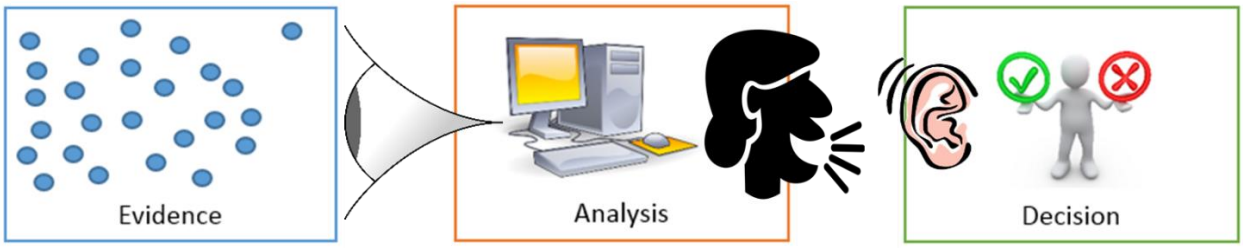
This paper examines a practical challenge in using data for policy whereby we may have the evidence for evidence-based policy but we are still unable to make the right decisions. It is argued that this is because, despite the emergence of Big Data, the importance of understanding statistical concepts by analysts/researchers and policymakers is still underappreciated. The process of analysing the evidence to produce research that informs policymakers implicitly relies on the understanding of statistical concepts to appreciate the limits of conclusions presented in research. There is an overemphasis on "dots" such as the mean/median which give a point value for forecasts, expected effects, etc. without an appreciation of the "distribution" of uncertainty and possible outcomes around this "dot". The paper talks about basic statistical concepts that still doggedly persist and result in misunderstandings on the conclusions of research including sampling, random variables and moments of a distribution. Basic statistical concepts and theory common in almost every undergraduate/high school statistics textbooks are used, highlighting a serious practical issue of analysts/researchers in trying to convey their research to policymakers in an understandable manner. A number of real-life examples are used to illustrate that this is a real problem even in developed economies. The paper concludes by discussing the author's views on why this problem persists and how it may be addressed.

Keywords: statistical theory, Big Data, evidence-based policy, statistical analysis

Introduction

Evidence-based policy aims to make optimal policy decisions based on an analysis of relevant and available evidence. Big Data is seen as a huge, untapped source of this evidence that may make evidence-based policy a reality. SAS Institute (2016), the provider of SAS statistical software, defines Big Data as "the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis." Given the high volume of this data, tools have been developed to ease the analysis of these large datasets; notably business intelligence. Business intelligence is using software applications to analyse large amounts of data, in order to better inform decisions. It enables the visualisation of data, create analyses and interactive reports/dashboards to provide a story analysing why something happened. The relatively straightforward user interface of business intelligence packages enables this data analysis to be opened up to more people, even those unfamiliar with statistical concepts.

However, this may result in problems linking to the basic computing concept of "Garbage In, Garbage Out". Business intelligence tools are powerful but the limitation is the human sitting at the desk clicking the buttons. This can be illustrated graphically as below:



Evidence-based decision-making relies on analysts interpreting the data (eye) and applying statistical concepts to produce robust analysis which is then communicated (mouth) to policy-makers that must understand the underlying statistical concepts (ear) in order to make the correct decision. A breakdown of the eye, mouth or ear may result in incorrect decisions being made regardless of the quality of evidence, analysis or decision-making process.

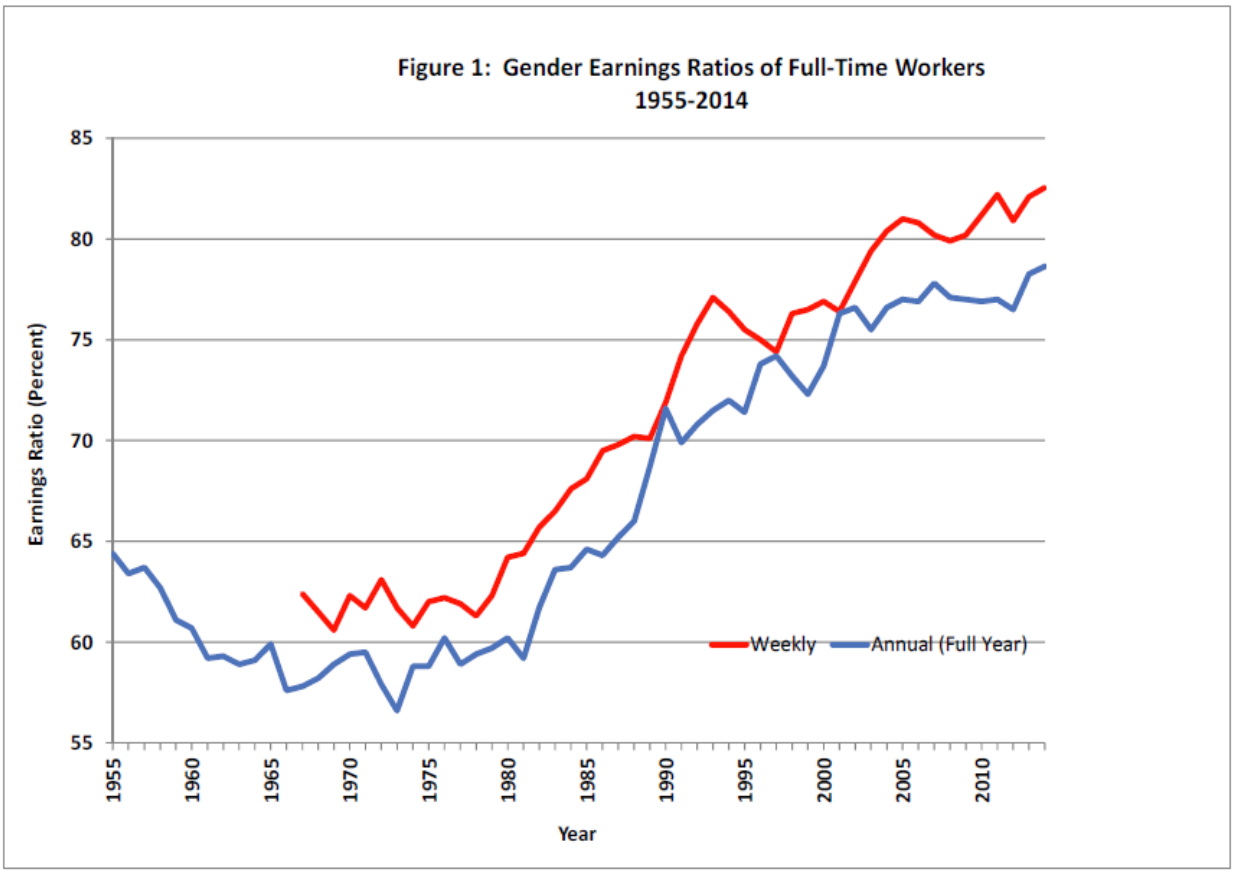
This brings us to the topic of this paper "From Dots to Distributions: Why a Statistician's Approach to Big Data Matters". By being able to make sense of a basic statistical concept of distributions, we can better understand the voluminous amounts of data (dots) needed for evidence-based policy.

Sample vs Population

The first concept to understanding a distribution, in my opinion, is understanding the difference between samples and populations. The population covers the entirety of all possible outcomes at all times. A sample just provides a snapshot of observed outcomes from the population over a specific period of time. Big Data for all its volume and detail is still just a sample. Continuing from this, a sample no matter how large it is will always have some data/information missing, i.e. the sample cannot fully explain the population but is a close approximation.

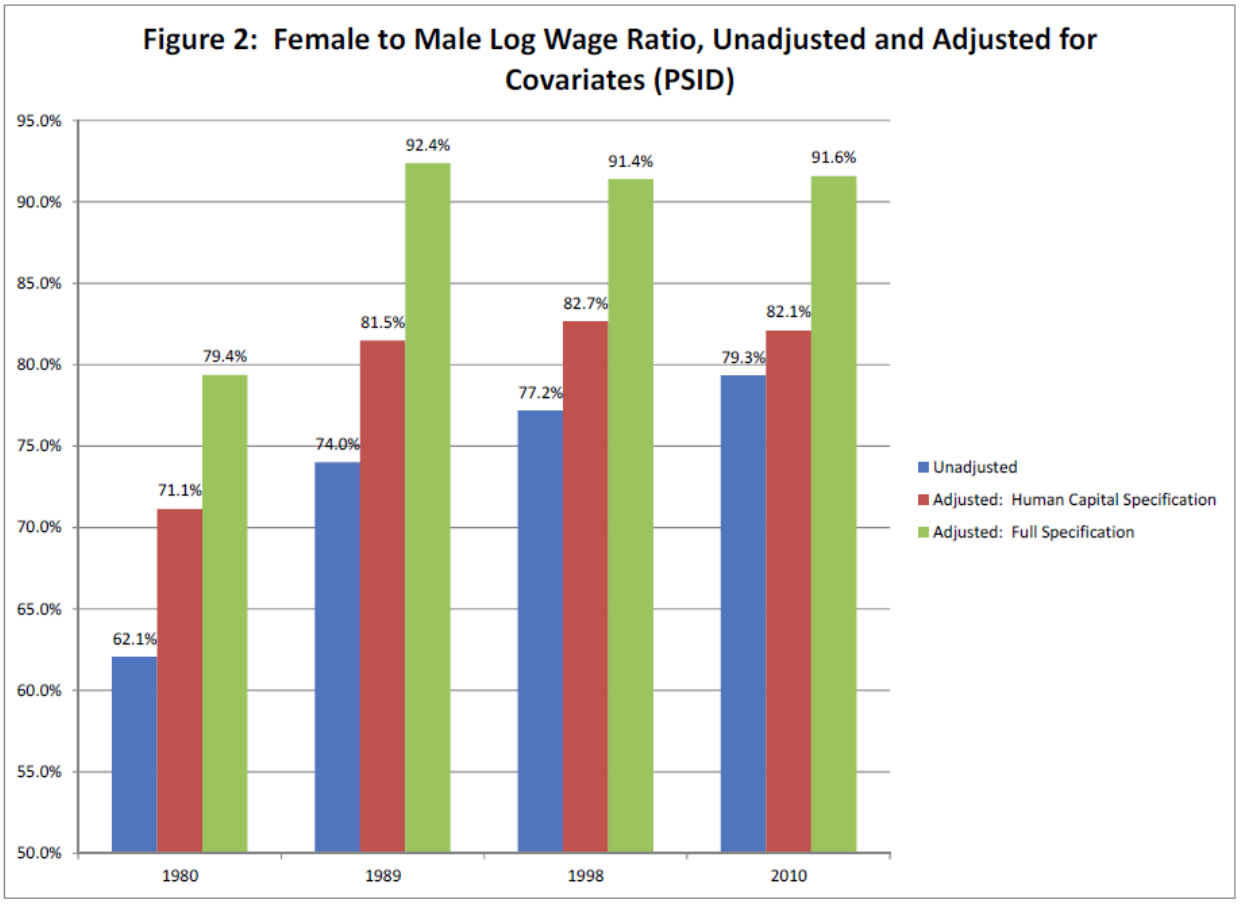
Thus, it is important for any analysis to start from a good sample. If we apply appropriate sampling techniques, we can ensure the sample is representative of the overall population. This then enables our analysis and conclusions to be extrapolated from the sample to be roughly applicable to the population. Even with correct analysis, an unrepresentative sample will result in erroneous conclusions.

An example of this is the much talked about US gender wage gap whereby a woman earns US$0.79-US$0.82 for every US$1 a man earns (Blau and Kahn, 2016).



Source: Blau and Kahn (2016)

Based on this evidence, policymakers may believe this is blatant sexual discrimination and propose a policy to give all female employees a 20% raise in order to correct for this problem. However, this analysis is focusing on the wrong sample. The authors constructed a more representative sample that adjusted for human capital (education, experience, race, region, metropolitan area residence) and full specification (industry, occupation, union coverage). Based on this adjustment, the gender wage gap shrinks to US$0.916:US$1. Blau and Kahn (2016) argue the remaining gap is likely due to sexual discrimination, women's workforce interruptions (e.g. due to pregnancy), and gender differences in psychological attributes/noncognitive skills. Based on this analysis, a one-time pay raise for female employees seems to be an incorrect decision.



Source: Blau and Kahn (2016)

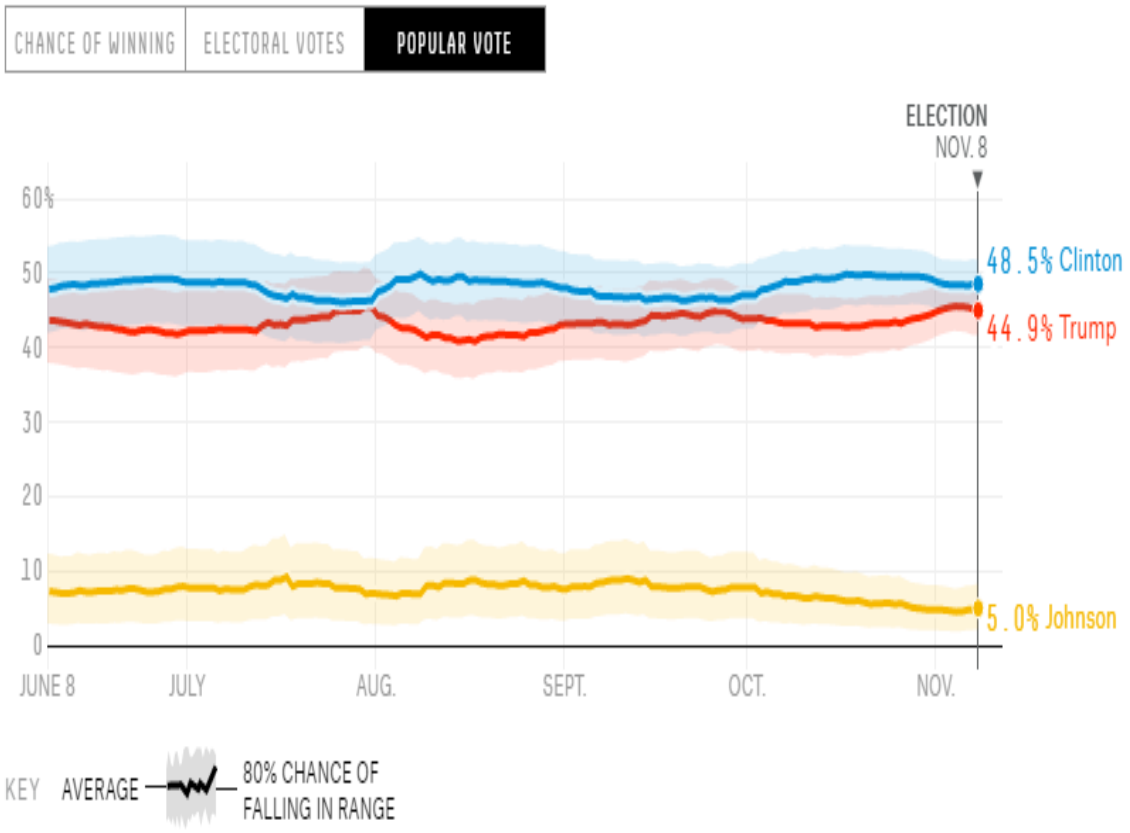
However, you may argue that policymakers would not make the mistake of choosing the wrong sample if the authors of research papers readily provide representative samples. The example below from Democrat Hillary Clinton would beg to differ. The study she was referencing was Chamberlain (2016) that found a gender wage gap of US$0.759:US$1 which shrinks to US$0.946:US$1 once statistical controls are put in. Thus, despite the efforts of researchers/analysts that produce robust analysis to inform policymaking, it can all be for nought if policymakers and the media don't understand sampling and choose to focus on the wrong part of their research papers.

Source: Twitter

Random variable/function

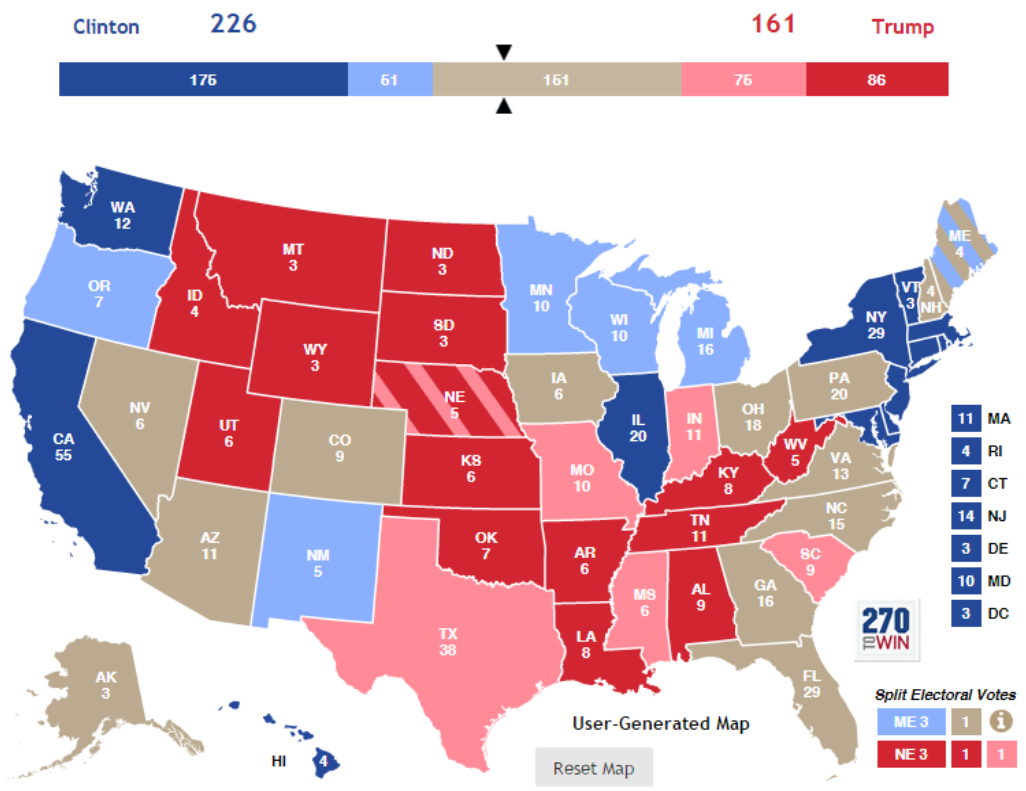
The next concept is the random variable/function. In statistics, this random variable/function is the underlying relationship that generates all the observed outcomes (data) for the population and the sample. By having a representative sample, researchers/analysts can use the sample data to try to estimate this random variable/function. But as mentioned previously, the sample is still a sample rather than the entire population. Thus, we cannot be sure our estimated relationship is the true relationship. We can only say it is a "true" relationship that can at least explain our sample data with an unclear idea whether this can accurately predict out-of-sample.

These "true" relationships are commonly applied to predict outcomes such as the recent US Presidential Election.



Source: FiveThirtyEight 2016 Election Forecast; Last updated 8 November 2016

The US Presidential Election is not based on a popular vote, meaning our "true" relationship should account for this or else risk being erroneous. The US President is chosen by an electoral college system meaning the location of voters matters. Voter turnout in US elections is typically around 60%, which further confounds predictions as we aren't sure which potion of a representative sample will actually vote. Thus, the US Presidential Election is a complicated process where each vote is a data point from the US population. An election represents a single instance generated from the random variable/function that explains the true relationship of how US citizens vote.



Source: 270towin; Last updated 8 November 2016

Source: FiveThirtyEight 2016 Election Forecast; Last updated 8 November 2016

One important thing to remember about random variable/function is the word "random" which implies probabilities. Going into Election night, Clinton had a lead in terms of probability to win but there are no 100% guarantees. A probability above 50% is not an assurance of an easy win. Trump's 28.6% chance to win is not low; by comparison if you pick a random day in the year then there's a 25% chance it'll be in winter or if you pick one random person from the world, there's a 20% chance they'll be from China. The only way to tell if the "true" relationship holds is to have more observations to verify the estimated probabilities are reasonably accurate. However, this is impossible in the case of the US Presidential Election that only occurs once with two specific candidates every 4 years.

Distributions

Now that we have some grounding in terms of sampling and random variable/function, we can move on to think about distributions. The moments of a distribution tell us about its shape and characteristics:

1. First moment: mean; the average value of the data
2. Second moment: variance; measure of how far data points are from the mean

3. Third moment: skewness; how symmetric is the distribution
4. Fourth moment: kurtosis; how fat are the tails relative to the Normal distribution

The first moment gives us a single datapoint summary of a set of data (**Dot**) and we need to go into higher order moments in order to see the **Distribution**. However, one common fallacy among the media and even policymakers is to focus on the first moment and never to see the underlying distribution.

Tall people 'more likely to develop cancer'

By James Gallagher
Health reporter, BBC News

21 July 2011 | Health

Share

Being tall has been linked to a greater risk of 10 common cancers by University of Oxford researchers.

For every four inches (10cm) above five feet a person was, the researchers said they had a 16% increased cancer risk.

The study of more than one million women, **published in The Lancet Oncology**, suggested chemicals that control growth might also affect tumours.

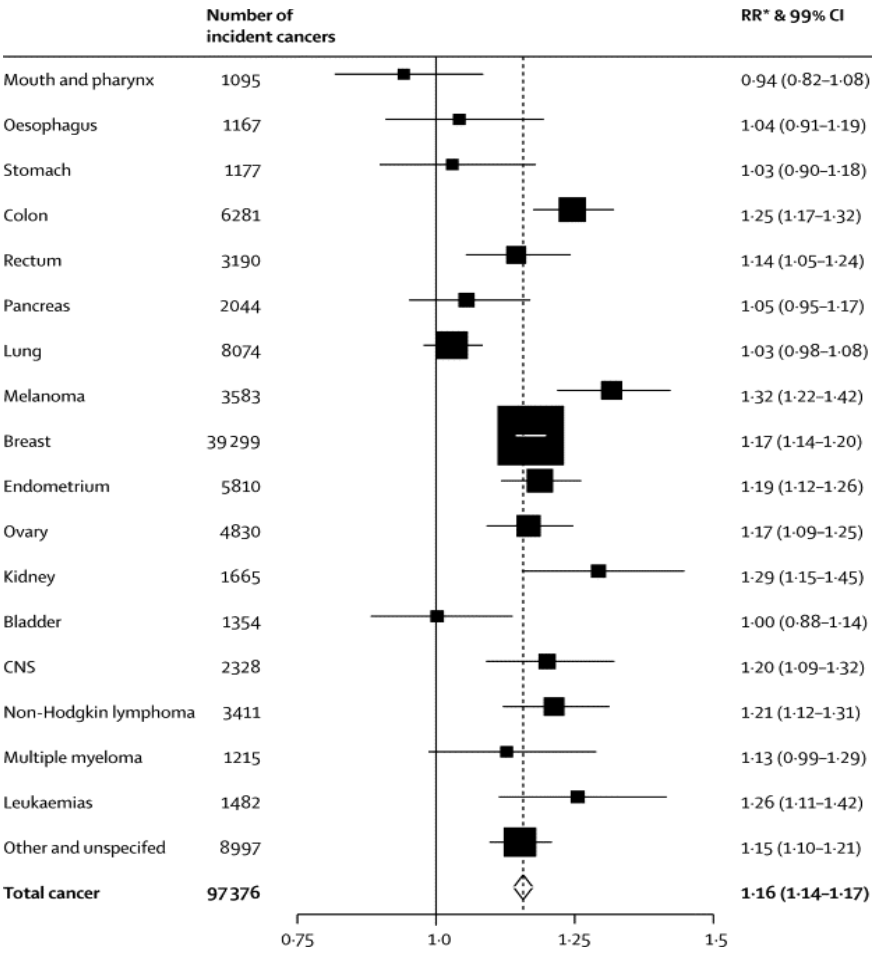
Cancer Research UK said tall people should not be alarmed by the findings.

The study followed 1.3 million middle-aged women in the UK between 1996 and 2001.

Height is linked to cancers, but the reason why is unknown.

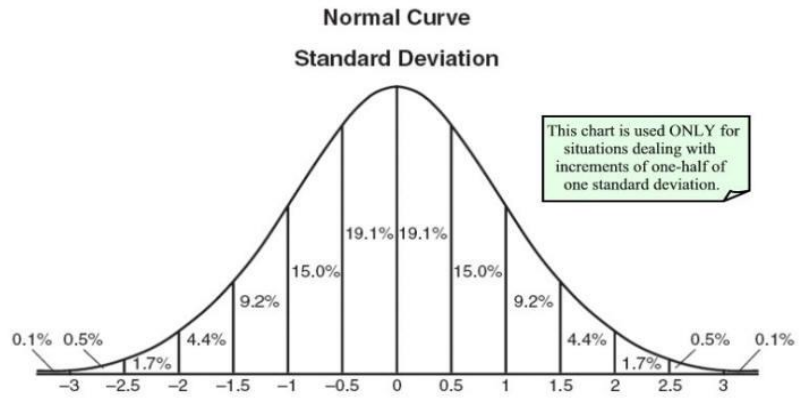
Source: BBC News

This news article suggests being 10cm taller gives you a 16% increased cancer risk. However, this does not tell me about the range of possibilities. The range for increased cancer risk could be 0%-32% or 15%-17%, both of which would result in a mean of 16%. The researchers provided a view into the distribution of this increased cancer risk via a box and whisker diagram. Essentially the diagram shows the range of possibilities lying between a confidence interval of 14% and 17%, with differing probabilities based on the specific types of cancer. A simple focus on “16% increased cancer risk” misses out on the wealth of information and detail arising from the efforts of the researchers.



Source: Green et al. (2011)

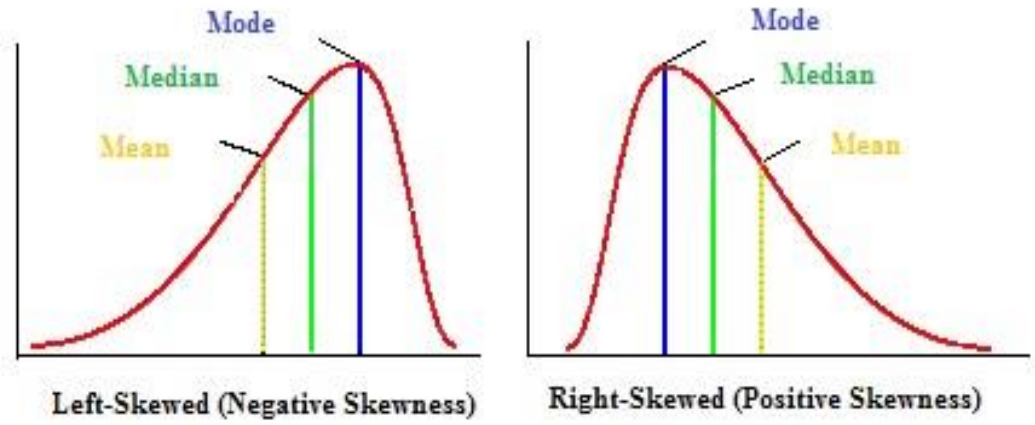
The Normal distribution is a good starting point to consider distributions. It is perfectly symmetrical with the mean and median at 0. The amount of standard deviations away from the mean gives a confidence interval providing probabilities of observing data in that range of values.



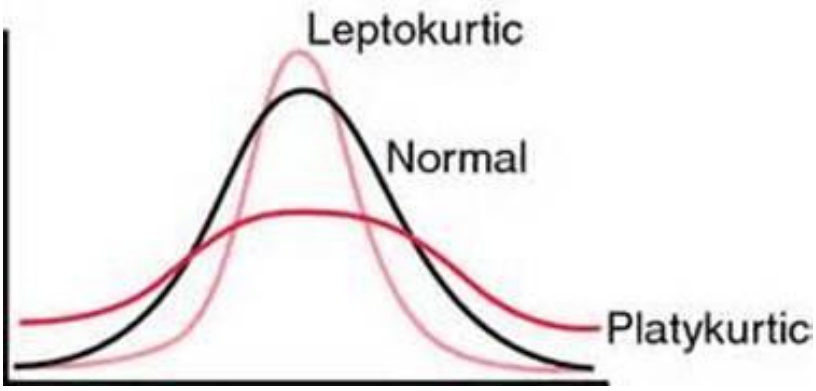
Source: Mathbits.com

Understanding the Normal distribution allows us to better grasp how the third and fourth moments influence the shape of a distribution. The third moment (skewness) tells us whether the bulk of our data

is above or below the mean. The fourth moment (kurtosis) tells us if we have fat tails, i.e. higher probability of extreme low/high events.



Source: StatisticsHowTo



Source: Social Science Research & Instructional Center

By knowing at least the four moments of a distribution we get a clear sense on how the distribution looks which tells us how the observed data is likely to be dispersed.

Forecasting

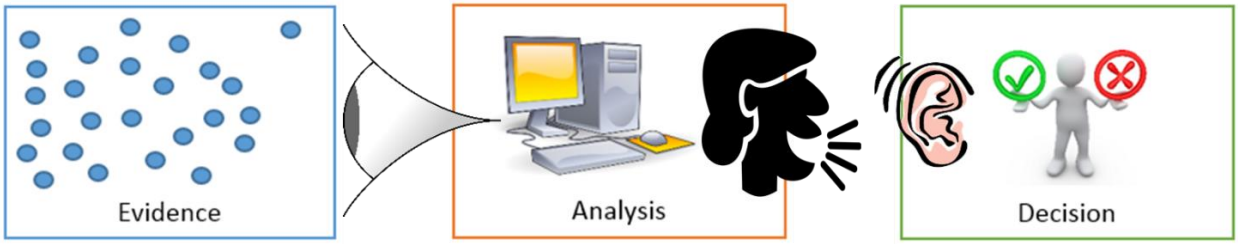
Data is frequently used to predict or forecast future events. As in the previous example of trying to predict the winner of an election or a meteorologist trying to predict the chance of rain tomorrow. Forecasts use sample data to estimate a "true" relationship of a random variable/function, which is hoped to be close to the true relationship of the population. Adding on to this the concept of distributions allows us to appreciate forecast uncertainty, which is the distribution around the forecast (**Dot**). This forecast uncertainty highlights the range of possibilities around the forecast. When a meteorologist says there is an 80% chance of rain tomorrow, what he/she isn't saying is the range of uncertainty which may be 75%-85%. If the actual chance of rain works out to be in fact 81% it doesn't mean that the meteorologist was wrong by saying 80%. This is because the 81% is within the range of possibilities that he/she assessed was reasonably possible.

Anchoring-and-adjustment heuristic

Renowed Nobel prize winning psychologists Tversky and Kahneman (1974) identified an anchoring-and-adjustment heuristic whereby people tend to fixate on starting information provided (anchor) whether it was relevant or not and then adjust from that value when asked to user that information to make a judgment/decision. This suggests that human judgment is significantly influenced by initial values given. Thus, for example informing a policymaker the mean/median estimate (dot) may influence them to overly fixate on that value without fully considering the uncertainty of the estimate (distribution).

Conclusion

The examples I explored above do not use Big Data, yet policymakers and the media still make mistakes in interpreting the research. Despite having good data/evidence and a robust methodology applied by a researcher, which can be used for evidence-based policy-making, it only takes a minor misinterpretation to result in incorrect policy advice. This highlights why understanding statistical concepts is essential if we wish to pursue evidence-based policy-making. Let us consider again the graphic from the introduction.



Generally, the “eye” is not the problem as peer review and strong technical training ensures researchers across a wide range of disciplines apply statistical concepts correctly. This follows on to the analysis section. Any shortcomings in these areas can be easily rectified with improved technical training. The problem areas seem to stem from the “mouth” and the “ear”.

How well do researchers/analysts communicate the results of their analysis? Can this be easily understood by a less technically proficient audience? Can policymakers understand the language used by researchers/analysts as well as the underlying statistical concepts applied? I believe the problem lies in both “mouth” and “ear” with a lost in translation problem. We, as researchers and analysts, need to improve how we communicate the results of our research (evidence-based) but at the same time policymakers need to improve their understanding of statistical concepts in order to appreciate the evidence and make informed policy decisions (policy-making). Without both being in sync then we will not have evidence-based policy-making. We need to all look beyond the dots and see the distributions hiding underneath.

References

Blau, F.D. and Kahn, L.M. (2016) The Gender Wage Gap: Extent, Trends, and Explanations. Institute for the Study of Labor, Discussion Paper No. 9656 (IZA DP No. 9656).

Chamberlain, A. (2016) Demystifying the Gender Pay Gap - Evidence From Glassdoor Salary Data. Glassdoor Research Report, March 2016

Green, J., Cairns, B.J., Casabonne, D., Wright, F.L., Reeves, G., and Beral, V. (2011) Height and cancer incidence in the Million Women Study: prospective cohort, and meta-analysis of prospective studies of height and total cancer risk. *The Lancet Oncology*, Vol. 12 (8), pp. 785 – 794.

SAS Institute (2016) Big Data What it is and why it matters. Available from: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html (Date accessed: 16 November 2016)

Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science*, Vol. 185, pp. 1124–1130.