CHAPTER 3

Data Analytics

Elaine TAN

"Information is the oil of the 21st century, and analytics is the combustion engine."

(Peter Sondergaad 2011)

What Is Data Analytics?

To measure. To understand. To improve.

These are the purposes of data analytics, the art and science of deriving patterns and relationships through numbers. Data analytics has had a long history.[1] What has changed in recent decades is the ever-expanding volume of available data, coupled with exponential growth in computing technology. These mean data analytics now has greater promise and wider reach. New economic roles and activities can be created through data collection, processing and analysis. Properly mined, data can be potent ingredients in improving how we work, play and live. There are, however, important limitations to quantitative analyses, so caution should be exercised both in their execution and interpretation.

For public officers, analytics can be undertaken by tapping on the many rich databases which are assembled in the course of their work. Such data are created each time a resident comes into contact with a public agency, e.g., exits or re-enters Singapore, drives through an ERP gantry, makes a CPF withdrawal, etc. These data may be studied to understand broad patterns and trends, guide

[1] For example, Charles Baggage's study in the early 1800s on British factory productivity and workers' skills used statistical observations from British and overseas factories. One hundred years later, the Nazis in Germany used punch-card technology and counting machines to identify minorities from individual data.

resource allocation and improve productivity. One example of such data mining is SingStat's study of the modes used by respondents in past surveys in order to improve Census 2010 processes, hence reducing manpower and increasing productivity (see **Box 1**).

BOX 1

Pattern Spotting and Higher Productivity

How Singstat used data tracking to adapt Census 2010 operations and raise productivity

Data collection for the decennial census is no small task. Census 2010 involved 200,000 households for which information on 58 survey items was required. Respondents selected for the Census were notified by post and given the option to reply via several modes: completing via Internet form; via computer-assisted telephone interview (CATI); or when field interviewers visit their homes. The last mode was labour intensive and made more challenging by difficulties in recruiting interviewers. SingStat's aim for Census 2010 was to increase the use of CATI and internet modes, and minimise face-to-face interviews.

To achieve that, patterns of household profiles and mode usage from previous surveys were derived. Census 2010 surveys were dispatched in 20 staggered batches. Thus, loads on CATI and internet modes were smoothed out to facilitate utilisation and avoid disruption. SingStat also tracked daily patterns in each batch's response rate to schedule CATI call reminders just as internet response was tapering off. Finally, only after two reminders were dispatched, field interviewers were sent to visit homes. Compared to Census 2000, internet response rose from 15% to 38% while face-to-face field interviews fell from 22% to 16%; response rate remained at 98% within 52 days after each batch, even as manpower used fell from 290 to 140.

Insights from an agency's internal micro-data can be supplemented with aggregated data from other agencies to obtain a more holistic picture. Aggregated textual data are available from the Government's open data portal (www.data.gov.sg) and the SingStat website (www.singstat.gov.sg). Geospatial data—which are useful when relative locations and distances are important information in policy decisions—can be easily obtained from OneMap (www.onemap.gov.sg).

Finally, data are also available through specialised surveys, which collect in-depth answers on respondents' attitudes, behaviour or choices unavailable in existing data. Examples include the Health Promotion Board's National Health Survey and the Panel Study on Social Dynamics launched recently by the Institute of Policy Studies. Another vehicle of data collection is the field experiment. In it, individuals may be asked to choose among several options, from which researchers infer underlying preferences. As residents' preferences could determine public reaction to—and hence overall effectiveness—of a policy, these field experiments can be useful for policymakers. One example is MTI's study, conducted with MOM, on low-income residents' relative preference for wages paid entirely in cash, or partially in CPF as is the case now. Their results showed 14% preferred jobs with cash-only wages, but around half had a strong preference for part of their incomes to be saved in CPF accounts, implying that the current system remains suitable.[2] Hence, populist calls for cash-only wages may not, in fact, be welcomed by the majority.

The first step in understanding a dataset is to describe its data items (variables) in broad terms using summary statistics, of which the mean, or average, is most popular.[3] However, the mean alone may not be adequate. Policies often target certain groups, and may affect individuals within these groups to varying degrees.[4] Likewise, individuals vary in their characteristics. This means that

[2] For details, see "Cash versus CPF? Understanding the Preferences of Low-Income Residents through a Field Experiment", Economic Survey of Singapore, Second Quarter 2012, pp. 13-21.

[3] The summary statistics can be easily calculated with any basic statistical software package.

[4] "Individuals" here refer to individual units, which may be persons, households, companies, etc.

data would show distributions of individuals affected by a policy, as well as individuals' costs and benefits from that policy. Distributions are thus best described with several values, including average, median and percentiles (read more about distributions at **Box 2**).

BOX 2 Distributions

A normal distribution, more commonly known as the "bell curve", shows the number or proportion of individuals on the vertical axis, while the horizontal axis shows the variable's values (see **Figure 1**). The bell curve can be divided into 10 equal areas, with cut-off values at the 10th to 90th percentiles; e.g.,10% of individuals are between the 70th and 80th percentiles, and 90% have values below the 90th percentile. In a normal distribution, mean (μ) is approximately equal to median (the 50th percentile) and mode,[5] while its spread may be measured by standard deviation (σ), with around two-thirds of individuals within one standard deviation on either side of the mean (see **Figure 1**). A small σ shows a distribution around a narrow band of values near the mean, while a large σ indicates a "stretched-out" distribution. For example, among one group of residents, if benefits from Policy A have a larger σ than Policy B, it implies that benefits are more evenly distributed under the latter. This statistic could aid in assessing the relative strengths of the two policies.

Skewness is indicative of how far a distribution is from normal. Positive skewness indicates that a single-mode distribution is right-skewed.[6] That

[5] Median is the middle value when observations are ranked from least to greatest, with 50% of individuals on either side of the median. Mode is the value shared by the most number of individuals.

[6] Conversely, negative skewness implies a long left-hand-side tail of a single-mode distribution.

is, the bell curve has a long right-hand-side tail, with very high 95th- and 99th-percentile values, which pulls up the mean. When distributions are skewed, medians may be more representative than means as the former are less affected by very high, or very low, values. One application is income distributions, which tend to have high positive skewness (with some very high earners) so the average is greater than the median. Hence, a large majority of individuals earn less than the average income. This jars with the psychological tendency of most people to think of themselves as average, or above-average, in positive attributes.[7] In such cases, communications with the public could be enhanced with greater sensitivity to statistical skewness.

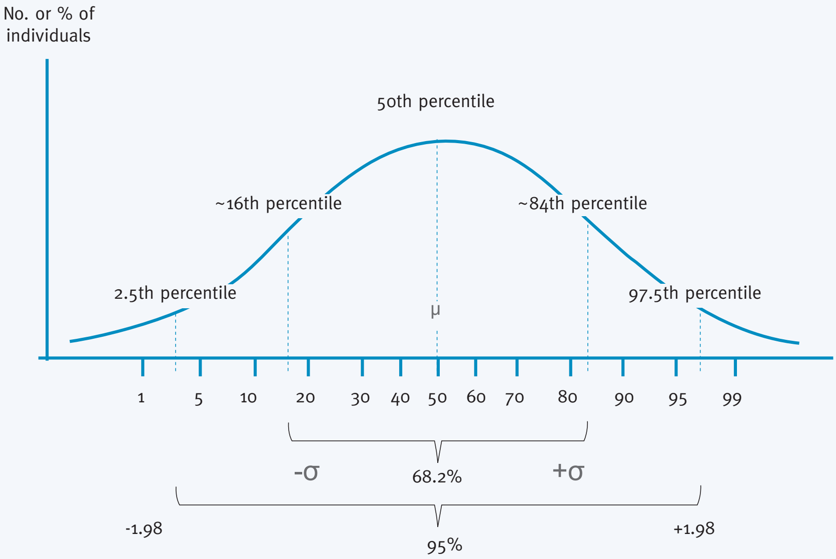

Figure 1: Percentiles in a Normal Distribution

[7] Psychologists term this cognitive bias "illusory superiority".

When Can Data Analytics Be Used?

Data analytics has a role to play at every stage of the policy cycle. Analytics need not be complex to be useful. For example, apart from describing a dataset, summary statistics can also provide a quantitative "feel of the ground" at the initial stages of policy development. They also paint a picture of the operating environment, as well as provide the basis for policymakers to derive desired policy outcomes and to track progress.

Problem Identification and Analysis

Data analytics can be used to identify and analyse problems before deciding what policy to implement. Suppose the intent of a hypothetical policy is to encourage housewives from low-income households (defined here as earning less than $1,000 per month) to increase their paid working hours to be similar to those of housewives from higher-income households. The policy affects individuals from eligible households ("treated"), but not housewives from households earning slightly more, e.g., $1,001—$1,100 per month ("controls"). A distribution plot of *current* working hours for the two groups can reveal the gap that the policy seeks to narrow (see **Figure 2**). Coupled with other quantitative evidence, and even qualitative information (e.g., interviews and case studies), researchers can infer factors behind the gap between the groups.[8]

Public officers may also quantify policy targets using desired distributions. For instance, should the desired outcome be a working-hours distribution of the treated that is similar to the controls? Or should it involve increasing the working hours of treated housewives who are already working at least 5 hours a week (see **Figure 2**)? By approximating a targeted outcome distribution, policymakers can use periodic data profiles of the two groups as one simple way to track policy progress.

[8] The distance between the means (or medians) of two distributions can be tested for statistical significance. For instance, the t-test, with or without adjusting for different standard deviations, can be used to test if the distance between the means of two distributions is statistically different from zero. For details on some of the more common tests, see Moore et al. (2009).

Before the initiation of a new policy, there may be a need to determine whether current policies are effective or where current policy gaps are. Such evaluation may be undertaken with administrative data or a survey.[9] A survey is particularly useful in obtaining a general sensing of an issue—to understand perspectives and opinions—which available data are unable to reflect. One example of this is an MTI-IAL telephone survey of manufacturing firms in Singapore on management practices in operations, monitoring, targets and human resources.[10] The survey found that Singapore's management was sixth best in the world, and noted that scores were higher for larger firms. It therefore recommended that interventions to improve management quality be focused on small and medium-sized enterprises.

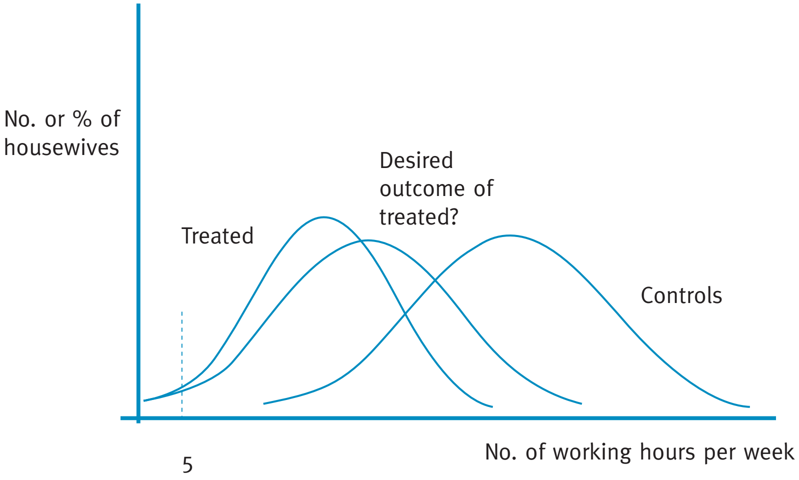

Figure 2: Outcome Distributions

[9] One example of a policy evaluation using administrative data is the study on how wages and decisions to enter into the workforce changed as a result of the Workfare Income Supplement scheme. See "The Impact of the Workfare Income Supplement Scheme on Individuals' Labour Outcomes", Economic Survey of Singapore, Second Quarter 2014, pp. 27–36.

[10] See "Management Practices in Singapore: Measuring and Explaining Management Practices across Manufacturing Firms in Singapore", A Report by the Policy, Research and Benchmarking Working Group of the National Productivity and Continuing Education Council (October 2013).

Exploring Alternatives, Predicting Effects

Data analytics can also provide critical inputs and guidance when formulating policy content. An appropriate tool is multivariate regression, which approximates statistical associations, or correlations, between each factor and the outcome that policymakers intend to influence (see **Box 3**).[11] Each factor that is correlated with the outcome points to a possible policy lever. For instance, regression analyses may show that, holding other factors constant, the greater the commuting time from home to workplace, the fewer the paid working hours done by housewives. Such a relationship suggests that one possible lever to encourage housewives to take up more paid employment would be to match them to jobs closer to their homes.

[11] A good starting reference is Gujarati (2011).

BOX

3

Multivariate Regressions

Are there relationships between variables in a dataset?

Regressions estimate the relationship between two variables. They can be run on a variety of data: cross-sectional, in which a group of individuals is measured for different variables at a certain time point; or time series, in which variables are measured over time but the composition of individuals differs at each time point. Longitudinal (or panel) data track the same individuals across time. A basic linear multivariate regression model with two independent or explanatory variables *(X)* may be expressed as:

$$\textit{Outcome (Y)} = \textit{Constant} + b_1.X_1 + b_2.X_2 + \textit{error}$$

Correlation, or statistical association, between independent variable X_1 (or X_2), and outcome is measured by b_1 (or b_2). Such correlations may be obtained using Ordinary Least Squares, which estimates the constants, b_1 and b_2, by minimising the squared-error terms summed up over all the individuals. The direction of the relationship between X_1 and outcome is indicated by the sign of b_1: a positive (negative) sign indicates that a rise in X_1 is associated with a rise (fall) in outcome. When data are in levels, b_1 is interpreted as the extent to which average value of the outcome changes with a one-unit change in X_1, holding all else constant.

These simple regressions can be run on many common statistical packages. It is worth noting that correlation is not causality. This is because a relationship with another variable outside the model may in fact be driving both the independent variable and outcome (omitted variable bias). Alternatively, the outcome could also be influencing the independent variable (reverse causality). If the intent of the policy research is to estimate causal links, then the researcher may need to employ certain quasi-experimental techniques (see **Box 4**).

Regression analyses also have the advantage of weighing the relative strengths of different correlations, providing insight into which policies are more likely to translate into desired outcomes. In the case of housewives' paid employment, a regression-based study may find that, while commuting time is a key factor, availability of childcare is even more important. Policy focus and content will then take a very different shape with this knowledge.

Apart from comparing various policy levers, regression correlations also provide a simple, but useful, basis for simulating outcomes under different scenarios of a proposed policy. Continuing with the hypothetical example of increasing paid working hours of targeted housewives: Policymakers planning locations of childcare centres may approximate, from geospatial and demographic data, how many targeted families each new centre could serve. Using scenario planning for different take-up rates at each centre, as well as regression correlations, they can derive a range of rough estimates of the likely increase in targeted housewives' paid working hours.[12] Such "broad strokes" approximations are one approach to assess the impact of different options *before* policy implementation.

One purpose of quantifying the impact of policy alternatives is to weigh them against policy costs, including opportunity costs of the funds involved. Where a policy option's costs clearly outweigh its benefits, alternatives or changes to policy content are needed. Even though the data may not include all aspects of a cost-benefit analysis—either because they are unobserved or take a long time before they can be measured—conducting data analytics prior to policy implementation can flag up those options that require deeper consideration. That alone makes the exercise worthwhile.

[12] For instance, suppose the regression coefficient between having a nearby childcare centre and working hours per week is 10, this means that adding a centre is likely to increase working hours of housewives by 10 hours a week on average. Suppose also that a new centre serves 1,000 families. Under the scenario that 10% of the 1,000 families earn below $1,000 per month, the likely average impact on targeted housewives is 10 hours for 100 families. The regression model may also include interaction terms to estimate different correlations for treated and controls.

Evaluating Policy Outcomes

After a policy has been implemented, data analytics can be used to evaluate it. First, it sheds light on whether the original policy intent is indeed met, and to what extent. Second, it provides more precise estimates of the beneficial impact of a policy, which policymakers can weigh against costs. Third, as Singapore's operating environment changes, what worked in the past may not work as well in the future. Quantifying the impact of specific policies is thus a window through which structural changes may be contextualised and understood. For instance, an evaluation study by MTI economists shows that less educated and older individuals were incentivised by Workfare Income Supplement payments to enter and stay in the workforce.[13] Their response to policy is also reflective of wider societal characteristics, such as smaller family sizes and relatively few elderly workers. Future interventions may have different effects as these characteristics change, i.e., if the number of elderly workers grows.

Evaluation, such as running a randomised controlled trial (RCT), may also be undertaken after a policy trial run on a small group of individuals. RCTs involve obtaining a representative sample and assigning policy treatment *randomly* to some individuals, who make up the treated group, but withholding it from the rest, who then make up the control group. The two groups are alike statistically in their characteristics, with policy treatment being the only real difference. Any statistical difference in the outcome may therefore be attributed solely to the policy.[14] Data analytical tools for RCT datasets are relatively straightforward, and include testing differences in means, medians or distributions between the treated and control groups.[15]

[13] For details, see "The Impact of the Workfare Income Supplement Scheme on Individuals' Labour Outcomes", Economic Survey of Singapore, Second Quarter 2014, pp. 27-36.

[14] When assessing RCT results, policymakers need to keep in mind their applicability when interventions are scaled up to a much larger, or national, level. As RCTs are conducted on a small scale, the implicit assumption is that the treated group has no information or effect on the control group, and vice versa. This may not be applicable when many more people are affected by the policy.

[15] See Moore et al. (2009) for a compendium of these statistical tests.

For policies which do not utilise RCTs, evaluation involves applying quasi-experimental statistical techniques (see **Box 4**). Broadly speaking, these techniques estimate policy effects by comparing observed outcomes of individuals who were affected by the policy with what those individuals would experience in the *absence* of that policy ("counterfactuals"). Unfortunately, counterfactuals cannot be measured as individuals are either affected by the policy, or not. Moreover, there are selection issues, in that some individuals may change their behaviour to avoid, or to qualify for, the policy under evaluation. Hence, these techniques seek to resolve such issues partially by choosing a control group which is as similar as possible to the counterfactual, in order to mimic a random experiment.[16]

[16] See World Bank (2011) for a non-technical introduction to quasi-experimental techniques.

BOX

4

Some Quasi-Experimental Techniques

Getting the control group as close as possible to the counterfactual

The causal effect of policy is the difference between the observed outcome and what-could-have-been in the absence of the policy, or the counterfactual. It is not straightforward to establish causal links. For example, the difference between what workers earned before and after a training programme may not be attributable to the training. This is because other factors could be at play, e.g., improving economic conditions, or that workers who go for training were already on a higher wage trajectory. As the counterfactual is unobserved—the individual is either treated by the policy or not— the three common techniques used to identify causal links include:

Difference-in-difference	**Propensity score matching**	**Regression Discontinuity Design**
Intuitively, this model compares the before-and-after outcomes for the treated (first difference), with the before-and-after outcomes of the controls (second difference). The first difference controls for factors which are constant over time in the treated group; the second difference controls for time-varying factors among the controls (which also apply to the treated). When the latter is subtracted from the former, what remains is the time-varying factor (the policy) on the outcome of the treated. Hence, the difference in *trends* is assumed to be due to policy.	Matching involves selecting as controls – among individuals who are not treated by a policy – those who are closest in characteristics to the treated. This is done by estimating "propensity scores", or the probability that an individual will be treated, given those characteristics, for both treated and non-treated. Then, those non-treated with the closest scores to the treated are chosen as controls. The difference between the matched controls and treated is the policy effect.	This model makes use of the eligibility criteria in many policies to compare those treated who just meet the criteria (e.g., income of $9,701–$10,000 with a cut-off of $10,000) with those controls who fall just outside the cut-off (e.g., $10,001–$10,300). The underlying assumption is that as the cut-off is exogenously determined by policy, individuals randomly fall on either side of the cut-off. Hence, controlling for other factors, the difference in their outcomes could be attributable to the policy.

Challenges and Limitations

Although data analytics holds much promise, there are limitations that practitioners need to be aware of. The main pitfalls—and their possible remedies—include:

Prioritising Size Over Representativeness

While very large datasets have the advantage of picking up small statistical associations, it may not mean that they are necessarily representative of the whole population. Data representativeness is a key condition for results to be generalised. For instance, data which depend on internet access, e.g., word searches on internet search engines, are unlikely to be representative of the Singapore population. This is because some segments of the population, such as the elderly, are less likely to be included and even those who are included may themselves not be representative of their segment. One solution is to limit studies' findings to only the groups they do represent and avoid over-generalising.

Sensitivity of Results to Variable Definitions

Although many data items have widely accepted definitions, some will be defined by the officer. For instance, in the hypothetical case of raising paid working hours of housewives from low-income households, the officer makes decisions on what is considered "low-income", as well as whether to count informal employment undertaken by the housewives (e.g., paid babysitting) as part of their paid working hours. The study's final results may be sensitive to these decisions.

One remedy is to perform robustness checks: repeat the analytics with different definitions to see if the original results hold. If findings differ, the officer could use them to gain additional insights into policy effects. In the hypothetical example, the policy may be less effective when housewives are already engaged in informal employment.

Sensitivity of Results to Model Choice

Just as researchers make decisions on how to define certain data items, they also choose the regression models or quasi-experimental techniques. Findings will typically depend on these choices. Testing across different models can assist in deciding which is more credible. More broadly, each model should be viewed as one estimate, and the researcher should aim for a consensus among different estimates.

Conclusion

The strength of data analytics is to provide a systematic, yet simplified, perspective through numbers. However, articles on data analytics tend to focus exclusively on "success" stories. An often cited example of data analytics success is US department store Target's data mining of past consumer shopping patterns, which it used to infer whether its female patrons were pregnant. Target then sent coupons for related products (e.g., baby cribs) to entice these women to shop with them. Its success is celebrated with the famous case of the retailer knowing that a teenage girl was pregnant even before her father did.[17]

Some healthy scepticism is useful. Cases when data analytics works should be weighed against cases when it does not. For example, how often did the store send coupons to women who were in fact not pregnant? Consumers of quantitative research should be alert to one-sided portrayals.

[17] For details on Target's use of customer data in the US, see http://charlesduhigg.com/new-york-times-magazine/.

In a complex environment, policymakers need to go beyond purely relying on quantitative methods to design effective policies. Data analytics needs to be seen as one of the tools that policymakers can draw on to measure, to understand and to improve the policymaking process, thereby pointing to what may work and what may not.

References

Gujarati, Damodar N. *Econometrics by Example*. Hampshire, UK: Palgrave Macmillan, 2011.

Ministry of Trade and Industry. "Cash versus CPF? Understanding the Preferences of Low-Income Residents through a Field Experiment". *Economic Survey of Singapore Second Quarter 2012,* 13–21.

Ministry of Trade and Industry. "The Impact of the Workfare Income Supplement Scheme on Individuals' Labour Outcomes". *Economic Survey of Singapore, Second Quarter 2014,* 27–36.

Moore, David S., George P. McCabe, and Bruce A. Craig. *Introduction to the Practice of Statistics*. 6th ed. New York: W. H. Freeman and Company, 2009.

Sondergaad, Peter. "Gartner Says Worldwide Enterprise IT Spending to Reach $2.7 Trillion in 2012". Press Release, October 17, 2011. Accessed November 2, 2015. www.gartner.com/newsroom/id/1824919.

Gertler Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. *Impact Evaluation in Practice*. Washington, DC: World Bank, 2011.