

# Singapore Society in Numbers

*Edited by Shannon Ang*

*Last updated 10 February 2020*

# Contents

<b>Preface</b>	<b>4</b>
Why I started this project . . . . .	5
How to contribute . . . . .	6
Acknowledgements . . . . .	7
About me . . . . .	7
<b>I Datasets for Social Science</b>	<b>8</b>
<b>1 Public Data</b>	<b>9</b>
1.1 Asian Barometer Survey (ABS) . . . . .	9
1.2 IPS Public Data . . . . .	10
1.3 Panel on Health and Ageing of Singaporean Elderly (PHASE) . . . . .	10
1.4 World Values Survey (WVS) . . . . .	11
<b>2 Restricted Data</b>	<b>12</b>
2.1 Department of Statistics (DOS) Microdata . . . . .	12
2.2 NVPC Individual/Corporate Giving Study . . . . .	13
2.3 National Youth Survey (NYS) . . . . .	13
2.4 Retirement and Health Study (RHS) . . . . .	14
2.5 Singapore Life Panel (SLP) . . . . .	14
2.6 Singapore Panel Study on Social Dynamics (SPSSD) . . . . .	14
2.7 Youth Survey on Transitions and Evolving Pathways in Singapore (Youth STEPS) . . . . .	15
<b>II Think Pieces</b>	<b>16</b>
<b>3 Thinking about Numbers</b>	<b>17</b>
<b>III Case Studies</b>	<b>18</b>
<b>4 Blown out of proportion</b>	<b>19</b>

<b>CONTENTS</b>	<b>3</b>
4.1 Support for the Watain ban . . . . .	19
4.2 Web-savvy Seniors? . . . . .	21
4.3 Technical excursus . . . . .	24
4.4 Conclusion . . . . .	25
4.5 Additional reading . . . . .	26
<b>5 Are we lonely?</b>	<b>27</b>
5.1 The lonely dichotomy . . . . .	27
5.2 Lonely by whose standard . . . . .	28
5.3 Same data, different results . . . . .	30
5.4 Conclusion . . . . .	32
<b>6 Size is not all that matters</b>	<b>34</b>
6.1 But it's too small! . . . . .	34
6.2 Size matters, but how so? . . . . .	36
6.3 Non-probability sample, now what . . . . .	40
6.4 Conclusion . . . . .	43
<b>7 Measuring societal change</b>	<b>44</b>
7.1 Three ways to understand change . . . . .	44
7.2 Declines in marriage and divorce . . . . .	45
7.3 Living longer than you expect . . . . .	53
7.4 Conclusion . . . . .	57
<b>8 You say; I say; Who confirm?</b>	<b>58</b>
8.1 SDP vs. MOM . . . . .	58
8.2 Really (just) golfing . . . . .	65
8.3 Conclusion . . . . .	69
<b>9 Your case study</b>	<b>70</b>

# Preface

## Note to Readers

This book is in Open Review. I want your feedback to make the book better for you and other readers. To add your annotation, select some text and then click the on the pop-up menu. To see the annotations of others, click the in the upper right hand corner of the page .

This online book is a compilation of resources aimed at advancing quantitative social science in Singapore. It is meant to be a ‘living document’, so it will be updated as frequently as possible. The main goal is to promote interest, rigour, and transparency in trying to understand Singapore society through quantitative lenses. It does so by:

1. **Providing information on Singapore-relevant datasets** that are currently used to answer research and policy questions (Chapter 1 and Chapter 2). This includes:
  - Descriptions of *publicly available* datasets and how to access them. This overview of the ‘data landscape’ will be helpful for social scientists to get started with research on Singapore, and prevent wasteful overlap in primary data collection across institutions.
  - A list of *restricted* or *non-publicly available* datasets that could be used to answer important research or policy questions if access was granted. If available, details on the dataset and reasons for data restriction will also be listed. It is hoped that this list will promote greater transparency in data sharing across research teams.
2. **Occasional think pieces by researchers** on best practices and on how to improve quantitative social science in Singapore (Chapter 3).
3. **Maintaining a repository of replicable case studies on Singapore society** (with annotated code, where possible) which can be used for illustrations in any quantitatively oriented college-level class (Chapter 4 onwards). These may be short summaries (blog-length) of published work, or side analyses that may not be appropriate for an academic journal but are useful for Singapore social science nonetheless.

I am actively looking for contributors (go here to see how you can con-

tribute). Readers with ideas on how to improve this resource (or who may wish to help me maintain it) may use the in-built annotation feature, or email me at shannon.ang@ntu.edu.sg.



This book is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

## Why I started this project

Quantitative research is not (and should not be) the only approach we take to understanding Singapore society, but constant appeals to “big data”<sup>1</sup> or claims of “evidence-based policy”<sup>2</sup> makes it ever more important for members of the public to **critically evaluate the use of numbers** in making arguments or in representations of social phenomena.

Educational institutions have an important role to play in this “data-driven” world. Every year, undergraduates studying the social sciences in our local universities take several courses in research methods to fulfil the requirements of their degrees. Part of this research methods sequence typically involves training in introductory statistics or “quantitative reasoning”. Quantitative courses in social science departments differ from those taught in the natural sciences because they are thought to be more applied - the focus is on the use of statistical methods to answer questions about society. Understanding and applying these methods **to the Singapore context** is crucial here - at this point, students learn about (and hopefully are inspired by) the kind of questions they can ask about the very society they live in, given the quantitative tools they are learning.

However, my first exposure to statistics as an undergraduate reading Sociology at NUS<sup>3</sup> was to textbooks containing examples from only Western societies (e.g., Agresti and Finlay, 2009; Treiman, 2009). While the use of these internationally-recognized textbooks may provide some assurance of quality education, sole reliance on foreign material often becomes a missed opportunity to inspire students to build on and improve Singapore social science. Without contextualization<sup>4</sup>, abstract statistical concepts (e.g., hypotheses testing, chi-squared tests) seem removed from everyday experience, and impede the ability to take these important concepts beyond the classroom and into public dialogue.

---

<sup>1</sup>See, for instance, <https://www.todayonline.com/singapore/business-big-data-singapore-has-built-cutting-edge>

<sup>2</sup>Government agencies such as the Ministry of Social and Family Development often use such a phrase.

<sup>3</sup>(the) National University of Singapore

<sup>4</sup>Notwithstanding the terribly unhelpful stereotype of social science students being “good at writing but bad at numbers”.

I started this book with the view to use it primarily *as a teaching tool*<sup>5</sup>, but it can be used in many other ways. In the long term, I hope that resources in this book will encourage quantitative literacy and research in Singapore by making it easier for interested parties to browse, use, and understand Singapore-relevant data. Social science researchers may use the dataset listings as a springboard for collaboration, or contribute their own interesting case studies for the benefit of the Singapore public. Others (such as journalists, civil servants, or non-profit organizations) may find value in these material as a gateway to quantitative research on Singapore society, and how to think carefully about pertinent issues surrounding such work.

*For Singapore social science.*

## How to contribute

### To list a dataset

I will progressively list any Singapore-related dataset that I know of, but my knowledge is far from exhaustive. To help me out, you can simply alert me to an unlisted dataset and let me know where to find information on it (and I will write it up). However, if you could do the following and send it to me, it would make my task much easier:

Write a short paragraph on the dataset which includes (but is not limited to) the following information:

- Basic details about the dataset (name, how was data collected, how many observations etc.)
- Name(s) of the Principal Investigator(s) (and links to their web-page/profiles, if possible)
- How to access the dataset (e.g., a website that allows a direct download or lists instructions to obtain the data)

*Publicly available* datasets are basically datasets that can be downloaded freely or for which access can obtained through a simple procedure (e.g., signing up for an online account, sending a form with a simple research topic). *Restricted* or *Non-publicly available* datasets are those that require more extensive clearance (e.g., a background check, use of a data enclave) to access. Data for which there is no information on access also fall under this criteria. Email me at shannon.ang@ntu.edu.sg to list a dataset.

### To write a think piece

Think pieces can be of any length (suggested length is 500-2000 words) and on any topic related to quantitative social science. For instance, pieces may comment on the state of quantitative social science in Singapore (e.g., what is

---

<sup>5</sup>For instance, the public repository of Singapore-oriented examples and illustrations may be used to supplement courses based on textbooks written by international scholars.

lacking, how we can do better) and/or provide tips for social scientists seeking to study Singapore (e.g., how to write grant applications, where to find data). That said, I am still mulling over whether this should be an invite-only section, or have pieces go through some kind of review mechanism (I do not wish to be the sole arbiter of what goes up here). Nevertheless, email me at shannon.ang@ntu.edu.sg if you think you can contribute a think piece (or have someone to suggest).

### To contribute a case study

Case studies are meant to illustrate a point about Singapore as a society or quantitative methods in general. These may include blog-length summaries of published research, or smaller side analyses that are useful for Singapore social science. I am hoping that most of these case studies include some form of data analysis, and/or relate a useful (quantitative) concept to the Singapore context. Replicable case studies are preferred (i.e., anyone familiar with some code<sup>6</sup> can reproduce results), but researchers unable to share code (and data) should not disqualify themselves from contributing. Further, this online book is written using the Rmarkdown language, so it would be helpful if contributors are familiar with it - but this is not a prerequisite. Email me at shannon.ang@ntu.edu.sg if you have an idea for a case study, and we can work together to make it happen.

## Acknowledgements

This book is being written through the **bookdown** package (Xie, 2019), which was built on top of R Markdown and **knitr** (Xie, 2015).

Contributors other than me include:

## About me

Little write-up about myself, which I will insert in due course... For now, refer to my personal website.

---

<sup>6</sup>Code can be in any language - R, Stata, Python, SAS, Mplus etc.

# **Part I**

# **Datasets for Social Science**

# Chapter 1

# Public Data

This section contains a list of public datasets available for social scientists to analyze. Datasets should be in disaggregated form in a way that is useful for academic work<sup>1</sup> and social research. For each one, a brief description of the dataset, the investigators, and details for access will be included. It is expected that this list will continue to grow.

Aggregated administrative data, like those that can be found on <https://data.gov.sg> or from government reports, are not really the focus here. *However*, information about data that can be linked to disaggregated datasets (e.g., data on neighbourhood characteristics, or other forms of contextual data) to improve analyses are welcomed.

Readers who believe I have provided wrong information on any of these datasets can contact me, so that I can correct it. Use the in-built annotation feature, or send me an email at [shannon.ang@ntu.edu.sg](mailto:shannon.ang@ntu.edu.sg) if you know about a dataset that should be featured in this list but is not included here.

## 1.1 Asian Barometer Survey (ABS)

The Asian Barometer Survey is a cross-country, longitudinal study (repeated cross-section) of public opinion in about 21 countries. Data was collected through face-to-face interviews with adults aged 21 and above. The survey assesses opinions on issues such as political values, democracy, governance, human security, and economic reform. Singapore has participated in 3 out of 4 waves of data collection - 2006 (N=1012), 2010 (N=1000), and 2014 (N=1039).

ABS is coordinated by the Hu Fu Center for East Asia Democratic Studies at the National Taiwan University.

---

<sup>1</sup>In essence, academics should be able to publish from it.

Instructions for data access can be found at <http://www.asianbarometer.org/data/data-release>.

## 1.2 IPS Public Data

The IPS Public Data Sharing Platform is not a single dataset, but a repository that houses data collected by the Institute of Policy Studies at the Lee Kuan Yew School of Public Policy (NUS). These data range from surveys on political information, perceptions of policies, to food prices by neighbourhood districts. The available datasets are listed below:

- 1) Survey on Political Traits and Media Use
- 2) Impact of New Media on General Election 2011
- 3) POPS (6): Perceptions of Singles on Marriage and Having Children
- 4) POPS (7): Perceptions of the Marriage & Parenthood Package (2013)
- 5) IPS Study on Perceptions of Singapore's history
- 6) POPS (8): IPS Post-Election Survey 2015
- 7) Perceptions of Governance Survey
- 8) Internet and Media Use During GE2015
- 9) The Makan Index – A Survey of Hawker Food Prices
- 10) Makan Index 2017: An indicator for Cost of Eating Out in Singapore

More information on these datasets and instructions for data access can be found at <https://lkyspp.nus.edu.sg/ips/research/surveys/public-data-sharing-platform>.

## 1.3 Panel on Health and Ageing of Singaporean Elderly (PHASE)

The Panel on Health and Ageing of Singaporean Elderly is a longitudinal panel study that tracks changes in the physical, social and mental health of Singapore residents. Respondents are aged 60 years and above in 2009, and three waves of data collection have been conducted so far. Wave 1 was conducted in 2009 (N=4990), Wave 2 in 2011 (N=3103), and Wave 3 in 2015 (N=1572).

PHASE includes data on physical health, mental health, social engagement (e.g., loneliness, social participation, social networks), income, employment, and housing. Anthropometric and performance measurements, including blood pressure, sitting and standing height, waist circumference, body weight and hand grip strength, were also conducted.

Principal Investigators for PHASE are Associate Professor Angelique Chan, Professor David Matchar, and Assistant Professor Rahul Malhotra at Duke-NUS Medical School.

Instructions for data access can be found at <https://www.duke-nus.edu.sg/care/>

research/dataset-codebook.

## 1.4 World Values Survey (WVS)

The World Values Survey is a global study to help social scientists understand changes in the beliefs, values and motivations of people across multiple countries. To date, there have been 7 waves of data collection (repeated cross-section, not panel) from almost 100 countries. The survey includes questions on societal trust, religion, work, security, and politics. Singapore participated in Wave 4 (2002, N=1512) and Wave 6 (2012, N=1972). Principal Investigators are Associate Professor Tan Ern Ser at NUS (2002), and Associate Professor Vincent CH Chua at SUSS (2012).

Data can be accessed at <http://www.worldvaluessurvey.org>.

# Chapter 2

## Restricted Data

This section contains a list of datasets that are potentially useful for social scientists to analyze, but for which access is restricted. The focus here is not on small studies (e.g., randomized controlled trials for a small subset of the population), but on large datasets that can be of use to social scientists across disciplines.

Other than a description of the data (as far as possible), this section will also include information on how restrictions may be lifted (i.e., how to gain access), if possible. It is hoped that listing them here will promote transparency in data sharing across research teams, and eventually prevent wasteful overlap in primary data collection across institutions. As in the public data section, disaggregated data that can be used for research is the focus here.

Readers who believe I have wrongly listed a dataset here (i.e., they are not restricted), or have more accurate information than provided, can contact me. Use the in-built annotation feature, or send me an email at [shannon.ang@ntu.edu.sg](mailto:shannon.ang@ntu.edu.sg) if you know about a dataset that should be featured in this list but is not included here.

### 2.1 Department of Statistics (DOS) Microdata

The Singapore Department of Statistics (DOS) regularly conducts a number of surveys that, if access to disaggregated data were provided, could advance Singapore social science greatly. These include the Household Expenditure Survey, the General Household Survey, as well as records from the Registry of Marriages to track marriage and divorce trends in Singapore. Disaggregated data (i.e., microdata) allow for deeper analysis than what is presented in the public domain - for instance, social scientists may examine age-period-cohort trends in marriage and divorce rather than be confined solely to period statistics (which may be

misleading), or to decompose within- and between- group variance in household income/expenditures for a better understanding of inequality in Singapore.

Unfortunately, the DOS does not share disaggregated data with researchers. When asked, the chatbot on the DOS website states:

The Department of Statistics (DOS) is unable to provide access to raw data or anonymised microdata in the public domain. You may wish to refer to our publications on the surveys conducted for the latest survey data. If the required aggregated data is not available in the reports, you could provide us with the table formats. Please send them to [info@singstat.gov.sg](mailto:info@singstat.gov.sg). We will assess the availability of the data and, if customised and aggregated data generation is required, we will let you know the applicable cost-recovery fees.

Further, under section 7 of the Statistics Act, “anonymised microdata” (i.e., disaggregated data) obtained using the powers of the Act (i.e., where individuals are legally compelled to provide information) can only be shared with other public agencies, or “any consultant commissioned by a public agency”. This means that social scientists outside of the government can only access these data by being appointed as a consultant by a public agency, unless the Statistics Act is amended to include other classes of persons<sup>1</sup>.

## 2.2 NVPC Individual/Corporate Giving Study

The NVPC Individual/Corporate Giving Study is a biennial, repeated cross-section survey that has been conducted since the year 2000. Therefore, 10 waves of data from 2000-2018 are potentially available. Its key aim is to study trends in volunteerism and philanthropic giving over time.

The study is administered by the National Volunteer and Philanthropy Centre (NVPC). Some details on the latest wave can be found at the NVPC website - <https://www.nvpc.org.sg/index.php?p=resources/individual-giving-study-2018-findings>. Melissa Kwee, the Chief Executive Officer of NVPC, has publicly stated her desire for open data sharing, but it is unclear how to access data from the NVPC Individual/Corporate Giving Study, or how to contact the research team. There are no indications of plans to make the dataset publicly available.

## 2.3 National Youth Survey (NYS)

The NYS is a cross-sectional survey of youth in Singapore that is conducted every three to five years. Its goal is to understand the aspirations, challenges, and attitudes of young Singaporeans aged 15-34. The survey has been conducted five times thus far, in 2002, 2005, 2010, 2013 (N=2,843), and 2016 (N=3,531).

---

<sup>1</sup> According to Section 7(2)(b)(ii), such classes may be specified in the Third Schedule.

NYS is administered by the National Youth Council (NYC)'s research team. Details can be found at the NYC website - <https://www.nyc.gov.sg/en/initiatives/resources/national-youth-survey/> - which hosts illustrated PDF summaries of the survey findings from selected years. It does not list any plans to make the dataset publicly available.

## 2.4 Retirement and Health Study (RHS)

The RHS is a longitudinal survey of Singapore residents' retirement and health-care needs and how they change over time. It is conducted by the Central Provident Fund Board (CPF). This is a panel study of individuals aged 45 to 85 in 2014, with the same individuals being interviewed once every two years (for ten years, beginning in 2014). The survey includes information on household expenses, employment, health, and financial status. This is a large study with potentially many uses - RHS purports to have reached out to more than 23,000 participants in the first two rounds of interviews<sup>2</sup>.

More information is available at [https://www cpf gov sg/Members/Others/member-pages/retirement-and-health-study-\(rhs\)](https://www cpf gov sg/Members/Others/member-pages/retirement-and-health-study-(rhs)). The RHS website does not list any plans to make the dataset publicly available, and is accessible only to government agencies or researchers working with/for government agencies. The RHS study team can be reached at [cpf\\_rhs@cpf.gov.sg](mailto:cpf_rhs@cpf.gov.sg).

## 2.5 Singapore Life Panel (SLP)

The Singapore Life Panel administers an internet-based monthly survey to approximately 11,000 Singaporeans aged 50 to 70 years. Information on income, expenditure, health, work and housing choices are solicited from panel members. The panel survey is one of the largest population-representative monthly surveys conducted in the world. This study is unique in Singapore because of the monthly frequency of surveys - the sheer amount of data collected over time is unprecedented.

The Singapore Life Panel is housed at the Centre for Research on the Economics of Ageing at SMU. The SLP website does not list any plans to make the dataset publicly available outside of their own research team. The SLP team can be reached at [crea@smu.edu.sg](mailto:crea@smu.edu.sg).

## 2.6 Singapore Panel Study on Social Dynamics (SPSSD)

The Singapore Panel Study on Social Dynamics was started in 2014 to understand challenges related to family cohesion and functioning. It is a longitudinal

---

<sup>2</sup>[https://www cpf gov sg/Assets/members/Documents/RHS\\_FAQ\\_Booklet pdf](https://www cpf gov sg/Assets/members/Documents/RHS_FAQ_Booklet pdf)

## 2.7. YOUTH SURVEY ON TRANSITIONS AND EVOLVING PATHWAYS IN SINGAPORE (YOUTH STEPS)15

panel study, starting with 5002 heads of households interviewed in 2014. Its purpose is to measure family dynamics, societal values and attitudes relevant to national identity and social mobility over time. Waves 1 to 4 of data collection have been completed<sup>3</sup>, with Wave 5 beginning in 2019. A strength of this study is its panel data - few Singapore datasets with rich data on family dynamics have this many waves collected from the same individuals over time.

The Principal Investigator for this study is Dr Natalie Pang at the Institute of Policy Studies, NUS. The SPSSD website does not list any plans to make the dataset publicly available, probably because of restrictions imposed by government funders. The SPSSD study team can be reached at ips.soclab@nus.edu.sg.

### 2.7 Youth Survey on Transitions and Evolving Pathways in Singapore (Youth STEPS)

Youth STEPS is a six-year longitudinal survey of Singaporean youth aged 17 to 24 that began in 2017 (N=4,041). It covers a wide range of topics, including respondents' educational and career trajectories, family relationships, civic participation, and attitudes regarding social mobility, family, marriage, and more. The survey is in its third wave (as of 2019) and is scheduled to be completed in 2022.

Youth STEPS is funded by the National Youth Council and administered by the Institute of Policy Studies (Social Lab), part of the Lee Kuan Yew School of Public Policy. The Principal Investigator is Dr Leong Chan-Hoong. The website does not list plans to make the data publicly available. Questions can be directed to ips.soclab@nus.edu.sg.

---

<sup>3</sup>See some basic descriptives here: [https://lkyspp.nus.edu.sg/docs/default-source/ips/2018-07-11\\_spssd-wave2\\_english.pdf](https://lkyspp.nus.edu.sg/docs/default-source/ips/2018-07-11_spssd-wave2_english.pdf)

## **Part II**

### **Think Pieces**

## **Chapter 3**

# **Thinking about Numbers**

Think pieces section. Empty for now, but would you like to contribute? Email me at shannon.ang@ntu.edu.sg.

## **Part III**

# **Case Studies**

# Chapter 4

## Blown out of proportion

Contributor: Shannon Ang

Date: 21 May 2019

Proportions (sometimes expressed in percentages) are commonly used in popular media to reflect public opinion. In fact, it is often the only type of statistic we get<sup>1</sup> to evaluate as “evidence”. For instance, a news article may state that “nearly 46 per cent of those aged 18 to 25 would allow extremist views that deem all other religions as enemies to be published”<sup>2</sup>, or that “59 per cent of Chinese find a Malay president acceptable”<sup>3</sup>. While these proportions are easy for the general public to understand, they can be misleading if not read carefully. This case study looks at two different news articles, showing how some claims can be exaggerated by careless use of numbers.

### 4.1 Support for the Watain ban

Swedish black metal band Watain was supposed to perform in Singapore on 7 March 2019. However, the gig was cancelled just hours before it was scheduled to begin, with the government citing concerns from the Christian community<sup>4</sup>. To evaluate public sentiment towards this incident, REACH<sup>5</sup> conducted a poll with 680 Singaporeans aged 15 and above. Of interest here is how results from this poll was represented in public discourse.

---

<sup>1</sup>Reports such as those released in the form of IPS working papers, sometimes include multivariable analysis, but often after many pages of crosstabulations

<sup>2</sup><https://www.todayonline.com/singapore/nearly-1-2-young-singaporeans-open-extremist-views-being-posted-online-survey-shows>

<sup>3</sup><https://www.straitstimes.com/singapore/majority-willing-to-accept-president-or-pm-of-another-race-but-prefer-one-of-their-own>

<sup>4</sup>See <https://www.channelnewsasia.com/news/singapore/watain-concert-cancelled-christian-community-reaction-shanmugam-11399434>

<sup>5</sup>The Singapore Government’s feedback unit

Our assessment of public sentiment turned out to be correct, because a subsequent REACH survey showed that, first of all, that 60% were aware of the cancellation. **Of those who were aware**, 86% of Christians agreed with the cancellation. That I think will be natural. But **64% of all who had heard about the cancellation**, Christian and non-Christian, also agreed with the cancellation. Twenty-eight percent thought that it should not have been cancelled.

Minister for Home Affairs K Shanmugam, 1 April 2019, emphasis mine

The quote above is taken directly from the Hansard, and is consistent with the results shown in REACH's press release. Note the phrases that I bolded for our purposes, which I will call "qualifiers".

Figure 4.1: Screenshot of online article on results from REACH poll. Retrieved May 21, 2019.

The next day, national newspaper The Straits Times ran a story headlined "Parliament: Two in three back move to ban Watain gig". Within the text of the article, it reads:

The Government decided to cancel the permit for Watain's concert last month when it received reports that mainstream Christians were very concerned and offended by the band, Home Affairs Minister K. Shanmugam said yesterday. **And a survey of Singaporeans by government feedback unit Reach found that two in three**

**supported the move, he noted.** Among Christians, 86 per cent were supportive of the move to disallow the concert, the Reach poll found.

Note the qualifier “among those who were aware” is neither in the headline (Figure 4.1) nor the body of the article<sup>6</sup>. *Why is this important?* Results from REACH show that 63% of respondents were aware, and *out of these respondents*, 64% supported the government’s ban. This means that out of *all* respondents to the survey, only about 40% reported supporting the ban. The qualifying phrase “among those who were aware” meaningfully changes the interpretation of the results - we shouldn’t be able to say that **the majority of Singaporeans** supported the ban when in fact only 40% of the survey respondents did so.

In effect, the Straits Times article is invoking a strong assumption here (see 4.3 for a more technical explanation) - that *if* those who were unaware were in fact able to express their support for the ban, the same proportion of respondents (among those who were aware, 64%) would also support the ban. But being aware of the ban is a *prerequisite* for support of the ban, which makes this assumption rather unreasonable. Even assuming this hypothetical scenario were possible, the actual figure could be higher or lower - it depends on how similar (or different) the unaware are to the aware. Those who were not aware may be less likely to care about black metal music (or simply too busy to keep up with current affairs) and simply base their support of the ban on their general sentiment toward government policies. This seemingly minor omission of the qualifier can lead to false conclusions pretty quickly. Let us look at another example.

## 4.2 Web-savvy Seniors?

Part of my research involves looking at how Internet use can improve the lives of older adults (see Ang and Chen, 2018). I was interested in what the overall situation was like in Singapore, and googled something like “internet use seniors”. One 2014 article in the Straits Times immediately caught my eye (see Figure 4.2).

Within the article, the reporter states:

**Also, 78 per cent of those aged 55 and older here access the Internet every day either via the traditional Web browser or smartphone apps, putting Singapore fifth in the world for having the most Internet-savvy seniors.**

I was skeptical. Over and above my anecdotal experience with Singapore older

---

<sup>6</sup>CNA ran a similar headline, but included the qualifier within the article. See <https://www.channelnewsasia.com/news/singapore/2-in-3-singaporeans-in-reach-poll-supported-government-s-11401066>



Figure 4.2: Screenshot of online article on web-savvy seniors. Retrieved May 21, 2019.

adults, past research in the United States<sup>7</sup> gave me reason to expect that the proportion of older people even using the Internet (everyday or not) should be much lower. Some results from Consumer Barometer are available online, so we can check for ourselves. Of interest in Figure 4.3 are numbers reflecting internet use in year 2014, which is when the news article was published. Note that the percent of Singaporeans aged 55 and above who use the internet daily in 2014 is **29%, not 78% as the article suggests**.

How then, did the reporter get things so wrong? While detailed statistics for 2014 doesn't seem available online anymore, a little investigation using 2017 figures shows how the reporter arrived at a number as high as 78%.

Looking at Figure 4.4, the crucial part is the footnote that says “base”, which tells us that the 82% figure for daily Internet usage in 2017 are **among those who use the internet**. We can easily calculate this 82% with the numbers in Figure 4.3 - note that  $\frac{43}{53}$ <sup>8</sup> is approximately 82% of 53<sup>9</sup>. That is,  $\frac{43}{53} \approx 0.82$ . This recovers the 82% figure that we see in Figure 4.4.

<sup>7</sup>For instance, see <https://www.pewinternet.org/2012/06/06/older-adults-and-internet-use/>

<sup>8</sup>The percent of Singapore residents aged 55+ using the internet in 2017, from Figure 4.3

<sup>9</sup>The percent of Singapore residents aged 55+ using the internet *daily* in 2017, from Figure 4.3



Figure 4.3: Screenshot of Consumer Barometer findings across time. Retrieved May 21, 2019.



Figure 4.4: Screenshot of Consumer Barometer results on 2017 internet use. Retrieved May 21, 2019.

Using the same strategy, we can recover the reporter's figure for 2014:  $\frac{29}{37} \approx 0.78$ .

*What does this mean?* This means that just like the reporter in the Watain example (4.1), this reporter left out an important qualifier - only 29% of all older adults in Singapore use the internet daily, but **78% of those who use the internet** use it daily. This vast discrepancy is highly consequential - the statement that "78 per cent of those aged 55 and older here access the Internet every day" is false, and the headline that "Majority of Singapore seniors are Web savvy" is misleading at best.

### 4.3 Technical excursus

Before we go into a more technical explanation of what went wrong in these two cases, let us first move from proportions to probabilities. The difference between a proportion and a probability is important here. Note that when Minister Shanmugam asserted the REACH poll provided evidence that the Government's "assessment of public sentiment turned out to be correct", he was not suggesting that 680 Singaporeans form the whole Singapore public. The underlying assumption was that since most survey respondents (who were aware) supported the ban, it is likely that most Singaporeans (who are aware) will also support the ban. That is, he was using the *proportion* of supportive survey respondents (a description of the sample), to infer the *probability* (a hypothetical quantity) of any one Singaporean supporting the ban.

The difference between a probability and a proportion may be simplified using a coin flip example. If I flip a fair coin 4 times, the proportion of heads may be 0, 0.25, 0.5, 0.75, or 1. However, since it is a fair coin, the probability of getting a heads is, by definition, 0.5. So the proportion may or may not equal the probability. What we know is that the more times I flip the coin, the more likely the proportion of heads will reflect the true probability of getting a heads. It is thus common to hear people say that the probability is the "long-run proportion of an event". Below is some code (in R) for you to try out the coin flip example.

```
# Set the number of trials to 4.
# You may change the number to see what happens.
n <- 4
# Get the proportion of heads after flipping a fair coin n times.
# Try this a few times.
sum(rbinom(n, 1, prob=0.5))/n
```

We have now established that the main reason why we are interested in proportions from a REACH poll is because they purport to tell us something about Singaporeans in general. That is, the REACH poll suggests that if we were to randomly pick a Singaporean from those who are aware of the ban, the probability of this person supporting the ban is about 0.64 (or 64%). The problem at hand then reduces to a trivial probability question, assuming that we all remem-

ber basic probability rules from secondary (primary?) school<sup>10</sup>. If the REACH poll is indeed representative of all Singaporeans, then we have the following quantities:

$$\begin{aligned}\Pr(\text{Aware of Ban}) &= 0.63 \\ \Pr(\text{Not Aware of Ban}) &= 1 - \Pr(\text{Aware of Ban}) = 0.37 \\ \Pr(\text{Support Ban} \mid \text{Aware of Ban}) &= 0.64\end{aligned}$$

$\Pr(\text{Support Ban} \mid \text{Aware of Ban})$  is a conditional probability, but the quantity that is being asserted in the news article is  $\Pr(\text{Support Ban})$ , which is the total probability. Using the law of total probability, we know that:

$$\begin{aligned}\Pr(\text{Support Ban}) &= \Pr(\text{Support Ban} \mid \text{Aware of Ban}) \cdot \Pr(\text{Aware of Ban}) \\ &\quad + \Pr(\text{Support Ban} \mid \text{Not Aware of Ban}) \cdot \Pr(\text{Not Aware of Ban})\end{aligned}$$

Plugging in the numbers that we have,

$$\Pr(\text{Support Ban}) = 0.64 \cdot 0.63 + \Pr(\text{Support Ban} \mid \text{Not Aware of Ban}) \cdot 0.37$$

we see that  $\Pr(\text{Support Ban}) = 0.64$  if and only if  $\Pr(\text{Support Ban} \mid \text{Not Aware of Ban})$  also equals 0.64. That said,  $\Pr(\text{Support Ban} \mid \text{Not Aware of Ban})$  is logically impossible, and should equal zero. Similarly, for the Web-savvy Seniors example,

$$\begin{aligned}\Pr(\text{Use Internet Daily}) &= \Pr(\text{Use Internet Daily} \mid \text{Use Internet}) \cdot \Pr(\text{Use Internet}) \\ &\quad + \Pr(\text{Use Internet Daily} \mid \text{Don't Use Internet}) \cdot \Pr(\text{Don't Use Internet}) \\ &= 0.78 \cdot 0.37 + \Pr(\text{Use Internet Daily} \mid \text{Don't Use Internet}) \cdot 0.63\end{aligned}$$

where  $\Pr(\text{Use Internet Daily} \mid \text{Don't Use Internet})$  is impossible and should be zero. In both cases, total probabilities are substantially different from the conditional probabilities, and there is no reason to believe they would be the same.

## 4.4 Conclusion

By now, it should be clear that qualifiers attached to proportions (and percentages) are critical. Without them, results from studies can be blown out of proportion. It is not wise to completely rely on assertions made by news articles (or other kinds of reports), even from supposedly credible agencies like

---

<sup>10</sup>Or that we can Google it if not

the Straits Times. As we have seen, social scientists should be comfortable with interpreting data from its source<sup>11</sup> in order to evaluate claims that are being made in public discourse today.

## **4.5 Additional reading**

Straits Times reporter Christopher Tan alleges (July 5, 2019) that the Land Transport Authority also makes this kind of misleading representation of support for legalizing onboard audio recordings in taxis and private hire cars: See <https://www.straitstimes.com/opinion/recording-in-taxis-for-hire-cars-consider-cost-consequences>. You may want to check the source material from LTA and REACH to judge for yourself if he is correct.

---

<sup>11</sup>This, however, first requires data to be made available for replication purposes.

# Chapter 5

## Are we lonely?

Contributor: Shannon Ang

Date: 25 May 2019

In population health research, there are a number of commonly used scales to measure psychosocial well-being. For instance, the Center for Epidemiologic Studies Depression Scale (CES-D) is widely used to measure depression, and the EQ-5D<sup>1</sup> to measure quality of life. These scales seldom have an intuitive interpretation - who knows what 10 points on the CES-D scale actually means in ‘real life’, versus 12 points? To address this, social scientists often choose a “cut-off” point to simplify the measure into two categories (e.g., either you are depressed, or you are not). Some of these cut-off points are well researched (such the cut-off point for mild cognitive impairment), while others are more arbitrary.

This case study looks at the prevalence of loneliness in Singapore older adults, and how these cut-offs can shape the way we think about it. The focus here is *not* to criticize researchers’ choices of cut-off points. Instead, this case study seeks to provide a way to evaluate claims that are based on these cut-offs, so that we understand how to compare claims across studies and/or reports.

### 5.1 The lonely dichotomy

Loneliness is a real issue for many people today, and it has been shown to have deleterious effects on health (Rico-Uribe et al., 2018). More of us are beginning to realize that social connections are key, especially for older persons. The Straits Times carried an article in 2018, with the headline “Senior citizens living with family, but still feeling lonely”.

---

<sup>1</sup>There’s no ‘full’ name for this, its just referred to as the EQ-5D.



Figure 5.1: Screenshot of online article on lonely seniors. Retrieved May 25, 2019.

But what does **feeling lonely** really mean? There seems to be a dichotomy being drawn here - either you feel lonely, or you don't feel lonely. There is little room for nuance like "I feel lonely when I ride the bus by myself". Of course, some level of simplification is needed to compare across groups - and this simplification is what we need to examine.

The news article highlights studies on loneliness (in Singapore) done by two different groups - a team at the National Healthcare Group (NHG), and a team at the Centre for Ageing Research and Education (CARE) in Duke-NUS Medical School<sup>2</sup>. As we will see, this lonely/not lonely dichotomy is drawn by researchers as well, but sometimes **in completely different ways**. Let us take a closer look.

## 5.2 Lonely by whose standard

The studies of interest<sup>3</sup> here are Wee et al. (2019), Ge et al. (2017), Lim and Chan (2017), and Chan et al. (2015). All of these studies use a variant of the 3-item UCLA<sup>4</sup> Loneliness Scale (Hughes et al., 2004). The 3-item UCLA

<sup>2</sup>Led by Associate Professor Angelique Chan

<sup>3</sup>All these studies are open-access articles, anyone can access them even without a library subscription.

<sup>4</sup>University of California, Los Angeles

Loneliness Scale consists of three questions, which all of these studies use:

1. How often do you feel you lack companionship?
2. How often do you feel left out?
3. How often do you feel isolated from others?

For each of the questions listed above, respondents were given a range of responses to choose from. These are listed in Figure 5.2. As you might notice, they were different across the studies. Wee et al. (2019) and Ge et al. (2017) had only 3 response categories (the right column), while Lim and Chan (2017) and Chan et al. (2015) had 5 different response categories (the left column). I put the point values for each response in parentheses. In all of these studies, researchers added up these points across the 3 questions, and came up with their own ‘cut-off’ point to determine who was “lonely”.



Figure 5.2: Summary of response categories

I added the **blue arrows** in Figure 5.2 to show how the 5 category option can be mapped to the 3 category option (but not vice versa). We don’t really know what kind of bias this will introduce<sup>5</sup>, but at face value I think this looks pretty reasonable. Why am I doing this? This “matching” allows us to do a little experiment with publicly available data<sup>6</sup> to answer the following question: **How does changing the criteria change our view of loneliness in Singapore?** The next section compares these different coding schemes.

<sup>5</sup>For instance, one might argue that respondents might choose differently when given more gradational categories.

<sup>6</sup>So you can try it yourself

### 5.3 Same data, different results

For this simple analysis, I use data from Wave 2 of the Panel on Health and Ageing of Singaporean Elderly (PHASE) (see 1.3), conducted in 2011. This is a nationally representative study of older adults aged 60 and above. It is essentially the same dataset used in Lim and Chan (2017) and Chan et al. (2015). Code provided is in R. For the sake of brevity, I leave out observations with any missing values on any of the loneliness items. A key concern of the news article in Figure 5.1 is that even older adults living with family members may be lonely, so we will look at a cross-tabulation of living arrangements with “loneliness”.

#### Coding scheme 1: Lots of loneliness

I first follow the coding scheme in Lim and Chan (2017) and Chan et al. (2015). I sum the items up (giving me a score that ranges from 0-12), and then dichotomize respondents into people who are “not lonely” (score of 0), and those who are “lonely” (score of 1-12)<sup>7</sup>. This cut-off point seems to have arisen from a “common-sense” approach rather than any kind of formal testing - that is, group the people who never experience loneliness in one group, and then put the rest who have had some experience of loneliness in another.

```
# Note: You first need to read in the data
# The data already contains a pre-coded version according to these criteria
lonely1_cat <- phase$w2_loneliness_yesno %>%
  factor(labels=c("Not lonely", "Lonely"))
# Make a table (proportions are weighted to account for survey design)
knitr::kable(
  GDAtools::prop.wtable(livingarr, lonely1_cat,
    dir=1, digits=3, w=phase$w2_weights, na=F, mar=F),
  caption = paste0('Crosstabulation using criteria in Lim and Chan (2017) ',
    'and Chan et al (2015). Note that these are row percentages.'),
  booktabs = TRUE)
```

Table 5.1 gives me a similar proportion as suggested in the news article - that is,

“[Associate Professor Chan’s study] found that half of Singaporeans over 60 felt lonely some or most of the time. But those who lived with spouses, or with spouses and children, did not.”

These numbers are indeed worrying. Older adults living alone are understandably lonely, but those who live with their children (but without their spouse)

---

<sup>7</sup>Note that Chan et al. (2015) further splits the “lonely” group into “sometimes lonely” and “mostly lonely”. Lim and Chan (2017), however, does not make this distinction. I have grouped them together since this is the way that it has usually been represented in public discourse (e.g., the claim that those who are lonely have a higher risk of mortality compared to those not lonely. See, for instance, <https://www.straitstimes.com/singapore/those-who-feel-lonely-more-prone>)

Table 5.1: Crosstabulation using criteria in Lim and Chan (2017) and Chan et al (2015). Note that these are row percentages.

	Not lonely	Lonely
Living alone	41.022	58.978
Living with spouse only	67.844	32.156
Living with child only	49.910	50.090
Living with spouse and child	71.680	28.320
Living with others only	44.791	55.209

are not that far behind (50.1%!). Even 28% of those who live with their spouse and child feel lonely, like the headline in the news article (Figure 5.1) suggests.

### Coding scheme 2: Not that much loneliness

We then arrive at the coding scheme used by Wee et al. (2019) and Ge et al. (2017). Summing the items gives me a score that ranges from 3-9, and I then dichotomize the group into people who are “not lonely” (score of 3-5), and those who are “lonely” (score of 6-9). Note that these cut-points are probably arbitrary - while the researchers cite a paper each to justify their use of the cut-point, the cited papers do not really provide evidence in support of the cut-point. The closest support for the cut-point in the cited papers that I could discern is in Steptoe et al. (2013), which states that they used the top quintile<sup>8</sup> to define loneliness. No reason was given as to why the top quintile was chosen. Table 5.2 shows the distribution of “loneliness” according to these criteria.

```
# Recode and sum the loneliness scores
lonely2 <- phase %>%
  select(w2_Q10_1_GV1, w2_Q10_2_GV1, w2_Q10_3_GV1) %>%
  mutate_all(funs(recode(., `0` = 1, `1` = 1, `2` = 2, `3` = 3, `4` = 3))) %>%
  rowSums()

# Categorize according to cut-off point
lonely2_cat <- if_else(lonely2 < 6, 0, 1) %>%
  factor(labels=c("Not lonely", "Lonely"))

# Show table
knitr::kable(
  GDAtools::prop.wtable(livingarr, lonely2_cat,
    dir=1, digits=3, w=phase$w2_weights, na=F, mar=F),
  caption = paste0('Crosstabulation using criteria in Wee et al (2019) ',
    'and Ge et al (2017). Note that these are row percentages.'),
  booktabs = TRUE)
```

<sup>8</sup>In their sample, not the Singapore one.

Table 5.2: Crosstabulation using criteria in Wee et al (2019) and Ge et al (2017). Note that these are row percentages.

	Not lonely	Lonely
Living alone	84.419	15.581
Living with spouse only	96.555	3.445
Living with child only	92.184	7.816
Living with spouse and child	97.863	2.137
Living with others only	93.046	6.954

Table 5.3: Comparison of absolute and relative differences

	Coding scheme 1	Coding scheme 2
(1) Living alone	59.0	15.6
(2) Living with child only	50.1	7.8
Difference [(1) - (2)]	8.9	7.8
Ratio [(1)/(2)]	1.2	2.0

What you will immediately realize is that these numbers are way lower than those when using coding scheme 1 (that is, the coding scheme of Lim and Chan (2017) and Chan et al. (2015)). These numbers are more consistent with the figures shown in Ge et al. (2017)<sup>9</sup>. Further, the difference in the proportion of those living alone and those living with their child (but without their spouse) is similar in absolute terms, but much larger in relative terms (see Table 5.3). While the proportion of those lonely among those who live alone is 1.2 times that of those living with only their children according to coding scheme 1, this ratio increases to 2 when using coding scheme 2. Based on these results, it seems that the overall loneliness situation is much less dire than before, and those who live with children don't seem as isolated as those who live alone after all.

## 5.4 Conclusion

You may be thinking, so which way of conceptualizing loneliness is “correct”? As I mentioned at the start of this case study, that is not the goal here. The process of figuring out a useful cut-off point is a long and tedious one that requires researchers to engage with each other<sup>10</sup>. Rather, the goal here has been to highlight that decisions like cut-off points may seem small, but are

<sup>9</sup>Note that the sample in Ge et al. (2017) is of all adults aged 21 and older, not just older adults, so the higher number seen here is expected. Note also that in the paper, the authors show column percentages instead of row percentages. Since we are comparing across living arrangements however, row percentages are more appropriate.

<sup>10</sup>And indeed, Singapore social science can perhaps use more of this.

critical in the way we interpret and talk about social phenomena. In this case study, these cut-off points essentially define who is considered lonely. **If these cut-off points are vastly different, we may not be even talking about the same people** when we talk about “lonely persons”. This means that when comparing or drawing conclusions from the findings of different studies, it is crucial that we understand the decisions that produced the numbers.

# Chapter 6

## Size is not all that matters

Contributor: Shannon Ang  
Date: 22 July 2019

In response to public surveys, it's quite common to hear people make comments like "they asked only 1000 people, how accurate can it be?". Of course, people tend to only say these things when the findings presented are incongruent with their perception of reality.

This case study looks at whether and how size makes a difference when we're trying to understand survey results from samples of the population. My hope is for you (the reader) to realize that often, *how* the sample is drawn (and how we quantify uncertainty) is more important than the size of the sample.

### 6.1 But it's too small!

We will first look at an example. Local newspaper The Straits Times recently reported the findings of a study done by Michael Page, alleging that **8 of out 10 people are “happy” with their work-life balance**.

Many people on social media spoke out against the findings (not unexpectedly), questioning how reliable these findings were. These commentators probably felt such findings could not possibly reflect reality, given their own experiences. Surely, they might believe, more people are unhappy with their work-life balance. To rationalize this incongruence, there will inadvertently be some attention drawn to the sample size.

The comment above (Figure 6.2) shows how people often rationalize their feelings toward studies that don't reflect their own perception<sup>1</sup>. How can 1328

---

<sup>1</sup>For studies that do agree with the layperson's perception, one probable response would be that it is "common sense" and that there is "no need to do a study to know this".



Figure 6.1: Screenshot of online article on work-life balance. Retrieved July 16, 2019.



Figure 6.2: Screenshot of comment on work-life balance article. Retrieved July 16, 2019.

survey respondents represent an entire population of approximately 5.9 million individuals?

The frustrating thing about the report by Michael Page is that it is **not transparent about its methodology**. Without this information, we cannot know for sure about how generalizable these findings are to the whole population. They likely cannot say much about *all Singaporeans* - the results may be limited to Singaporeans who took the survey, who may be a very different group of Singaporeans compared to those who did not take the survey.

This is not an isolated case - it often is difficult to find out exactly how surveys like these (that is, surveys aimed at generating public interest and reported by the media) were conducted. As we will see, the key questions are these: **How were respondents recruited, and how is uncertainty being estimated?**

## 6.2 Size matters, but how so?

It is true that size matters. However, people tend to paint some variant of this false dichotomy - results are accurate if the sample size is arbitrarily large (e.g., ~50k individuals, or some percentage of the population size), and completely inaccurate otherwise. But more data doesn't necessarily mean appropriate data. For instance, having data on all Singaporean Chinese doesn't really help me understand Singaporean Malays better. The dichotomy is also terribly unhelpful, because what we want is to be able to draw big conclusions from small data<sup>2</sup>. In fact, we do this all the time - we conduct interviews to hire people, we date others before entering into a serious relationship, we administer written tests to measure area-specific knowledge, we audition actors before casting them in a role, and more. While understanding that these processes are not perfect, we do expect them to provide us with sufficient information on which to base our bigger decisions.

**But how certain can we be?** This question is key. First, we need the following information: what are the sampling probabilities for each respondent in the sample? In other words, how likely is each member to be selected into the sample? If they were selected at random<sup>3</sup>, there are established methods to calculate uncertainty around our findings<sup>4</sup>.

---

<sup>2</sup>The obsession with big data is sometimes misguided, especially if this suggests ‘small data’ is useless or inferior. Big data is mostly useful for other reasons rather than its ‘bigness’ (i.e., the sheer number of observations). See Matthew Salganik’s subsection on this in his book *Bit by Bit* for a good treatment of this subject: <https://www.bitbybitbook.com/en/1st-ed/observing-behavior/characteristics/big/>

<sup>3</sup>Random here means that every person in the population has an equal (or known) chance of being selected. It does not mean I randomly approach someone on the street, since certain types of people are more likely to be selected with this method (e.g., people who live close to that street, people who are around at that time, etc.). Usually, in Singapore, researchers who want a random sample get a list of household addresses from the Department of Statistics. It is assumed that the Department of Statistics provides a random sample of all household addresses, but I am not privy to what happens exactly.

<sup>4</sup>Via the Central Limit Theorem. If the quantity of interest is a proportion/probability,

Let's do a small simulation exercise here to demonstrate how these calculations work, without having to read the math<sup>5</sup>. Code here is written in R. A simulation means that we can determine what is the 'truth' in a hypothetical population, and then use the computer to draw samples from this population<sup>6</sup> to see how sampling (and sample sizes) work. Our quantity of interest is the proportion of people who like durian. We define the population such that 65% of people (in our population) like durian (i.e.,  $\Pr(\text{Likes Durian}) = 0.65$ )<sup>7</sup>.

```
# Determine the sample size
n <- 100

# Draw samples from a population, where Pr(Likes Durian) = 0.65
sample <- rbinom(n, 1, p=0.65)

# Calculate the proportion of people who report liking durian
sum(sample)/n
```

```
## [1] 0.61
```

The number produced above is 0.61, but if you run the code above multiple times, you will realize the number produced is different each time you run it (i.e., each time you draw a new sample). This makes sense, since samples are drawn randomly. Since we know the true value of  $\Pr(\text{Likes Durian})$ , we also know how far off these sample estimates of  $\Pr(\text{Likes Durian})$  tend to be. The next step in this exercise is to draw a large number of samples of the same size to show how much these sample proportions can vary.

```
# Determine the sample size
n <- 100

# Draw 10000 samples of the same size
sample_no <- 1:10000
samples <- lapply(sample_no, function(x) rbinom(n, 1, p=0.65))
proportions <- lapply(samples, function(x) sum(x)/n)

# Visualize the proportions with a histogram
hist(unlist(proportions),
```

---

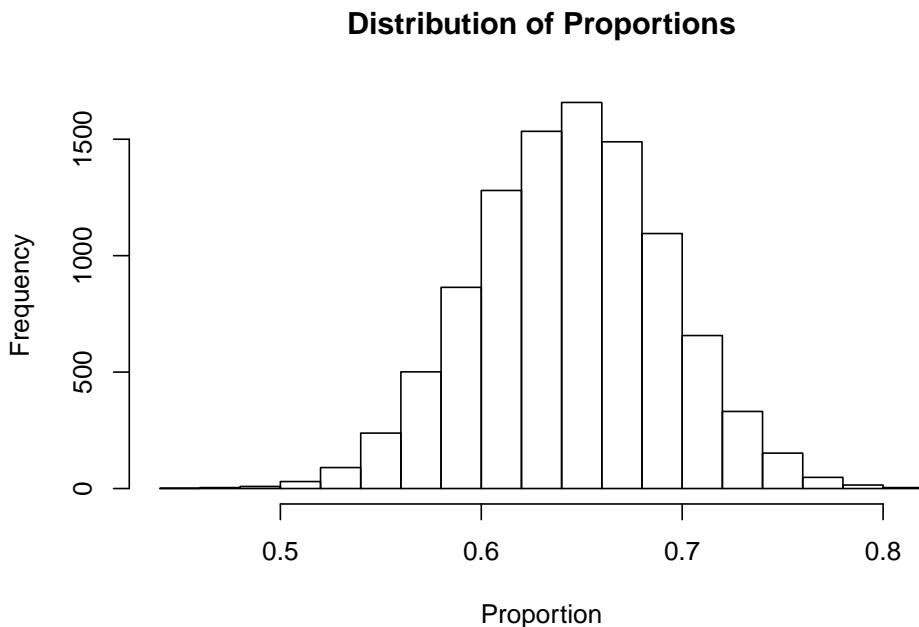
then the normal approximation property (and its assumptions) also needs to be invoked. The point here, however, is not to demonstrate the Central Limit Theorem, but to demonstrate how sample size matters. I will therefore avoid getting into the technicalities of the Central Limit Theorem.

<sup>5</sup>If you already understand what  $\lim_{n \rightarrow \infty} \Pr \left[ \frac{\sqrt{n}(S_n - \mu)}{\sigma} \leq x \right] = \Phi(x)$  means, you may skip ahead to the next section

<sup>6</sup>The code here utilizes the Monte Carlo method, which essentially is a computational way to see how samples perform under realistic data conditions with a known probability distribution.

<sup>7</sup>Look at Chapter 4.3 for a short explanation of how we went from proportions to probabilities

```
main="Distribution of Proportions",
xlab="Proportion")
```



You can see that the distribution of proportions from these samples looks (somewhat) like a bell curve<sup>8</sup>. Let's now summarize this distribution with some numbers.

```
# Get the mean of all proportions,
# rounded to 2 decimal places
round(mean(unlist(proportions)), 2)

## [1] 0.65

# Get the 2.5th and 97.5th percentiles
quantile(unlist(proportions), prob=c(0.025, 0.975))

## 2.5% 97.5%
## 0.56 0.74
```

**What do these numbers tell us?** If you realize, the mean of the distribution of sample proportions is, in fact, the true value of  $\Pr(\text{Likes Durian})$ . This demonstrates that even with relatively small samples, it is possible to recover the true population value through repeated sampling. The only difficulty is that we can seldom get so many samples, so the quantile values are of more help. The quantile values tell us that 95% of all the sample proportions will fall between

---

<sup>8</sup>Or, in statistics speak, it is approximately normally distributed. This will remind you of the Central Limit Theorem, if you're familiar.

0.57 and 0.74. This gives us some measure of uncertainty. In other words - if I take a single sample of size 100, 95% of the time the proportion of those who like durian calculated from this sample will fall between 0.57 and 0.74 (given that the true value is 0.65)<sup>9</sup>. In other words, for this particular scenario (i.e., 100 observations, true proportion = 0.65), sample estimates will differ from the true value by at most 9 percentage points 95% of the time. This is the fundamental basis of claims by researchers that a survey has a certain “margin of error”.

**So what happens when we increase the sample size?** Let us run the same code again, but increase the sample size to 1000.

```
# Determine the sample size
n <- 1000

# Draw 10000 samples of the same size
sample_no <- 1:10000
samples <- lapply(sample_no, function(x) rbinom(n,1,p=0.65))
proportions <- lapply(samples, function(x) sum(x)/n)

# Get the 2.5th and 97.5th percentiles
quantile(unlist(proportions), prob=c(0.025, 0.975))

## 2.5% 97.5%
## 0.620 0.679
```

You will realize here that the interval for the quantile values have narrowed<sup>10</sup>. This shows that with a bigger sample size, our uncertainty around the true estimate also shrinks.

Our little simulation exercise has shown a number of things:

1. The basic intuition is correct - larger sample sizes in fact produce estimates that are less susceptible to sampling error (that is, estimates generally fall closer to the true value).
2. However, this basic intuition matters less than we often think. Even a relatively small sample size of 1000 (as asserted by the commentator in Chapter 6.1) produces estimates quite close to the true value (in this case, within 3 percentage points 95% of the time).

What is important here, however, is that these estimates can only be calculated in this manner if we have a probability sample - that is, if we know the sampling probabilities for each respondent in the sample<sup>11</sup>. **If we do not know how respondents in a sample were recruited, there is no way to determine**

---

<sup>9</sup>Other than the sample size, uncertainty also increases if the true value of Pr(Likes Durian) is closer to 0.50.

<sup>10</sup>If you calculate the mean of proportions again, you will also realize also that it recovers the true value of 0.65

<sup>11</sup>The easiest way is to get a simple random sample (where everyone has an equal chance of selection), but there are also ways to adjust for more complex cases (e.g., multistage sampling)

**whether purported findings actually reflect the population, and how much uncertainty there is around them.**

This is not to say that there is no way to calculate a “margin of error” or bounds of uncertainty for non-probability samples. Researchers have been developing new and innovative ways to recover reliable estimates even with non-probability samples, as “big data” comes to the forefront. Unfortunately, people are often careless and apply methods meant for probability samples to non-probability samples, expecting them to work in the same way. We will explore this in the next section.

### 6.3 Non-probability sample, now what

Probability samples are hard to come by, and while in theory they are the “gold standard”<sup>12</sup>, there are a number of drawbacks when trying to implement them:

1. Poor response rates. Not all respondents in your probability sample will agree to respond to your survey or participate in your study. While researchers have tried various methods (e.g., offering a monetary “token of appreciation”) to increase response rates, low response rates threaten the validity of a probability sample, especially if those who refuse are systematically different from those who eventually participate.
2. High cost. Partly because of strategies to increase response rates through “tokens of appreciation”, the cost of implementing a high-quality survey (especially through traditional face-to-face interviews) with a probability sample have increased substantially.

Some researchers, therefore, have turned to online surveys (which are much cheaper to conduct). One recent study commissioned by the Lee Kuan Yew School of Public Policy on Singaporeans’ perception of class markers did this.

The newspaper article reports the following:

While Dr Dodgson acknowledged that the sample size was relatively low, with a 4.25 per cent margin of error, she told TODAY that the open-ended nature of the survey “gave greater nuance and accuracy” than multiple-choice questionnaires.

The study report also contains the following ‘infographic’, which confirms what the news article reported.

As you may recognize, this “margin of error” is something we had talked about in the previous section. I simulated the results to show the basic reasoning behind sampling theory, but you can in fact easily find online calculators purporting to calculate this “margin of error” for you<sup>13</sup>. Let’s use a random one I found.

---

<sup>12</sup>Apart from having information on the entire population, through a census or administrative government data.

<sup>13</sup>These calculators use closed-form formulas to produce a ‘margin of error’, rather than through the more numerical/computational approach I used.

**TODAY Singapore** MENU ▾ 

## How to tell if someone is high or low class? Poll gets interesting replies from Singaporeans

By NG JUN SEN



Illustration: Kenneth Choy

A study commissioned by the Lee Kuan Yew School of Public Policy found that around 66 per cent of those who took part in a poll believed that what determines class identity is income.

Published 19 JULY, 2019  
UPDATED 19 JULY, 2019

Figure 6.3: Screenshot of TODAYonline article on class perceptions. Retrieved July 16, 2019.



A total of 538 responses were collected, giving a margin of error of around 4.25% with a confidence level of 95%.

Figure 6.4: Screenshot of TODAYonline article on class perceptions. Retrieved July 16, 2019.

Just change the margin of error to 4.25 and the population size to ~3.5 million (Singapore resident population size) and you get back a recommended sample size of ~530<sup>14</sup>. But remember, **these calculations are for a probability sample!**<sup>15</sup>

Is the study in question conducted with a probability sample? The details provided are rather vague<sup>16</sup>, but my general sense is that they were not using a probability sample. The report states that:

Responses were solicited via social media and via a panel responses service. The goal with this was to attract a mixture of opinions, both from people with a pre-existing interest in the topic (whose opinions are generally weighted more heavily in political discussion, simply because they are more likely to assert them and eventually to take action on the same basis) and from those with no immediate interest (whose opinions are often discounted in political discussion, but which can have startling effects at the ballot box). Similarly, the demographics of the respondents were tracked with the aim of creating a sample that would be broadly representative of the general population.

In what seems like implicit acknowledgement that the study did not use a probability sample, some effort has been made to ensure adequate ‘representation’ through varied modes (social media and panel responses service), and through mimicking the country’s demographic composition (probably by ensuring sufficient numbers of minorities)<sup>17</sup>. It is worth noting, however, that in no way does this make it equivalent to a probability sample of *all Singaporeans*<sup>18</sup>. **It is therefore inappropriate to calculate and use a margin of error as if it were a probability sample of all Singaporeans.**

This does not mean that there is no way to use a non-probability sample to produce accurate estimates. But methods to derive estimates from non-probability samples are often context dependent (i.e., they have to be customized for each case, depending on the quantity of interest) and have different requirements beyond sample size. It requires more effort from researchers, but this is not necessarily a bad thing. Because of the practical drawbacks of probability sampling (highlighted above), statistical innovation to draw inferences from non-probability samples (often done through post-sampling adjustments) is much needed in the world of research.

‘Big data’ plays an important role here. (Matthew Salganik gives one of the best overviews on the characteristics of ‘big data’ in his book *Bit by Bit*, which

<sup>14</sup>The report itself does not state how they calculated the margin of error. It is my guess that this (or a variant of this calculation) is what they did. If you are aware that a different method was used, please write to me so I can correct it.

<sup>15</sup>More specifically, a simple random sample.

<sup>16</sup>For instance, how did the panel responses service recruit participants?

<sup>17</sup>This appears similar to quota sampling, which is a non-probability sampling method.

<sup>18</sup>If only it were that simple.

I highly recommend). Data that is collected constantly and from many people at once allows for timely and cost-effective results, if we use the right approach to analyze them. Wang et al. (2015) (available here) use multilevel regression and poststratification (affectionately called “Mister P”) to show that accurate forecasts of the 2012 US presidential election could be obtained using data from Xbox<sup>19</sup> users. Two points are of note here. First, the sample was large - 750,148 interviews were conducted through Xbox polls, with 345,858 unique respondents, and over 30,000 respondents completed five or more polls. Second, having strong covariates<sup>20</sup> to adjust for non-response bias (Gelman et al. (2016), available here) and subgroup level characteristics (if more granular subgroup level estimates are of interest) are crucial to good estimation<sup>21</sup>. **Neither of these was true in the study by the Lee Kuan Yew School of Public Policy discussed above.** Of course, this does not mean *all* of the claims in that study are invalid. There are probably still insights worth learning from. However, it does tell us that we need to constantly evaluate the claims of studies against their methodology - understanding the *how* is often key.

## 6.4 Conclusion

We have discussed sampling - what probability samples and non-probability samples are. It is true that sample size matters, but we can quantify the uncertainty in smaller samples by calculating bounds/intervals around our estimates. This is quite simple to do if we have a probability sample, and in such cases, estimates are quite accurate even with a relatively small sample size (e.g., 1000 observations).

In many studies reported in the media, however, non-probability samples are used. Non-probability samples can be a good alternative to probability samples especially given that non-response rates are increasing. **However, we cannot use formulas meant for probability samples and apply them to non-probability samples as if they are the same.** Non-probability samples require different (and slightly more complex) methods to produce equally accurate estimates.

What does this case study highlight? First, we need to be transparent<sup>22</sup> about how studies recruit members into their sample. **How the sample is recruited is usually much more important than how large the sample is**, but we are seldom given enough detail about it. Second, we need to be aware if appropriate methods are being used. Using a non-probability sample can seem more practical, but we become overconfident and may end up with the wrong conclusions if we simply act as if it is a probability sample.

---

<sup>19</sup>A gaming platform

<sup>20</sup>Other characteristics of the sample, beyond basic demographic information.

<sup>21</sup>See also <https://statmodeling.stat.columbia.edu/2013/10/09/mister-p-whats-its-secret-sauce/>

<sup>22</sup>And we need to demand transparency.

# Chapter 7

## Measuring societal change

Contributor: Shannon Ang  
Date: 17 September 2019

We hear various statistics that imply social change all the time. For instance, we hear that suicides are rising in Singapore, or that Singaporeans are living longer, or that the number of babies born is declining.

As always, people have many comments and theories about reasons behind what is happening. But before we can evaluate these comments, we need to first understand - what do these numbers really mean, and what kind of change is it really capturing? This case study looks at several ways we talk about societal change in public, and provides a better understanding of these commonly used numbers for higher quality discourse.

### 7.1 Three ways to understand change

Before we evaluate real examples, it will be helpful to understand different kinds of societal change. Let us work with a toy example. Suppose the newspaper (your favourite one) headline tomorrow states: “Average walking speed of Singaporeans slower than 10 years ago”. There are three possible explanations for this (maybe more, this is not a perfect analogy).

1. Older people walk slower, and Singapore is an ageing society (average age is rising), so its average walking speed will fall as there are more older people in society.
2. Those born later (e.g., Millennials, Gen Z) are walking slower than those born earlier (e.g., Baby Boomers), bringing down our average walking speed.
3. There was an alien invasion this year and many aliens suddenly came to

live among us. This suddenly caused overpopulation, since our system is only prepared for 6.9 million. People in Singapore thus have to lower their walking speed this year because it has become very crowded compared to last year.

The last one is slightly facetious, but as you can see, the three explanations point to three **quite different** reasons for change. We will look at each one, in order.

1. Change with **age**: Age effects refer to changes over the life course that apply to all individuals as they grow older. This is best demonstrated in terms of physical and biological changes that occur as we age (e.g., our metabolism slows down, we get wrinkles etc.) - such changes tend to be common to all of us, regardless of when we were born, or what era we live in.
2. Change with **cohort**: Cohort effects are probably easiest to think about as *generational* change (i.e., change with *birth cohorts*), even though more generally they are used to refer to a group of individuals that have gone through shared experiences (e.g., school cohorts). When you hear remarks from people such as “*young people nowadays* are [such strawberries / so spoilt / have no manners]”, this is usually the kind of change they are suggesting has happened.
3. Change with **period**: Period effects are changes due to contextual factors that affect all living persons at a particular point in time. For instance, significant events such as wars, epidemics, or natural disasters affect everyone who is exposed to it regardless of age.

Social scientists have long recognized the need to think clearly about these three distinct ways of understanding social change<sup>1</sup>. Popular discourse, however, tends to ignore the meaningful differences between these concepts - a common error is making conclusions about cohort change based off period data. We will examine two instances of this.

## 7.2 Declines in marriage and divorce

Marriage has been a key preoccupation of the Singapore government for many reasons (e.g., increasing fertility rates, shaping housing policy). Trends in marriage/divorce/fertility are thus often closely watched. The Straits Times recently published this article on marriage/divorce rates in 2018.

The article is based on a report on Marriages and Divorces, produced by the Department of Statistics<sup>2</sup>. Among other things, the article states:

---

<sup>1</sup>Although there is an equally long-running debate about how to estimate these effects, given linear dependence between the terms, i.e.  $Cohort = Period - Age$ . See Fosse and Winship (2019) for a great discussion.

<sup>2</sup>We will focus on discussing non-Muslim marriages here, but the same principles can be



Figure 7.1: Screenshot of online article on marriage rates. Retrieved September 17, 2019.

A total of 27,007 marriages were registered last year - the lowest in five years and 4.3 per cent fewer than the 28,212 marriages registered in 2017.

... Meanwhile, 7,344 marriages ended in a divorce or annulment last year, a 3.1 per cent drop from the 7,578 marital dissolutions in 2017.

... Given the volatile economy, observers said financial considerations could have contributed to the falling number of marriages and divorces last year.

Now, given that we've talked about age, cohort, and period changes, one question that should come to mind is: *what kind of measure is this?* Well, this is a **period measure** - it records the number of marriages in a certain year. Unnamed 'observers' cited by the article seem to understand this. They point out that the economy last year may have had a part to play, which is a plausible reason for period changes that suits the nature of the measure<sup>3</sup>. However, often times, people provide *cohort* explanations for changes in *period* measures.

The next two subsections will look at the question - **How reasonable is it to adopt cohort explanations for changes in period measures of marriage?**

applied to Muslim marriages as well.

<sup>3</sup>The general marriage rate, which is adjusted for population size of unmarried residents aged 15-49, also seems to tell this story.

### 7.2.1 Evaluating explanation 1

We first take a look at the following paragraph on changes in median age at first marriage:

The median age at first marriage for grooms rose from 29.8 years in 2008 to 30.2 years last year, while for brides, it went up from 27.3 years in 2008 to 28.5 years last year.

This is because more people spend a longer time getting an education and building up their careers before they settle down, said National University of Singapore sociologist Tan Ern Ser.

Note here that median age at marriage is, once again, a **period measure** - of all people who got married that year, half of them got married before age 28.5.

NUS sociologist Tan Ern Ser says that the median age at first marriage is rising because “people” spend more time in education and early-career building before they “settle down” (i.e., get married). But who are these “people”, exactly? Since most members of older cohorts do not usually go back into education and early-career building, there seems to be an implicit reference to younger cohorts (or, in layman terms, “*people nowadays*”)<sup>4</sup>.

**How do we evaluate whether this explanation is reasonable?** I advocate taking **two steps**:

1. Figure out if trends observed reflect cohort change, rather than age or period change.
2. Figure out if the explanation given for cohort change is supported by evidence.

Frustratingly, marriage statistics by cohort are not available from the government, so we cannot directly assess whether trends are really due to cohort change. The reason why period measures are used is often a matter of practicality (as we will discuss in the next section), but in other countries, academics use longitudinal panel survey data to estimate and project what might happen in younger cohorts. In Singapore, such survey data is scant and/or tend not to be shared with researchers outside of the government. This makes cohort change especially difficult to study. **But we will work with what available data we can find.**

#### Step 1

For a start, let us try to do Step 1 and make sure that changes in period measures are *not* due to the changing population age structure (i.e. age effects). Age-specific marriage rates are adjusted for the size of age groups and give us a picture of period change that is not driven by changes in age composition. They

---

<sup>4</sup>Of course, there is now a great drive towards lifelong learning and mid-career switches, but in the context I think it is reasonable to assume this is not the group of people being referred to.

are thus very helpful here, and have been provided in the report on Marriages and Divorces by the Department of Statistics, where we find Figure 7.2.



Figure 7.2: Screenshot from Department of Statistics report titled "Statistics on Marriages and Divorces, 2018". Retrieved September 17, 2019.

You can see that the mountain-like shape in Figure 7.2 is moving to the right, telling us that in 2018, the marriage rate is lower among younger age groups (15-29) and higher among older age group (30+), compared to 2008. **The figure thus suggests that people are getting married at later ages, and this is not just because our population as a whole is getting older.** The combination of age and period data suggest that the presence of cohort change remains plausible (Step 1).

### Step 2

Step 2 is to verify the cohort explanation itself. NUS sociologist Tan Ern Ser says that younger cohorts are delaying marriage in view of education and career building, so we should check that younger cohorts are indeed spending longer in education/early-career. The Ministry of Education happens to provide us data on Primary 1 cohorts (i.e., people who attended Primary 1 in the same year are in the same cohort), which tend to approximate birth cohorts closely (Figure 7.3).

The chart above shows that younger cohorts are indeed spending longer in education, since a larger percentage of them end up in post-secondary courses overall (see the rising trend across years). It does seem like younger cohorts are spending more time in education, so the cohort explanation therefore seems reasonable *up to this point*.

**However**, note that the report by the Department of Statistics states the following on page 7:

Between 2008 and 2018, the median age at first marriage among brides with secondary and below qualification rose considerably, narrowing the gap with graduate brides. For grooms, the differences in

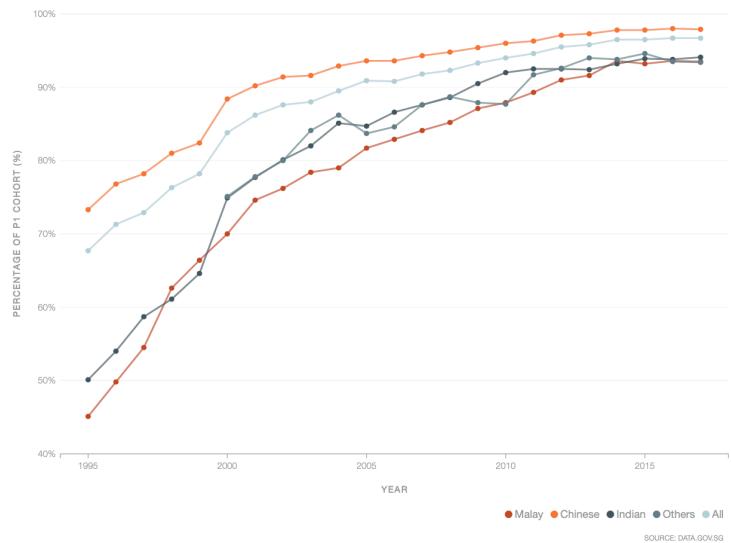


Figure 7.3: Screenshot from Data.gov.sg on "Percentage of P1 Cohort that Progressed to Post-Secondary Education". Retrieved September 17, 2019.

median age at first marriage across the educational groups also diminished over the same period.

This considerably undermines NUS sociologist Tan's explanation that younger cohorts are delaying marriage in view of education and early career building, since it seems that even though younger cohorts *are* getting more education, median age of marriage is rising **even among those without** post-secondary education (and at a faster rate than those more highly educated). We have not considered the "early career building" part yet, but it shows that this explanation requires more work to validate.

### 7.2.2 Evaluating explanation 2

Let us now look at another comment that was made about marriage rates. The news article states:

...Singapore Management University professor of sociology (practice) Paulin Straughan said they also seem to be spending a longer time looking for the right partner. She said: "People believe that marriage is forever and unless they are very sure they have found a life partner, they wouldn't marry."

SMU Professor (Practice) Paulin Straughan surmises that median age at first marriage is rising because "people" wait to be very sure of their partner. While there may be some kernel of truth here (although no data is provided to support

her point), it is quite unclear what she really means by this. Is she saying that everyone is delaying marriage because Singaporean culture has changed over time and people across all generations are now more careful about who to marry compared to the past<sup>5</sup> (i.e., a period effect)? Or is she saying that younger cohorts are now more careful about who to marry (i.e. a cohort effect)?

Regardless, a natural follow-up question would be - does this also explain the reason why divorce rates are falling (as the news article headline states)? If the ‘waiting to be very sure’ explanation holds weight, it should follow that since people have become more careful to marry, marriages are probably less likely to end up in divorce<sup>6</sup>! People “waiting longer” also would then explain why divorce numbers are falling.

*Suppose we adopt the following explanation: Divorce numbers are falling because younger cohorts are more careful (and therefore taking longer) about choosing their partners<sup>7</sup>.*

Remember that the divorce numbers mentioned in the news article are a **period** measure<sup>8</sup>. Just like median age of marriage, the median age at divorce is also rising, according to the report by the Department of Statistics, which states on page 15:

With an ageing population and rising divorce rates for older persons, the proportion of divorcees aged 45 years & over rose from 32.7 per cent in 2008 to 44.0 per cent in 2018 for males, and from 21.7 per cent to 29.9 per cent for females. As a result, the median age at divorce rose over the last decade.”

*Also remember the 2 steps outlined in the previous sub-section:*

1. Figure out if trends observed reflect cohort change, rather than age or period change.
2. Figure out if the explanation given for cohort change is supported by evidence.

### Step 1

The quote above already hints at trends being driven by age composition (“ageing population”)<sup>9</sup>. This already tells us the government believes the period

<sup>5</sup>Think about what this really means though. How possible is it to observe this?

<sup>6</sup>It could also be that *even though* people are more careful, they still make choices that end up in divorce, but this is somewhat an awkward assumption to make given a plain reading of the comment.

<sup>7</sup>This is not what SMU Professor (Practice) Paulin Straughan explicitly said, but something that we have inferred for heuristic purposes, after making several assumptions about carefulness in partner choice and stability of marriages.

<sup>8</sup>You may also ask, since we’re dealing with absolute numbers, isn’t it obvious that if there are less marriages, then there would be less divorces too? This is why absolute numbers are seldom helpful when examining societal change. At a minimum, we should attempt to look for rates that are adjusted for the size of the group at risk.

<sup>9</sup>For those more technically oriented, Martin (2009) and Luo and Hodges (2020) have

trend is unlikely to be driven by cohort change. However, we still can look at age-specific rates to replicate what we did earlier with marriage figures (Figure 7.4).



Figure 7.4: Screenshot from Department of Statistics report titled "Statistics on Marriages and Divorces, 2018". Retrieved September 17, 2019.

Once again we see from Figure 7.4 that in 2018, divorce rates are lower among younger age groups (20-34) and higher in older age groups (35+), compared to 2008. This is the same situation as before (with marriage numbers), and there we temporarily concluded cohort change is possible, though not certain. So why does the Department of Statistics not connect period trends with cohort change?

Data on divorce trends by birth cohort are (again) not available, but all is not lost. Weirdly enough, there exists some information on divorce by "marriage cohorts" (where each cohort is defined by the year they got married - i.e., if you got married in the same year as I did, you are in my marriage cohort) in a report by the Ministry of Social and Family Development (found here). What kind of change marriage cohort trends really capture is quite hard to say, but for heuristic purposes we may assume they (very) roughly approximate<sup>10</sup> birth cohort trends. Figure 7.5 shows us the proportion of divorces (y-axis) that happen before the couples'  $x^{th}$  anniversary (shown on the x-axis), by marriage cohort (indicated by the color and shape of the lines).

What can be observed from Figure 7.5 is that **younger (marriage) cohorts seem to be getting divorced earlier**<sup>11</sup>. For instance, you may notice that a

intriguing discussions about age\*period interactions as cohort effects.

<sup>10</sup>Think about who gets married in a certain year. What kind of people are they composed of? What characteristics do they share? It is hard to say. Other than sharing the same year of marriage, diversity within marriage cohorts is likely to be huge. We can only assume that younger birth cohorts will dominate most of the marriages in later years, but it remains to be seen whether this is accurate.

<sup>11</sup>It remains to be seen if lifetime divorce rates will be higher, since we cannot yet observe the full trajectory of union dissolutions for the younger cohorts (or even for older cohorts, since not all of them have dissolved their union or passed away).



Figure 7.5: Screenshot from Ministry of Social and Family Developopment report titled "Dissolutions of Marriages Among Marriage Cohorts, 1987-2015". Retrieved September 17, 2019.

higher proportion of those married in 2011 (light pink line) dissolve their unions before their 5th anniversary compared to those married in 2001 (dark pink line) or earlier<sup>12</sup>.

The picture painted here seems more bleak<sup>13</sup> compared to what we expected from period trends (and the news article). While period trends seemed to suggest that people were getting more careful with partner selection (since divorce rates were falling overall and median age of marriage and divorce has increased), **cohort trends seem to be telling the opposite story** - younger cohorts seem to be getting divorced earlier into their marriages!

### Step 2

The cohort explanation we adopted for this section was that “Divorce numbers are falling because younger cohorts are more careful about choosing their partners”. We have already found, however, younger cohorts (albeit marriage cohorts, not birth cohorts) are in fact getting divorced earlier. This directly contradicts the cohort-based explanation.

What is key here is to note that **while period measures and cohort mea-**

<sup>12</sup>Note that this does not really line up with the 2018 and 2008 measures we were comparing before, but there really is no easy way to do this.

<sup>13</sup>Assuming divorce is always a bad thing, which may not always be the case. A rise in divorce rates may possibly reflect changes in power dynamics within couples, with women becoming more independent and more willing to leave abusive situations.

sures can sometimes tell similar stories about societal change, there are many times when they do not. When we say “marriages and divorces are falling”, we need to be clear what the data allows us to say, and what kinds of explanations we give. If cohort trends are not the same as period trends, than providing cohort explanations for period change becomes meaningless.

We must therefore be very careful how we make claims about societal change, and what data we use to back it up. In the next section, we will look at life expectancy - a highly misunderstood measure in popular discourse.

### 7.3 Living longer than you expect

We've heard this so many times - “life expectancy is increasing”. “Singapore has one of the highest life expectancies in the world.” **But what does the term “life expectancy” mean?**



Figure 7.6: Screenshot of online article on life expectancy. Retrieved September 17, 2019.

A typical article (such as this one written by a physician) will use estimates of life expectancy to say, for instance, “Singaporean residents can expect to live up to 83.2 years<sup>14</sup>”. This reading of life expectancy is quite common.

<sup>14</sup>The figure for the year 2017.

The obvious question here is - Who can expect to live up to 83.2 years? My grandma? Me? My child? Do all Singaporeans now expect to live to the same age<sup>15</sup>?

The figure most often cited in the media is life expectancy *at birth*. A more precise way to interpret the life expectancy in 2017 is “Singaporean residents **who are born in 2017** can expect to live to 83.2”. This other news article in the Straits Times rightly refers to the life expectancy estimate as “expected lifespan at birth”.

**But this still does not fully capture how life expectancy is calculated.** To do this, we need to go back to our understanding of age, period, and cohort measures. Life expectancy, as it is reported yearly, is a **period measure**.

Without going into full detail (interested readers can refer to resources such as this), I will explain briefly how this is calculated. A typical life table from the Department of Statistics (taken from this report) looks like Figure 7.7 (it shows just the top part, the entire table is a few pages long).

Table A1: Complete Life Table for Singapore Resident Population, 2017

Age x (Years)	Probability of dying between exact age x and age x+1	Number of survivors at exact age x	Number of deaths between exact age x and age x+1	Number of person-years lived between exact age x and age x+1	Total person- years lived after exact age x	Expectation of life at exact age x
						$e_x$
$q_x$		$l_x$	$d_x$	$L_x$	$T_x$	
0	0.00238	100,000	238	99,802	8,318,903	83.2
1	0.00012	99,762	12	99,756	8,219,101	82.4
2	0.00012	99,750	12	99,744	8,119,345	81.4
3	0.00011	99,738	11	99,733	8,019,601	80.4
4	0.00010	99,727	10	99,722	7,919,868	79.4
5	0.00009	99,717	8	99,713	7,820,146	78.4
6	0.00007	99,709	7	99,706	7,720,433	77.4
7	0.00007	99,702	7	99,699	7,620,728	76.4
8	0.00007	99,695	7	99,692	7,521,029	75.4
9	0.00008	99,688	8	99,684	7,421,338	74.4

Figure 7.7: Screenshot of Complete Life Table from the Department of Statistics.  
Retrieved September 17, 2019.

1. Age-specific mortality rates (not pictured, but often referred to as  $m_x$ ) for a specific year are obtained from the national mortality database and turned into<sup>16</sup> the probability of dying in a specific age-interval ( $q_x$ , pictured).
2. You start with a hypothetical base population, usually 100,000, and you put them through the mortality rates of each age until there are no more survivors ( $l_x$ ). In other words, the probability of dying at an age-interval

<sup>15</sup>If this is true, then if life expectancy increases faster than one year annually, does this mean we can all expect to be immortal? You can see how ludicrous this sounds.

<sup>16</sup>Some assumptions need to be made here about a quantity called  $a_x$ , which captures when most deaths in a certain age-interval happen within the time period, but usually this calculation is straightforward.

becomes certain ( $q_x = 1$ ) at some point (usually 100+) since no one is immortal.

3. From there, you calculate a number of other things (e.g.,  $d_x$ ,  $L_x$ ,  $T_x$ , using simple math formulas). These numbers then help you derive the life expectancy (basically,  $e_x = \frac{T_x}{l_x}$ ).

You may not yet fully understand what is happening here, but just note one key thing - the key quantities here that affect everything else in the whole process are the age-specific mortality rates ( $m_x$ ). And these age-specific mortality rates used in calculating life expectancy are a **period** measure, which capture what is happening *that particular year*.

To understand what this means for how we interpret life expectancy, we return to our previous interpretation - that “Singaporean residents **who are born in 2017** can expect to live to 83.2”.

Let’s now make it more precise and say that “Singaporean residents **who are born in 2017 and who are exposed to mortality rates of 2017 throughout their lives** can expect to live to 83.2”.

This sounds highly awkward, I know. You would have to imagine some kind of time machine allowing you to stay stuck in 2017, continuing to experience its mortality rates even as you grow older. But it is in fact how the measure is calculated. **Life expectancy, as often cited in the media, is a period measure that is constructed by stitching together the experiences of multiple cohorts** (often called a *synthetic cohort*).

A more intuitive measure of life expectancy is called **cohort life expectancy**. Cohort life expectancy follows every person who was born in a certain year (e.g., 1965), and is calculated using the observed/projected rates of all persons in that cohort. With cohort life expectancy, we can truly say that “Singaporean residents born in year XXXX live XX.X years on average”. So why don’t we do this? The answer is simple - to obtain observed rates, we need to observe these deaths for each cohort! We would therefore need to wait for all (or most<sup>17</sup>) of each cohort to pass away before we can calculate such numbers.

Period life expectancy is therefore practical to use as a measure of how well our health system is doing from year to year, but we must acknowledge its limitations. In order to examine how volatile period rates of life expectancy can be, we can simply look at older rates of life expectancy.

*But first, a small excursus* - if you look back at Figure 7.7, one thing you will realize is that there are life expectancy estimates for each age. Therefore, life expectancy at age 40 is not the same as life expectancy at birth, and they do not usually add up (i.e., Life expectancy at 40  $\neq$  Life expectancy at birth - 40). Knowing this, we can now do a little thought experiment with life expectancy estimates.

---

<sup>17</sup>Using a mixture of observed and projected rates is possible.

Let us now look at life expectancy at birth in the year 1965. Using the Department of Statistics Table Builder (go here) to retrieve the data, it seems life expectancy at birth for Singapore residents back in 1965 was 64.5. Using the common way of interpreting these data, we might say, “we can expect those born in 1965 to live up to 64.5 years old.”

Fast forward to 2017, and those born in 1965 are now 52 years old. If those numbers in 1965 are correct, it follows that we might expect life expectancy at 52 to be  $64.5 - 52 = 12.3$  years. Let’s go back to the Complete Life Tables we were looking at before to see how long more those born in 1965 can expect to live (Figure 7.8).

Table A1: Complete Life Table for Singapore Resident Population, 2017 (Continued)

Age x (Years)	Probability of dying between exact age x and age x+1	Number of survivors at exact age x	Number of deaths between exact age x and age x+1	Number of person-years lived between exact age x and age x+1	Total person- years lived after exact age x	Expectation of life at exact age x
	$q_x$	$l_x$	$d_x$	$L_x$	$T_x$	$e_x$
40	0.00066	98,860	66	98,827	4,340,416	43.9
41	0.00073	98,794	73	98,758	4,241,589	42.9
42	0.00082	98,721	81	98,681	4,142,831	42.0
43	0.00092	98,640	91	98,595	4,044,151	41.0
44	0.00104	98,549	102	98,498	3,945,556	40.0
45	0.00115	98,447	114	98,390	3,847,058	39.1
46	0.00128	98,333	126	98,270	3,748,668	38.1
47	0.00142	98,207	140	98,137	3,650,398	37.2
48	0.00160	98,067	157	97,989	3,552,261	36.2
49	0.00180	97,910	176	97,822	3,454,273	35.3
50	0.00201	97,734	196	97,636	3,356,451	34.3
51	0.00222	97,538	216	97,430	3,258,815	33.4
52	0.00245	97,322	239	97,203	3,161,385	32.5
53	0.00273	97,083	265	96,951	3,064,182	31.6
54	0.00303	96,818	293	96,672	2,967,232	30.6

Figure 7.8: Screenshot of Complete Life Table from the Department of Statistics. Retrieved September 17, 2019.

Wait. We see now that those who aged 52 in 2017 can in fact expect to live 32.5 more years. That means they are expected to live up to  $52 + 32.5 = 84.5$  years old! That seems way different from the 64.5 years projected by mortality statistics in 1965. Why is this so? This is because period life expectancy does not account for future improvements in mortality driven by better healthcare, hygiene, health behaviors etc. Period life expectancy is therefore often an *underestimate* of the true cohort life expectancy<sup>18</sup>. The discrepancy we see between the 1965 life table and the 2017 life table is likely due to great improvements in mortality rates since independence.

Having studied how life expectancy ‘works’, let us conclude by evaluating this paragraph from an article on the Dollars and Sense website, aimed at helping Singaporeans plan their retirement:

The next step of the retirement planning equation is to understand

<sup>18</sup>Unless mortality rates spike due to events such as wars, famines etc.

how long we can expect to spend in retirement. The longer we spend in retirement, the more we need to sustain ourselves.

**How Long Do We Spend In Retirement?** The simplest way to derive this figure is to use the average life expectancy of Singaporeans – which stands at 83.1 years today. This figure is only going to go up as the latest report from the World Health Organisation (WHO) revealed that people here can expect to live up to 85.4 years by 2040.

What this tells us is that if we retire at the moment we are able to receive our CPF LIFE payouts – at 65 – and live to the average life expectancy in Singapore – to 83.1 – we will have close to 18 years of retirement.

**What are the misinterpretations of life expectancy present here?** First, the writer talks about average life expectancy as if it applies to everyone. As we discussed, however, the number he uses likely reflects the life expectancy *at birth*. Is he speaking to newborns and asking them to plan for retirement? A better approach would be to go to the Complete Life Tables published by the Department of Statistics and refer to life expectancy at 50 (or some other reference age that his target audience can identify with). Second, he is correct that life expectancy will probably continue to rise, but the 18 years (obtained supposedly using  $83.1 - 65 \approx 18$ ) is likely an underestimate of how long “retirement” will be (apart from the error already made in the first point). Plan harder, people.

## 7.4 Conclusion

Age, period, and cohort effects are important ways to understand changes in society over time. People most commonly mix up period and cohort effects, such as offering cohort explanations for changes in period measures, or mistaking one kind of change for the other. In many cases, period and cohort rates tell the same overall story. Where they tend to diverge is when there are large societal changes within short periods of time (like in Singapore), as we have seen in the case of life expectancy. Under such circumstances, period measures are usually bad estimates of cohort change.

We often assume that experts and social scientists quoted in newspapers are making statements about societal change based off empirical evidence and research that they have done. But sometimes experts may just be speculating about something there is no real data to show for (e.g., no cohort data, and no research done to project cohort trends). In such a case (and with the application of some common sense), anyone’s speculation may be as good as theirs. Where these speculations happily coincide with reality they may be of some use, but readers should carefully evaluate whether it is suitable for certain numbers to be used to make certain points.

# Chapter 8

## You say; I say; Who confirm?

Contributor: Shannon Ang

Date: 06 February 2020

People argue about how to interpret statistics. Often, parties quickly resort to inflammatory language and mudslinging in public, accusing others of being “disingenuous” and “dishonest”. A dichotomy is typically drawn - essentially, “my” interpretation is absolutely correct and you are “manipulating the numbers”. There are indeed cases where this can be a legitimate claim, such as erroneously interpreting a conditional probability as a total probability (see Chapter 4.3). But most cases are not so clear cut, and cannot be reduced to errors of technicality.

The interpretation of statistics is not context-free - this means that some ways of looking at numbers support certain arguments better than others. Both statistical rigour and domain expertise are needed to establish a strong link between evidence and conclusion. Simply drawing a dichotomy and taking an adversarial approach is very unhelpful when the disagreement is about what certain statistics tell us about larger phenomena (e.g., earth is getting warmer, people have more liberal attitudes).

In this case study, we will look at two very public cases of disagreements around statistics. I will suggest some ways in which we can move past the (unnecessarily) incendiary language being used, and ask the questions that really matter.

### 8.1 SDP vs. MOM

In an application of the Protection from Online Falsehoods and Manipulation Act (POFMA), sitting Minister for Manpower Josephine Teo issued three corre-

tion directions against the Singapore Democratic Party (SDP)<sup>1</sup> for propagating online falsehoods. The SDP appealed against the order, and eventually lost in the High Court (see related news article).

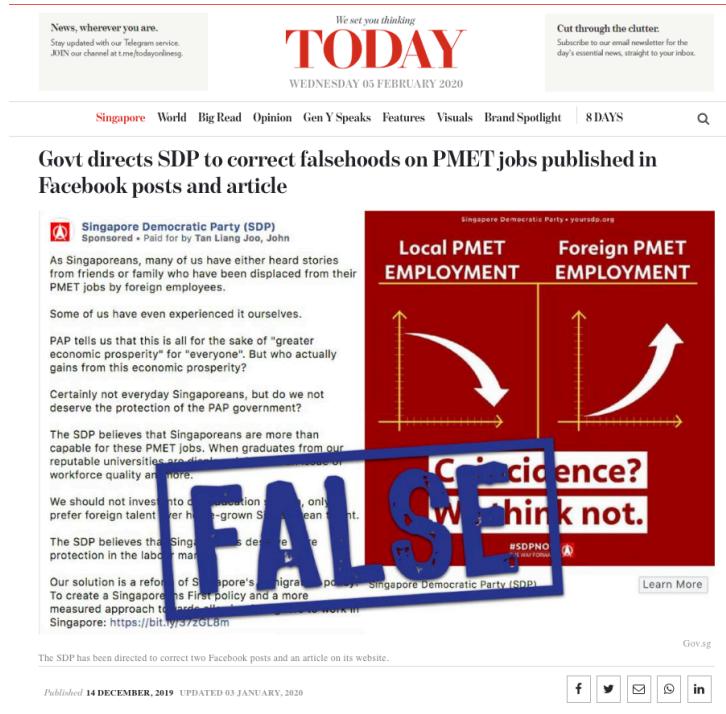


Figure 8.1: Screenshot of online article on MOM's correction directive. Retrieved February 5, 2020.

The point of this case study is not to say whose legal case was made out, which the court has already ruled on<sup>2</sup>. Rather, I will focus on one specific portion of the case, and highlight how we can ask better questions and demand better arguments from people who use statistics.

We will look at the first correction direction (CD-1), related to one of SDP's earlier articles (found here). The point we will examine here is aptly described in the government's Factually article<sup>3</sup>:

<sup>1</sup>A local political opposition party.

<sup>2</sup>Note that Minister for Law Shanmugam said the following in Parliament on 15 August 2016: “Academics or media outlets publish commentaries on a judge’s decision, even where an appeal is pending. These are not prohibited. Commenting on a judge’s reasoning is unlikely to pose a real risk of prejudicing the appeal outcome. We have not changed the law. Such commentaries are allowed today. If the Bill gets passed, they will continue to be allowed under the Act.”

<sup>3</sup>Justice Ang Cheng Hock ultimately decided that the facts provided by Factually were

**Falsehood #2 - The claim that Singaporean PMET retrenchment has been going up.** The SDP website states that the party’s “proposal comes amidst a rising proportion of Singaporean PMETs getting retrenched”. This statement is false.

**Fact #2a:** There has been no rising trend of local retrenchments since 2015.

**Fact #2b:** There has been no rising trend of local PMET retrenchment since 2015. The number of retrenched local PMETs has declined from 6,460 in 2015 to 5,360 in 2018, the lowest since 2014.

In response, the SDP provided their own take on the Ministry’s figures, and showed an overall increasing linear trend from 2010 to 2018. To bypass all this truncation of data, I decided to just plot all the data available. For numbers of *all* retrenchments (including locals and foreigners), you can find the figures I used from here. For numbers of *local* retrenchment, I got them by directly emailing the Manpower Research and Statistics Department at the Ministry of Manpower (MOM).

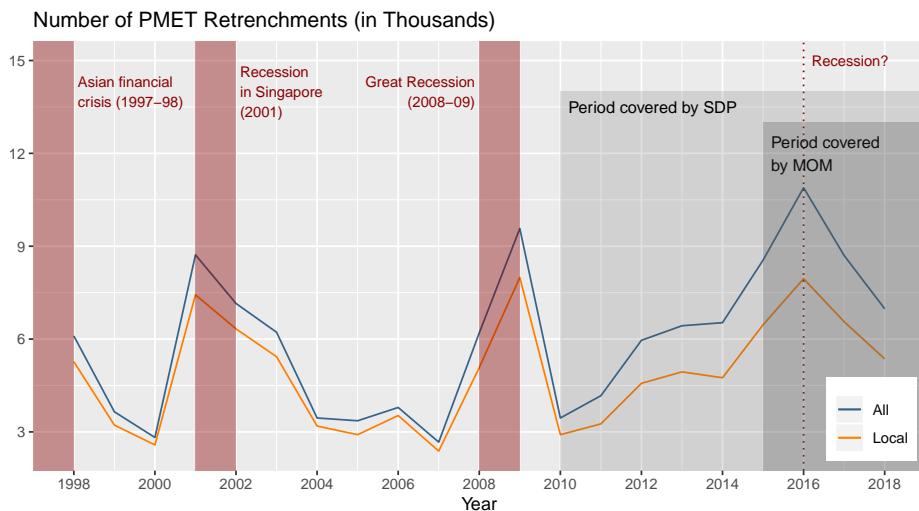


Figure 8.2: Chart of retrenchment rate over time

As we can see from the figure, there are vastly different trends depending on where you want to start from and how long of a trend you want to track. I have highlighted (in grey and dark grey) the portions that were used by SDP

not directly relevant to this protracted falsehood, because it was not fair to interpret SDP’s statement on *proportions* as applying to *absolute numbers* (*Singapore Democratic Party v Attorney-General [2020] SGHC 25* in [80] and [92]). He did, however, consider the arguments made on the grounds of absolute numbers, which we will also consider here.

and MOM respectively. If you look across the whole figure, you will naturally find that in some parts of the graph the trend is increasing, and in other parts the trend is decreasing<sup>4</sup>. But does this mean that choosing a starting point from *anywhere* is arbitrary? Based on my reading of his judgement, Justice Ang Cheng Hock seems to suggest this:

Third, even if one takes the appellant's [i.e., SDP] case at its highest and relies on the data from 2010 to 2018, it bears note that the appellant's reference to 2010 as a starting point is somewhat arbitrary. **Understandably, any timeframe may, to some extent, be criticised as being arbitrarily selected.** However, the appellant's own reasoning could warrant a starting point of 2009, or even 2008. **Using 2009 as a starting point would show an overall decrease in unemployed local PMETs from 2009 to 2018, and this illustrates how using 2010 as a starting point is itself quite arbitrary.**

*Singapore Democratic Party v Attorney-General [2020] SGHC 25 in [96], emphasis mine*

I would respectfully disagree with Justice Ang's reasoning here<sup>5</sup>, if he is implying that any kind of starting point can be criticized as being *equally* arbitrary. In my view, it is possible to discriminate between worse and better time frames/trends to look at. I provide **two reasons** how this is so:

### 1. The length of trends matter

Long-term trends and short-term trends may move in different directions, but parties should be expected to justify *why* looking at either type of trend is more relevant to the argument they are trying to make. This kind of justification should go beyond what the "reasonable person's" default time frame is when reading an/any article, and look instead at exactly what sort of argument is being made (i.e., the topic at hand) to decide if long or short term trends are in view. **While it may seem like people commonly have only "the most recent" trends in mind, there are instances when people naturally think about long term trends. It is easy to see why this is the case with some practical examples.**

One relevant example would be investing in the stock market. Many investors (or financial advisors) use this adage: "It's all about time in the market, not timing the market". Essentially, what this suggests is that in the short term, the value of products such as exchange traded funds (ETFs) will fluctuate up or down, but in the long term (5-10+ years) it will generally appreciate. Go here and play around with the interface to see how trends of the Straits Times Index (STI) change depending on the time frame you pick. Most Singaporeans who

---

<sup>4</sup>You can also find that as the Factually article states, the proportion of local PMET retrenchments among all PMET retrenchments is decreasing.

<sup>5</sup>Even though he qualifies these statements with "somewhat" and "to some extent".

have some form of financial literacy will understand this, and naturally think long-term trends when ETFs are talked about.

Another example is climate change. Scientists have pointed out again and again that tracking short-term trends in temperature *cannot* tell us much about long-term global climate change. In one article, the researchers note that we only start to see meaningful patterns when you look at 15-year averaged trends! Any argument made about climate change therefore needs to address long-term changes (e.g., how has the temperature changed in the past decade), not just short-term ones (e.g., how has the temperature changed from 2-3 years ago).

**There are different benefits to thinking about long and short term trends.** Long term trends are less susceptible to fluctuations due to random variation, and can provide a clearer picture of trends that short-term trends cannot detect (e.g., climate change). On the other hand, short term trends can alert us if something unusual is happening, e.g., a spike in flu-like symptoms may signal a virus spreading, or an increase in rural-urban migration rates in China may signal it is time for Chinese New Year.

My point, therefore, is this - there are good arguments to be made for whether a longer or shorter time-frame should be used. They are not *equally* arbitrary. **Parties who are trying to make their arguments based on these choices (i.e., short or long term) should justify these choices, because it matters substantively.**

## 2. Not all start/end points are made equal:

Just as there are arguments to be made for longer or shorter trends, there are also arguments to be made for what are good “starting points” to use. In Figure 8.2 above, I **highlighted in red** the times when Singapore was facing a recession. As one would naturally conclude from the figure, there are spikes in retrenchment when a recession occurs. A “spike” here simply means that there is a rapid increase, and then a rapid decrease in the numbers, usually due to a exogenous (or endogenous) shock to the usual course of affairs<sup>6</sup>.

The question to ask is this - **how meaningful is it to interpret (whether as a result of policy or not) the increase in retrenchments before a recession, or the decrease in retrenchments after a recession?** I believe it would raise a red flag if someone decided to use data from 2007-2009 to show that retrenchments have increased, or to use data from 2009-2010 to show a decline. I requote from the judgement here:

Using 2009 as a starting point would show an overall decrease in unemployed local PMETs from 2009 to 2018, and this illustrates how using 2010 as a starting point is itself quite arbitrary.

---

<sup>6</sup>Our Ministers have been very consistent in pointing out that Singapore’s economy is highly dependent on the global economy, so assuming recessions are usually exogenous may not be too far off the mark

*Singapore Democratic Party v Attorney-General [2020] SGHC 25 in [96]*

I would respectfully disagree with Justice Ang here. **Using 2009 to illustrate the arbitrariness of using 2010 as a starting point is, in my view, not a fair comparison.** This is because 2009 was the peak of the Great Recession in Singapore, and retrenchments are much more inclined to fall after a recession ends (somewhat of a “regression to the mean” phenomena). To say that retrenchment has decreased since 2009 effectively means that “we have fewer retrenchments now compared to during the Great Recession”, which is a little circular and doesn’t really tell us much. To say that retrenchment has increased from 2010, however, would mean that “we have more retrenchments now compared to since we first recovered from the Great Recession”. Various reasons could be given in both cases, but you should be able to see that there are *meaningful* (and not simply arbitrary) differences between starting from 2009 and 2010. This applies to other periods of recession as well. **How we account for these ‘shocks’, and what meaning we ascribe to them, is important here.**

So these are my two points again:

1. The length of trends matter.
2. Not all start/end points are made equal.

Something that I noted in the course of writing this article was the ‘mini-spike’ at 2016, which looks a little like the other recession spikes. I googled a little and found that there were some **signs of an impending recession in that year** (see, for instance, this article). The question that then came to my mind, given the considerations I’ve pointed out above, was this: **how meaningful is it, then, to interpret the “most recent” 4-year trend provided by MOM? Why should we look at this trend instead of a longer term trend?**

The court judgement seems to focus on one word when determining the appropriate time-frame - “amidst”. I was confused when I first read this, because the judgement suggests that the use of “amidst” is a reasonable basis to focus on a more recent time-frame (*Singapore Democratic Party v Attorney-General [2020] SGHC 25 in [99]*) contemporaneous to the release of the offending SDP article.

Perhaps there is a legal definition I am missing, but my own plain reading of amidst means something like “in the middle of” or “surrounded by” without any specific time-frame (long or short) attached to it. When used with trend data, I assume it means something like “co-occurring”, or “co-varying”. Both short-term trends and long-term trends can covary - e.g., if I say that “corporate greed has increased amidst globalization”, it is likely to refer to long-term trends covarying with each other. But can short-term situations covary with long-term trends? One example may be to say: “Greta Thunberg’s call to action comes amidst accelerating global climate change”. This suggests urgency and contemporaneity of the call to action, but intuitively refers not just to trends of

short-term climate change but also long-term climate change (both short- and long-term trends reinforce each other). So perhaps SDP might have had a better case if they were able to show that their policies were explicitly made to address long-term trends, but this is just speculative.

I will categorically say that I take no sides in this case. **What I am hoping for, however, is that when parties come together to use statistics to make arguments, they can do so with less heat and more light.** News articles on this case reported that Deputy Attorney General Hri Kumar called SDP's arguments "hopelessly flawed", "disingenuous", and a "blatant attempt to manipulate the available statistics in order to ignore the obvious downward trend in recent years and to artificially demonstrate an upward trend". Despite the strong words being used here, the arguments of the Attorney General's Chambers were highly unsatisfactory to me in terms of a concrete statistical justification (and for the record, so were the SDP's arguments), for reasons I've highlighted above.

**What, then, do I expect?** In this particular case, I wished to see more of the following from both parties:

1. **Justification for why long or short term trends matter**, especially with regard to retrenchment figures. Since what is in view is how *policy* affects the labor market - can we see short-term declines in retrenchment after a shock to the market (e.g., a recession) as a result of policy? How do we evaluate the effects of policy in the first place? Do we access them via long-term trends, or short term trends?
2. **Some explanation as to why we were seeing such trends** - i.e., what are these increases/decreases driven by (e.g., composition of term contract/permanent employees)? How do recessions figure into the discussion and affect the way we interpret trends?
3. **More direct engagement to reduce confusion** - As Justice Ang alludes to in *Singapore Democratic Party v Attorney-General* [2020] SGHC 25 in [80] and [92], if one party talks about proportions, the other should engage and provide relevant numbers for proportions, not make a separate argument using absolute numbers.
4. **More statistical rigour in general around topics like this.** It's one thing to look at a few numbers and argue whether the line has gone up or down. But it would be another if the MOM or the SDP could have statistically modelled, for instance, the probability that employers choose locals over foreigners, all else being equal (e.g., education, experience, etc.). This would directly address the allegations of locals being passed over for foreigners.

Ideally, POFMA is a law that pursues and protects truth. Justice Ang is worth quoting in full:

I pause here to highlight that both parties attempted to cast aspersions on each other's intentions and motivations, with labels such as

“disingenuous” and “dishonest” being bandied about. I underscore that the POFMA necessitates an objective approach based on the wording of the material in question. The issues are whether the subject statement(s) are borne out by the words and/or depictions in the communicated material, and then whether those subject statement(s) are true or false. The intentions of the parties in relation to the POFMA are thus, *sensu stricto*, irrelevant when there is no question before the Court of any criminal liability.

*Singapore Democratic Party v Attorney-General* [2020] SGHC 25 in [128]

**In this spirit, my wish is for the court to consider the use of *amicus curiae* when dealing with issues requiring statistical expertise, especially when the determination of “true” and “false” is made somewhat on the interpretation of statistics.** We now move on to a second example.

## 8.2 Really (just) golfing

In January, NUS researchers released a study to show the following:

1. After land sales announcements, some developers play golf with other developers more frequently.
2. Developers who play golf tend to acquire land parcels at lower prices.

While the press release did not include the words “*insider trading*”, nor did a later version of the manuscript that I found dated 3 Dec 2019 (before the media report, which was released a month later), an earlier version of the manuscript dated 29 April 2019 did so. It seems that the earlier version of the manuscript was circulated to reporters, instead of the later version<sup>7</sup>.

The phrase “*insider trading*” triggered the Real Estate Developers’ Association of Singapore (REDAS), who came out strongly against what they saw as an attack against them (see here). In it, we see quite a bit of inflammatory language without any real form of statistical justification. Let us examine the critique as it stands.

REDAS takes these allegations seriously. We question the validity of the methodology used and reserve the right to examine the findings to address the faulty assumptions in the paper.

*Real Estate Developers’ Association of Singapore (REDAS) statement, dated 17 January 2020*

---

<sup>7</sup>Media reports seem to suggest that the paper was edited after REDAS lashed out at the researchers, but I don’t think this is correct - as I mentioned above, I had already found versions of the article without the claim of “*insider trading*” *before* REDAS said anything about it.



Figure 8.3: Screenshot of online article on golfing study. Retrieved February 6, 2020.

**First**, REDAS points out “faulty assumptions” and questions the “validity of the methodology”. We have then, to ask, **what faulty assumptions?** The paper details very clearly the methodology that the researchers used, which includes:

1. How they constructed the dataset through matching golf records, land bidding results, etc.; and
2. Their statistical model, which is a difference in difference (DID) regression.

Briefly, a difference in difference regression is a commonly used method for causal inference. It basically uses the change in an outcome before and after an event (after - before; the first difference), and compares that change across groups (group A - group B; the second difference). Informally, the idea can be represented as such:

$$\text{DID estimate} = (\text{After}_A - \text{Before}_A) - (\text{After}_B - \text{Before}_B)$$

The full model allows for covariates to account for other factors, and can allow the effect to vary based on how long ago something happened. Now, we return to our question: *what faulty assumptions?* At the very least, we should go through each part of the methodology to see what assumptions have been made<sup>8</sup>.

With regard to their **data**, we could ask questions such as: Who is in the dataset and who is not? Is this representative of the people who bid for land parcels? Are there land sales records that are missing that would alter the finding substantially? These questions all get at **selection bias** - basically, whether the data captures information in a way that represents the people we want to study.

We could also ask questions with regard to their **method** (i.e., DID). In addition to the usual assumptions for linear regression models, the DID regression model has a well known assumption that would make its findings questionable if violated - the “parallel trends assumption”. This means that the trends of groups being compared (e.g., A and B) should be parallel in the absence of the event under study (see this for a fuller explanation). **But this is an assumption that is explicitly checked within the study.** The authors report that “[the] results show no violation of the parallel trend assumption.”

**So what “faulty assumptions” are REDAS referring to?** Without any kind of elaboration, the accusations here fall flat.

We are appalled by the lead researcher’s unsubstantiated assertion and the conclusions drawn by the authors are misleading.

*Real Estate Developers’ Association of Singapore (REDAS) statement, dated 17 January 2020*

---

<sup>8</sup>Sometimes, we can search for the word “assume” (or similar) within the manuscript to see what assumptions have been explicitly discussed.

**Second**, REDAS says they are “appalled by the lead researcher’s unsubstantiated assertion and the conclusions drawn by the authors are misleading”. As above, we have to ask: *how so?* **How are assertions “unsubstantiated” and conclusions drawn “misleading”?** As we already discovered above, the authors are very clear on how they arrived at their conclusions. They quite robustly substantiate their findings with statistical evidence. What they find, in my view, is what I stated in the beginning:

1. After land sales announcements, some developers play golf with other developers more frequently.
2. Developers who play golf tend to acquire land parcels at lower prices.

These are rather straightforward findings. Perhaps there is a technical definition of “insider trading” that paints a misleading picture, but to a reader like me, whether or not the actual term “insider trading” is used is besides the point. The finding, as the authors state, is that “social interactions enable developers to realize higher profits, while the government loses land sale revenues”. This means there may be a form of bias in the land bidding system that privileges people with important social connections (i.e., social capital). This is really *not* a contentious finding at all - researchers have found that social capital can lead to finding better jobs or health outcomes, and it remains common sense that maintaining key network connections are advantageous for life in general.

My view is that REDAS looked at these findings and felt attacked because they believe the researchers were making a moral judgment on their professionalism. But even if things *look bad* on the surface, there may be legitimate reasons for observed trends. For example, Minister for Trade and Industry Chan Chun Sing said recently that permanent residents (PRs) took up a disproportionate amount of new jobs created. He reasoned this was because they wouldn’t be taken in as PRs in the first place if they couldn’t get jobs (see this article). The implication is that there is no malice and/or discrimination against locals, even though on the surface it looks like there is. Therefore in the “golfing case”, REDAS should seek to understand the researchers’ findings more thoroughly. Even better, they can even work with them to figure out why we see this happening, and whether they can do anything to address it if it leads to unfair advantages.

In the interest of the community and the nation at large, researchers should publish papers which provide objective and balanced perspective and contribute constructive comments to ensure a stable and sustainable property market.

*Real Estate Developers’ Association of Singapore (REDAS) statement, dated 17 January 2020*

**Finally**, we examine the statement that “researchers should publish papers which provide objective and balanced perspective and contribute constructive comments to ensure a stable and sustainable property market.” As we have already discussed, there is no reason to reject the researchers’ findings as not “objective or balanced”. I have read the paper and found that by most stan-

dards, they present a good case (with regard to the findings I list above) and are transparent about their discussion. The onus is on REDAS to show how these standards are not met, *before* casting aspersions on their results. What is worrying is the final phrase that suggests perspective and comments should only be given “to ensure a stable and sustainable property market”. Surely research is not at service of the property market, contrary to what REDAS is trying to suggest. **If a well-conducted study reveals a phenomenon that upsets the property market, in no way does this automatically mean that it is not balanced and objective. It is unreasonable to want statistics to always operate in your favour, and/or become combative when it doesn’t.** The key question to ask is: why do the numbers say this?

### 8.3 Conclusion

As we have seen, there often are heated arguments about how to interpret statistics. Understandably, people want to interpret statistics in their own favour. It is tempting to then devolve into saying that everything about interpreting statistics is arbitrary, or to paint a dichotomy between who is absolutely right and who is being “dishonest”. But this is quite unhelpful in the pursuit of truth. **There are statistical interpretations that support some arguments better than others, and we have to carefully evaluate each one.** My hope is that instead of precipitously turning to mudslinging and name-calling, people can actually look at the statistics and evaluate the evidence carefully. This will take a high amount of statistical literacy *and* substantive expertise in each domain, but is quickly becoming more important as we move towards a world of more data.

# **Chapter 9**

## **Your case study**

**Contributor:**

**Date:**

Another case study goes here. Do you wish to contribute? Send me an email at [shannon.ang@ntu.edu.sg](mailto:shannon.ang@ntu.edu.sg)

# Bibliography

- Agresti, A. and Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Pearson Education, 4 edition.
- Ang, S. and Chen, T.-Y. (2018). Going Online to Stay Connected: Online Social Participation Buffers the Relationship Between Pain and Depression. *The Journals of Gerontology: Series B*.
- Chan, A., Raman, P., Ma, S., and Malhotra, R. (2015). Loneliness and all-cause mortality in community-dwelling elderly Singaporeans. *Demographic Research*, S15(49):1361–1382.
- Fosse, E. and Winship, C. (2019). Analyzing Age-Period-Cohort Data: A Review and Critique. *Annual Review of Sociology*, 45(1):467–492.
- Ge, L., Yap, C. W., Ong, R., and Heng, B. H. (2017). Social isolation, loneliness and their relationships with depressive symptoms: A population-based study. *PLOS ONE*, 12(8):e0182145.
- Gelman, A., Goel, S., Rivers, D., and Rothschild, D. (2016). The Mythical Swing Voter. *Quarterly Journal of Political Science*, 11(1):103–130.
- Hughes, M. E., Waite, L. J., Hawkley, L. C., and Cacioppo, J. T. (2004). A Short Scale for Measuring Loneliness in Large Surveys: Results From Two Population-Based Studies. *Research on Aging*, 26(6):655–672.
- Lim, K. K. and Chan, A. (2017). Association of loneliness and healthcare utilization among older adults in Singapore. *Geriatrics & Gerontology International*, 17(11):1789–1798.
- Luo, L. and Hodges, J. S. (2020). The Age-Period-Cohort-Interaction Model for Describing and Investigating Inter-cohort Deviations and Intra-cohort Life-course Dynamics. *Sociological Methods & Research*, page 0049124119882451.
- Martin, S. (2009). Comment: Distinguishing Cohort Effects from Age Period Effects on Non-Marital Fertility. *Social Forces*, 87(3):1481–1489.

- Rico-Uribe, L. A., Caballero, F. F., Martín-María, N., Cabello, M., Ayuso-Mateos, J. L., and Miret, M. (2018). Association of loneliness with all-cause mortality: A meta-analysis. *PLOS ONE*, 13(1):e0190033.
- Steptoe, A., Shankar, A., Demakakos, P., and Wardle, J. (2013). Social isolation, loneliness, and all-cause mortality in older men and women. *Proceedings of the National Academy of Sciences*, 110(15):5797.
- Treiman, D. (2009). *Quantitative Data Analysis: Doing Social Research to Test Ideas*. Jossey-Bass.
- Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.
- Wee, E. L., Tsang, Y. T., Yi, H., Toh, A. S., Lee, L. G., Yee, J., Lee, S., Oen, K., and Koh, C. G. (2019). Loneliness amongst Low-Socioeconomic Status Elderly Singaporeans and its Association with Perceptions of the Neighbourhood Environment. *International Journal of Environmental Research and Public Health*, 16(6).
- Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.
- Xie, Y. (2019). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.13.