Singapore Society in Numbers

Edited by Shannon Ang

Last updated 06 July 2019

Contents

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.

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"[1] or claims of "evidence-based policy"[2] 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[3] 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[4], 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.

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

[2] Government agencies such as the Ministry of Social and Family Development often use such a phrase.

[3] (the) National University of Singapore

[4] 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*[5], 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 webpage/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

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

comment on the state of quantitative social science in Singapore (e.g., what is 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[6] 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.

[6] 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[1] 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 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.

[1] 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 if you know about a dataset that should be featured in this list but is not included here.

2.1 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).

NYS is administered by the National Youth Council (NYC)'s research team. More detail 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.

2.2 Retirement and Health Study (RHS)

The RHS is a longitudinal survey of Singapore residents' retirement and healthcare 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[1].

More information is available at 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.

2.3 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.

[1] https://www.cpf.gov.sg/Assets/members/Documents/RHS_FAQ_Booklet.pdf

2.4 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 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[2], 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.

2.5 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.

[2]See some basic descriptives here: 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[1] 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"[2], or that "59 per cent of Chinese find a Malay president acceptable"[3]. 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 Media claim 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[4]. To evaluate public sentiment towards this incident, REACH[5] conducted a poll with 680 Singaporeans aged 15 and above. Of interest here is how results from this poll was represented in public discourse.

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

[2] https://www.todayonline.com/singapore/nearly-1-2-young-sporeans-open-extremist-views-being-posted-online-survey

[3] https://www.straitstimes.com/singapore/majority-willing-to-accept-president-or-pm-of-another-race-but-prefer-one-o

[4] See https://www.channelnewsasia.com/news/singapore/watain-concert-cancelled-christian-community-reaction-shann

[5] The Singapore Government's feedback unit

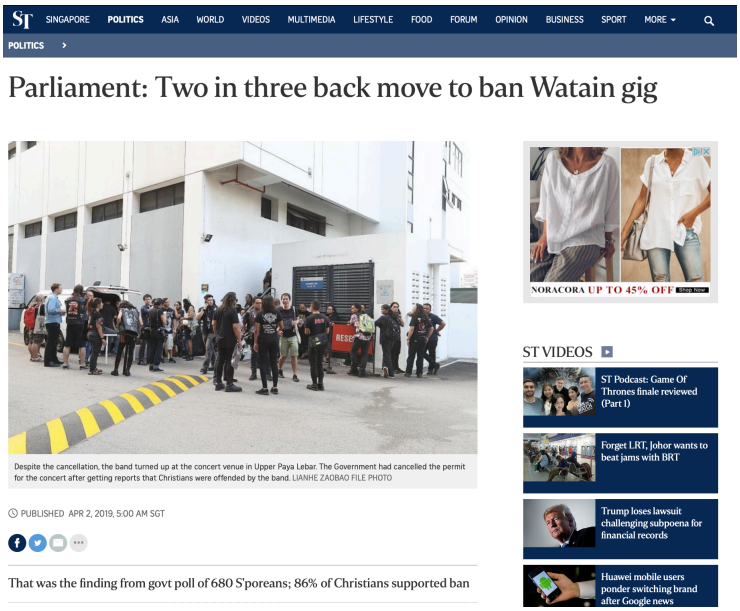

Figure 4.1: Screenshot of online article on results from REACH poll. Retrieved May 21, 2019.

> 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".

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[6]. *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 Media claim 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

[6] 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

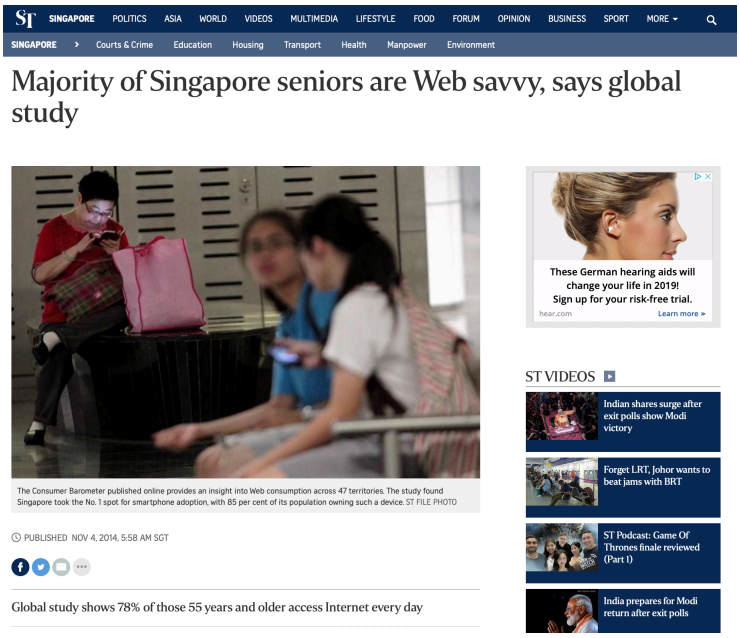

Figure 4.2: Screenshot of online article on web-savvy seniors. Retrieved May 21, 2019.

adults, past research in the United States[7] 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 43[8] is approximately 82% of 53[9]. That is, $\frac{43}{53} \approx 0.82$. This recovers the 82% figure that we see in Figure 4.4.

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

[8]The percent of Singapore residents aged 55+ using the internet in 2017, from Figure 4.3

[9]The percent of Singapore residents aged 55+ using the internet *daily* in 2017, from Figure 4.3

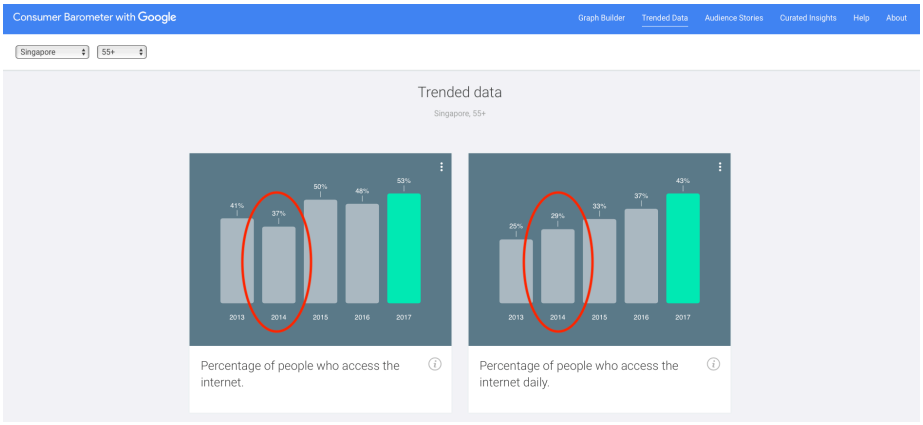

Figure 4.3: Screenshot of Consumer Barometer findings across time. Retrieved May 21, 2019.

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 Appendix

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 remember basic probability rules from secondary (primary?) school[10]. 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(Support Ban | Aware of Ban) is a conditional probability, but the quantity that is being asserted in the news article is Pr(Support Ban), which is the total probability. Using the law of total probability, we know that:

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

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(Support Ban) = 0.64 if and only if Pr(Support Ban | Not Aware of Ban) also equals 0.64. That said, Pr(Support Ban | Not Aware of Ban) is logically impossible, and should equal zero. Similarly, for the Web-savvy Seniors example,

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

where Pr(Use Internet Daily | Don't Use Internet) is impossible and should be zero. In both cases, total probabilites 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

[10] Or that we can Google it if not

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

Additional reading: Straits Times Reporter Christopher Tan alleges (July 5, 2009) 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.

[11] 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[1] 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".

[1] There's no 'full' name for this, its just referred to as the EQ-5D.

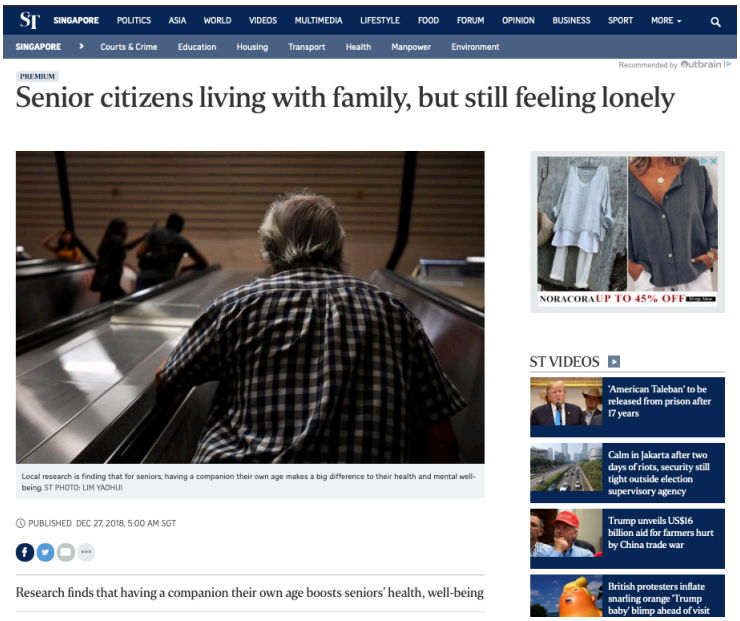

Figure 5.1: Screenshot of online article on lonely seniors. Retrieved May 25, 2019.

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[2]. As we will see, this lonely/not lonely dichotomy is drawn by reseachers as well, but sometimes **in completely different ways**. Let us take a closer look.

5.2 Lonely by whose standard

The studies of interest[3] 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

[2]Led by Associate Professor Angelique Chan

[3]All these studies are open-access articles, anyone can access them even without a library subscription.

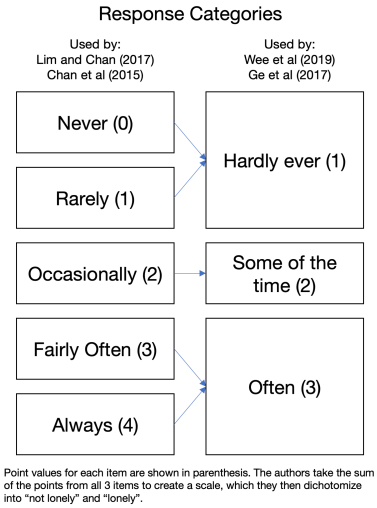

Figure 5.2: Summary of response categories

the 3-item UCLA[4] Loneliness Scale (Hughes et al., 2004). The 3-item UCLA 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".

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[5], 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[6] 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.

[4] University of California, Los Angeles

[5] For instance, one might argue that respondents might choose differently when given more gradational categories.

[6] 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)[7]. 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.

```
# Load required libraries
library(dplyr)

# 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

[7]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

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) 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[8] 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)',
```

[8] 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	15.6	59.0
(2) Living with child only	7.8	50.1
Difference [(1) - (2)]	7.8	8.9
Ratio [(1)/(2)]	2.0	1.2

```
            'and Ge et al (2017). Note that these are row percentages.'),
  booktabs = TRUE)
```

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)[9]. 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 smaller in relative terms (see 5.3). While the proportion of those lonely among those who live alone is 2 times that of those living with only their children according to coding scheme 1, this ratio reduces to 1.2 when using coding scheme 2. Based on these results, it seems that the overall loneliness situation is much less dire than before.

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

[9]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.

requires researchers to engage with each other[10]. Rather, the goal here has been to highlight that decisions like cut-off points may seem small, but are 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.

[10] And indeed, Singapore social science can perhaps use more of this.

Chapter 6

Case study 3

Contributor:
Date:

Another case study goes here. Do you wish to contribute? Send me an email at shannon.ang@ntu.edu.sg

This is an example of in-line code annotation and output.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Figures can be referenced, e.g., see Figure 6.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 6.1.

```
knitr::kable(
  head(iris, 5), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

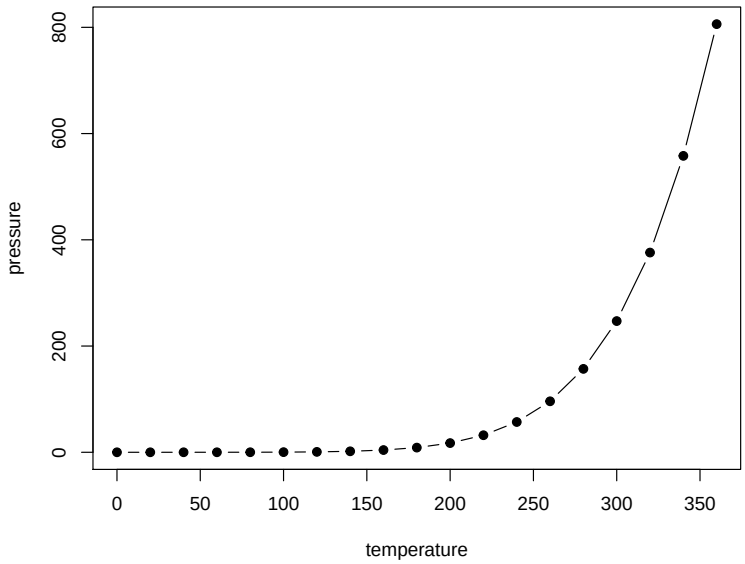

Figure 6.1: Here is a nice figure!

Table 6.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa

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.

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.

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.

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.

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 0.11.