Singapore Society in Numbers

Edited by Shannon Ang

Last updated 22 May 2019

Contents

Preface

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.

Readers with ideas on how to improve this resource (or who may wish to help me maintain it) may 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)

[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.

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.

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

Instructions (tbc) on how to list a dataset, contribute a case study, or write a think piece for this page.

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 include:

About me

Little write-up about myself, which i will insert…

[3](the) National University of Singapore

[4]Notwithstanding the terribly unhelpful stereotype of social science students being "good at writing but bad at numbers".

[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.

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.

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 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 here.

1.2 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

[1] In essence, academics should be able to publish from it.

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 here.

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. Other than a description of the data (as far as possible), this will also include information on how restrictions may be lifted (i.e., how to gain access). 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.

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 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 here. The RHS website does not list any plans to make the dataset publicly available. The RHS study team can be reached at cpf_rhs@cpf.gov.sg.

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

Part II

Think Pieces

Chapter 3

Thinking about Numbers

Think pieces section

3.1 Think piece 1

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

> 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**

[1] https://www.todayonline.com/singapore/nearly-1-2-young-sporeans-open-extremist-views-being-posted-online-survey-shows
[2] https://www.straitstimes.com/singapore/majority-willing-to-accept-president-or-pm-of-another-race-but-prefer-one-of-their-own
[3] See https://www.channelnewsasia.com/news/singapore/watain-concert-cancelled-christian-community-reaction-shanmugam-11399434
[4] The Singapore Government's feedback unit

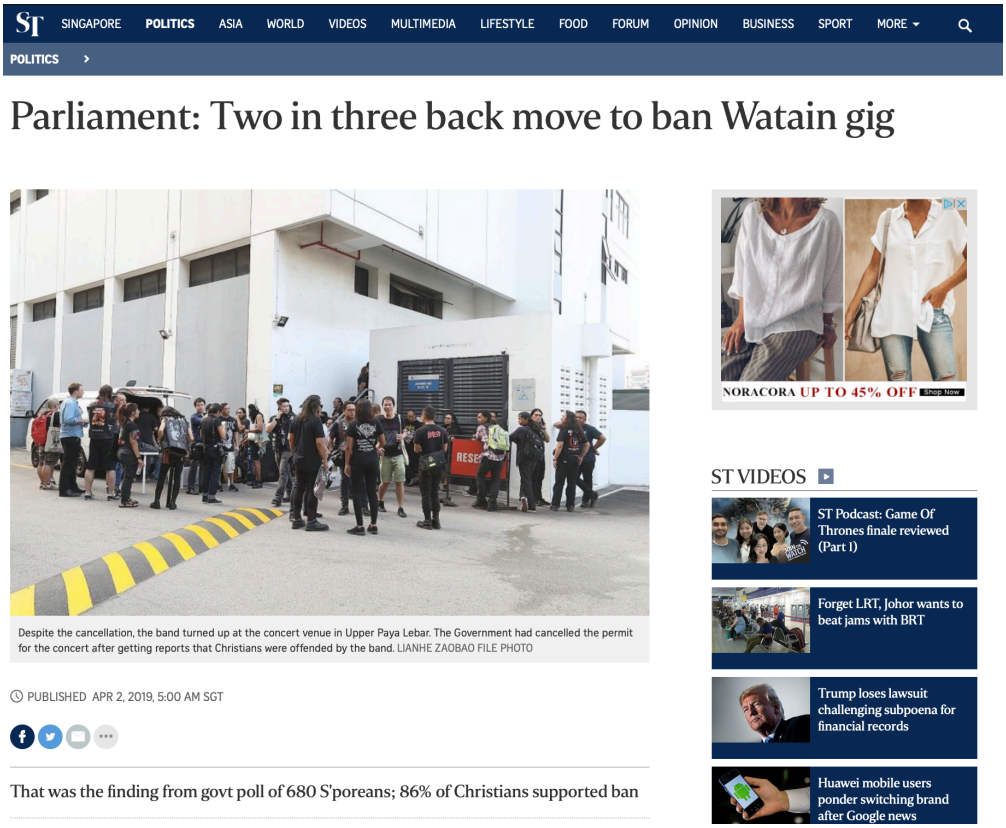

Figure 4.1: Screenshot of online article on results from REACH poll. Retrieved May 21, 2019.

> **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[5]. *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. This means that 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 adults, past research in the United States[6] 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[7] is approximately 82% of 53[8]. That is, $\frac{43}{53} \approx 0.82$. This recovers the 82% figure that we see in Figure 4.4.

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

[5]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

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

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

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

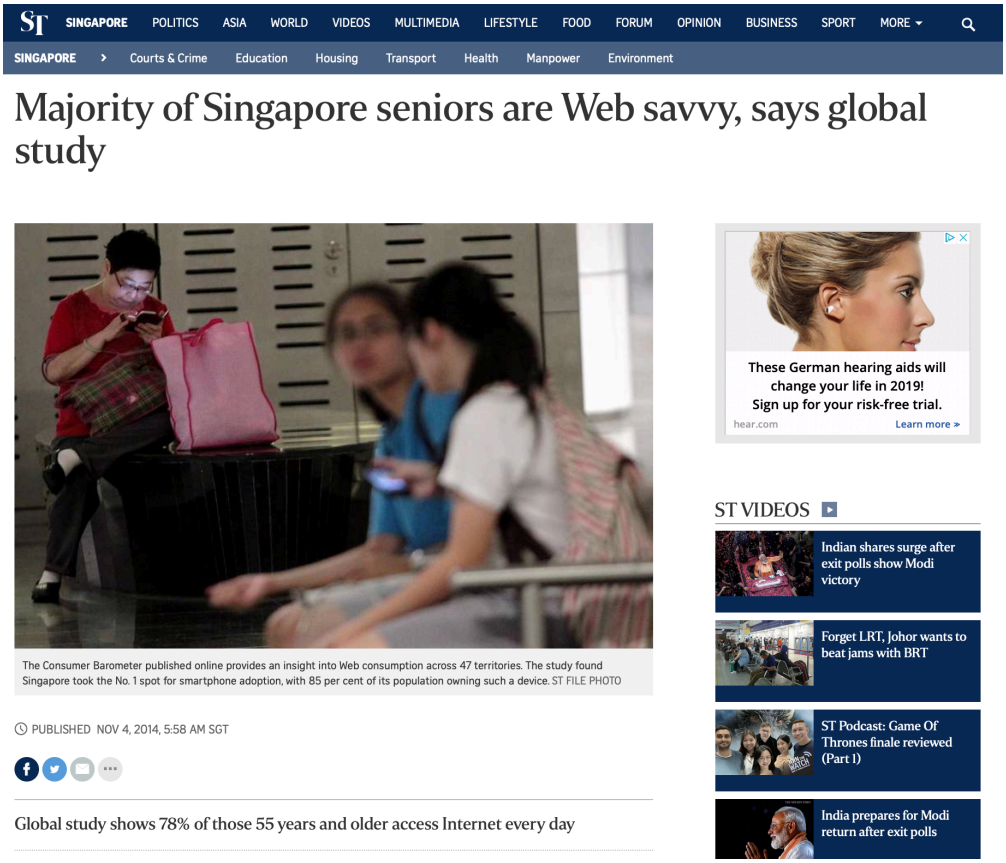

Figure 4.2: Screenshot of online article on web-savvy seniors. Retrieved May 21, 2019.

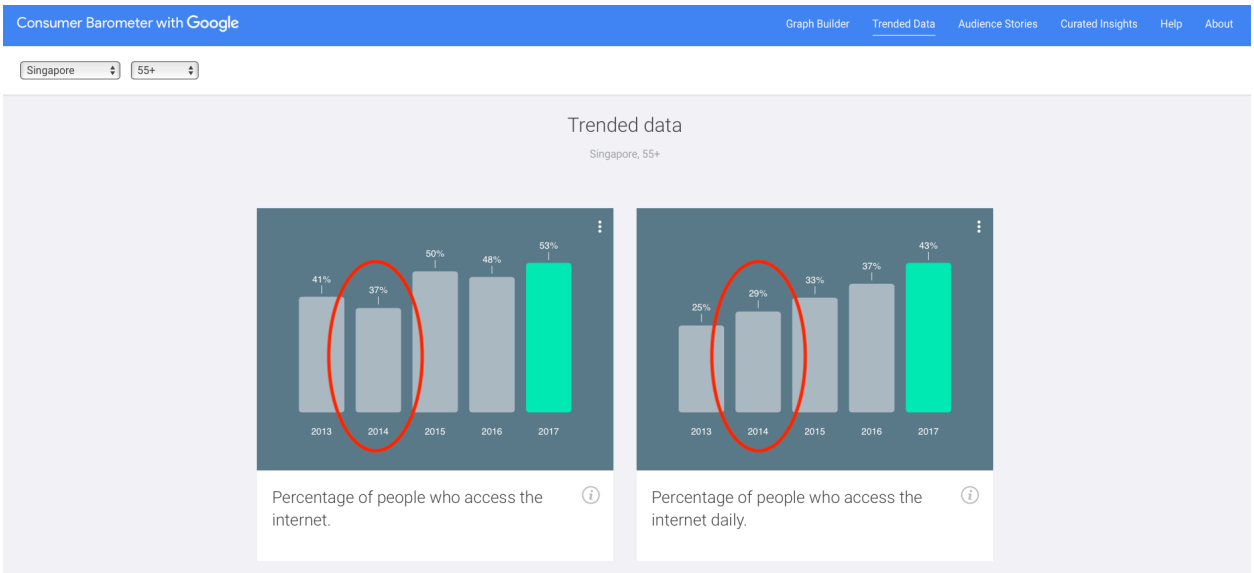

Figure 4.3: Screenshot of Consumer Barometer findings across time. Retrieved May 21, 2019.

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

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[9]. 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:

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

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 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 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[10] in order to evaluate claims that are being made in public discourse today.

[9] Or that we can Google it if not

[10] This, however, first requires data to be made available for replication purposes.

Chapter 5

Case study 2

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 5.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 5.1.

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

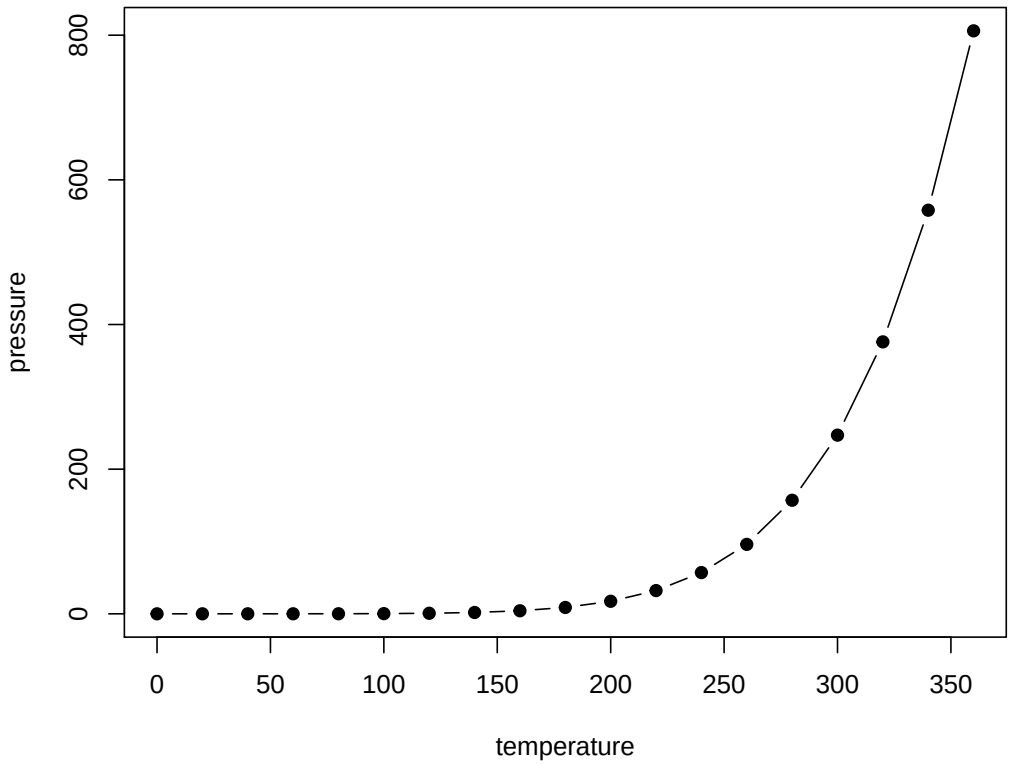

Figure 5.1: Here is a nice figure!

Table 5.1: 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*.

Treiman, D. (2009). *Quantitative Data Analysis: Doing Social Research to Test Ideas*. Jossey-Bass.

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 0.10.