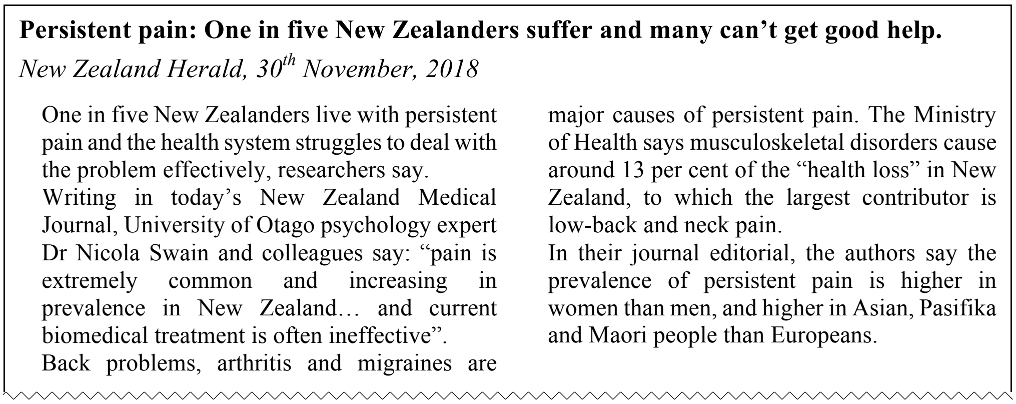
MAKING SENSE OF CATEGORICAL DATA – QUESTION CONFUSION

Stephanie Budgett and Malia Puloka
The University of Auckland, New Zealand
s.budgett@auckland.ac.nz

When students encounter categorical data, lessons often focus on computing probabilities from two-way tables. These computations may involve simple, joint, and conditional probabilities, and the calculation of relative risk. However, little attention has been given to the questions posed. The purpose of this paper is to explore the questions that undergraduate students pose of categorical data, and their reasoning with a variety of representations of categorical data. Results from a small pilot study suggest that when the questions posed involved making comparisons, students were often confused as to whether they should compare proportions between conditions, or compare proportions within a condition.

INTRODUCTION

In today's information-driven society, we are constantly bombarded with data-derived information. In order to interpret this information successfully, statistical literacy is essential. Data on societal topics such as migration, social inequality, health and safety, and education are increasingly available to the general public and can be easily misinterpreted. Consider the following abbreviated excerpt (Figure 1) from a New Zealand media outlet:



Persistent pain: One in five New Zealanders suffer and many can't get good help.

New Zealand Herald, 30th November, 2018

One in five New Zealanders live with persistent pain and the health system struggles to deal with the problem effectively, researchers say.

Writing in today's New Zealand Medical Journal, University of Otago psychology expert Dr Nicola Swain and colleagues say: "pain is extremely common and increasing in prevalence in New Zealand… and current biomedical treatment is often ineffective".

Back problems, arthritis and migraines are major causes of persistent pain. The Ministry of Health says musculoskeletal disorders cause around 13 per cent of the "health loss" in New Zealand, to which the largest contributor is low-back and neck pain.

In their journal editorial, the authors say the prevalence of persistent pain is higher in women than men, and higher in Asian, Pasifika and Maori people than Europeans.

Figure 1. Media excerpt

The data that forms the basis of the journal editorial comment regarding the prevalence of persistent pain referred to in Figure 1 are taken from the New Zealand Health Survey (Ministry of Health, 2017), a nationally representative household survey carried out every year since 2011/2012. Data relating to chronic or persistent pain are taken from adult respondents (aged 15+ years) who answered the question "Do you experience chronic pain?", with the answer options of Yes or No. Chronic or persistent pain was clearly defined to all respondents. A comparison of the incidence of chronic or persistent pain across different ethnic groups was of interest to the authors of the journal editorial. The survey report states that prevalence of chronic pain in 2016/2017 was 21.8% for European/Other, 22.8% for Maori, 14.3% for Pasifika, and 10.8% for Asian. The author of the media excerpt has misinterpreted the journal article which states that "Persistent pain differentially affects ethnic groups in New Zealand. Pasifika and Asian populations are less likely to report pain than Europeans. Prevalence for Maori was complex, having a higher rate when sociodemographic factors were taken into account" (Swain, et al., 2018, p. 6).

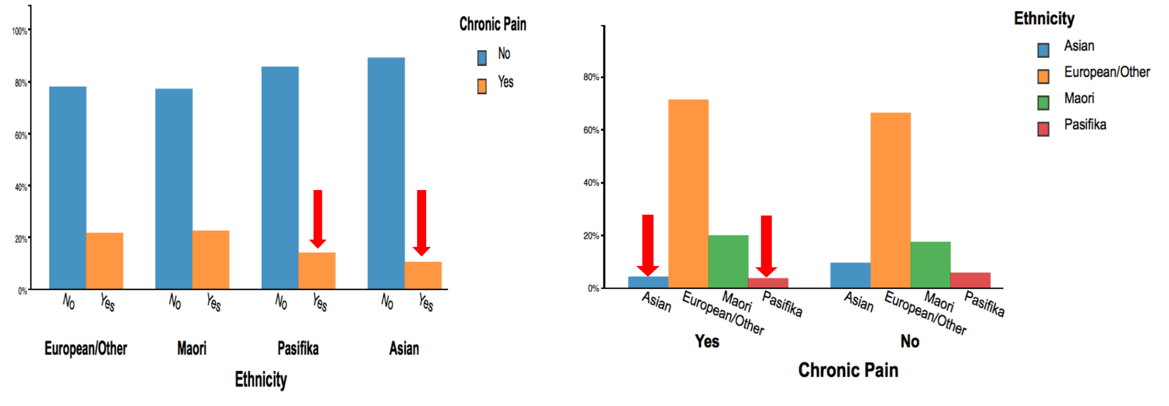
Suppose that we are interested in the following question: *Of people living in New Zealand, who is more likely to report chronic pain, Pasifika people or Asian people?* Rather than the more traditional questions asked of categorical data displayed in two-way tables, this is a more natural question that students are likely to ask (Puloka & Pfannkuch, 2018). To answer this question requires analysis of the relevant data from the New Zealand Health Survey (see https://www.health.govt.nz/publication/annual-update-key-results-2016-17-new-zealand-health-survey). Information on ethnicity and presence of

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. ©2019 ISI/IASE iase-web.org/Conference_Proceedings.php

persistent or chronic pain is provided via two categorical variables. The first variable, Ethnicity, has four possible outcomes: European/Other, Maori, Pasifika, Asian, and the second variable, Chronic Pain, has two possible outcomes: Yes, No. Three representations of the New Zealand Health Survey information for these two variables in 2016/2017 are provided in Figure 2.

		Asian	European/Other	Maori	Pasifika	Total
Chronic Pain	Yes	142	2225	626	120	3113
	No	1176	7980	2121	723	12000
	Total	1318	10205	2747	843	15113

(a)



(b) (c)

Figure 2: Representations of Chronic Pain and Ethnicity

In order to answer the question above, do we compare the proportion of Pasifika people who report chronic pain with the proportion of Asian people who report chronic pain? If so, the relevant comparison is between 120/843 and 142/1318, or 0.142 and 0.108, a comparison *between* the Ethnicity conditions Pasifika and Asian. The answer is then that Pasifika people are more likely to report chronic pain than Asian people (see Figure 2 (b)). Or do we compare the proportion of people who report chronic pain who identify as being Pasifika with the proportion of people who report chronic pain who identify as being Asian? If so, the relevant comparison is between 120/3113 and 142/3113, or 0.039 and 0.046, a comparison *within* the condition Chronic Pain. The answer is then that those who report chronic pain are more likely to be Asian than Pasifika (see Figure 2 (c)). Depending on the approach used, the answers are different because the conditioning is different. Is there anything in the question that suggests which way around the conditioning should be?

BACKGROUND LITERATURE

Making sense of the data described in the two examples above requires statistical literacy, critical thinking, contextual knowledge and proportional reasoning. Both situations involve categorical data, or data that can be divided into groups. Categorical data is typically presented in tables of counts or in bar graphs. Determining which proportions are relevant versus which are misleading can be subtle. Reasoning with categorical data is problematic for students and teachers alike, with lack of proportional reasoning being a major contributing factor (e.g. Batanero et al., 1996; Böcherer-Linder et al., 2018; Watson & Callingham, 2014). Furthermore, according to Konold, Finzer and Kreetong (2017), research on the use of data in tables has been neglected possibly because tables are so ubiquitous and the ability to interpret them has been taken for granted.

Research has documented many of the difficulties encountered by students when reasoning with conditional probabilities (e.g. Diaz, Batanero, & Contreras, 2010). In particular, the base rate fallacy and confusion of the inverse are two prevalent misconceptions associated with Bayesian-type problems (Bar-Hillel, 1980; Villejoubert & Mandel, 2002). However, in Bayesian-type problems the conditioning is usually made explicit. For example, in Eddy's (1982) influential study, 100 doctors were given some information regarding the accuracy of mammography. They were then asked to estimate the probability that a woman with a positive mammogram actually has breast cancer, thus making the condition (positive mammogram) unambiguous. However, the comparison questions described above do not follow the same format and the condition may be unclear.

METHOD

The data that forms the basis of this paper comes from a small pilot study designed to explore undergraduate students' questioning of categorical data and their reasoning with a variety of representations of categorical data. The study comprised two first-year university statistics students, with pseudonyms Sera and Tara, who volunteered to participate and worked together on several tasks over two 2-hour sessions. The participants were provided with scenarios involving two categorical variables, and encouraged to pose questions of the data. They were then asked to either construct their own representations of the data, or to select pre-prepared representations of the data, in order to assist in answering their questions. A 'think-aloud' protocol was used, whereby the participants were encouraged to verbalise their thoughts and actions. Occasionally the researchers would intervene to clarify what the participants were thinking.

During the first 2-hour session, the participants were provided with information about an online survey completed voluntarily by first year introductory statistics students. Without being given access to the data, they were asked to pose investigative questions relating to the two categorical variables Gender (Male, Female) and Student Loan (Yes, No). After several questions had been posed, the participants and the researchers collectively grouped them by question-type: simple, joint, conditional and comparison questions. The participants were asked to create representations to display Gender and Student Loan both separately and combined, and to choose which of these representations, if any, could be used to answer their questions. They were then presented with pre-prepared representations of the Student Loan and Gender information gathered in the online survey. They were first asked to interpret the representations before deciding which, if any, would answer their questions.

In the second 2-hour session, the participants were asked to explore the variables Gender and typical Social Media usage per day (none, < 1 hour, 1-3 hours, 3-6 hours, > 6 hours). Following a similar sequence to that in the first session, investigative questions were posed and then grouped into question-type in collaboration with the researchers. Instead of creating their own representations, the participants interpreted pre-prepared representations and were asked to choose which, if any, would answer their questions.

RESULTS

The focus of this paper will be on the comparison questions posed by the participants in both sessions (Figure 3), and their selections and interpretations of the pre-prepared representations used to answer these questions.

Session One	Session Two
Who is more likely to have a student loan, males or females?	Are females more likely to spend 3-6 hours on social media than males?

Figure 3. Comparison questions posed by students

Session One

The representations forming the basis of the participants' reasoning as they answered the comparison question posed in Session One are given in Figure 4. These representations consist of bar graphs and eikosograms, often referred to as mosaic plots, created using an updated version of a prototype software tool developed at the University of Auckland (Pfannkuch & Budgett, 2016).

When the participants attempted to answer the question "who is more likely to have a student loan, males or females?", they were confused as to which representation to use. Sera, referring to Figure 4(a), initially said that males were more likely to have a student loan. She reasoned that the bar representing the proportion of males with a student loan was higher than the corresponding bar representing the proportion of females with a student loan. However, when she considered the representation in Figure 4(b), she decided the opposite was true:

> *If you compare these two together* [indicating the bars representing those with student loan who are females and those who are males in Figure 4(b)] *then it's more likely for females to have a student loan, and then compare these two* [comparing the proportion of males with a student loan with the proportion of females with a student loan in Figure 4(a)] *to say that it's more likely for males than females to have a student loan.*

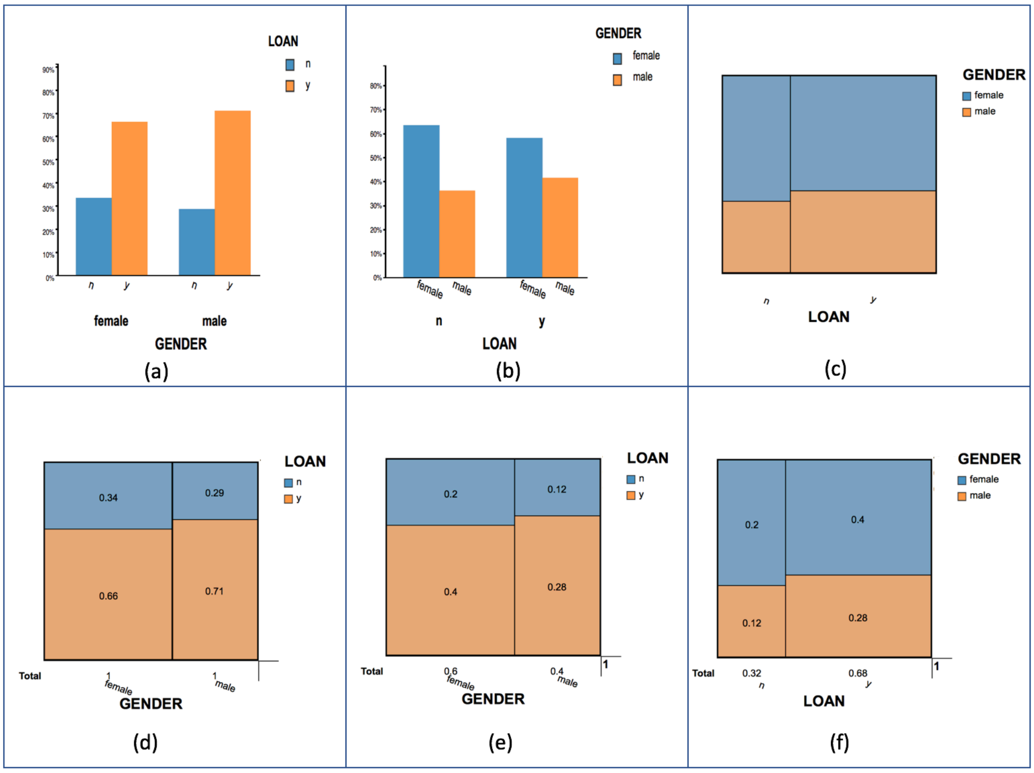


Figure 4. Representations of Gender and Student Loan

When Sera then looked at the eikosogram representation in Figure 4(c), she seemed convinced that females were more likely to have a student loan than males.

Tara referred to the eikosogram representations in Figure 4(d) and (e) in order to answer the question. She reasoned:

> *I like things out of 100 instead of technically 60* [opting to use Figure 4(d) rather than Figure 4(e)] *because if you go out of 100 then you can use percentages and turn it into a decimal, whereas 0.4 out of 0.6 is kind of … I was going to say it is 71% of males have a student loan*

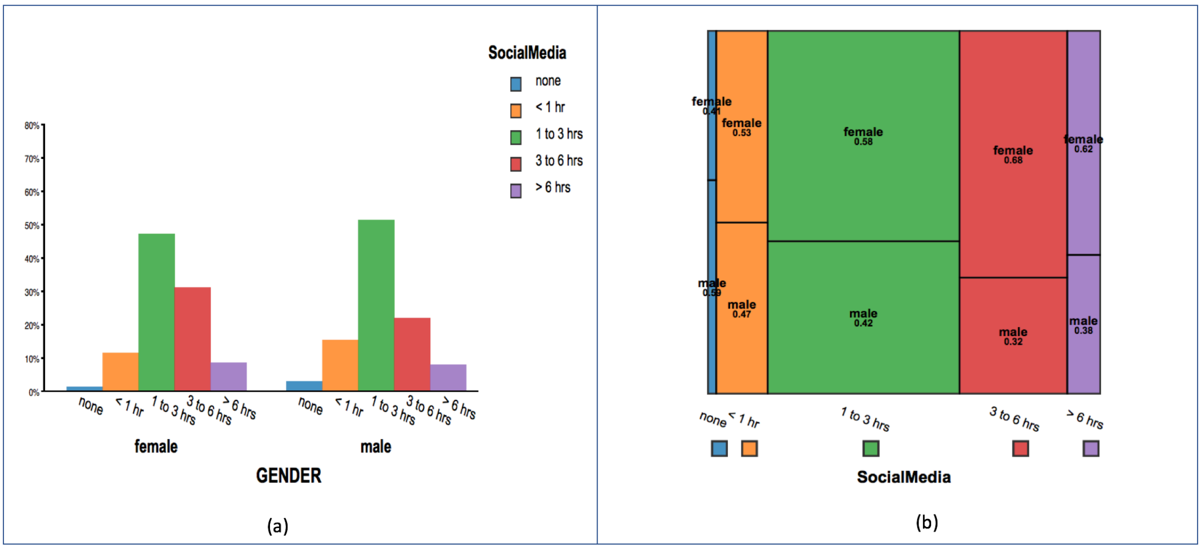
However, when Tara noted the eikosogram in Figure 4(f) she decided that females were more likely to have a student loan:

> *That's so weird but that makes sense because you go this way so you are comparing 0.4 with 0.28. You can move across… comparing them down and so yeah, females would be more likely to have a student loan.*

Both Sera and Tara arrived at conflicting answers to the question *Who is more likely to have a student loan, males or females?* which appears to be a consequence of the representation that they use to support their answer. For Sera, Figure 4(a) suggests that males are more likely, while Figures 4(b) and (c) suggest that females are more likely. For Tara, Figure 4(d) suggests that males are more likely, while Figures 4(e) and (f) suggest females are more likely. The answer to the question varies according to the conditioning used. If the comparison is made *between* the conditions male and female, males are more likely to have a student loan than females. However, if the comparison is made *within* the condition of those having a student loan, then there are more females than males.

Session Two

The variables explored in the second session were Gender and Social Media usage. The pre-prepared representations forming the basis of the participants' reasoning as they answered the comparison question posed in Session Two are provided in Figure 5.



(a)

(b)

Figure 5. Representations of Gender and Social Media usage

Figure 5(a) illustrates the distribution of Social Media Usage for females and males separately. When prompted to interpret the red bars in Figure 5(a), Tara stated:

> *Out of the three to six-hour group females are more likely to spend three to six hours in comparison to males, but only out of the three to six-hour group.*

When asked what question could have been posed to result in this comparison, Sera responded: *Are females more likely to spend three to six hours on social media compared to males?* Later in the session, she referred to Figure 5(b) to answer the same question, with the explanation:

> *Yes, females are more likely to spend three to six hours* [on social media] *because out of the people who spend three to six hours on social media, 68% of them are females.*

Again, the representations used by the participants to support their interpretation of the situation influenced their responses.

DISCUSSION

Throughout the two sessions, the participants used different representations to answer the same question. It was not clear to them which was the more appropriate representation for a given question. When posing 'more likely' questions (Figure 3), the participants were confused as to whether the comparison should be between conditions, or within a condition. In the Student Loan example the participants reasoned, by referring to representations conditioned on Gender (Figs. 4(a), (d)), that males were more likely to have a loan. However, when they considered representations conditioned on Student Loan (Figs. 4(b), (c), (f)) they decided that females were more likely to have a loan. Although the supporting representations for the Social Media example did not change the participants' answer to the question posed, the information conveyed in Figure 5(a) is conditioned on Gender, while the information conveyed in Figure 5(b) is conditioned on Social Media usage.

Much of the literature documenting misunderstandings about conditional probability relates to confusion in identifying the appropriate conditioning variable despite the conditioning variable seemingly being made explicit (e.g. Eddy, 1982). The examples used in this study differ in that the questions are framed using more natural language. It may be that when more naturally formed questions are posed, the conditioning variable may be ambiguous. Therefore, how *should* these questions be answered? The first author presented the following question, similarly phrased to questions in Figure 3, individually to eight colleagues, comprising two statisticians and six teachers of undergraduate statistics:

"*Who's more likely to get lunch from the tuck shop, boys or girls?*", accompanied by Table 1. The aim was to discover if these 'experts' would reach consensus in identifying the conditioning variable.

Table 1. Two-way table of information on gender and lunch

		Gender	
		Boy	Girl
Lunch from:	Tuck shop	6	8
	Home	4	7

Seven responded by comparing the proportion of boys who got lunch from the tuck shop (0.60) with the corresponding proportion of girls (0.53) – a comparison between conditions – thereby answering 'boys' while one of the statisticians noted that 8 of the 14 children who got lunch from the tuck shop were girls – a comparison within a condition – thereby answering 'girls'. The statistician who had originally made a comparison between conditions was later asked why he chose not to make a comparison within a condition and commented that either response was valid. Thus the 'experts' did not reach consensus in answering what, on the face of it, is a natural question to ask.

Given the confusion experienced by the two participants in this small exploratory study, the anecdotal lack of consensus described above, the research documenting the problems with interpreting conditional probabilities (e.g. Diaz et al., 2010) and the suggestion that the interpretation of information in two-way tables has been taken somewhat for granted (e.g. Konold et al., 2017), future research is warranted, particularly on posing questions related to categorical data.

REFERENCES

Bar-Hillel, M. (1980). The base rate fallacy in probability judgments. *Acta Psychologica, 44*, 211–233.

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education, 27*(2), 151–169.

Bocherer-Linder, K., Eichler, A., & Vogel, M. (2016). The impact of visualization on understanding conditional probabilities. *Proceedings of the 13th International Congress on Mathematical Education*, (pp. 1–4). Hamburg.

Diaz, C., Batanero, C., & Contreras, J. M. (2010). Teaching independence and conditional probability. *Boletin de Estadistica e Investigacion Operativa, 26*(2), 149–162.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: heuristics and biases* (pp. 249–267). Cambridge: Cambridge University Press.

Konold, C., Finzer, W., & Kreetong, K. (2017). Modeling as a core component of structuring data. *Statistics Education Research Journal, 16*(2), 191–212.

Ministry of Health. (2017). *Annual Update of Key Results 2016/2017: New Zealand Health Survey.* Wellington: Ministry of Health.

Pfannkuch, M., & Budgett, S. (2016). Reasoning from an Eikosogram: an exploratory study. *International Journal of Research in Undergraduate Mathematics Education*, 1–28.

Puloka, M. S., & Pfannkuch, M. (2018). Year 13 students' reasoning from an eikosogram: An exploratory study. In M. A. Sorto, A. White, & L. Guyot (Ed.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics, Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute.

Swain, N., Parr-Brownlie, L. C., Lennox Thompson, B., Darlow, B., Mani, R., & Baxter, D. (2018). Six things you need to know about pain. *New Zealand Medical Journal, 131*(1486), 5–8.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: an account of deviations from Bayes theorem and the additivity principle. *Memory & Cognition, 30*, 171–178.

Watson, J. M., & Callingham, R. (2014). Two-way tables: issues at the heart of statistics and probability for students and teachers. *Mathematical Thinking and Learning, 16*(4), 254–284.