An Excel-based Introduction to Analysing Buyer Behaviour Using Consumer Panel Data

Bruce G. S. Hardie
London Business School

November 2022

The writing of this note was supported by the London Business School's Research and Materials Development Fund.

This document and the associated data files and spreadsheets can be found at `http://brucehardie.com/notes/042/`.

Contents

Chapter 1

Introduction

Several leading market research firms (e.g., Kantar, IRI, Nielsen) collect so-called consumer panel data and sell reports and analyses based on these data to interested parties. A large number of households are recruited to join the panel and they record all their grocery/HBA purchases. The idea of a panel is that we have repeated observations of the same people. The ability to track what individual households are purchasing over time can give important insights into what behavioural changes lie beneath observed changes in aggregate sales data.[1] When consumer panel operations were originally founded in the 1940s, panellists recorded their purchases in paper diaries that were sent to the market research firm (for processing, etc.). These days, it is common for panellists to record their purchasing by scanning product barcodes via an app on their smartphones.

The objective of this note is to provide an introduction to basic analyses we can undertake using panel data. All the analysis will be undertaken in Excel.[2] After briefing describing the data we will be working with (Chapter 2), we present some preliminary aggregate-level analyses (Chapter 3). Next we introduce some basic brand performance measures and consider the simple analyses that describe the variation we observe in buyer behaviour in a given time period (Chapter 4). This analysis focuses on one brand at a time; in Chapter 5 we consider some basic analyses that describe consumers' buying of multiple brands in a category. We finish (Chapter 6) with some basic analyses that describe how buyer behaviour evolves over time, both for established products and new products.

This note is written with three audiences in mind:

- *Newcomers to panel data.* Most marketing students are not exposed to consumer panel data and have no idea as to how it can be used to gain

[1] Throughout this note we use the terms panellist and household interchangeably.

[2] It is more efficient to perform these analyses using other software, with R and Python being obvious free options. However, the advantage of performing the analyses at least once in Excel is that it makes the process completely transparent to all.

insights into the customer behaviour that lies beneath the aggregate sales numbers reported by the firm. This note attempts to fill (some of) that gap.

- *Current "consumers" of panel data reports.* A surprising number of people who use panel data reports in their jobs do not fully understand what the various reports actually represent. Years of teaching experience suggests that the process of creating a report from scratch — if only once — deepens the user's understanding of what the report actually represents, and can lead them to go beyond the stock reports, requesting custom reports that give even greater insight into buyer behaviour.

- *People with no interest in FMCG (fast-moving consumer goods) but who have access to customer transaction databases.* Whereas consumer panel data gives us information on the purchasing of a sample of customers for the whole category, the data in a firm's transaction database gives us complete information on the purchasing of our products (but not those of our competitors). At a fundamental level, the types of reports developed by market research firms over the past 60+ years are a good starting point for the types of reports a firm should create as it seeks to understand the buying behaviour of its customers. As with the second audience above, the process of creating various reports from scratch is a good way of really understanding what information the report is (and is not) conveying.

The reports considered here do not represent the universe of possible reports. It is hoped that, after working through this note, you will be able to think of additional reports that would be useful. Furthermore, you will have the basic set of skills to be able to produce them for yourself using Excel.

It is recommended that you work your way through this note starting with a blank Excel workbook. The "solution" spreadsheets should be viewed as model answers, against which you can compare your analyses. (Simply looking through them is no substitute for creating all the reports for yourself.)

It is assumed that you have some familiarity with pivot tables in Excel. (If not, help can be found on the Microsoft Office support website or via various tutorials on YouTube.) Some of you will be able to look at a completed report and quickly work out how it was created. For the rest of you, this note provides detailed step-by-step instructions. If you are a relative Excel novice, the process of creating all the reports for yourself should help develop your Excel skills, independent of the application context considered here.

Typographical Conventions

The following typographical conventions are used in this note:

Bold
: Used for the names of worksheets.

Underlined
: Used for the name of a column heading in a worksheet.

`Constant width`
: Used for formulas and file names.

Chapter 2

Data

A traditional consumer panel works in the following manner.

- When an individual first joins the panel, they complete a detailed questionnaire. A section of this questionnaire focuses on the demographics of their household. (This is typically updated once a year.)
- After each shopping trip, each panellist records their purchases, scanning the barcode associated with each product and recording other information such as where the purchase was made (store or channel), the price paid, and the use of promotional deals. Twenty years ago, this would have been done using a custom handheld barcode scanner provided by the market research firm. These days, it is more common to use a smartphone app.[1,2]
- These data are uploaded to the research firm's servers and merged with the purchases records of the other panellists. Each barcode is matched with detailed product information (e.g., category, brand, size, flavour) and this information is also stored in the database. The analyst can then create (typically product-category-specific) datasets for further analysis that tell us what each panellist purchased, when and where, and associated transaction- and/or product-specific data that may be of interest.

[1] If a standalone handheld scanner is used for data entry, a booklet with custom barcodes is provided for items that do not have barcodes, such as fruit and vegetables. This is not necessary if a smartphone app is used to collect the data, as this information can be entered manually using the app.

[2] There are several companies that collect data by getting panellists to take photographs of their purchasing receipts via an app on their smartphones and then use "AI" to extract information on product purchasing. Some market research firms are combining such data collection methods with their traditional panels. A discussion of the pros and cons of the data provided by these receipt panels is beyond the scope of this note.

- Panellists drop out of the panel all the time, and the research firm will be recruiting replacement households on a regular basis. When creating a dataset for further analysis, it is generally desirable to work with a so-called static sample of panellists, which comprises all those panellists active using the time period of interest; new panellists, as well as those that dropped out during the given time period, are excluded.

2.1 Description of Datasets

We will make use of two datasets as we explore the basic types of summaries of buyer behaviour that can be created using consumer panel data. The first contains data on the purchasing of established brands in a mature product category, while the second contains data on the purchasing of a new product. Both datasets were created using static samples.

While these are small datasets and contain a subset of the information available in the research firm's databases, they are more than sufficient to convey the logic of creating the key summaries of buyer behaviour.

Neither dataset includes data on the demographics of each panellist. As such, we will not consider how to create reports that explore how behaviours differ across demographic groups (e.g., by age or geography). However, anyone comfortable with the analyses undertaken in this note should be able to work out how to create such reports for themselves.

2.1.1 Dataset 1

The file `edible_grocery.csv` contains two years of data from a panel of 5021 households on their purchasing in an unnamed edible grocery product category. (We intentionally do not identity the category and the associated brand names.) There are 119 SKUs[3] in this category. 91 SKUs are associated with the four largest brands in the category, which we have named Alpha, Bravo, Charlie, and Delta. The remaining SKUs belong to very low-share brands and we grouped them under the brand Other.

Each record in this file consists of seven fields:

panel_id	A unique identifier for each household.
trans_id	A unique identifier for the purchase occasion.
week	The week number in which the purchase occurred. Week 1 corresponds to the calendar week starting on January 1, 20xy. Week 53 corresponds to the calendar week starting on December 31, 20xy.
sku_id	The SKU code.

[3] A stock-keeping unit (SKU) is a unique combination of the attributes (e.g., brand, package type, package size, flavour) that define the products in the category.

units	The number of units purchased on the particular purchase occasion.
price	The price per unit paid at the point of purchase.
brand	The brand associated with the SKU purchased.

The associated file `sku_weight.csv` gives us the weight (in grams) of each SKU. There are two fields: sku_id and weight.

2.1.2 Dataset 2

"Kiwi Bubbles" is a masked name for a shelf-stable juice drink, aimed primarily at children, which is sold as a multipack with several single-serve containers bundled together. Prior to national launch, it underwent a year-long test conducted in two of IRI's BehaviorScan markets.[4] The file `kiwibubbles_trans.csv` contains purchasing data for the new product, drawn from 1300 panellists in Market 1 and 1499 panellists in Market 2. (The purchasing of other brands in the category has been excluded from the dataset.)

Each record in this file consists of five fields:

ID	A unique identifier for each household.
Market	1 or 2.
Week	The week in which the purchase occurred.
Day	The day of the week in which the purchase occurred. (The product was launched on day 1 of week 1.)
Units	The number of units of the new product purchased on the particular purchase occasion.

[4] BehaviorScan was a panel-data-based service run by IRI used primarily for advertising testing and new product tests.

Chapter 3

Preliminaries

With the exception of the new products analysis section in Chapter 6, all of our analysis will make use of the first dataset. Before we start analysing household-level behaviour, let us first get a sense of the general sales patterns observed in this category.

Our initial objective is to create Figures 3.1 and 3.2, which plot weekly revenue for Alpha and the overall category, respectively, and Figure 3.3, which plots weekly (volume) market shares for Alpha and Bravo.

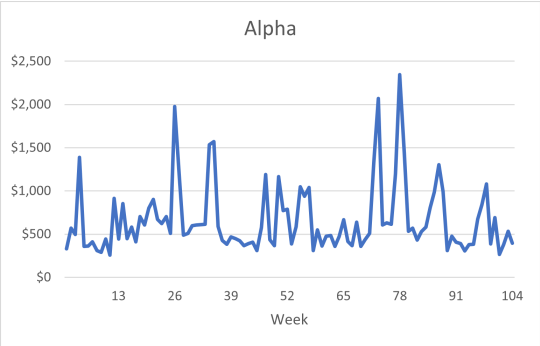

Figure 3.1: Plot of weekly revenue for Alpha

In order to create Figures 3.1 and 3.2 (and equivalent plots for the other brands), we first need to create a table that reports total weekly revenue for each brand and for the category. Category revenue is obviously the sum of the revenues associated with each brand. (This should be a reasonably simply exercise, so see if you can create this summary table for yourself before reading any further.)

- We start by opening `edible_grocery.csv` in Excel. We immediately save it as an Excel workbook, say `chapter_3.xlsx`.
- Looking at units, we see that sometimes more than one unit of a SKU was purchased on a given transaction. We create a spend variable,

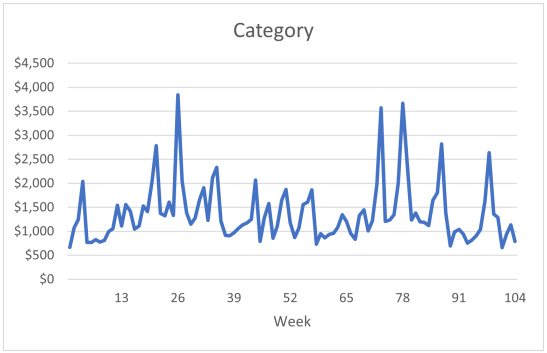

Figure 3.2: Plot of weekly category revenue

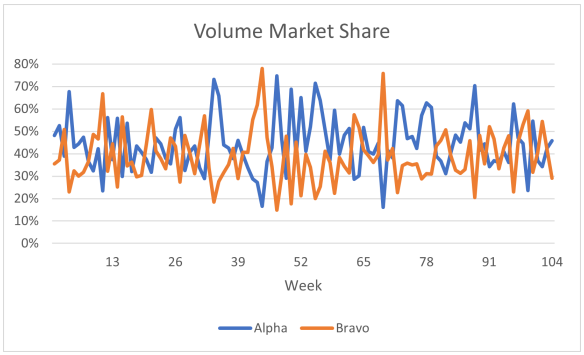

Figure 3.3: Plot of weekly (volume) market share

which is simply the product of units and price. We first enter `spend` in cell H1. Next we enter `=E2*F2` in H2 and copy this formula down to H43557.

- In order to create the weekly summary of brand and category revenue, we insert a pivot table where Rows is week and Columns is brand, and Values is (Sum of) spend. This creates the **Sheet1** worksheet.
- We copy the contents of the pivot table (A5:G108) into a new worksheet, which we will call **Revenue**, leaving the first row empty. We add the following variables names in cells A1:G1: `Week Alpha Bravo Charlie Delta Other Category`.
- Figure 3.1 is simply the (2-D Line) plot of Alpha and Figure 3.2 the plot of Category.

In order to create Figure 3.3, we first need to create a table that reports total weekly *volume* sales for each brand and for the category.

- The weight of each SKU is given in `SKU_weight.csv`. We copy the contents of this file into cells M1:N120 of **edible_grocery**.

- We want to report in column I the weight of the SKU associated with each row. We need to identify the SKU id and extract the associated weight from the table in cells M1:N120. We will make use of the `VLOOKUP` function. We first enter `weight` in cell I1. Next, we enter `=VLOOKUP(D2,$M$2:$N$120,2)` in I2 and copy this formula down to I43557.

- We create a volume variable, which is simply the product of units and weight, expressing the answer in kilograms. We first enter `volume` in cell J1. Next we enter `=E2*I2/1000` in J2 and copy this formula down to J43557.

- Next, we insert a pivot table where Rows is week and Columns is brand, and Values is (Sum of) volume. This creates the **Sheet3** worksheet.

- We copy the contents of the pivot table (A5:G108) into a new worksheet, which we will call **Volume**, leaving the first row empty. We add the following variables names in cells A1:G1: `Week Alpha Bravo Charlie Delta Other Category`.

- Next, we need to compute each brand's volume market in each week. We add the following headings in cells I1:M1: `Alpha Bravo Charlie Delta Other`. Next we enter `=B2/$G2` in cell I2, formatting it as a percentage, and copy this formula across and down to M105.

- Figure 3.3 is simply the plot of Alpha and Bravo.

- There appears to be a high level of competition between these two brands. What is the correlation in their market shares? Assuming the Data Analysis ToolPak is installed, we compute the correlations by going to Data > Data Analysis, selecting the Correlation option and clicking OK. The input range is I1:M105, and we tick the "Labels in first row" option. This creates the **Sheet5** worksheet. We observe that there is a strong negative correlation between the shares of Alpha and Bravo: an increase in one brand's share is associated with a corresponding decrease in the share of the other brand.

Let us finish this preliminary analysis of the data by computing the annual sales of each brand.

- Going back to **edible_grocery**, we want to create a year variable which indicates whether each record is associated with the first or second year of the observed data. We first enter `year` in cell K1. Next we enter `=IF(C2<=52,1,2)` in K2 and copy this formula down to K43557.

- We insert a pivot table where Rows is year and Columns is brand, and Values is (Sum of) spend. This creates the **Sheet6** worksheet.
- We compute each brand's annual *value* (or dollar) market share by entering `=B5/$G5` in cell J5, formatting it as a percentage, and copying this formula across and down to N6. We add the relevant row and column labels.
- We note that Alpha's revenue grew by 5%, even as category revenue dropped by 2%. Alpha's market share grew by 3.3 percentage points, a 7.1% increase in its annual value market share.

Having performed some basic aggregate-level analysis, we now turn our attention to characterising the household-level variation in buyer behaviour in a given time period.

Chapter 4

Exploring Variation in Buyer Behaviour

The objective of this chapter is to explore how to perform the basic analyses that describe the variation we observe in buyer behaviour in a given time period. We will continue to work with the edible grocery dataset used in the previous chapter, exploring both purchase frequency and spend.

4.1 Getting Started

Before we can doing any analysis, we need to create some summary datasets. The first will summarise how many times each panellist purchased each brand as well as in the category. The second will summarise how much each panellist spent on each brand and in the category.

We start by copying the **edible_grocery** worksheet (from the workbook used in the previous chapter) to a new workbook.[1] (Don't forget to save this new workbook, say as `chapter_4.xlsx`)

Let us consider rows 13 to 18:

```
3102021 844   4  5 1  3.49  Alpha
3102021 844   4  5 1  3.49  Alpha
3102021 844   4  5 1  3.49  Alpha
3102021 844   4 15 1  3.49  Alpha
3102021 844   4 15 1  3.49  Alpha
3102021 844   4 89 1  2.49  Delta
```

On this one shopping trip (trans_id = 844), panellist 3102021 purchased a total of six items: three packs of SKU 5, two packs of SKU 15, and one

[1] Why don't we continue working in the same Excel workbook? These spreadsheets can become rather large and (at the time of writing) cumbersome to work with unless you have a computer with a decent amount of memory.

pack of SKU 89. They purchased two different brand, Alpha and Delta.[2]

By convention, this purchase occasion is recorded as one category transaction, one Alpha transaction, and one Delta transaction. The number of units of Alpha purchased is five. When we say that the panellist made one category transaction, we mean they purchased at least one item in the category on that shopping trip. When we say that the panellist made one Alpha transaction, we mean they purchased at least one item associated with the brand on that shopping trip.

In order to analyse buyer behaviour in terms of transactions, we need to know the number of brand and category transactions for each person. We cannot work directly with the dataset we have been using; we effectively need to collapse the five rows associated with Alpha into one. This will require some intermediate analysis, which we undertake in the following manner.

- We insert a pivot table where Rows is trans_id and Columns is brand, Filters is year (selecting just year 1), and Values is (Max of) panel_id. This creates **Sheet2**.
- Our goal is to create a transaction summary table which indicates whether or not each brand was purchased on each transaction. (Obviously a category transaction occurred.) Working in **Sheet2**, we enter the following eight variable names in row 4, starting at I4:
  `trans_id panel_id Alpha Bravo Charlie Delta Other Category`
- The transaction id is in column A, so we enter `=A5` in cell I5.
- The panellist id associated with each transaction is given in columns B–G. We enter `=G5` in cell J5.
- We want to create an indicator of whether or not each brand was purchased on each transaction. A non-zero entry in columns B–F indicates that a transaction did occur. We enter `=1*(B5>0)` in cell K5 and copy the formula across to P5.
- We copy this row of formulas down to row 20034.

We now have a dataset where there is just one row per transaction, with a binary indicator of whether or not each brand was purchased. We can create the desired panellist-level transaction summary in the following manner:

[2]Why are three lines for SKU 5 and not one line with units = 3. This is simply a function of how the items were scanned at the checkout. Some checkout operators will scan the three items separately; this would result in three lines in the transaction file, each with units = 1 . Others will scan the item and press "3" on their till, resulting in one line in the transaction file with units = 3.

- With the active cell somewhere in this new table, we insert a pivot table where Rows is panel_id and Values is (Sum of) Alpha, Bravo, Charlie, Delta, Other, Category. This creates the **Sheet3** worksheet.
- You will see a button at the RHS of cell A3. Click on it and "Sort Smallest to Largest".
- We create a cleaned-up version of this table by copying A4:G4577 to a new worksheet, which we call **panellist x brand (trans)**, and we add the following seven variable names to row 1:
  `panel_id Alpha Bravo Charlie Delta Other Category`

We also need to create a similar table that summarises how much each panellist spent on each brand (and in the category) during year 1. This is much easier to create.

- Going back to **edible_grocery**, we insert a pivot table where Rows is panel_id, Columns is brand, Filters is year (selecting just year 1), and Values is (Sum of) spend. This creates the **Sheet5** worksheet.
- You will see a button at the RHS of cell A4. Click on it and "Sort Smallest to Largest".
- We create a cleaned-up version of this table by copying A5:G4578 to a new worksheet, which we call **panellist x brand (spend)**, and we add the following seven variable names to row 1:
  `panel_id Alpha Bravo Charlie Delta Other Category`

While we are at it, let us create a similar table that summarises each panellist's volume purchasing in year 1.

- Going back to **edible_grocery**, we insert a yet another pivot table where Rows is panel_id, Columns is brand, Filters is year (selecting just year 1), and Values is (Sum of) volume. This creates the **Sheet7** worksheet.
- You will see a button at the RHS of cell A4. Click on it and "Sort Smallest to Largest".
- As above, we create a cleaned-up version of this table in a new worksheet, which we call **panellist x brand (volume)**.

Question

> Looking at **panellist x brand (spend)** and **panellist x brand (volume)**, we note that for each row Category equals the sum of the brand numbers, as we would expect. But this is not always the case in **panellist x brand (trans)**. Why is the sum of the brand-specific numbers sometimes greater than the associated category number?

4.2 Examining Purchase Frequency

Two standard brand performance metrics that summarize purchasing behaviour are penetration and purchases per buyer (PPB). Penetration is the percentage of households buying the product/category at least once in the given time period. Purchases per buyer is the average number of times (*separate shopping trips*) the product/category was purchased (in the given time period) by those households that made at least one product/category purchase (in the given time period).

Looking at **panellist x brand (trans)**, we see that there are 4574 rows in this table, meaning that we have summary data on the purchasing of 4574 households in year 1. But there are 5021 households in the panel. What has happened to the remaining 447 households? They did not make any category purchase during the year. (But they will have purchased in other categories.)

Recall that penetration is the percentage of households buying the product at least once in the given time period. In order to compute this, we need to know the number of panellists buying the product at least once in the time period of interest and the size of the panel during this period. Similarly, recall that purchases per buyer (PPB) is the average number of times the product was purchased (in the given time period) per buyer. This is computed as the total number of purchase occasions on which the product was purchased by the panellists (in the time period of interest) divided by the number of panellists that purchased the product at least once (in the time period of interest).

We compute these quantities in the following manner.

- Staying in **panellist x brand (trans)**, we enter the following labels `Alpha Bravo Charlie Delta Other Category` in cells K1:P1.
- We can compute the number of panellists who purchased each brand at least once in year 1 counting the number of entries in the associated column in the worksheet that are greater than zero. We can determine this using the `COUNTIF` function in Excel. We compute the number of buyers of Alpha by entering `=COUNTIF(B2:B4575,">0")` in cell K2. We see that 2624 households made at least one purchase of Alpha in year 1. Copying the formula across to cell P2 computes this quantity for the other brands and the category.
- On how many purchase occasions did the 2624 buyers of Alpha buy Alpha? We compute this simply by summing up the entries in the column associated with Alpha. We enter `=SUM(B2:B4575)` in cell K3 and copy the formula across to cell P3.
- Penetration is simply the number of brand buyers divided by the number of panellists, expressed as a percentage (via number formatting).

Having entered the number of households in the panel in cell K5, this is computed for Alpha by entering `=K2/$K5` in cell K7 and formatting the cell as a percentage. We copy the formula across to P7.

- Purchases per buyer (PPB) is simply to total number of purchases occasions divided by the number of buyers. We enter `=K3/K2` in cell K8 and copy the formula across to cell P8.

We see that 91% of the households in the panel purchased in the category at least once in year 1. (This is a widely purchased product category.) On average, they purchased in the category 4.4 times that year. Looking at Alpha, we see that 52% of the households in the panel purchased the brand at least once, purchasing it on average 3.5 times.

Penetration and PPB are in fact summary measures of an important but frequently overlooked summary of buyer behaviour: the distribution of purchase frequency. We first explore how to create this summary of category purchasing and then consider how to create such a summary of brand purchasing.

4.2.1 Distribution of Category Purchase Frequency

Our goal is to create Figure 4.1, from which we see that 9% of the panellists made no category purchases, 13% of the panellists made one category purchase, ..., and 1% of the panellists made at least 15 category purchases in year 1.

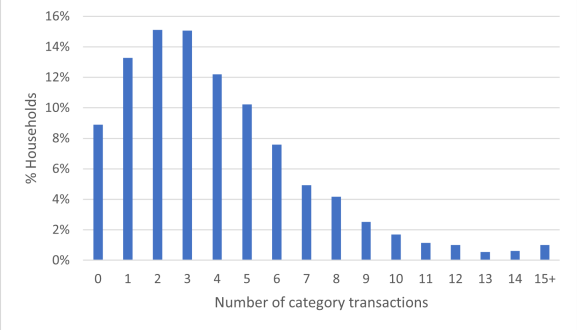

Figure 4.1: Distribution of category purchasing in year 1

Looking at the **panellist x brand (trans)** worksheet, the distribution of category purchasing is determined by counting how many households made one category purchase (panellists 3102011, 3102046, etc.), two category purchases (panellists 3102012, 3102021, etc.), and so on. We do this in the following manner.

- With the active cell somewhere in the main table, we insert a pivot

table where Rows is Category and Values is (Count of) panel_id. We rename the worksheet **Trans distribution – Category**

- In order to have a complete summary of category purchasing, we need to account for those households that made no category purchases the year. We copy the frequency bins observed in the pivot table to column D and add a 0 bin. The associated frequencies for 1, 2, 3, ... category purchases are copied from the pivot table to column E. The frequency of zero category purchases is simply the total number of panellists (5021) minus the number of panellists that made 1, 2, 3, ... category purchases. We enter `=5021-SUM(E4:E26)` in cell E3.

- In column F, we compute the proportion of panellists that made 0, 1, 2, ... category purchases. We enter `=E3/SUM($E$3:$E$26)` in cell F3, format the cell as a percentage, and copy the formula down to F26. (We immediately see that penetration is simply 100% minus the percentage of households making zero category purchases.)

If we wish to create a visual representation of this distribution, it is tempting to simply plot the data in column F. However, the resulting plot would be misleading as some purchase frequencies are missing in the data. In particular, we see that no one made 21 category purchases; ditto for 23, 24, and 26. One solution is to insert manually the missing number of purchases levels with 0 frequencies and then plot the data. However, the observed (relative) frequencies in the right tail are so small that they do not show up in a plot. We can therefore create a *right-censored* distribution. We do so in columns H–J. Here we have chosen 15 as the (right) censoring point[3]; we see that 51 households (or 1% of the panellists) made 15 or more category purchases in year 1.

[Optional] The Relationship Between the Mean and PPB

What is the average number of times the category was purchased by a household in year 1?

- Recall from your introductory statistics course that the mean, which we denote by $E(X)$, is given by

$$E(X) = \sum \frac{x f_x}{n},$$

where f_x is the frequency with which x occurs in the dataset, n is the sample size, and the summation is over all possible values of x.

[3]There is nothing magical about our choice of 15. Generally, the choice of censoring point is a function of how many bins you wish to display and the height of the right-most bar with which you feel comfortable.

- Denoting the relative frequency with which x occurs (i.e., f_x/n) by $P(X = x)$,[4] the mean is given by

$$E(X) = \sum xP(X = x)\,.$$

- We compute this in cell I20 using the `SUMPRODUCT` function: `=SUMPRODUCT(D3:D26,F3:F26)`.

We see that average number of category purchases is 4.0. Why is this different from the 4.4 purchases per buyer computed above? The mean we have just computed includes those households that made zero purchases, whereas PPB is the average among those households that made at least one (in this case category) purchase.

We can derive the relationship between these two quantities in the following manner:

$$\begin{aligned}
\text{PPB} &= \sum_{x=1}^{\max x} \frac{xf_x}{n - f_0} \\
&= \sum_{x=0}^{\max x} \frac{xf_x}{n - f_0} \\
&= \sum_{x=0}^{\max x} \left(\frac{xf_x}{n - f_0}\right)\left(\frac{n}{n}\right) \\
&= \sum_{x=0}^{\max x} \left(\frac{xf_x}{n}\right)\left(\frac{n}{n - f_0}\right) \\
&= \left(\frac{n}{n - f_0}\right) \sum_{x=0}^{\max x} \left(\frac{xf_x}{n}\right) \\
&= \frac{E(X)}{1 - P(X = 0)}\,.
\end{aligned}$$

In other words, PPB is the mean divided by penetration. Computing this in cell I21 (`=I20/(1-F3)`) gives us 4.4.

Note that while the right-censored distribution (H2:J18) is useful when creating a summary figure or table, it is of limited value beyond that. For example, it is not possible to compute the mean purchase frequency (and therefore PPB) from this summary.

4.2.2 Distribution of Brand Puchase Frequency

Let us now create the distribution of purchase frequency for Alpha, which is given in Figure 4.2. With the one exception noted below, the logic follows that associated with creating the distribution of category purchasing.

[4] More formally, we can say that $P(X = x)$ is the empirical probability that a randomly chosen household made x purchases.

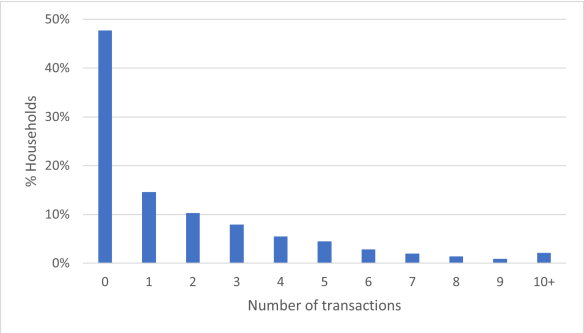

Figure 4.2: Distribution of purchase frequency for Alpha in year 1

- With the active cell somewhere in the main table in **panellist x brand (trans)**, we insert a pivot table where Rows is Alpha and Values is (Count of) panel_id. We rename the worksheet **Trans distribution – Alpha**
- We copy cells A4:B22 to D4:E22.
- In contrast to the pivot table output associated with our summary of category purchasing, this pivot table does contain a zero category. However, we must be careful in our interpretation of the associated frequency. We see that 1950 category buyers did not buy Alpha in year 1. However, in order to have a complete summary of brand purchasing, we should also account for those 447 households that made no category purchases that year. The number of panellists making zero purchases of Alpha is the total number of panellists (5021) minus the number of panellists that made 1, 2, 3, ... purchases. We enter `=5021-SUM(E5:E22)` in cell E4.[5]
- In column F, we compute the proportion of panellists that made 0, 1, 2, ... Alpha purchases, expressed as a percentage (via number formatting).
- When plotting the data, we choose to right-censor the distribution at 10 — see columns H–J and the associated plot.

4.3 Exploring the Distribution of Spend

We now turn our attention to creating summaries of total spend. Our initial goal is to create Figure 4.3, which is a histogram of category spend (in dollars) across those panellists that made at least one purchase in the category

[5] What is Alpha's penetration of category buyers? $100 \times (1 - 1950/4574) = 57\%$.

in year 1. In this plot, the raw total spend data have been binned in bins with a width of \$5. We see that 16% of category buyers spent up to \$5 in the category during this one-year period; 23% spent between \$5 and \$10; ... and 2% spent more than \$50.

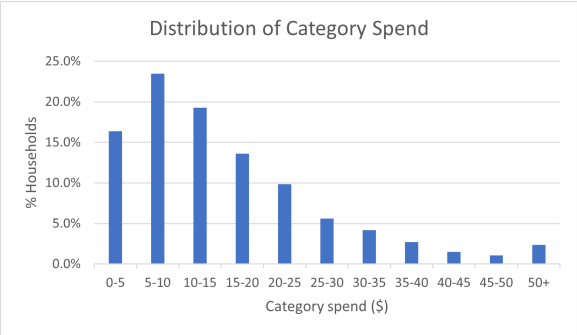

Figure 4.3: Distribution of category spend across category buyers

We create this plot in the following manner.

- We copy the panel_id and Category columns from **panellist x brand (spend)** into a new worksheet, which we rename **Spend distribution – Category**.
- Before deciding on what bin width to use when creating the histogram, let us first compute some basic descriptive statistics.
  - The minimum total spend is computed in cell E2:
    `=MIN(B2:B4575)`
  - The maximum total spend is computed in cell E3:
    `=MAX(B2:B4575)`
  - The average total spend is computed in cell E4:
    `=AVERAGE(B2:B4575)`
  - The median total spend is computed in cell E5:
    `=MEDIAN(B2:B4575)`
- There appears to be quite a bit of variability in category spend. To get further insight into the distribution of total category spend across the panellists, we use Excel's percentile function to determine the total spend quantities associate with the various percentile levels.
  - We enter 0.05 in cell D8, `=D8+0.05` in cell D9, and copy this formula down to D26. Cells D8:D26 are then formatted as percentages.

  - To quote Microsoft documentation, the `PERCENTILE.INC` function "returns the k-th percentile of values in a range." We enter `=PERCENTILE.INC($B$2:$B$4575,D8)` in cell E8 and copy this formula down to E26.

- We see that 5% of the category buyers spent \$2.69 or less in the category during the year, 10% spent \$3.39 or less, and so on. The heaviest 5% of buyers each spent more than \$39.72 in the category during the year.

- Looking out this output, we conclude that a bin size of \$5 is probably about right.

- How many bins do we go with? This is an empirical question. Since 5% of the panel spent more than \$39.72, we certainly want to go above \$40 in order to get a sense of how they are spread between \$39.72 and the maximum of \$166.70. We will go with \$50. If too many panellists have spent more than \$50, we can always add more bins. If too few panellists fall into this bin, we can always combine the bins we have created.

- We will use Excel Histogram data analysis tool to determine how many panellists' total spend fell into each bin with a width of \$5.

  - We first specify the range of histogram bins we wish to use. We enter 5 in cell G8, `=G8+5` in cell G9, and copy this formula down to G17. This gives us 10 bins with a width of \$5.
  - Clicking on Data > Data Analysis > Histogram, we specify cells B2:B4575 as the input range, cells G8:G17 as the bin range, and select I7 as the output range.
  - The resulting histogram is given in I7:J18. We see that 750 panellists spent \$5 or less in the category during year 1, 1073 spent between \$5 and \$10, ..., and 109 spent more than \$50.
  - We convert the raw counts into percentages and plot these percentages as a bar chart.

The general shape of this distribution (i.e., an interior mode, median less than the mean, right-skewed with a long right-tail) is what we typically observe when we look at spend data.[6]

Let us now create the distribution of spend on Alpha. We will follow the same basic process as for the distribution of category spend with a few minor changes.

[6]This, of course, depends on the bin width we choose when summarising the data. If we had chosen a bin width of \$10, the left-most bar would be the highest bar and we would no longer observe an interior mode.

- We copy the panel_id and Alpha columns from **panellist x brand (spend)** into a new worksheet, which we rename **Spend distribution – Alpha**.
- We immediately notice that there are a number of empy cells in the Alpha column. While these panellists purchased in the category during the year, they did not purchase any of Alpha's SKUs. The first thing we need to do is remove these observations.
- One way to do this is to first sort this table on Alpha from "Largest to Smallest". The last entry for Alpha occurs in cell B2625. We delete the contents of A2626:A4575 (1950 cells), giving us a table that contains only those panellists that purchased at least one Alpha SKU in year 1.
- Having looked at some basic descriptive stats (as above), we will summarise the data using bins of width \$2, right censoring at \$40.
- Following the process used in our analysis of category spend, we arrive at Figure 4.4. (There's room to improve the formatting of the x-axis labels.)

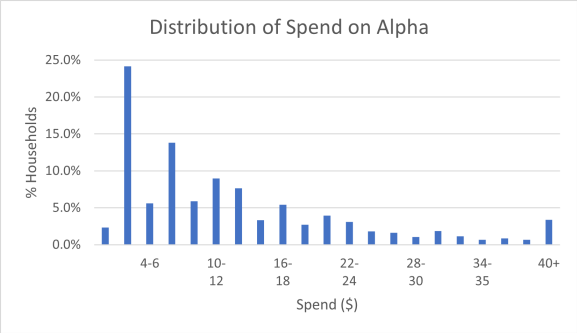

Figure 4.4: Distribution of spend on Alpha across brand buyers

The general observation made above about the shape of the distribution of spend holds. It is not so smooth, but this is a function of the smaller bins sizes. There is, however, one obvious aberration — the large spike for \$2–\$4. What is going on here? The average price of most Alpha SKUs is between \$2 and \$4.[7] We recall from Figure 4.2 that 48% of the panel made zero purchases of Alpha and 15% made one transaction. This implies that $0.15/(1 - 0.48) = 29\%$ of Alpha buyers made just one purchase of the brand. Assuming they only purchased only one unit of one Alpha SKU on

[7] How can we work this out? Going back to **edible_grocery**, we insert a pivot table where Rows is sku_id, Columns is brand, Filters is year (selecting just year 1), and Values is (Average of) price — see **Average SKU price**.

that purchase occasion,[8] we would expect a large number of Alpha buyers to spend between \$2 and \$4; 24% is not unrealistic.

Exercise

Create the distribution of volume purchased for both the category and Alpha.

4.4 Performing a Basic Decile Analysis

We have observed that there is a lot of variability in purchase frequency and spend, be it at the product or category level. A useful way of expressing the variability is via a decile analysis. As the name suggests, this sees us dividing the customer base into 10 equally sized groups and then summarising each group's buying behaviour. We will consider two versions of this analysis.

We will first focus on creating Table 4.1. Recall that 4574 households made at least one category purchase in year 1. Having sorted these households by total category spend, we create 10 equally sized groups. Decile 1 is the 10% of households that spent the most in the category during year 1, decile 2 is the next largest 10% of spenders, and so on.

Decile	% HHs	% Spend	% Trans	Spend/HH	Cat trans/HH	Avg spend/trans	# unique brands
1	10%	28%	24%	\$44.65	10.6	\$4.21	1.8
2	10%	17%	16%	\$27.15	7.1	\$3.80	1.6
3	10%	13%	13%	\$20.96	5.9	\$3.56	1.6
4	10%	11%	11%	\$16.91	4.9	\$3.47	1.5
5	10%	9%	9%	\$13.81	4.2	\$3.33	1.5
6	10%	7%	8%	\$11.14	3.4	\$3.27	1.4
7	10%	6%	7%	\$8.98	2.9	\$3.05	1.4
8	10%	4%	5%	\$6.74	2.1	\$3.16	1.2
9	10%	3%	4%	\$4.56	1.6	\$2.82	1.2
10	10%	2%	2%	\$2.70	1.0	\$2.65	1.0

Table 4.1: Decile analysis of category buying behaviour (where each decile equals 10% category buyers)

Reading across the row associated with Decile 1, we see that they accounted for 28% of category spend and 24% of total transactions. On average, they spent \$44.65 in the category across an average of 10.6 purchase occasions. This corresponds to an average category spend per category transaction of \$4.21. On average, these households purchased 1.8 different brands during the year. Contrast this to the 10% of category buyers that spent the least in the category. They account for 2% of category spend and transactions, making on average one category purchase. And so on.

We create this table in the following manner.

[8] How would you determine the validity of this assumption? This is left as an exercise for the interested reader.

- We insert a new worksheet, which we call **Decile data (i)**, and copy the panel_id and Category columns from **panellist x brand (spend)** into columns A and B. We rename column B spend.
- We copy the Category column from **panellist x brand (trans)** into column C of **Decile data (i)** and rename this column trans.
- We would like to count the number of unique brands purchased by each panellist in year 1.
  - We enter `# unique brands` in cell D1.
  - The number of unique brands purchased by each customer is simply the number of brands in **panellist x brand (trans)** that have a non-zero number of transactions.
  - We perform this calculation by entering `=COUNTIF('panellist x brand (trans)'!B2:F2,">0")` in cell D2.[9]
  - We copy this formula down to D4575.
  - Finally, we copy cells D2:D4575 and "paste values" onto these same cells. (We need to do this as we will sort the table at a later stage and keeping the formula "live" would introduce errors.)
- Before we go any further, let's make a copy of this worksheet, calling it **Decile data (ii)**.
- Returning to **Decile data (i)**, the next step is to determine the decile bin into which each panellist falls. There are several ways to do this. Here's one approach:
  - We sort the table of data by spend from "Largest to Smallest".
  - Next we create a rank variable. Having entered `rank` in cell E1, we enter 1 in cell E2, `=E2+1` in E3, and copy the formula down to E4575.
  - We convert the rank number into a decile number by entering `=INT(10*(E2-1)/E$4575)+1` in F2 and copying the formula down to F4575. (Make sure you understand the logic of the formula. Why is -1 in the formula?)
  - We enter `decile` in cell E1.

[9] A note to the novice Excel user: When entering this formula, we do not actually type `=COUNTIF('panellist x brand (trans)'!B2:F2,">0")`. Having typed `=COUNTIF(` in cell D2 (but not pressing <Enter>.), we then click on the worksheet **panellist x brand (trans)**, highlight cells B2:F2, and continue typing in the formula before pressing <Enter>.

- We insert a pivot table where Rows is decile and Values is (Count of) panel_id, (Sum of) spend, (Sum of) trans, and (Average of) # unique brands. This creates the **Sheet16** worksheet.

- We create the various columns of Table 4.1 in the following manner. For decile 1, we enter the following formulas:

  cell H4 `% HHs` is `=B4/B$14` (formatted as a percentage)
  cell I4 `% Spend` is `=C4/C$14` (formatted as a percentage)
  cell J4 `% Trans` is `=D4/D$14` (formatted as a percentage)
  cell K4 `spend/HH` is `=C4/B4`
  cell L4 `Cat trans/HH` is `=D4/B4`
  cell M4 `Avg spend/trans` is `=C4/D4`
  cell N4 `# unique brands` is simply a copy of relevant entry in column E, `=E4`

- We copy this row of formulas down to row 13.

This is a very basic decile table. Some additional information that could be reported includes the average number of units purchased per transaction and the average number of unique SKUs purchased in the year. We leave this as an exercise for the interested reader.

The decile analysis we have just presented uses deciles that represent 10% of the category buyers. An alternative approach is to create deciles that represent 10% of category spend. The only change to what we have done above is how we create the decile variable.

- Turning to **Decile data (ii)**, we sort the table of data by spend from "Largest to Smallest".

- Next we create a cumulative spend column. We enter `Cum spend` in cell E1, `=B2` in E2, `=E2+B3` in E3, and copy the formula down to E4575.

- The next step is to determine the cumulative total spend numbers that correspond to the boundaries between each decile.

  - We first need to know the value of overall category purchasing by these panellists. This lurks in cell E4575. So as to make our analysis a little more transparent, we copy this value to cell J1 (using the formula `=E4575`) and enter `Total category spend` in cell I1).
  - One-tenth is this is computed by entering `=J1/10` in cell J2; we label this `Size of 10th equal split` (cell I2).
  - We compute the cutoffs between the first and second, second and third, ..., and ninth and tenth deciles by entering `=J2` in cell J5, `=J5+$J$2` in cell J6, and copying this formula down to cell J13.

We then enter the decile labels 1–9 in cells I5:I13. We label these columns by entering `Decile` in cell I4 and `Cutoff` in cell J4.

- We now assign each panellist to a spend decile by comparing the cumulative total spend number associated with their row in the table to the various decile cutoff values. While we could use a series of nested `IF` functions, it is cleaner to use the `IFS` function. We enter `=IFS(E2<=$J$5,1,E2<=$J$6,2,E2<=$J$7,3,E2<=$J$8,4,E2<=$J$9,5,E2<=$J$10,6,E2<=$J$11,7,E2<=$J$12,8,E2<=$J$13,9,E2>$J$13,10)` in cell F2, copy the formula down to F4575, and label this column `decile` (cell F1).

We then create the decile table as above, giving us Table 4.2

Decile	% HHs	% Spend	% Trans	Spend/HH	Cat trans/HH	Avg spend/trans	# unique brands
1	2%	10%	8%	$63.77	13.4	$4.74	1.9
2	4%	10%	9%	$42.33	10.5	$4.03	1.8
3	5%	10%	9%	$33.97	8.7	$3.89	1.8
4	6%	10%	9%	$28.21	7.4	$3.83	1.6
5	7%	10%	10%	$23.69	6.5	$3.65	1.6
6	8%	10%	10%	$20.01	5.7	$3.53	1.6
7	9%	10%	10%	$16.66	4.8	$3.47	1.5
8	12%	10%	11%	$13.49	4.1	$3.31	1.5
9	16%	10%	11%	$10.06	3.1	$3.21	1.4
10	32%	10%	12%	$4.89	1.7	$2.93	1.2

Table 4.2: Decile analysis of category buying behaviour (where each decile equals 10% of category spend/revenue)

Reading across the row associated with Decile 1, we see that the top 2% of households accounted for 10% of category spend. On average, they spent $63.77 in the category, across an average of 13.4 purchase occasions. This corresponds to an average category spend per category transaction of $4.74. On average, these households purchased 1.9 different brands during the year.

4.5 [Optional] Creating Lorenz Curves

A Lorenz curve is a common graphical tool for visualising "concentration" or "inequality" in the distribution of a quantity of interest (e.g., income, buying behaviour). It show the proportion of the overall quantity (e.g., number of transactions, spend) associated with the bottom x% of the unit of observation associated with the distribution (e.g., households). When analysing buyer behaviour, this sees us lining up all customers in ascending order of their level of purchasing and computing the share of total purchasing accounted by each person. The Lorenz curve is created by plotting the cumulative percentage of customers (x-axis) against the cumulative percentage of total purchasing (y-axis).

The Lorenz curve for transactions associated with Alpha is given in Figure 4.5 and interpreted in the following manner. We see that that x = 80% roughly corresponds to y = 54%, which means the 80% of the buyers of Alpha (when sorted from least to most frequent buyers) account for 54% of all the buying of Alpha in year 1. This implies the top 20% of buyers (in terms of purchase frequency) account for 46% of total purchases. The "rule" of 80/20 does not hold here; rather, it is 46/20.

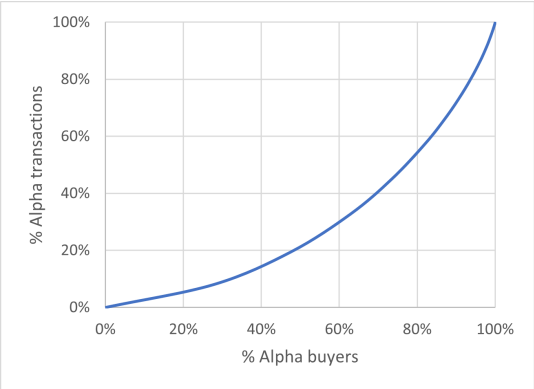

Figure 4.5: Lorenz curve for Alpha transactions

We create this Lorenz curve in the following manner:

- We make a copy of **Trans distribution – Alpha**, renaming it **Lorenz – Alpha (trans)**, and delete all the content right of column C.
- We are only interested in those panellists that made at least one purchase of Alpha in year 1. We copy the numbers in cells A5:B22 of the pivot table to D5:E22.
- We compute in cell F5 the total number of transactions made by those customers who made one purchase of Alpha; this is obviously `=D5*E5`. We copy this formula down to F22. With reference to cell F6, we see that the 517 households that purchased Alpha twice made a total of 1034 transactions.
- What percentage of the total number of buyers of Alpha are the 733 households that made one purchase? The 517 households that made two purchases? What percentage of total purchasing is the 733 purchases made by those households that made one purchase? The 1034 purchases by made by those households that made two purchases? We first compute the total number of Alpha buyers (`=SUM(E5:E22)` in cell E23) and the total amount of purchasing by this group (`=SUM(F5:F22)` in cell F23), and then compute the associated percentages in columns G and H (e.g., `=E5/E$23` in cell G5 and `=F5/F$23` in cell H5). We see

that 28% of Alpha buyers made just one purchase and their purchasing accounted for 8% of all Alpha purchasing. Similarly, we see that 20% of buyers made two purchases and their purchasing accounted for 11% of all Alpha purchasing.

- After entering zeros in cells I4 and J4, we compute the cumulative percentage of both quantities in columns I and J (e.g., `=I4+G5` in cell I5 and `=J4+H5` in cell J5). We see, for example, that 48% of all buyers made two or fewer purchases and accounted for 20% of all the purchases of Alpha.

- We create the Lorenz curve by plotting these data using chart type "Scatter with Straight Lines", setting the maximum of both axes to 1, and manually adjusting the shape so that plot is (approximately) square.

As noted above, we see that x = 80% roughly corresponds to y = 54%. We can compute the exact number by interpolation in the following manner:

- We see that 73.4% of buyers accounted for 45.0% of purchases, and that 82.1% of buyers accounted for 57.5% of purchases.

- Entering `=(0.8-I8)/(I9-I8)` in cell L8, we see that 80% falls 76% of the way between 73.4% and 82.1%. Therefore, the associated percentage of total purchases will lie 76% of the way between 45.0% and 57.5%. Entering `=J8+L8*(J9-J8)` in cell M8 (formatted as a percentage), we see this is 54.5%.

- As the "bottom" 80% account for 54% of total purchasing, it follows that the "top" 20% account for 46% of total purchasing.

A related quantity of interest is the percentage of buyers that account for half of total purchasing. This can be read off the Lorenz curve in the following manner.

- We see that 73.4% of purchasers account for 45.0% of total purchases, and that 82.1% of purchasers account for 57.5% of total purchases.

- Entering`=(0.5-J8)/(J9-J8)`in cell L9, we see that 50% lies 40% of the way between 44.0% and 57.5%. Therefore, the associated percentage of purchasers lies 40% of the way between 73.4% and 82.1%. Entering `=I8+L9*(I9-I8)` in cell M9 (formatted as a percentage), we see this is 76.9%.

- As the "bottom" 77% of buyers account for half of total purchasing, it follows that the "top" 23% also account for half of total purchasing, which we can write as 50/23. This quantity is easy for most people grasp.

Purchase frequency is a discrete quantity and we created the Lorenz curve off the distribution of transactions. We now consider how to create a Lorenz curve when the quantity of interest in continuous. We will focus on creating the Lorenz curve for the spend associated with Alpha (Figure 4.6).

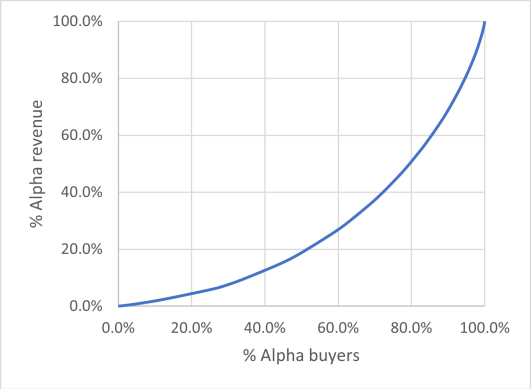

Figure 4.6: Lorenz curve for Alpha spend

- We start by making a copy of **Spend distribution – Alpha**, renaming it **Lorenz – Alpha (spend)** and delete all the content right of column C. We then sort this table by Alpha, from "Smallest to Largest" and insert a row above row 2. (Row 2 is now empty.)

- We start by calculating the percentage of total Alpha purchasing accounted for by each panellist. We first compute in cell B2627 the total spend on Alpha (`=SUM(B3:B2626)`). Next we enter `=B3/$B$2627` in cell C3, formatting the result as a percentage, and copy the formula down to cell C2626.

- Next, we compute the cumulative percentage of buyers number. The total of Alpha buyers is determined by entering `=COUNT(A3:A2626)` in cell A2627. We enter `=(ROW()-2)/$A$2627` in cell D2, format the result as a percentage, and copy the formula down to cell D2626.

- In order to compute the cumulative percentage of spend numbers, we first enter 0 in cell E2 and format the cell as a percentage. Next we enter `=E2+C3` in cell E3, formatting the result as a percentage, and copy the formula down to cell E2626.

Looking at row 2100, we see that the "bottom" 80% of Alpha buyers account for 50.8% of its revenue in year 1, which is equivalent to 49/20. This is higher than the 46/20 we observed for transactions.

We create the Lorenz curve by plotting the data in cells D2:E2626 using chart type "Scatter with Straight Lines", setting the maximum of both

axes to 1, and manually adjusting the shape so that plot is (approximately) square.

Chapter 5

Exploring Multibrand Buying Behaviour

So far, we have explored purchasing at the level of the individual brand or overall category. We now step back and consider panellists' purchasing of multiple brands (in a given time period).

To set the scene, let us first determine the percentage of category buyers that bought 1, 2, 3, 4, or 5 different brands in the course of year 1. (Note that we only have five brands in the dataset, with Other being an aggregation of several very small share brands). We want to create Table 5.1.

# Brands	1	2	3	4
% Category buyers	65%	28%	6%	1%

Table 5.1: Distribution of the number of separate brands purchased by category buyers in year 1

- We start by copying the **panellist x brand (trans)** worksheet from the workbook used in the previous chapter and save it in a new workbook (say `chapter_5.xlsx`). We delete the content to the right of column H.
- We insert a new worksheet (**# brands**) and copy column A from **panellist x brand (trans)**.
- We compute the number of brands purchased by each household by counting the number of non-zero entries in each row of the **panellist x brand (trans)** table. We enter `=COUNTIF('panellist x brand (trans)'!B2:F2,">0")` in cell B2 and copy the formula down to B4575. We label column this column `# brands` (cell B1).
- Inserting a pivot table where Rows is # brands and Values is (Count of) panel_id gives us the distribution of the number of brands pur-

chased in year 1 by those panellists that made at least one category purchase that year. Expressing these counts as percentages gives us Table 5.1.

We note that over two-thirds of the category buyers only ever bought one brand. This is despite the fact that 86% of category buyers made two or more category purchases in the course of the year.[1] No household purchased all five brands that year.

How does the number of different brands purchased in the year vary as a function of the number of category purchases made during the year?

- We make a copy of the worksheet **# brands** and rename it **# brands x cat purchasing**. In column C we report the number of category purchases made by each household; we simply copy column G from **panellist x brand (trans)**.
- We insert a pivot table where Rows is Category, Columns is # brands, and Values is (Count of) panel_id.

Looking at the resulting pivot table output (**Sheet5**), we see that there quite a high level of sole-brand loyalty (i.e., only buying one brand), even as the number of category purchases increases. We have two panellists that made 20 category purchases, all with the same brand. That's some level of loyalty!

We can compute the average number of different brands purchased for each level of category purchasing by entering `=SUMPRODUCT($B$4:$E$4, B5:E5)/F5` in cell H5 and copying the formula down to H27. We note that the average number of brands purchased does increase as a function of category purchasing. This should not come as a surprise, as more category purchases equals more opportunities to buy different brands. We saw in Chapter 3 that two brands (Alpha and Bravo) had a combined value market share of 86%. As such, the relatively low number of different brands purchased in the year is not too surprising.

We now consider three common analyses designed to give insight into the nature of multibrand buying behaviour.

Warning: The rest of the chapter assumes basic familiarity with matrix multiplication.

[1] We know from Figure 4.1 that 8.9% of the panel made zero category purchases and 13.2% made one category purchase. $1 - 0.132/(1 - 0.089) = 0.855$, which means 86% of year 1 category buyers made more than one category purchase that year.

5.1 Duplication of Purchase

We know from our analysis of penetration and PPB in Chapter 4 that 52% of households purchased Alpha and 51% of households purchased Bravo. What percentage of Alpha buyers also purchased Bravo during the year (and vice-versa)? The answer to these (and similar) questions is provided by a duplication of purchase analysis.

Before performing this analysis on our dataset, let us consider a toy example. The following table summarizes the purchasing of four brands by six households.

	A	B	C	D
HH01	1	0	2	0
HH02	0	1	0	0
HH03	1	3	0	0
HH04	0	0	1	4
HH05	1	1	0	1
HH06	0	0	0	1

We see that three households made at least one purchase of brand A, three households made at least one purchase of brand B, and so on. How many brand A buyers also purchased brand B? Two (HH03 and HH05). How many brand A buyers also purchased brand C? One (HH01). Repeating this for all brands gives us the following table, which we will call a duplication count table.

	A	B	C	D
A	3	2	1	1
B	2	3	0	1
C	1	0	2	1
D	1	1	1	3

For any given row, the number in each cell is the number of buyers of the brand associated with that row that also purchased the brand associated with that column. (The diagonal is obviously the number of buyers of each brand.) Looking at the row for brand A, we see that three households purchased that brand. Two of these households (67%) also made at least one purchase of brand B, and one of these three households (33%) also made at least one purchase of brand C. These row percentages are reported in the following table, which we call the duplication of purchase table. (By convention, we leave the diagonal blank.)

	A	B	C	D
A		67%	33%	33%
B	67%		0%	33%
C	50%	0%		50%
D	33%	33%	33%	

How can we create this table efficiently when we have a large number of panellists and it is therefore not practical to do so by hand? One approach, which makes use of matrix multiplication, is as follows:

- We create that we will call an "ever buyers" matrix, which is a matrix of size (number of panellists) $\times$ (number of brands), where each cell takes on a value of 1 if the panellist (row) ever purchased the brand (column) in the period of interest; 0 otherwise:

	A	B	C	D
HH01	1	0	2	0
HH02	0	1	0	0
HH03	1	3	0	0
HH04	0	0	1	4
HH05	1	1	0	1
HH06	0	0	0	1

→

"Ever buyers"

	A	B	C	D
HH01	1	0	1	0
HH02	0	1	0	0
HH03	1	1	0	0
HH04	0	0	1	1
HH05	1	1	0	1
HH06	0	0	0	1

- Pre-multiplying the "ever buyers" matrix by its own transpose

	HH01	HH02	HH03	HH04	HH05	HH06
A	1	0	1	0	1	0
B	0	1	1	0	1	0
C	1	0	0	1	0	0
D	0	0	0	1	1	1

A	B	C	D	
1	0	1	0	HH01
0	1	0	0	HH02
1	1	0	0	HH03
0	0	1	1	HH04
1	1	0	1	HH05
0	0	0	1	HH06

gives us the duplication count table we created above. (You should verify this for yourself.) Dividing each cell by the number of buyers of the brand associated with each row gives us the duplication of purchase table.

The duplication of purchase table for year 1 is given in Table 5.2. We see that 34% of those panellists that purchased Alpha in year 1 also made at least one purchase of Bravo that year. We see that 15% of Alpha buyers also bought Charlie, whereas 50% of Charlie buyers also bought Alpha. This asymmetry is not surprising given the relative size of the two brands.

This table is created in the following manner:

- The first thing we need to do is to create the "ever buyers" matrix. We insert a new worksheet called **Ever buyers**. We copy in both column A and cells B1:F1 from **panellist x brand (trans)**. We want to populate this table with zeros and ones, depending on whether or not each household made at least one purchase of the brand in question. We enter `=1*('panellist x brand (trans)'!B2>0)` in cell B2, and copy the formula across and down to cell F4575.

		who also purchased				
		Alpha	Bravo	Charlie	Delta	Other
% people who purchased	Alpha		34%	15%	9%	3%
	Bravo	35%		15%	5%	4%
	Charlie	50%	47%		14%	3%
	Delta	63%	37%	31%		3%
	Other	39%	63%	15%	6%	

Table 5.2: Duplication of purchase table for year 1

- We insert a new worksheet and rename it **Duplication of purchase**. The first thing we do is compute the total number of buyers of each brand. (These numbers are simply the column totals of the "ever buyers" matrix.) We enter the brand names in cells C1:G1. Next we enter `=SUM('Ever buyers'!B2:B4575)` in cell C2 and copy the formula across to cell G2.

- The duplication count table is created by pre-multiplying the "ever buyers" matrix by its own transpose. Excel has a series of functions for basic matrix operations, including transpose and matrix multiplication. Having entered the brand names in cells C5:G5 and B6:B10, we type `=MMULT(TRANSPOSE('Ever buyers'!B2:F4575),'Ever buyers'!B2:F4575)` in cell C6. At the time of writing, using the latest version of Office 365, the rest of the table is automatically filled when <Enter> is pressed. (If you are using an older version, the cell should display #VALUE! when you press <Enter>. Copy this formula across and down to cell G10. (Cells C6:G10 should all display #VALUE!.) With cells C6:G10 highlighted, press <F2> and then <Ctrl> + <Shift> + <Enter>.)

- To make the final step of the calculation easy, we want the totals given in cells C2:G2 in cells I6:I10. We do this using the Excel's transpose function. We type `=TRANSPOSE(C2:G2)` in cell I6. At the time of writing, using the latest version of Office 365, cells I6:I10 are automatically filled when <Enter> is pressed. (If using an older version of Excel , you will need to copy the formula down to cell I10. With cells I6:I10 highlighted, press <F2> and then <Ctrl> + <Shift> + <Enter>.)

- The duplication of purchase table is created by dividing each entry of the duplication count table (cells C6:G10) by the total number of buyers of the brand associated with each row (cells I6:I10). We first enter the brand names in cells C13:G13 and B14:B18. Next we enter `=IF($B14=C$5,"",C6/$I6)` in cell C14, format as a percentage, and copy the formula across and down to cell G18.

5.2 Share of Category Requirements

The duplication of purchase table tells use that 34% of Alpha buyers also purchased Bravo, 15% also purchased Charlie, and so on. This "polygamous purchasing" leads to an obvious question: How "loyal" are they to Alpha? The answer to this question obviously depends on what we mean by "loyal". One commonly used measure of loyalty is *share of category requirements* (SCR), which is the percentage of category volume that the brand represents among its buyers.

The SCR numbers for each brand are reported in Table 5.3. We see that Alpha has a share of category requirements of 69%. This means that 69% of the total category volume purchased by buyers of Alpha goes to Alpha. Contrast this to Delta, which has an SCR of 40%. This means that 40% of the volume purchased in the category by buyers of Delta goes to that brand.

Alpha	Brave	Charlie	Delta	Other
69%	68%	45%	40%	29%

Table 5.3: Share of category requirements for year 1

We compute these numbers in the following manner:

- We first copy the **panellist x brand (volume)** worksheet from the workbook used in the previous chapter into the workbook we are using for this chapter's analysis.
- We should first check that the rows in this worksheet correspond to rows in **Ever buyers**. We can check this by entering `=1-(A2='Ever buyers'!A2)` in cell H2 of **panellist x brand (volume)** and copying the formula down to H4575. If the same panellist is associated with each row in both worksheet, which is what we want, the sum of this column is 0.
- Inserting a new worksheet and renaming it **SCR**, we enter the five brand names in cells B2:F2.
- We compute in row 3 the total purchasing of each brand, which is simply the sum of the relevant column in **panellist x brand (volume)**. We enter `=SUM('panellist x brand (volume)'!B2:B4575)` in cell B3 and copy the formula across to cell F3.
- The next step is to compute the total amount of category purchasing conditioned on the fact that at least one purchase of the brand of interest was made. Recall that the "ever buyers" matrix (**Ever buyers**) indicates whether or not each panellist ever purchased each brand. For any given brand, we want to sum up total category purchasing (column G from **panellist x brand (volume)**) across those

panellists for whom the ever-buy indicator is 1. We can do this by multiplying each cell in the relevant column of the "ever buyers" matrix by the associated panellist's cell in the Category column in **panellist x brand (volume)** and then summing across panellists. We enter `=SUMPRODUCT('Ever buyers'!B2:B4575,'panellist x brand (volume)'!$G$2:$G$4575)` in cell B4 and copy the formula across to cell F4.

- SCR is the ratio of brand purchasing to category purchasing. We enter `=B3/B4` in cell B5, format the result as a percentage, and copy the formula across to cell F5.

5.3 Cross Purchasing

We have just seen that 69% of total category volume purchasing by the buyers of Alpha went to that brand. We know from the duplication of purchase analysis that 15% of Alpha buyers also purchased Charlie. How much of their category volume purchasing went to Charlie? This is answered via cross purchase analysis (sometimes called a combination purchase analysis).

To illustrate the logic of the associated calculations, let us revisit the toy problem introduced in Section 5.1. We have the following summary of the purchasing of four brands by six households and the associated "ever buyers" matrix. We will assume the matrix on the left reports volume purchasing in kilograms.

	A	B	C	D	Total
HH01	1	0	2	0	3
HH02	0	1	0	0	1
HH03	1	3	0	0	4
HH04	0	0	1	4	5
HH05	1	1	0	1	3
HH06	0	0	0	1	1

"Ever buyers"

	A	B	C	D
HH01	1	0	1	0
HH02	0	1	0	0
HH03	1	1	0	0
HH04	0	0	1	1
HH05	1	1	0	1
HH06	0	0	0	1

We see that buyers of brand A purchased 3 kg of brand A and a total of 10 kg in the category (i.e., SCR = 30%). We see that they also purchased 4 kg of brand B, 2 kg of brand C and 1 kg of brand D. The associated numbers for all brands are given in the following table:

	A	B	C	D
A	3	4	2	1
B	2	5	0	1
C	1	0	3	4
D	1	1	1	6

The sum of the elements of each row gives us the total amount of category purchasing by buyers of the brand of that row. Dividing each row entry by

the sum of that row's elements gives us the following cross purchasing table, the diagonal of which is obviously SCR.

	A	B	C	D
A	30%	40%	20%	10%
B	25%	63%	0%	13%
C	13%	0%	38%	50%
D	11%	11%	11%	67%

How do we create the table efficiently when we have a large number of panellists and it is therefore not practical to do so by hand? One approach is to pre-multiply the panellist × brand volume purchasing summary table by the transpose of the "ever buyers" matrix:

	HH01	HH02	HH03	HH04	HH05	HH06	A	B	C	D	
A	1	0	1	0	1	0	1	0	2	0	HH01
B	0	1	1	0	1	0	0	1	0	0	HH02
C	1	0	0	1	0	0	1	3	0	0	HH03
D	0	0	0	1	1	1	0	0	1	4	HH04
							1	1	0	1	HH05
							0	0	0	1	HH06

(You should verify for yourself that this results in the table give above.)

The cross purchasing analysis for year 1 is reported in Table 5.4. We see that for those panellists that purchased Alpha at least once in year 1, 69% of their category volume purchased went to Alpha, 18% went to Bravo, 8% to Charlie, and so on.

		% total volume purchased of				
		Alpha	Bravo	Charlie	Delta	Other
Purchasers of	Alpha	69%	18%	8%	4%	1%
	Bravo	21%	68%	7%	2%	1%
	Charlie	26%	24%	45%	5%	1%
	Delta	31%	14%	15%	40%	1%
	Other	23%	41%	5%	2%	29%

Table 5.4: Cross purchasing analysis for year 1

We perform this analysis on the edible grocery dataset in the following manner:

- Looking at **panellist x brand (volume)**, we see that each panellist's lack of purchasing of any brand is indicated by an empty cell. These empty cells cause a problem when using `MMULT`. We need a lack of purchasing to be indicated by 0. We could use the Excel feature that allows us to fill empty cells with zeros or we can replicate the table in the following manner. We enter `=B2` in cell J2 and copy the formula across and down to N4575.

- We insert a new worksheet, renaming it **Cross purchase**, and enter the brand names in cells C2:G2 and B3:B7.

- We type=`MMULT(TRANSPOSE('Ever buyers'!B2:F4575),'panellist x brand (volume)'!J2:N4575)` in cell C3 and press Enter. As noted above, the rest of the table should automatically be populated if you are using the latest version of Excel.

- Having entered the brand names in cells C10:G10 and B11:B15, we enter `=C3/SUM($C3:$G3)` in cell C11, formatting the result as a percentage. We then copy this formula across and down to G15, giving us the cross purchase table. (The diagonal of this table gives us the SCR numbers computed in the worksheet **SCR**.)

For a given brand, we can plot the associated row entries as a pie chart — see, for example, the worksheet **Importance of competition plot**, which plots the numbers in cells C11:G11 in the **Cross purchase**.

As previously noted when we computed SCR, we see that Alpha accounts for 69% of category purchasing by the buyers of Alpha. We see from the cross purchase analysis see that Bravo accounts for 18% of their category purchasing. Is this large or small? One way of answering this question is to compare actual purchasing against expectation given general purchasing patterns in the category.

- We insert a new worksheet, renaming it **Importance against expectation**, and enter the brand names in cells B1:F1.

- We first enter the cross purchasing percentages for Alpha in row 2: we enter `='Cross purchase'!C11` in cell B2 and copy the formula across to cell F2. Next, we compute the *volume* market share numbers for each brand in row 3. We enter `=SUM('panellist x brand (volume)'!B2:B4575)/SUM('panellist x brand (volume)'!$G$2:$G$4575)` in cell B3, format the result as a percentage, and copy the formula across to cell F3.

- In row 5, we compute the share of residual purchasing amongst Alpha buyers (i.e., the percentage of the category purchasing not accounted for by Alpha that goes to each of the other brands). We enter `=C2/(1-$B$2)` in cell C5 and copy the formula across to cell F5.

- In row 6, we compute the residual share of category purchasing (across *all category buyers*) once Alpha is removed (i.e., when we exclude Alpha, what percentage of (the remaining) category purchasing goes to each of the other brands). We enter `=C3/(1-$B$3)` in cell C6 and copy the formula across to cell F6.

- Let us consider Charlie. With reference to cell D5, we see that it accounted for 26% of the category purchasing by Alpha buyers that did not go to Alpha. If the purchasing of Alpha buyers was consistent with overall market patterns (as reflected in the volume market shares), we would expect Charlie to account for 19% of their purchasing (cell D6). We can therefore say that Charlie's share of purchasing amongst the buyers of Alpha is above expectation (when expectation is based on overall patterns of buying behaviour).

- In row 7 we create an index against expectation. We enter `=100*C5/C6` in cell C7 and copy the formula across to cell F7. We plot these numbers using a horizontal bar chart.

We see that, relative to market share, Bravo is less of threat to Alpha than we would expect (index = 84). Relative to market share, Charlie and Delta are purchased more by buyers of Alpha than we would expect.

We can repeat these analyses using spend rather than volume purchasing. Using the **panellist x brand (spend)** worksheet from the workbook and following the logic outlined above, we create **Cross purchase (spend)**. We see, for example, that buyers of Alpha spent 71% of their category spend on Alpha. This is in contrast to the 69% of their category volume requirements satisfied by Alpha. We can create a spend-based importance against expectation plot, using *value* market share as the reference.

Chapter 6

Exploring Dynamics in Buyer Behaviour

Up to now, we have been characterising buyer behaviour in a given time period (i.e., one year), be it focusing on one brand (Chapter 4) or multiple brands (Chapter 5). We now consider some standard analyses that give insight into the dynamics of buyer behaviour from period to period. We first consider the case of established products and then turn our attention to the analysis of new product buying behaviour.

6.1 Established Products

We are interested in summarizing how buyer behaviour varies across consecutive periods. We first consider how temporal variations in total sales can be understood by decomposing total sales. Next we explore temporal variations in customer-level purchasing by examining how the distribution of purchasing in one period varies as a function of the level of purchasing in the previous period. Finally, we consider a summary measure of period-to-period purchasing called the repeat rate.

6.1.1 Understanding Temporal Variations in Sales

Most firms have systems that will report sales over time. As we try to make sense of any observed changes, it is helpful to note a fundamental (multiplicative) sales decomposition. For any time period,

$$\begin{aligned}\text{Sales} = {} & \#\text{ households (HHs) in the country} \\ & \times \text{proportion of HHs buying the brand (penetration)} \\ & \times \#\text{ purchase occasions per buyer (PPB)} \\ & \times \#\text{ packs per purchase} \\ & \times \text{weight or price per pack}\end{aligned}$$

There is nothing magical about this specific decomposition. We can create variations on a theme that are more relevant for the specific analysis setting at hand. For example, suppose we are doing an analysis at the brand level, where the SKUs associated with the brand come in different sizes. Furthermore, suppose the time period is sufficiently small that households make only one purchase per period, if at all. (In other words, PPB = 1). A more relevant decomposition would be

$$\begin{aligned}\text{Sales (\$)} = \#\ &\text{households (HHs) in the country}\\ &\times \text{proportion of HHs buying the brand (penetration)}\\ &\times \text{average volume per purchase}\\ &\times \text{average price per unit of volume}\end{aligned}$$

The product of the last two quantities is often called average order value.

Recall the plot of Alpha's revenue we created in Chapter 3, which we repeat in Figure 6.1

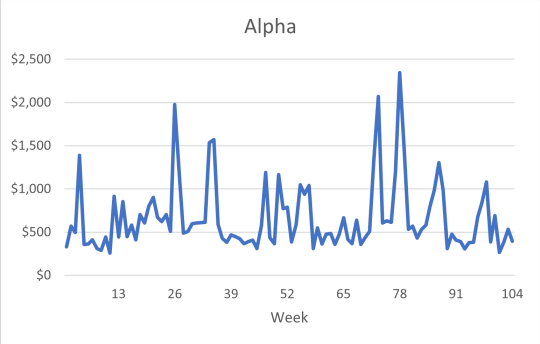

Figure 6.1: Plot of weekly revenue for Alpha

We observe some weeks where this is a massive increase in revenue. How much of this is due to an increase in penetration versus, say, buyers simply buying more product on a given purchase occasion? Let us explore this using the second decomposition given above.

We start by computing Alpha's *weekly* penetration numbers.

- We copy the **edible_grocery** worksheet from the workbook used in Chapter 4 to a new workbook. (Don't forget to save this new workbook, say as `chapter_6a.xlsx`.)
- The first thing we will do is check whether PPB = 1. We insert a pivot table where Rows is trans_id, Filter is brand (selecting just Alpha) and Values is (Average of) week and (Average of) panel_id. This gives us a table (**Sheet2**) that reports the panellist id and week associated with each transaction.

- We copy the contents of this table (cells A4:C18303) to adjacent cells (e.g., E4:G18303) and add the relevant column labels `trans_id week panel_id` in row 4.
- With the active cell somewhere in this new table, we insert a pivot table where Rows is week and Values is (Count of) trans_id and (Count of) panel_id. We note that the number of transaction ids associated with each week equals the number of panellist ids associated with each week. In other words, no panellist made more than one purchase in any given week, which means PPB = 1, as assumed above.
- We finish by copying the week numbers and panellist counts into a new worksheet (which we call **Weekly summary**), adding the labels `week` and `# panellists` in cells A1 and B1, respectively.

The next step is to add the weekly volume and revenue numbers for Alpha to **Weekly summary**. We computed these numbers in Chapter 3 and so we copy the relevant columns from the associated workbook, adding the labels `volume` and `revenue` in cells C1 and D1, respectively.

We now have all the data to create the numbers associated with the revenue decomposition.

- Weekly penetration is simply the number of panellists active in any given week divided by the size of the panel. We enter `penetration` in cell E1, enter `=B2/5021` in cell E2, format as a percentage, and copy the formula down to E105.
- Average order value is simply total revenue for any given week divided by the number of panellists that made at least one purchase of Alpha in that week. Having entered `avg order value` in cell F1, we enter `=D2/B2` in cell F2 and copy the formula down to F105.
- As noted above, this quantity can be decomposed into average order volume and average price per unit volume (in this case, kg).
- Average order volume is simply (total) volume sold in any given week divided by the number of panellists that made at least one purchase of Alpha in that week. Having entered `avg order volume` in cell G1, we enter `=C2/B2` in cell G2 and copy the formula down to G105.
- Average price per unit volume is simply total revenue for any given week divided by (total) volume sold in that week. Having entered `avg price/kg` in cell H1, we enter `=D2/C2` in cell H2 and copy the formula down to H105.

Having computed these four quantities, we use the correlation option in the Data Analysis ToolPak to compute the correlations between weekly revenue and the components of its (multiplicative) decomposition.

Figure 6.2 is a plot of the penetration numbers. Comparing this with Figure 6.1, we see that the fluctuations in revenue go hand-in-hand with fluctuations in penetration.

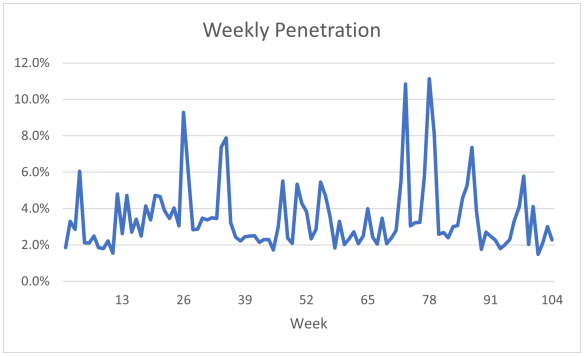

Figure 6.2: Plot of weekly penetration for Alpha

To make this comparison clearer, we overlay the two series of numbers in Figure 6.3. The correlation between these to quantities is 0.98. It would appear that the key driver of revenue increases is simply more people buying the brand that week.

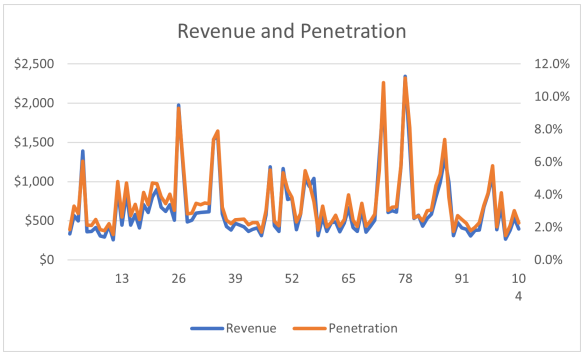

Figure 6.3: Plot of weekly revenue and penetration for Alpha

However, the lack of a perfect correlation means there is some variability in average order value that is not highly correlated with penetration. We plot this quantity in Figure 6.4. (The correlation between penetration and average order value is 0.39.) In order to get a sense of what lies behind the variability in average order value, we plot in Figures 6.5 and 6.6 weekly average order volume and average price per kg, respectively.

Looking at Figure 6.6, there was much less variability in price/kg in weeks 1–52 compared to weeks 53–104. It would appear that there was some change in promotion policy between years 1 and 2. We do not have

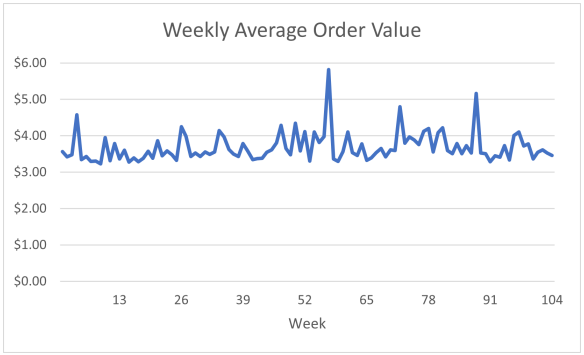

Figure 6.4: Plot of weekly average order value for Alpha

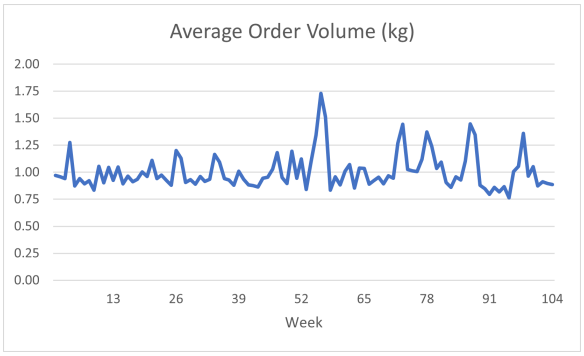

Figure 6.5: Plot of weekly average order volume for Alpha

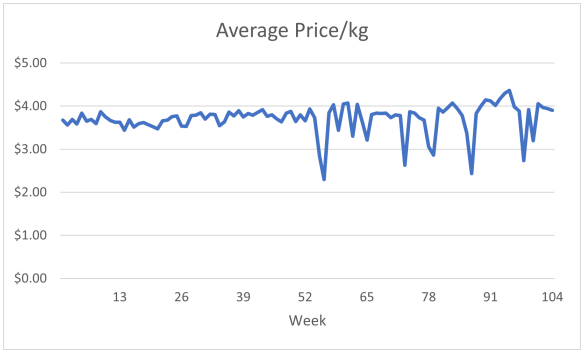

Figure 6.6: Plot of weekly average price per kg for Alpha

the data to explore this further.[1]

This same logic can be used to analyse change across longer time periods.

[1]However, a curious reader might want to compute the correlations between weekly revenue and the components of its (multiplicative) decomposition separately for each year and reflect on the various differences observed.

For example, we note from our analysis in Chapter 3 that Alpha's revenue increased from \$33,571 in year 1 to \$35,251 in year 2, a 5% increase. As we are looking a static sample (i.e., the same group of panellists over the two years), the increase cannot be due to population growth. Are we observing this growth because more households are buying the product, and/or because those that are buying it are buying it more often, and/or because the average spend per transaction has increased?

In order to explore this, we will use the following (multiplicative) sales decomposition:

$$\begin{aligned}\text{Annual revenue} = {} & \#\text{ households (HHs) in the panel} \\ & \times \text{proportion of HHs buying the brand (penetration)} \\ & \times \#\text{ purchase occasions per buyer (PPB)} \\ & \times \text{average order volume} \\ & \times \text{average price per unit of volume}\end{aligned}$$

Our goal is to create Table 6.1.

	Year 1	Year 2	% change
Penetration	52%	55%	5%
PPB	3.5	3.3	−3%
Average order value	\$3.71	\$3.82	3%
Average order volume	1.0	1.1	11%
Average price/kg	\$3.66	\$3.41	−7%

Table 6.1: Comparing the revenue decomposition for year 1 with that for year 2.

The first thing we need to compute is the number of panellists that made at least one transaction in year 1 and year 2.

- Going back to the **edible_grocery** worksheet, we insert a pivot table where Rows is panel_id, Columns is year, Filter is brand (selecting just Alpha) and Values is (Sum of) panel_id. This gives us a new worksheet, **Sheet5**.
- We enter `=1*(B5>0)` in cell E5 and copy the formula down to E3146. This column of 0s and 1s indicates whether or not each panellist made a purchase in year 1. We sum this up by entering `=SUM(E5:E3146)` in cell E4 and see that 2624 households purchased Alpha at least once in year 1.
- Copying E4:E3146 across to F4:F3146 gives us the equivalent information for year 2.

Next, we compute total revenue and volume for Alpha by year. Going back to the **edible_grocery** worksheet, we insert a pivot table where Rows is year, Filter is brand (selecting just Alpha) and Values is (Sum of) spend and (Sum of) volume. This gives us a new worksheet (**Sheet6**).

We add a new worksheet, which we call **Annual Summary**. We enter the column labels `Year 1` and `Year 2` in cells B1:C1 and the row labels `# buyers # transactions revenue total volume` in cells A2:A5. The relevant values are copied from **Sheet5** and **Sheet6**. Recalling that **Sheet3** contains a weekly count of the number of transactions associated with Alpha, we compute the total number of transactions for years 1 and 2 by entering `=SUM(Sheet2!B4:B55)` and `=SUM(Sheet2!B56:B107)` in cells B3 and C3, respectively.

We can now compute the components of the revenue decomposition.

- Year 1 penetration is simply the number of households that made at least purchase of Alpha in the year divided by the size of the panel. We enter `=B2/5021` in cell B7.
- Year 1 PPB is the total number of transactions associated with Alpha divided by the number of households that made at least purchase of Alpha in the year. We enter `=B3/B2` in cell B8.
- Year 1 average order value is annual revenue divided by the total number of transactions associated with Alpha. We enter `=B4/B3` in cell B9.
- Year 1 average order volume is annual total volume divided by the total number of transactions associated with Alpha. We enter `=B5/B3` in cell B10.
- Year 1 average price/kg is revenue divided by total volume. We enter `=B4/B5` in cell B11.
- Copying cells B7:B11 across to C7:C11 gives us the equivalent numbers for year 2.
- Computing the percentage changes gives us Table 6.1.

We note that the 5% increase in revenue from year 1 to year 2 is associated with a 5% increase in the number of households making at least one purchase of Alpha in the year. While PPB drops, average order value increases, with these two changes effectively cancelling out each other. (The product of these two quantities changes by one cent between the two years.) While the average price/kg drops from year 1 to year 2, this is more than compensated by the increase in average order volume, resulting in a 3% increase in average order value between the two years.

6.1.2 Temporal Variation in Customer-level Purchasing

We have observed that the number of buyers has increased, yet the average number of transactions has dropped. Does this mean the "new" buyers are light buyers? Or are the existing buyers buying less?

In order to dig deeper, we need to examine temporal variation in customer-level purchasing. A natural starting point is to examine the joint distribution of purchasing for two consecutive periods, such as that given for Alpha in Table 6.2. We now describe how to create such a table.

		# transactions in year 2										
		0	1	2	3	4	5	6	7	8	9	10+
# transactions in year 1	0	1879	342	105	39	18	9	4	1	0	0	0
	1	259	201	128	79	40	14	6	3	2	0	1
	2	83	120	108	80	75	27	11	9	3	1	0
	3	25	60	78	83	65	54	21	9	3	1	1
	4	8	28	62	45	54	34	26	8	3	6	3
	5	5	13	28	31	49	46	23	20	5	3	4
	6	1	6	15	17	24	31	20	13	11	3	3
	7	0	2	7	5	15	15	16	14	8	4	15
	8	1	1	3	5	9	9	16	6	10	8	3
	9	1	0	3	4	4	4	7	8	4	5	7
	10+	0	1	3	0	3	8	12	14	12	5	49

Table 6.2: Joint distribution of the purchasing of Alpha in years 1 and 2.

- We make a copy of **Sheet2**, and delete rows 1–2 and columns A–D.
- We create a year variable which indicates whether transaction is associated with the first or second year. We first enter `year` in cell D1. Next we enter `=IF(B2<=52,1,2)` in D2 and copy this formula down to D18301.
- We insert a pivot table where Rows is `panel_id`, Columns is year, and Values is (Count of) trans_id.
- We enter the column headings `panel_id year_1 year_2` in cells F4:H4.
- We enter `=A5` in cell F5 and copying the formula across and down to H3146.
- With the active cell somewhere in this new table, we insert a new pivot table where Rows is year_1, Columns is year_2, and Values is (Count of) panel_id. We rename the resulting worksheet **y1 vs. y2 joint distrib)**.
- The pivot table is very sparse for high transaction counts in both the first and second periods. We therefore create a right-censored version of the table (replacing 10 with 10+). Having entering 0, 1, . . . , 10+ in

cells Z4:AJ4 and Y5:Y15, we enter `=B5` in cell Z5 and copy the formula across and down to AI14.

- Note that we have been working with a subset of the original dataset that only contains the purchasing of those that bought Alpha at least once in the two years. Looking at the bottom-right cell of the pivot table, we see that there are 3142 such households. The panel contains 5021 panellists. Therefore the correct number of households that made zero purchases of Alpha in years 1 and 2 is obtained by entering `=5021-3142` in the (0,0) cell of the table (Z5).

- Next, we compute the 10+ numbers by first entering `=SUM(B15:B23)` in cell Z15 and copying the formula across to AI15. Next, we enter `=SUM(L5:S5)` in cell AJ5 and copy the formula down to AJ14. Finally, the number of panellists that made 10+ purchases in both years is computed by entering `=SUM(L15:S23)` in cell AJ15. This gives us Table 6.2.

- How do we read this table? Cells Z6:AJ6 tell us how many people who bought Alpha once in year 1 bought Alpha 0, 1, 2, ... times in year 2. For example, 259 households didn't buy Alpha in year 2, 201 bought Alpha once in year 2, and so on.

- Having computed the row (cells AL5:AL15) and column (cells Z17:AJ17) totals, we compute the marginal distribution of purchasing of Alpha in the first year in cells AM5:AM15, and the marginal distribution of purchasing in the second year in cells Z18:AJ18.

Given this joint distribution of transaction counts for years 1 and 2, it is a simple exercise to compute the row percentages, giving us the *conditional* distributions of transaction counts — see Z22:AJ32. How do we interpret this table? Looking at Z23:AJ23, we see that 35.3% of the panellists that made one purchase of Alpha in year 1 made no purchases of Alpha in year 2, 27.4% purchased Alpha once, 17.5% purchased Alpha twice, and so on.

Let us made two immediate observations:

- The distribution of purchasing in the first year (cells AM5:AM15) is reasonably similar to that for the second year (cells A18:AJ18). Some differences that stand out are the smaller percentage of households making zero purchases in year 2, which corresponds to the higher penetration, and the smaller percentage of households buying Alpha ten or more times in year 2.

- When first seeing a table that shows the distribution of year 2 purchasing broken down by the level of year 1 purchasing, many expect there to be a strong diagonal in the table and are alarmed by the fact

that this is rarely the case. It is important to realize that buying behaviour is not deterministic. From the perspective of the analyst, it can be viewed as-if random, bouncing around each person's underlying propensity to buy the product. Someone who makes one purchase in year 1 is possibly a light buyer and so is the fact that 35.3% did not buy the product again in the second year that surprising? It does not mean that they are "lost"; most of them will buy Alpha again sometime the following year (year 3). Similarly, are those who did not buy the product in period 1 but did in period 2 new customers? Some probably are. But, for any established product category, most are probably people who have purchased the product in previous years and who, for whatever reason, did not buy it that year.[2]

6.1.3 Repeat Rates

In many situations, reports such as Table 6.2 are too detailed. One common summary measure is the *repeat rate* (or repeat-buying rate), which is defined as the percentage of the brand's customers in a given period who also purchase the product in the following period. We now explore how to compute quarter-by-quarter repeat rates. We will doing this for Alpha, and our goal is to create Figure 6.7.

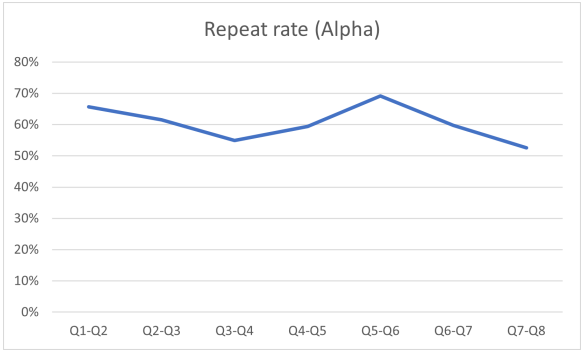

Figure 6.7: Plot of quarterly repeat rates for Alpha

- The first step is to create a table that indicates whether or not each panellist made at least one purchase of Alpha in each of the eight quarters in our dataset (i.e., an ever-buyers (by quarter) table for Alpha).
- Having made a copy of **Sheet2**, we want to create a variable that reports the quarter associated with the week of each transaction. We

[2]The key to analysing such tables is to compare them to a benchmark that assumes stable underlying purchasing patterns. See for example, Greene, Jerome D. (1982), *Consumer Behavior Models for Non-Statisticians*, New York: Praeger, pp. 55–58.

first enter `quarter` in cell H3. Next we enter `=INT((F4-1)/13)+1` in H4 and copy this formula down to H18303.

- We insert a pivot table where Rows is `panel_id`, Columns is quarter, and Values is (Sum of) trans_id. We rename this worksheet **Repeat rates**.

- Having entered in the quarter labels (`Q1`... `Q8`) in cells L4:S4, we create the ever-buyers (by quarter) table by entering `=1*(B5>0)` in cell L5 and copying the formula across and down to S3146. An entry of 1 means the panellist made *at least* one purchase of Alpha in the quarter; 0 means no purchase (of Alpha) occurred.

- Recall that the repeat rate is the percentage of a brand's customers in a given period who also purchase the product in the following period. How many households bought Alpha in Q1? It is simply the sum of the numbers under Q1. (This is the denominator.) If you purchased Alpha in both periods, you will have a 1 for Q1 and a 1 for Q2. If we multiple these two columns of numbers together, only those customers that purchased Alpha in both periods will have a 1. Any other combination of purchasing will result in a zero. Summing up this product gives us the number of households that purchased Alpha at least once in both periods. (This is the numerator.)

- Entering `=SUMPRODUCT(L5:L3146,M5:M3146)/SUM(L5:L3146)` in cell U5 gives us the Q1-Q2 repeat rate. We see that 66% of those households that purchased Alpha in Q1 bought the brand again at least once in Q2.

- Copying this formula across to AA5 gives us the repeat rates for the other quarters. Plotting these numbers gives us Figure 6.7.

We see that there is some variability in the repeat rate. Is it possibly in decline? One thing to realise is that the variability can be driven by the firm's promotional activities. Aggressive promotions may attract a segment of consumers that only buy on promotion. If the firm promotes heavily in one quarter and has fewer promotions the following quarter, we could expect to a drop in the repeat rate. (The interested reader could explore this by looking at how Alpha's promotional activity (as reflected in price/kg) varied across quarters.)

6.2 New Products[3]

Central to diagnosing the performance of a new product is the decomposition of its total sales into trial and repeat sales. A given overall aggregate sales history could be the realization of very different purchasing scenarios. For example, a low sales level for a new product could be the result of (i) many consumers making a trial (i.e., first-ever) purchase but few of them making a repeat purchase (because the product does not meet their expectations), or (ii) a low trial rate but a high level of repeat purchasing amongst the triers (because the product meets a real need among a relatively small set of buyers). Without a trial/repeat sales decomposition, it is impossible to determine which of these scenarios best describes the sales data. (And of course it is impossible to make such decompositions without access to household-level purchasing data (i.e., consumer panel data).)

An example of such a sales decomposition is given in Figure 6.8, which shows the week-by-week sales of Kiwi Bubbles (dataset 2) broken down into trial and repeat sales (i.e., first-ever purchases of the new product by a panellist vs. subsequent purchases of the new product by a panellist).

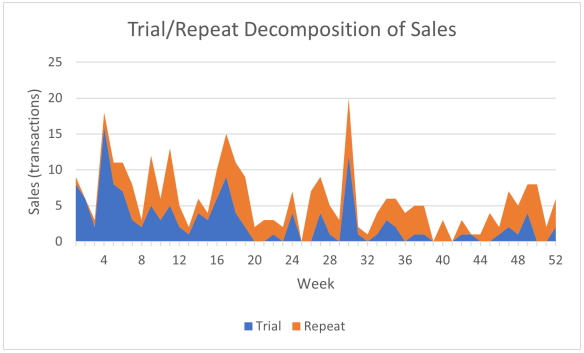

Figure 6.8: A trial/repeat decomposition of Kiwi Bubble sales.

Another variant of the sales decomposition is given in Figure 6.9, which plots *cumulative* sales and separates repeat sales into first repeat (second-ever) purchases and additional repeat (second repeat (third-ever) purchases + third repeat (fourth-ever) purchases + ...). For this plot, we see that 45% of Kiwi Bubble's year 1 sales came from trial purchases and 38% were due to additional repeat purchases.

Given the data summary from which these two figures are created, we can create further plots that give us insight into buyer behaviour. For example, we can plot cumulative trial (Figure 6.10), which shows the percentage of

[3]This section draws on material in the "Creating a Depth-of-Repeat Sales Summary Using Excel" note by Pete Fader and Bruce Hardie (`http://brucehardie.com/notes/006/`).

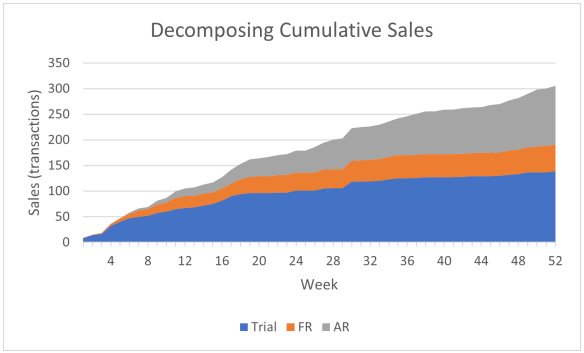

Figure 6.9: Decomposing cumulative sales into its trial, first repeat, and additional repeat components.

households in the market that have made a trial purchase by any week in the new product's first year on the market. (We see that just under 10% of households have tried the new product by year end. This still appears to be growing but at a far slower rate than earlier on in the year.)

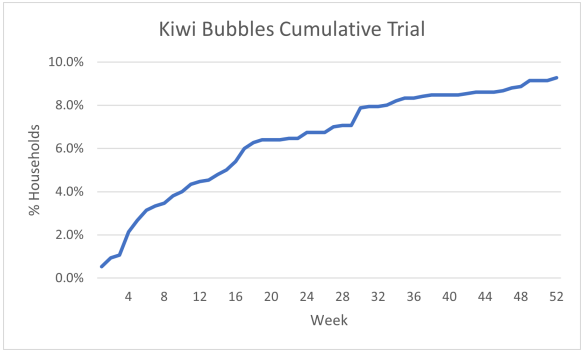

Figure 6.10: Growth in cumulative trial over time.

People can be induced to try a new product by promotions, etc. but the key to success is repeat purchasing. The first step is making a first-repeat purchase. One useful metric to track is "percent triers repeating" — the percentage of panellists that made a trial purchase that have gone on to make a (first) repeat purchase. The associated plot for Kiwi Bubbles is given in Figure 6.11.

We now explore how to create these basic plots (and one more) using consumer panel data.

6.2.1 Basic Analyses of New Product Performance

Our initial goal is to create a summary of new product purchasing from which Figures 6.8–6.11 can be created.

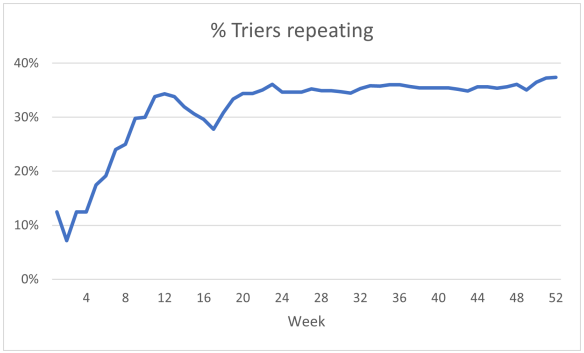

Figure 6.11: Evolution of the percentage of triers that have made a repeat purchase.

- We start by opening`kiwibubbles_tran.csv` in Excel. We immediately save it as an Excel workbook, say `chapter_6b.xlsx`).
- We will just focus on purchasing by those panellists in Market 2. We make a copy of the **kiwibubbles_tran** worksheet (renaming it **Market2**), delete the records associated with market 1 (rows 2–552) and the column corresponding to the Market field.
- The next step is to determine the so-called "depth of repeat" (DoR) level associated with each transaction. Is it a trial purchase (DoR = 0), a first repeat purchase (DoR = 1), a second repeat purchase (DoR = 2), etc. Labelling column E `DoR` (cell E1), we enter `=IF(A2<>A1,0,E1+1)` in cell E2 and copy the formula down to E307. (Note that this formula assumes the records are sorted chronologically for each panellist.)
- The next step is to create a table that tells us how many trial, first repeat, etc. purchases (columns) occurred in each week (rows). We want there to be 52 rows, one for each week of the test. However, it turns out that this panel of 1499 households only purchased the test product in 49 weeks; no purchases occurred in weeks 25, 39, and 41. How can we create a table that will contain zeros in the rows corresponding to these three weeks? At the bottom of column B, we add the numbers 1, 2, ..., 52. For these new records, we assign a depth-of-repeat level of −1 (cells E308:E359).
- We insert a pivot table where Rows is Week, Columns is DoR, and Values is (Count of) ID, and rename the resulting worksheet **DoR by week**.[4] We create a cleaned-up version of this table to the right of the pivot table output, adding meaningful column names.

[4]Note that, in order to make life simple for ourselves at this stage, we will assume

We now consider the basic set of plots that we can create using this data summary:

- We create Figure 6.8, which breaks weekly sales into its trial and repeat components, in the following manner:
  - We insert a new worksheet, renaming it **TR decomposition**, and add the week numbers in column A.
  - The weekly trial numbers are extracted from the worksheet **DoR by week** by entering `='DoR by week'!R5` in cell B2 and copying the formula down to B53. We label this column `Trial` (cell B1).
  - Repeat sales can be computed by either summing up the R1, R2, R3, ... numbers or by subtracting trial sales from total sales. We enter `='DoR by week'!AD5-'DoR by week'!R5` in cell C2 and copy the formula down to C53. We label this column `Repeat` (cell C1).
  - Given these two columns of data, we create the basic trial/repeat decomposition of weekly sales plot using the chart type "2-D Stacked Area".
  - We can also used the chart type "2-D 100% Stacked Area" to create a plot that shows the percentage of weekly sales due to trial versus repeat purchasing.
- We create Figure 6.9, which breaks *cumulative* sales into its trial, first repeat (FR), and additional repeat (AR) components, in the following manner:
  - We insert a new worksheet, renaming it **Cum. sales decomposition**, and add the week numbers in column A.
  - The weekly trial and first-repeat numbers are extracted from the worksheet **DoR by week** by entering `='DoR by week'!R5` in cell B2 and `='DoR by week'!S5` in cell C2, and copying these formulas down to row 53.
  - The AR sales numbers are computed by entering `=SUM('DoR by week'!T5:AC5)` in cell D2 and copying the formula down to D53.
  - We enter the column headings `Trial FR AR` in cells B1:D1.
  - Having copied the week numbers from column A to column F and entered the column headings `Trial FR AR` in cells G1:I1, we create the cumulative trial, first-repeat, and additional-repeat

that only one unit is purchased per purchase occasion. If we want Figures 6.8 and 6.9 to report unit sales, we would need Values to be (Sum of) Units. The pivot table we have just created is needed for Figures 6.10 and 6.11.

sales numbers by first entering `=B2` in cell G2, `=G2+B3` in cell G3, copying this second formula down to G53, and then copying the formulas in cells G2:G53 across columns H and I.

- Given these three columns of data, we create the trial/FR/AR decomposition of cumulative sales plot using the chart type "2-D Stacked Area".

- We create Figure 6.10, which plots the growth in the *percentage* of households that have made a trial purchase, in the following manner:
  - We insert a new worksheet, renaming it **Cum. trial**, and add the week numbers in column A.
  - We need to extract the cumulative trial numbers from the worksheet **Cum. sales decomposition**. We do so by entering `='Cum. sales decomposition'!G2` in cell B2 and copying the formula down to B53.
  - Recall that there are 1499 panellists in Market 2. We compute the % cumulative trial numbers by entering `=B2/1499` in cell C2 (formatting the answer as a percentage) and copying the formula down to C53.
  - We plot the numbers in C2:C53.

- We create Figure 6.11, which plots the evolution of the percentage of panellists that made a trial purchase that have gone on to make a (first) repeat purchase, in the following manner:
  - We insert a new worksheet, renaming it **% Triers repeating**, and add the week numbers in column A.
  - We enter `='Cum. sales decomposition'!G2` in cell B2 and `='Cum. sales decomposition'!H2` in cell C2, and copy these formulas down to row 53. This gives us the trial and first repeat numbers from the worksheet **Cum. sales decomposition**.
  - The % triers repeating numbers are computed by entering `=C2/B2` in cell D2 and copying the formula down to D53.
  - We insert a line chart which plots the numbers in D2:D53.

6.2.2 Exploring Time to First Repeat

Looking at Figure 6.11, we see that the appeal of the new product is such that just over 37% of those customers that made a trial purchase in the first year it was on the market ended up making a repeat purchase of the new product in that year. Note that these % triers repeating numbers plotted in Figure 6.11 are in calendar time (or, more precisely, time since the launch of the product). We can compute a related measure that tells

us how many weeks after their trial purchase a panellist makes their (first) repeat purchase — see Figure 6.12.[5]

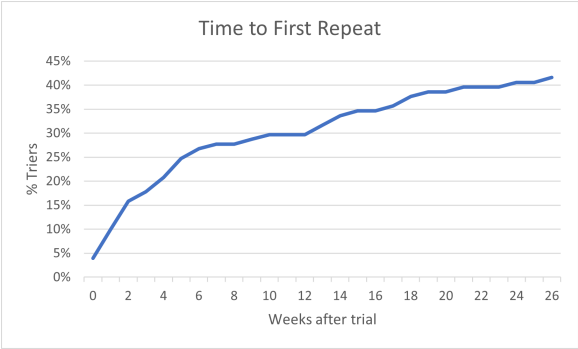

Figure 6.12: First repeat curve

We now consider how to create this plot. As a first step, we create a report that tells us the cumulative percentage of triers that have made a (first) repeat purchase so many weeks after their trial purchase, broken down by time of trial.

- We make a copy of the worksheet **Market 2**, renaming it **Cum. FR by trial class (a)**, and delete the numbers we added below row 307.
- We create a `Trial week` variable (cell F1) that equals the value of Week if this is a trial purchase and -99 otherwise. We enter the formula `=IF(E2=0,B2,-99)` in cell F2 and copy it down to F307.
- Next, we create a `FR delta` variable (cell G1) that tells us how many weeks after the trial purchase the panellist's first repeat purchase occurred (assuming it was observed). We enter the formula `=IF(E3=1, B3-F2,-99)` in cell G2 and copy down to cell G307; -99 indicates that a first repeat purchase was not observed for this panellist in the 52-week observation period.[6]
- A pivot table where Rows is Trial week, Columns is FR delta, and Values is (Count of) ID will give us a table that shows how many weeks after the trial purchase the first repeat purchase occurred, broken down by week of trial purchase. However, it will not be a 52×53 matrix

[5] We see that 42% of triers have made a first repeat purchase within a year of their trial purchase. This is higher than the 37% of triers that have made a repeat purchase by the end of week 52 reported in Figure 6.11. Why is this the case?

[6] Strictly speaking, FR delta $= 0$ if the first-repeat purchase occurs in the same *calendar* week as the trial purchase, 1 if the first-repeat purchase occurs in the *calendar* week immediately after that in which the trial purchase occurred, etc.

of numbers,[7] as trial purchases did not occur on every week and we do not observe first repeat purchases being made in all the possible weeks after trial. In order to ensure that we automatically generate this 52×53 matrix of numbers we first enter 0 to 51 in cells G308:G359 and -99 in cells F308:F359, and then 1 to 52 in cells F360:F411 and -99 in cells G360:G411.

- We now insert a pivot table where Rows is Trial week, Columns is FR delta, and Values is (Count of) ID. We rename the resulting worksheet **Cum. FR by trial class (b)**.
- We now create a cleaned-up version of this table.
  - We first insert a new worksheet, renaming it **Cum. FR by trial class (c)**.
  - We enter the "Trial week" numbers in cells B3:B54. The number of triers in each week are extracted from the relevant row totals in the pivot table we just created. We enter `='Cum. FR by trial class (b)'!C6` in cell C3 and copy the formula down to C54.
  - The "Weeks after trial" numbers (0–51) are entered in D2:BC2, and the main entries from the pivot table are extracted by entering`='Cum. FR by trial class (b)'!C6` in cell D3 and copying the formula across and down to BC54.
- Notice the greyed-out cells. These are by definition 0. As we have 52 weeks of data, the only time someone who made a trial purchase in week 52 could make a repeat purchase is in week 52. (One week after trial would be week 53, which we do not observe.)
- Looking at row 3, we see that eight panellists made a trial purchase during the first week Kiwi Bubbles was on the market. One panellist purchased the product again that same week. Another made their first repeat purchase in two weeks later. Summing cells D3:BC3, we see that five of the eight week-1 triers (62.5%) had made a first repeat purchase by the end of the 52-week observation period.
- We want to create a version of this table in which the rows report the cumulative percentage of triers that have made a first repeat purchase so many weeks after their trial purchase. We enter the "Trial week" numbers in cells B58:B109 and the corresponding number of triers numbers in cells C58:C109. (Enter `=C3` in cell C58 and copy the formula down to cell C109). The "Weeks after trial" numbers (0–51) are entered in cells D57:BC57. Next, we enter `=IF(`

[7] Why 53? Because we have 52 week numbers and -99 (to indicate that a first repeat purchase is not observed).

`$B58>52-D$57,"",IF($C58>0,SUM($D3:D3)/$C58,0))` in cell D58, format the result as a percentage, and copy the formula across and down to BC109. (Notice how this formula automatically suppresses the entries that were previously greyed out.)

A useful visualisation of this table is a plot of the cumulative percentage of triers that have made a first repeat purchase so many weeks after their trial purchase. This is a weighted average of the rows of the table we have just created, where there weights are the number of triers associated with each row.

While we could create a weighted average of all 52 rows, it would be misleading in that we would be mixing groups of triers with different numbers of weeks in which they could have made a repeat purchase. (A week-51 trier has only two weeks in which to make a first repeat purchase: weeks 51 and 52.) Given a 52-week observation period and desire to compute the percentage of triers making a first repeat purchases 0, 1, ..., x weeks after their trial purchase, we can only consider those triers that made a trial purchase in the first $52 - x$ weeks the product was on the market. For this illustrative example, we will create a plot of the cumulative percentage of triers that have a made a first repeat purchase within 26 weeks of their trial purchase.

- We enter the "Weeks after trial" numbers (0, 1, ..., 26) in cells D112:AD112.
- We compute the number of panellists that made a trial purchase in the first 26-weeks in cell C113 (`=SUM(C58:C83)`).
- We enter `=SUMPRODUCT($C$58:$C$83,D58:D83)/$C$113` in cell D113, format the result as percentage, and copy the formula across to AD113.
- We can plot these numbers to give us an empirical distribution of the time from trial to first repeat purchase (Figure 6.12).

We can perform similar analyses for the time from first repeat to second repeat, second repeat to third repeat, and so on. An undocumented analysis of first repeat to second repeat is given in the worksheets **Cum. 2R by FR class (a)**, **Cum. 2R by FR class (b)**, and **Cum. 2R by FR class (c)**.

Chapter 7

Further Reading

We have explored how to perform basic analyses of buyer behaviour using data collected via a consumer panel. The following readings may be of interest to the reader who wishes to step back and reflect on the bigger picture.

- Sudman and Ferber (1979) and Sudman and Wansink (2002) provide a basic overview of consumer panels and explore the issues surrounding the management of a panel.
- Charan (2015, Chapter 7) provides an overview of some basic reports generated using data from a consumer panel.
- Jones and Slater (2003, Chapter 5) provides an introduction to the nature of repeating buying for mature (FMCG) products.
- Sharp (2010) explores a series of empirical regularities regarding buyer behaviour, many of which are derived from analyses of consumer panel data.
- Fader, Hardie, and Ross (2022) illustrates how these types of analyses can be applied to data from a firm's own customer transaction databases.

References

Charan, Ashok (2015), *Marketing Analytics: A Practitioner's Guide to Marketing Analytics and Research Methods*, Singapore: World Scientific Publishing.

Fader, Peter S., Bruce G. S. Hardie, and Michael Ross (2022), *The Customer-Base Audit: The First Step on the Journey to Customer Centricity*, Philadelphia, PA: Wharton School Press.

Jones, John Philip, and Jan S. Slater (2003), *What's in a Name?: Advertising and the Concept of Brands*, 2nd edition, London: Routledge.

Sharp, Byron (2010), *How Brands Grow: What Marketers Don't Know*, Oxford: Oxford University Press.

Sudman, Seymour, and Robert Ferber (1979), *Consumer Panels*, Chicago, IL: American Marketing Association.

Sudman, Seymour, and Brian Wansink (2002), *Consumer Panels*, 2nd edition, Chicago, IL: American Marketing Association.