Survey Data to NBD-Dirichlet

Hume Winzar

Table of contents

1 Converting Survey-Style Data into NBD–Dirichlet Inputs

Although the NBD–Dirichlet model ideally uses **panel data** (exact counts of purchases per brand per consumer), it is still possible to approximate the required inputs using **survey-style self-report data**.

This section explains how to convert typical survey responses into the frequency-based quantities needed by the model.

1.1 Survey Data We Might Receive

Surveys often collect behavioural data in simplified or banded form, for example:

1. **Category frequency bands**
    - "In the past 3 months, how often did you buy this category?"
        - 0 times
        - 1–2 times
        - 3–5 times
        - 6–10 times
        - 11+ times

My preference is to get exact numbers from respondents, but often this is infeasible, and respondents can give only a *best guess* of such behaviour.

2. **Brand repertoire checklist**
    - "Which of the following brands have you bought in the past 3 months?" (Multiple selections allowed.)

Similar data could be gathered from:

3. **Brand purchase proportions**
    - "Of all your purchases in this category, approximately what **percentage** went to each brand?" (Must sum to 100%.)

As a last resort we could use:

4. **Primary / secondary brand indicators**
    - "Which brand did you buy most often?"
    - "Which brand did you buy second most often?"

While imperfect, these can be transformed into approximate **purchase counts**, enabling useful NBD–Dirichlet teaching and exercises.

1.2 Converting Category Frequency Bands Into Numeric Counts

The NBD requires a **distribution of category purchase counts**.
If we only have banded responses, then we can map each band to an approximate numeric value.

Example mapping (customisable):

Band	Assigned numeric value	Rationale
0 times	0	Exact
1–2	1.5	Midpoint
3–5	4	Midpoint
6–10	8	Midpoint
11+	12	Conservative lower bound

From these converted counts, compute: - Mean category frequency μ
- Zero-class proportion $P(0)$ - fit or solve for k using

$$P(0) = \left(\frac{k}{k+\mu}\right)^k$$

This yields the **NBD parameters**.

1.3 Deriving Brand-Level Purchase Data from Survey Responses

1.3.1 Option A — Brand repertoire only (0/1 per brand)

If respondents indicate which brands they bought, but **not how often**, we can:

- Treat each indicated brand as **at least one purchase**.
- Convert repertoire size into a **minimal counts vector**, e.g.:

Apple	Banana	Cherry	Durian	Total (approx.)
1	0	1	0	2

This yields: - **Brand penetrations**

- **Duplication rates**
- Rough **brand shares** (all selected brands treated equally)

Good for teaching duplication laws; less accurate for frequency modelling.

1.3.2 Option B — Purchase proportions (sums to 100%)

If respondents report approximate **share of the category** spent on each brand, e.g.:

Brand	Reported share
Apple	50%
Banana	30%
Cherry	20%
Durian	0%

Then we can multiply by their estimated category frequency to obtain **pseudo-counts**:

$$\text{pseudo-count}_i = \text{category frequency} \times \text{proportion}_i$$

Example: 4 category purchases $\times$ 0.30 = 1.2 $\rightarrow$ round to 1.

This yields: - **Brand purchase frequencies** - **Brand shares** s_i - A usable customer-by-brand matrix

---

1.3.3 Option C — Primary / secondary brand questions

If respondents identify their **main brand** and optionally a second brand:

- Assign **main brand = 70%**, **second brand = 30%**, or similar rule-of-thumb
- Multiply by estimated category frequency to create purchase counts
- Brands not listed receive 0

This produces reasonable approximations of: - Brand shares

- Penetrations
- Duplication patterns

1.4 Building an Approximate Customer–Brand Matrix

After converting responses using any of the above methods, we aim to produce a table like:

Customer	Apple	Banana	Cherry	Durian	Total
R01	2	1	0	0	3
R02	0	4	0	0	4
R03	1	1	1	0	3

From this table we compute:

1.4.1 NBD inputs

- μ = mean row total
- $P(0)$ = proportion of respondents with 0 buys
- k = derived from μ and $P(0)$

1.4.2 Dirichlet inputs

- s_i = column total / grand total
- Brand penetrations p_i = proportion of rows with count > 0
- Duplication matrix = cross-brand overlaps
- Estimate α_0 from duplication or adopt a reasonable value (5–40)

1.5 Summary

> **Key Point**
>
> Survey responses must be transformed into **approximate purchase counts**.
> The NBD–Dirichlet model needs **behavioural frequencies**, not attitudes.

Even approximate pseudo-counts can provide: - Category frequency distribution → NBD

- Brand shares and duplication → Dirichlet
- Predicted penetrations, loyalty patterns, and growth effects

Enough to support robust **diagnostic reasoning**, and simplified **market simulations**.

2 Determining the Base Population and Computing $P(0)$ in the NBD–Dirichlet Model

A common point of confusion is **who counts as a "category buyer"** and therefore who belongs in the **NBD population**.

This matters because the NBD uses the proportion of consumers who made **zero** purchases in the observation period:

$$P(0) = \Pr(0 \text{ category purchases in the time window}).$$

However, only certain types of "zero" belong in this calculation.
This section provides a simple decision process to determine the correct base population.

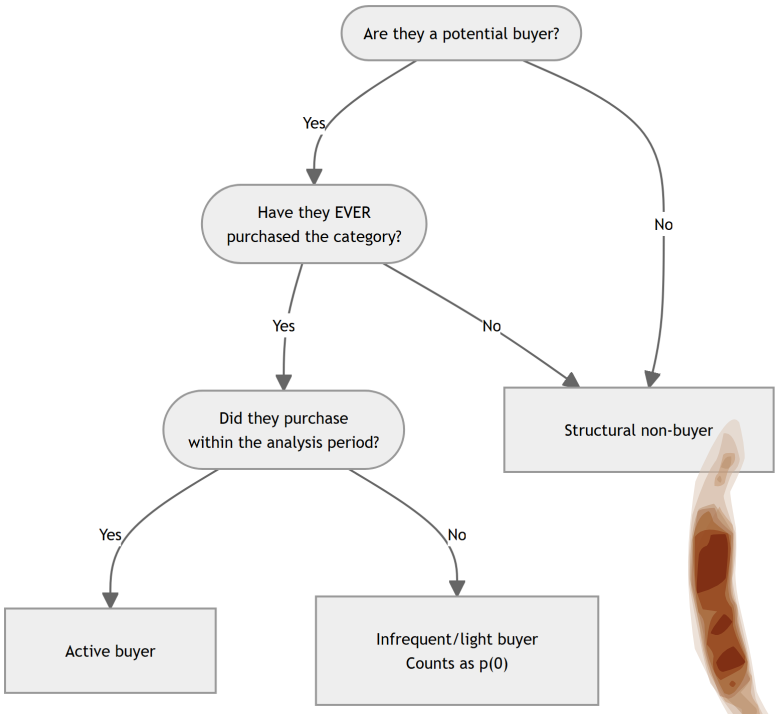
2.1 Three Types of Consumers

For almost any category, individuals fall into one of three groups:

Type of consumer	Behaviour	Include in NBD population?	Contributes to $P(0)$?
1. Active category buyers	Purchased in the period	**Yes**	**No**
2. Infrequent / light buyers	Did not purchase *this period* but do purchase occasionally	**Yes**	**Yes**
3. Structural non-buyers	Never buy the category (e.g., lactose-intolerant for cheese)	**No**	**No**

Only types **1** and **2** are part of the **market for the category** and behave according to the NBD.
Type **3** has a true purchase rate of zero and must be excluded.

2.2 Decision Tree: Who Belongs in the NBD Population?



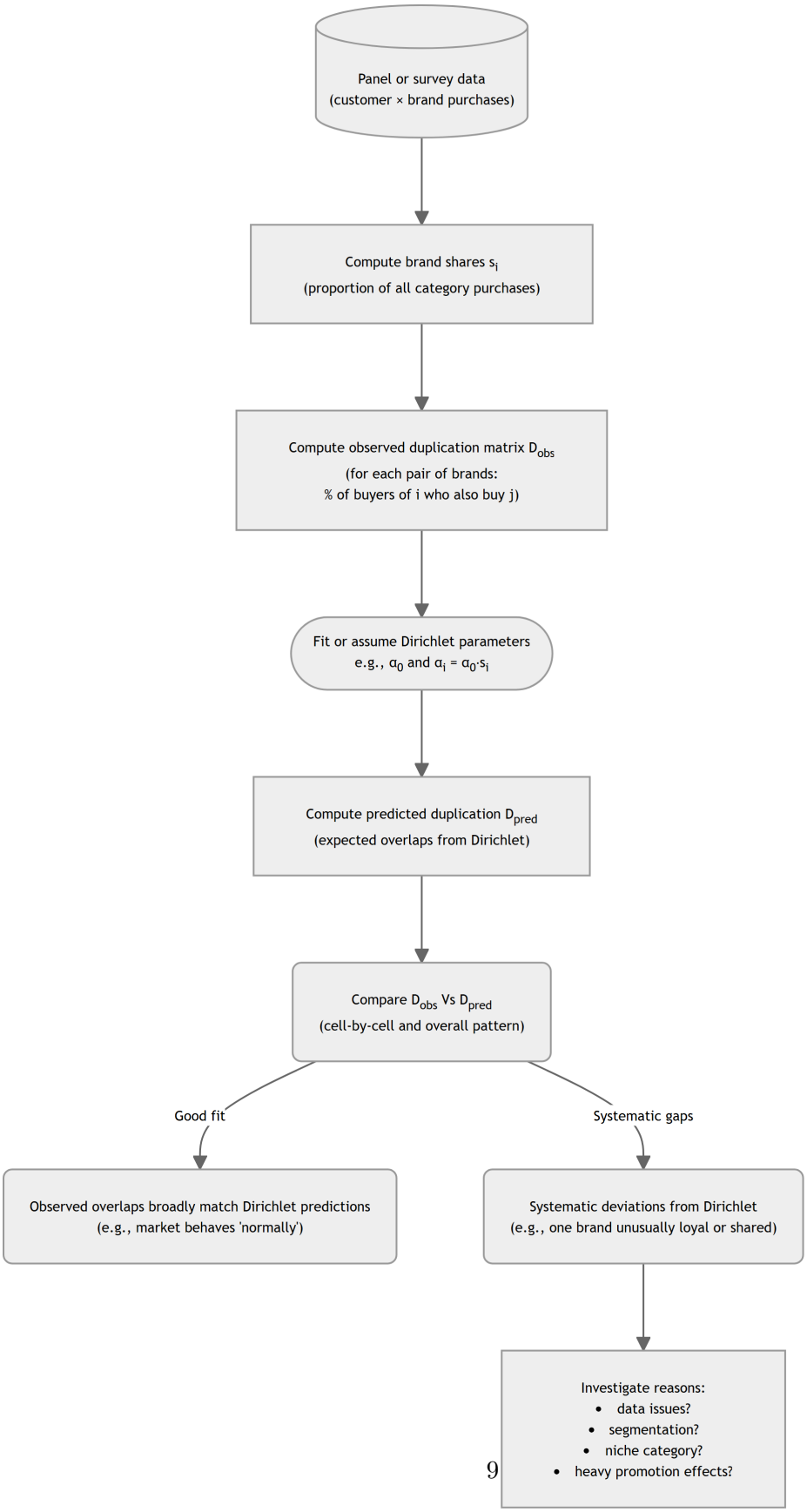
Structural non-buyers

If a person has never purchased the category across an appropriate historical window

- e.g., 12 months for groceries

they are almost certainly a structural non-buyer

2.3 How Dirichlet Predictions Relate to Observed Brand Overlaps



2.4 Constructing the Category Frequency Distribution

Once the base population is defined (i.e., only **active category buyers** and **infrequent buyers**, not structural non-buyers), the next step is to create the **category frequency distribution**.

2.4.1 Steps

1. **Count the number of category purchases** each person made in the analysis period (e.g., 3 months, 6 months, 12 months).
2. **Tabulate** how many people fall into each purchase-frequency group.
3. Ensure the table includes the zero-buyers who are *active* category buyers overall.

2.4.1.1 Example Frequency Table

Purchases in period	Number of consumers
0	8
1	10
2	7
3	5
4+	2

From this, we compute:

- **Mean purchase rate**

$$\mu = \frac{\sum_i (\text{purchases}_i)}{N}$$

- **Zero-class proportion**

$$P(0) = \frac{\text{n of consumers with 0 purchases}}{N} = \frac{8}{32} = 0.25$$

- **Dispersion parameter** k (solved numerically)

  The NBD gives the probability of zero purchases as:

$$P(0) = \left(\frac{k}{k+\mu}\right)^k$$

  Rearranging is not analytically simple, so k is typically found using:

  – Goal Seek (as in the extant *Excel* worksheet)

- Newton–Raphson method
- Grid search

These computed values (μ, $P(0)$, and k) fully specify the **NBD component** of the NBD–Dirichlet model.

2.5 Practical Guidelines

2.5.1 A. Use an "active category buyers" denominator

A robust rule of thumb for panels:

> Include only those consumers who have purchased the category **at least once in the previous 12 months**.

This minimises the inclusion of structural non-buyers.

2.5.2 B. Exclude structural non-buyers

Individuals who *never* buy the category: - Should **not** be included in the NBD population

- Should **not** count toward $P(0)$
- Otherwise, they inflate $P(0)$ and distort k

2.5.3 C. Shorter time windows produce more zero-buyers

This is expected: - A 4-week period may have 40–60% zero-buyers - A 12-month period usually has very few

The choice of period should match: - the category purchase cycle

- the analytical purpose
- standardisation across brands

2.5.4 D. Infrequent buyers are legitimate zero-buyers

Consumers who buy the category occasionally but **did not buy this period** are *exactly* the people the NBD is designed to model.

They must be **included** in the base population and count toward $P(0)$.

2.6 Summary Box

How to Compute $P(0)$ Correctly

- Define the **category-active population** first.
- Exclude **structural non-buyers** (people who *never* buy the category).
- Among the active population, count those with **zero purchases in the analysis period** — this is $P(0)$.
- Use the resulting frequency distribution to compute μ and solve for k.