TECHNICAL PAPER

NBD 2.0: Harmonizing Data to Make It Really Sing

JULY 2017

IRi
Growth delivered.

ABOUT THE AUTHOR

Norbert Schumacher, PhD is a principal of research and development, IRI Consumer and Shopper Marketing. He can be reached at **Norbert.Schumacher@iriworldwide.com**.

ACKNOWLEDGMENTS

Daniel Pliske, PhD is a principal of research and development, IRI Consumer and Shopper Marketing. He can be reached at **Daniel.Pliske@iriworldwide.com**.

ABOUT IRI

IRI is a leading provider of big data, predictive analytics and forward-looking insights that help CPG, OTC health care organizations, retailers and media companies to grow their businesses. With the largest repository of purchase, media, social, causal and loyalty data, all integrated on an on-demand cloud-based technology platform, IRI helps to guide its more than 5,000 clients around the world in their quests to remain relentlessly relevant, capture market share, connect with consumers and deliver market-leading growth. A confluence of major external events—a revolution in consumer buying, big data coming into its own, advanced analytics and automated consumer activation—is leading to a seismic shift in drivers of success in all industries. Ensure your business can leverage data at www.iriworldwide.com.

ABOUT IRI CONSUMER AND SHOPPER MARKETING CENTER OF EXCELLENCE

The IRI Consumer and Shopper Marketing Center of Excellence focuses on leveraging the IRI shopper marketing cloud of vast data to develop deep shopper insights, segment planning, opportunity sizing and activation strategies that empower its customers to win the sale and the shopper. Our solutions help marketers connect with consumers and shoppers one household and one store at a time across as many touchpoints as possible along the new path-to-purchase.

Smoothing the Rough Spots

Gathering, integrating and analyzing big data is no easy feat. With vast discrepancies in consistency, modes of measurement and coverage, using the data to glean an accurate picture of the consumer packaged goods (CPG) market and its consumers requires solid technology and elite analytic knowhow. IRI uses point-of-sale (POS) and National Consumer Panel (NCP) data in a complementary fashion to provide clients with a well-rounded view of attitudes and behaviors, down to the household level.

To ensure maximum accuracy and minimize innate weaknesses, IRI adjusts and aligns the data through complex statistical processes. One such process is known as negative binomial distribution (NBD) adjustment. This paper provides greater detail on the current and evolving statistical process behind IRI's data sets.

Painting a Picture of the Shopper Journey

As one of the original innovators in big data, IRI integrates the world's largest set of otherwise disconnected purchase, media, social, causal and loyalty data to help CPG, retail, over-the-counter health care and media companies grow their businesses. Among the most long-standing and basic data in the IRI library are POS data NCP (also called panel) data.

Because panel and POS form an integral foundation on which CPG companies base marketing and overall business decisions, IRI has invested heavily to eliminate innate weaknesses within the data and capitalize on the benefits each has to offer.

POS data is the gold standard of marketplace measurement of product sales and share trends, measuring actual levels of product sales within and across retail channels and geographies (see exhibit 1). It also provides valuable insight into distribution and price and promotion considerations. Because it is so comprehensive, POS data is ideal for trend analysis.

Panel data relies on a panel of U.S. shoppers who agree to scan their grocery purchases and periodically participate in surveys to share their opinions, beliefs and behaviors. Panelists agree to participate for a specific time period and are rewarded for their participation on a points system.

Through panel data, CPG marketers can glean insights into a variety of household-level measures, including:

- Trial and repeat behaviors
- Penetration
- Purchase frequency
- Cross-purchase behavior
- Differences in purchase behavior across demographics

Furthermore, because panelists provide basic demographic data and answer surveys, panel data provides the ability to link these demographics, behaviors and attitudes to product purchases. Panel data also provides important consumer-level statistics, such as brand penetration; buy rate; brand loyalty (an intra-purchase correlation); and the correlations among categories, subcategories and brands. Taken together, these measures provide valuable insights into "the why behind the buy."

EXHIBIT 1

POS Data

- Marketplace measurement of sale and share levels and trends
- Sales tracking
- Distribution
- Price
- Volume drivers: price and promotion

Consumer Network Panel Data

- Consumer insights into household (HH) purchase behavior
- HH trial, purchase frequency and transaction size
- Demographic, behavioral and attitudinal profiling
- HH spending
- Attitude and usage drivers

The Sum is Greater Than Its Parts

The two data sets work in complementary fashion, with each painting an important and longitudinal part of the consumer shopping behavior picture. They overlap in some areas, such as the ability to provide insights into mean sales per household, but in other ways they are quite different – and importantly, complementary. Because POS data does not allow visibility into the demographics of the individual making the purchase, it provides answers to only a very limited set of questions around data aggregates. For instance, it does not provide an opportunity to correlate sales and share insights to the household level.

The robust data and insights that can be gleaned through panel data are also limited by several important shortcomings. First of all, panel data represents a sample, not a census, and therefore, it is subject to greater variability where small sample sizes are present. Also, the sample of panelists is recruited in an unrestrained manner in the interest of gaining a very large sample, which increases the precision of estimates. As a result, the panelist pool does not represent an accurate demographic cross-section of the U.S. population.

Additionally, panelist scanning behaviors are not perfect. Panelists may completely miss or underreport some purchase occasions, such as immediate-consumption snacks (e.g., chips and soda), purchases intended for out-of-home consumption (e.g., treats or beverages purchased on the way to a party) and purchases of personal and/or "vice" categories (e.g., sexual health, tobacco).

Where underreporting is significant, a full set of rich measures can become inaccurate. For instance, because panel coverage of the convenience channel is weak, verticals that sell a lot of their volume through convenience stores (e.g., beer/wine/spirits) pose additional adjustment complexities. These types of missed-reporting situations cause observed penetration to be unduly low. Where coverage is low, trends become more volatile.

At the other end of the spectrum is overreporting. Because panel participants are more likely than others to try new products, for instance, new product introductions tend to have inaccurately high penetration measurements.

In short, panel data cannot be assumed to reflect the absolute sales levels that are captured through POS measurement. To account for the discrepancy, IRI routinely makes adjustments to the data. The two noteworthy adjustments are demographic and negative binomial distribution (NBD).

Bringing the Numbers into Line

Mentioned above, the wide-net approach to recruiting panelists for NPC results in a panel pool that is large, but not representative of the U.S. population. IRI corrects this misalignment through an algorithm known as iterative proportional fitting (IPF). Simply stated, the methodology allocates more weight to demographically underrepresented panelists and less weight to overrepresented panelists. The resultant adjusted data is an accurate reflection of the joint probability distribution of U.S. demographics, including age, income, household size and ethnicity.

Rising to the Next Level of Accuracy: NBD

With demographic adjustments complete, IRI moves to the next level of accuracy through the use of a complementary process known as NBD adjustment. NBD is a widely used probability distributional model that aligns observed count data, or data that counts rather than ranks. Examples of count data include the frequency of purchases made by a given household, number of children in a household, etc. By applying NBD after panel data are weighted demographically, coverage becomes consistent and year-over-year variability is eliminated. And by using NBD to adjust penetration and buy rate, the internal consistency of the data is maintained.

In order to resolve discrepancies between panel data and POS because of panelists' inconsistent scanning, NBD is aligned to POS targets, actual movement data that are available for most channels, and uses POS data to calculate the panel-to-store coverage needed to align the panel data. For this reason, when confronted with a discrepancy between POS and panel mean sales per household, POS becomes the de facto gold standard.

NBD has been shown to be well-suited for adjusting household purchase data. Because household purchases tend to follow a negative binomial distribution, or a relative frequency of certain outcomes, it stands to reason that when the household randomly fails to report some items, the remaining reported items will still follow that negative binomial distribution pattern. Though the resultant NBD has a diminished mean and variance relative to the true unobserved distribution, it is still representative of the household's behaviors. So, for example, if a distribution of the panelists' heights is known, it can be assumed that the panelists' pants inseam measurements would follow a similar normal distribution.

The NBD adjustment process requires three statistics: current observed penetration, current observed buy rate and current coverage.

The first two statistics are gained directly from panel data.[1] Coverage is gained by comparing total purchases on the panel with the total purchases implied by POS (the resulting ratio of the two). Using these three statistics, the NBD adjustment "fills in" the unobserved panelist statistics such as penetration, buy rate, repeat purchase behavior and brand-switching prevalence, resulting in a clear and complete picture of actual customer behavior.

In the CPG world, IRI's NBD adjustment methodology resolves any dissonance between panel and POS data that is caused by misreporting panelists, ensuring that various household-level statistics, including buy rate and penetration, are consistent with POS mean sales. And, since mean sales is equivalent to buy rate times penetration, the product of the two is perfectly aligned to the mean sales implied by POS. Similarly, statistics such as repeat purchasers and cross-purchasers, post-NBD adjustment, should likewise be consistent with POS.

[1] IRI uses observed penetration and buy rate, though actually any two statistics associated with NBD could serve as alternatives. For example, mean and variance of purchases is popular in the statistics community. Another very common strategy to fitting NBD associated with the observed data is the maximum likelihood approach.

Harmonizing these data sets allows the emergence of a complete and accurate measure of product sales and consumer-level details associated with those sales, bringing to life not just the story of products being purchased, but also the consumers making those purchases and the influencers of those purchase behaviors (see exhibit 2). Discrepancies in POS trends subside, and variation of longitudinal statistics that emanate from small panel sample sizes are greatly diminished. In short, NBD adjustment ensures the granularity of panel data with the accuracy of POS.

IRI is continually working to appropriately address ongoing complexities inherent in NBD, including those related to reduced coverage in the beer/wine/spirits sector and the convenience store geography. Our statisticians are actively testing the fundamental assumptions of zero-inflation and heterogeneity adjustment methodologies to understand the robustness of these models in low-coverage situations and assessing extended versions of NBD that allow coverage to be estimated.

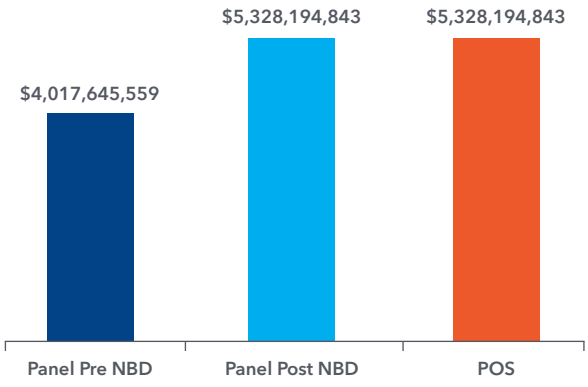
The team is also exploring the utility of new and alternative data sources, such as convenience channel frequent shopper program (FSP) and shipment data, as a means of improving the adjustment of panel data, and analyzing the multivariate time-series aspect of the data to incorporate the past, to inform the future and fuse multiple data sources to better stabilize estimates.

In other words, we are hard at work to ensure that NBD also means "now, better data!"

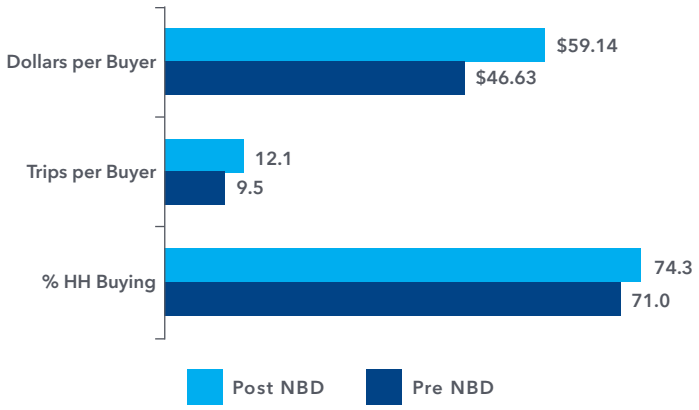
EXHIBIT 2

What's to Come?

Yogurt Dollar Sales



Yogurt Penetration & Buy Rate



Source: IRI POS and IRI Consumer Network™, 52-week data

Appendix

FAQ: What portion of the adjustment is attributed to the increased penetration versus the increase in buy rate?

First, note that there is an increase over the observed penetration and buy rate after the NBD adjustment only when the panel coverage is less than 100 percent. It is possible, but much less common, that panel coverage exceeds 100 percent, in which case the NBD adjustments to penetration and buy rate both experience a decrease relative to observed statistics.

It can be shown that the distribution for the observed purchase occasions when the coverage is p is given by the negative binomial distribution (NBD)

$$P(X = x) = \frac{\Gamma(x + k)}{\Gamma(k)\,(x)!} \left(\frac{\mu_o}{\mu_o + k} \right)^x \left(\frac{k}{\mu_o + k} \right)^k$$

and the distribution for the actual purchase occasions is given by the NBD distribution

$$P(N = n) = \frac{\Gamma(n + k)}{\Gamma(k)\,(n)!} \left(\frac{\mu_o}{\mu_o + kp} \right)^n \left(\frac{kp}{\mu_o + kp} \right)^k$$

where $0<p<1$ is the coverage and μ_o is the observed mean frequency.

The dispersion parameter $k>0$ is specific to the particular category, subcategory or brand. Typically, k is less than one, but can exceed one for household staple categories like soup, toothpaste, soap, batteries, laundry detergent and toilet tissue. In these instances, $1.2<k<1.4$. On the other hand, k is very small for categories exhibiting a great deal of heterogeneity from household to household; for instance, for diapers $k\approx0.09$, baby food $k\approx0.05$ or cigarettes $k\approx0.03$. Some households buy quite a lot within these categories, while others buy none at all.

The mean of the observed purchase occasions is μ_o and the mean of the actual purchase occasions is μ_o/p.

The probability of no observed purchase occasions is simply

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

and the probability of no actual purchase occasions is

$$P(N = 0) = \left(\frac{kp}{\mu_o + kp} \right)^k .$$

It follows that the ratio of actual to observed penetration because of the NBD adjustment is therefore

$$\frac{1 - P(N = 0)}{1 - P(X = 0)} = \frac{1 - \left(\frac{kp}{\mu_o + kp} \right)^k}{1 - \left(\frac{k}{\mu_o + k} \right)^k} .$$

This expression can be explicitly made time-dependent through the equation

$$\frac{1 - P(N = 0)}{1 - P(X = 0)} = \frac{1 - \left(\frac{kp}{\mu_o T + kp} \right)^k}{1 - \left(\frac{k}{\mu_o T + k} \right)^k} .$$

Using the expression for sales = (total # of households × penetration) × buying rate and letting B_a and B_o be the adjusted and observed buy rate respectively and S_a and S_o be the adjusted and observed sales, we find the ratio of adjusted sales to observed sales

$$\frac{\mu_a}{\mu_o} = \frac{1}{p} = \frac{1 - P(N = 0)}{1 - P(X = 0)} \frac{B_a}{B_o}$$

and thus the ratio of adjusted buy rate and observed buy rate is

$$\frac{B_a}{B_o} = \frac{1 - P(X = 0)}{p(1 - P(N = 0))} = \frac{1}{p} \frac{1 - \left(\frac{k}{\mu_o T + k} \right)^k}{1 - \left(\frac{kp}{\mu_o T + kp} \right)^k} .$$

It can be shown that

$$\lim_{T \to 0} \frac{1 - \left(\frac{k}{\mu_o T + k} \right)^k}{1 - \left(\frac{kp}{\mu_o T + kp} \right)^k} = p$$

and it is easy to see that

$$\lim_{T \to \infty} \frac{1 - \left(\frac{k}{\mu_o T + k} \right)^k}{1 - \left(\frac{kp}{\mu_o T + kp} \right)^k} = 1.$$

Thus for very small time periods represented by T, the adjusted-to-observed ratio of penetrations equals $1/p$, whereas the ratio of adjusted-to-observed buy rates is essentially one. However, when T is very large, the converse is true: the ratio of penetration is unity and the actual-to-observed ratio of buy rate equals $1/p$.

So the answer to the question, "What portion of the NBD adjustment is distributed toward buy rate versus penetration?" can be summarized in the pithy statement, "If the time period is small, it's all penetration. If the time period is large, it's all buy rate."