

Store-Level Weights for Analyses Using InfoScan Data

FCSM Research & Policy Conference

October 24, 2023

Saki Kinney

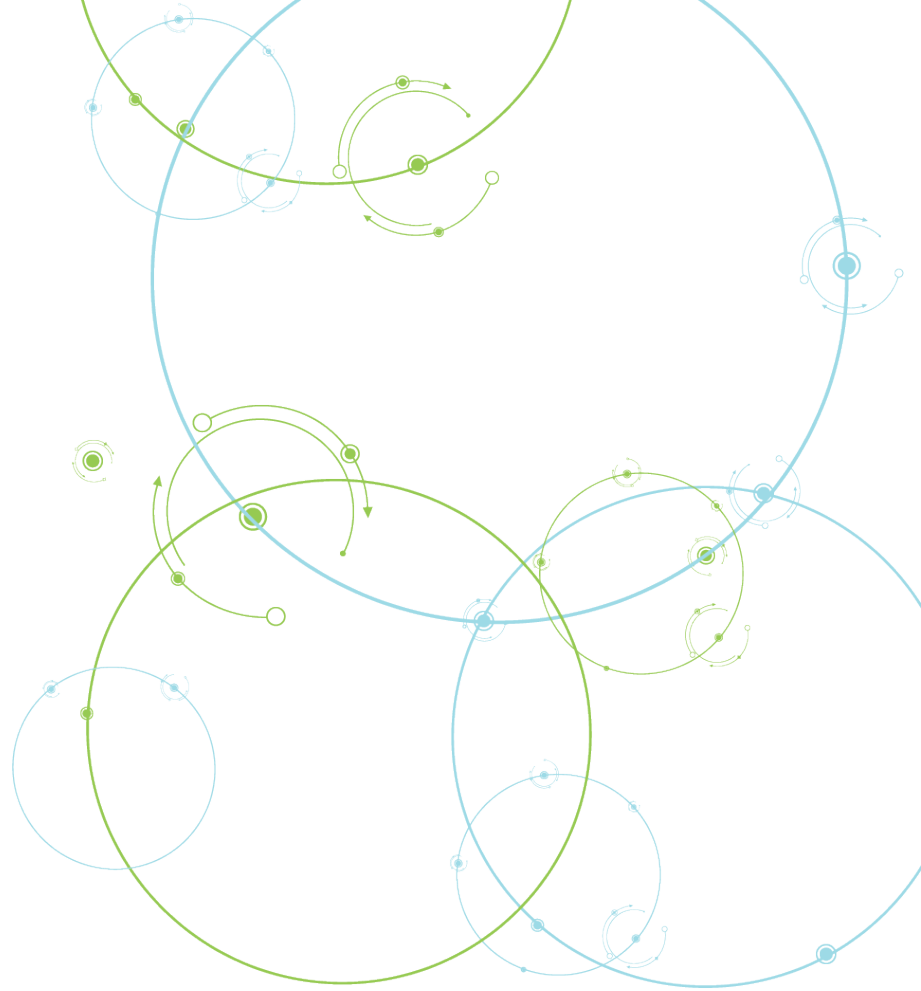


Acknowledgement

- This work was funded under Contract No. GS-00F-354CA between RTI International and USDA's Economic Research Service (ERS).
- The views expressed are those of the authors and should not be attributed to ERS, USDA, or IRI.

Overview

- Motivation
- Nielsen TDLinx population data
- Weight construction
- Use of weights



OmniMarket Core Outlets (InfoScan) sample characteristics

- Retail point-of-sale data: we aggregated all food and beverage products sales to store level
 - Unlike Consumer Network household panel, Circana (IRI) does not provide weights for retail scanner data.
- Nonprobability sample of stores participating in InfoScan 'census' component
 - Predominately comprised of large chain stores
 - We estimate that the sample represents about 20% of stores in the population but 50-60% of retail food and beverage sales (some variation by year)
 - Limited coverage outside of urban areas; varies by state
 - Zero stores in a few states

Why weights?

- Store-level InfoScan data are useful for characterizing the food environment, but the lack of weights prevents calculation of national or subnational sales estimates.
 - E.g. per-capita purchases used to estimate demand, market share of WIC formula purchases
- Weighted estimates will better approximate population values since they account for differences between sample and population.
 - Since smaller stores are underrepresented, they will tend to have higher weights than larger stores.

Population information source: TDLinx

- Nielsen data product that approximates a census of U.S. retail food and beverage stores with \$1 million or more in annual sales

TDLinx	Census of Retail Trade (CRT)
Annual	Every 5 years
ERS already purchases this data	Requires restricted-use access to FSRDC
Store classifications align well with InfoScan	Store classifications align poorly with InfoScan
Over 97% of IRI stores linked by ERS	Additional record linkage effort, low match rate
Good estimates of total store sales	Self-reported total store sales
Does not report food and beverage sales separately	Sales by product category but sparse, quality issues

Weight construction



Thus far weights have been created for 2012-2020 InfoScan data.

1. Prepare store-level data from InfoScan
2. Compute population estimates (control totals) of store counts and sales by geography and channel using TDLinx
3. Use generalized raking procedure to create weights such that weighted counts and sales totals match control totals from step 2
4. Generate replicate weights for variance estimation

InfoScan data preparation

- Aggregate product-level sales to annual store-level food and beverage sales
- Allocate RMA sales to individual stores proportionally based on TDLinx total sales
- Imputation of missing random weight and private label sales
 - Nominal missing data rates, except for drug store PL sales 2013-on
 - TDLinx and Census ACS tract characteristics used as imputation model predictors.

Control total computation

- Merge InfoScan, TDLinx, and American Community Survey tract-level data (eg median income, urbanicity)
- Impute food and beverage sales for non-IRI stores using multiple imputation
- For each geographic region and channel, average total food and beverage sales across imputations
 - Geographic regions = Top 10 metro areas plus remainder by Census Region
 - Top 10 metro areas: New York, Los Angeles, Chicago, Houston, Dallas, Miami, Atlanta, Philadelphia, Detroit, Boston
 - Channels: Club & Mass merchandiser, Drug, Grocery, Convenience, Dollar (2012-16)
- Store count totals are computed directly from TDLinx

Linked InfoScan and TDLinx data, 2012

Channel	IRI Count	TDLinx Count	% of TDLinx
Club store	597	1,236	48%
Mass merchandiser	6,944	7,956	87%
Dollar	7,436	25,361	29%
Drug	19,492	39,947	49%
Grocery	12,641	46,079	27%
Convenience	9,348	148,269	6%
Total	56,458	268,848	21%

Total Sales Benchmarks – 2012

IRI Channel	Total Sales (\$M)		# Stores	
	TDLinx	CRT	TDLinx	CRT
Grocery, Convenience	793,648	1,066,491	194,348	184,178
Drug, Dollar, Club/Mass merch.	579,082	804,530	74,500	73,758
Total	1,372,730	1,871,021	268,848	257,936

Food and Beverage Sales Benchmarks - 2012

Census Division	Census of Retail Trade		Imputed TDLinx	
	Food & Beverage Sales (\$M)	% of Total Sales	Control Total (\$M)	% of Total Sales
New England	26,662	6.8	24,312	6.9
Middle Atlantic	60,211	15.4	53,149	15.0
East North Central	50,332	12.9	47,260	13.4
West North Central	23,944	6.1	20,243	5.7
South Atlantic	77,456	19.8	73,270	20.7
East South Central	17,308	4.4	16,302	4.6
West South Central	38,494	9.9	31,561	8.9
Mountain	24,260	6.2	23,904	6.7
Pacific	71,963	18.4	63,393	17.9
Total	390,629	100.0	353,394	100.0

Food and Beverage Sales Benchmarks – 2012, 2017

IRI Channel	2012 Food & Bev. Sales (\$M)		% of Total Sales	
	CRT	Imputed	CRT	Imputed
Grocery, Convenience	453,144	430,691	71.1	70.2
Drug, Dollar, Club/Mass merch.	184,217	182,648	28.9	29.8
Total	637,361	613,339	100.0	100.0
IRI Channel	2017 Food & Bev. Sales (\$M)		% of Total Sales	
	CRT	Imputed	CRT	Imputed
Grocery, Convenience	570,840	550,729	72.1	69.6
Drug, Dollar, Club/Mass merch.	221,122	240,506	27.9	30.4
Total	791,961	791,234	100.0	100.0

Raking procedure

- Use generalized raking procedure to create weights such that weighted counts and sales totals computed from InfoScan sample match control totals
 - Involves iterative estimation of adjustments until they converge to values satisfying the constraints imposed by the control totals. Adjustments are bounded to limit variance inflation.
- Commonly used to adjust sampling weights when sample characteristics are known to be different from the population, correct for nonresponse bias, and to adjust nonprobability samples
- Process is repeated on the replicate weights

Variance inflation

- Replicate weights were created using bootstrap to allow variance estimation that takes into account weighting procedure
- Variance estimates using weights can and should be higher than what would be obtained under simple random sampling.
 - Overall unequal weighting effects (UWE) for weights range from 3.8 to 4.8 over years
 - Club/Mass Merchandiser and Convenience Stores have UWEs close to 1 while Drug and Grocery by geography can be much higher.
 - Bounds on adjustments/weights are applied during raking to limit variance inflation, but involves tradeoffs with bias, and over-constraining the raking model may prevent convergence.

Limitations

- For both weighted and unweighted analyses, be realistic about the extent to which results generalize to the population.
 - InfoScan is dominated by large chain stores with geographic coverage limitations. Weights reduce but do not eliminate bias.
 - Weights can be used to generate representative estimates at the metro/region geographic area and industry channel. InfoScan coverage limitations prevents calculation of weights for smaller areas.
 - Some weights are less than one.

For more information: <https://www.rti.org/publication/user-documentation> (more current information is forthcoming)



Thank you

Contact: Saki Kinney | email: skinney@rti.org