Sampling and estimation in business surveys: Introduction and overview of basic issues

Li-Chun Zhang

University of Southampton (L.Zhang@soton.ac.uk)

& Statistics Norway (lcz@ssb.no)

Outline

- Business contrast household surveys

- The **unit** problem: **delineation** and **classification**

- Business sample survey

  - basic design: **take-none, -some & -all**

  - two perennial problems: **skewness** & **outliers**

- Some examples

- *More on estimation issues: has to be another time*

Example of business survey: CPI (consumer price index)

- Collect price data from businesses (mostly) and households (e.g. rent); consumption data from households

- Features of sample design & data collection

  - **distinct** sampling frames for businesses & households
  - often **fixed representative** goods for price collection
  - **statistical population $\neq$ sampling frame**
  - **scan data**: bless or curse? **transaction data**? etc.

- What are the **inclusion probabilities?**

Example of business survey: PRODCOM

- Statistical Unit: a list of **products** by EU regulation

- Measure: Amount produced and sold of each product

(NB. Some products also what is retained for use of production)

- Some examples of sampling design:

  - Eurostat: 90% total covered; **cutoff** sample by no. employees

  - ONS: "stratified random sample" (ref. http://ons.gov.uk)

  - Japan: frame based on Census instead of Business Register

Example of business survey: R&D

- Yearly research and development expenditure

  - **rare** characteristics; skewed & truncated distribution
  - **lack** of efficient frame information

- Some methodological elements

  - cutoffs (e.g. 10 employees) or **take-nones**
  - threshold sample of **surprise** units (e.g. Norway)
  - yearly questionnaire *vs.* file-away reporting?
  - measurement interaction with innovation survey?

Business *vs.* household surveys

	Business surveys	Household surveys
Unit	birth, death, drifters	person etc.
(population)	complex organization	(NB. cohabitation)
Frame	Business Register	register/area-frame
Classification	SIC/NACE	household type
	measure-of-size	demographic
Annual	structural, R&D, etc.	mostly
Short-term	turnover, price, etc.	LFS
Measure	continuous & categorical	mostly categorical
	truncated; skewed; outlier	(NB. income, etc.)
Theo. framework	national account	n/a (NB. SAM)

- **Business Register (BR)**

  - **statistical** *vs.* **administrative** register

  - **basis** statistical register: person, business, immobility

- **Business**: engaged in production of goods & services

e.g. enterprise, farm, government department, non-profit organization, etc.

- **Distinguish** between, among others (!),

  - **sampling** unit = unit-in-frame

  - **statistical** unit $\neq$ business unit; e.g. job, goods, service

  - **response/contact** unit for data collection

More on BR: statistical vs. real world

- Regarding **units**

  - **birth, death & drifter**

  - **frame = snapshot of an evolving mass**

- Most important **classification**

  - **NACE**; **measure-of-size** (e.g. no. employees, turnover, etc.)
  - **type** of business units
    - SNA 2008: establishment $\subseteq$ local unit $\subseteq$ enterprise
    - SNA 2008: establishment $\subseteq$ kind-of-activity unit $\subseteq$ enterprise
    - Eurostat: local unit/establishment $\subseteq$ enterprise $\subseteq$ enterprise group
    - Eurostat: local kind-of-activity unit $\subseteq$ kind-of-activity unit $\subseteq$ enterprise

More on BR: progressive nature of measurement

Multiple sources; potential **lag** and **error** of each

e.g. VAT, PAYE, D&B, CISTATS, DEFRA, Companies house for IDBR at ONS

e.g. see Hedlin et al. (2006) for lag-caused coverage errors in IDBR

Illustration: t = statistics time point, $(s, s+6)$ = measurement time points

(n, N) = (sample, population) size, h = stratum, U = business population

	$U(t;s+6)$					$U(t;s+6)$			
$U(t;s)$	$h=1$	$\cdots$	$h=H$	Death	$U(t;s)$	$h=1$	$\cdots$	$h=H$	Death
n_1	n_{11}	$\cdots$	n_{1H}	n^*_{10}	N_1	N_{11}	$\cdots$	N_{1H}	N^*_{10}
$\vdots$		$\vdots$			$\vdots$		$\vdots$		
n_H	n_{H1}	$\cdots$	n_{HH}	n^*_{H0}	N_H	N_{H1}	$\cdots$	N_{HH}	N^*_{H0}
Birth	-	$\cdots$	-	-	Birth	N_{01}	$\cdots$	N_{0H}	-

- *Implications for sampling design and estimation?*

More on BR: classification illustrated (Smith, 2013, Box 5.1, p. 172)

Product	Sales	Input of Materials	Value Added
Cheese	200	150	50
Scallops	60	0	60
Smoked salmon	150	50	120

NACE classification of an establishment

- by sales: NACE **1051** ("diaries ...")
- by value added: NACE **1020** ("fish ...")
- in survey or according to registration on birth: **?**

- Similarly between an enterprise and its local units

- Zhang (2012): *partial-classification* causes *identification error*

Sampling design and measure-of-size: take-none & take-all

- **Probability sampling**: *all* units are *take-somes*
- **Take-nones** and **take-alls**
  - **common feature** of business surveys
  - **requires measure-of-size** as frame information
  - cutoff = take-nones; self-representing = take-alls
  - **cutoff (or purposive) sample** if **no** take-somes
  - **cutoff sampling**: if there are take-somes in design
- e.g. Haziza et al. (2010); Benedetti et al. (2010); Kanub (2011)
- NB. cutoff by design or cutoff due to inaccessibility?

- Impediments

- violation of design-based inference framework
- potential drifters and outliers
- *pseudo*-inference common in practice

- Motivations

- imperfect frame e.g. PRODCOM
- efficiency (Brewer, 1963; Royall, 1970; model-based approach necessary)
- cost for response and process; non-sampling errors
- take-none outliers: effects curtailed by design; bias vs. robustness

Perennial problem (I): skewed & truncated distribution

- **Skewed** distribution: **asymmetry around mean**
- **Truncated** distribution: 0 **or n/a most common**

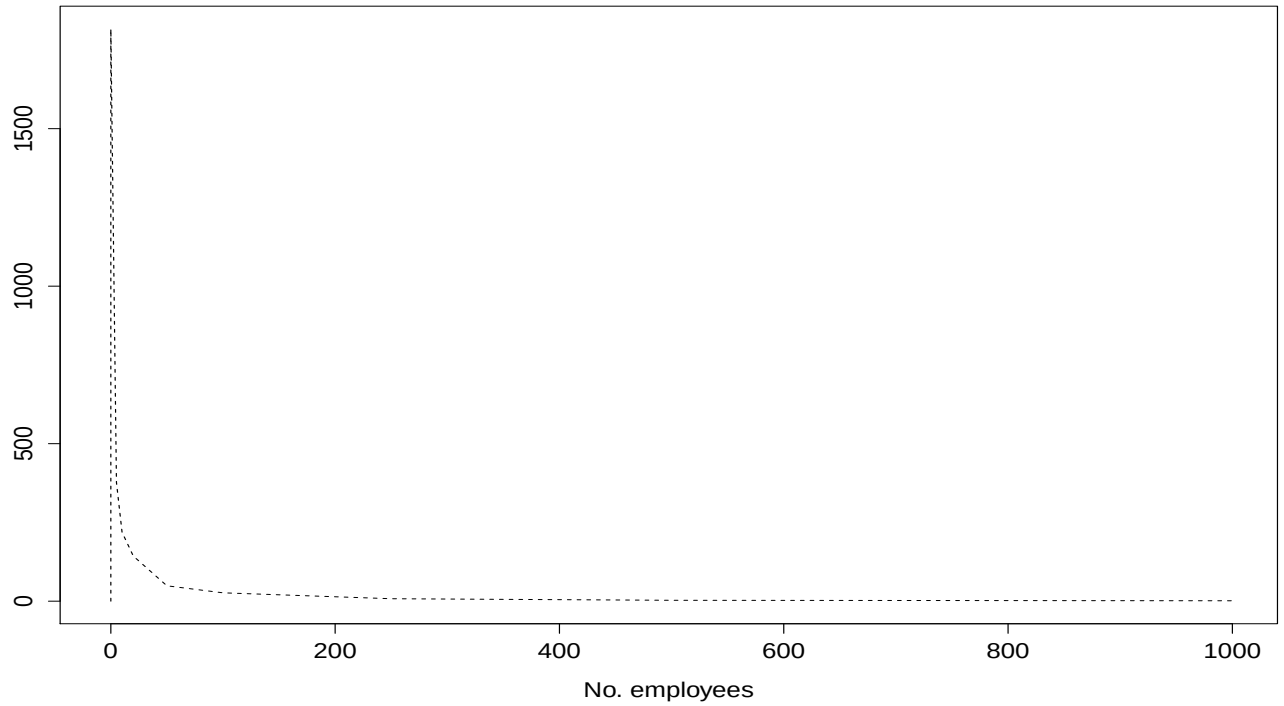


Figure: Distribution of local units (in 1000) by no. employees: IDBR 2008

- **Stratification** & disproportionate sample **allocation**

- **Instrumental approach** to finite-population variance

  • e.g. *ad hoc* remedy for truncated measure-of-size x_i

  1 if truncated; add 1 if not

  • construct **instrumental** measure-of-size d_i

$$\begin{cases} d_i = y_i - \bar{Y} & \text{if truncated } x_i \\ d_i = y_i - Rx_i & \text{if not truncated } x_i \end{cases}$$

design with variance of d_i instead of variance of y_i

Perennial problem (II): outliers

- Outliers are

  - **not** the units with **very large measure-of-sizes**

  - **extreme** despite **comparable measure-of-sizes**

- A **characterization** of outliers (Chambers, 1986)

  - **representative**: correct observations; similar ones may exist out of sample – issue for design & estimation

  - **non-representative**: observations with gross errors; do not reflect true variation in data – issue for editing

Threshold sample of observed outliers

- **Threshold sample** of "surprise units" (Kish, 1965, Sec. 12.6C): the ones that were observed to exceed a given value or **threshold** in the previous survey (or surveys)

- Seems intuitive if these *remain* extreme over time. Still,

  - how large can the threshold sample be allowed for compared to the probability sample? choice of threshold

  - what about the likely large contribution of a "surprise unit" to the **change estimator**?

A result for threshold sample (Zhang and Hagesæther, 2011)

Use of threshold sample is more efficient than not provided

$$\theta < (1 - \phi)\xi f$$

- **sampling fraction** f: incl. both threshold & random samples
- **catch rate** ξ: proportion of threshold sample above the threshold
- **prevalence** θ: proportion of population units above the threshold
- **variance factor** ϕ: of the units below the threshold

Or: high catch rate; low prevalence; small variance factor

NB. prevalence θ must be lower than ξf

Norwegian R&D Survey (NRDS): an illustration

Self-representing, threshold and probability sub-samples of NRDS 2003.

	Number of Units			R&D-Value ($\times 10^6$)	
Sub-sample	Total	Above threshold	Catch rate (%)	Total	Average
Self-representing	1737	558	**32.1**	9685	5.576
Threshold	187	158	**84.5**	993	5.310
Probability	2510	228	**9.1**	1085	0.432

Combined use of threshold-sample design and smooth domain estimation

($\xi = 0.8$)	Summary of domain RE			Number of Domains		
Threshold value	Minimum	Median	Maximum	RE < 1	RE $= 1$	RE > 1
1×10^6	.080	.287	1	50	5	0
5×10^6	.221	.640	1	42	10	3
$\theta = 0.05$	.107	.421	1	49	6	0
$\theta = 0.2$	.086	.317	1	50	5	0

Dealing with potential outliers that can be identified in frame

- Introduce **measure-of-activity** variable

  - require **additional information** to measure-of-size

  - e.g. previous **turnover** from administrative sources

- Form **threshold stratum** by measure-of-activity

  - *across* strata formed by measure-of-size

  - *across* detailed NACE classification

  - sampling fraction in threshold stratum: up to 100% for **efficiency**; or e.g. 50% to allow **sample rotation**

Double stratification by measure-of-size and measure-of-activity

- **Double stratification** = (activity, size)-strata

  - population size $N = N_0 + N_1$; sample size $n = n_0 + n_1$
  - put finite-population variance $(\tilde{S}^2, \tilde{S}_0^2, \tilde{S}_1^2)$ for overall, activity-threshold- and size-stratum, respectively

- **Relative efficiency** of double stratification

$$\mathrm{RE} = \frac{N_0^2(\frac{1}{n_0} - \frac{1}{N_0})\tilde{S}_0^2 + N_1^2(\frac{1}{n_1} - \frac{1}{N_1})\tilde{S}_1^2}{N^2(\frac{1}{n} - \frac{1}{N})\tilde{S}^2}$$

NB. investigate how RE varies with N_0; trial-and-error to choose sensibly

NB. apply instrumental approach to finite-population variance

- **Tripartition**: take-none, take-all & take-somes

- **Instrumental approach & double stratification**

- Supplement sample by **CV-maximum** (e.g. at NACE-3 level)

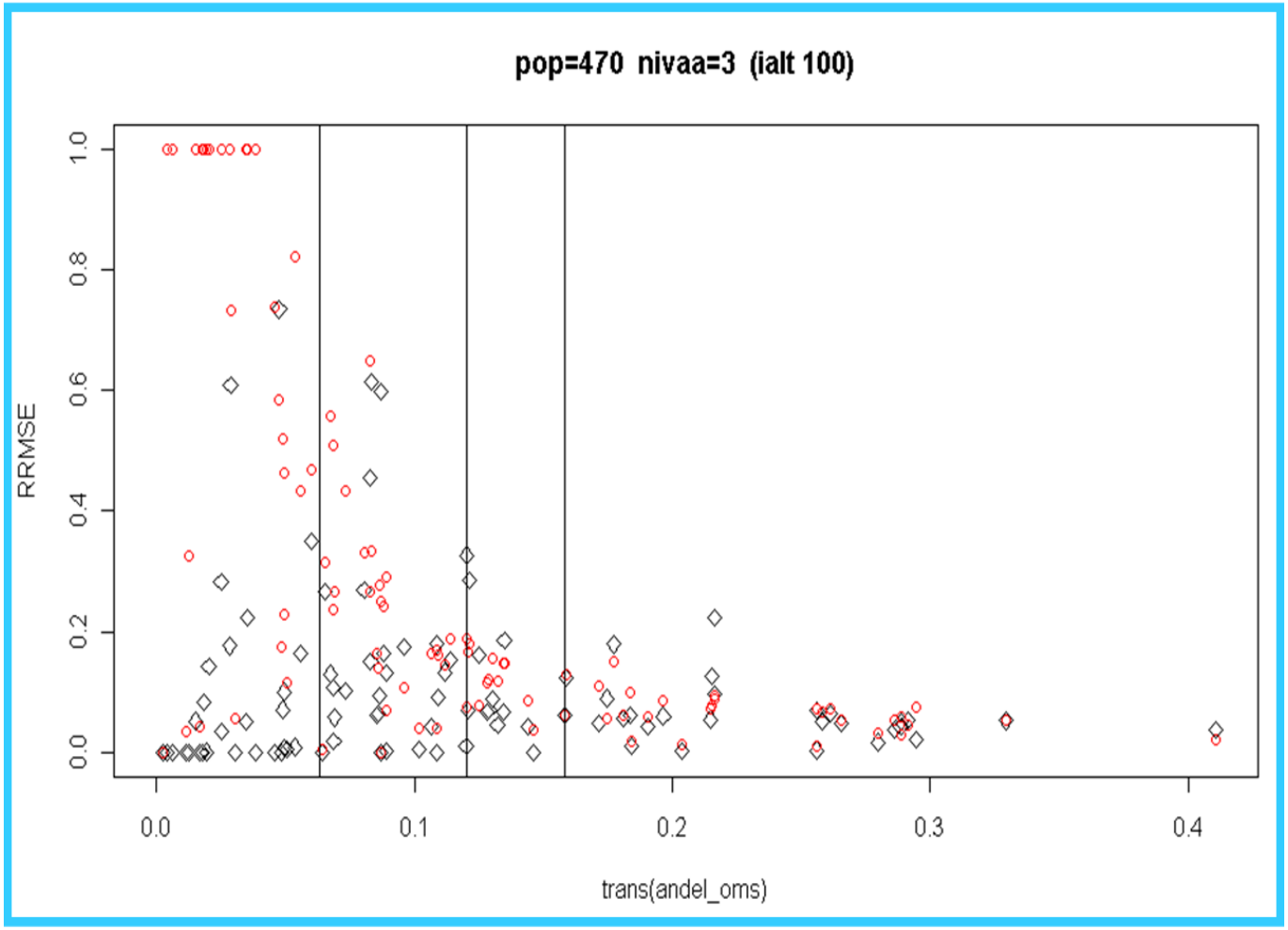
Proportion by activity	$20-100\%$	$10-20\%$	$5-10\%$	$1-5\%$	$0-1\%$
Maximum CV	**0.05**	**0.1**	**0.2**	**0.5**	**0.99**

- Possible **threshold sample** of representative outliers

NB. To be developed and implemented for

- **change** estimation for **short-term statistics**
- **price index** surveys (Zhang, 2010)

Illustration: redesign of Norwegian Structural Business Surveys



Redesign of Norwegian Structural Business Surveys: Single-establishment enterprises, situation 11.01.2012

NACE classification	**Population size**	**Sample size**		RRMSE (%)	
		Before	**After**	Before	After
Travel	11 410	1 259	785	2,46	2,88
Land transport	18 848	1 206	775	3,74	2,31
ICT	16 441	920	475	3,61	3,31
Shipping & Air	2 485	775	604	9,80	4,05
Retail	48 637	3 679	1 754	1,88	2,02
Construction	49 222	1 634	969	5,86	2,80
Service	110 126	3 488	1 703	9,34	3,60
Industry	17 580	2 383	1 266	1,12	1,10
Environmental	1 063	112	108	10,83	4,52
Total	**275 812**	**15 456**	**8 439**	-	-

References

[1] Benedetti, R., Bee, M. and Espa, G. (2010). A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, vol. 26, pp. 651 - 671.

[2] Brewer, K. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, vol. 5, pp. 93 - 105.

[3] Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, vol. 81, pp. 1063 - 1069.

[4] Haziza, D., Chauvet, G. and Deville, J.C. (2012). Sampling and estimation in the presence of cut-off sampling. *Australian and New Zealand Journal of Statistics*, vol. 53, pp. 303 - 319.

[5] Hedlin, D., Fenton, T., McDonald, J.W., Pont, M. and Wang, S. (2006). Estimating the under-coverage of a sampling frame due to reporting delays. *Journal of Official Statistics*, vol. 22, pp. 53 - 70.

[6] Kish, L. (1965). *Survey Sampling*. New York: Wliey.

[7] Knaub, J.R. (2011). Cut-off sampling and total survey error (letter to the editor). *Journal of Official Statistics*, vol. 27, pp. 135 - 138.

[8] Royall, R. (1970). On finite populations sampling theory under certain linear regression models. *Biometrika*, vol. 57, pp. 377 - 387.

[9] Smith, P. (2013). Sampling and estimation for business survey, in *Designing and Conducting Business Surveys*, eds. G. Snijkers, G. Haraldsen, J. Jones and D.K. Willimack. John Wiley & Sons, Inc.

[10] Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, vol. **66**, pp. 41-63.

[11] Zhang, L.-C. and Hagesæther, N. (2011). A domain outlier robust design and smooth estimation approach. *The Canadian Journal of Statistics*, vol. **39**, pp. 147 - 164.

[12] Zhang, L.-C. (2010). A model-based approach to variance estimation for fixed weights and chained price indices, in Carlson, Nyquist and Villani (eds), *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn*, pp. 149-166. Available at officialstatistics.wordpress.com