HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Whither Statistics?
Paradigm shift in Survey Statistics?

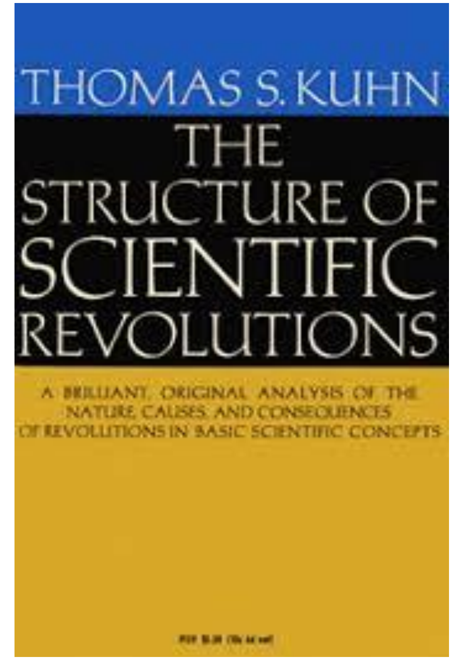
Risto Lehtonen
University of Helsinki

University of Helsinki, Statistics Seminar 28.9.2016

Outline

- Concepts & definitions
- Past debates
- Recent debates
- Conclusion

Paradigm shift - Kuhn



- Thomas Kuhn (1962) *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.
- “A *paradigm* constitutes an accepted way of interrogating the world and synthesizing knowledge common to a substantial proportion of researchers in a discipline at any one moment in time.”
- “Periodically, Kuhn argues, a new way of thinking emerges that challenges accepted theories and approaches.”
(Kitchin)
- This process is called *paradigm shift*

Rob Kitchin (2014) Big Data, new epistemologies and paradigm shifts.
Big Data & Society. April-June 2014: 1–12
DOI: 10.1177/2053951714528481

Gray's paradigm shifts

- Thomas Kuhn:

  "Paradigm shifts occur because the dominant mode of science cannot account for particular phenomena or answer key questions, thus demanding the formulation of new ideas." (Kitchin)

- Jim Gray (computing & software scientist at Microsoft):

  "The evolution of science has proceeded through four broad paradigms."

  "Transitions are founded on advances in *forms of data* and the development of *new analytical methods*. "

  "Science is entering a *fourth paradigm* based on the growing availability of Big Data and new analytics." (Kitchin)

Rob Kitchin Big Data & Society 2014

Table 1. Four paradigms of science (by Jim Gray)

Paradigm	Nature	Form	When
First	Experimental science	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical science	Modelling and generalization	pre-computers
Third	Computational science	Simulation of complex phenomena	pre-Big Data
Fourth	Exploratory science	Data-intensive; statistical exploration and data mining	Now

Compiled from Hey et al. (2009).

Commentary

- Gray's scheme clearly different from the basic Kuhnian recipe

- Main drivers in Gray:

- Technological infrastructure, data infrastructure

- Progress in infrastructures open completely new avenues for scientific research

3 points

- Relevance of paradigms and paradigm shifts approach?

- The notion of paradigms is problematic BUT it can have potential in framing the current debates on Big Data and their consequences

- Anyway, paradigm shifts only can be verified afterwards...

Scenario

Some disciplines adapt and survive

Others do not

Some disciplines not affected (?)

Two questions

Signs of paradigm shift in statistical science?

Signs of paradigm shift in survey statistics?

Statisticians as ACTORS in making scientific paradigms alive?

Kitchin's conclusions

Rob Kitchin Big Data & Society 2014

- Rob Kitchin concludes:

  "There is little doubt that the development of Big Data and new data analytics offers the possibility of reframing the epistemology of science, social science and humanities."

  "Such a reframing is already actively taking place across disciplines."

- Kitchin's slogans

  A fourth paradigm in science?

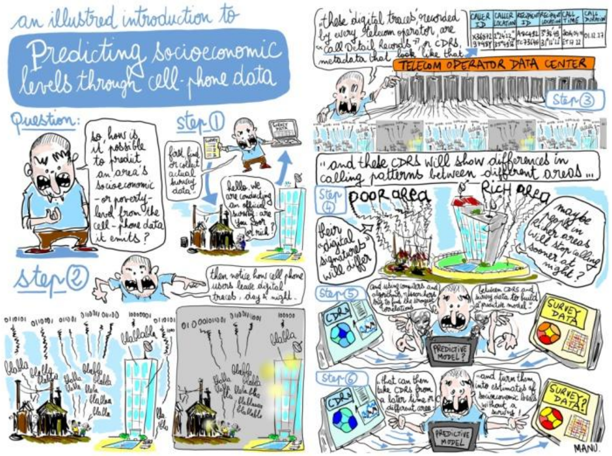
  The end of theory: Empiricism reborn

  Data-driven science

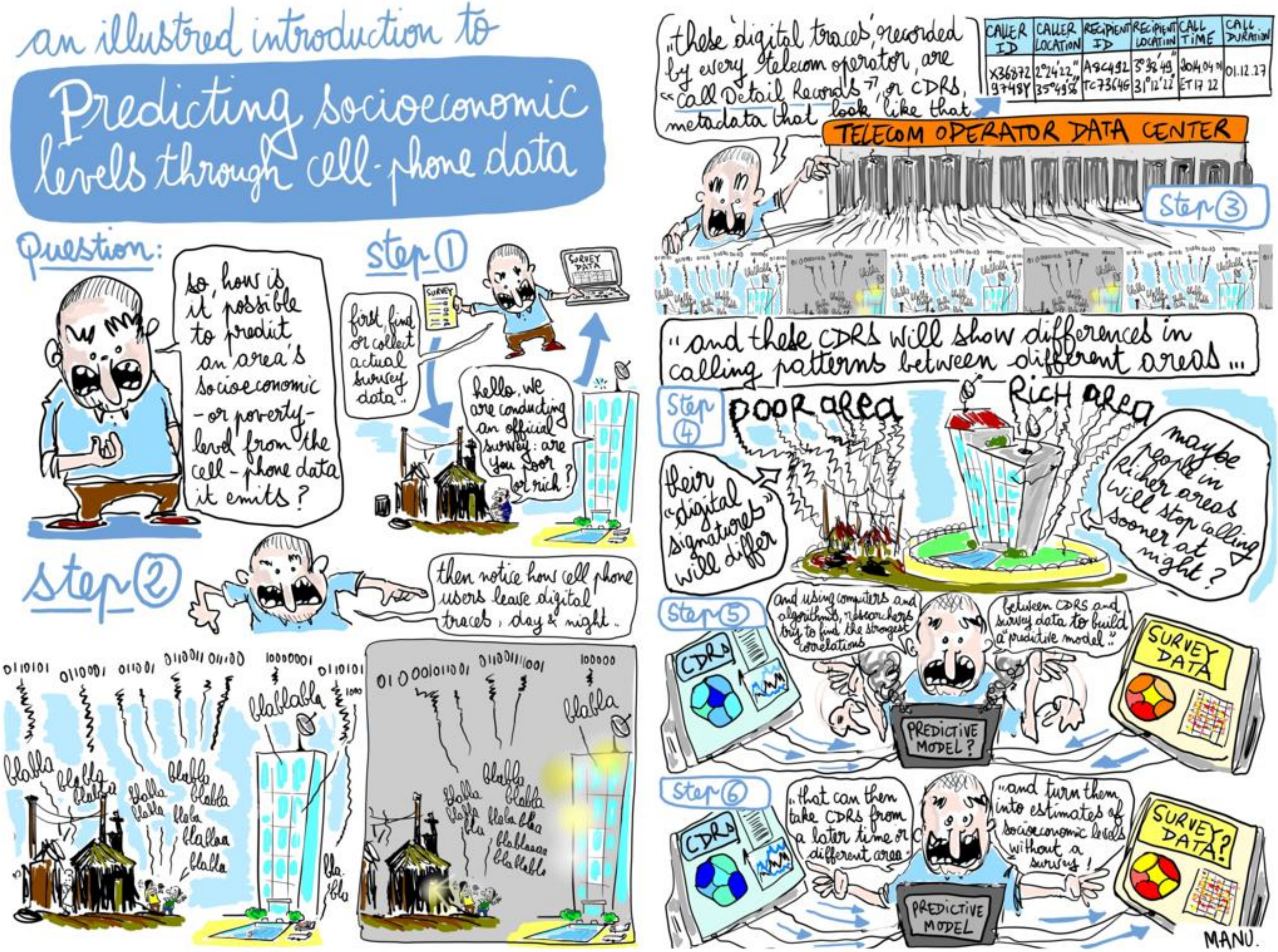
  Computational social sciences

  Digital humanities

Figure 3
Predicting socioeconomic levels through cell phone data.
Credit: Emmanuel Letouzé



Borrowed from: SciDevNet webpages at:
http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html


an illustred introduction to
Predicting socioeconomic levels through cell-phone data
Question:
so, how is it possible to predict an area's socioeconomic - or poverty - level from the cell-phone data it emits?
step 1
first find or collect actual survey data...
SURVEY
SURVEY DATA
hello, we are conducting an official survey: are you poor or rich?
step 2
then notice how cell phone users leave digital traces, day & night...
"these 'digital traces', recorded by every telecom operator, are «Call Detail Records», or CDRS, metadata that look like that
CALLER ID | CALLER LOCATION | RECIPIENT ID | RECIPIENT LOCATION | CALL TIME | CALL DURATION
X36872 97484 | 2°24'22" 35°49'56" | A8C492 TC73646 | 3°38'49" 31°12'22" | 2014.04.01 ET 17 22 | 01.12.27
TELECOM OPERATOR DATA CENTER
Step 3
"and these CDRS will show differences in calling patterns between different areas...
Step 4
POOR AREA
RICH AREA
their "digital signatures" will differ
maybe people in richer areas will stop calling sooner at night?
Step 5
and using computers and algorithms, researchers try to find the strongest correlations
between CDRS and survey data to build a "predictive model."
CDRS
SURVEY DATA
PREDICTIVE MODEL?
Step 6
that can then take CDRS from a later time or different area
"and turn them into estimates of socioeconomic levels without a survey!
CDRS
SURVEY DATA?
PREDICTIVE MODEL
MANU.

Areas of survey statistics: Example

- *Journal of Survey Statistics and Methodology*
- Focus:
- Statistical and methodological issues for sample surveys, censuses, administrative record systems, and other related data

- Topics of interest include:
  - Survey sample design
  - Statistical inference
  - Nonresponse
  - Measurement error
  - The effects of modes of data collection
  - Paradata and responsive survey design
  - Combining data from multiple sources
  - Record linkage
  - Disclosure limitation
  - …and other issues in survey statistics and methodology

  - My focus: Sampling design and inference

Some features

- **Survey statistics**
  - Sub-area of statistical science
  - Developed mainly outside academic statistics
  - Strongly related to official (public) statistic production
- **Focus**
  - Methodologies for survey taking from real populations
  - Statistical inference based on the collected data
- **Past times**
  - Isolation. Debates took place in research communities within the area
- **Recently**
  - Debates spread over discipline barriers
  - More integration to mainstream statistics
  - Increased input from academic statistics
- **Future**
  - New technology & data infrastructures challenge survey statistics and statistical science in general

Tentative scheme for periods

with reference to official statistics

- **Pre-1900**
  Period of complete enumeration

- **From 1900 to 1930’s**
- Debate between complete enumeration, purposive sampling and probability sampling

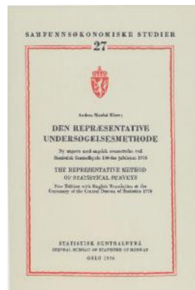
- **From 1930’s to 1970’s**
  Golden era of probability sampling and randomization inference for fixed (finite) populations
  - Equal & unequal probability sampling designs
  - Horvitz-Thompson estimator
  - Stratified multi-stage sampling designs in large government surveys

- **From 1970’s to 2000’s**
  Debate between randomization (design-based) inference and prediction-based (model-based) inference for fixed (finite) populations
  - Reinforcement of design-based inference as the prevailing paradigm in official statistics
  - Model-assisted estimation and calibration weighting

- **2000’s**
  New approaches emerge
  - Random balanced sampling
  - Model-based and Bayesian methods in small area estimation
  - Methods for big data

Early paradigm shift: 1900 – 1930's

- Debate between complete enumeration, purposive sampling and probability sampling
- **Anders Kiær (1897)**
  - "Partial investigation" instead of complete enumeration
  - *Representative method* based on *balanced sampling* with *purposive* (non-random) selection
  - Early features of balanced sampling



Kiær A. (1897) Den repræsentative Uildersøgelsesmethode. Kristiania: Statistisk Sentralbyrå, Sammfunsøkonomiske Studier 27.

- **Arthur Bowley** (1926)
  - Stratified random sampling with proportional allocation
  - *Representative sample* with equal inclusion probabilities
- ISI meeting 1925: Acceptance to both randomization and purposive sampling
- Coexistence lasted until 1934
- The next two decades: Tendency for randomization to become mandatory
- **Jerzy Neyman** (1934) Randomization inference Unequal inclusion probabilities Confidence interval

Details: Ken Brewer (2013), Yves Tillé (2011), J.N.K. Rao (2011), Jelke Bethlehem (2009), Vesa Kuusela (2009)

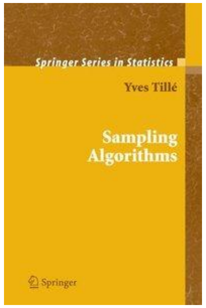
Comment

- **Representative method in survey sampling**
- No uniquely accepted definition of representativeness or representative sampling
- William Kruskal and Frederick Mosteller (International Statistical Review 1979, 1989): nine different definitions of representative sampling found in scientific literature

- Proposal by Hájek (1981): Strategy: a couple of sampling design and estimation design
- Representative strategy: strategy that estimates the totals of auxiliary variables exactly (without error)
- A recent proposal:
- Yves Tillé (2011): Balanced sampling design with Horvitz-Thompson estimator is a representative strategy

Tillé Y. (2011) Ten years of balanced sampling with the cube method: An appraisal. Survey Methodology 37 (215-226, 201).

Balanced sampling



- **Balanced probability sampling**
- Incorporation of auxiliary data in the sampling design
  - Jean-Claude Deville & Yves Tillé (2004)
  - Yves Tillé (2011)

Auxiliary data vector $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})'$

known for all population units $k \in U$

Generate inclusion probabilities π_k for $k \in U$ that satisfy

balancing equation: $\sum_{k \in s} \mathbf{x}_k / \pi_k = \sum_{k \in U} \mathbf{x}_k$ (1)

Computation: Cube method (Deville & Tillé 2004)

NOTE: Probability sample balanced to several auxiliary variables

NOTE: π_k do not depend on sample s

NOTE: difference to balanced purposive sampling

Neyman 1934 landmark paper (Rao 2011)

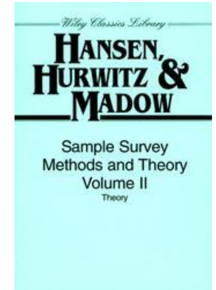
- Relaxing the condition of equal inclusion probabilities
- Theory of stratified random sampling and optimal sample size allocation
- Normal theory confidence intervals for large samples
- Randomization-based inference
- Concept of design unbiasedness
- Demonstration of vulnerability of purposive (balanced) sampling if the model assumptions are violated
- Robustness: no model assumptions

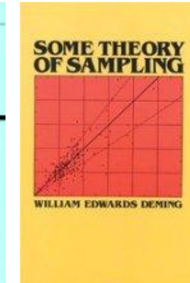
- Neyman showed theoretically and with practical examples the benefits of probability sampling over purposive sampling
- Since Neyman's paper, probability sampling and randomization inference have had a dominant role, especially in the production of official statistics

- Neyman J. (1934). On the two different approaches of the representative method: The method of stratified sampling and the method of purposive selection. *JRSS* 97 (558–606).

From 1930's to 1970's: Golden era

- Probability sampling reached maturity and dominance





Unequal probability sampling with inclusion probability

$\pi_k > 0$ for population element $k \in U$ (Neyman 1934)

Stratified multi-stage sampling designs (Hansen & Hurwitz 1943)

"Golden era" of U.S. Census Bureau. Influence of Morris Hansen

Horvitz-Thompson (HT) estimator (1952) for sample s

$$\hat{t}_{HT} = \sum_{k \in s} y_k / \pi_k \qquad (2)$$

Design unbiased for population total $t = \sum_{k \in U} y_k$

Variance estimation (Yates & Grundy 1952, Sen 1952)

Drivers of development (by J.N.K. Rao 2011)

- Much of the basic sampling theory was developed by official statisticians or those closely associated with official statistics
- Theory was driven by the need to solve real problems
- Often theory was not challenging enough to attract academic researchers to survey sampling
- As a result, university researchers paid little attention to survey sampling
- Few exceptions e.g., Iowa State University under the leadership of Cochran, Jessen and Hartley

Rao J.N.K. (2011) Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. *Statistical Science* 26 (240–256).

Debate on inference: Royall (1970)

- Debate between randomization (design-based) inference and prediction-based (model-based) inference for finite populations

- Royall R.M. (1970). On finite population sampling theory under certain regression models. *Biometrika* 57 (377-387).

- Brewer K. (2013) Three controversies in the history of survey sampling. *Survey Methodology* 39 (249-262).

- Brewer (2013) wrote:

  "It came as a considerable shock to the finite population sampling establishment when Royall (1970) issued his highly readable call to arms for the reinstatement of purposive sampling and prediction-based inference."

Royall's (1970) prediction-based estimator

Consider superpopulation model

$$y_k = \beta x_k + \varepsilon_k \tag{3}$$

$$E(\varepsilon_k) = 0, \quad E(\varepsilon_j, \varepsilon_k) = 0,\ j \neq k, \quad E(\varepsilon_k^2) = \sigma^2 x_k$$

Prediction estimator for population total $t = \sum_{k \in U} y_k$

$$\hat{t}_{PRED} = \sum_{k \in s} y_k + \sum_{k \in U-s} \hat{y}_k = \sum_{k \in U} \hat{y}_k + \sum_{k \in s} (y_k - \hat{y}_k) \tag{4}$$

where $\hat{y}_k = \hat{\beta}_{BLUE} x_k$ and $\hat{\beta}_{BLUE} = \sum_{k \in s} y_k / \sum_{k \in s} x_k$ is the BLU estimator

NOTE: Randomization-based estimator under the same model:

$$\hat{t}_{RAND} = \sum_{k \in U} \hat{y}_k + \sum_{k \in s} (y_k - \hat{y}_k) / \pi_k \tag{5}$$

where $\hat{y}_k = \hat{\beta}_{WLS} x_k$ and $\hat{\beta}_{WLS} = \dfrac{\sum_{k \in s} y_k / \pi_k}{\sum_{k \in s} x_k / \pi_k} = \dfrac{\hat{t}_{HTy}}{\hat{t}_{HTx}}$

Variance estimator

Variance estimator of superpopulation parameter estimate $\hat{\beta}_{BLUE}$ under model (3):

$$\hat{v}(\hat{\beta}_{BLUE}) = \frac{\hat{\sigma}^2}{\sum_{k \in s} x_k}$$

where $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{k \in s} \frac{(y_k - \hat{y}_k)^2}{x_k}$ and $\hat{y}_k = \hat{\beta}_{BLU} x_k$

Optimal (minimum variance) strategy under model (3):

Select the n population elements with largest x_k values

NOTE: Purposive sampling

NOTE: Prediction estimator (4) not necessarily design unbiased (not even design consistent). Problems occur if assumed model (3) deviates from the actual true (unknown) model

NOTE

- Prediction-based (model-based) framework
  - Inference is based on stochastic structure generated by the assumed model
  - Estimators finite population parameters (such as totals) suffer from design bias under model failure
  - Bias declines and efficiency improves under model improvement
  - Much of discussion has focused (and still focuses) on methods to protect against model misspecification

- Randomization-based (design-based) framework:
  - Inference is based on stochastic structure generated by the sampling design
  - Estimators of finite population parameters remain design unbiased under model specification but efficiency weakens under model failure
  - Efficiency improves under model improvement

Debate on inference: Royall (1973, 1976)

- By 1973: Royall had withdrawn the most extreme of his recommendations
- In later articles from 1973 Royall suggested that the chosen sample be *balanced* such that sample moments of x-variables coincide with population moments
- NOTE: Balancing here does not involve random selection!
- Sample balancing formalized the much earlier (e.g. Kiær 1897) notion that samples should be chosen *purposively* to be *representative* and resemble the population in miniature

- Royall R.M. and Herson J. (1973). Robust estimation in finite population I. *JASA* 68 (880-889).
- Royall R.M. and Herson J. (1973). Robust estimation in finite population II: Stratification on a size variable. *JASA* 68 (890-893).
- Royall formalized further the BLU approach in 1976 paper
- Royall, R. (1976). The linear least squares prediction approach to two-stage sampling. *JASA* 71 (657-664).

Balancing and inference revisited

- Recent results on balanced sampling in prediction-based (model-based) inference

- Random balanced sampling with several auxiliary x-variables in finite population sampling
- Nedyalkova D. and Tillé Y. (2012) Bias robustness and efficiency in model-based inference. *Statistica Sinica 22 (*777-794).

- Nedyalkova D. and Tillé Y. (2008) Optimal sampling and estimation strategies under the linear model. *Biometrika* 95 (521-537).

- Important result for combined randomization & prediction inference:

- Under random balanced sampling design, with inclusion probabilities proportional to the standard deviations of the errors of the model (and under certain other conditions), the BLU estimator is the Horvitz-Thompson estimator (Nedyalkova and Tillé 2012)

- **This kind of an optimal (minimum MSE or variance) strategy reconciles the randomization and prediction approaches**

Debate on inference: Reactions to Royall

- “Sampling establishment” reaction
- Cassel C.-M., Särndal C.-E. and Wretman J.H. (1977) *Foundations of Inference in Survey Sampling*. Krieger Publishing Company
- Särndal C.-E., Thomsen I., Hoem J.M., Lindley D.V., Barndorff-Nielsen O. and Dalenius T. (1978) Design-based and model-based inference in survey sampling. SJS 5 (27-52).
- Cassel C.M., Särndal C.-E. and Wretman J.H. (1979) Prediction theory for finite populations when model-based and design-based principles are combined. SJS 6 (97-106).
- Hansen M. H., Madow W. G. and Tepping B. J. (1983) An evaluation of model-dependent and probability sampling inferences in sample surveys. *JASA* 78 (776–793)
- Brewer K.R.W. and Särndal C-E. (1983) Six approaches to enumerative survey sampling. *Incomplete data in survey sampling*, 3, Session VIII, 363-368. Academic Press.
- Theoretical verification and empirical demonstration of problems in prediction approach
- **First (?) attempts to combine ideas of randomization inference and prediction inference**

Brewer and Särndal (1983)

- **Six approaches to sampling inference** by Brewer & Särndal (1983)
- Progressively less dependent on model assumptions

1. The model-based Bayesian approach

- Inference is based on an assumed model and on the specified prior distributions of its parameters
- Sampling design does not play a role

2. The model-based non-Bayesian approach

- As 1 but without priors
- Sampling design does not play a role

3. Robust model-based non-Bayesian approach

- As 2 but uses balanced non-random sampling

4. Probability sampling with modelling

- Model-based or design-based

5. Classical randomization-based probability sampling

6. Inference without exchangeability

Adapted from:
Kangas A. & Maltamo M. (Eds.) (2006) *Forest Inventory: Methodology and Applications*. Springer. (p. 49)

The controversies discussed in Brewer (2013) and a conclusion

- Brewer K. (2013) Three controversies in the history of survey sampling. Survey Methodology 39 (249-262).
- First controversy:
  Kiær's "Representative Method"
- Second controversy:
  The exclusive use of randomization as a means for selecting samples, as advocated by Neyman (1934)
- **Third controversy:**
  Sampling inference: Model-assisted or model-based?"

- Brewer's conclusion:
  "…since there were merits in both the design-based (or randomization-based) and the model-based (or prediction-based) approaches, and that since it was possible to combine them, the two should be used together."
- "The third controversy is still in progress and it is not altogether clear as to how it will turn out."

Comment: Models

- Models have played a central role through the history of model-based (prediction-based) survey statistics

- Models have entered gradually in the design-based (randomization-based) survey statistics machinery
  - Design-based model-assisted survey statistics

Attempts to formalize a combined inference framework – Brewer and Särndal in 2000's

- Brewer K.R.W. (2002) *Combined Survey Sampling. Weighing Basu's Elephants*. Oxford University Press.
- Särndal C.-E. (2011) Combined inference in survey sampling. Pak. J. Statist. 27 (359-370).
- Brewer K.R.W. (2011) Remarks on the paper on "Combined inference in survey sampling" by Carl-Erik Särndal. Pak. J. Statist. 27 (567-572).

- NOTE: At that stage, these two authors seem to reach an agreement on potentials of combined randomization-based and prediction-based inference (although some points for discussion still remains)

Chambers (2011): Proposal for a unified approach

- Chambers R.J. (2011) Which sample survey strategy? A review of three different approaches. Pak. J. Statist. 27 (337-357).
- Abstract of the paper:
- "We review the essential characteristics of the three different approaches to specifying a sampling strategy; the design-based approach, the model-assisted approach and the model-based approach. "
- "We then describe a unified framework for survey design and estimation that incorporates all three approaches, allowing us to contrast them in terms of their concepts of efficiency as well as their robustness to assumptions about the characteristics of the finite population. "
- "Our conclusion is that although no one approach delivers both efficiency and robustness, the model-based approach seems to achieve the best compromise between these typically conflicting objectives."

Other lines of discussion - briefly

- Chambers R. and Clark R. (2012) An Introduction to Model-Based Survey Sampling with Applications. Ray Chambers R. and Clark R. (2012) Oxford Statistical Science Series.

- Rao J.N.K. (2011) Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. Statistical Science 26 (240–256).

- Little R.J.A. (2013) Survey sampling: past controversies, current orthodoxies, and future paradigms. In Lin et al. (Eds.) Past, Present and Future of Statistical Science, COPSS 50th Anniversary Volume X. CRC Press.

- Arjas E. (2011) On future directions in statistical methodologies - Some speculations. SJS 38 (185–194).

Note on the role of auxiliary information

- **Phase 1:** Golden era of probability sampling
- Incorporation of auxiliary data in the **sampling design**
- Prevailing paradigm in official statistics: Design-based strategies with unequal probability sampling and Horvitz-Thompson estimator
- Use of auxiliary data
  - Stratification
  - PPS sampling
  - Multi-stage designs
- Estimation design did not use auxiliary data

- **Phase 2:** 1970's - 2000's
- Incorporation of auxiliary data in the **estimation design**
- Design-based calibration and model-assisted methods
- Unequal probability sampling if needed
- Prevailing paradigm in official statistics

- **Phase 3** (Current / Future)
- Incorporation of auxiliary data in the sampling design by random balanced sampling
- Incorporation of aux.data in the estimation design by calibration and model-asststed methods
- Current discussion: Ex-ante or ex-post use of auxiliary data – or both?

Prevailing paradigm (in official statistics)

- Deville J.-C. and Särndal C.E. (1992) Calibration estimators in survey sampling. *JASA* 87 (376-382).
- Design-based model-free calibration methods
- No explicit model statement

- Särndal C.-E., Swensson B. and Wretman J. (1992) Model Assisted Survey Sampling. Springer.
- Design-based model-assisted methods
- Generalized regression estimation (GREG) using linear fixed-effects model
- NOTE: The role of auxiliary data is crucial in both calibration and GREG methods

Paradigm shift?

- Based on on-going discussion, are there signs of paradigm shift in inference for finite populations?
- What about the role of ”big data”?
- Paradigm shift in official statistics world?

Kott's (2005) proposal for a new paradigm

Randomization-Assisted Model-Based Survey Sampling

Phillip S. Kott

The Model-assisted paradigm presently dominates survey sampling. Under it, randomization-based theory is treated as the only true approach to inference. Models are helpful only when choosing between randomization-based methods. We propose an alternative theoretical paradigm. Model-based inference, which conditions on the realized sample, is the focus of this approach. Randomization-based methods, which focus on the set of hypothetical samples that could have been drawn, are employed solely to provide protection against model failure. Although the choices made under the randomization-assisted model-based paradigm are often little different from those recommended by Särndal et al. (1992), the motivation is clearer. Moreover, the approach proposed here for variance estimation leads to a logically coherent treatment of finite-population and small-sample adjustments when they are needed.

Rod Little: Calibrated Bayes - 1

- Little R.J.A. (2006) Calibrated Bayes: A Bayes/frequentist roadmap. *Amer. Statist.* 60 **(**213–223).
- Little R.J. (2012) Calibrated Bayes: an alternative inferential paradigm for official statistics (with discussion and rejoinder). *Journal of Official Statistics 28* (309–372).
- Little R.J. (2015) Calibrated Bayes, an inferential paradigm for official statistics in the era of big data. Statistical Journal of the IAOS 31 (555–563). DOI 10.3233/SJI-150944

.

Rod Little: Calibrated Bayes - 2

- Little R.J. (1015)
- “In CB, *all* inferences are explicitly Bayesian and hence model-based, but models are sought to yield inferences that are well calibrated in a frequentist sense; specifically, models are sought that yield posterior credibility intervals with (approximately) their nominal frequentist coverage in repeated sampling.”
  …
- ”However, CB models need to incorporate explicitly design features like stratification, weighting and clustering, since models that ignore these features are vulnerable to model misspecification.”

Revisiting Brewer and Särndal (1983): Tentative update

Six (most visible) recent approaches to (descriptive) finite population inference

1. Model-based Bayesian
- Small area estimation (Malay Ghosh)

2. Robust model-based Bayesian
- Donald Rubin, Rod Little
- Calibrated Bayes
- Robustness against model failure
- "Models should be chosen to have good frequentist properties"

3. Model-based (non-Bayesian)
- Ken Brewer
- Combined survey sampling
- Ray Chambers
- "Unified framework"

4. Robust model-based (non-Bayesian)
- Phil Kott
- Model-based design-assisted

5. Probability sampling with modelling
- Carl-Erik Särndal
- Design-based model-assisted Wu & Sitter
- Design-based model calibration
- J.N.K. Rao
- SAE framework

6. Classical randomization based
- Yves Tillé
- Random balanced sampling
- Carl-Erik Särndal
- Design-based model-free calibration

Thank you for your attention!