Data Integration – A New Statistical Frontier for Official Statistics

BY DR SIU-MING TAM

HONORARY PROFESSORIAL FELLOW, UNIVERSITY OF WOLLONGONG

PRINCIPAL – TAM DATA ADVISORY SERVICES

PRESENTED TO

STATS CAFÉ ON DATA INTEGRATION – 16 SEPTEMBER

Outline

- Data integration – why, what and how
- Busting a Big Data myth – Big does not necessarily mean good
- The ABC of Big Data
- Repairing Big Data by mass imputation
- Linkage error correction for integrated data sets
- Conclusion

Data integration –Why

- Direct data collection is an expensive and increasingly unsustainable business model for NSOs
- Declining budgets, increasing demands for more frequent and richer statistics, declining response rates etc. are strong drivers for NSOs to look at reusing/repurposing existing data sets e.g. adm data, Big Data
- The analytical value of integrated data sets will be significantly higher than each of its component data sets, e.g. combining census data with migration data to assess how different migration cohorts settled in Australia

Data integration - What

- Type I Data integration (DI) – Micro level integration
  - defined by the UNECE as "the activity when at least two different sources of data are combined into a dataset. This dataset can be one that already exists in the statistical system or ones that are external sources (e.g. administrative dataset acquired from an owner of administrative registers or web-scraped information from a publicly available website)"
- Type II DI – Macro level integration
  - defined in the statistical literature as borrowing strength from a non-random data set, say B, to improve the statistical value of estimates from a random sample, say A.
- In this talk, I share some cutting edge research undertaken in Australia on using integrated data set to make valid statistical inference for both types of DI
  - Explain the ideas rather than the maths, to suit a broad audience

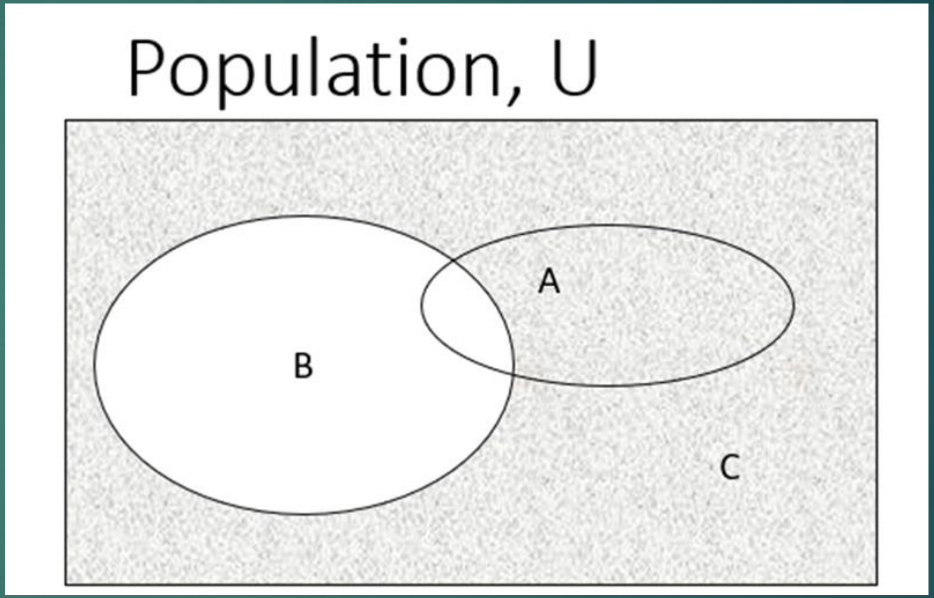
Data Integration - How

- Felligi-Sunter (FS) algorithm for probabilistic matching
- FS steps
  - Compare the linkage variables from a record in Source A with a record in Source B, and use “1” to denote a match, “0” a non-match and “-1” as missing – the string of “1” , “0”and “-1” for each the linkage variables is called an Agreement Vector
  - Use the FS algorithm to calculate the FS weight for this record pair, based on the logarithm of the ratio of “true positive” (aka “m”) and “false positive” (aka “u”) probabilities
  - These probabilities are determined by the Expectation-Maximisation (EM) algorithm
  - Repeat the above for all possible record pairs
  - Choose n record pairs that have the highest FS weights, with n determined based on a priori knowledge, or outputs from the FS algorithm

  - Note - For official statistics, most of the DI takes place with administrative data. I shall use adm data and Big Data interchangeably throughout the talk.

Busting a Big Data myth – Big does not necessarily mean it is good

- Suppose the population U comprises 500K males, and 500K females; and the Big Data B comprises 500K males and 400K females
- The population estimate of the proportion of males in U = 50%
- The Big Data (“sample” fraction f of 90%) estimate = 55.56%
- Why is there a bias in the Big Data estimate? Because the proportion of males included in B does NOT equal to the proportion of females included in B. The difference 5.56% is defined as the response bias, denoted by b (in the next slide)



Busting a Big Data myth – Big does not necessarily mean it is good

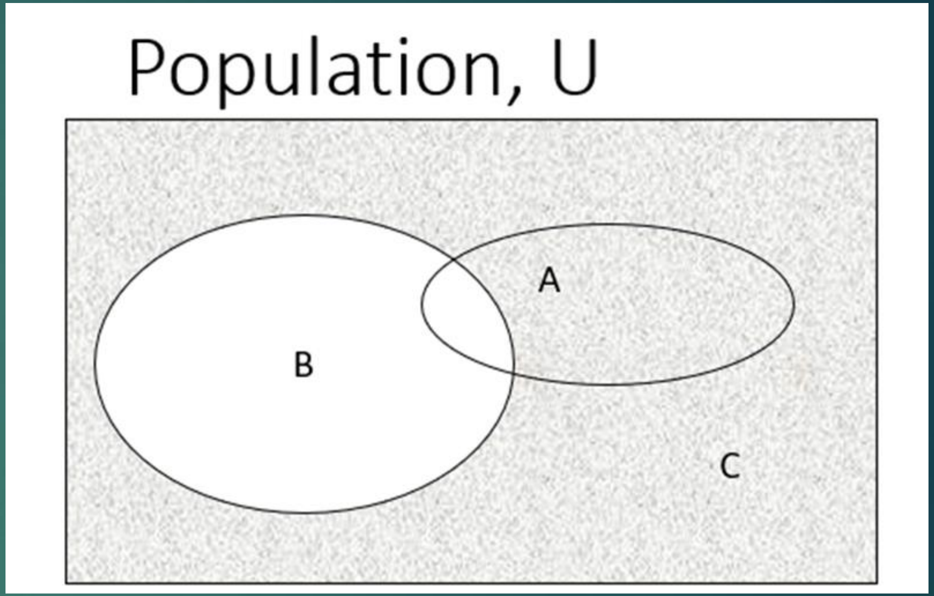
- The inferential value of a Big Data set will be approximately **inversely proportional** to the extent of response bias in the data set – Fundamental Theorem of Estimation Error
- b = difference between the proportion of English speakers and non-English speakers in the Big Data set (response bias)
- f = Big Data "sample" as a proportion of the total
- Effective sample size means having the same Mean Squared Error as a probability sample
- Data from the 2016 Australian Census show the effective sample size is minuscule as compared with Big Data size

Effective sample size to estimate the proportion of English speakers at home, with different values of f and b

Big Data fraction, f	Big Data size	Response bias, b		
		1%	5%	10%
1/10	2,340,189	507	20	5
1/4	5,850,473	3,171	127	32
1/3	7,722,624	5,525	221	55
1/2	11,700,946	12,684	507	127

The ABC of Big Data – Type II DI

- The Big Data set, B, generally suffers from under-coverage, as shown by the missing data in C
- Total (U) = Total (B) + Total (C)
- Total (C) can be estimated by the random sample, A intersects C
- The above simple equation can be rewritten as a calibration equation, to calibrate the weights in A to match population counts in B, C and the total in B
- The above calibration insight allows us to extend the above method to address incompatible definition of response variables in B and A, and non-response in A
- The resultant estimator is called a Regression Data Integrator (RDI), and the estimated total for U is called RDI total. Note that we shall have one RDI total for one response variable.
- For multi-purpose surveys, we therefore have optimum estimates of all response variables. Such a sweet spot cannot be achieved with weighting.



8 & 13 fold improvement in efficiency in ABS Ag survey estimates using the Ag Census as Big Data

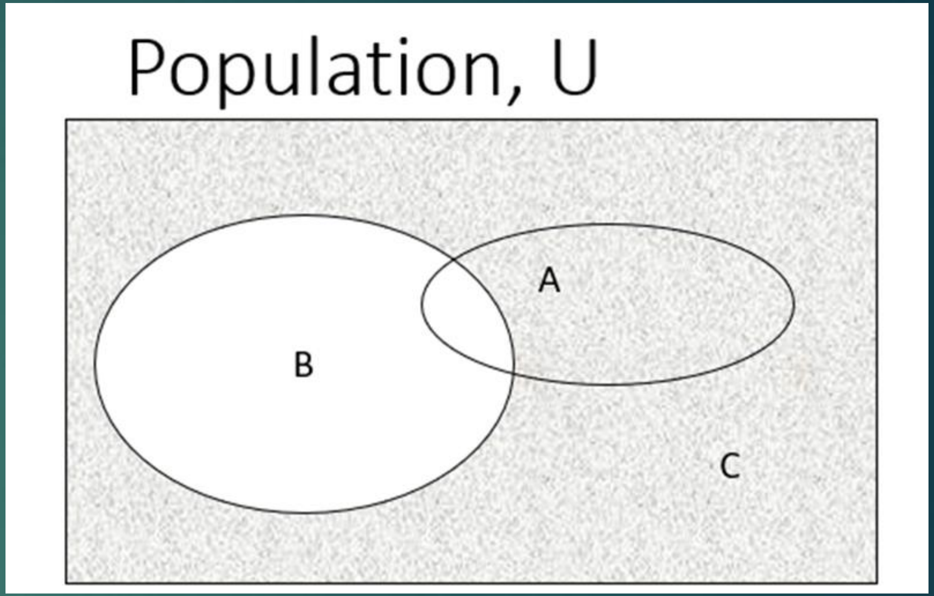
Table: Bias, Variance and Mean Squared Error of Selected Agricultural Commodities

Variable	Estimator from	Bias ($x10^3$)	Var ($x10^9$)**	MSE ($x10^9$)
DAIRY	REACS only (A)	0.00	6.19	6.19
	Agricultural Census only (B)*	-362.45	0	131.37
	(A) and (B)	0.00	0.43	0.43
BEEF	REACS only (A)	0.00	85.00	85.00
	Agricultural Census only (B)*	-2,389.53	0	5,709.86
	(A) and (B)	0.00	6.79	6.79
WHEAT	REACS only (A)	0.00	171.29	171.29
	Agricultural Census only (B)*	-2,043.52	0	4,176.00
	(A) and (B)	0.00	20.83	20.83

* Estimated by the difference between the total from B and the published ABS estimate from the Agriculture Census adjusted for non-response.

Repairing defective Big Data

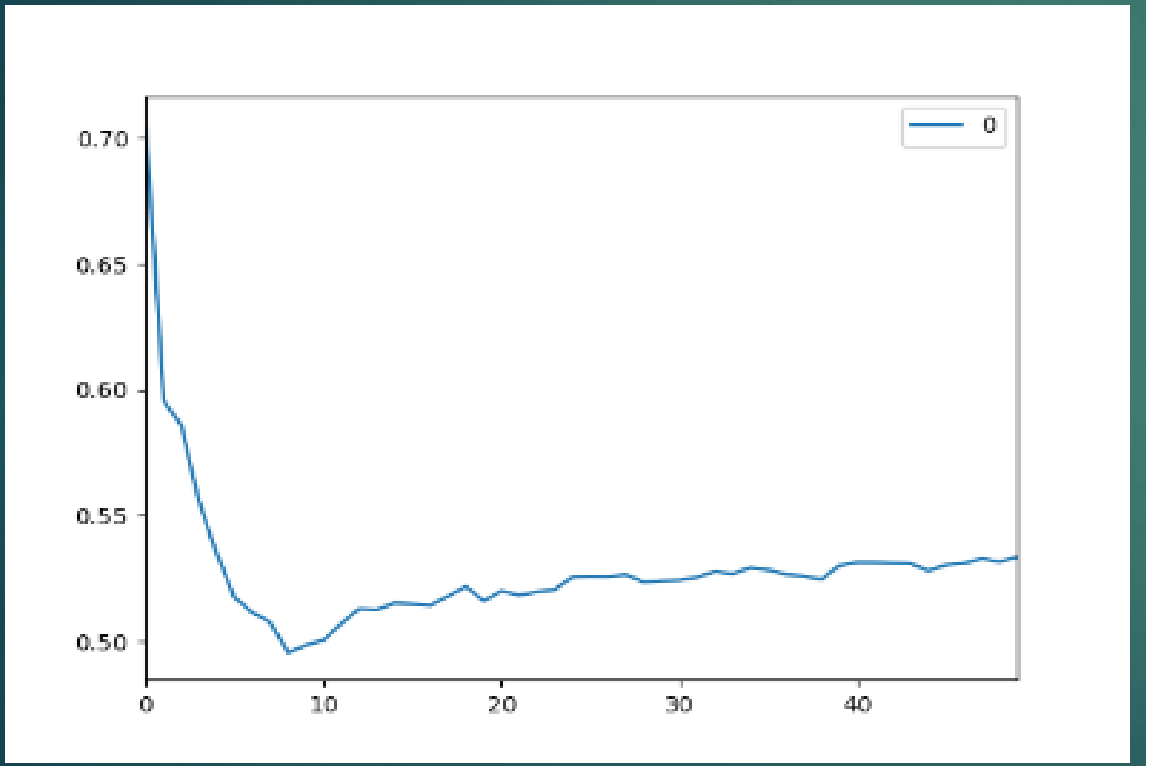
- By mass imputing the missing data in C, using the random sample A to "train" and "test" a machine learning (ML) algorithm
- A PhD student of mine is currently undertaken this research with simulated data with 6 continuous and categorical variables and 6 features, with B being a NMAR data set.
- The idea is this:
  - Train and test KNN (K Nearest Neighbour algorithm) using A to determine optimum K for each response variable. An optimum K is one with the least prediction errors for the testing sample
  - Determine K_C which minimises the prediction error for ALL response variables
  - Prediction for a data point = weighted average of the K_C NN, where the weights are calibrated so that sum of all predicted data points equals to the RDI total



Optimum K Determination for continuous variables

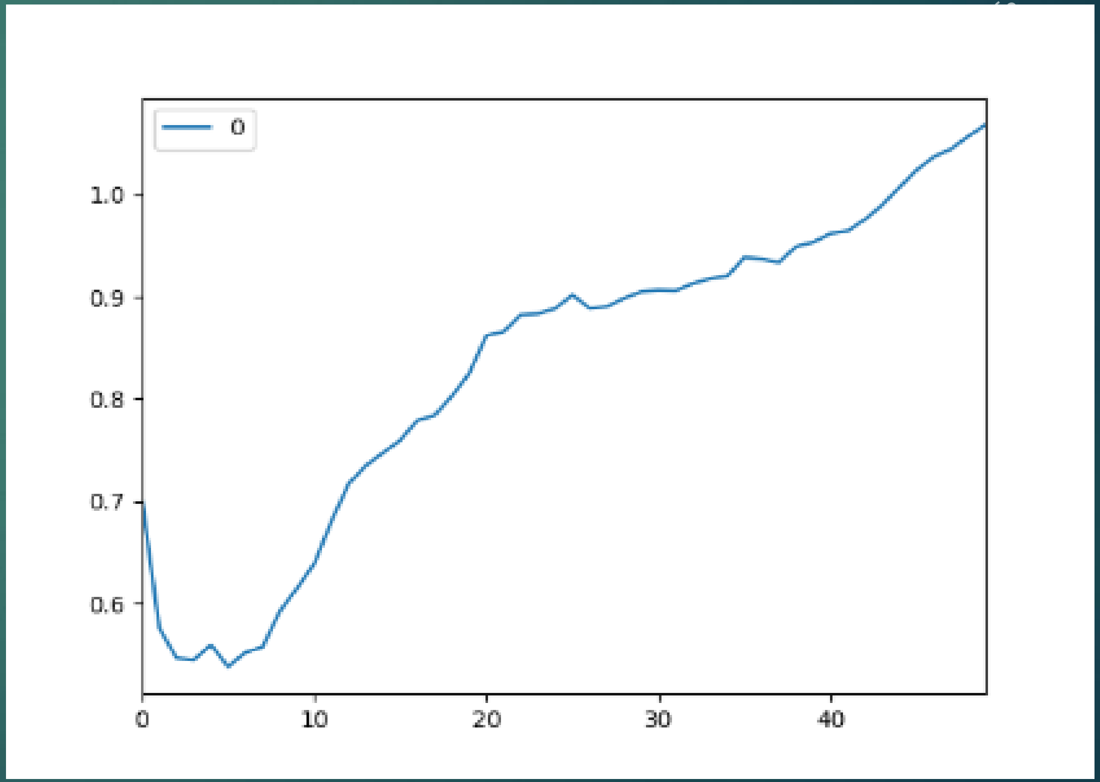
Y1 : Optimum k: 15

RMSE :0.52455



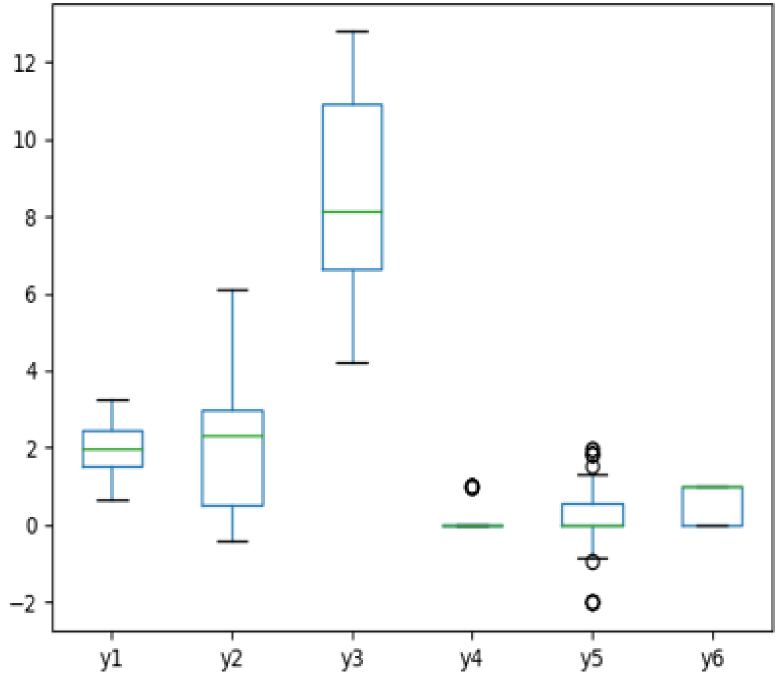
Y2 : Optimum k: 6

RMSE :0.45977

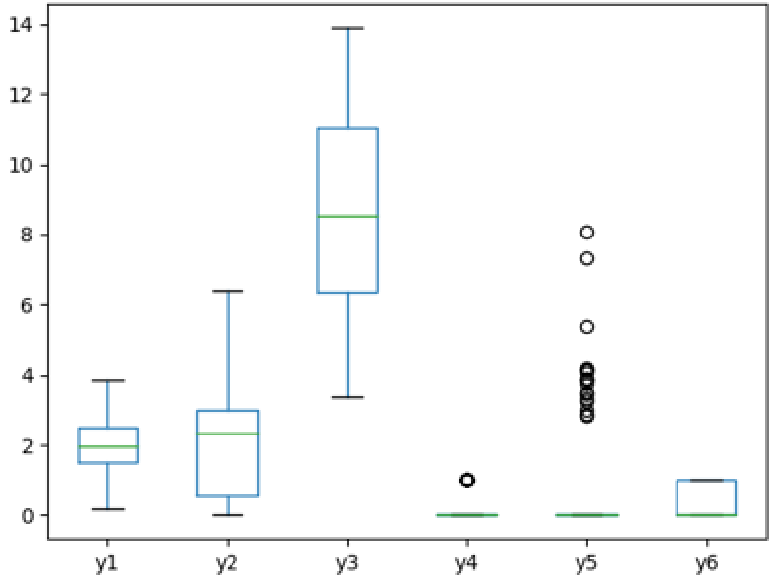


Box Plots

RDI KNN predictors



Original data points



Assumptions for FS algorithm to work – Type II DI

- Linkage variables are statistically independent
- There are no linkage errors
  - However, such errors may exist when sources A and B don't have the same collection standard
  - The errors can adversely and substantially impact analytical inferences
  - Bias correction is therefore paramount. Another PhD student of mine has developed an adjustment method which can apply to all types of estimators, e.g. regression coefficient, contingency tables, variances etc
  - As an example, let's consider the correlation coefficient of a linked data set to look at the relationship between national law test (LSAT), and undergraduate grade point average (GPA)
    - Data came for the Bootstrap book by Efron and Tibshirani (1993).

Correlation Coefficient of a finite population LSAT and GPA data –

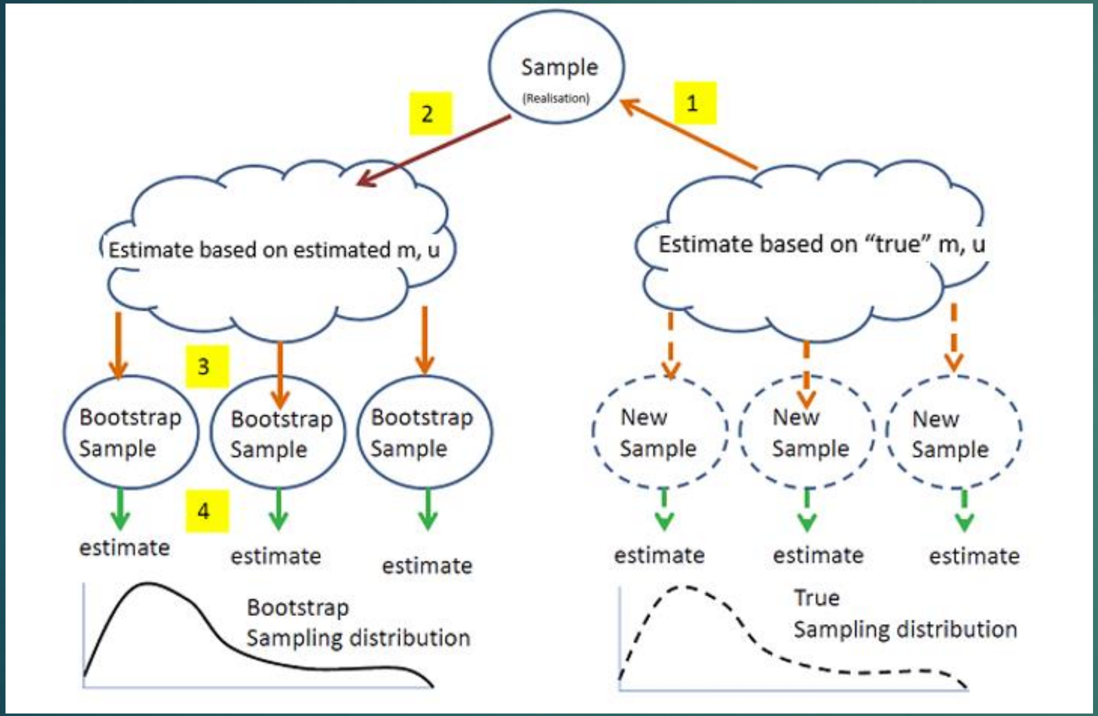
lsat	gpa	lsat	gpa	lsat	gpa	lsat	gpa	last	gpa
606	3.09	627	3.41	617	3.03	637	3.23	575	3.20
628	3.10	608	3.04	572	3.08	542	2.83	603	3.23
615	3.40	578	3.03	635	3.30	579	3.24	546	2.92
547	2.91	545	2.76	575	3.15	553	2.97	512	3.38
587	3.15	645	3.27	666	3.44	641	3.30	555	3.00
611	3.33	586	3.11	570	3.15	599	3.23	580	3.07
594	2.99	560	2.93	595	3.19	646	3.47	632	2.82
614	3.19	631	3.21	615	3.15	598	3.03	570	2.92
581	3.01	597	3.32	558	2.81	704	3.36	605	3.45
662	3.39	621	3.24	611	3.16	477	2.57	622	2.74

Figures highlighted represent incorrect links

	Value
$\hat{\theta}(\chi)$:	0.76
$\hat{\theta}(\chi_1)$:	0.51
$\tilde{\theta}(\chi_1)$:	0.73

- Bias correction table above
  - First row = correct estimate with no linkage errors
  - Second row = estimate from the integrated data set with linkage error
  - Linkage bias-corrected estimate

How did we do it?



- The idea is to use Parametric Bootstraps, i.e. binomial distributions with m and u
  - In simple terms, treat the given integrated data set as the "population"
  - Create bootstrap samples by simulating the Agreement Vectors to create replicated integrated data sets
    - The key point is replicating the Agreement Vectors
  - Difference between the bootstrap and "population" estimates constitute a bias estimate
  - If one uses 1 cycle of replication, it is a single bootstrap; 2 cycles is a double bootstrap etc.

An example of double bootstrap adjustment

- Hormone data and length of time wearing the hormone device (data also from Efron and Tibshirani's book)
- Interest is in estimating the intercept and slope parameters of the regression line

Beta coefficient comparison:

	Intercept	Slope
Non-bias corrected estimators:	31.55	-0.042
Single bootstrap corrected adjusted estimators:	33.42	-0.053
Double bootstrap corrected adjusted estimators:	33.91	-0.056
Monte Carlo estimate of $\theta(F)$:	34.20	-0.057

Conclusion

- Data integration will be the future of official statistics because integrated data sets significantly increase the public value of the data
- Need to be mindful that Type I DI may provide misleading analytical inferences, which could significantly affect policy formulation and evaluation
  - Bias adjusted estimation provides a method to address linkage errors
- Combining Big Data and survey data can significantly improve the statistical efficiency of finite population estimates; and
- also be used to mass impute missing data to correct for under-coverage bias in the Big Data
    - Using RDI and RDI KNN methods

References

- JK Kim and S Tam (2020). Data integration by combining big data and survey sample data for finite population inference. Submitted for publication.
- S Tam and A Holmberg (2020). New sources for official statistics – a game changer for survey statisticians? *The Statistician*, 81, p21-35.
- S Tam, D Tao, J Chipperfield and B Loong (2020). On linkage error correction using the bootstrap for analytical inferences. Subitted for publication.
- S Tam and G Van Halderen (2020).The five V's, seven Virtues and ten Rules of Big Data Engagement for Official Statistics. *Statistical Journal of the IAOS*, 36, p423 -433.
- S Tam and JK Kim (2020). Big Data ethics and selection-bias: An official statistician's perspective. *Statistical Journal of the IAOS*, 34 p577 – 588.
- S Tam, JK Kim, L Ang and H Pham (2021). Mining the new oil for official statistics. In *Big Data Meets Survey Science (Edited by CA Hill et al)*. John Wiley and Sons