INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

SYNTHETIC DATA FOR COMPLEX SURVEY SAMPLING DESIGNS

AN ILLUSTRATION USING DATA FROM THE U.S. ECONOMIC CENSUS

STRATOS Workshop,
Leiden University, September 19, 2024

Jörg Drechsler
Hang Kim (University of Cinncinnati)
Katherine J. Thompson (U.S. Census Bureau)

DISCLAIMER

Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. The Census Bureau has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied. (Approval ID: CBDRB-FY19-B00001).

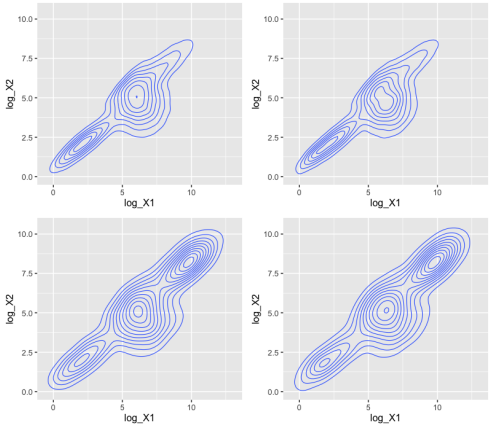
SYNTHETIC DATA FOR THE SOCIAL SCIENCES

- Synthetic data typically used for broadening access to confidential data
- Data typically collected through surveys
- Several properties of survey data make the synthesis process challenging
  - Data contain many attributes
  - Complex relationships and constraints between the variables
  - Samples are drawn using complex sampling designs
- Additional problems arise as the targeted users of the data are very heterogeneous

COMPLEX SAMPLING DESIGNS

- Surveys are never based on simple random samples from the population
- Common design features of surveys
  - Combination of stratification and cluster sampling
  - Often drawn in multiple stages
  - Probability of selection varies between units
  - Probabilities of selection can be data dependent
- Implies that analysts need to take the sampling design into account
- Statistical agencies typically provide survey weights to simplify the analysis task
- In its simplest form survey weights are the inverse of the probability of selection

ILLUSTRATION – EFFECTS OF COMPLEX SAMPLING DESIGNS



Bivariate contour plots from the simulation study displayed on the log scale: the simulated finite population $\boldsymbol{Y}_N$ (top left), a sample $\boldsymbol{Y}_n$ randomly drawn from $\boldsymbol{Y}_N$ (bottom left) which is the input file for data synthesis, a synthetic population modeled using $\boldsymbol{Y}_n$ as input (top right), and a synthetic sample modeled using $\boldsymbol{Y}_n$ as input (bottom right).

OUR USE CASE: BUSINESS MICRODATA

- We address several of the survey related challenges in the context of business data

- Data on businesses are considered highly sensitive
- Access to business data strictly regulated
- Often no data dissemination

- Generating synthetic data as a viable strategy for dissemination of highly sensitive data
- We do not aim at offering any formal privacy guarantees

CHALLENGES WITH BUSINESS SYNTHESIS

- Data are highly skewed with irregular distributions
- Raw data are often subject to edit constraints:
  - Range restrictions for each item, e.g., $L_1 \leq y_1 \leq U_1$
  - Ratio edits for some pairs of items, e.g., $L_{12} \leq y_1/y_2 \leq U_{12}$
- Stratified sampling design needs to be taken into account
- Two types of potential data users
  - Policymakers, trade unions, businesses, etc. mostly interested in design based inference
  - Economists and other academic researchers mostly interested in fitting complex models to the data

CHALLENGES WITH BUSINESS SYNTHESIS

- Data are highly skewed with irregular distributions
- Raw data are often subject to edit constraints:
  - Range restrictions for each item, e.g., $L_1 \leq y_1 \leq U_1$
  - Ratio edits for some pairs of items, e.g., $L_{12} \leq y_1/y_2 \leq U_{12}$
- Stratified sampling design needs to be taken into account
- Two types of potential data users
  - Policymakers, trade unions, businesses, etc. mostly interested in design based inference
  - Economists and other academic researchers mostly interested in fitting complex models to the data

DEALING WITH IRREGULAR DISTRIBUTIONS

- We rely on DP mixture of multivariate normals
- Bayesian version of a normal mixture model
- Flexible tool for modeling irregular multivariate distributions
- The model is given by

$$f(\boldsymbol{Y}_n|\{\Theta_k\}) = \prod_{i=1}^{n}\sum_{k=1}^{K}\eta_k \mathsf{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

ACCOUNTING FOR EDIT CONSTRAINTS

- Truncated version of the DP mixture model to allow accounting for edit constraints
- The likelihood for the data under this model can be expressed as

$$f(\boldsymbol{Y}_n|\{\Theta_k\}) = \prod_{i=1}^{n} c_1(\mathcal{Y}, \{\Theta_k\}) \sum_{k=1}^{K} \eta_k \mathsf{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(\boldsymbol{y}_i \in \mathcal{Y})$$

- We incorporate a data augmentation step (O'Malley and Zaslavsky, 2008) to avoid calculating the normalizing constant and evaluating the truncated normal distribution
- MCMC methods can be employed to generate synthetic data based on this model
- Problem: approach does not account for complex sampling design

USING THE PSEUDO LIKELIHOOD

- Sampling design for business data is informative
- Likelihood in the sample differs from the likelihood in the population
- We propose using the pseudo likelihood approach to account for the sampling design
- The pseudo likelihood is given by

$$\mathcal{L}^{\text{ps}}(\Theta; \boldsymbol{Y}_n) = \prod_{i=1}^{n} f(\boldsymbol{y}_i|\Theta)^{w_i},$$

- The pseudo likelihood for the truncated DP model is

$$f^{\text{ps}}(\boldsymbol{Y}_n|\{\Theta_k\}) = \prod^{n} \left[c_1(\mathcal{Y}, \{\Theta_k\}) \sum^{K} \eta_k \, \text{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(\boldsymbol{y}_i \in \mathcal{Y}) \right]^{w_i}$$

SIMULATION STUDY

- We conduct a repeated simulation design consisting of four steps
  1. Generating a population with distributional features appropriate for business data
     - four variables + measure of size variable
     - 12 ratio edit constraints
  2. Drawing stratified samples mimicking a sampling design typically used for business surveys ($n = 6,300$)
  3. Synthesizing the data using different strategies
  4. Evaluating the quality of the synthetic data
- Steps 2–4 are repeated 400 times

SYNTHESIZING THE DATA

- We look at three different synthesis strategies
  1. *SyntSample1*: uses the truncated DP mixture model but ignores the sampling design
  2. *SyntSample2*: incorporates the sampling design by generating synthetic data conditional on the measure of size variable.
  3. *SyntPop*: uses the pseudo likelihood approach to generate synthetic data of size N
- For each synthesis method we generate $m = 10$ datasets

QUALITY EVALUATIONS – ANALYTICAL VALIDITY

- We evaluate analytical validity from a design based and a model based perspective
- Design based evaluations: compare estimated totals with true totals in the population
- Model based evaluations: compare regression coefficients from the following model

$$\log Y_1 = \beta_0 + \beta_1 \log Y_2 + \beta_2 \log Y_3 + \beta_3 \log Y_4 + \varepsilon.$$

RESULTS – DESIGN BASED ANALYTICAL VALIDITY

		Y_1	Y_2	Y_3	Y_4
Population	θ	177.9	44.7	220.9	716.8
SyntSample1	$E(\hat{\theta})$	1445.3	218.6	1008.6	5819.9
	$V(\hat{\theta})$	232893.9	1885.6	97514.8	8967013.0
	$E(\hat{V}(\hat{\theta}))$	18187.9	148.2	5986.0	570918.9
SyntSample2	$E(\hat{\theta})$	184.0	46.1	223.0	745.3
	$V(\hat{\theta})$	31.1	7.2	64.2	739.7
	$E(\hat{V}(\hat{\theta}))$	12.7	1.3	34.5	599.2
SyntPop	$E(\hat{\theta})$	178.7	46.6	228.1	685.0
	$V(\hat{\theta})$	35.3	4.8	60.7	588.4
	$E(\hat{V}(\hat{\theta}))$	46.5	5.4	53.2	569.8

unit: 10^6 for point estimates and 10^{12} for MSE and variance estimate

RESULTS – MODEL BASED ANALYTICAL VALIDITY

$$\mathrm{E}(\log Y_1) = \beta_0 + \beta_1 \log Y_2 + \beta_2 \log Y_3 + \beta_3 \log Y_4$$

		β_0	β_1	β_2	β_3
Population	θ	0.91	-0.08	0.19	0.69
SyntSample1	$\mathrm{E}(\hat{\theta})$	0.80	0.03	0.22	0.60
	$\mathrm{MSE}(\hat{\theta})$	1.29	1.12	0.15	0.89
SyntSample2	$\mathrm{E}(\hat{\theta})$	0.88	-0.06	0.19	0.68
	$\mathrm{MSE}(\hat{\theta})$	0.32	0.24	0.03	0.24
SyntPop	$\mathrm{E}(\hat{\theta})$	0.90	-0.11	0.22	0.69
	$\mathrm{MSE}(\hat{\theta})$	0.28	0.14	0.10	0.04

unit: 10^{-2} for MSE

APPLICATION TO THE ECONOMIC CENSUS 2012

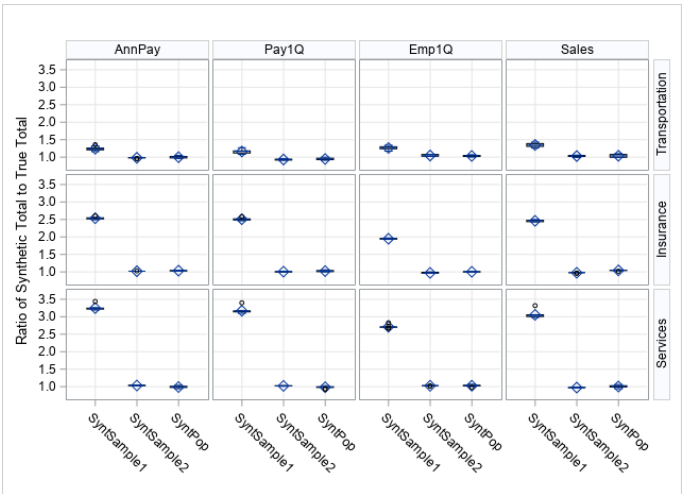
- U.S. Economic Census is not a full census
- Most sectors sampled using stratified simple random sampling
- We select data from three industries

Industry	488330	524210	544110
Sector	Transportation	Insurance	Services
Description	Navigational services to shipping	Insurance agencies and brokers	Offices of lawyers
# sampled units	545	29,682	37,670
% certainty units	0.64	0.51	0.51
Range of weights	1.00 – 1.50	1.00 – 10.00	1.00 – 20.00

APPLICATION TO THE ECONOMIC CENSUS 2012

- We study four variables:
  - Annual Payroll (AnnPay)
  - 1st Quarter Payroll (Pay1Q)
  - 1st Quarter Employment (Emp1Q)
  - Sales/Receipts (Sales)
- We use the three synthesis methods from the simulation study
- Apply the same ratio edits which are used for the Economic Census
- Some additional range edits are applied to avoid generating unrealistic outliers
- We generate 10 trials of $m = 10$ synthetic datasets
- We compare results to results obtained from the original data

RESULTS – DESIGN BASED VALIDITY



RESULTS – MODEL BASED VALIDITY

$$\log(AnnPay) \sim \beta_0 + \beta_1 log(EmpQ1) + \beta_2 log(Sales)$$

Industry		Truth	SyntSample1	SyntSample2	SyntPop
Transportation	β_0	1.52	**1.77**	1.53	**1.62**
	β_1	0.41	**0.48**	**0.45**	**0.44**
	β_2	0.52	**0.47**	0.50	0.50
Insurance	β_0	0.67	0.65	**0.79**	0.67
	β_1	0.66	**0.60**	0.64	0.65
	β_2	0.56	**0.59**	0.55	0.56
Services	β_0	0.87	**1.01**	**0.82**	0.87
	β_1	0.68	0.67	**0.62**	0.68
	β_2	0.53	0.53	0.55	0.53

SUMMARY DISCLOSURE RISK EVALUATIONS

- We ran disclosure risk evaluations for the simulation study and the application
- Risks for SyntPop approach generally smaller than for SyntSample2
- Problems can arise if the data contain some units that
  - are outliers compared to the rest of the data
  - are very homogenous to each other
- Records will always end up in the same mixture component
- Will be replicated in the synthetic data
- Lessons learned: even fully synthetic data are not free from risks

CONCLUSIONS

- Truncated DP mixture model useful for modeling the complex distributions found in business data
- Important to account for the sampling design
- Pseudo likelihood approach suitable for dealing with this problem
- Additional advantage: sampling design can be ignored when analyzing the synthetic data
- Synthetic data are not free from risk

THANK YOU FOR YOUR ATTENTION

Jörg Drechsler (joerg.drechsler@iab.de)

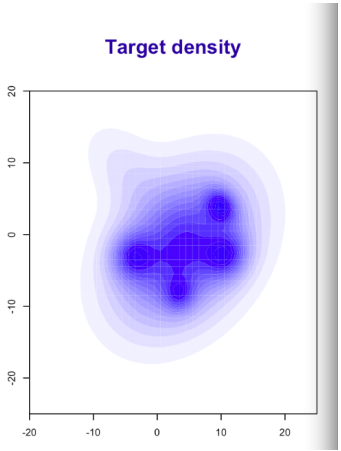
VARIANCE ESTIMATION BASED ON THE SYNTHETIC DATA

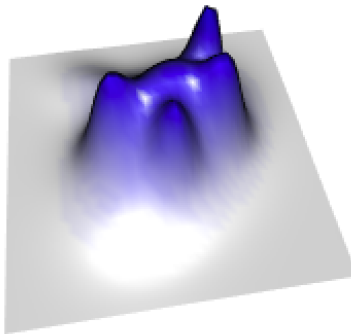
- Estimating the uncertainty straightforward for SyntSample1 and SyntSample2
  - Use variance estimate for fully synthetic data for SyntSample1
  - Use variance estimate for partially synthetic data for SyntSample2
- Appropriate variance estimate not obvious for SyntPop
- We show that variance can be estimated by

$$\hat{V}(\hat{\theta}) = (1 + 1/m)b_m, \text{ with } b_m = \sum_{i=1}^{m} (\hat{\theta}^{(i)} - \bar{\theta}_m)^2/(m-1)$$

- Very attractive for the user since sampling design no longer needs to be taken into account

DEALING WITH IRREGULAR DISTRIBUTIONS





GENERATING SYNTHETIC DATA

- The pseudo likelihood for the data under the truncated DP mixture model is

$$f^{\text{ps}}(\boldsymbol{Y}_n|\{\Theta_k\}) = \prod_{i=1}^{n} \left[c_1(\mathcal{Y}, \{\Theta_k\}) \sum_{k=1}^{K} \eta_k \, \text{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) I(\boldsymbol{y}_i \in \mathcal{Y}) \right]^{w_i}$$

- Introducing a latent mixture component membership indicator $\boldsymbol{z}_n$, the likelihood is

$$f(\boldsymbol{Y}_n|\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}, \boldsymbol{z}_n) = \prod_{i=1}^{n} \left[\, c_2(\mathcal{Y}, \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \, \text{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) I(\boldsymbol{y}_i \in \mathcal{Y}) \, \right]^{w_i},$$

$$f(\boldsymbol{z}_n|\boldsymbol{\eta}) = \prod_{i=1}^{n} \left[\eta_1^{I(z_i=1)} \cdots \eta_K^{I(z_i=K)} \right]^{w_i}.$$

GENERATING SYNTHETIC DATA

- We run a Gibbs sampler for T iterations to obtain draws for all the parameters from their posterior distributions
- To draw m synthetic datasets, we implement the following steps every T/m iteration.

  Let $c_{\text{synt}} = 0$.

  1. Draw z^* from a categorical distribution, $f(\boldsymbol{z}^*|\boldsymbol{\eta}) = \eta_1^{I(z^*=1)} \cdots \eta_K^{I(z_i=K)}$.
  2. Draw $\boldsymbol{y}^*$ from $f(\boldsymbol{y}^*|\boldsymbol{z}^*, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}) \sim \text{N}(\boldsymbol{\mu}_{z^*}, \boldsymbol{\Sigma}_{z^*})$.
  3. If $\boldsymbol{y}^* \in \mathcal{Y}$, let $c_{\text{synt}} = c_{\text{synt}} + 1$, $\boldsymbol{y}_{c_{\text{synt}}} = \boldsymbol{y}^*$ and $z_{c_{\text{synt}}} = \boldsymbol{z}^*$.
- Repeat the process until $c_{\text{synt}} = n_{syn}$.

COMPLETING THE BAYESIAN MODEL

- The stick-breaking representation of a truncated Dirichlet process (Sethuraman 1994; Ishwaran and James 2001) is assumed as a flexible prior distribution for $\{\eta_k\}$,

$$\eta_k = \nu_k \prod_{g=1}^{k-1}(1-\nu_g) \text{ for } k = 2, \ldots, K, \quad \eta_1 = \nu_1,$$
$$f(\nu_k|\alpha) \sim \text{Beta}(1, \alpha) \text{ for } k = 1, \ldots, K-1, \quad \nu_K = 1,$$
$$f(\alpha) \sim \text{Gamma}(a_\alpha, b_\alpha),$$

where the prior mean of the Dirichlet process concentration parameter α is a_α / b_α.

COMPLETING THE BAYESIAN MODEL

- For the other parameters, we assume the following prior distributions:

$$f(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) \sim \mathrm{N}\left(\boldsymbol{\mu}_0, \frac{1}{h_0}\boldsymbol{\Sigma}_k\right), \quad k = 1, \ldots, K,$$
$$f(\boldsymbol{\Sigma}_k|\boldsymbol{\Phi}) \sim \mathrm{InverseWishart}(\zeta_0, \boldsymbol{\Phi}),$$

  where $\boldsymbol{\Phi}$ is a diagonal matrix with diagonal components ϕ_j, which follow $f(\phi_j) \sim \mathrm{Gamma}(a_\phi, b_\phi)$ for $j = 1, \ldots, p$.

- We set all hyperparameters to standard values recommended in the literature:
  - $a_\alpha = b_\alpha = a_\phi = b_\phi = 0.25$
  - $\zeta_0 = p - 1$
  - $\boldsymbol{\mu} = \mathbf{0}$
  - $h_0 = 1$

THE GIBBS SAMPLER: STEP 1

- For each $k = 1, \ldots, K$, update $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\mu}_k$ by drawing from

$$
\begin{aligned}
f(\boldsymbol{\Sigma}_k | \cdots) &\sim \text{InverseWishart}\left(\zeta_k, \boldsymbol{\Phi}_k\right) \\
f(\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k, \cdots) &\sim \text{N}\left(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*\right),
\end{aligned}
$$

$$
\begin{aligned}
\text{where} \quad \zeta_k &= \zeta_0 + N_k, \qquad N_k = \sum_{i=1}^{n_{\text{aug}}} I(z_i = k) w_i \\
\boldsymbol{\Phi}_k &= \boldsymbol{\Phi} + \boldsymbol{T}_k + (N_k h_0)/(N_k + h_0)\left(S_k/N_k - \boldsymbol{\mu}_0\right)\left(S_k/N_k - \boldsymbol{\mu}_0\right)^\top \\
\boldsymbol{T}_k &= \sum_{i=1}^{n_{\text{aug}}} I(z_i = k) w_i \left(\log \boldsymbol{y}_i - S_k/N_k\right)\left(\log \boldsymbol{y}_i - S_k/N_k\right)^\top \\
S_k &= \sum_{i=1}^{n_{\text{aug}}} I(z_i = k) w_i \log \boldsymbol{y}_i, \qquad \boldsymbol{\mu}_k^* = (S_k + h_0 \boldsymbol{\mu}_0)(N_k + h_0) \\
\text{and} \quad \boldsymbol{\Sigma}_k^* &= \boldsymbol{\Sigma}_k/(N_k + h_0).
\end{aligned}
$$

THE GIBBS SAMPLER: STEPS 2 AND 3

- Set $\nu_K = 1$ and update ν_k for $k = 1, \ldots, K-1$ by drawing from

$$f(\nu_k | \cdots) \sim \text{Beta}\left(1 + N_k, \alpha + \sum_{m=k+1}^{K} N_m\right)$$

  and set $\eta_1 = \nu_1$ and $\eta_k = \nu_k \prod_{m=1}^{k-1}(1 - \nu_m)$ for $k = 2, \ldots, K$.

- Update α by drawing from

$$f(\alpha | \cdots) \sim \text{Gamma}(a_\alpha + K - 1, b_\alpha - \log \eta_K).$$

THE GIBBS SAMPLER: STEPS 4 AND 5

- For $j = 1, \ldots, p$, draw ϕ_j from

$$f(\phi_j | \cdots) \sim \text{Gamma}\left(a_\phi + (K\zeta_0)/2,\ b_\phi + \sum_{k=1}^{K} \sigma_{k,j}^{-2}/2\right),$$

  where $\sigma_{k,j}^{-2}$ is the jth diagonal element of $\mathbf{\Sigma}_k^{-1}$.

- For each $i = 1, \ldots, n$, update z_i by setting

$$f(z_i | \cdots) = \eta_{i1}^{*I(z_i=1)} \cdots \eta_{iK}^{*I(z_i=K)}$$

  where $\eta_{ik}^* = [\eta_k \, \text{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)] / [\sum_{g=1}^{K} \eta_g \, \text{N}(\log \boldsymbol{y}_i; \boldsymbol{\mu}_g, \mathbf{\Sigma}_g)]$.

THE GIBBS SAMPLER: STEP 6

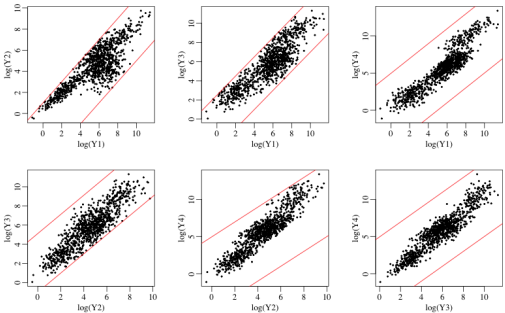
- Sample the auxiliary values for data augmentation. Let $c_{\text{pass}} = c_{\text{fail}} = 0$.

  1. Draw z^* from the categorical distribution, $f(z^*|\boldsymbol{\eta}) = \eta_1^{I(z^*=1)} \cdots \eta_K^{I(z^*=K)}$.
  2. Draw $\boldsymbol{y}^*$ from $f(\boldsymbol{y}^*|z^*, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}) \sim \text{N}(\boldsymbol{\mu}_{z^*}, \boldsymbol{\Sigma}_{z^*})$.
  3. If $\boldsymbol{y}^* \in \mathcal{Y}$, let $c_{\text{pass}} = c_{\text{pass}} + 1$.
     If $\boldsymbol{y}^* \notin \mathcal{Y}$, let $c_{\text{fail}} = c_{\text{fail}} + 1$, $\boldsymbol{y}_{n+c_{\text{fail}}} = \boldsymbol{y}^*$, $z_{n+c_{\text{fail}}} = z^*$, and $w_{n+c_{\text{fail}}} = 1$.

  Repeat the process until $c_{\text{pass}} = n$ (or $c_{\text{fail}} = n$).

GENERATING THE POPULATION ($N = 100,000$)

- Population consists of four variables and a measure of size variable
- Generated from a mixture of normal distributions on the log scale
- Twelve ratio-edit constraints are imposed



DRAWING THE SAMPLE

- We employ stratified simple random sampling without replacement using Neyman allocation
- The certainty strata boundary is determined with the Lavallée- Hidiroglou stratification algorithm (Lavallée and Hidiroglou, 1988)
- Other stratum boundaries determined via *cum-root-f* rule (Dalenius and Hodges Jr, 1959)
- Setting target coefficient of variation to 0.02 resulted in five strata and $n = 6,300$

Table: Stratum sizes and sampling rates for the simulation study

Stratum	1	2	3	4	5
n_h	1410	476	109	286	4019
n_h/N_h	1.000	0.235	0.062	0.061	0.045

QUALITY EVALUATIONS – RISK OF DISCLOSURE

- Difficult to measure the disclosure risk for fully synthetic data
- No link between the original data and the synthetic data
- Looking at risk of re-identification does not make sense
- We evaluate risk of attribute disclosure instead

$$ARD_j = \left|\hat{L}_j - L_j\right| / L_j$$

with L_j largest value in the original data for item j

$\hat{L}_j$ the average of the largest values from each of the m synthetic datasets, i.e., $\hat{L}_j = \sum_{l=1}^{m} \max_i(\tilde{y}_{ij}^{(l)})/m$

- Commonly used measure for risk assessment

RESULTS – DISCLOSURE RISK

- Median of the ARD measure across 400 simulation runs

	Y_1	Y_2	Y_3	Y_4
SyntSample1	2.13	0.54	1.96	1.01
SyntSample2	0.41	0.08	1.16	0.21
SyntPop	2.21	5.59	17.77	0.60

RESULTS – MODEL BASED VALIDITY

- We also look at simple linear regression models

$$Y_i = \beta X_i + \varepsilon_i, \text{ where } \varepsilon_i \sim (0, X_i\sigma^2).$$

- Least square estimate is $\hat{\beta} = \sum_i Y_i / \sum_i X_i$

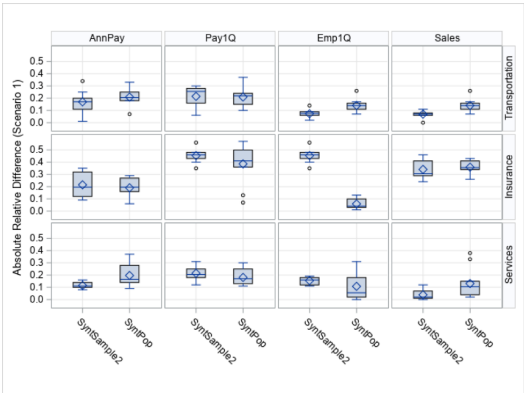
- Such ratios are important measures of economic activity
  - **AnnPay/Emp1Q**: indicator for economic growth/decline over the year if ratio larger/less than 4
  - **Sales/AnnPay**: indicator for industry growth
  - **AnnPay/Emp1Q**: measure for labor conditions within industries.

- Weighted least squares regression are used for SyntSample2 and the original data

RESULTS – MODEL BASED VALIDITY

Industry	Method	AnnPay/Emp1Q	AnnPay/Pay1Q	Sales/AnnPay
Transportation	Truth	68.60	4.12	3.64
	SyntSample1	67.38	4.36	3.97
	SyntSample2	64.62	4.34	3.84
	SyntPop	66.40	4.33	3.81
Insurance	Truth	59.63	3.91	2.82
	SyntSample1	77.61	3.96	2.74
	SyntSample2	62.56	3.97	2.71
	SyntPop	61.65	3.97	2.85
Services	Truth	77.40	4.37	2.80
	SyntSample1	92.65	4.48	2.63
	SyntSample2	78.00	4.42	2.65
	SyntPop	74.56	4.40	2.84

RESULTS – DISCLOSURE RISK



- Risks typically lowest for SynthPop
- Exception: Emp1Q for insurance and services industries.

EXPLANATION FOR THE HIGH RISKS

- Both industries contain a small set of certainty units that
  - have very large values of employment
  - have relatively small ratios of AnnPay to Sales
  - are outliers compared to the rest of the data
  - are very homogenous to each other
- These units will always be grouped in the same cluster, with very small uncertainty
- Ad-hoc fix for now: use micro aggregation within that cluster
- Has negligible effect on utility measures but substantially increases risk measure