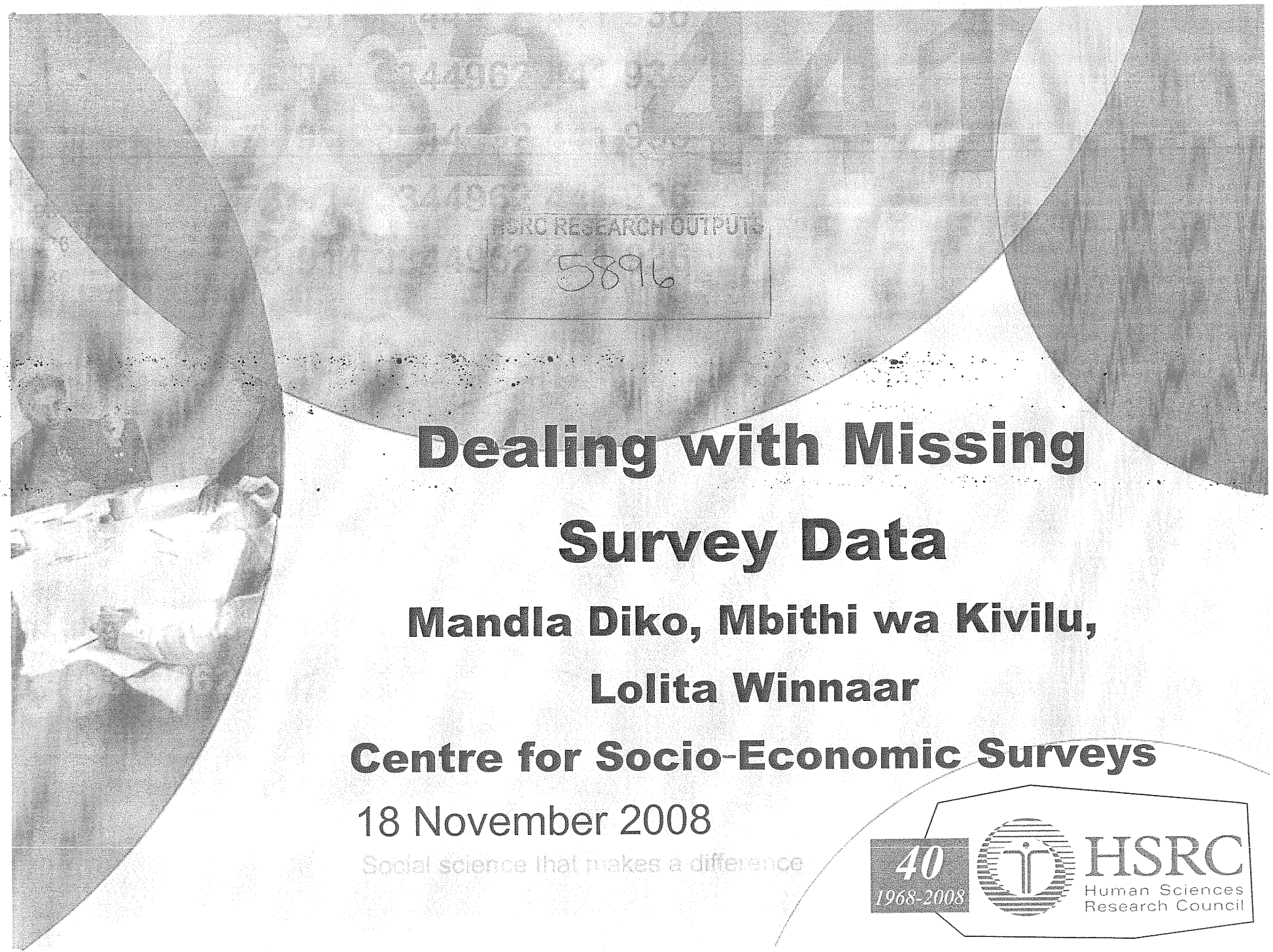

Dealing with Missing Survey Data

Mandla Diko, Mbithi wa Kivilu,

Lolita Winnaar

Centre for Socio-Economic Surveys

18 November 2008

Social science that makes a difference

40 1968-2008

HSRC Human Sciences Research Council

Knowledge Systems

Topics for Discussion

- Introduction-Survey methodologies Project
- What determines the choice of a strategy to deal with missing survey data
- Simulation of data sets with specific population parameters
- Evaluation criteria of strategies for dealing with missing survey data
- Strategies for dealing with missing survey data
- Conclusion

HSRC
Human Sciences
Research Council

Centre for Socio-Economic Surveys

Knowledge Systems

- Provide technical support in research design and statistical analysis and modelling.
- Conduct national surveys (including household surveys)
- The focus of one of our projects "Survey Methodologies" is to identify survey methods/techniques that could be used to enhance the quality of survey data and hence the inferences drawn from such data.

Survey Methodologies Project

- This year our theme for the project is **Missing Survey Data** and strategies of dealing with the problem.
- Missing data are a part of almost all research, and we all have to decide how to deal with it from time to time.
- Data may be missing because equipment malfunctioned, the weather was terrible or people got sick or data was not entered correctly.
- Missing data can occur during data collection and data processing phases of the research process.

Knowledge Systems

What determines the choice of a strategy for dealing with missing survey data?

1. Why data is missing?
2. How much is missing?
3. Distribution of missingness
4. Pattern of missing data

Knowledge Systems

HSRC
Human Sciences Research Council

What determines the choice of a strategy for dealing with missing data?

Knowledge Systems

1. **Why data is missing?**
   - A respondent refuses to answer,
   - A respondent answered "Don't Know",
   - A respondent made a valid skip,
   - A respondent skipped due to interviewer error.
   - Our focus will be on missing data due to "Refuse to answer" and Interviewer error.

What determines the choice of a strategy for dealing with missing data?

2. How much is missing?

- How much is missing versus How it is missing?
- Few values that are not randomly missing are still a problem.
- They will affect the representiveness of the sample and hence generalizability of the results.
- Large amount of randomly missing values are not usually a problem since deleting cases with missing values still leaves a random sub-sample

What determines the choice of a strategy for dealing with missing survey data?

3. Distribution of missingness

3.1. Missing Completely at Random (MCAR)

- **Missing completely at random (MCAR),** we mean that the probability that an observation (Xi) is missing is unrelated to the value of Xi or to the value of any other variables.
- Thus data on family income would not be considered **MCAR** if people with low incomes were less likely to report their family income than people with higher incomes.
- Analysis of MCAR data remains unbiased, we may lose power for our design, but the estimated parameters are not biased by the absence of the data.

What determines the choice of a strategy for dealing with missing data?

Knowledge Systems

3.2. Missing at Random (MAR)

- Data can be considered to be missing at random **(MAR)** if the data meets the requirement that missingness does not depend on the value of Xi after controlling for another variable.

- Persons score on depression is MAR if his/her missing value on depression does not depend on his/her level of depression but to other variables e.g. poverty status.

What determines the choice of a strategy for dealing with missing data?

3.2. Missing not at Random (MNAR)

- Probability of missingness for a variable depends on the variable itself.
- A standard example is individual responses to income questions, where high income people are more likely to refuse to answer survey questions about income and other variables in the data set cannot predict which respondents have high income.

What determines the choice of a strategy for dealing with missing data?

Knowledge Systems

4. Pattern of missing data

We will discuss three such patterns i.e. Univariate pattern, Monotone pattern and Arbitrary pattern.

4.1. Univariate Pattern:

- Missing values occur on one item but a set of (p) other items is completely observed.

4.2. Monotone Pattern:

- If the (j^{th}) item is missing for a unit then the following (j+1,……,p) items are missing as well eg. Longitudinal data.

4.3. Arbitrary Pattern:

- Any set of variables may have missing values.

HSRC Human Sciences Research Council

Methodology

- Datasets were **Simulated** using SAS.
- One thousand samples of size 50 were drawn from a bivariate normal population.
- Each variable had a mean of 125 and standard deviation of 25.
- Correlation between the variables was 0.6 .
- The slope for the regression of Y on X was also 0.6 .

Methodology (cont.)

- **Setting missing data conditions**

  - **MCAR**: a random sample of cases was selected from each sample. The value of Y was made missing for each selected case
  - **MAR**: value of Y was made missing if X>= 140
  - **MNAR**: value of Y was made missing if Y>=140.

Evaluation criteria for strategies of dealing with missing survey data.

- **Bias:** distance between the average of the parameter estimates and the true population value. We want bias to be small.
- **Root mean square error:** average of the distances between the estimates and their true values. We want RMSE to be small
- **Coverage:** proportion of confidence intervals that cover the true population value. Should be 90% or more
- **Distance:** Average of the distances between confidence interval limits should be narrow.

Strategies for dealing with missing data

1. **Deleting Variables**
2. **Deleting Cases**
3. **Mean Substitution**
4. **Single Imputation**
5. **Multiple Imputation**

Strategies for dealing with missing data

Knowledge Systems

1. Deleting variables

- This strategy is applied when the pattern of missing data is **Univariate** or **Monotonic**, and variables with missing data are not part of the analysis,
- Should not be used when missing values are spread throughout the variables and cases.

Strategies for dealing with missing data (cont.)

Knowledge Systems

2. Deleting cases

2.1. Listwise deletion

- A case will be excluded from all analysis if it has at least one piece of missing information.
- This leads to sample size reduction.
- Standard errors for parameter estimates are inflated.
- The level of significance is reduced.
- The statistical test becomes conservative.
- Type 2 error is increased and hence Power is reduced.
- Causes Bias in parameter estimates when MCAR assumption is not met.
- Some effects are exaggerated and others underestimated.

Knowledge Systems

Case Deletion with types of missing data			
	MCAR	MAR	MNAR
MEAN	125.179	**143.75**	**155.847**
RMSE	6.3112	19.7937	31.0601
CI (coverage)	95.4	**20.5**	**0**
Ave dist between CI	27.6683	26.5967	15.3537
STD	24.7352	**21.2015**	**12.3515**
	4.766	5.9898	13.1043
	95.3	91.6	**22.1**
	21.2134	20.7966	11.9861
CORR	0.5731	**0.3183**	**0.343**
	0.1852	0.3907	0.3619
	95.8	**83.1**	**84.4**
	0.718	0.9778	0.9649
BETA	0.6024	0.6054	**0.2243**
	0.2325	0.599	0.421
	94.4	95.5	**44.4**
	0.9842	2.4154	0.7671

Strategies for dealing with missing data (cont.)

2.2. Pairwise deletion

- Excludes cases only if they are missing the data required for a specific analysis.
- Less cases are lost compared to listwise.
- Different parts of the model are based on different sub-samples of participants. ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$)
- Difficult to calculate the degrees of freedom
- May produce a correlation matrix that is not invertible (invertible matrices are necessary conditions for estimating regression equations.

Strategies for dealing with missing data

3. Estimating missing values:

3.1. Mean Substitution

- Substitutes the mean value for the variable in every missing value.
- Variance for the variable is attenuated (reduced).
- Correlations between this variable and other are underestimated.
- Does very bad in all types of missing data (MCAR, MAR, MNAR).
- Regression Substitution. Use Linear regression to predict what the missing score should be on the basis of other variables that are present. We increased sample size and reduced standard error

Knowledge Systems

Knowledge Systems

Mean Substitution with types of missing data			
	MCAR	MAR	MNAR
MEAN	122.574	**137.988**	**150.609**
RMSE	3.106	13.093	25.6308
CI (coverage)	**72.9**	**0**	**0**
Ave dist between CI	7.8206	6.3995	3.9529
STD	**13.7591**	**11.259**	**6.9545**
	11.6122	13.9872	18.1748
	1.5	**0.1**	**0**
	5.6522	4.6252	2.8569
CORR	**0.3207**	**0.2901**	**0.3226**
	0.3074	0.3606	0.3108
	32.8	**32.9**	**33.2**
	0.4945	0.495	0.4921
BETA	**0.1868**	**0.1368**	**0.0974**
	0.4242	0.4727	0.5059
	2.1	**0**	**0**
	0.3025	0.2475	0.1517

Strategies for dealing with missing data

3.2. Single imputation using the Expectation Maximization Algorithm (EM).

- Creates a new dataset in which all missing values are imputed with maximum likelihood values.
- Omits possible differences between multiple imputations.
- Underestimates Standard Errors.
- Overestimate the level of precision
- Tends to increase the level of significance
- Tends to reduce the likelihood of the type 2 error and exaggerate power.

Knowledge Systems

Single Imputation with types of missing data			
	MCAR	MAR	MNAR
MEAN	125.481	124.704	**151.948**
RMSE	3.3744	3.4181	27.011
CI (coverage)	95.9	96.6	0
Ave dist between CI	14.0901	14.3745	7.6355
STD	24.7892	25.2896	**12.9586**
	2.4806	2.6869	12.1396
	95.7	93.5	**0**
	10.1834	10.389	5.3233
CORR	0.5906	0.5913	**0.3991**
	0.097	0.0925	0.2372
	94.4	95.5	**57.3**
	0.367	0.367	0.465
BETA	0.5986	0.6122	**0.2131**
	0.117	0.1208	0.3943
	95.2	93.9	**0.1**
	0.4651	0.4747	0.2767

Strategies for dealing with missing data

3.4. Multiple Imputation

- Imputes a missing value 5 to 10 times to create 5 to 10 datasets.
- Imputations are generated from their predictive distribution with parameters randomly selected from Bayesian posterior distribution with non-informative prior.
- Does statistical analysis separately for each of the 5 to 10 datasets
- Computes pooled estimates of the parameters and the standard errors using 5 to 10 individual solutions.

Strategies for dealing with missing data

Knowledge Systems

3.4. Multiple Imputation (cont.)

- Takes into account differences between multiple imputations and therefore tries to deal with this source of variability.
- Offer reasonable standard errors that do not confer a false sense of precision compared to EM single imputation.
- Does well in MCAR, MAR conditions.

Knowledge Systems

Multiple Imputation with types of missing data			
	MCAR	**MAR**	**MNAR**
MEAN	125.266	124.944	**151.729**
RMSE	2.801	2.7444	26.7628
CI (coverage)	100	100	100
Ave dist between CI	17.7553	18.993	55.4198
STD	24.8417	25.1045	**13.0478**
	1.8138	1.984	12.0087
	100	100	100
	14.9408	14.9964	25.2364
CORR	0.6954	0.6945	**0.4545**
	0.1655	0.1476	0.1819
	99.3	99.3	97.9
	0.7797	0.7457	0.7938
BETA	0.5986	0.6056	**0.2228**
	0.0787	0.0927	0.3811
	99.3	99.5	**1.2**
	0.6535	0.6263	0.3862

Conclusion:

- The aim of the study was to prove that Multiple Imputation was ideal when the distribution method are MCAR and MAR.
- Determining the distribution of missingness is difficult, if not impossible when looking at survey data. Thus use of MI is recommended

Conclusion (cont.):

- **Listwise or Case deletion** is well suited if the data is Missing completely at random (MCAR). Cannot tell most of the time
- This strategy works when the sample size is large.
- From the example using **Mean substitution**; we see that it is a poor strategy for MCAR, MAR as well as MNAR.
- **Single Imputation** works well with both MAR and MCAR however it omits possible difference that occur between imputations.

Conclusion (cont.):

- Multiple Imputation on the other hand, takes into account differences between imputations and therefore tries to deal with the source of variability.
- Pooled estimates of the parameters and standard errors are computed
- From the example it has been proven that Multiple imputation works well if the distribution of missingness is MCAR and MAR.
- None of the strategies work if data is missing not at random (MNAR).

Knowledge Systems

Knowledge Systems

THANK YOU!!

HSRC
Human Sciences
Research Council