PhD project:

Dealing with Imprecision in Data

Andrea Wiencierz

Department of Statistics, LMU Munich

September 6, 2010

Motivation

Possible deficiencies of data often occuring in practice:

1. data containing measurement errors,
2. rounded or heaped values,
3. interval observations or multiple categories,
4. missing values.

Motivation

Possible deficiencies of data often occuring in practice:

1. data containing measurement errors,
2. rounded or heaped values,
3. interval observations or multiple categories,
4. missing values.

Such imperfect data can also be conceived as imprecise data in the sense that it is only known that the true value lies in a subset of the possibility space.

Motivation

Possible deficiencies of data often occuring in practice:

1. data containing measurement errors,
2. rounded or heaped values,
3. interval observations or multiple categories,
4. missing values.

Such imperfect data can also be conceived as imprecise data in the sense that it is only known that the true value lies in a subset of the possibility space.

Plan and goal: Review and compare different theoretical approaches to modelling and analysing imprecise data in order to find suitable methods for statistical analyses in practice.

One Considered Situation: Socio-Economic Survey Data

There are some strategies to correct the effects of the data's deficiencies on the estimates of a statistical model, for example:

- Measurement or classification error models or
- Sheppard's correction for rounded values.

These methods imply assumptions about which kind of imprecision is present in the data and about the prevailing error or rounding processes.

One Considered Situation: Socio-Economic Survey Data

There are some strategies to correct the effects of the data's deficiencies on the estimates of a statistical model, for example:

- Measurement or classification error models or
- Sheppard's correction for rounded values.

These methods imply assumptions about which kind of imprecision is present in the data and about the prevailing error or rounding processes.

Idea: Modify the data collection process instead and use inferential tools for data analysis that can deal with this kind of data.

One Considered Situation: Socio-Economic Survey Data

There are some strategies to correct the effects of the data's deficiencies on the estimates of a statistical model, for example:

- Measurement or classification error models or
- Sheppard's correction for rounded values.

These methods imply assumptions about which kind of imprecision is present in the data and about the prevailing error or rounding processes.

Idea: Modify the data collection process instead and use inferential tools for data analysis that can deal with this kind of data.

Questions:

- How to collect data such that the observations (precise or / and imprecise) are reliable?
- Which data analysis tools can deal with imprecise observations and how do they perform compared to correction methods?

Example: Income Distribution - 1/4

In practice there are different ways of asking for this information:

Example: Income Distribution - 1/4

In practice there are different ways of asking for this information:

- Open question: *"How much do you earn per month on average?"*

  Problems: missing values, rounding, heaping, measurement errors.

Example: Income Distribution - 1/4

In practice there are different ways of asking for this information:

- Open question: *"How much do you earn per month on average?"*

  Problems: missing values, rounding, heaping, measurement errors.

- Question with answer categories: *"In which of the given intervals lies your monthly income?"*

  Loss of information: precise answers impossible, coarse information.

$\ldots,$ $[1250, 1375[,$ $[1375, 1500[,$ $[1500, 1750[,$ $\ldots,$ 7500 or more

Example: Income Distribution - 1/4

In practice there are different ways of asking for this information:

- Open question: *"How much do you earn per month on average?"*

  Problems: missing values, rounding, heaping, measurement errors.

- Question with answer categories: *"In which of the given intervals lies your monthly income?"*

  Loss of information: precise answers impossible, coarse information.

- Open question with option to answer in categories: coarse information.

$\ldots$, $[1250, 1375[$, $[1375, 1500[$, $[1500, 1750[$, $\ldots$, 7500 or more

Example: Income Distribution - 1/4

In practice there are different ways of asking for this information:

- Open question: *"How much do you earn per month on average?"*

  Problems: missing values, rounding, heaping, measurement errors.
- Question with answer categories: *"In which of the given intervals lies your monthly income?"*

  Loss of information: precise answers impossible, coarse information.
- Open question with option to answer in categories: coarse information.

Idea: Income question as a combination of open question with option of answering according to a randomly presented category scheme.

$\Rightarrow$ Less missings, reliable precise or imprecise observations with overlapping intervals.

Example: Income Distribution - 2/4

Data simulation in two steps:

1. sample true income data from an inverse Gaussian with mean $\mu = 10.000$ and skewness parameter $\lambda = 15.000$

Example: Income Distribution - 2/4

Data simulation in two steps:

1. sample true income data from an inverse Gaussian with mean $\mu = 10.000$ and skewness parameter $\lambda = 15.000$
2. modify original data in order to obtain three (partly) imprecise data sets:
   - "coarse mix" data set: 60% of the data are coarsened in using 3 different category schemes
   - "coarse same cat" data set: all observations are categorized with the same category scheme
   - "round mix" data set: 60% of the data are rounded according to 3 different rounding mechanisms

Example: Income Distribution - 2/4

Data simulation in two steps:

1. sample true income data from an inverse Gaussian with mean $\mu = 10.000$ and skewness parameter $\lambda = 15.000$
2. modify original data in order to obtain three (partly) imprecise data sets:
   - "coarse mix" data set: 60% of the data are coarsened in using 3 different category schemes
   - "coarse same cat" data set: all observations are categorized with the same category scheme
   - "round mix" data set: 60% of the data are rounded according to 3 different rounding mechanisms

Data analysis:

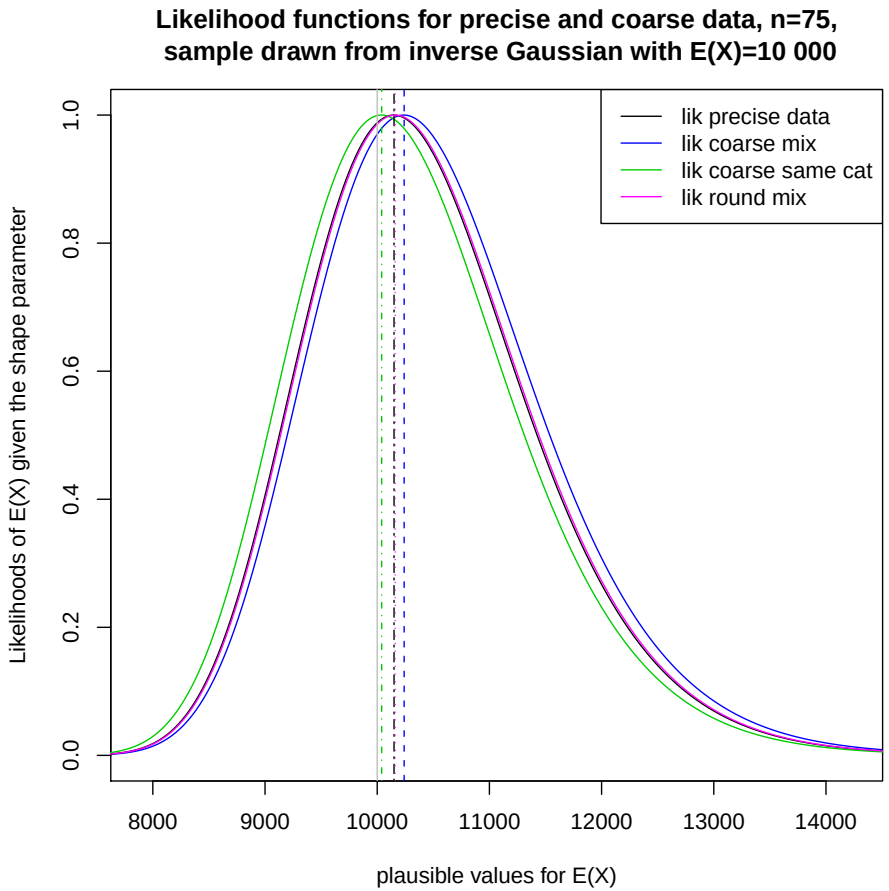
- Parametric method: Comparison of likelihoods
- Nonparametric method: Empirical distributions

Example: Income Distribution - 3/4

Parametric Inference:
Likelihood functions of $\mu = E(X)$ for different data sets with different kinds of imprecision.

$$L_\lambda(\mu, \mathbf{X}) = P_{invG(\mu,\lambda)}(\mathbf{X}),$$

$$\mu \in [X_{min}, X_{max}]$$



Example: Income Distribution - 4/4

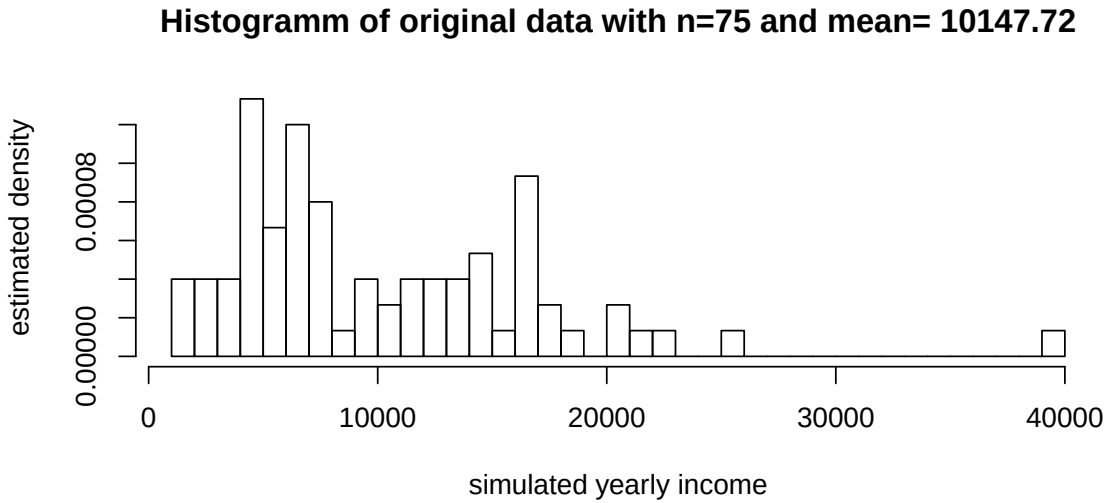
Nonparametric Inference: Empirical distribution.

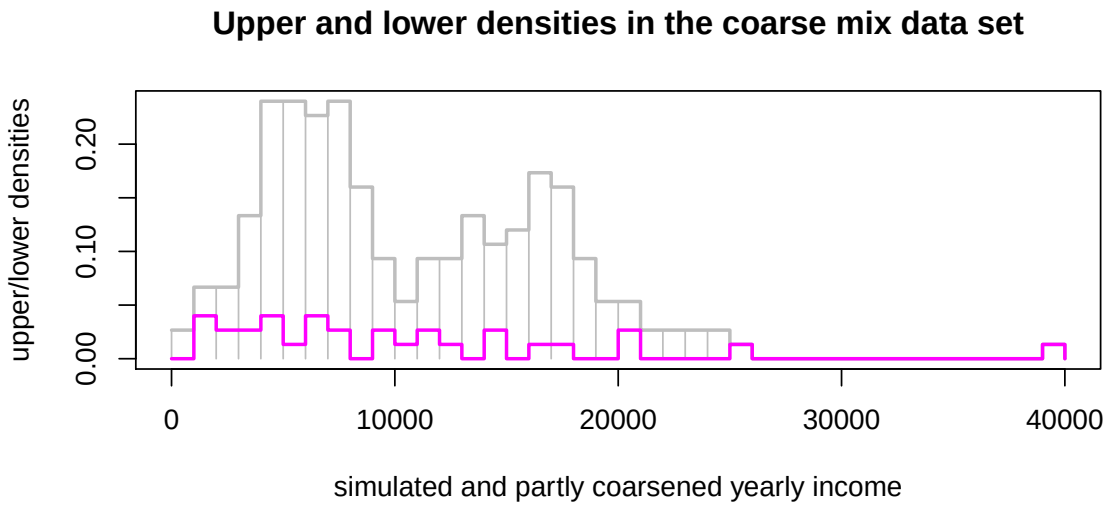
Upper/lower density of A:

$$\overline{f}(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \cap A \neq \emptyset\}}$$

$$\underline{f}(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \subseteq A\}},$$

A_j, $j \in \{1, \ldots, 40\}$ equal sized intervals, partition of $\mathcal{X} = [1, 40000]$.





Theoretical Concepts to be Considered

Conventional Statistics:

- Likelihood based inference
- Correction models for measurement, classification, or rounding errors
- Nonparametric methods

Theoretical Concepts to be Considered

Conventional Statistics:

- Likelihood based inference
- Correction models for measurement, classification, or rounding errors
- Nonparametric methods

Generalized Approaches:

- Random Set Theory
- Fuzzy Set Theory
- Imprecise Probability Theory