# Statistical Methods, part 1

## Module 2: Latent Class Analysis of Survey Error

## Models for measurement errors

Dan Hedlin
Stockholm University
November 2012

Acknowledgement: **Paul Biemer**

Stockholm University

# The most interesting error

- $\hat{Y} - Y$
- That is, the difference between estimate and what you desire (truth, value obtained with the ideal method, population parameter, or however the desired aim is envisaged)
- Is this difference regularly published?
- What is the established term for this difference?

# Why variance?

- Why do we base our statistical theories on concepts like variance, bias etc, which are advanced constructs?

# Why in this course?

- Measurement errors can rarely be observed
- You have to draw conclusions (make inference) about something unobservable
- Models play a crucial role in this inference process
- Quantitative research usually faces measurement errors
- Rather neglected in practical work

# Scenarios

1. True values/gold standard values of a random subsample of sample
2. Dependent or independent measurements of a random subsample of sample
3. One sample with several variables measured once, although with measurement error (most common and least favourable situation)

# Gold standard

- Gold standard (error-free) measurements
  - In-depth reinterviews with probing
  - Assumption: error in second measurement is negligible or relatively inconsequential
  - Record check studies
  - Direct observation (or close to it)
  - However, gold standard has sometimes been shown to be ‘silver standard’ at best. See references in Biemer’s book, page 67

- Direct estimation of measurement bias requires true values or gold standard measurements
- If you have a sample of values with measurement errors, $y_i$, and true scores for each, $\tau_i$, then the difference $y_i - \tau_i$ is like a new variable. The variance of the difference is the same as the variance of $y_i$ (why?)

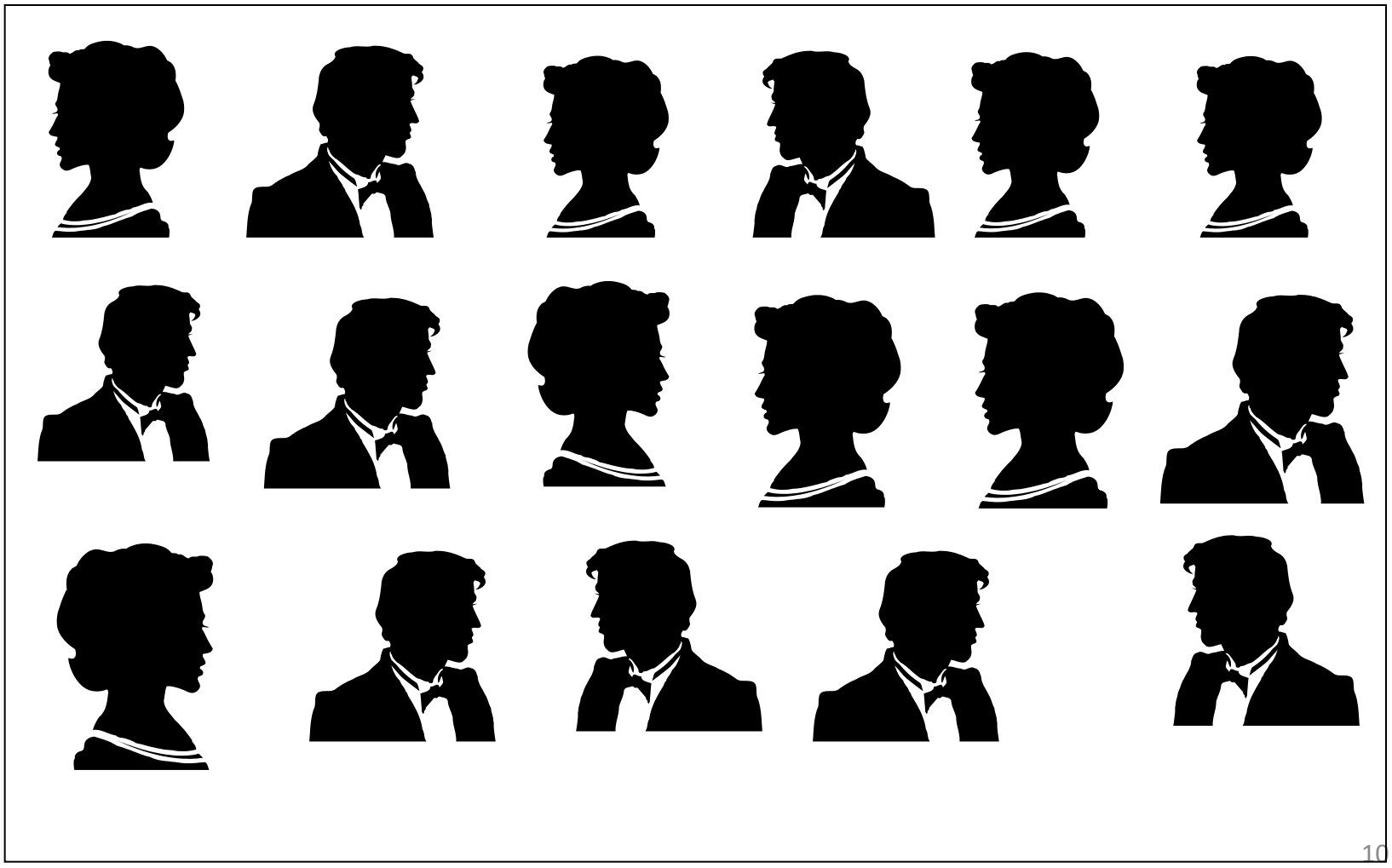
# Classical Test Theory

- The following pages describe the ‘classical test theory model’
- From psychometrics
- No gold standard required
- Used in surveys
- In other applied areas of statistics other models are more popular (ANOVA type of models)

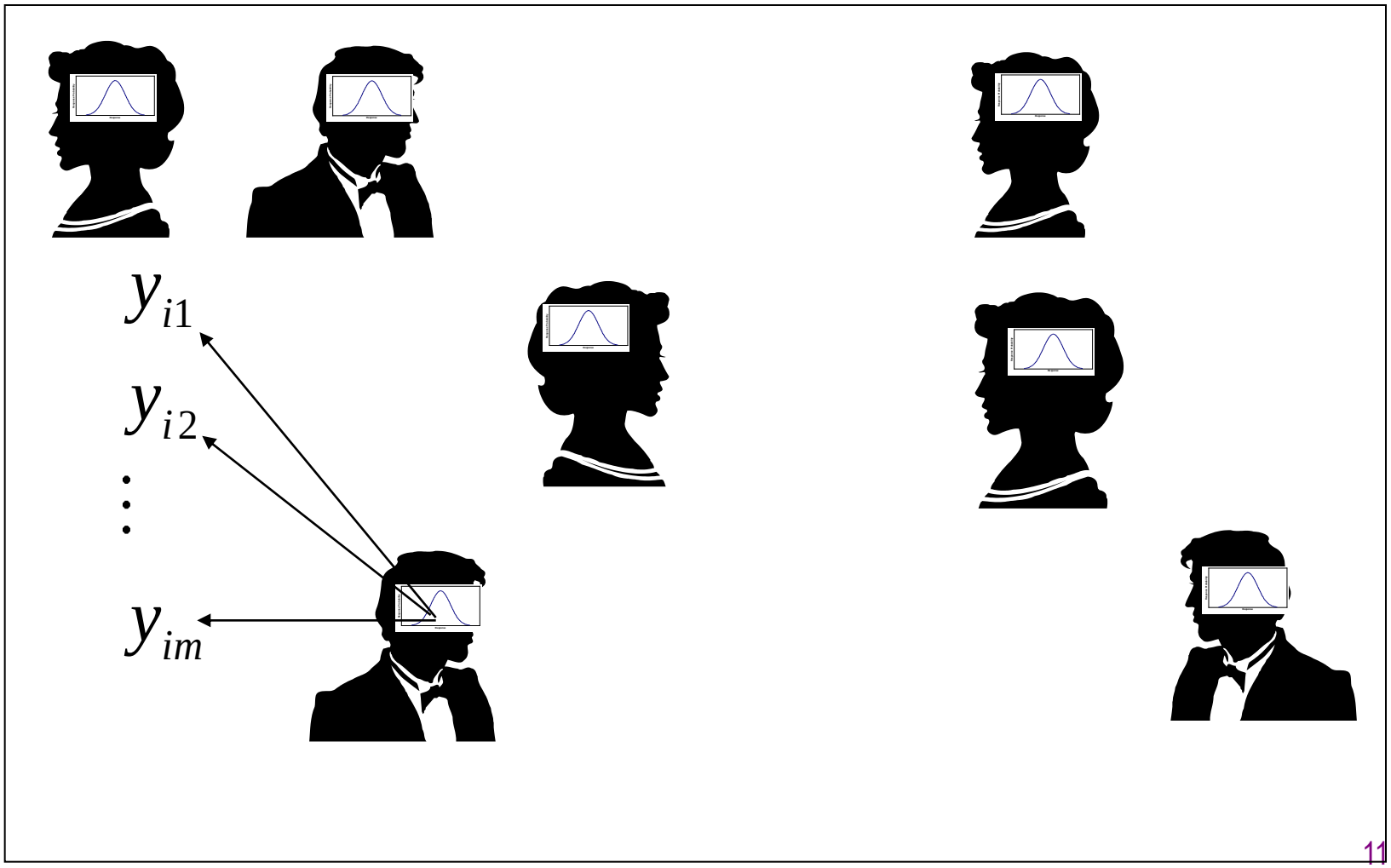
# Conceptual Development

- Hypothetical distribution of responses for each individual in the population
- Individuals represent “clusters” of “potential” responses
  - Analogous to 2-stage sampling
- The response process is analogous to two-stage cluster sampling using SRS at each stage
- The first-stage design can be more complex than SRS, classical test theory will still be useful

# Clusters are Persons
# Responses are Nested within Persons



# Sample Persons and then Response from within Persons



# ‘True score’

- Not the same true value
- True score is average of responses from individual i
- Can be influenced by for example change of mode of interview or by pictures on the questionnaire

- Individuals may be viewed as equal size clusters of potential responses to a question; i.e., Primary Sampling Units (PSUs)
- $n$ = number of persons in the sample
- $m$ = number of observations made on each person
- A response to an interview question essentially selects a response from an individual randomly and independently ($m$ = 1 response)
- For an interview-reinterview survey, cluster sample size is $m$ = 2

# A Measurement Model Based Upon Two-Stage Cluster Sampling: "Census Bureau" Model

- First stage = individual in the population, i = 1, …, N
- Second stage, infinite number of possible measurements (or trials) on the individual
- SRS at both stages (can be relaxed)
- Negligible sampling fraction at second stage (i.e., m/M <<1 or essentially unlimited number of hypothetical responses within person)

# ‘Parallel measures’

- All measurements (ie what people may say as an answer to a question) are indicators of the same construct (ie same variable)
- They are taken from the same distribution

- Then they are independent and identically distributed (iid). Tall order.

# A Measurement Model Based Upon Two-Stage Cluster Sampling: "Census Bureau" Model (cont'd)

- Want to estimate $\bar{Y} = \sum_{i=1}^{N} \tau_i$

- where $\tau_i$ is average of the infinite number of responses from individual *i*
- (Well, we would have wanted the true value…)

# Review of Formulas for Two-Stage Cluster Sampling (Cochran, 1977, Chapter 10)

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm} = \frac{\sum_{i=1}^{n} \bar{y}_i}{n}$$

$n$ = number of clusters

$m$ = cluster (sample) size

$$E(\bar{\bar{y}}) = \bar{\bar{Y}}$$

For negligible sampling fraction at the second stage

$$Var(\bar{\bar{y}}) = (1-f)\frac{S_1^2}{n} + \frac{S_2^2}{nm}$$

$$v(\bar{\bar{y}}) = (1-f)\frac{s_1^2}{n} + f\,\frac{s_2^2}{m} \cong \frac{s_1^2}{n}, \text{ if } f<<1$$

# Review of Formulas for Two-Stage Cluster Sampling (cont'd)

$$S_1^2 = \sum_{i=1}^{N} \frac{(\bar{Y}_i - \bar{\bar{Y}})^2}{N-1}$$

$$S_2^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y}_i)^2}{N(M-1)}$$

$$s_1^2 = \frac{\sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}})^2}{n-1}$$

$$s_2^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \bar{y}_i)^2}{n(m-1)}$$

# For Measurement Error Applications

The formulas on previous slide can be directly applied to the measurement error problem.

$$\mathrm{Var}(\bar{\bar{y}}) = (1\text{-}f)\frac{S_1}{n} + \frac{S_2}{nm} = (1\text{-}f)\frac{\mathrm{SV}}{n} + \frac{\mathrm{SRV}}{nm}$$

SV $= S_1^2$ i.e., "sampling variance

SRV $= S_2^2$ i.e., "simple response variance"

$$Var(\bar{\bar{y}}) \hat{=} v(\bar{\bar{y}})$$

# Some implications

$$\mathrm{Var}(\bar{\bar{y}}) = \frac{\mathrm{SV}}{n} + \frac{\mathrm{SRV}}{nm}$$

- The variance of a mean or proportion from a SRS contains variation due to both sampling variance and response variance (note what happens when $m = 1$)
- Sampling variance decreases as $n$ increases, i.e., precision inversely proportional to sample size

- Measurement variance decreases as both $n$ and $m$ increases; i.e., better precision with multiple measurements on each unit

# Reliability ratio

$$R = \frac{SV}{SV + SRV}$$ is the reliability ratio

If SRV=0 then $R$=1 (i.e. maximum value)

The smaller $R$ is, the more the estimate will be improved by repeated measures, because then SRV is larger

# Estimation of SRV

- Why do we want to estimate the *SRV* when this component is already accounted for in the usual estimate of Var($p$)?
  - *SRV* increases Var($p$)
  - *SRV* has implications for other analysis as well

  E. g. measurement error may have implications for estimation of coefficients in some models

# Special Formulas for Proportions

$$P = \bar{\bar{Y}}; \quad Q = 1 - P; \quad P_i = \bar{Y}_{i;} \quad Q_i = 1 - P_i;$$

$$p = \bar{\bar{y}}; \quad q = 1 - p; \quad p_i = \bar{y}_i; \quad q_i = 1 - p_i$$

$$S_1^2 = \sum_{i=1}^{N} \frac{(P_i - P)^2}{N - 1} \qquad S_2^2 = \sum_{i=1}^{N} \frac{P_i Q_i}{N}$$

$$s_1^2 = \frac{\sum_{i=1}^{n} (p_i - p)^2}{n\text{-}1} \qquad s_2^2 = \frac{m}{n(m-1)} \sum_{i=1}^{m} p_i q_i$$

# Total Mean Square Error of *P* for *m* = 1

$$\mathrm{MSE}(p) = [\mathrm{Bias}(p)]^2 + \mathrm{Var}(p)$$

$$\mathrm{Bias}(p) = \mathrm{E}(p) - \pi \quad \leftarrow \text{True proportion}$$

$$= P - \pi \quad \leftarrow \text{Lack of validity}$$

Thus,

$$\mathrm{MSE}(p) = (P-\pi)^2 + \mathrm{Var}(p)$$

$$= (P-\pi)^2 + (1-f)\,\frac{S_1^2}{n} + \frac{S_2^2}{n}$$

This can be rewritten as

$$\mathrm{MSE}(p) = (P - \pi)^2 + (1 - f)\,\frac{SV}{n} + \frac{SRV}{n}$$

# Estimation of Simple Response Variance (*m*=2)

Suppose $m = 2$ for all $n$ (example: an interview followed by a reinterview for all cases)

$$p_i = \frac{(y_{i1} + y_{i2})}{2}$$

Then,

$$p_i q_i = [\frac{y_{i1} + y_{i2}}{2}][1 - \frac{y_{i1} + y_{i2}}{2}]$$

It can be shown that

$$p_i q_i = \frac{(y_{i1} - y_{i2})^2}{4}$$

$$\text{Hint}: y^2 = y \text{ for dichotomous variables}$$

$$\therefore \sum_{i=1}^{n} p_i q_i = \sum_{i=1}^{n} \frac{(y_{i1} - y_{i2})^2}{4}$$

$$\therefore SRV = s_2^2 = \frac{1}{2n} \sum_{i=1}^{n} (y_{i1} - y_{i2})^2$$

Note also that

$$Var(y_i) = E(y_i - P)^2 = E(y_i^2) - P^2 = P - P^2$$
$$= PQ$$

Thus,

$$SV + SRV = PQ \hat{=} pq$$

# Some ‘science thinking’

- Is the model that the brain is a “random machine” realistic and credible?
- First, model misspecification, would that result in. As for the iid assumption, suppose the two measurements are correlated. How is the estimated variance affected?
- If the identical distribution assumption is violated?

- If the assumptions are mildly violated, is the test theory model useful anyway?