

Inference-based Measures of Utility for Contingency Tables

Ashish Sanil

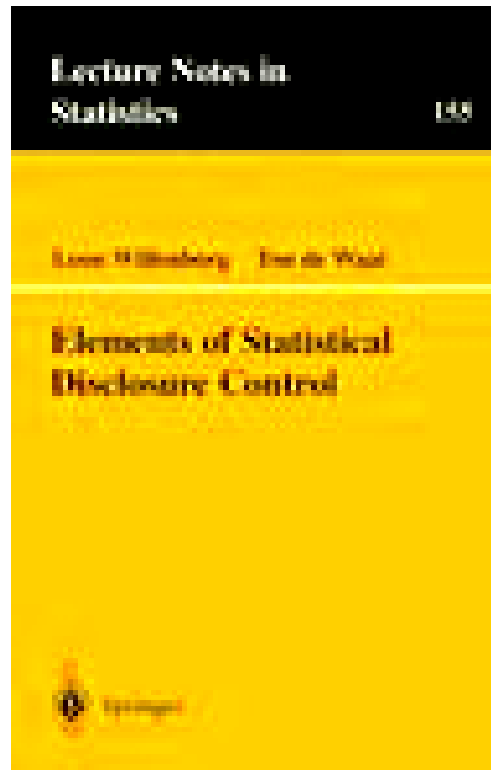
National Institute of Statistical Sciences

[part of work done with Adrian Dobra, Steve Fienberg, Shanti Gomatam, and Alan Karr]

August 12, 2004

“... the paradigm of statistical confidentiality: to modify unsafe data in such a way that safe (enough) data emerge, with minimum information loss.”

Willenborg and de Waal
“Elements of Statistical Disclosure Control”

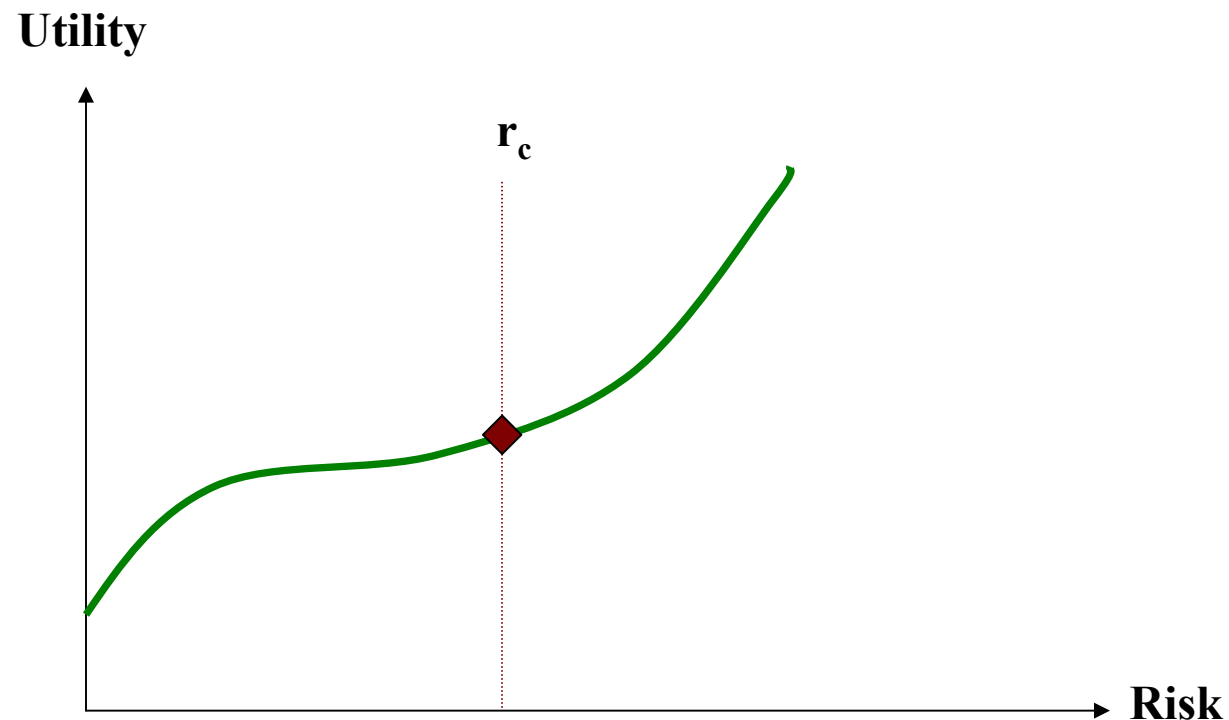


Leon Willenborg and Ton de Waal
“Elements of Statistical Disclosure Control”

“Once it has been decided how to measure safety and information loss, the production of safe data from unsafe data, using a particular protection technique, is often a matter of solving an optimization problem.”

Willenborg and de Waal
“Elements of Statistical Disclosure Control”

$$\begin{aligned} \max U(\alpha) \\ \text{s.t. } R(\alpha) \leq r_c \end{aligned}$$



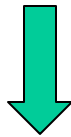
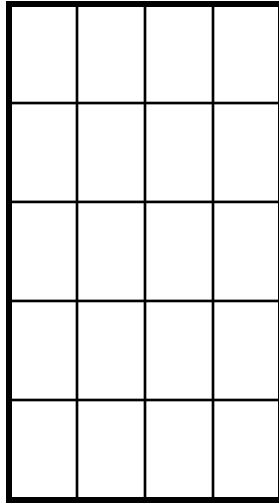
“We view tables very much as final products, ready for use in the form they are delivered to customers... We are not assuming that [the customer] will want to input such a table into a statistical package to analyze its structure.. Therefore we shall not deal with the statistical consequences of applying SDC techniques to tables.”

Willenborg and de Waal

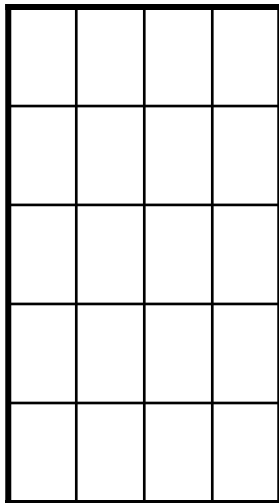
“Elements of Statistical Disclosure Control”

Distortion Measures as (Dis)-Utility Measures

Original (O)



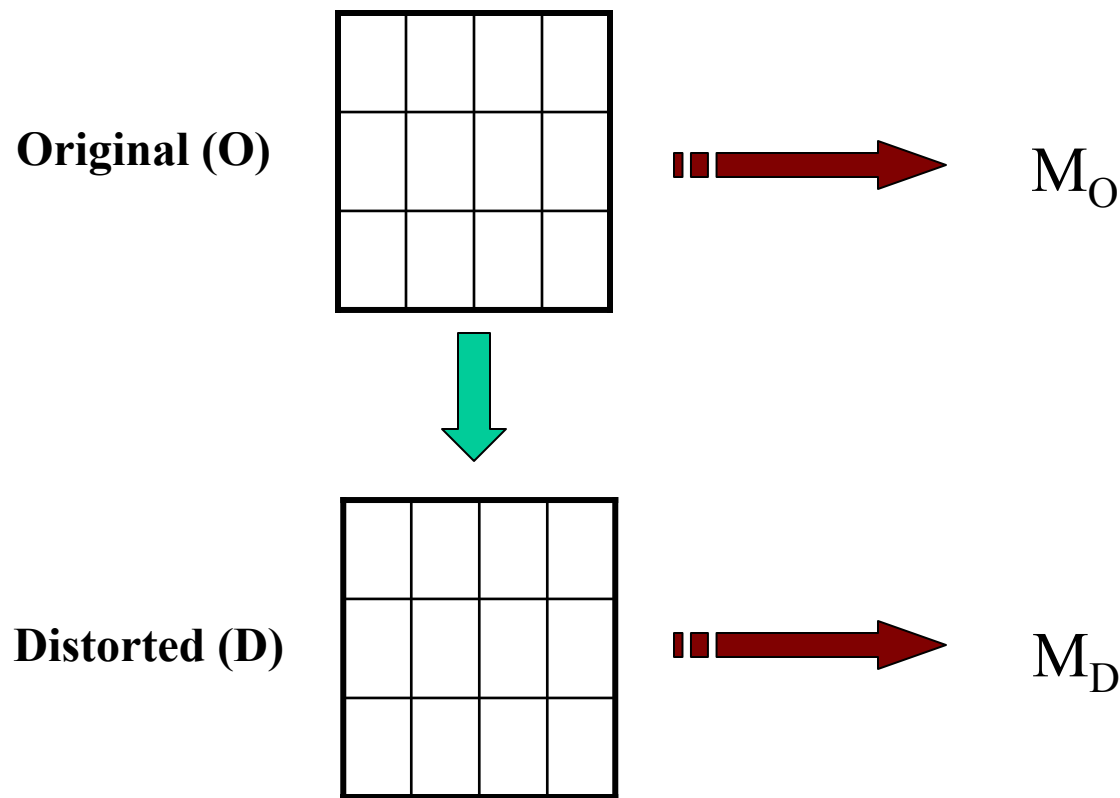
Distorted (D)



Distortion/Dis-Utility
measures used are
usually of the form

$$d(O, D) \sim \sum_{i \in \text{Cells}} d(O_i, D_i)$$

Examples: Hellinger
distance, χ^2 -type, sum-of-
squared-error, etc.



If the released table is going to be subjected to statistical analyses, should we be using $d(M_O, M_D)$ instead of $d(O, D)$ as measures of utility/loss.

Characteristics of $d(M_O, M_D)$

Inferences should be similar to those with original data

- Multivariate analyses performed should yield similar results
- Associations between variables preserved as much as possible

Why use $d(M_O, M_D)$?

- Good models for the data are a lot less complex than the full data (“saturated model”). So there is potential for applying more aggressive SDL procedures without distorting the message of the data model.
- Conversely, small perturbations of the full data might yield data sets that do not support the model for the original data.

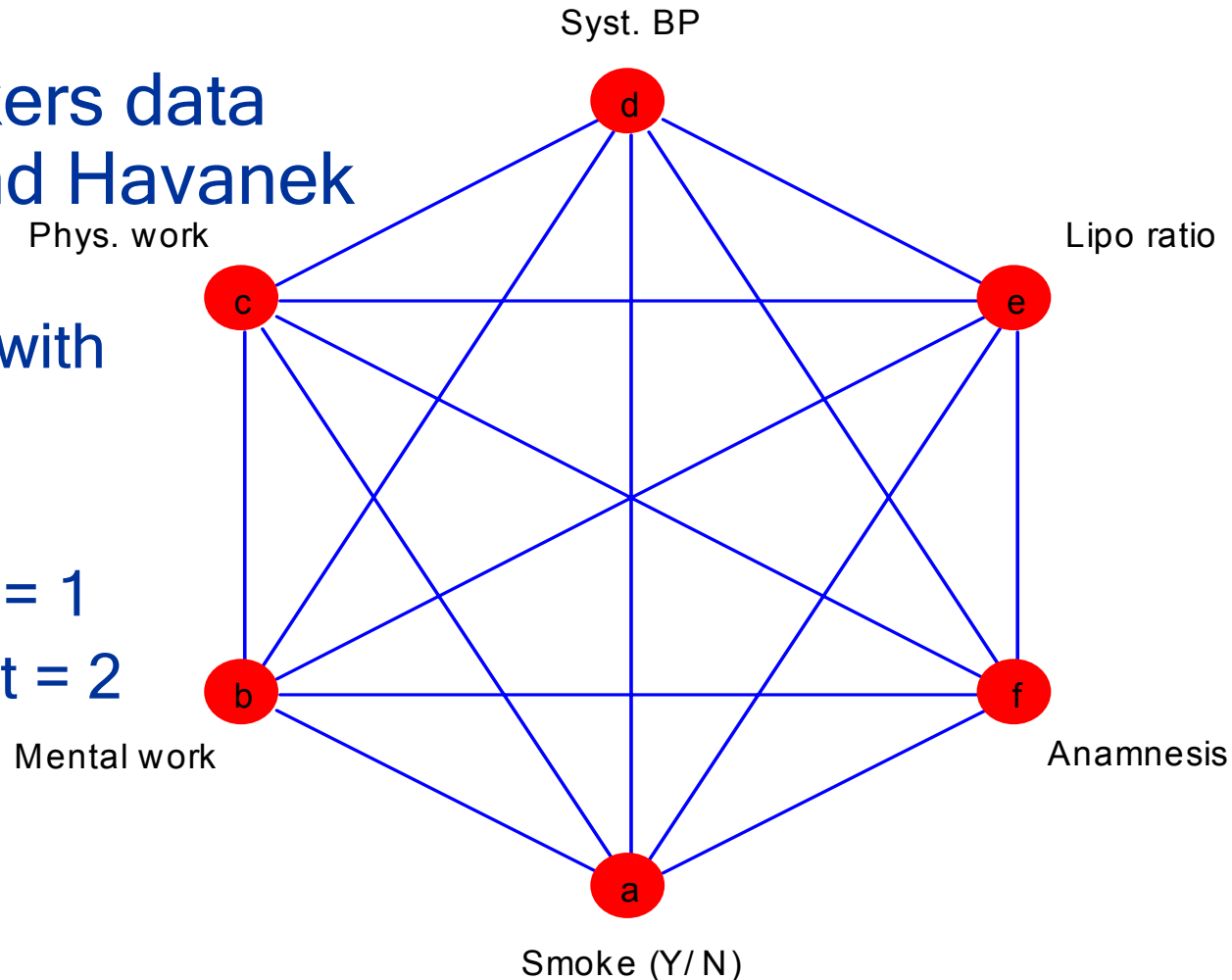
General Strategy to Evaluate the Effect of SDL

- Determine a good model for the data
- After applying a SDL technique, check to see if the altered data support the model
 - Determine if the altered data “contain” the model (e.g., if MSS are preserved)
 - Apply a model selection procedure and see if we recover the original data model
 - Apply a model selection procedure and see if we get close to the original data model (E.g., number of n-way interactions preserved, K-L distance, ability to predict some variable(s) given the others)
 - Evaluate the “goodness-of-fit” of the altered data w.r.t. the original data model

Example 1: Risk Factors for Coronary Heart Disease

- Czech auto workers data from Edwards and Havanek (1985)

- Population data with $n=1841$
- 2^6 table
- 1 cell with count = 1
- 2 cells with count = 2



Example 2: The Data

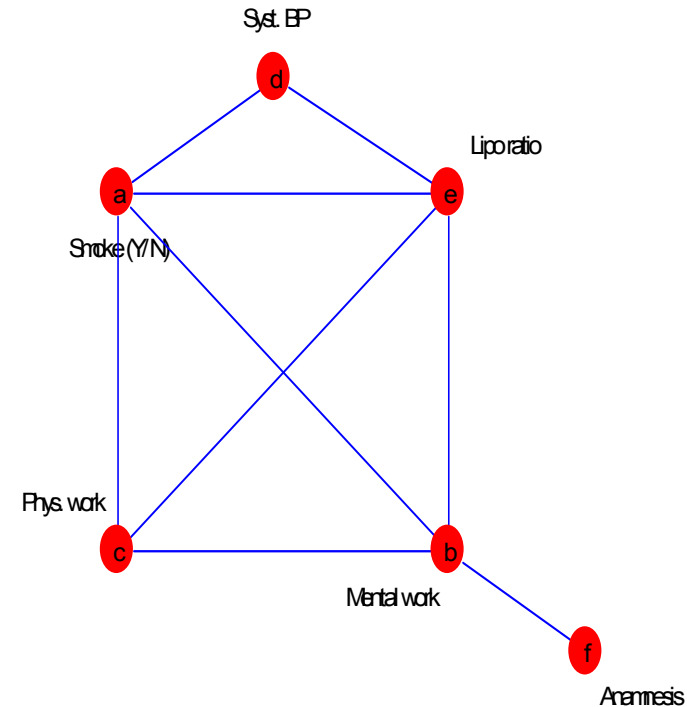
F	E	D	C	B	no		yes	
				A	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		>= 140	no		35	12	80	33
	>= 3	< 140	yes		109	67	7	9
			no		23	32	70	66
		>= 140	yes		50	80	7	13
pos	< 3	< 140	no		24	25	73	57
			yes		51	63	7	16
		>= 140	no		5	7	21	9
	>= 3	< 140	yes		9	17	1	4
			no		4	3	11	8
		>= 140	yes		14	17	5	2
	>= 3	< 140	no		7	3	14	14
			yes		9	16	2	3
		>= 140	no		4	0	13	11
		>= 140	yes		5	14	4	4

Example 1: “Best” Model

Consider the graphical model

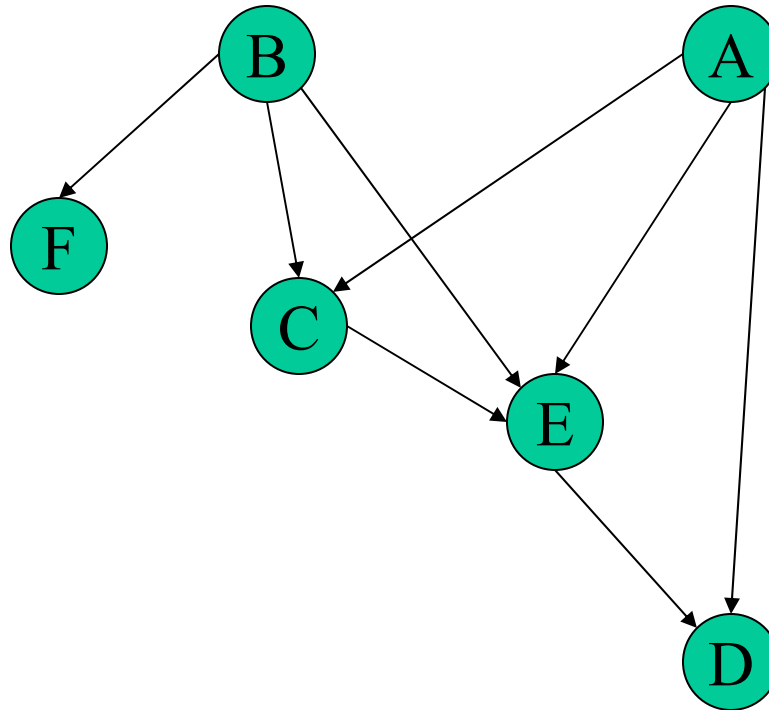
[ADE][ABCE][BF] :

- Corresponds to decomposable graph
- Several published analyses of this data set favor this model for the data set (Edwards and Havanek, Whittaker, Fienberg, ..)
- AIC-based model selection also picks this model as the best graphical model



Example 1: “Best” DAG Model

$$P(A,B,C,D,E,F) = P(F|B) * P(C|B) * P(E|A,B) * P(E|C) * P(D|E,A) * P(A) * P(B)$$



Example 1: Safe release

- The set of marginals [BCDEF], [ABCEF], [ABCDF], [ACDE] was considered safe to release (in the sense that none of the designated sensitive cells were closely bound (Dobra, Karr, Sanil and Fienberg, IJUFKS 2002))
- Includes the MSS for the favored model :
[ADE][ABCE][BF]
- The released marginals are sufficient for model selection procedures to select the “favorite” model
- *Can do correct log-linear model-based analyses*
- Can also do perturbations (via Grobner bases) that preserve these marginals, and release perturbed table

Example 1: Distortion Minimizing release

- Performed perturbation to minimize

$$\sum_{i \in \text{Cells}} |o_i - d_i|$$

Subject to 1-d marginal constraints

- AIC-based model selection picked [ABCEF][ADE] as the best graphical model

Example 2: CPS 8-way

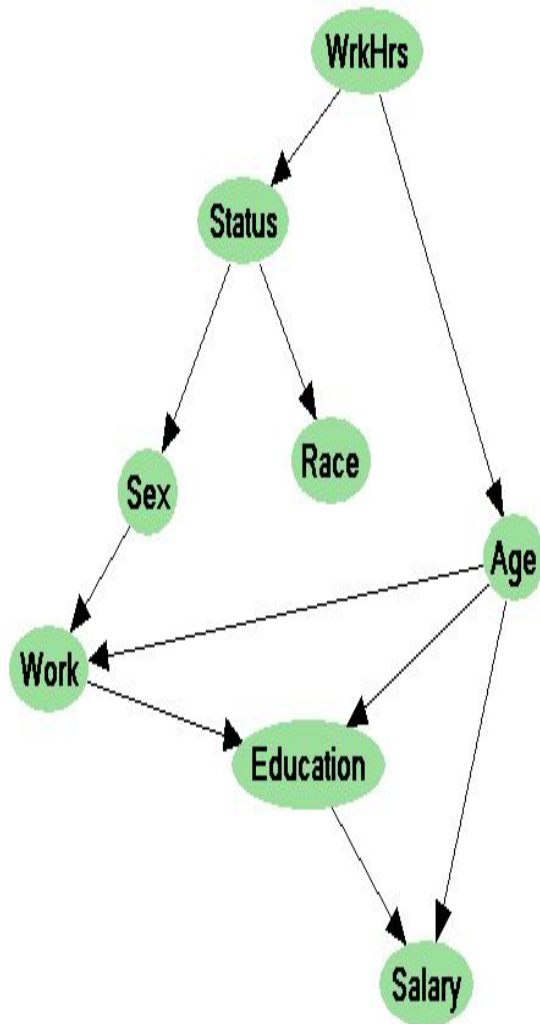
Excerpt from 1993 CPS data

- 48,842 data records
- 8 categorical attributes: Age, Sex, Race, Education, Employment, Hrs Worked, Marital Status, Salary
- Goal is to study the effect of data swapping on the 8-d structure of the data
- In this talk we look at seven swapped data files where **Salary** is swapped (5% swap rate) within a fixed value of one of the other **variables** (e.g., swap salary with fixed race \Rightarrow take two records with same race and swap their salary values)

Example 2: “Best” model

- Very difficult model selection problem
 - Enormous search space
 - BIC-based search criterion has many (weak) local maxima
- Fitted a “bayes net” model for the data (i.e., a directed acyclic graph representation for the 8-d probability distribution of the data)
- Will use data simulated from this model

Example 2: “Best” model

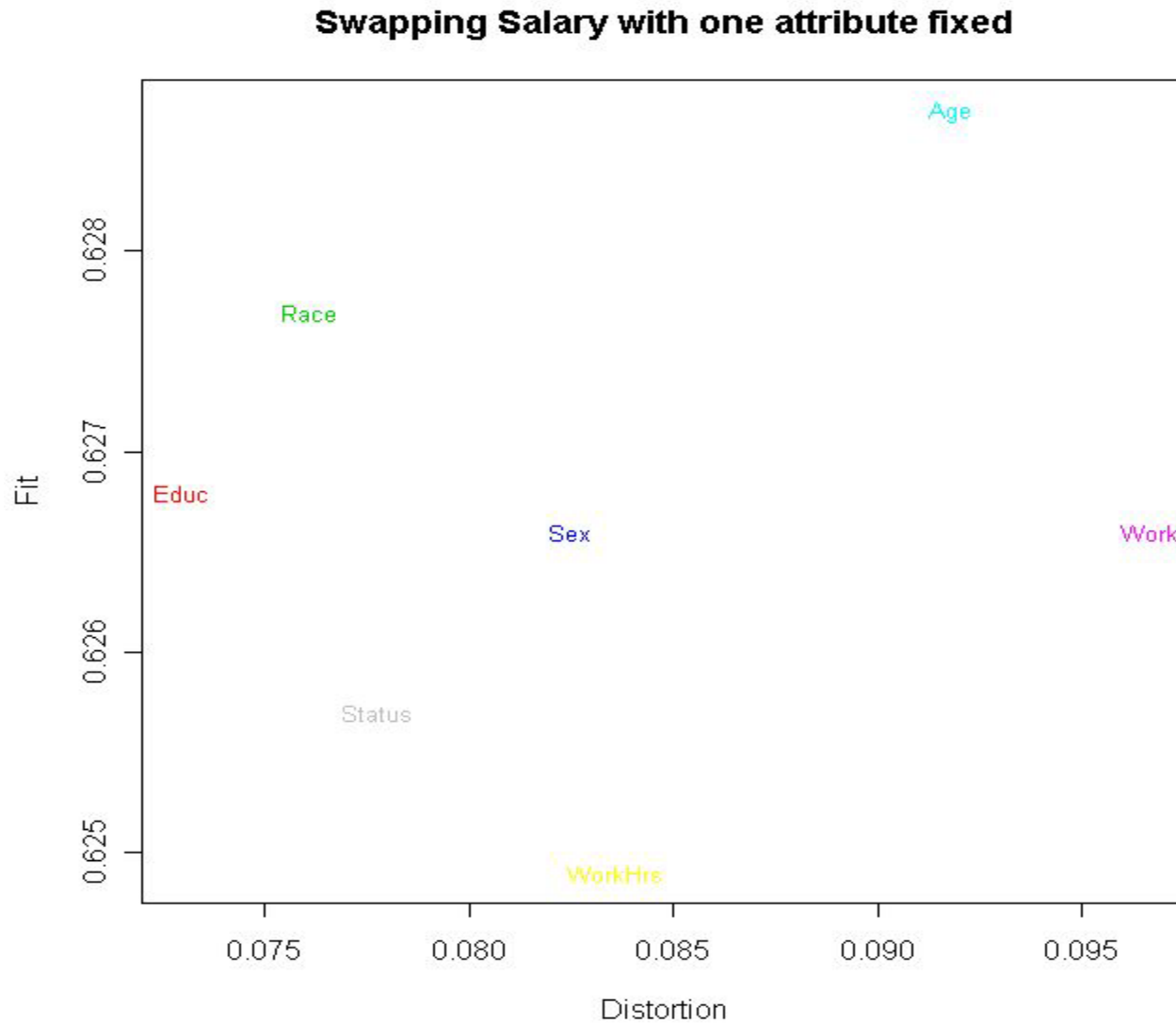


$$\begin{aligned} P(\text{Age, Edu, Work, Sex, Race, WrkHrs, Status, Salary}) = & \\ P(\text{Salary} \mid \text{Age, Edu}) * P(\text{Edu} \mid \text{Age, Work}) * & \\ P(\text{Work} \mid \text{Age, Sex}) * P(\text{Age} \mid \text{WrkHrs}) * & \\ P(\text{Sex} \mid \text{Status}) * P(\text{Race} \mid \text{Status}) * P(\text{WrkHrs}) & \end{aligned}$$

General Strategy to Evaluate the Effect of SDL

- ✓ Determine a good model for the data
- After applying a SDL technique, check to see if the altered data support the model
 - Determine if the altered data “contain” the model (e.g., if MSS are preserved)
 - Apply a model selection procedure and see if we recover the original data model
 - Apply a model selection procedure and see if we get close to the original data model
- ✓ Evaluate the “fit” of the altered data w.r.t. the original data model

Example 2: Assessing Utility



Research Issues

- Measures of $d(M_O, M_D)$ need to be developed and their properties should be investigated
- How much extra is revealed to potential intruders if the released data comes with a guarantee of statistical structure being preserved?
- Automatic/objective procedures for determining the “best” model
- Computational scalability

Summary

- Presented some general principles for assessment of statistical utility of released data
- Two examples illustrating various aspects of the process
- Highlighted some of the challenges
- **Basic message: After applying SDL procedures, you should check to see if the inferential properties of the data are preserved**

References

- <http://www.niss.org/dg> : papers, references on cell-bounds and many other things
- Leon Willenborg and Ton de Waal:
 - “Statistical Disclosure Control in Practice” (1996), Springer
 - “Elements of Statistical Disclosure Control” (2000), Springer