Data Quality - Edits

Hans J. Lenz, Freie Universität Berlin

September 2007

Edits

Objective:
Data Cleansing at the Data Entry
to assert semantic consistency of data with nominal, ordinal and metric scales.

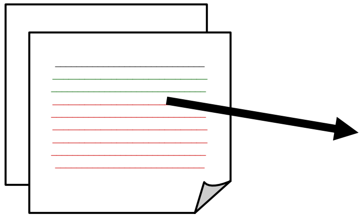
„Numbers don‘t mind where they come from“ *(LORD*)

Agenda

1. Examples
2. Definitions
3. Simple Edits
4. Logical Edits
5. Numerical Edits
6. Statistical Edits
7. Fuzzy Edits
8. Evaluation of Statistical Edits vs. Fuzzy Edits
9. MCMC Simulation

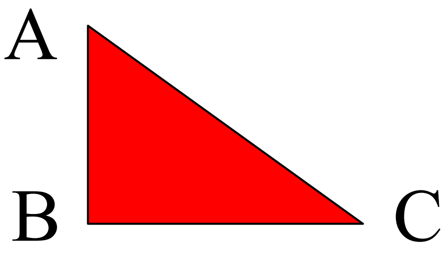
1. Examples

Ex. 1: Triangle Data



x = (2,4,3) consistent?

Frame of discernment:
x is a list of lengths from a triangle with a right angle at B, and AB=2, AC=4, BC=3.



+ Pythagoras =

Constraint

$BC^2+AB^2=AC^2$

$4 + 9 \neq 16$

Ex. 2: Relational Data

- Is the tupel

x = (0010, 015, ´elementary´, ´child´, ´single´) consistent?

- DB Schema as frame of discernment:

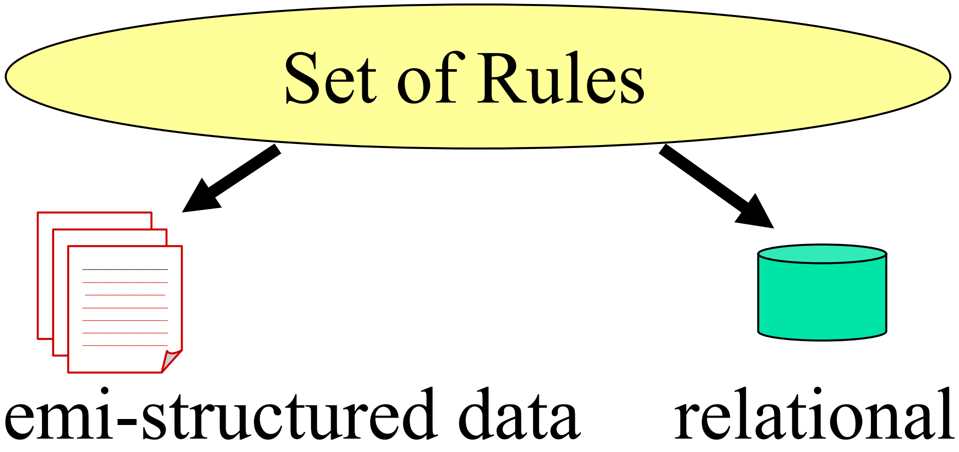
questionnaire (oid, age, school_type, household_status, marital_status)

2. Definitions

Edits (Rules, Checks)

DEF.:

Edits are (normalized) rules (or formulas)
 - applicable to each object of a data set and
 - generated by relationships which exist
   between attributes.



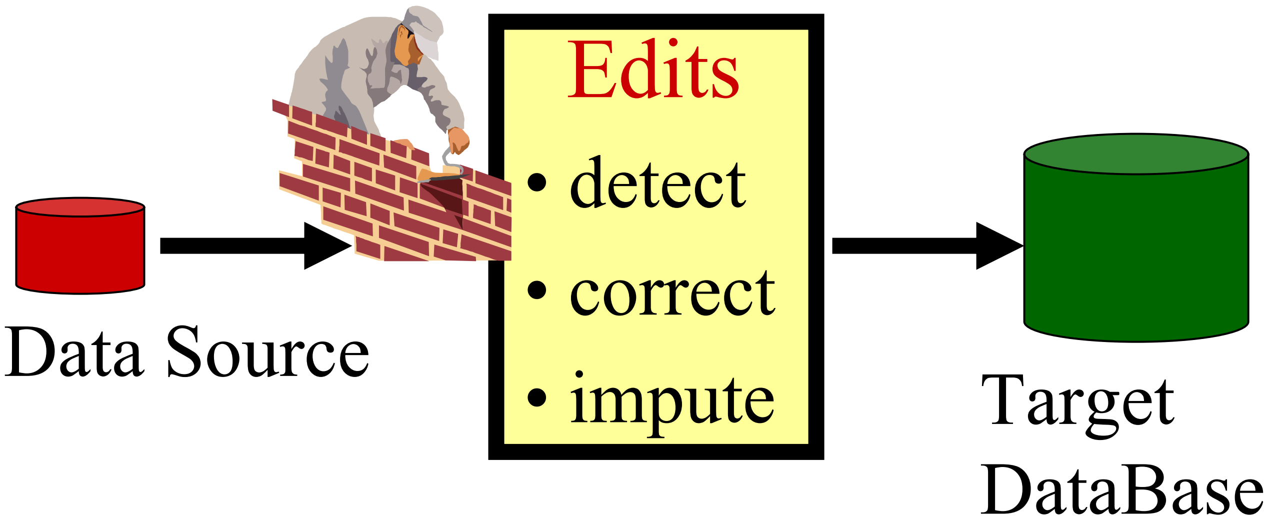
Set of Rules

Semi-structured data relational data

Editing Tasks

Given a database **D** on a universe **U**

- *detect* semantic inconsistencies
- *correct* for incoherent data
- *impute* data in case of missing values.



Data Source

Edits

- detect
- correct
- impute

Target DataBase

Objects at Data Entry

- **fix-formatted Data**
  - **Relations**
  - **Sets**
  - **Lists**
  - **Files**
  - **Records**
  - **Fields / Attributes / Variables**
- **semi-structured Data**
  - **HTML**
  - **XML**

Types of Edits

- simple edits
- logical edits
- numerical edits
- probabilistic edits
- statistical edits
  (based on Probability Theory)
- fuzzy edits (based on Fuzzy Logic)

3. Simple Edits

☞ Conceptually the simplest edits are those applied to single field or attribute with respect to

- Data type
- Length
- Subset Constraints
- Scale
- Dimension

3. Simple Edits

Examples

Syntax:

<attribute name> <predicate> <argument>

- Age type cardinal
- code length 4
- date between (01.07.06-06.07.06)
- Consumption rate scale metric
- costs unit €/year

4. Logical Edits

(cf. Categorical rules)

Example 1:

for all u $\in$ **U**

if le(Age,15)

    or school(elementary)

then not household_status(head)

    and marital_status(single)

Generalization:

if x_1 *is* A_1 *and* x_2 *is* A_2 *and*...x_p *is* A_p

then y *is* B

☞ Fellegi and Holt (1973, 1976),
Mamdani and Assilian (1975)
Winkler (2004)

4. Logical Edits

Theorem „Normal Form Edit”

(Fellegi and Holt (1973)

Let f, g,h* $\in$ $\boldsymbol{C}$ be compound clauses and A,B,C,..., P,Q,R,... $\in \mathscr{P}$ predicates of 2order-type. Then

$$f(A,B,C,\ldots) => g(P,Q,R,\ldots)$$

$$<=> \bigwedge_{s} \mathbf{h}^{*}_{s} = .\text{false}.$$

where s is a subset of the set **S** of attributes

☞ Fellegi and Holt (1973, 1976),
Mamdani and Assilian (1975), Boskovits (2007)

4. Normal Forms of Edits

Examples

- le(Age, 15)∧ household _status(head)=.false.
- le(Age, 15)∧ not marital_status(single)=.false.
- school(elementary) ∧
  household _status(head)=.false.
- school(elementary) ∧
  not marital_status(single)=.false.

4. Logical Edits - Algorithms

Algorithms exist for:

- normalising edits into NF
- deciding whether an edit is new
- construct the complete set of essentially different edits
- identifying attributes to be most likely in error.

Sources: Fellegi, Holt (1976), Greenberg and Surdi (1984), Wetherill and Gerson (1986)

5. Numerical Edits

- *Imprecise Values* of attributes not allowed
- Defined only for data types *integer, cardinal, real, decimal*
- "Constraint Programming" by LP:

$A\ x \geq b$	(numerical constraints)
$x \geq 0$	(non negativity)
$x \in \mathbf{X}$	(for all attributes x)

5. Numerical Edits

Example

- Fact: The former Student s
  is now 29 years of age (x_1),
  stayed 6 years at elementary school (x_2),
  stayed 7 years at high-school (x_3),
  studied 5 years at an university (x_4), and
  is employed since 2 years (x_5).

- *LP: x solves* $a'x \geq b$ with
  $a' = (1,-1,-1,-1,-1)$
  $x' = (x_1,x_2,x_3,x_4,x_5)=(29,6,7,5,2)$
  $b = 6$
  $\mathbf{X} = \Pi \text{ range}(x_i)$

5. Numerical Edits - Objectives

- Detect Redundancy
- Detect Inconsistency
- Error Location (misprint, transcription error, misspelling etc.)
- model-based Imputation

Algorithms for detection & location by:
Sadiq (1986), ...

6. Probabilistic Edits

- Let f $\in$ **C** be a clause over a set of predicates A,B,C,...

- Instead of a numerical edit:

  f(A,B,C,...)=.false.

- Now a probabilistic edit:

  f(A,B,C,...)=.false. with Prob$\geq$1-α

6. Probabilistic Edits - Example

- surprise (=low probability) for any person$_A$, person$_B$:
  if sex(person$_A$ male) *and*
  sex(person$_B$, female) *and*
  married(person$_A$,person$_B$) *and*
  le(*diff*(age(person$_A$),age(person$_B$)),-10).

- surprise *fo*r company x
  if growth(profit, 30%) *and*
  loss(sales, 20%) *and*
  equal(year(profit), year(sales)).

7. Statistical Edits Example

- Data (six variables):

Capital = 60 ± 1	Profit = 10 ± 2
Sales = 55 ± 20	Expenditures = 45 ± 20
ROI% = 10 ± 5	Margin% = 50 ± 5

•Constraints (four equations based on definitions):

Profit = Sales - Expenditures
ROI% = 100 * Profit / Capital
Margin% = 100 * Profit / Sales
Turnover% = 100 * Sales / Capital

Note: Margin% = 50% > 100*Profit / Sales=18% !

7. Statistical Edits

Syntax:

<variable> = <value> ± <abs. error>/

< % error>/

<stdv>

Example:

Profit = 10 ± 2 with confidence level 1-α

Sources: Schmid (1979), Lenz and Rödel (1991)

7. Statistical Edits Modelling

- Model Specification
  - random variables

- Ranges
  - confidence intervals using Gaussian distribution

- Parameter Estimation / Learning
  (☞ only one value per variable measured!)
- Inference (Detection, Correction, Imputation)

7. Statistical Edits - Model

- State Space Model:

x observed state vector

ξ unobservable (error-free) state vector

ζ dependent vector

$\zeta = H \xi$ balance equation system

z observed vector of ζ

v, w vectors of measurement errors

state space equation: $x = \xi + v$

observational equation: $z = H \xi + w$

7. Statistical Edits

- State Space Model:

u‘ = (v,w) with E(u) = 0 , i.e. no bias!

E(v,w) = 0 due to partial information from independent data sources!

$$E(u\, u') = \mathbf{Q} = \begin{pmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}$$

state space equation: $x = \xi + v$

observational equation: $z = H\,\xi + w$

Source: Schmid (1978), Schneeweiss (1989), Lenz and Rödel (1991, 2000), Rödel (1999)

7. Statistical Edits - Problem

- *Two Problem Classes:*

1. Imputation:

   Given H, P, R and x
   estimate ξ, $\zeta = H\,\xi$ and predict z.

2. Correction:

   Given H, P, R and x, z
   estimate ξ, $\zeta = H\xi$

state space equation: $x = \xi + v$

linear observational equation: $z = H\,\xi + w$

7. Statistical Edits - Estimation

- General Least Squares – Estimator of ξ:

  Let $y' = (x,z)$ and $J = (\mathbf{I}, H)$.

  $$\min \| y - J\,\xi \|_{\mathbf{Q}^{-1}}$$
  $$\text{s.t. } \zeta = H\,\xi$$

☞ $\zeta = H\,\xi$ may be generalized to $\zeta = H(\xi)$

7. Statistical Edits - Estimators

1. Imputation:

Given H, P, R and x
estimate ζ= Hξ and predict z:

$$\hat{\zeta} = \mathrm{Hx} \text{ and } \hat{\mathrm{z}} = \mathrm{Hx}$$

2. Correction:

Given H, P, R and x, z
estimate ξ, ζ= H ξ :

$$\hat{\xi} = x + K(z - Hx) \text{ and } \hat{\zeta} = H\hat{\xi}$$

$$\mathrm{K} = \mathrm{PH}'(\mathrm{HPH}' + \mathrm{R})^{-1}$$

Note: GLS estimators called Kalman Filter Equations and Kalman Gain

7 Statistical Edits - Example

- Data: $x_1 = 30 \pm 20$

  $x_2 = 30 \pm 10$

  $z = 50 \pm 10$

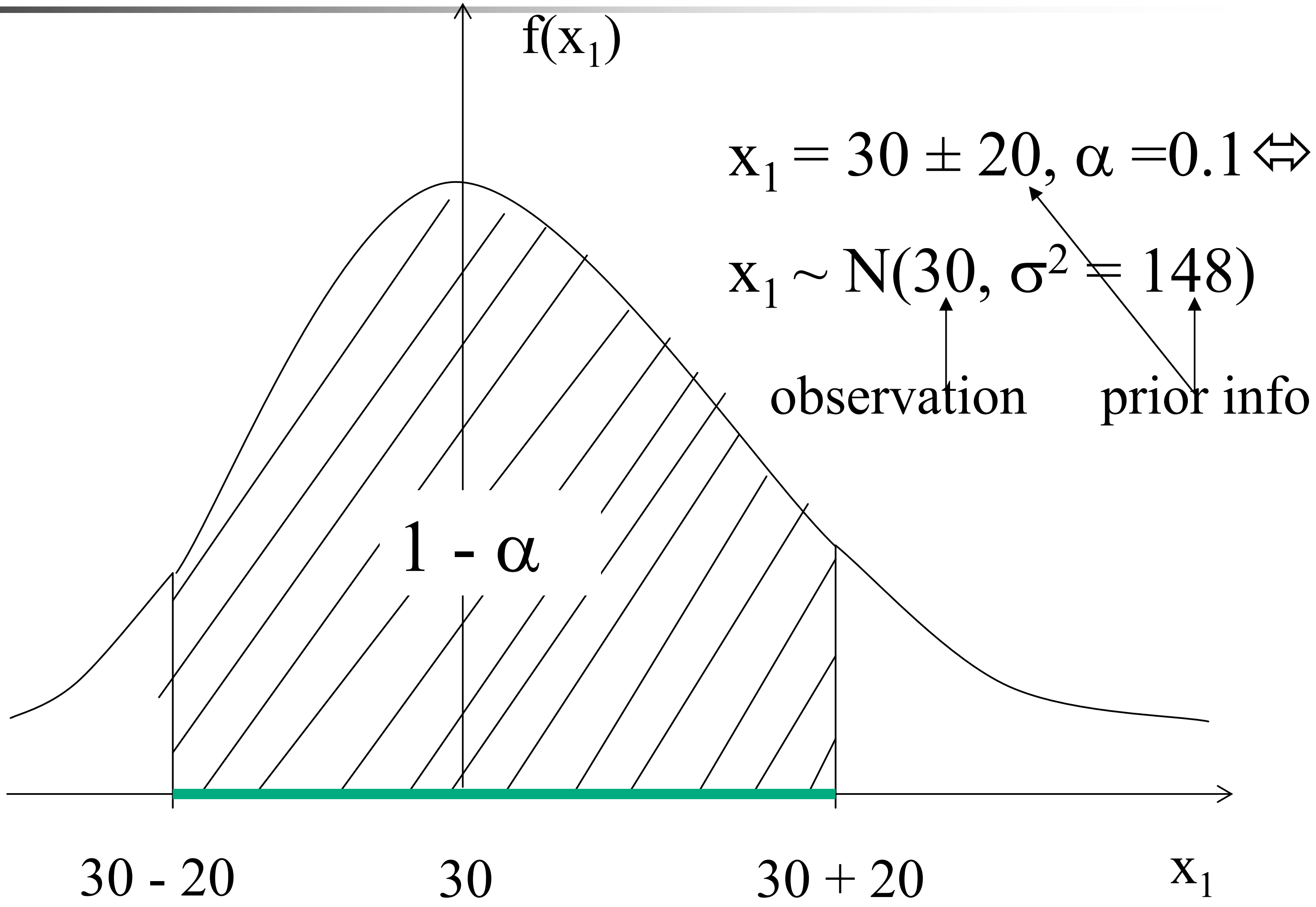
- Model: $\zeta = \xi_1 + \xi_2 = H\,\xi$

  Error Distributions (α=0,1):

  $$E(vv') = P = \begin{pmatrix} 148 & 0 \\ 0 & 37 \end{pmatrix}$$

  $$E(ww') = R^2 = 37$$

7 Statistical Edits Transformation errors into confidence intervals



$x_1 = 30 \pm 20,\ \alpha = 0.1 \Leftrightarrow$

$x_1 \sim N(30,\ \sigma^2 = 148)$

observation

prior info

7. Statistical Edits - Example

Data: $x_1 = 30 \pm 20$

$x_2 = 30 \pm 10$

$z = 50 \pm 10$

Estimates: $\hat{\xi}_1 = 23 \pm 12$

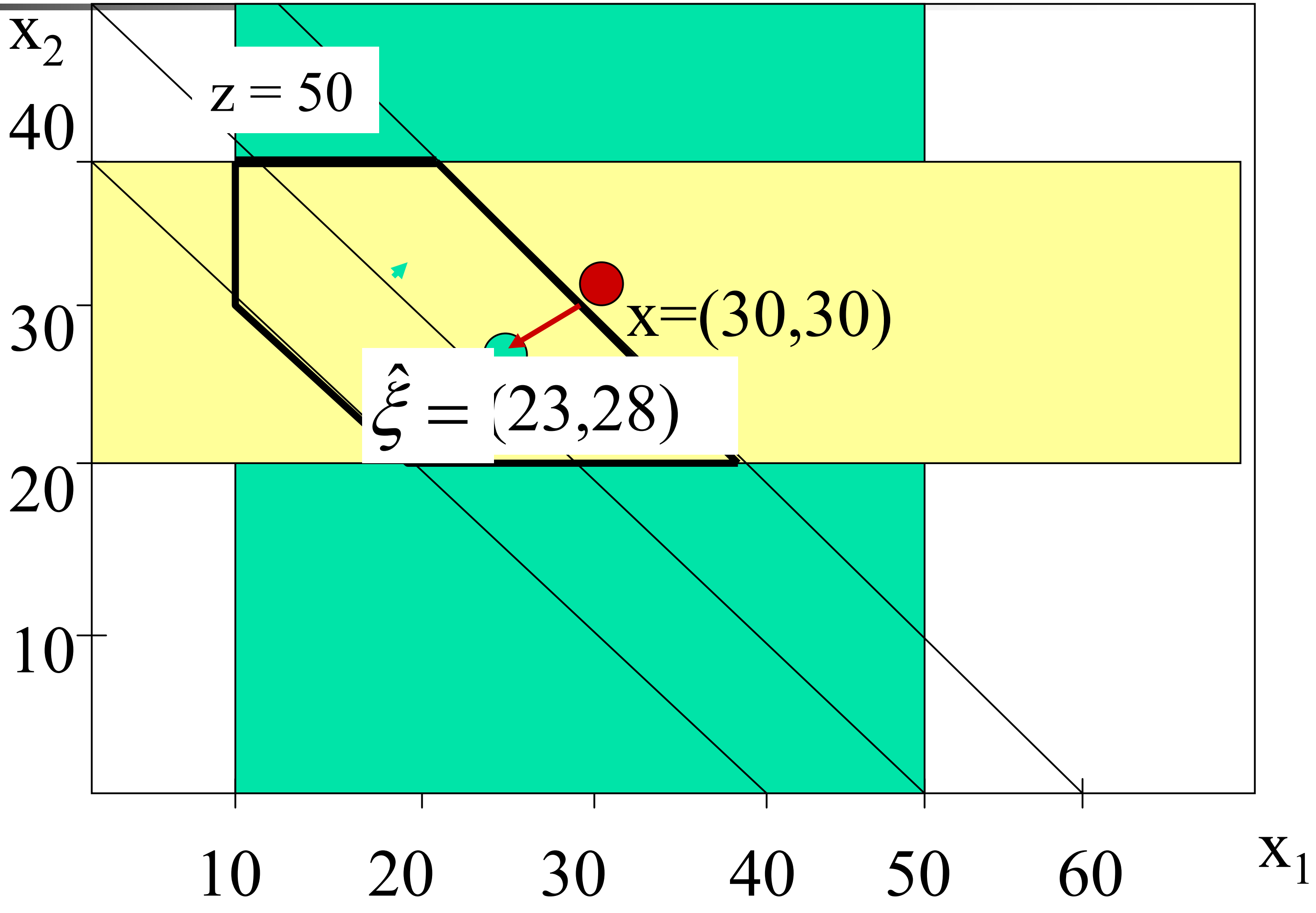
$\hat{\xi}_2 = 28 \pm 9$

$\hat{\zeta} = 51 \pm 9$

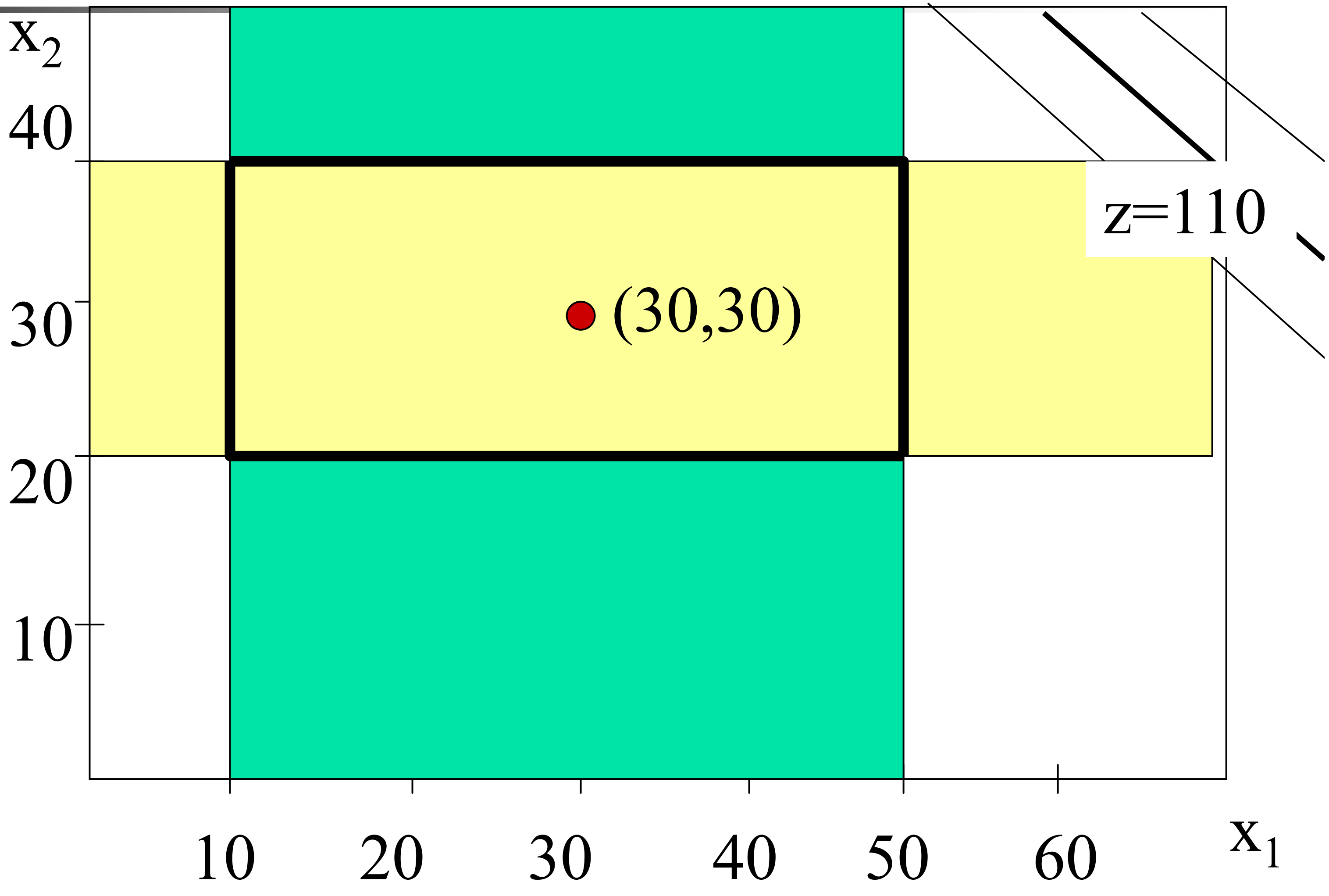
Effects: - Constraints satisfied

- Error Reduction ~ error variance
- Corrections right-shifted

7 Statistical Edits Weak Inconsistency



7. Statistical Edits StrongInconsistency



7. Statistical Edits Algorithms

- PRTI: Schmid (1978) at ETH Zürich
- PRTI II: Schmid and Müller (~1983)
- QUANTOR: Müller and Schürer (~1990), Daimler-Benz AG

8. Fuzzy Edits

Why Fuzzy Edits?

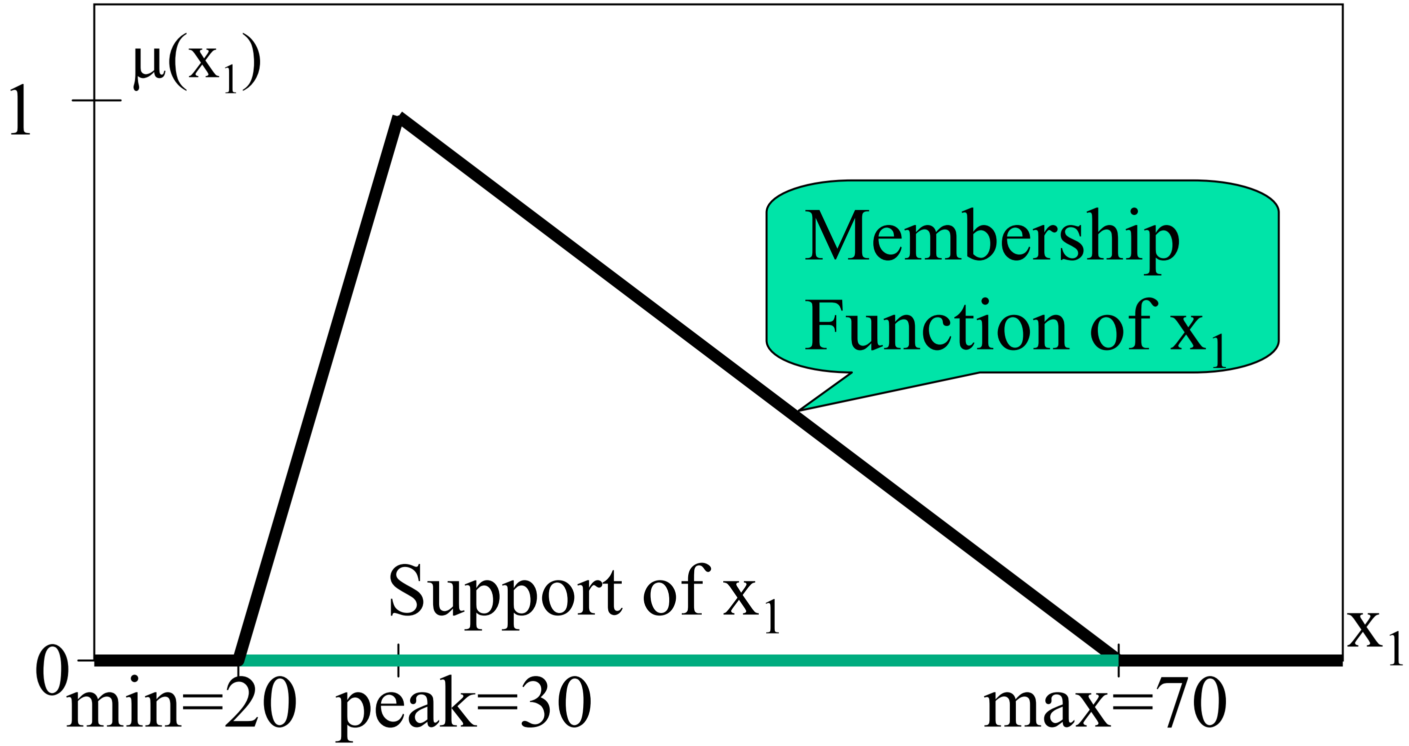
- *Statistical Edits assume*:
  - Gaussian Distribution of errors
  - Linearity (Gaussian Distribution not closed under non-linear transformations!)
  - Correlation almost unknown
- *Fuzzy Edits*
  closed under non-linear transformations
  (Extension Principle of Zadeh)

8. Fuzzy Edits Modelling

- Each variable x_1, x_2,…

  is treated as a fuzzy set (variable).
- A triangle type of membership function is assumed – may be changed.
- Error bounds from expert knowledge is used for specifying the corresponding supports of each variable.
- RHS of balance equations are separable

8. Fuzzy Edits Modelling



Statement: $x_1 = 30 \quad (-10 , +40)$

8. Fuzzy Edits FuzzyCalc® Algorithm

x_0 vector of variables (observed /missing v.)

X product-space spanned by variables x

G(x) = 0 system of p nonlinear equations

F set of p fully specified fuzzy sets on **X**

$$\tilde{x} \in L = \mathop{\times}_{i=1}^{p} \text{support}_i$$

$\text{support}_i := \text{support}_{xi0} \cap \text{support}_{x|g1}$

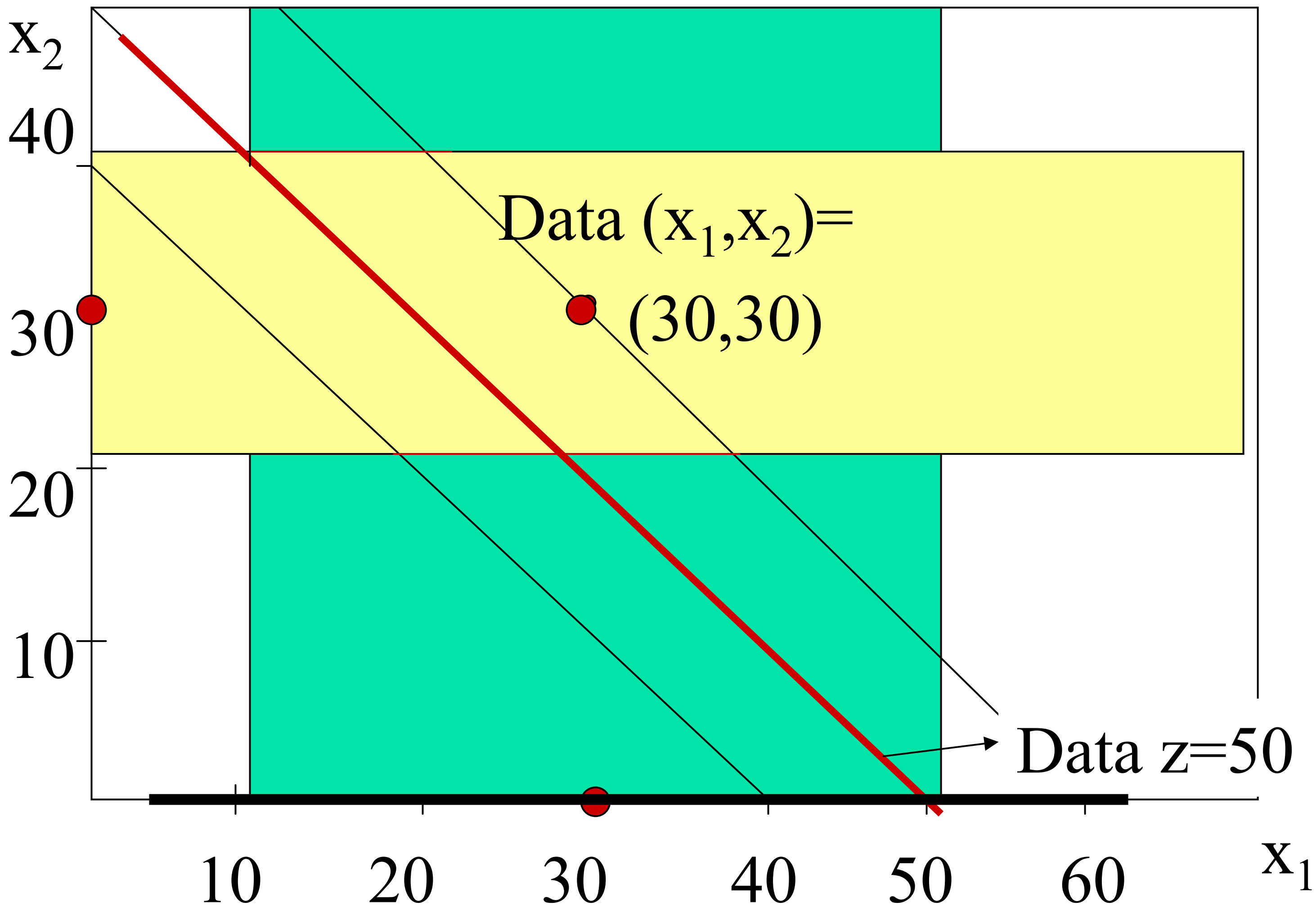
$\cap\ \text{support}_{x|g2} \cap \ldots \cap \text{support}_{x|gk}$

and $G(\tilde{x}) = 0$.

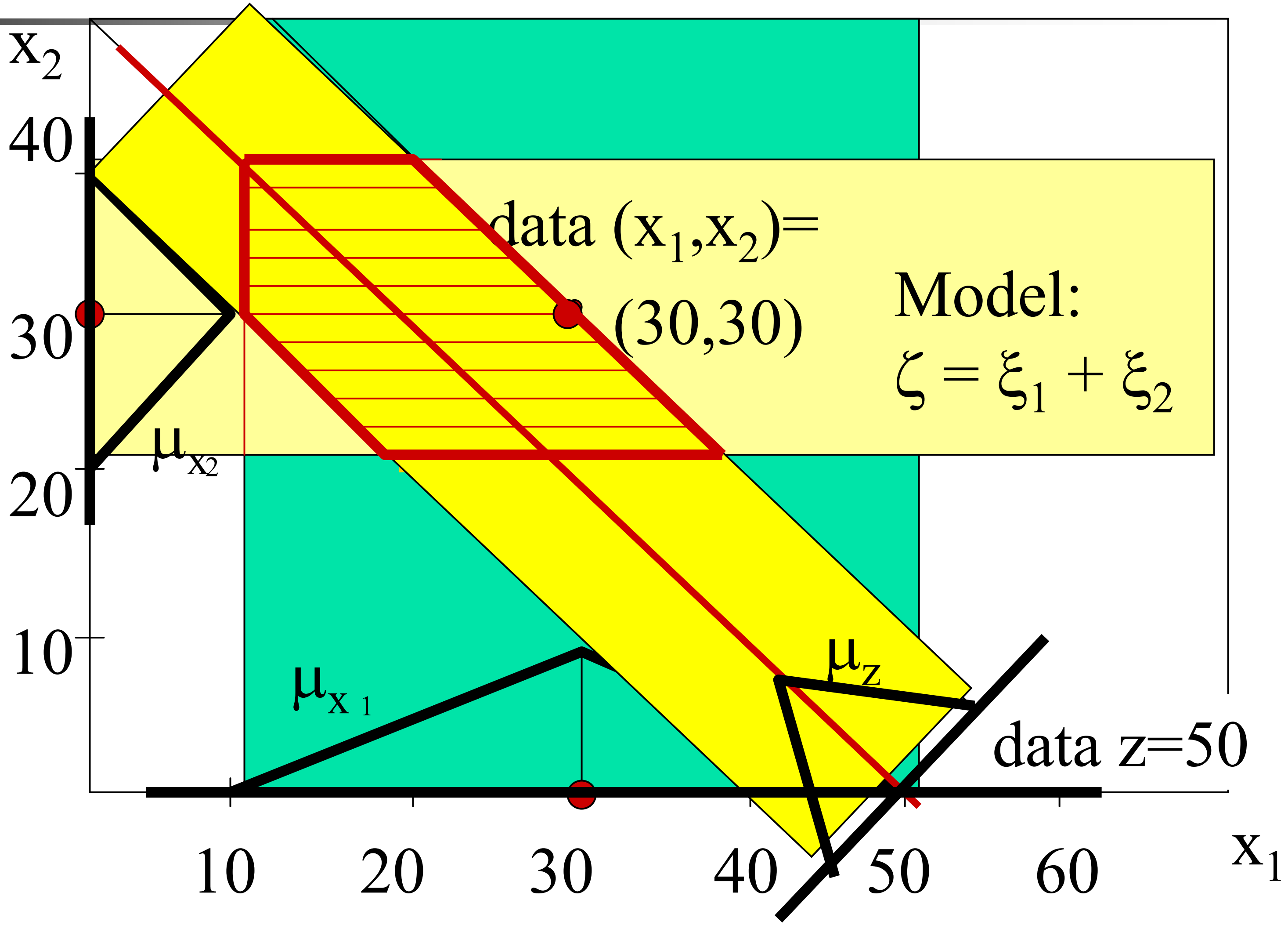
SOLD: R. Kruse (1989) Statistics on Linguistic Data;

FuzzyCalc® H.-J. Lenz and R. Müller (1999, 2000)

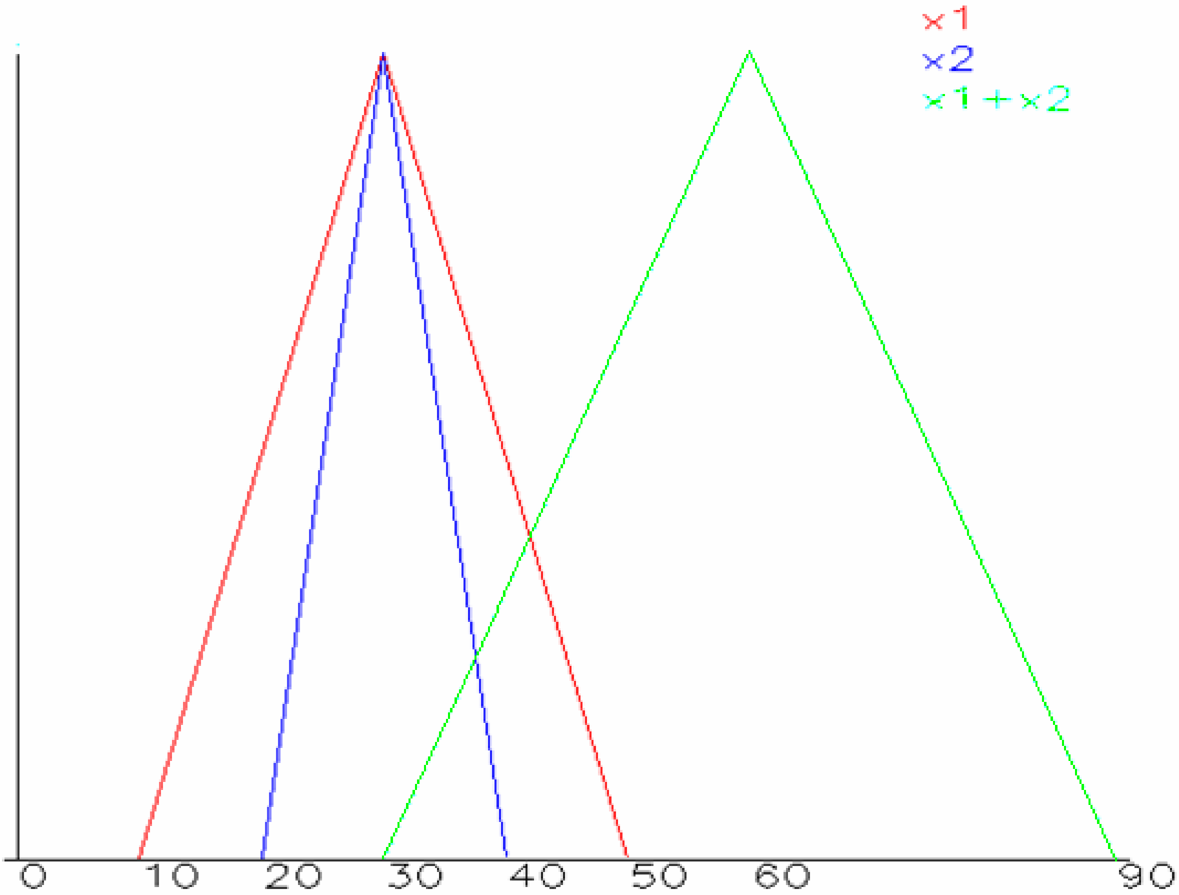
8. Fuzzy Sets Weak Inconsistency



8. Fuzzy Sets Weak Inconsistency



8. Fuzzy Edits Example 1

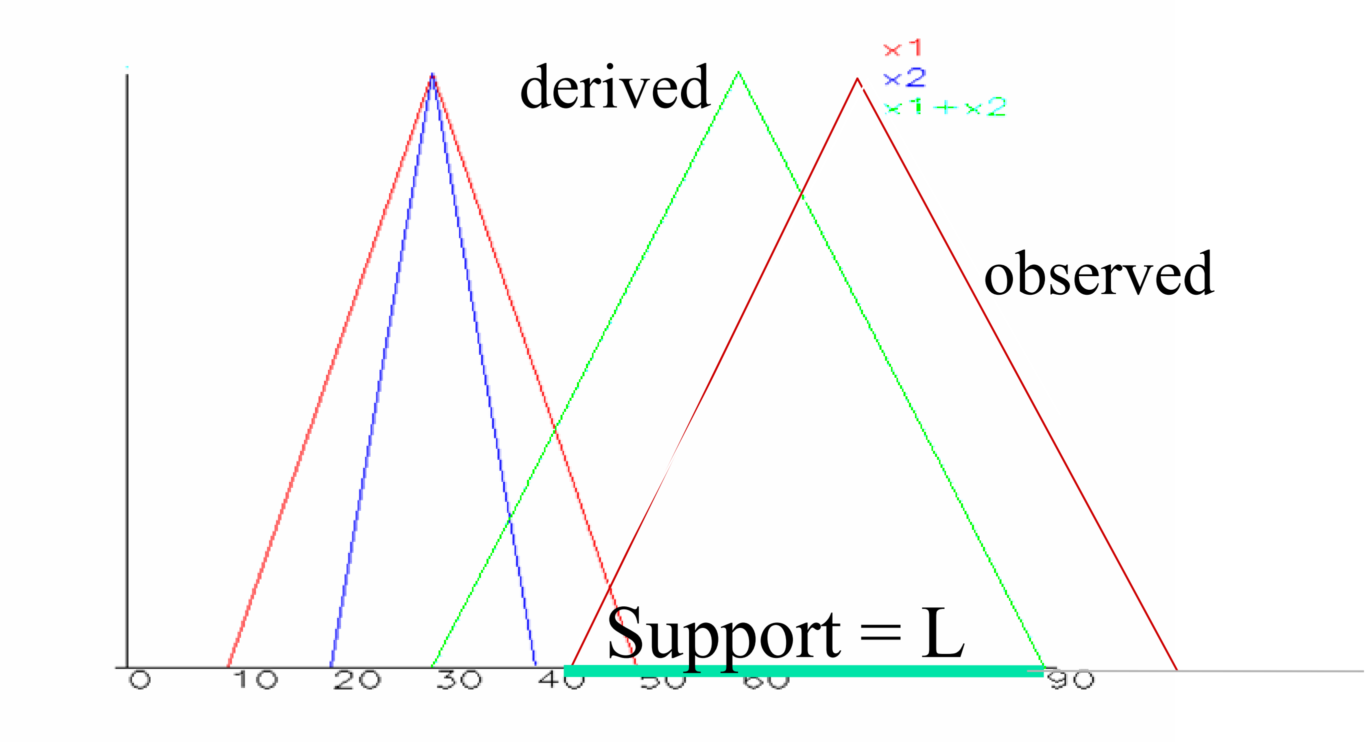


data: $x_1 = 30$, $x_2 = 30$ additive model: $\zeta = \xi_1 + \xi_2$

derived: $\widetilde{z} = 60$

8. Fuzzy Edits Example 2

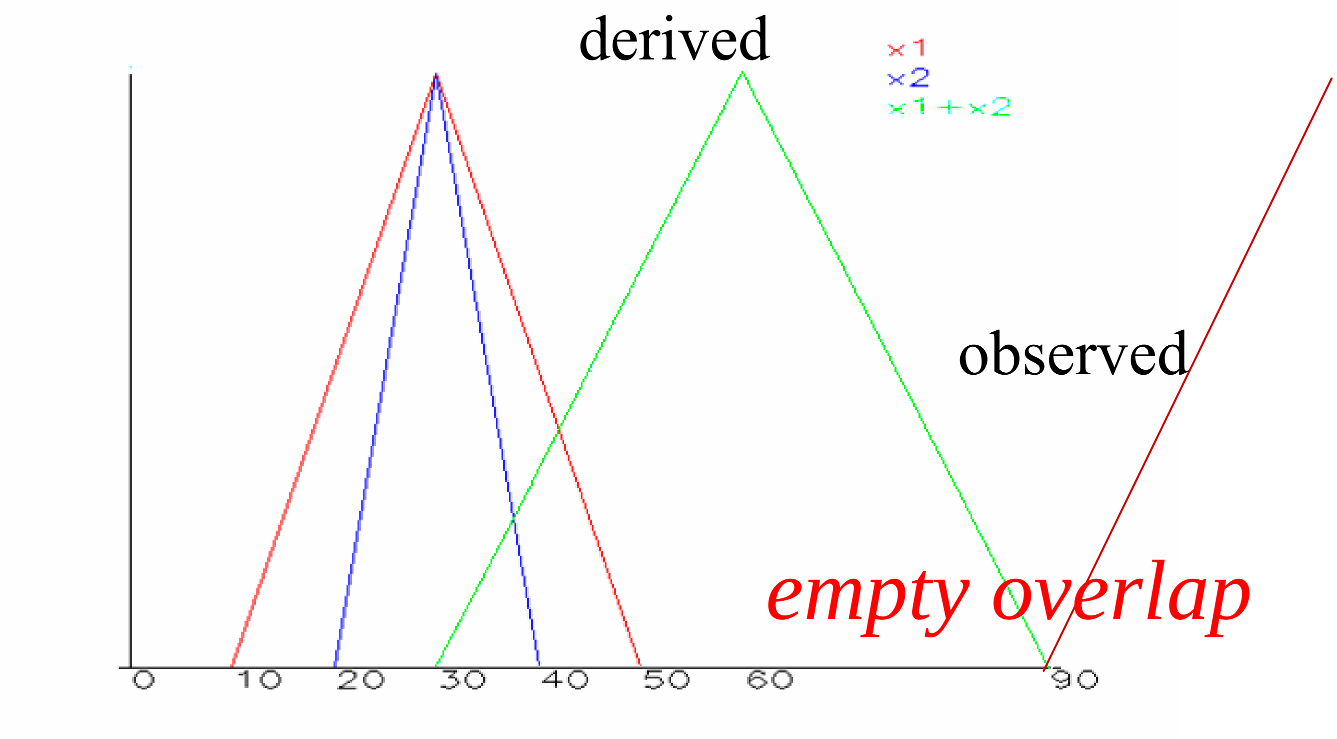
Weak Inconsistency of Data



observed: x_1, x_2, z derived: $\tilde{z} = x_1 + x_2$

8. Fuzzy Edits Example 3

Strong Inconsistency of Data



observed: x_1, x_2, x_1+x_2 derived: $\tilde{z}$ =x_1+x_2

8. Fuzzy Set Theory

Zadeh's Extension Principle

THEOREM 1: Let $p \in \boldsymbol{N}$ and $\mu_1, \mu_2, \ldots, \mu_p, \eta$ (normalized) fully specified membership functions, and $\Phi: \boldsymbol{R}^p \to \boldsymbol{R}$ a mapping with $\Phi \in \boldsymbol{T}$.

Then

$$\eta(x) := \sup \{\mu(x_1, x_2, \ldots, x_p) \mid \Phi(x_1, x_2, \ldots, x_p) = x, \quad (x_1, x_2, \ldots, x_p) \in \boldsymbol{R}^p \}$$

for all $x \in \boldsymbol{R}$ with

$$\mu(x_1, x_2, \ldots, x_p) = \min \{\mu_1(x_1), \mu_2(x_2), \ldots, \mu_p(x_p)\}$$

8. Fuzzy Set Theory

Zadeh‘s Extension Principle Example

Let p=2 and μ_1, μ_2 normalized & fully specified Membership functions, and Φ: $\boldsymbol{R}^2 \rightarrow \boldsymbol{R}$ a mapping with $\Phi = \oplus$.

Then

$$\mu(y) := \sup_{(x_1, x_2) \in \boldsymbol{R}^2} \{\min \{\mu_1(x_1), \mu_2(x_2)\} \mid x_1 + x_2 = y\}$$

for all x $\in$ $\boldsymbol{R}$.

8. Fuzzy Sets

Test Data

No	Sales	Cost	Capital	Profit	Pr		
1	100 ± 5	80 ± 4	80 ± 4	-	-	-	-
2	100 ± 10	80 ± 8	80 ± 8	-			
3	100 ± 50	80 ± 40	80 ± 40	-			
4	-	80 ± 4	80 ± 4	20 ± 1			
5	-	80 ± 8	80 ± 8	20 ± 2			
6	-	80 ± 40	80 ± 40	20 ±10			
7	100 ± 10	80 ± 8	80 ± 8	20 ± 2			
8	100 ± 10	80 ± 8	80 ± 8	30 ± 3			
9	100 ± 10	80 ± 8	80 ± 8	40 ± 4			
10	100 ± 10	80 ± 8	80 ± 8	-			
11	100 ± 10	80 ± 8	80 ± 8	-			
12	100 ± 10	80 ± 8	80 ± 8	-		0,5 ± 0,05	-
13	100 ± 5	80 ± 4	80 ± 4	30 ± 1,5	0,2 ± 0,01	0,4 ± 0,02	-
14	100 ± 10	80 ± 8	80 ± 8	30 ± 3	0,2 ± 0,02	0,4 ± 0,04	-
15	100 ± 50	80 ± 40	80 ± 40	30 ± 15	0,2 ± 0,10	0,4 ± 0,20	-

1-3:Profit missing
4-6:Sales missing
1-6: Deviations 1:2:10

7-9:Profit observed
10-12:Profitability observed

13-15:Full Data Set
(except for CapTurnover)
inconsistency decreasing

Base Model: DuPont Model

Quantor - FuzzyCalc

Benchmark : DuPont-Model

No	Sales	Costs	Cap	Profit	Profitab	ROI	Turnover	M
1	100±5	80±4	80±4	*20* *±9*	*0.2* *±0.1*	*0.25* *±0.1*	*1.25* *±0.1*	***FC***
				20 ±6.4	*.2±0.1*	*0.3 ±0.1*	*1.3 ±0.1*	***Q***
12	100±10	80±8	80±8	*35* *±3*	*0.32* *±0.04*	*0.5* ±0.05	*1.4* *±0.2/0.1*	***FC***
				37 ±*5.2*	*0.33* ± *0.04*	*0.48* ±*0.05*	*1.44* ±*0.16*	***Q***

Interval Length!

Asymmetry!

Legend: FC FuzzyCalc

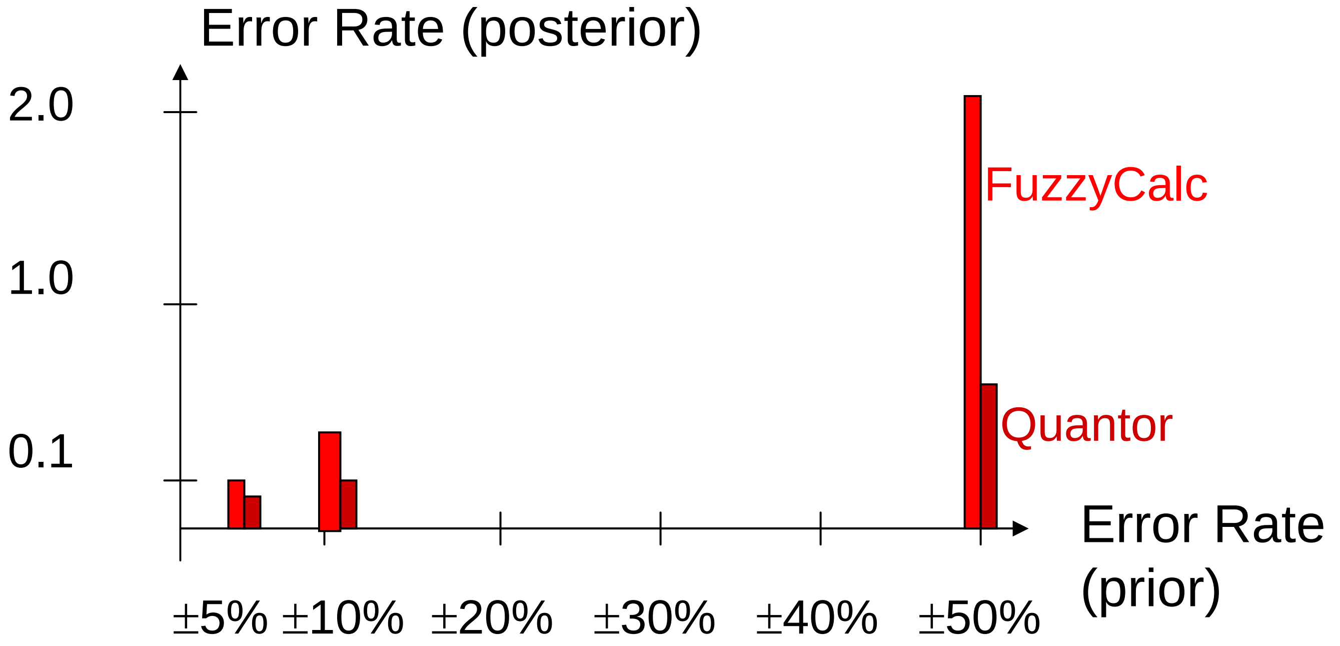
Q Quantor

Items: 100 ± 5

Measurement Errors

Effects I: Precision

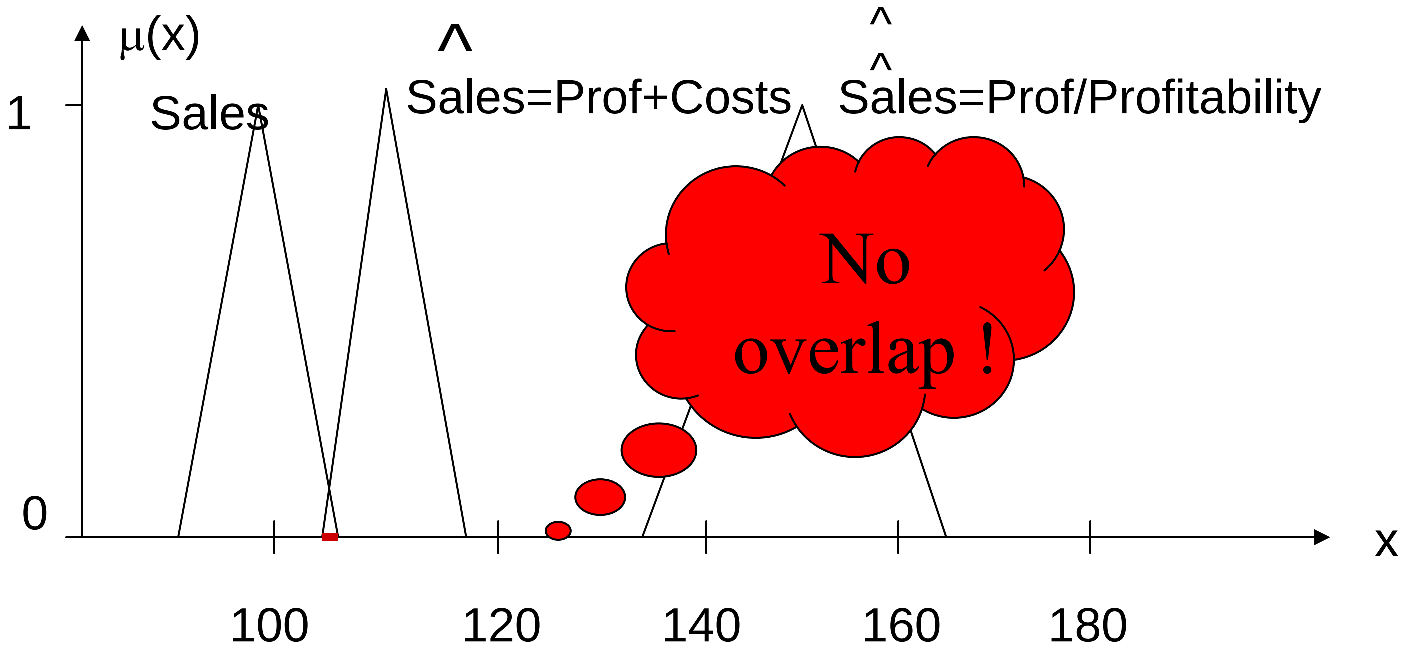
No	Sales	Costs	Capital	Profitability (**F**uzzy)	(**Q**uantor)
1	100±5	80±4	80±4	0.2 ± 0.1	0.2 ± 0.06
2	100±10	80±8	80±8	0.2 ± [0.18, 0.22]	0.2 ± 0.11
3	100±50	80±40	80±40	0.2 ± [1.6, 2.0]	0.2 ± 0.44



Measurement Errors

Effects II: Strong Inconsistency

No	Sales	Costs	Capital	Profit	Profitability	ROI	Turnover	
13	100±5	80±4	80±4	30±1.5	0,2±0,01	0,4±0,02	?	**Q**
	110±3	85±2,8	72±3	26±0,9	0,2±0,01	0,36±0,02	1,5 ±0.07	
	-	-	-	-	-	-	-	**FQ**



Computational Aspects

- *Contraction* of prior supports
- *Intuitive correct shifts* of peaks
- *Shift Distance* ~ length of prior support =>no shift of singletons
- *Invariance Property* w.r.t. model consistent data
- *Strong Inconsistency* of data <=> *empty intersection* of supports
- Corrected (=estimated) data fulfill balance equations (=model)
- Bad Scalability → Use edits at data entry!

9 MCMC Simulation

Example: Business Figures

Annual Report

Sales	Profit	Costs	ROI	Capital
55	10	45	10	60

± 20	± 2	± 20	± 5	± 1

- *Balance Equations*
- Sales = Profit + Costs (linear equation)
- ROI = 100 * Profit / Capital (nonlinear equation)

- *Measurement Errors Model*

$$x = \xi + u \quad with\ u \sim N(0, \Sigma_{uu})$$

MCMC Simulation & Estimation

Given

M structural equation system (balance equations) of fully separable variables

x observed state vector (missing values allowed)

f_x fully specified density function of x (given mean and standard deviation)

C_{xx} correlation matrix

Wanted

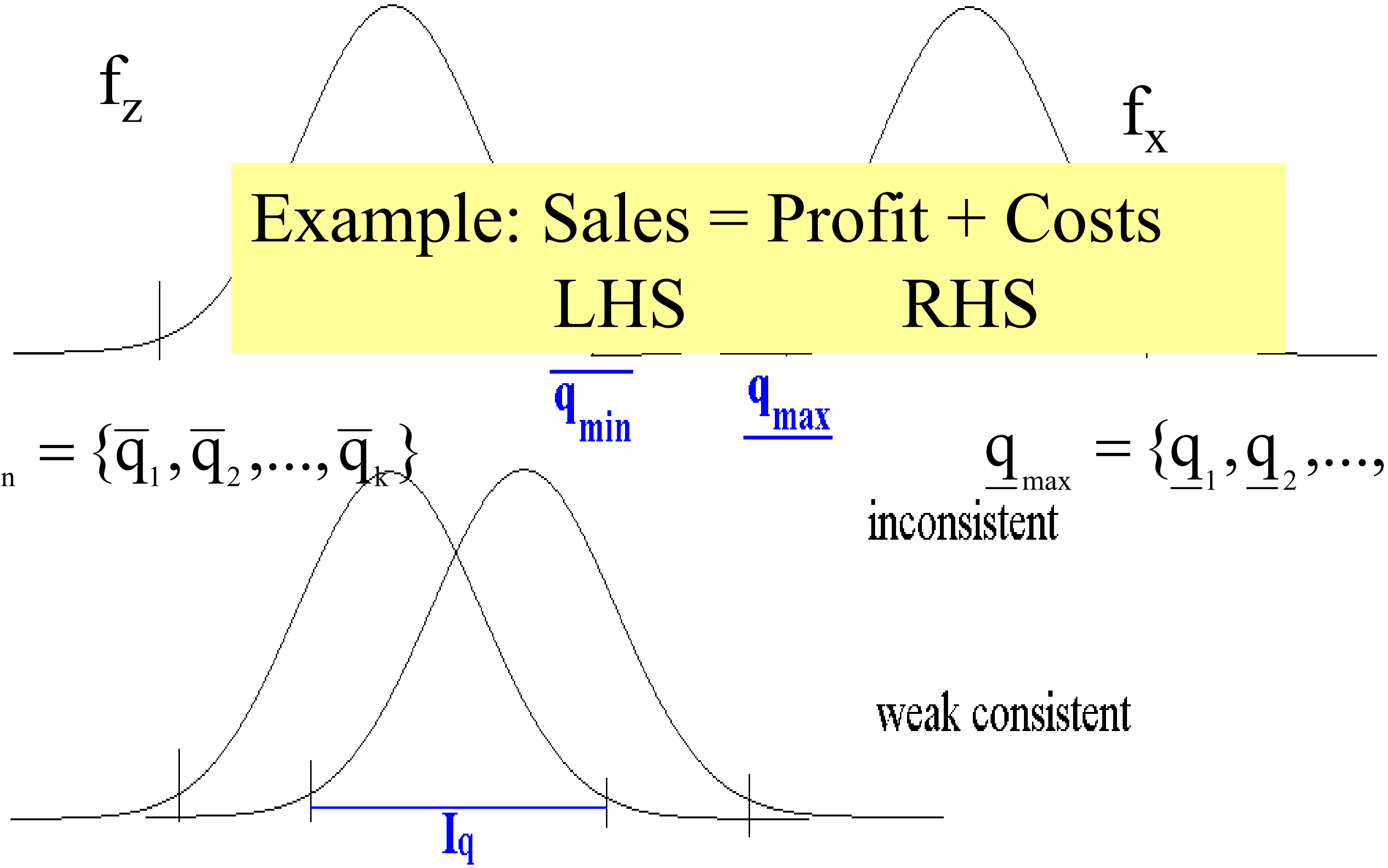
$\hat{f}, \hat{\xi}, \hat{\sigma}$ density, mean and standard deviation for each state variable

SamPro-Algorithm

Resoving, Sampling, Estimation and Projection

```
begin
resolve (set) LHS ≡ RHS (x) for all variables of all
        equations
sample from the joint density function of all RHS
        variables for each LHS (z)
estimate f, $\mu$, $\sigma$ of all LHS variables
estimate $\alpha/2$, 1-$\alpha/2$-quantiles $\underline{q}_{max}$ ,$\overline{q}_{min}$ for each variable,
            i=1,2,…,p .
if $\overline{q}_{min} < \underline{q}_{max}$ flag “M -inconsistency” else
compute the distribution $\hat{f}_{xz}$ restricted to the subspace
     x-z = 0, and $\mu_x$ and $\sigma^2$ (in case of k= 2 equations).
end
```

M-Inconsistency (k = 2)



Example: Sales = Profit + Costs
LHS RHS

$$\overline{q}_{min} = \{\overline{q}_1, \overline{q}_2, \ldots, \overline{q}_k\}$$

$$\underline{q}_{max} = \{\underline{q}_1, \underline{q}_2, \ldots, \underline{q}_k\}$$

$$\boxed{I_q = [\underline{q}_{max}, \overline{q}_{min}]}$$

6 Experiments and Analysis

Experimental Group A

(missing values allowed)

1. Single Effect of skewness
2. Effect of correlation
3. Interaction Effect of Skewness and Correlation

Experimental Group B

(complete data sets, Gaussian distributions, correlation between variables)

1. Effect of M-consistency
2. Effect of M-inconsistency

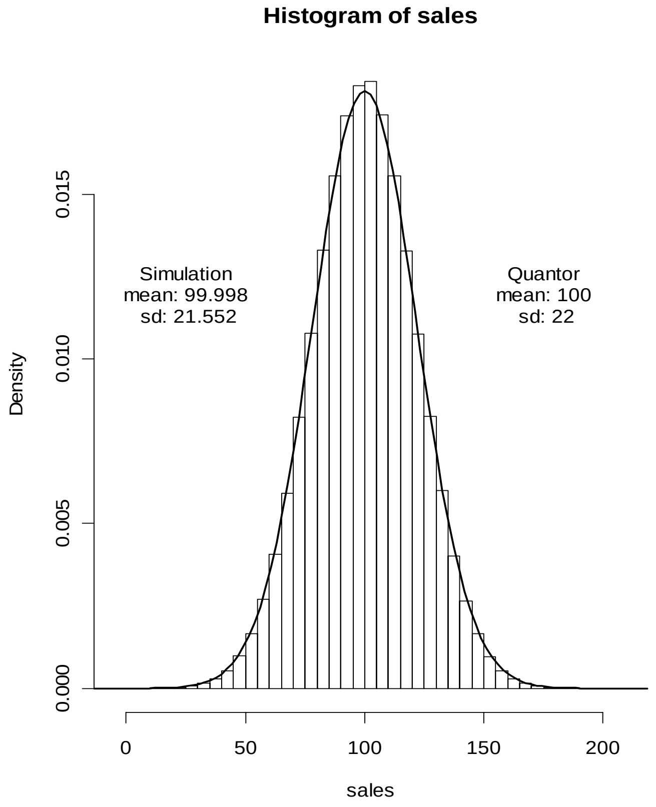
Exp.-Group 1: Scenario 1

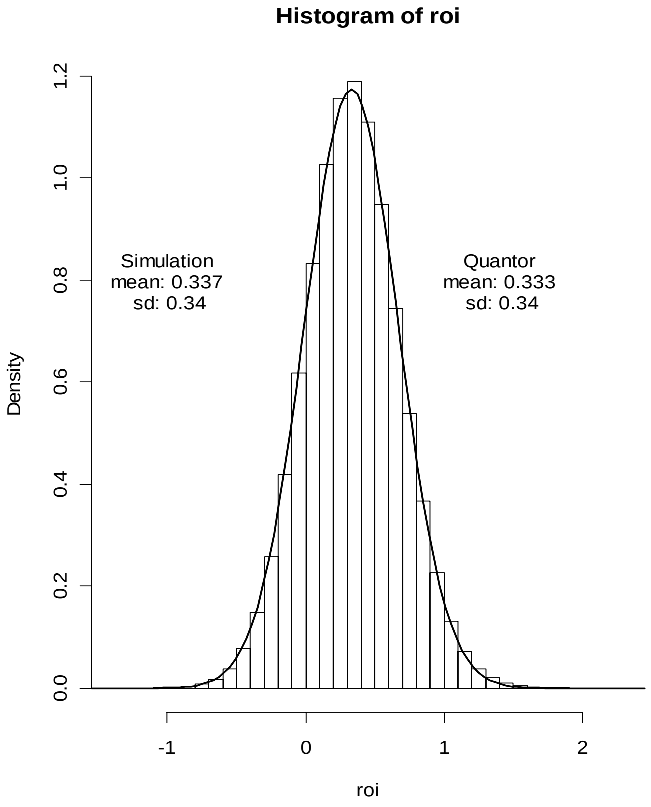
Normality - no Correlation

***Specification of the distributions**:*

Profit ~ N(20,2^2); Costs ~ N(80, 8²); Capital ~ N(60, 6²)

***Missing values**: Sales, ROI*



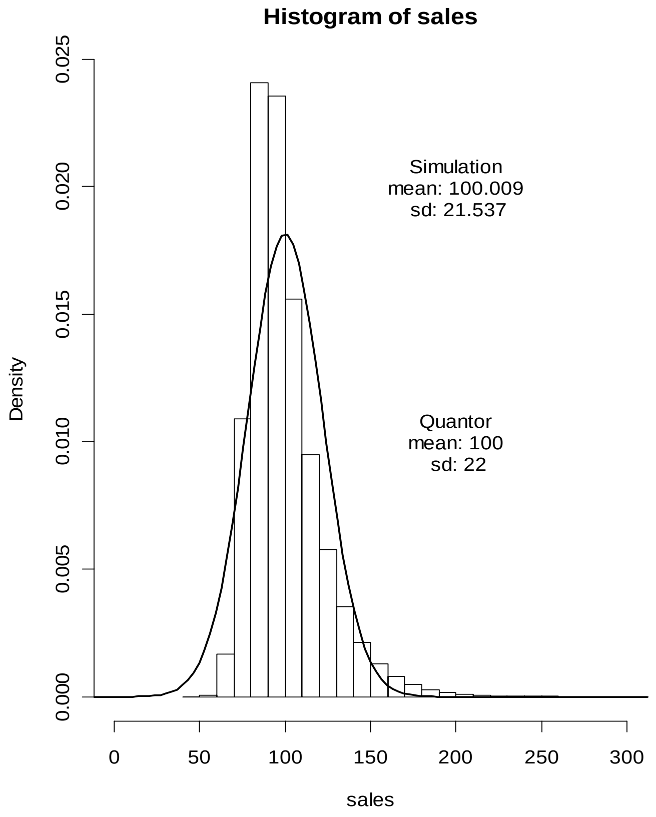


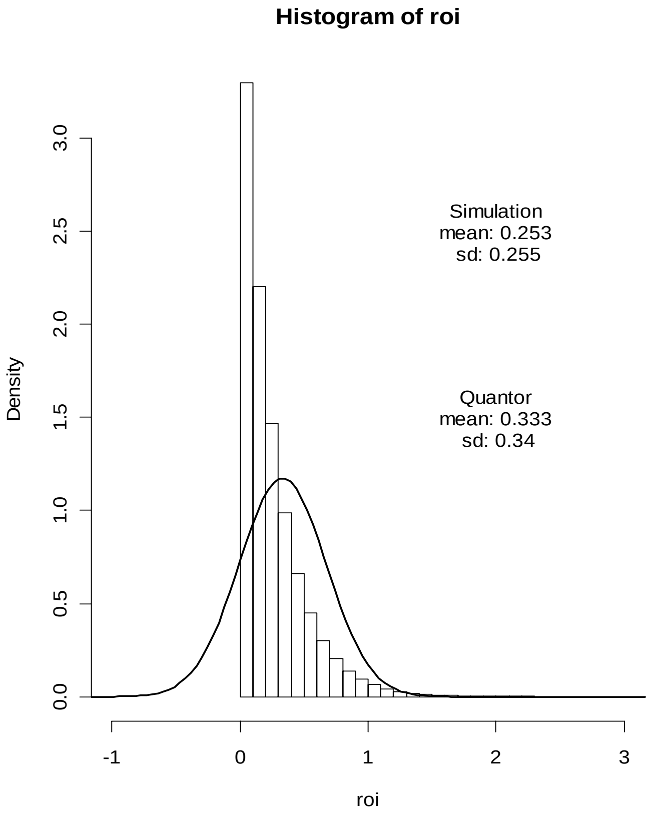
Scenario 2

Skewness - no correlation

***Specification of the distributions**: Profit ~ Exp(1/20) vs. N(20, 20^2); Costs ~ N(80, 8^2); Capital ~ N(60, 6^2)*

***Missing values**: Sales, ROI*



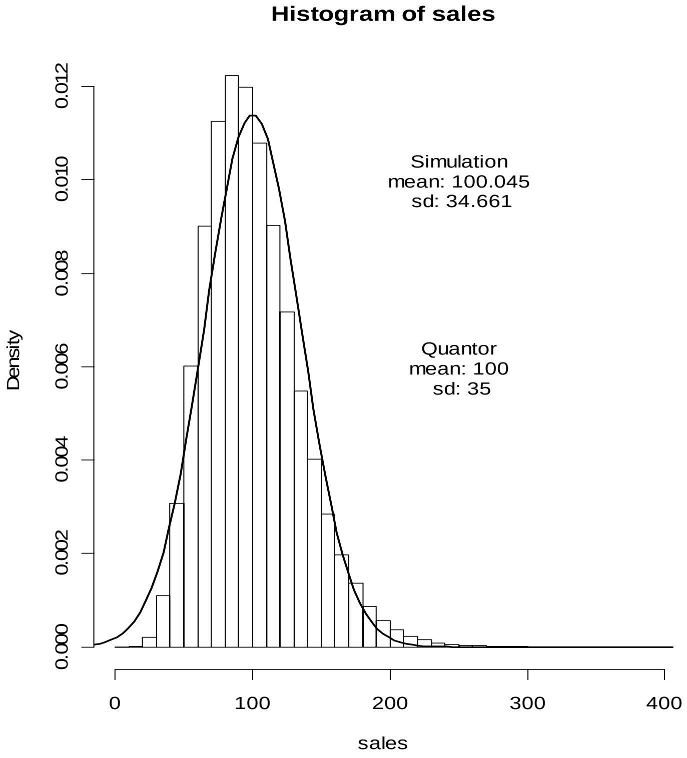


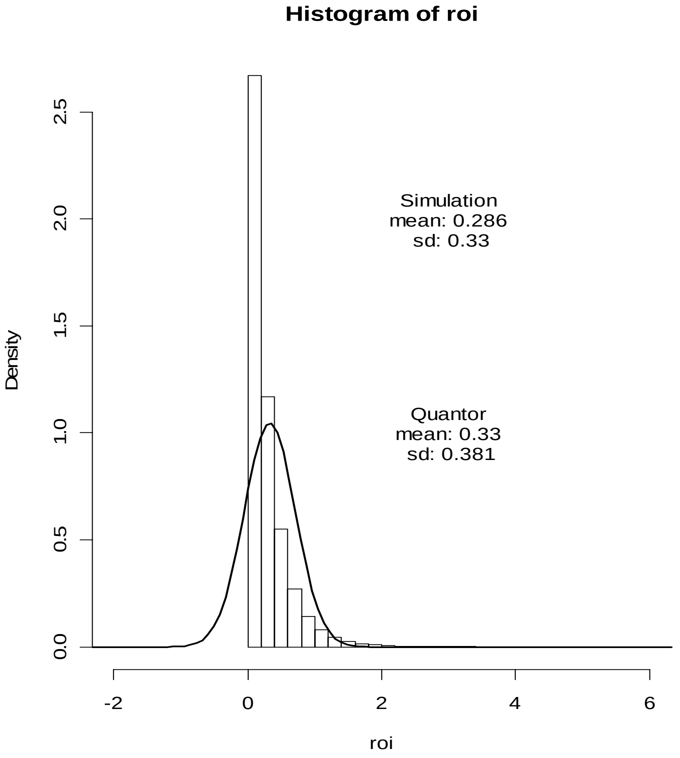
Scenario 3

Skewness - no Correlation

***Specification of the distributions**: Profit ~ Exp(1/20) vs. N(20, 20^2); Costs ~ Gamma(8,0.1) vs. N(80, 28^2); Capital ~ Gamma(15, 0.25) vs. N(60, 8.6^2)*

***Missing values**: Sales, ROI*


Histogram of sales


Histogram of roi

Scenario 4

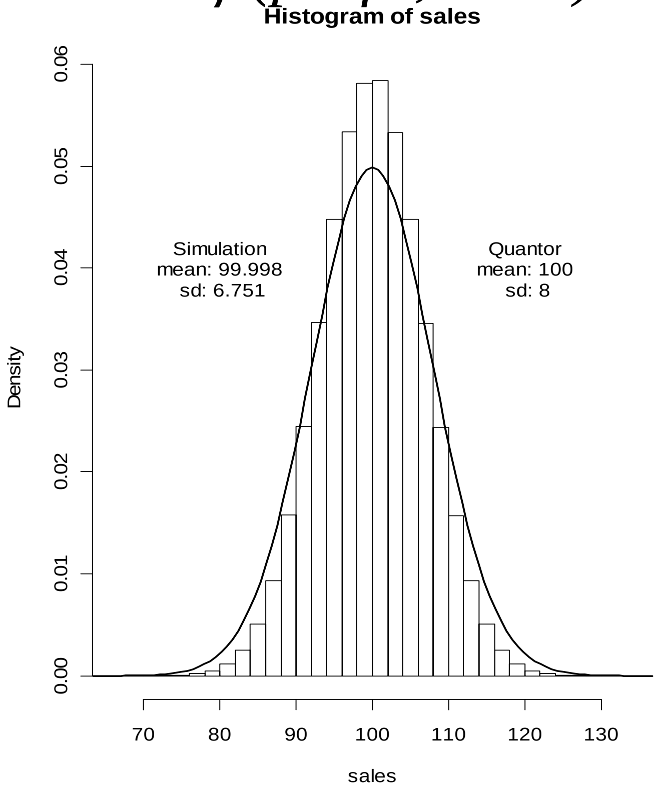
Normality - neg. Correlation

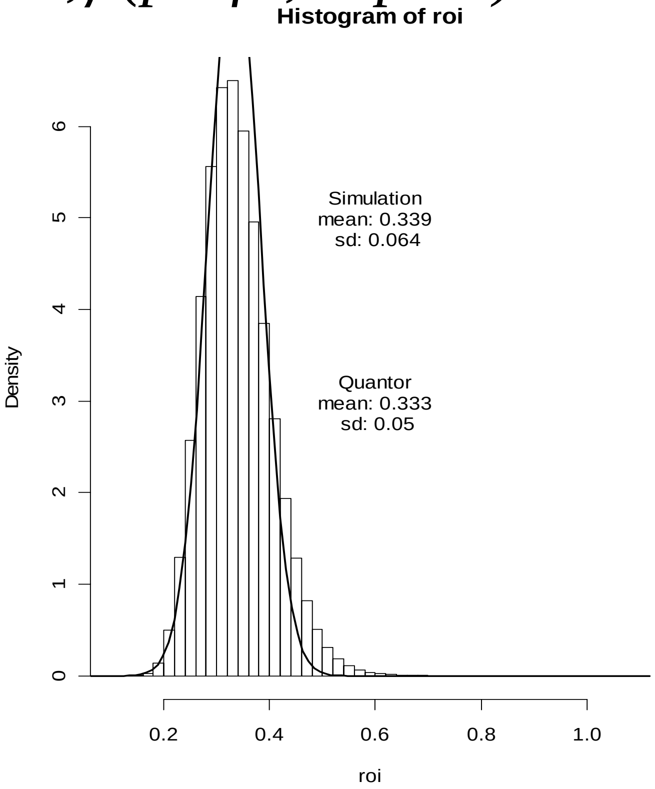
***Specification of the distributions**:*

Profit ~ N(20, 2^2); Costs ~ N(80, 8^2); Capital ~ N(60, 6^2)

***Missing values**: Sales, ROI*

***Correlation:** ρ(profit, costs) = -0.7;ρ(profit, capital) = -0.7*





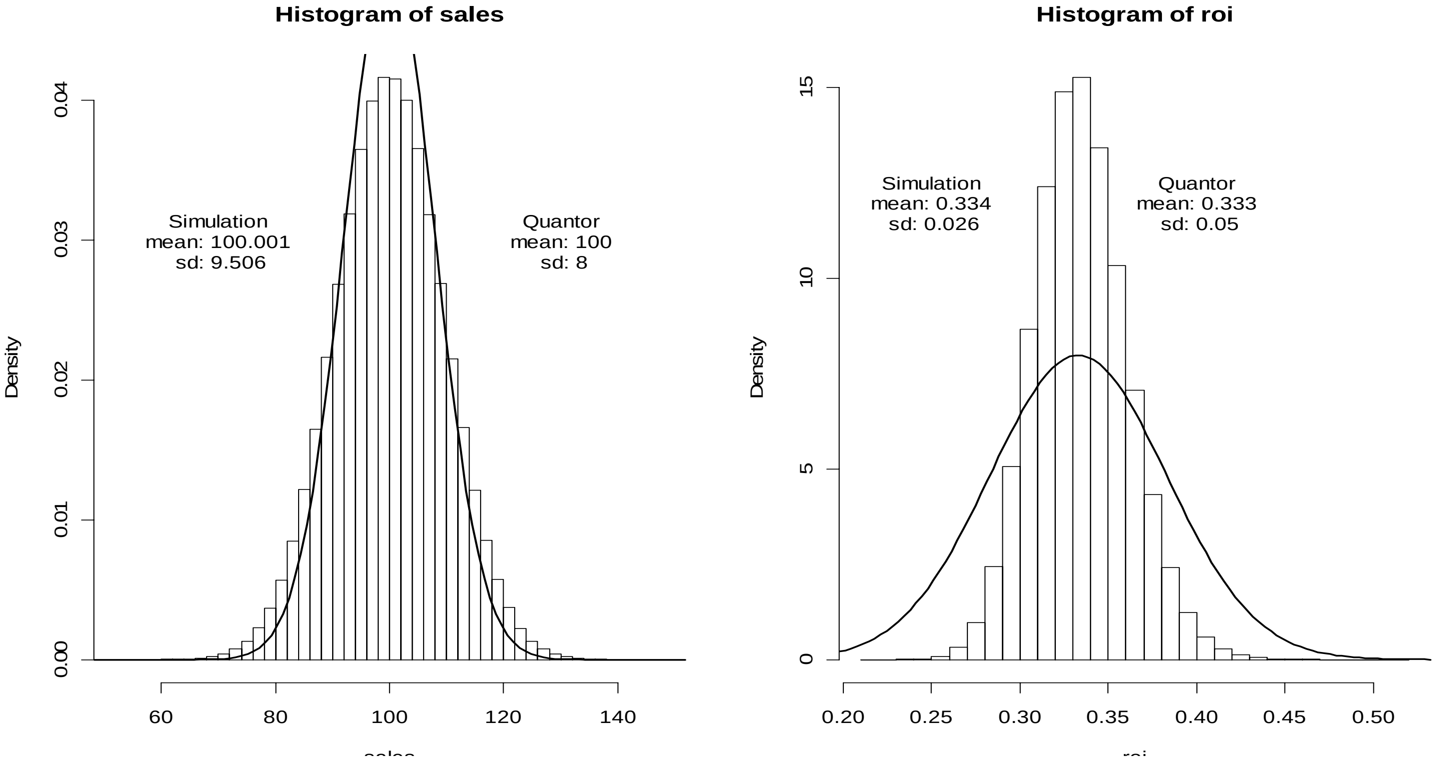
Scenario 5

Normality - pos. Correlation

***Specification of the distributions**: Profit ~ N(20, 2^2); Costs ~ N(80, 8²); Capital ~ N(60, 6²)*

***Missing values**: Sales, ROI*

***Correlation:** ρ(profit, costs) = 0.7; ρ(profit, capital) = 0.7*



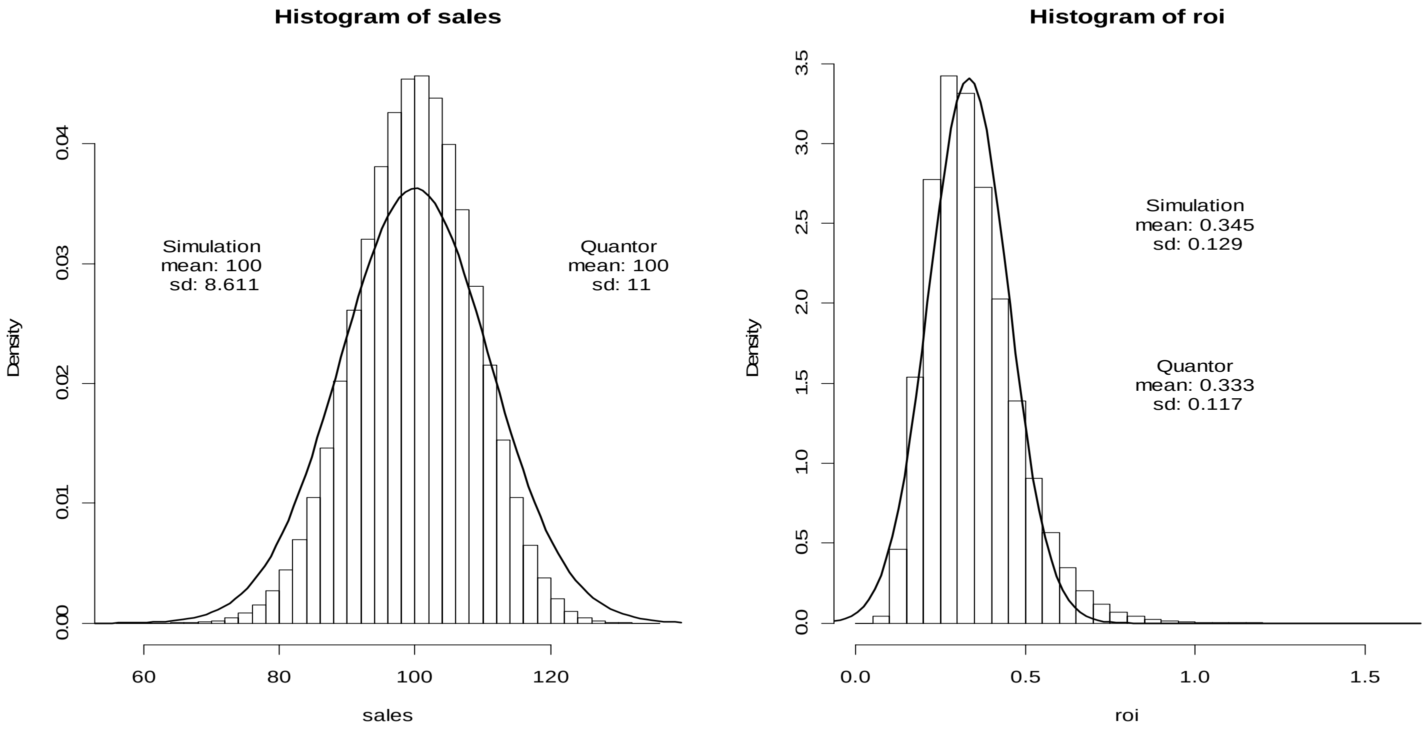
Scenario 6

Skewness - neg. Correlation

***Specification of the distributions**: (Profit, Costs, Capital) ~ Dir(10, 40, 30) vs. Profit ~ N(20, 5.9²); Costs ~ N(80, 8.9²); Capital ~ N(60, 8.6²)*

***Missing values**: Sales, ROI*

***Correlation:** ρ(profit, costs) = -0.4; ρ(profit, capital) = -0.3*



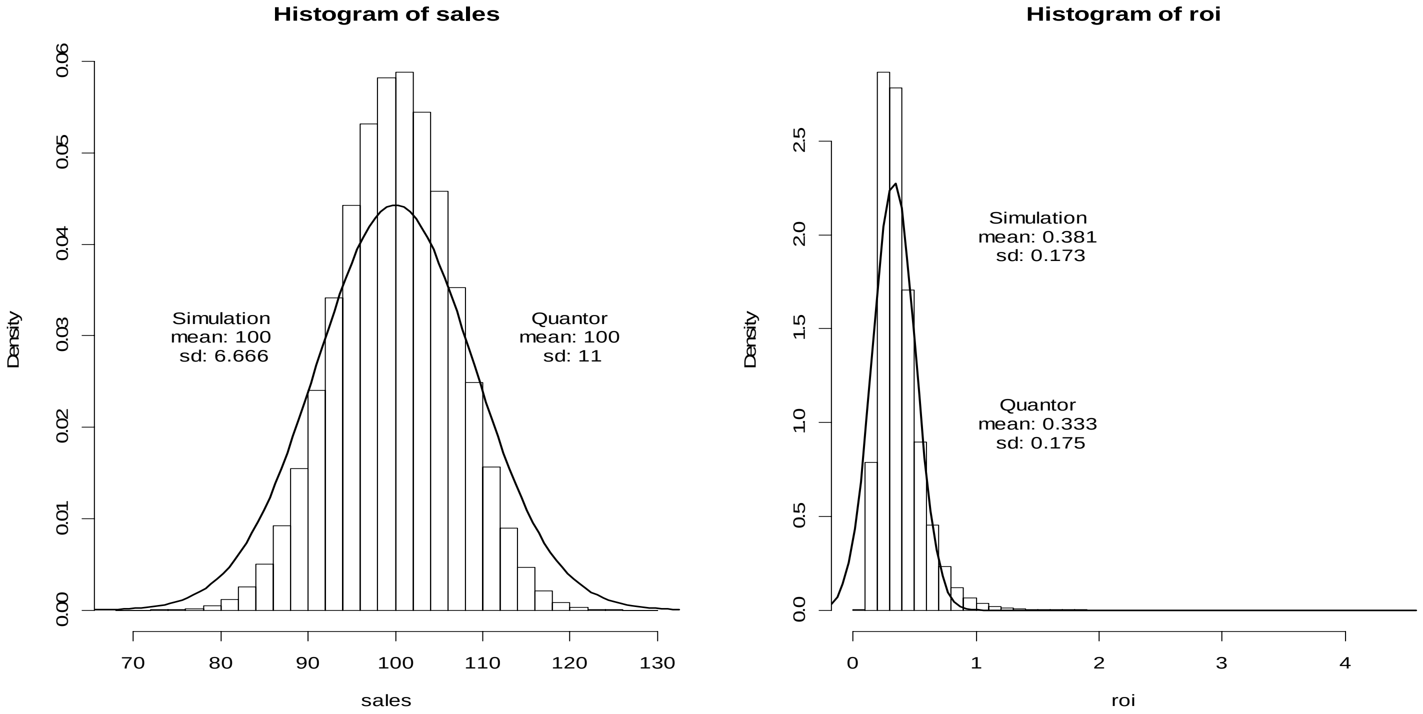
Scenario 7

Skewness - pos. Correlation

***Specification of the distributions**:(Profit, Costs, Capital) ~ Dir(30, 40, 8) vs. Profit ~ N(20, 2.8²); Costs ~ N(80, 8.8²); Capital ~ N(60, 20²)*

***Missing values**: Sales, ROI*

***Correlation:** ρ(profit, costs) = -0.8; ρ(profit, capital) = -0.3*



Exp.-Group B:
Normality, Correlation

Prior Information:
Sales and ROI M-consistent vs. M-inconsistent

Missing values*: no*

Correlation Matrix (ρ ∈{-0.4, 0.0, + 0.4})

$$R = \begin{bmatrix} 1 & \rho & 0 & -\rho & -\rho \\ \rho & 1 & 0 & -\rho & \rho \\ 0 & -\rho & 1 & 0 & \rho \\ -\rho & -\rho & 0 & 1 & -\rho \\ -\rho & \rho & \rho & -\rho & 1 \end{bmatrix}$$

Scenario 1

Normality, Correlation

Specification of the distributions: *Profit ~ $N(20,2^2)$; Costs ~ $N(80, 8^2)$; Capital ~ $N(60, 6^2)$; Sales ~ $N(100, 10^2)$; ROI ~ $N(0.333, 0.0333^2)$*

Missing values*: no*

Special Set-up:

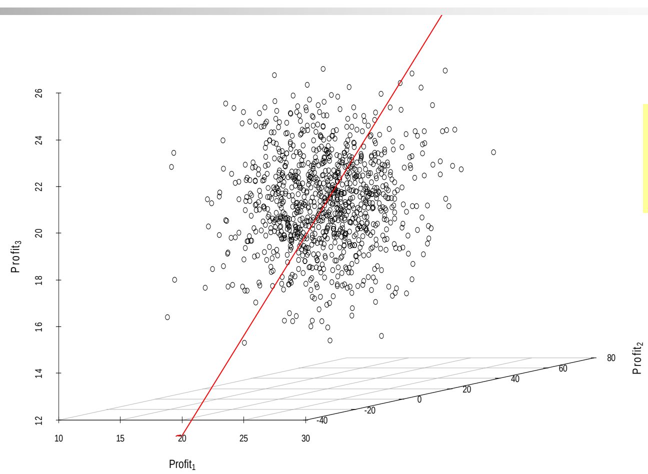
- M-consistent observations of Sales, ROI
- $\rho = 0$

Variable	**Mean**	**Stdv**
profit	19.97	1.58
costs	79.95	6.29
capital	59.81	4.89
sales	99.96	6.37
ROI	0.33	0.03

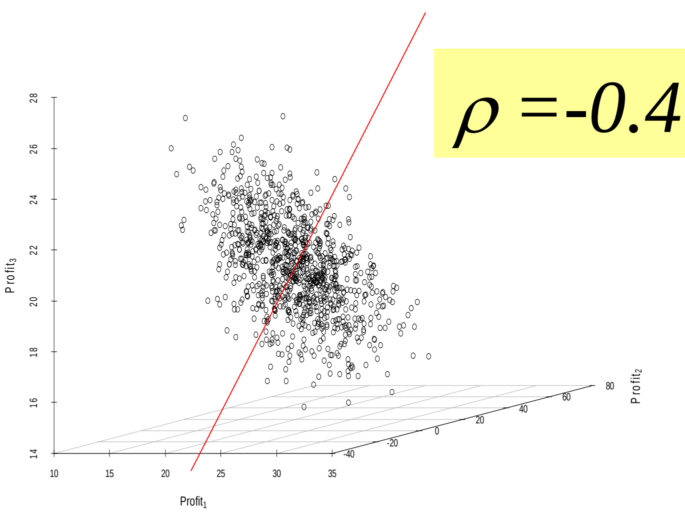
Scenario 1

3D scatter plots of Profit

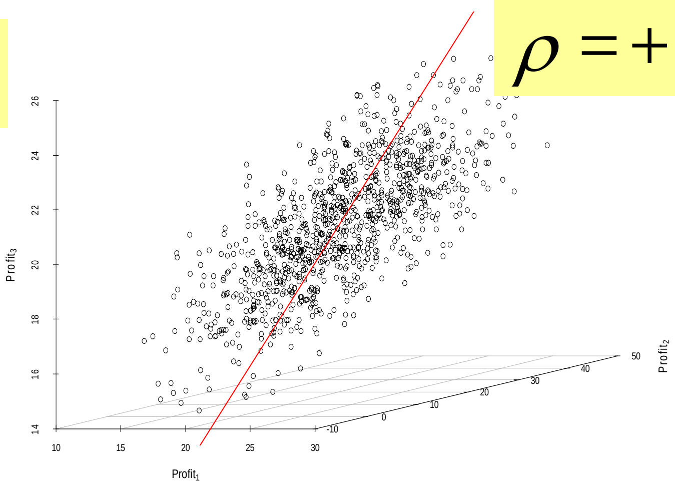
1. Profit $\sim N(20,2^2)$
2. Profit = Sales -Costs
3. Profit = Capital * ROI


Common Distribution of Profit

$\rho = 0$


Common Distribution of Profit

$\rho = -0.4$


Common Distribution of Profit

$\rho = +0.4$

Scenario 2

Normality, Correlation

Specification of the distributions: Profit ~ N(30,3^2); Costs ~ N(80, 8^2); Capital ~ N(60, 6^2); Sales ~ N(100, 10^2); ROI ~ N(0.333, 0.333^2)

Missing values: No

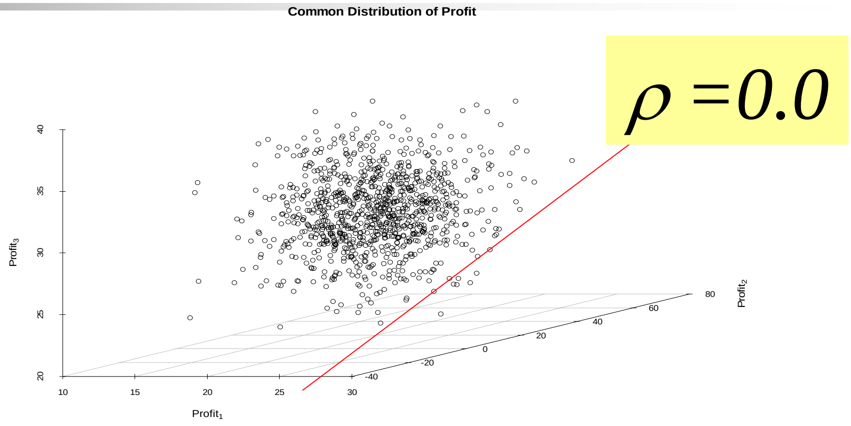
Special Set-up:

- M-inconsistent observations of Sales, ROI
- $\rho = 0$

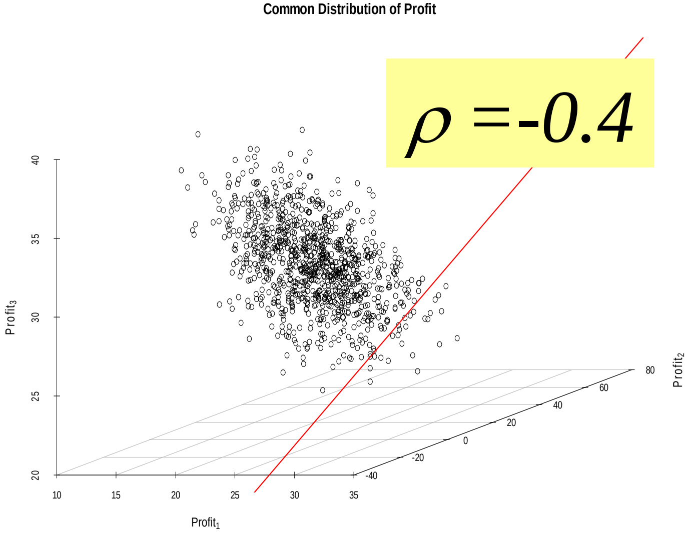
Variable	**Mean**	**Stdv**
profit	24.84	2.18
costs	76.27	6.36
capital	67.09	5.05
sales	105.74	6.50
ROI	0.37	0.03

3D Scatterplots of Profit

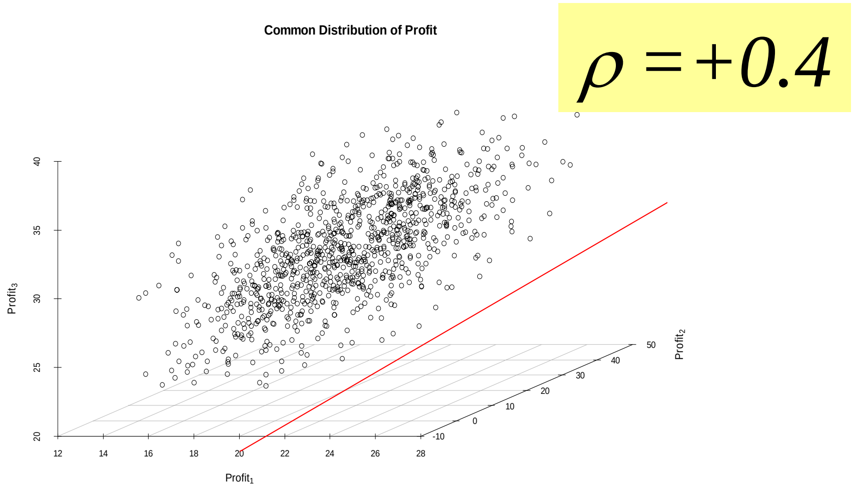
1. Profit ~ $N(30,3^2)$
2. Profit = Sales - Costs
3. Profit = Capital * ROI



$\rho = 0.0$



$\rho = -0.4$



$\rho = +0.4$

strong M-inconsistency !

Summary

1. No correlation:
   (the means of) the simulated quantities are about the same as the GLS estimates under a Gaussian hypothesis.

2. Skewness of distributions:
   mostly has only a small effect on the estimates.

3. Positive cross-correlations of the variables:
   can lead to severe problems: The balance equation system may become M-inconsistent !

Basic Literature

1. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methods and Techniques. Heidelberg: Springer Verlag (2006)
2. Dombrowski, Erik und Lechtenböger, Jens: Evaluation objektorientierter Ansätze zur Data-Warehouse-Modellierung, Datenbank-Spektrum 15/2005
3. Naumann, Felix: Datenqualität, Informatik-Spektrum_30_1_2007
4. Naumann, Felix: Quality-driven Query Answering for integrated Information Systems, LNCS 2261, Springer, Heidelberg, 2002

Edits – A Firewall against inconsistent Data Entry

- Thank you!

- hjlenz@wiwiss.fu-berlin.de
- http://www.wiwiss.fu-berlin.de