StatisticalData Editing

Anders Norberg,
Statistics Sweden (SCB)

2010-11-23

Papers by our colleague Leopold Granquist

- Granquist (1984). On the role of editing. Statistical Review 2
- Granquist (1997). The New View on Editing. International Statistical Review
- Granquist and Kovar (1997). Editing of Survey Data: How Much is Enough? In Survey Measurement and Process Quality. Wiley.

If …

- we only want information from businesses that we know they have,
- and we ask for that information so they understand,
- and we motivate them to deliver as good quality in data as possible,
- and we help them to avoid accidental errors in answering questionnaires,
- then editing would be a minor process!

The role of editing

- Quality Control of the measurement process
  - Find errors (efficient controls)
  - Consider every identified error as a problem for the respondent to deliver correct data by our collection instrument
  - Identify sources of error (process data)
  - Analyse process data – communicate with cognitive specialists
- Contribute to quality declaration
- Adjust *(change/correct)* significant errors

Types of errors

- Obvious errors / Fatal errors
  - Non-valid values
  - Item non-response
  - Data structure- or model errors, total ≠ sum of components
  - Contradictions

- Suspected data values
  - Deviation errors (Outliers)
    - Suspiciously high/low values, data outside of predetermined limits
  - Definition errors (Inliers)
    - Many respondent miss-understand a question in the same way
    - Many respondents fetch data from info-systems with other definitions

Suspected data values

Deviation errors

- Manual follow-up takes time and is expensive
- Few deviation errors have impact on output statistics (low hit-rate, many changes in data have very little impact)

Editing must have impact on the output!

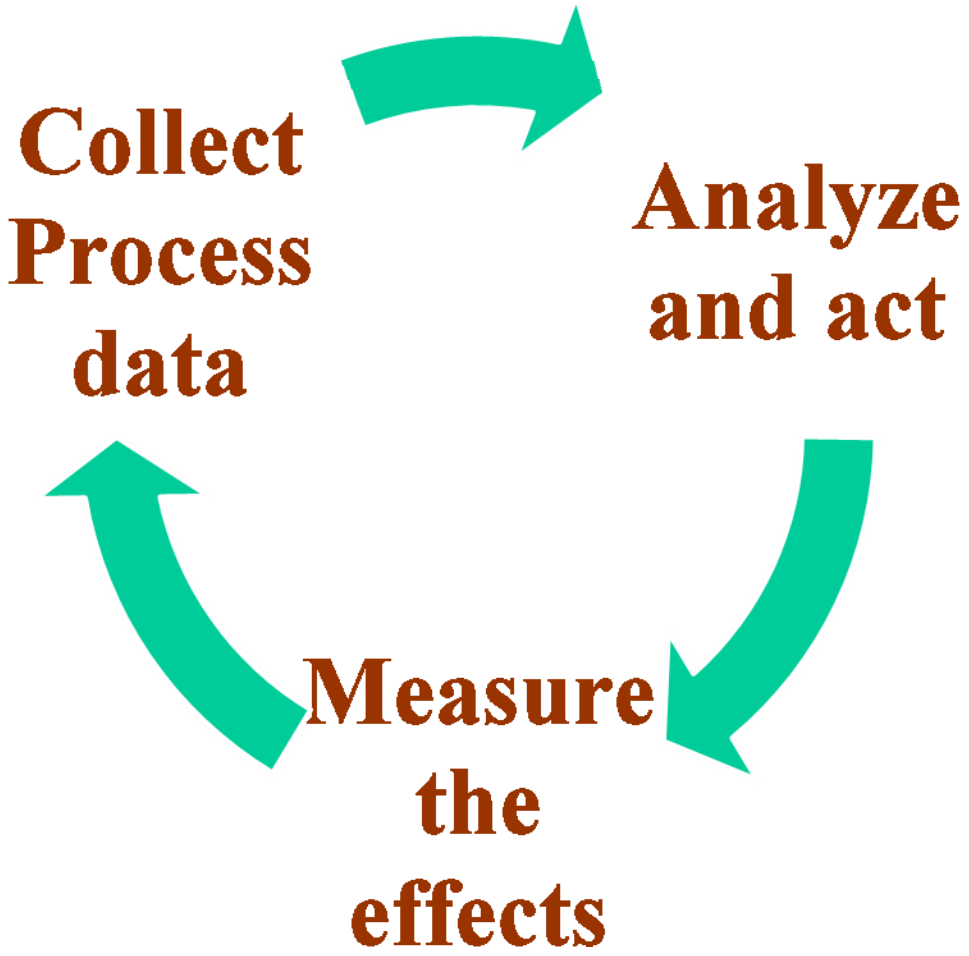
Remember response burdon !

Suspected data values

Definition errors (Inliers)

- Difficult to find
- Ways to find them:
  - Combined editing for several surveys
  - Deep interviews in focus groups
  - Use statistics from FEQ and from re-contacts with respondents
  - High proportions of item non-response
  - Graphical editing
  - Good examples

The Process Perspective



The Process Perspective

- Audit and improve data collection (measurement instrument and collection process)
- and the editing process itself

Un-edited data must be saved in order to produced important process indicators, as hit-rate and impact on output!

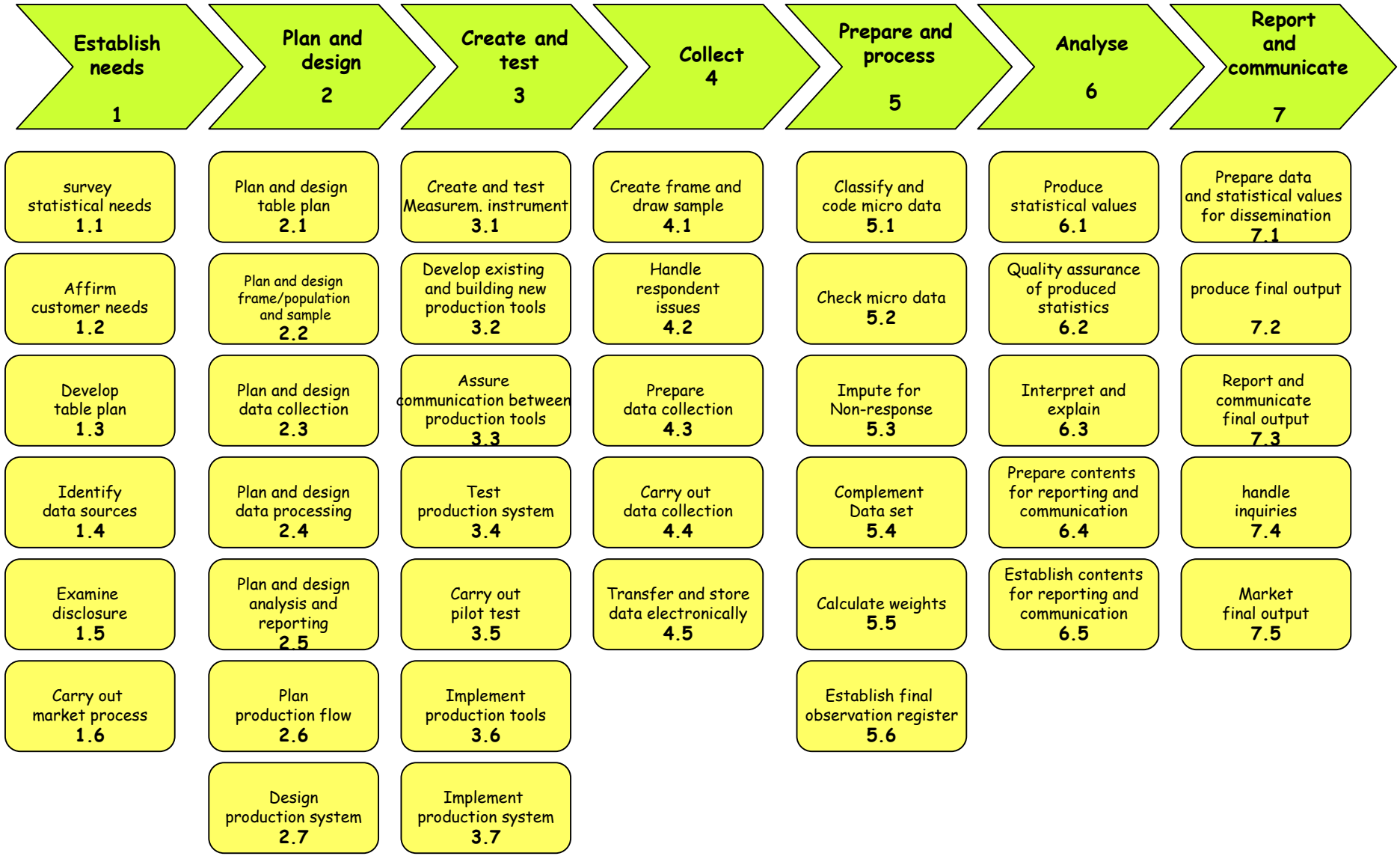
Process data

- Sources of errors = problem for the respondents
- Suspicions
- Error codes
- Manuel actions (accept / amended values)
- Automatic actions

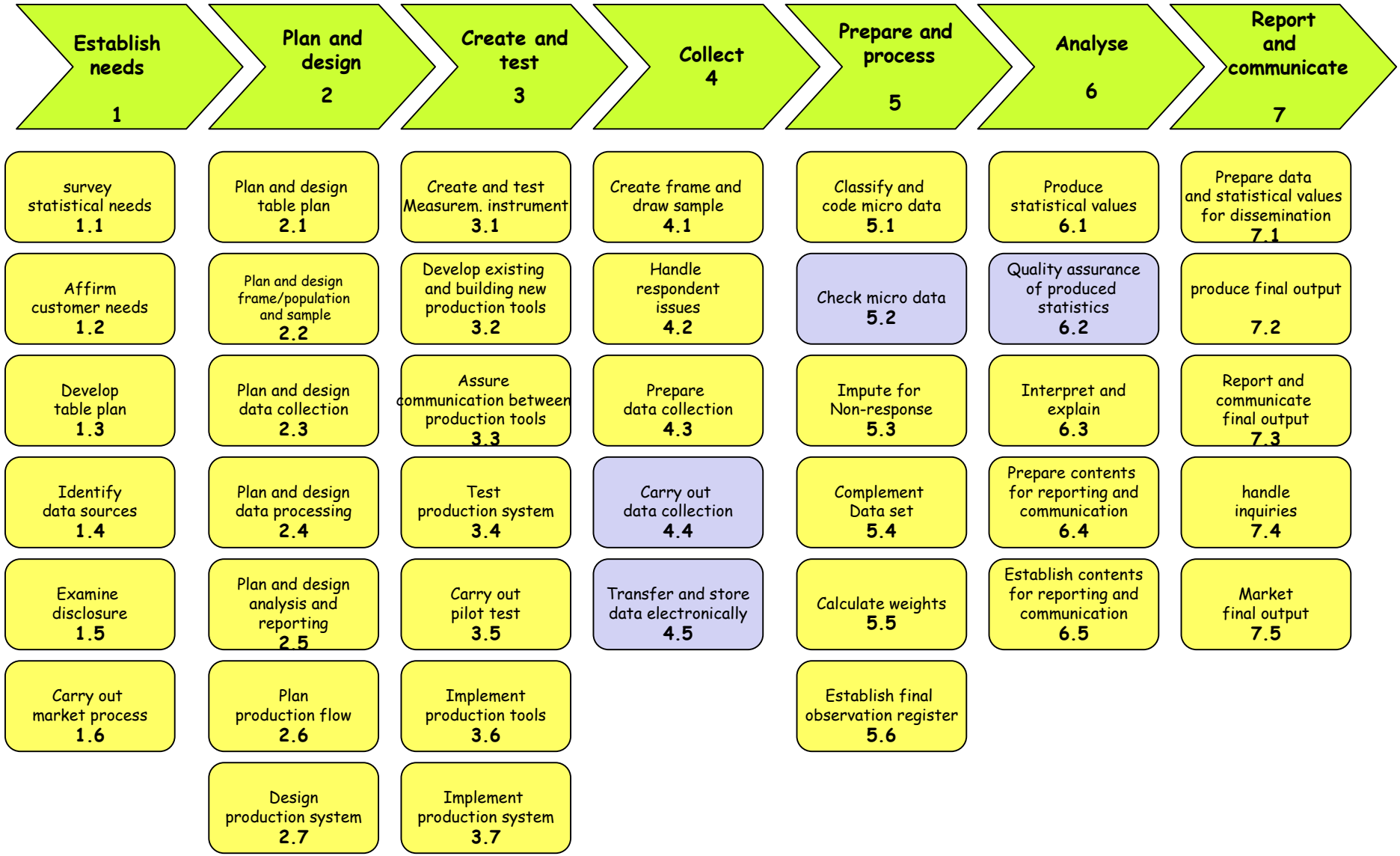
Process indicators

- Sources of errors (problem for the respondents)
- Prop. of flagged units and variables
- Prop. of manually and automatically reviewed units and variables
- Prop. of amended values and impact of the changes, per variable
- Hit-rate for edits

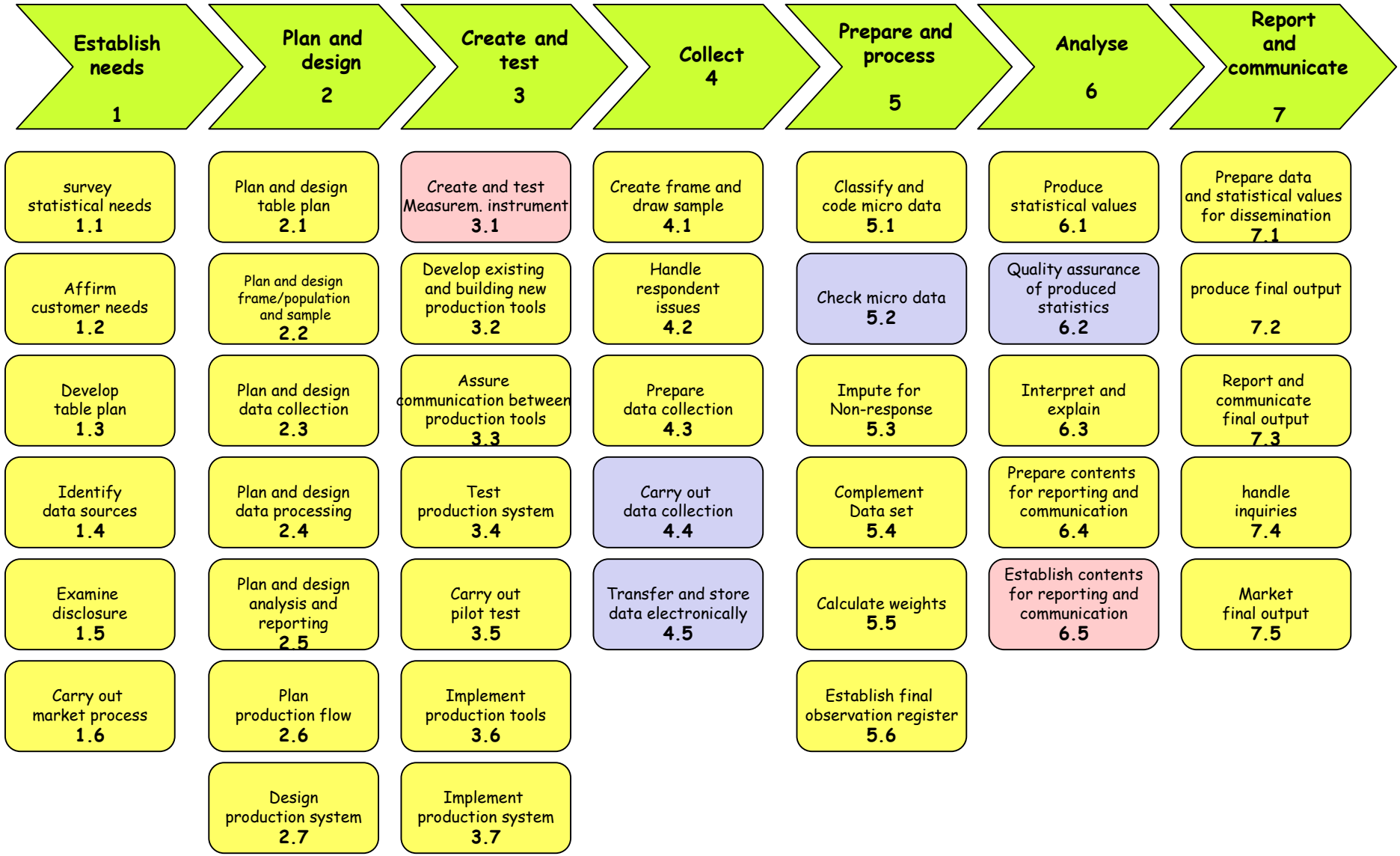
Statistical Production Process



Statistical Production Process



Statistical Production Process



Average proportions of costs of sub-processes 2004

Process	Proportion of total cost (%)		
	All products	Short-period	Annual surveys and periodic
Respondent service	3.3	3.3	3.4
Manual pre-editing	4.4	3.9	5.1
Data-registration editing	5.6	5.1	6.5
Production editing	**15.3**	12.7	18.9
Output editing	3.9	3.4	4.8
Total editing cost	**32.6**	28.3	38.6

Web data collection

Demands:

- High hit-rate in electronic questionnaires
- System that can measure hit-rate?

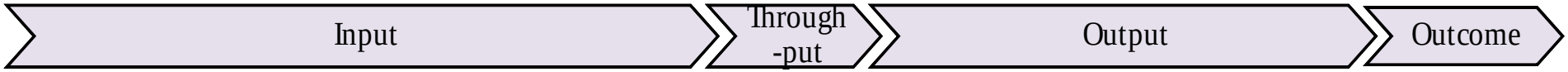
Question:

- Can it be a goal for us to move all editing to electronic data collection?

Expectation on the production editing process at Stat. Sweden

- Generic IT-tools
  - Less IT-maintenance
  - Easier planning of work and personnel at Data collection units
  - Better working environments
  - Methodology studies
- Efficient editing methods
  - Selective/significance editing
  - Better working environments
  - Less response burden
- Collection and analysis of process data
  - Continuous improvement of data collection and editing processes
  - Information for quality declaration of statistics

Input, throughput, output



Sample object (k), respondent

Observed object (l)	Background variable			Measurement variable (j)		
	Industry	Gender	Occup.	1	2	3
1						
2	B	M	2		y_{jkl}	
3						
4						

-coding
-editing
-imputation
-estimation

Sum of variable j by Industry

Industry	
A	
B	
C	
D	
E	
F - Z	

Sum of variable j by profession and sex

	Gender		
Occupa-tion	Men	Women	Sum
1			
2			
3			
4			
Sum			

-decision making
-information

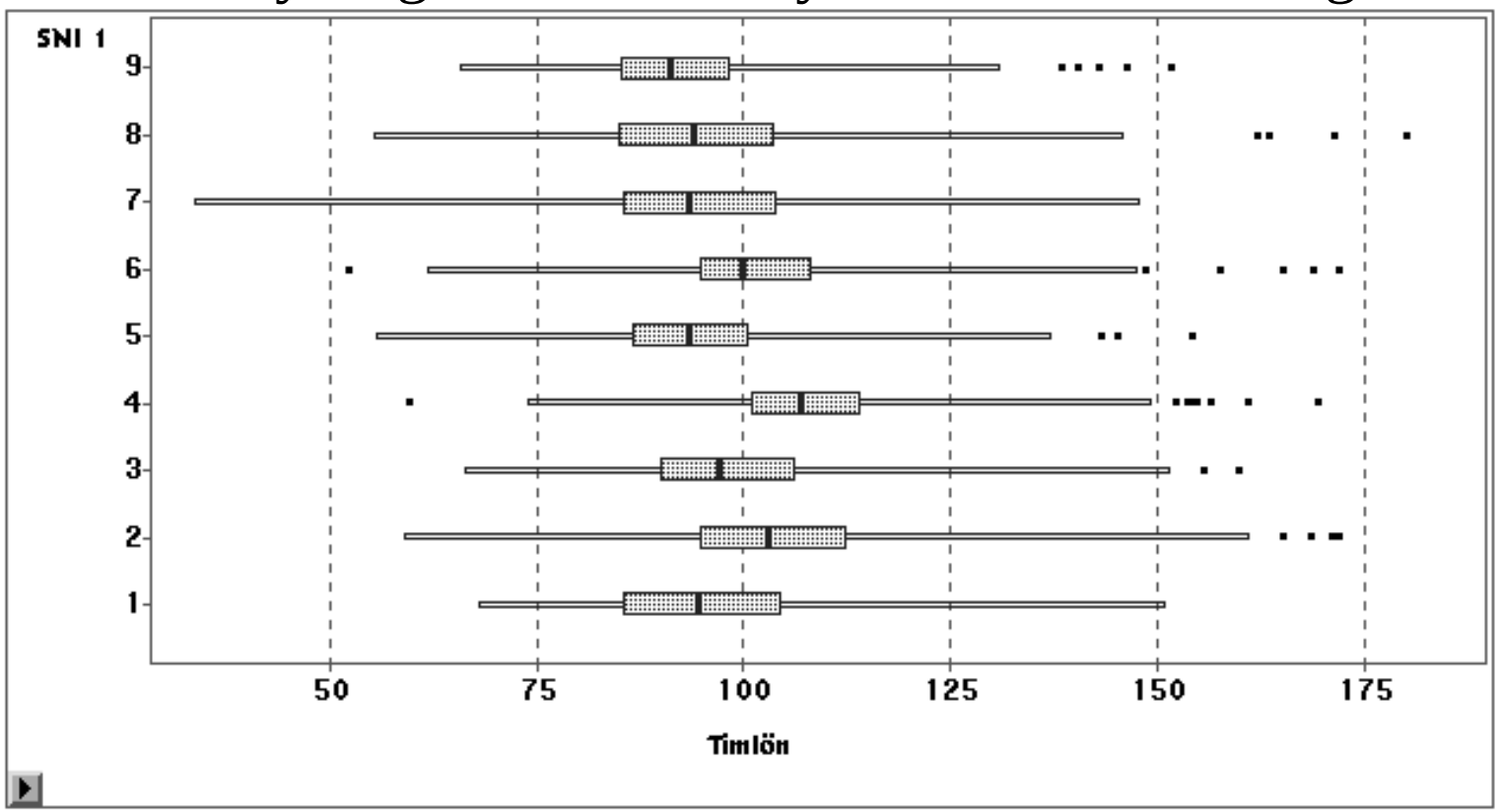
Impact

- **Actual impact** = **w (y_une – y_edi)** for observation **k** is the impact on domain-total **T** if **y_une** is kept instead of making a review to find **y_edi.**
- **Potential impact = w (y_edi – y_pred)** is a proxy for actual impact to be used in practice, as **y_edi** will not be known until review. **y_pred** is a prediction (expected value) for **y_edi.**
- **Anticipated (expected) impact** (per domain, variable, observation) is the product of suspicion and potential impact.

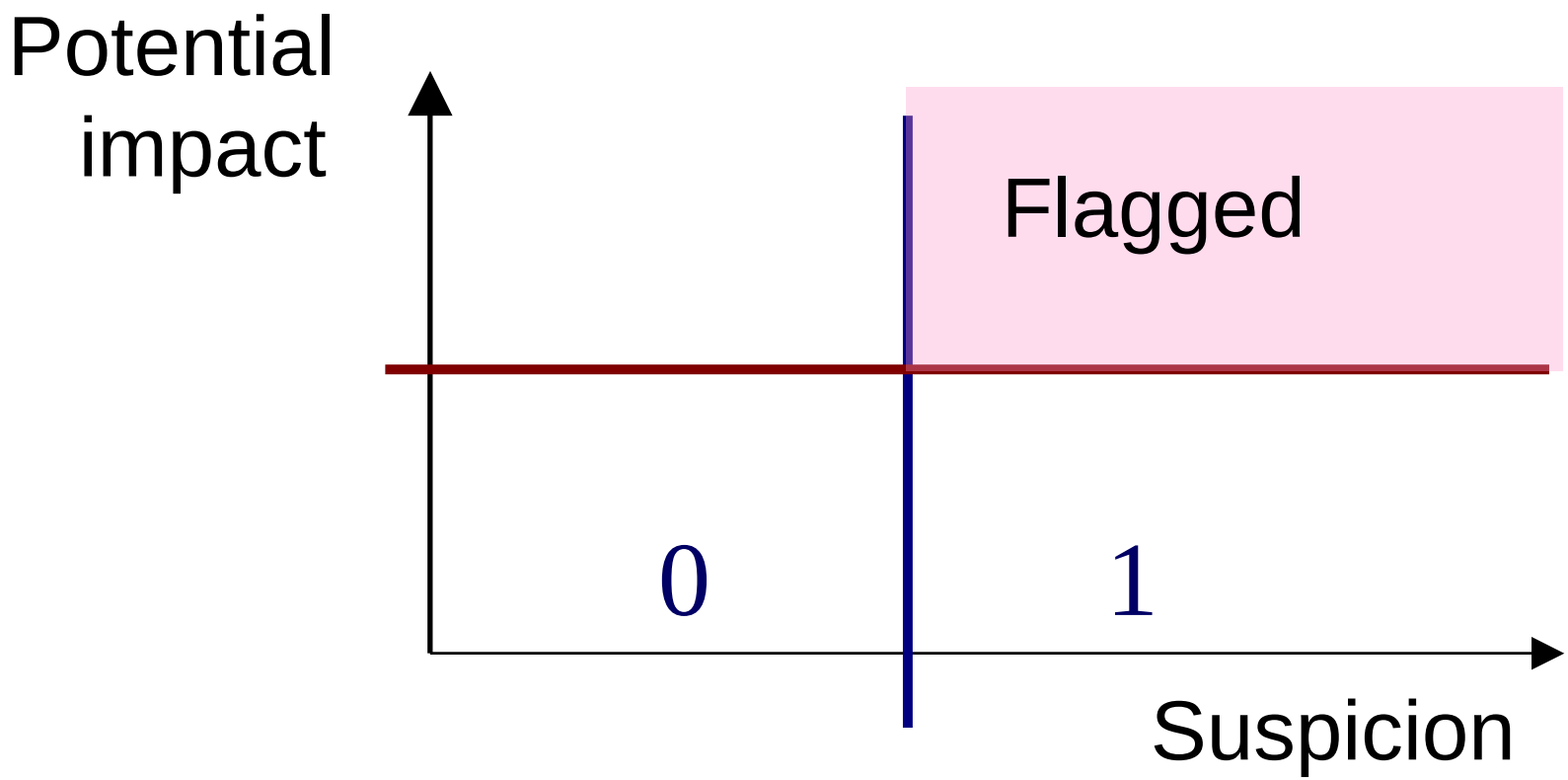
Suspicion: Traditional edits

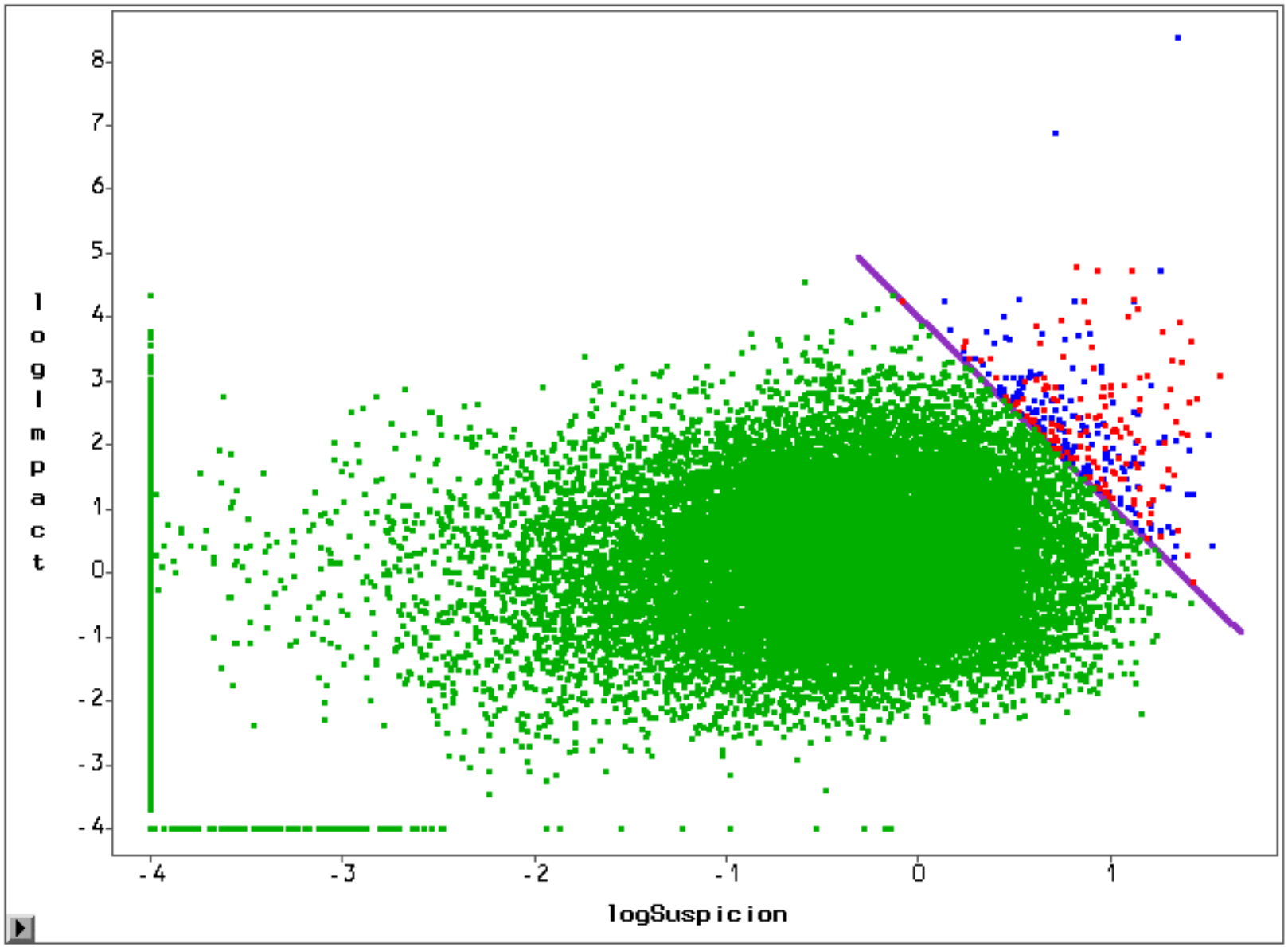
Finding acceptance limits: Data from previous survey rounds

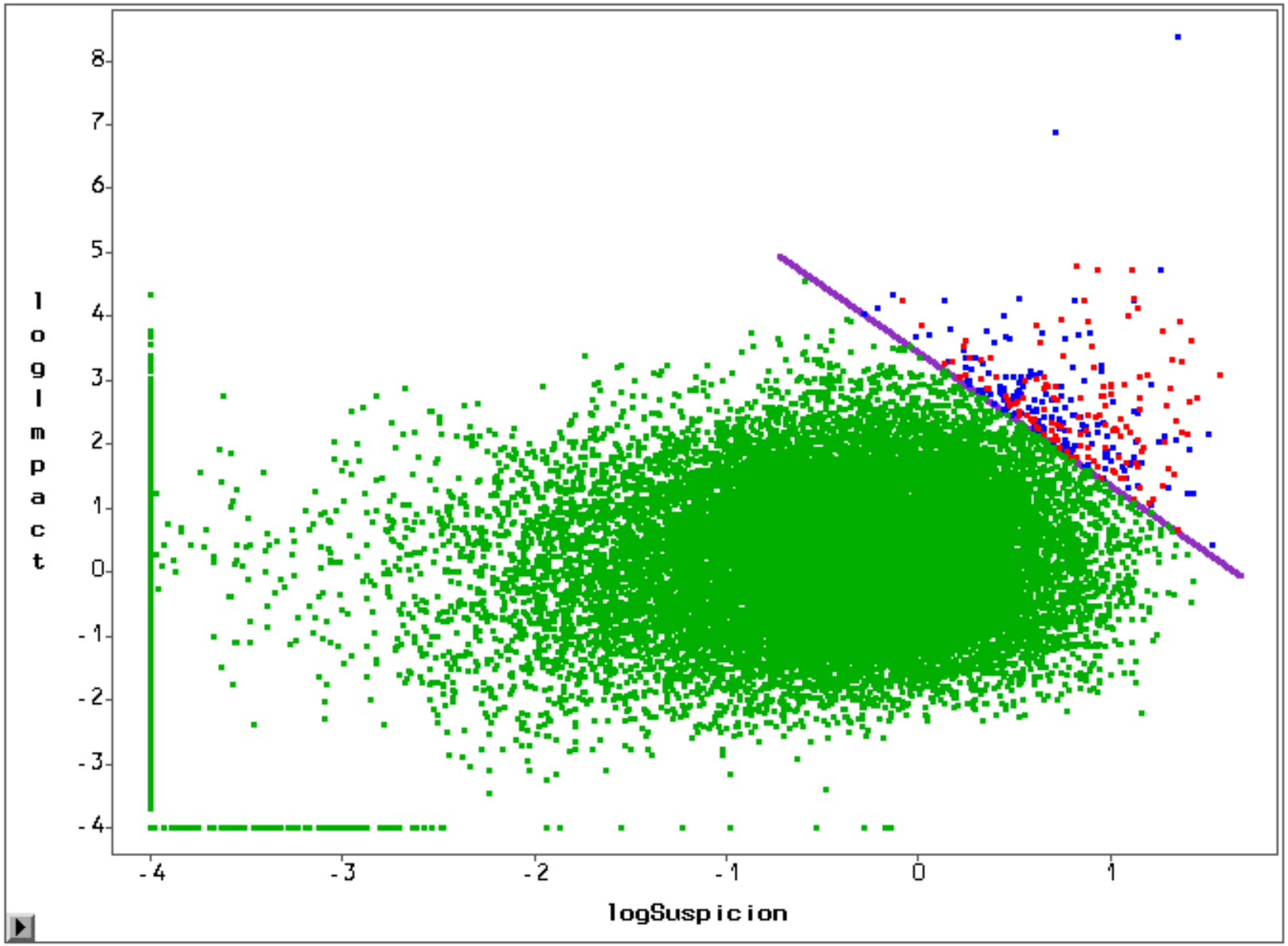
Hourly wage distributed by SNI code at one-digit level.

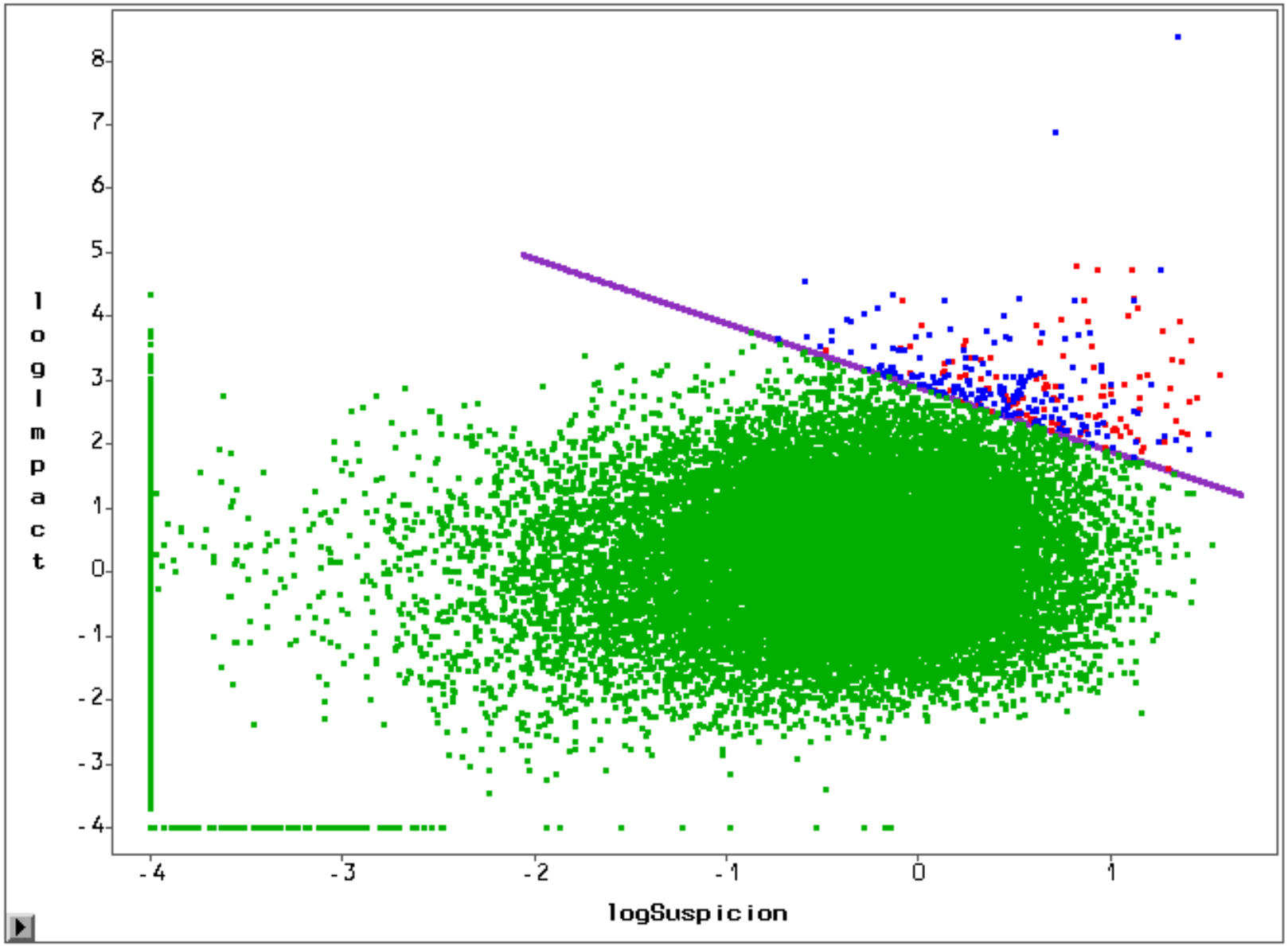


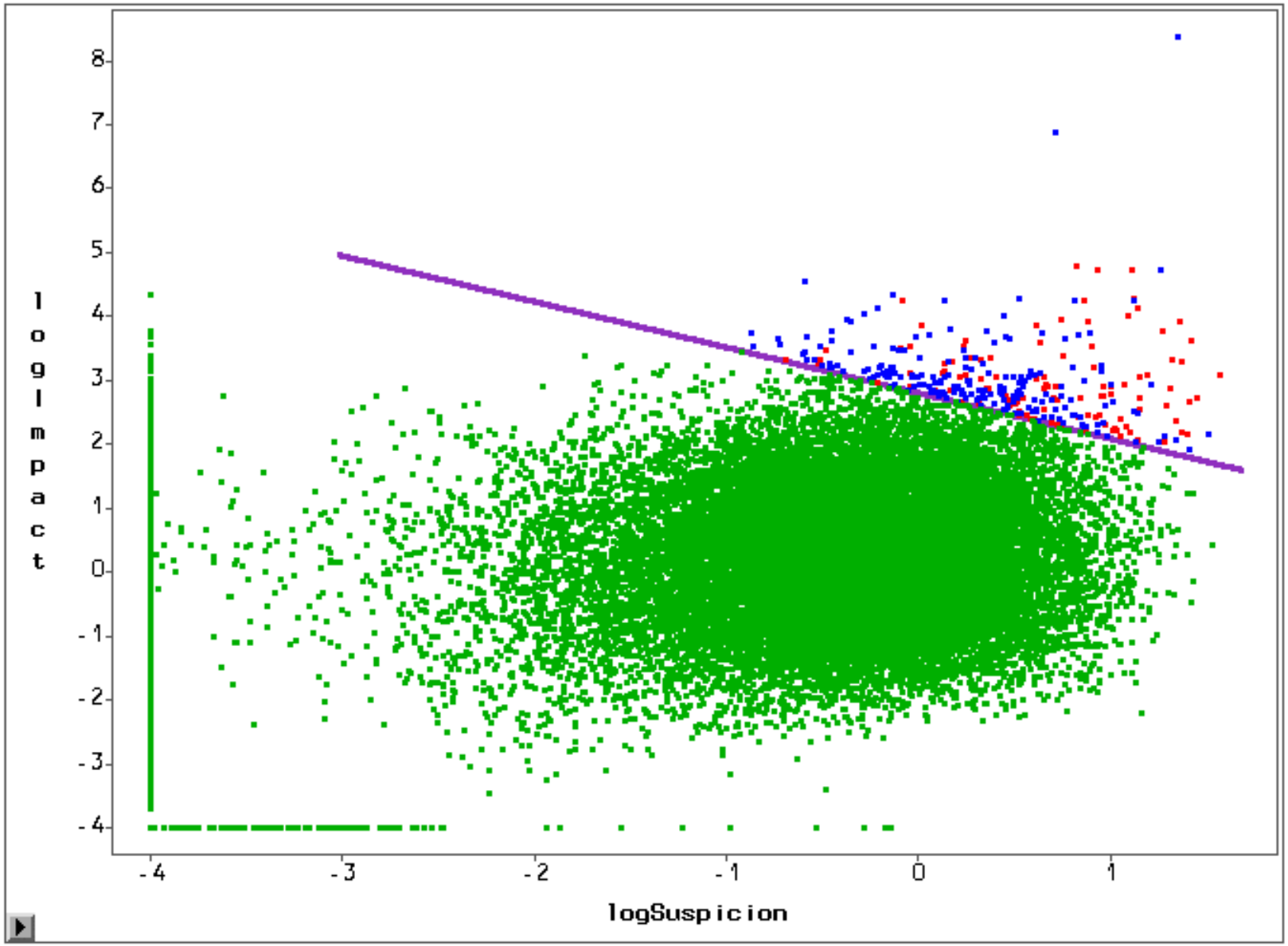
Selective editing

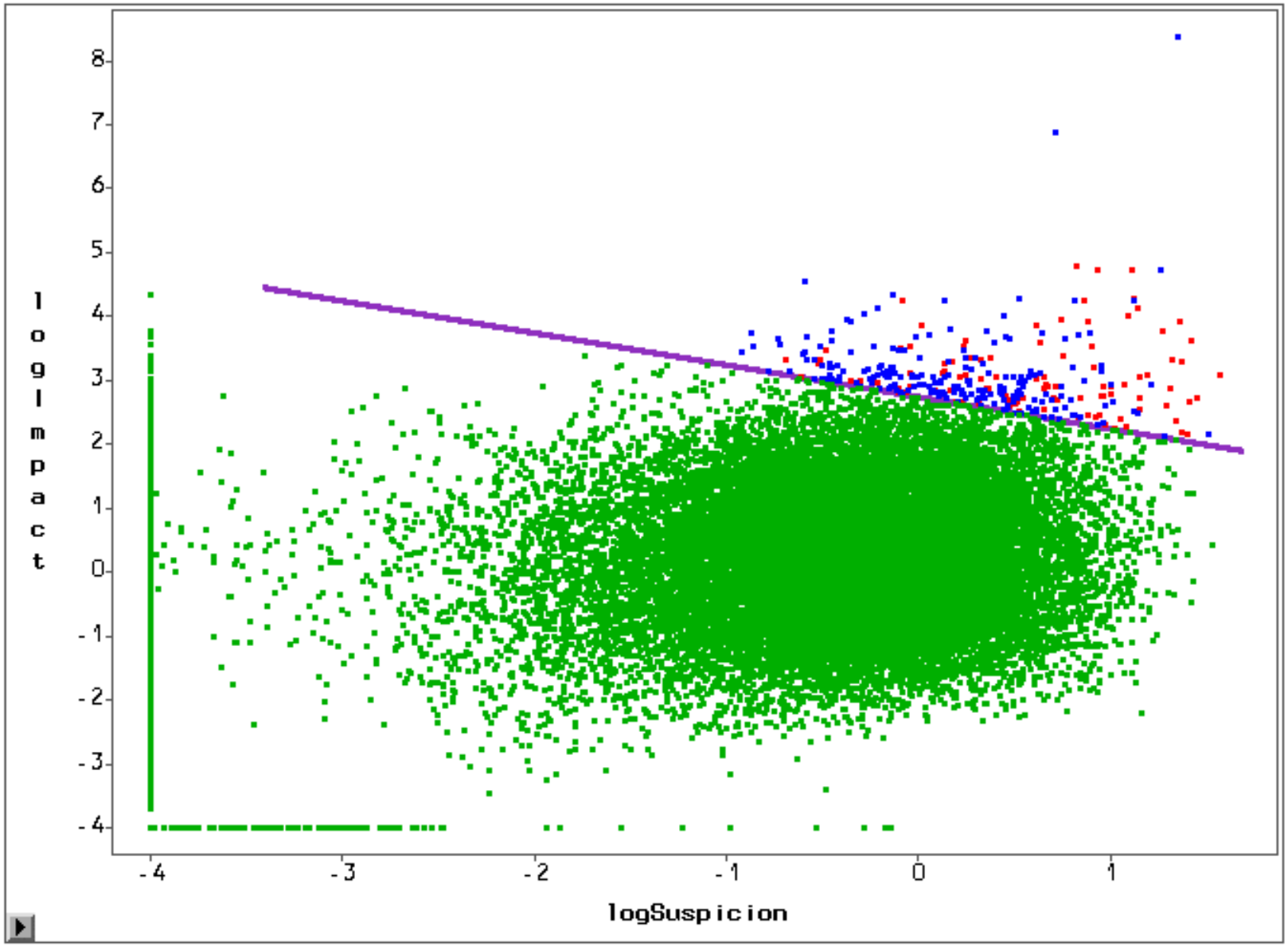



8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
logImpact
-4
-3
-2
-1
0
1
logSuspicion


logImpact
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
logSuspicion


logImpact
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
logSuspicion


logImpact
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
logSuspicion

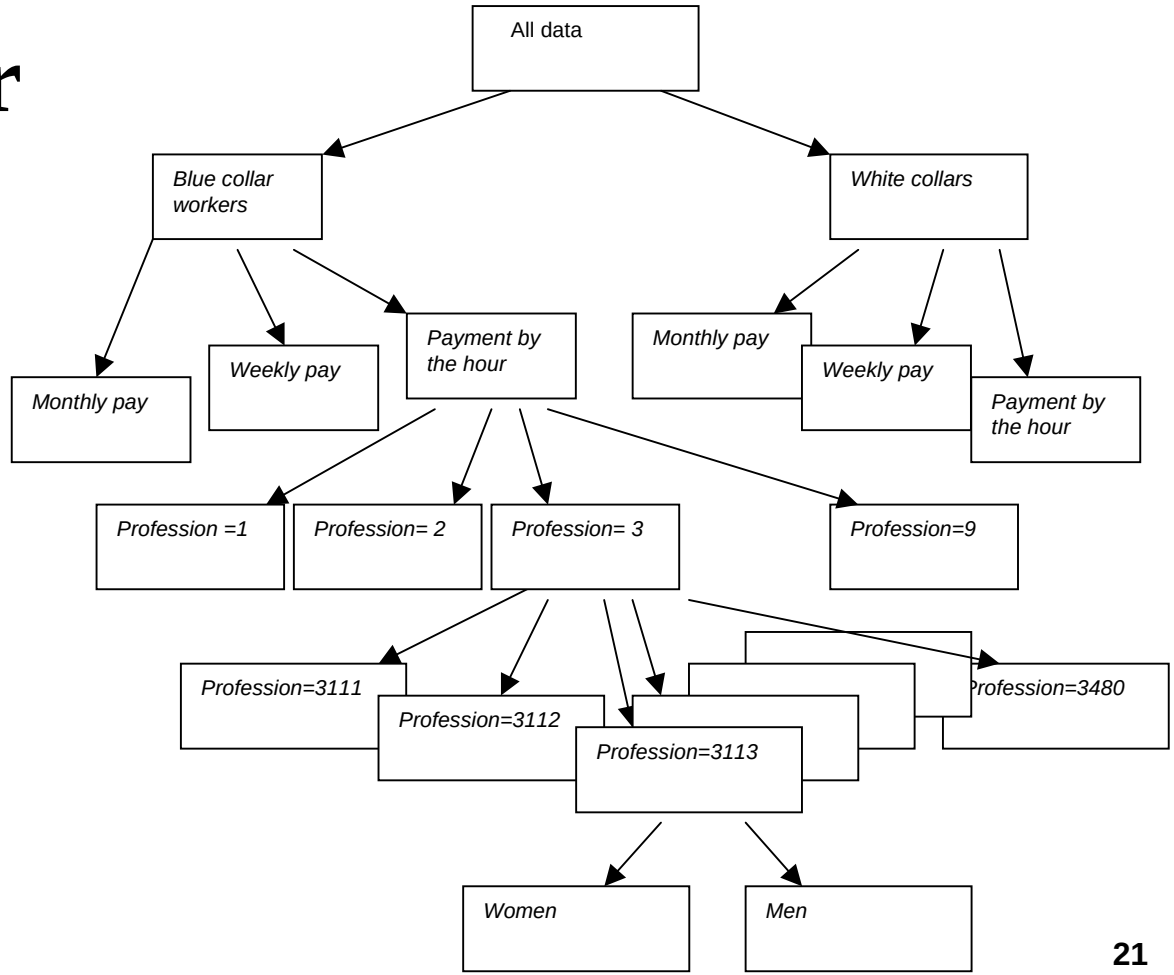

logImpact
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
logSuspicion

Predicted (expected) values

Data / predictor

- Time series
  - Previous value
  - Forecast
- Cross section
  - Mean/standard error
  - Median/quartile

Edit groups



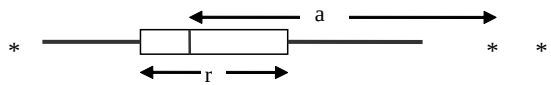
Suspicion

$$R=\begin{cases}\left(\tilde{z}_{j,k,l}-KAPPA\left(\tilde{z}_{j,k,l}-\tilde{z}^{L}_{j,k,l}\right)-z_{j,k,l}\right)/(\tilde{z}^{U}_{j,k,l}-\tilde{z}^{L}_{j,k,l}) & if\ z_{j,k,l}<\tilde{z}_{j,k,l}-KAPPA\left(\tilde{z}_{j,k,l}-\tilde{z}^{L}_{j,k,l}\right)\\ 0 & if\ \tilde{z}_{j,k,l}-KAPPA\left(\tilde{z}_{j,k,l}-\tilde{z}^{L}_{j,k,l}\right)<z_{j,k,l}<\tilde{z}_{j,k,l}+KAPPA\left(\tilde{z}^{U}_{j,k,l}-\tilde{z}_{j,k,l}\right)\\ \left(z_{j,k,l}-\tilde{z}_{j,k,l}+KAPPA\left(\tilde{z}^{U}_{j,k,l}-\tilde{z}_{j,k,l}\right)\right)/(\tilde{z}^{U}_{j,k,l}-\tilde{z}^{L}_{j,k,l}) & if\ \tilde{z}^{U}_{j,k,l}<\tilde{z}_{j,k,l}+KAPPA\left(\tilde{z}^{U}_{j,k,l}-\tilde{z}_{j,k,l}\right)\end{cases}$$

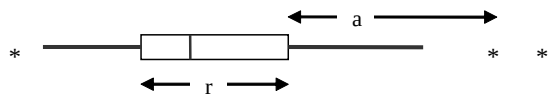
Suspicion=R/(TAU+R)

KAPPA = 0. The ratio *R* is the distance between *t* and the centre $\tilde{t}$ divided by the dispersion range $r = \tilde{t}^{(U)} - \tilde{t}^{(L)}$,

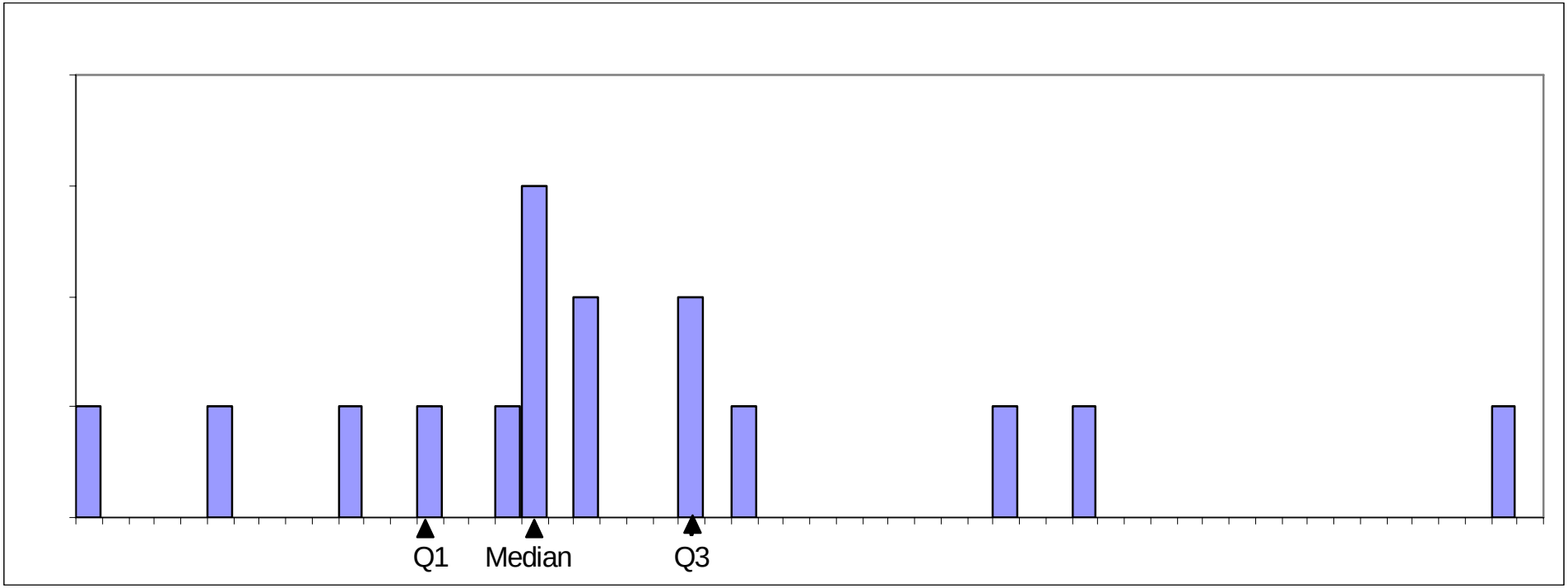
R = a/r:

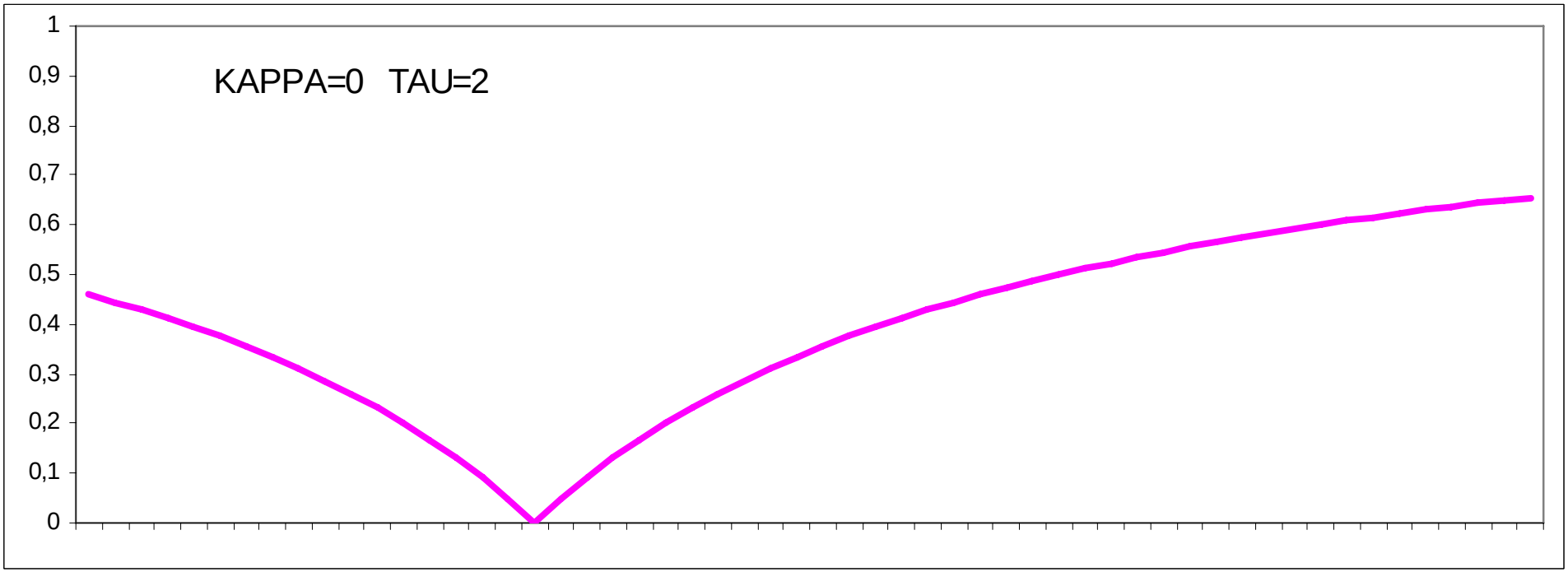


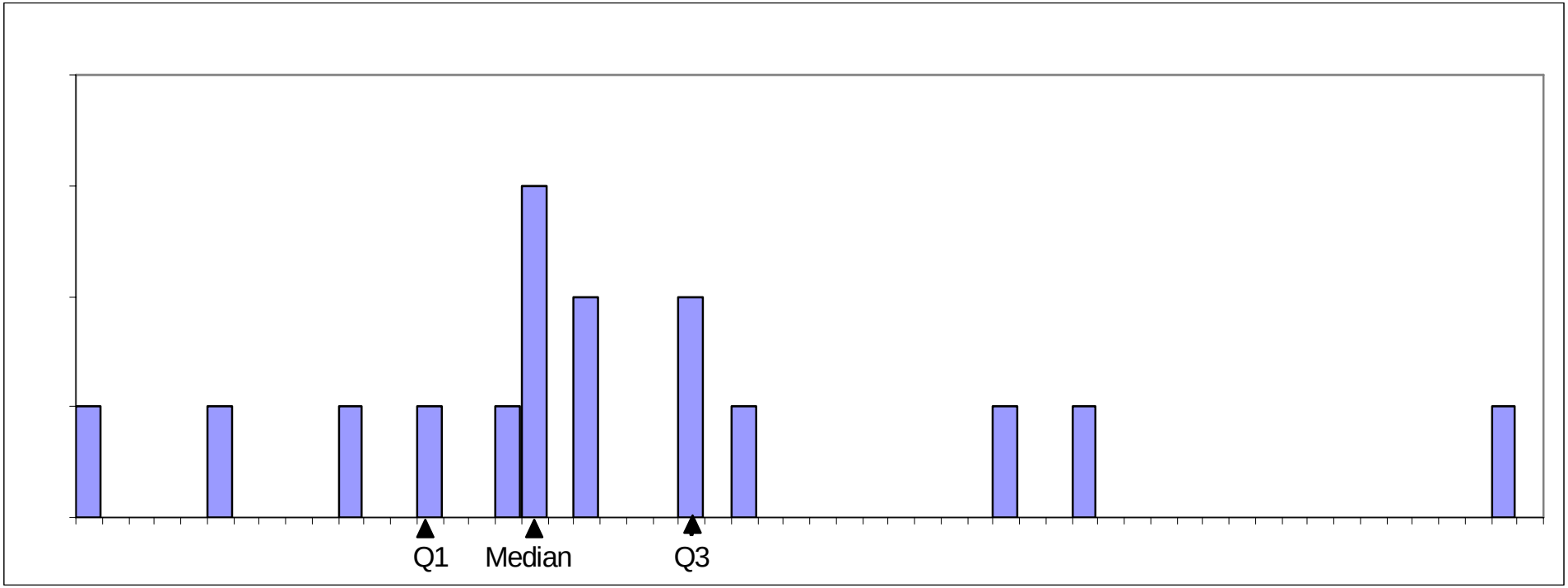
KAPPA = 1. The ratio *R* is the distance from the nearest range limit divided by the range. Hence *R* = a/r. For data between the lower and upper limits of the dispersion range the suspicion is zero.

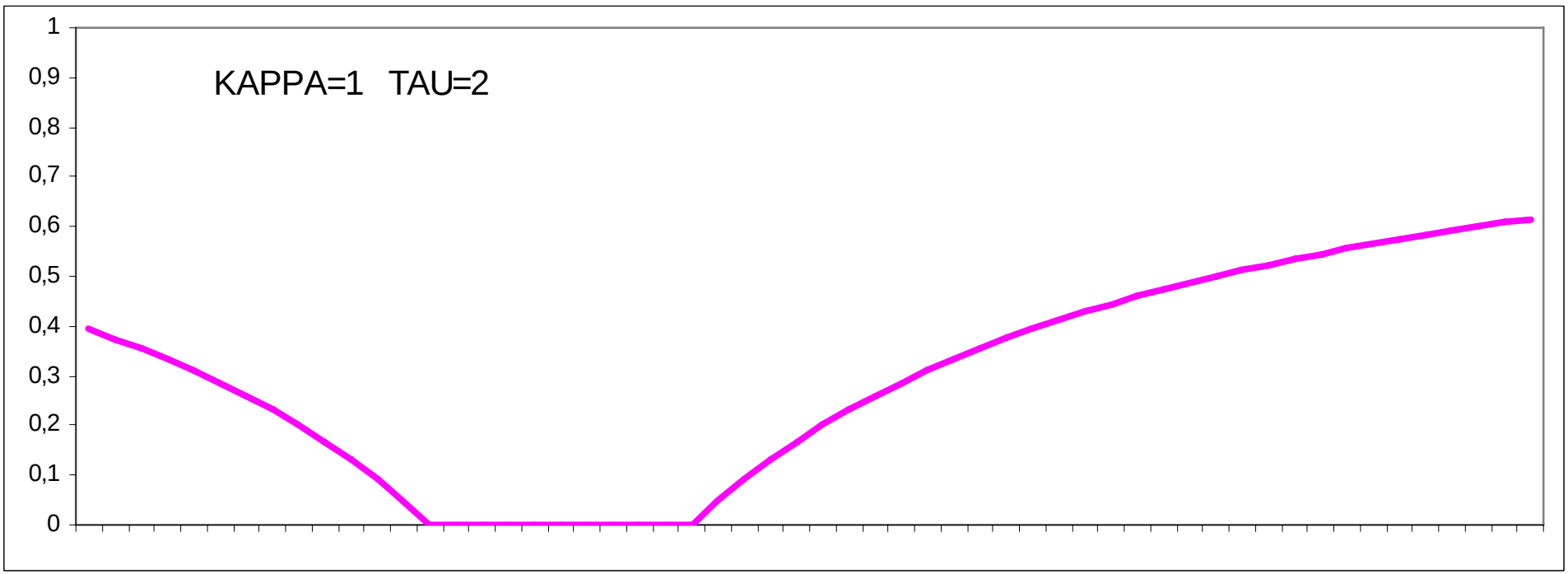


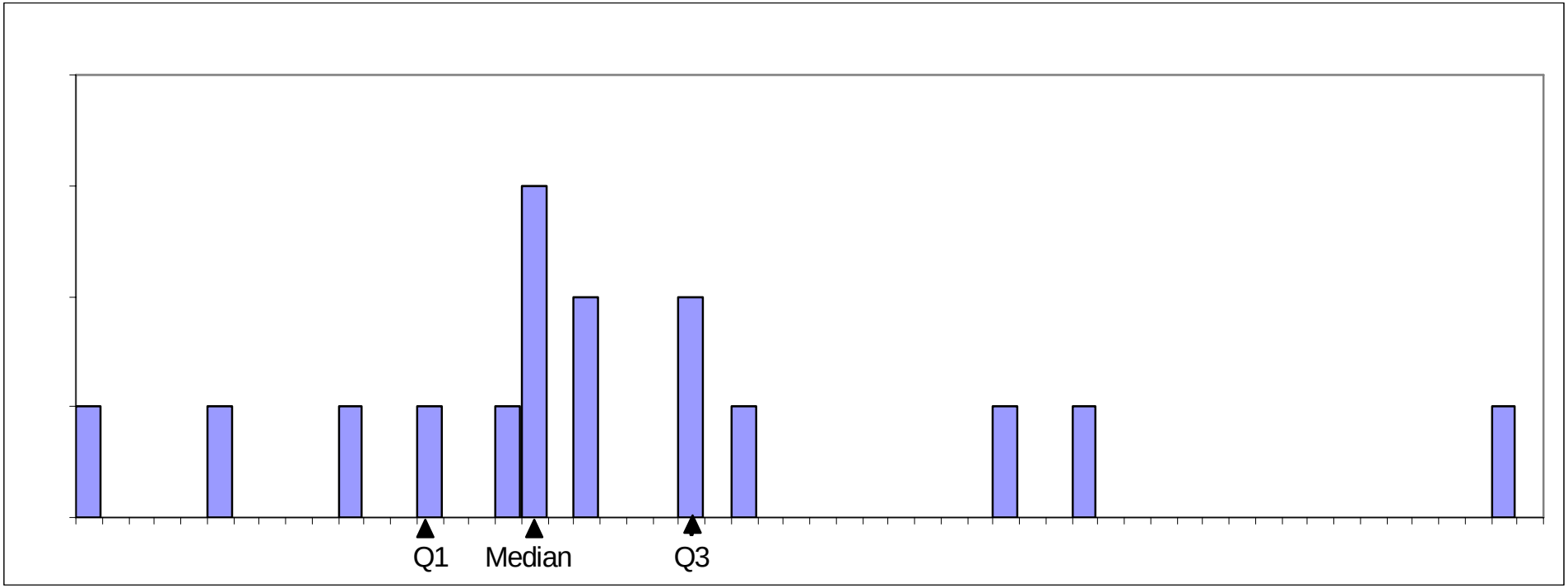
Susp

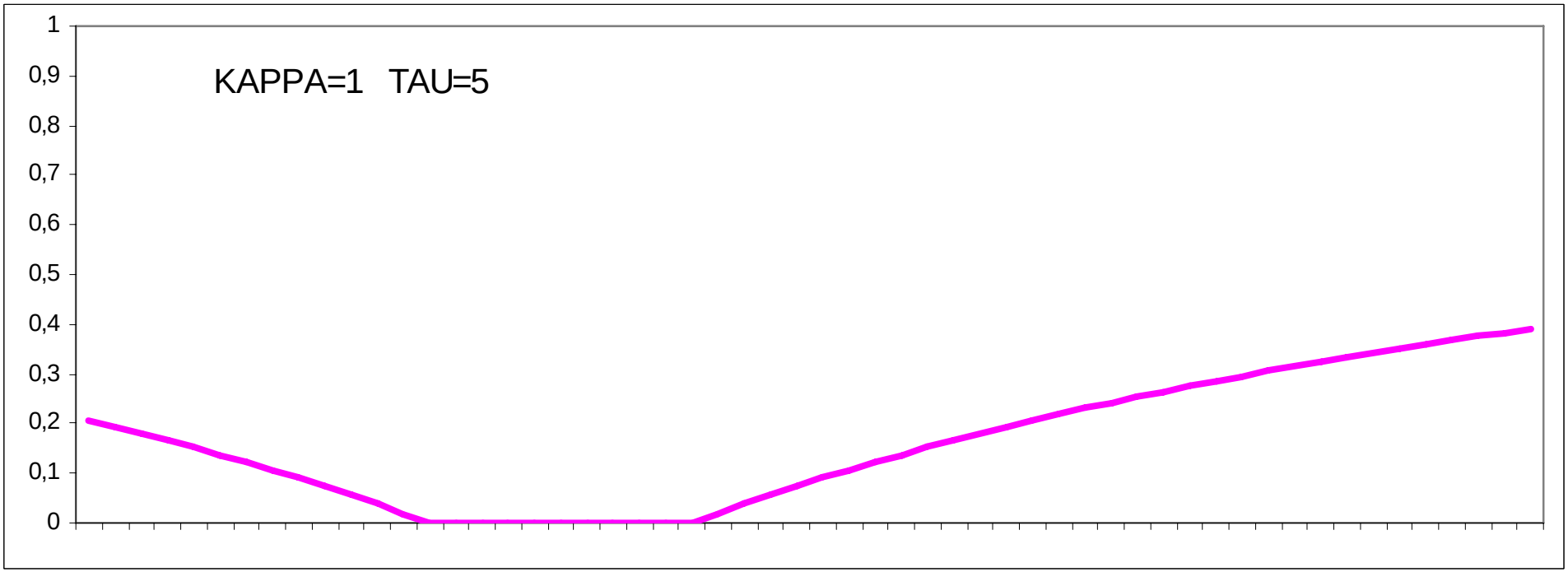

Q1
Median
Q3

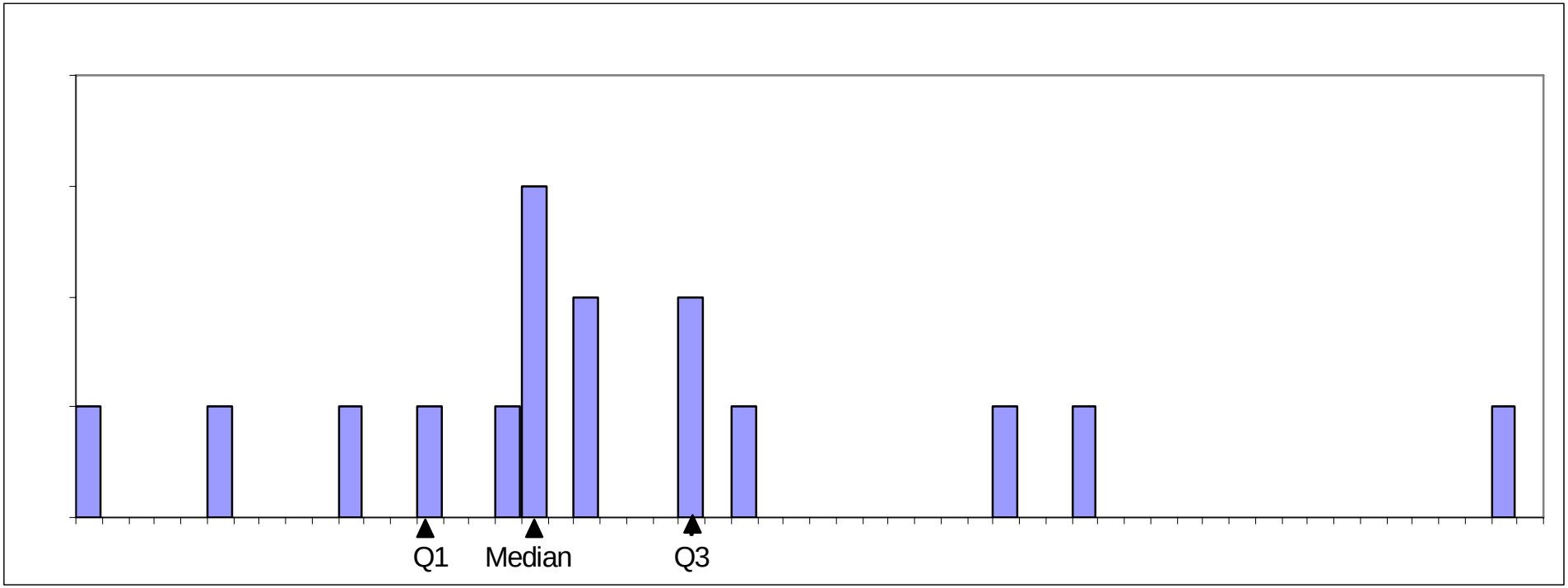

KAPPA=0 TAU=2
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0

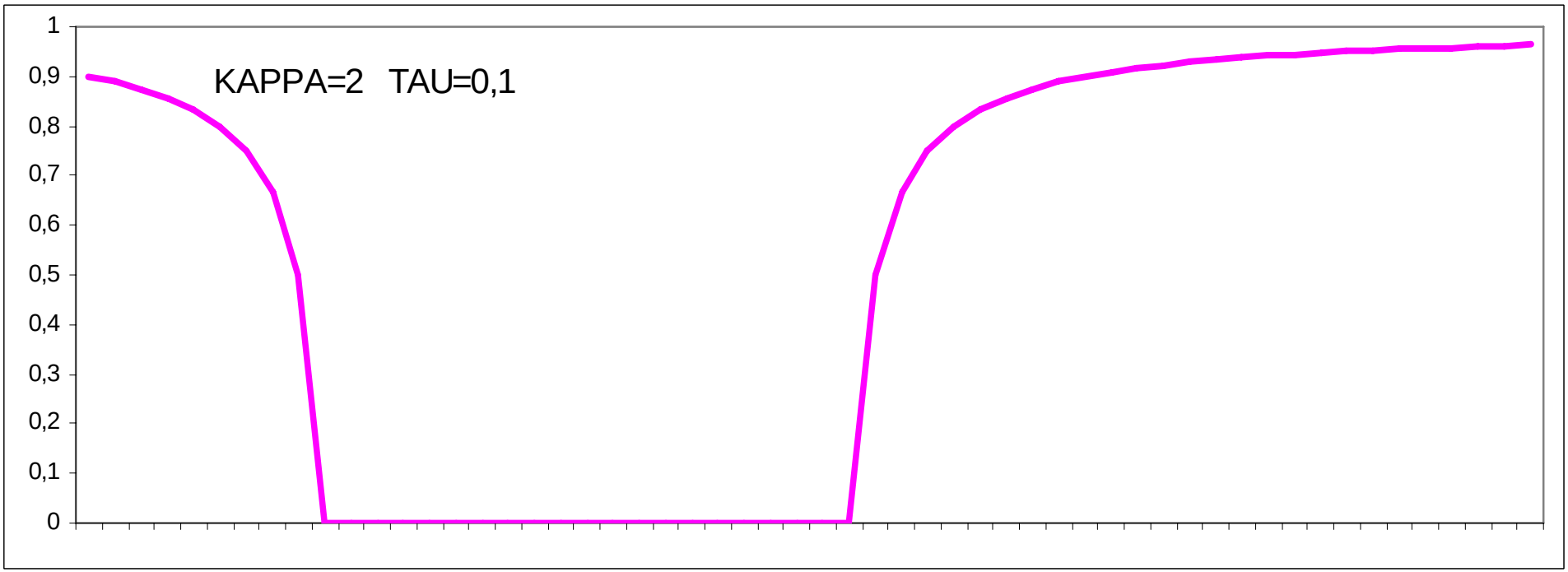

Q1
Median
Q3


KAPPA=1 TAU=2
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0


Q1
Median
Q3


KAPPA=1 TAU=5
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0


Q1
Median
Q3


KAPPA=2 TAU=0,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0

Score function

- Local score, by domain d, variable j & observed unit k,l is the anticipated impact related to an appropriate measure of size for the domain/variable, say standard error of estimate.
- $VIOLIN_j$ = weights for variables (j)
- $CLARINET_c$ = weights for classifications (domains) c(d)
- $OBOE_j$ = adjustment for size of estimated total or its standard error (j)
- $LScore_{d,j,k,l}$ = $Suspicion_{j,k,l}$ x(Potential impact) x $CELLO_{d(c),j}$

$$CELLO_{d(c),j} = \frac{VIOLIN_j \cdot CLARINET_{c(d)}}{\left(maximum\left\{ALFA_j \cdot \hat{T}_{d,j,t-1}, SE\left(\hat{T}_{d,j,t-1}\right)\right\}\right)^{OBOE_j}}$$

Score function

- Global scores are aggregated local scores by domain, variable and possibly second stage units to one score for each primary unit or respondent.
- Methods: sum, sum of squares, sum of local scores truncated by local thresholds, maximum etc.

$$\mathrm{GScore}_k = \left(\sum_l \left(\max\left\{ 0, \mathrm{LScore}_{k,l} - \mathrm{BETA} \right\} \right)^{\mathrm{LAMBDA}} \right)^{1/\mathrm{LAMBDA}}$$

Evaluation

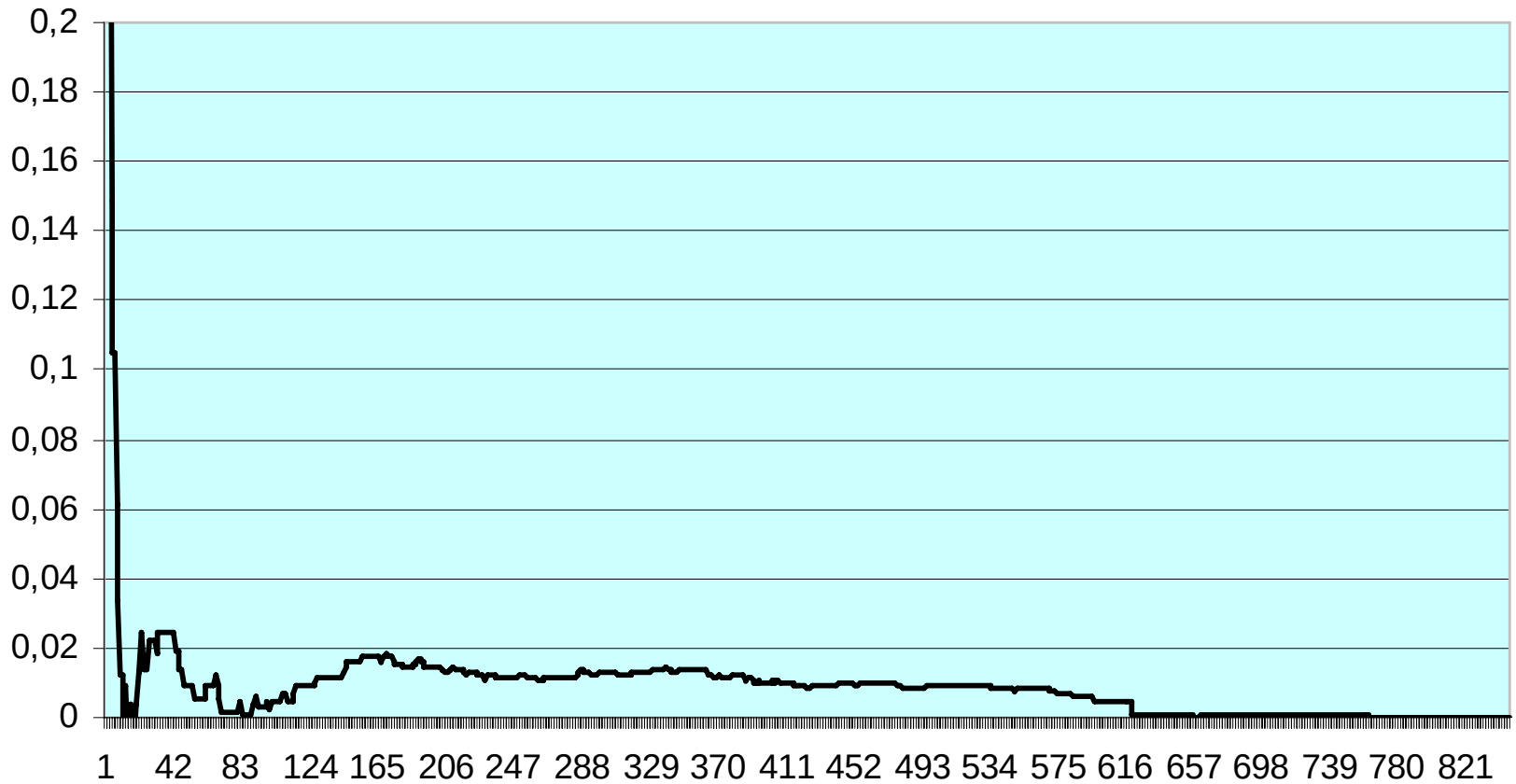
Relative pseudo-bias is a measure of error in output due to incomplete data review

$$\mathrm{RPB}(q) = \frac{\left|\hat{T}_q - \hat{T}_{100}\right|}{\mathrm{SE}\left(\hat{T}_{100}\right)}$$

Evaluation

Psedobias for PPI relative to the overall price index.

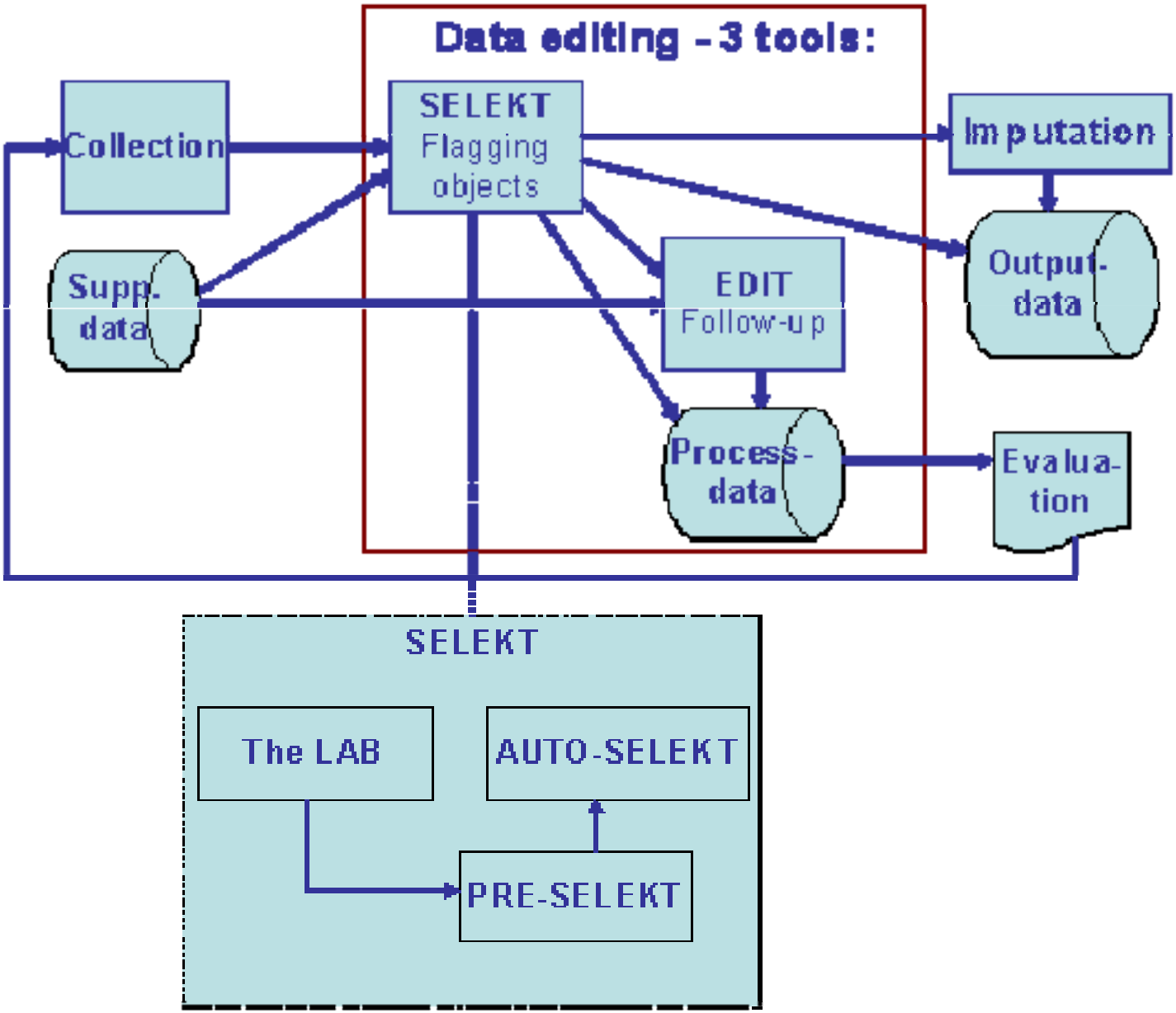
Observation units ordered in descending order of impact.



Editing – remaining methodology issues

- Fatal errors
  - Classifying variables
  - Survey variables
- Confidence (respondents and clients)
- New and old respondents
- Edited in earlier processes
  - Web-questionnaires
  - Scanned paper questionnaires
- Data and methods for computing predicted values etc.
- Homogenous groups
- Priorities; variables, domains (from the clients perspective)
- Score functions
- How to decide threshold values
- Sampling below threshold
  - Inference
  - Data for evaluation

Data editing process



SELEKT Parameters (in)

Parameter_group	Parameter	Value
A. Titles	TITLE1	Simulated quarterly salary
A. Titles	TITLE2	SELEKT ver 1.0 2010-02-04
B. Data to be edited	EDIT_DATA	Adap_Ent_Inflowdata
B. Data to be edited	EDIT_T1_VALUE	2009
B. Data to be edited	EDIT_T2_VALUE	1
C. Data structure	ID2_TYPE	
C. Data structure	ID2_VAR	EntNr
C. Data structure	ID2_WGT	PopEnt/SampleEnt
C. Data structure	ID3_VAR	
C. Data structure	ID3_WGT	1
C. Data structure	T1_VAR	Year
C. Data structure	T2_VAR	Quarter
D. Y-, X- and Test-Variables	EDI_X1	1
D. Y-, X- and Test-Variables	EDI_X2	1
D. Y-, X- and Test-Variables	EDI_Y1	Last_PopEmp
D. Y-, X- and Test-Variables	EDI_Y1_T1	Last_PopEmp/Frame_PopE
D. Y-, X- and Test-Variables	EDI_Y1_T2	Last_PopEmp/First_Turnover
D. Y-, X- and Test-Variables	EDI_Y2	Last_Turnover
D. Y-, X- and Test-Variables	EDI_Y2_T1	Last_Turnover/Pre_Turnover
D. Y-, X- and Test-Variables	UNE_X1	1
D. Y-, X- and Test-Variables	UNE_X2	1
D. Y-, X- and Test-Variables	UNE_Y1	First_PopEmp
D. Y-, X- and Test-Variables	UNE_Y1_T1	First_PopEmp/Frame_PopE
D. Y-, X- and Test-Variables	UNE_Y1_T2	First_PopEmp/First_Turnover
D. Y-, X- and Test-Variables	UNE_Y2	First_Turnover
D. Y-, X- and Test-Variables	UNE_Y2_T1	First_Turnover/Pre_Turnover
E. Cold deck. Cross section	CS_EXPECT	MEDIAN
E. Cold deck. Cross section	CS_HIST_DATA	Adap_Ent_Surveydata(where
E. Cold deck. Cross section	CS_MIN_OBS	7
E. Cold deck. Cross section	CS_SPREAD	QUANTILES
E. Cold deck. Cross section	CS_T1_END_VALUE	2008
E. Cold deck. Cross section	CS_T1_START_VALUE	2005
E. Cold deck. Cross section	CS_T2_END_VALUE	4
E. Cold deck. Cross section	CS_T2_START_VALUE	1
E. Cold deck. Cross section	HG_DIG1	1
E. Cold deck. Cross section	HG_DIG2	1
E. Cold deck. Cross section	HG_DIG3	1
E. Cold deck. Cross section	HG_DIG4	8
E. Cold deck. Cross section	HG_LENGTH	8
E. Cold deck. Cross section	HG_MAX	4
Parameter_group	Parameter	Value
E. Cold deck. Cross section	HG_VAR1	EnterpriseSize

Give AUTO-SELEKT the parameters:

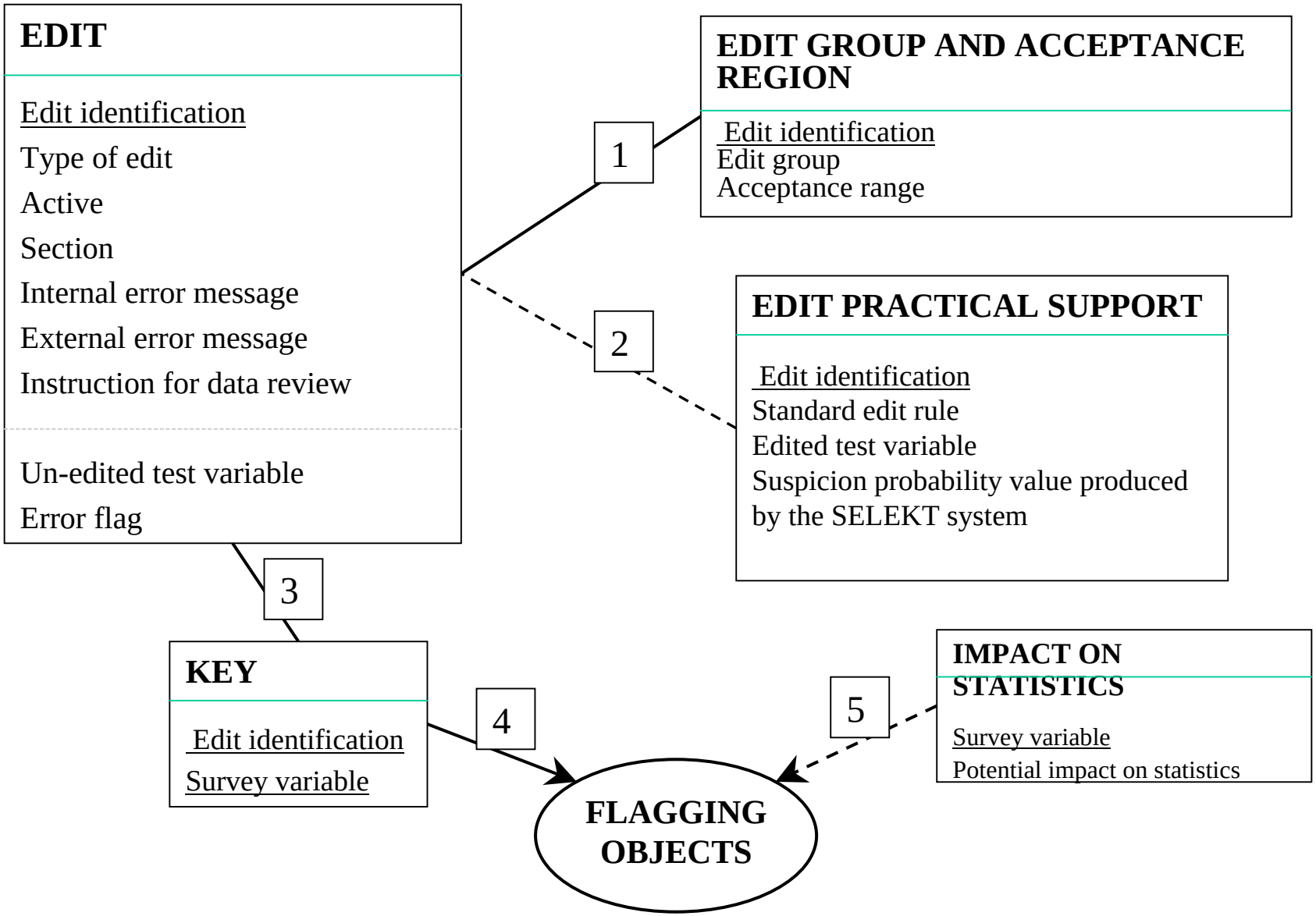
```
%let PATH_sys=C:\SELEKT\1.0;
%let PATH_app=C:\SELEKT\Prod\Demo1_Enterprise;
%let EDIT_parms=Ent_Parms1;
%let EDIT_data=Demo1_Adap_Ent_Inflowdata;
%let EDIT_T1_Value=2009;
%let EDIT_T2_Value=1;
```

By the parameter table &EDIT_parms AUTO-SELEKT knows what to do.

SELEKT Error list (out)

- **Identification:**
- Column name = Variable
- Id1 = Identity for respondent[optional]
- Id2 = Identity for primary sampling unit (PeOrgnr, CfarNr etc.)
- Id3 = Identity for observational unit (Social security number, CN8 for products)
- EditNumber = Edit identification, if the edit flags for suspicions or obvious error.
- Timestamp = Time when the questionnaire passes **SELEKT**
- **Process data:**
- EditFlag = 0 = accepted, 1-5 = error flagged
- EditSuspicion = Suspicion generated by continuous edits.
- Score1 = (Local) Score for respondent[optional]
- Score2 = (Local/Global) Score for primary sampling unit [optional]
- Score3 = (Global) Score for observational unit [optional]
- N_Obs = Number of observations, which have gone through the edit round.
- N_Obs_Flagg = Number of error flagged observations in the PSU, on this list
- N_PSU = Number of PSU for the respondent, which have passed the edit round
- N_PSU_Flagg = Number of error flagged primary sampling units, on this list

Edits


EDIT
Edit identification
Type of edit
Active
Section
Internal error message
External error message
Instruction for data review
Un-edited test variable
Error flag
1
EDIT GROUP AND ACCEPTANCE REGION
Edit identification
Edit group
Acceptance range
2
EDIT PRACTICAL SUPPORT
Edit identification
Standard edit rule
Edited test variable
Suspicion probability value produced by the SELEKT system
3
KEY
Edit identification
Survey variable
4
FLAGGING OBJECTS
5
IMPACT ON STATISTICS
Survey variable
Potential impact on statistics