

Modern granskning

Anders Norberg, SCB

2009-10-14

Om modern granskning



- 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.
- Massor av artiklar i tidskrifter

God svars kvalitet förutsätter att uppgiftslämnaren...

- **har kapacitet att ta fram data**
- **förstår frågans innebörd**
- **är motiverad att ta fram korrekta uppgifter**
- **får hjälp att undvika slarvfel**

Granskningens roll:

- Hitta fel (**Bra kontroller**)
- Identifiera felkällor (**Processdata**)
- Analysera processdata – framställ indikatorer – kommunicera med mättekniker
- Bidra till kvalitetsdeklarationen
- Åtgärda betydelsefulla fel (**Ändra/rätta**)

Granskning = Kvalitetskontroll

Vilka fel ska granskningen hitta?

■ Uppenbara fel

- Otillåtet värde, validitetsfel
- Värde saknas men måste finnas, partiellt bortfall
- Datastruktur- eller modellfel (totalen $\neq$ summan av komponenterna)
- Motsägelser

■ Misstänkta fel

- **Avvikelsefel (Outliers)**
 - Värde för stort eller för litet
- **Definitionsfel (Inliers)**
 - Många missuppfattar på samma sätt
 - Hämtar data från info-systemet med andra definitioner

Misstänkta fel

Avvikelsefel

- Utredning dyrbar och tidsödande
- Få avvikelsefel med effekter (låg träffsäkerhet, många ändringar ger inte några effekter på statistiken)

Granskningen ska ha effekt på skattningarna !

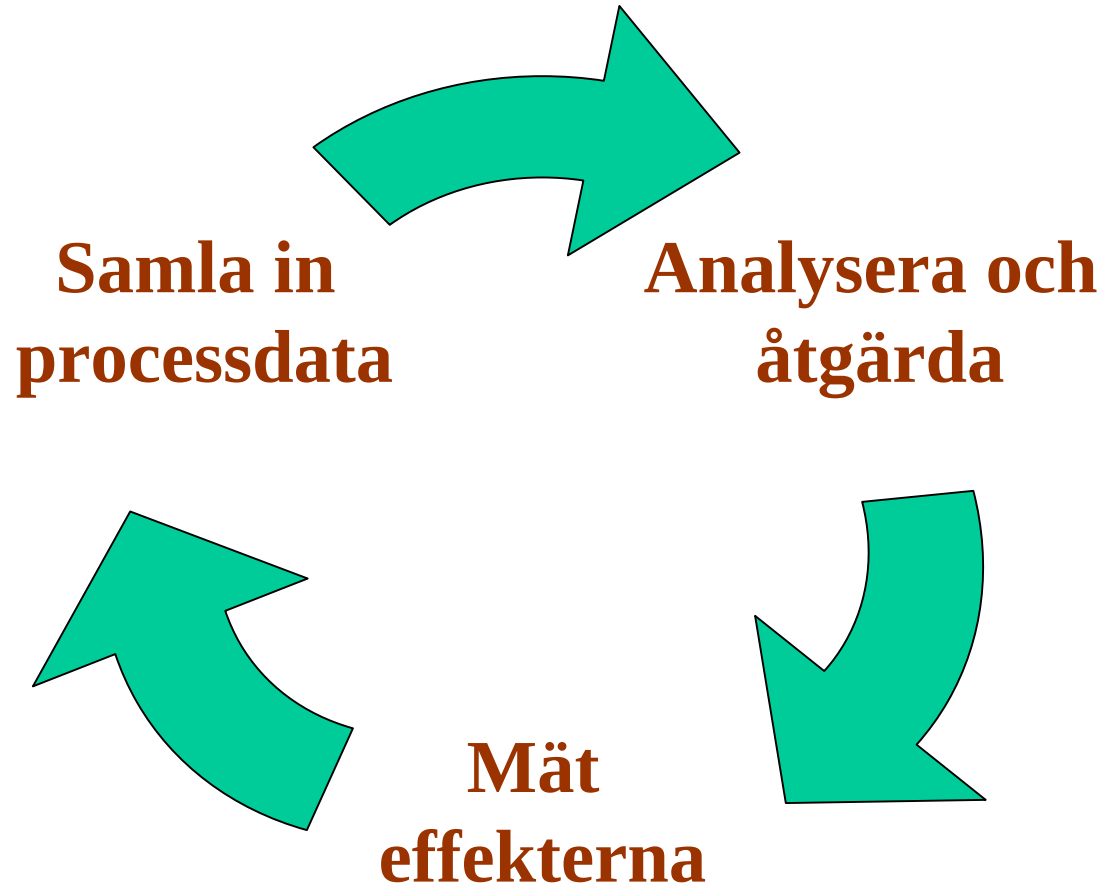
Tänk på uppgiftslämnarbördan !

Misstänkta fel

Definitionsfel (Inliers)

- Svåra att hitta
- Sätt att hitta dem kan vara:
 - samgranskning
 - djupintervjuer i fokusgrupper
 - utnyttja statistik från FEQ och från kontakter med uppgiftslämnarna
 - högt partiellt bortfall
 - grafisk granskning
 - goda exempel

Processperspektivet



Processdata

- Övervaka och förbättra mätningen (mätinstrument och insamlingsprocess)
- samt granskningsprocessen självt

Ogranskade data måste sparas så att viktiga processindikatorer, som bl.a. träffsäkerhet och granskningens effekter, kan beräknas!

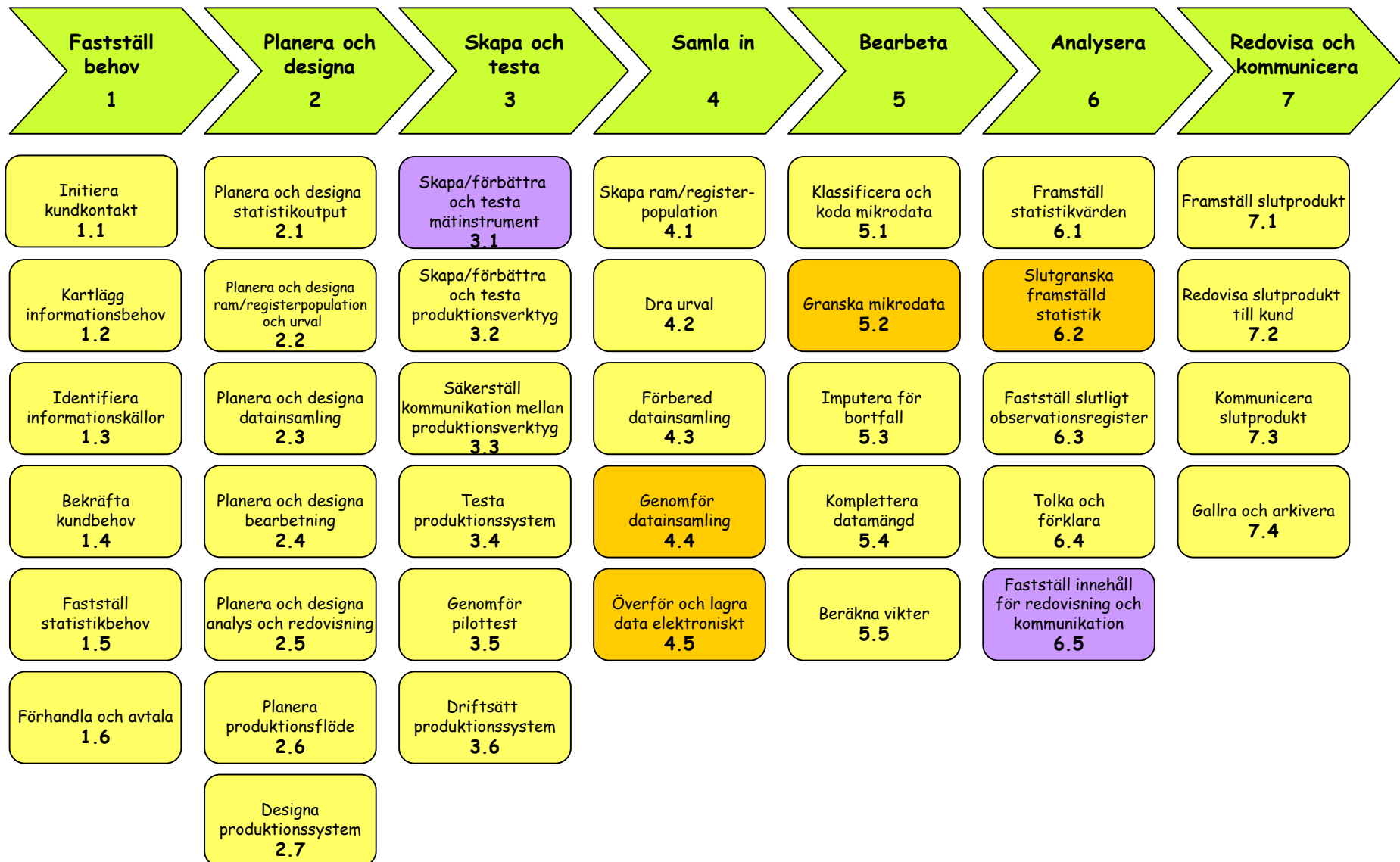
Processdata

- Felkällor (uppgiftslämnarnas problem)
- Misstankar
- Felsignaler
- Manuella åtgärder (accept / ändrade värden)
- Automatiska åtgärder

Processindikatorer

- Felkällor (uppgiftslämnarnas problem)
- Andel felsignalerade objekt och variabler
- Andel manuellt och automatiskt åtgärdade objekt och variabler
- Andel förändrade värden och effekten av förändringarna per variabel
- Kontrollernas träffsäkerhet

Statistikproduktionsprocessen



53 produkter	Andel av totala kostnader, %
Uppgiftslämnarservice	3
Manuell förgranskning	4
Dataregistreringsgranskning	6
Produktionsgranskning	15
Outputgranskning	4
Alla steg	33

Webbinsamlingar

Krav:

- Hög träffsäkerhet i elektroniska mätinstrument
- System som kan mäta träffsäkerheten

Fråga:

- Kan det bli ett mål för elektronisk datainsamling att alla kontroller skall kunna utföras i samband med uppgiftslämnandet ?

Förväntningar på granskning

- Generella verktyg
 - Mindre IT-underhåll
 - Lättare arbetsplanering
 - Bättre arbetsmiljö
 - Metodstudier blir lättare att göra
- Effektiv granskning → Selektiv granskning
 - Bättre arbetsmiljö
 - Mindre belastning på uppgiftslämnare
- Systematisk insamling och analys av processdata
 - Förbättring av datainsamling och granskning
 - Underlag för kvalitetsdeklaration

Input och output

Input

Through
-put

Output

Outcome

Sample object (k), respondent

-coding
-editing
-imputation
-estimation

Sum of variable j by Industry

-decision making
-information

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

Industry	
A	
B	
C	
D	
E	
F - Z	

Sum of variable j by profession and sex

Occupation	Gender		
	Men	Women	Sum
1			
2			
3			
4			
Sum			

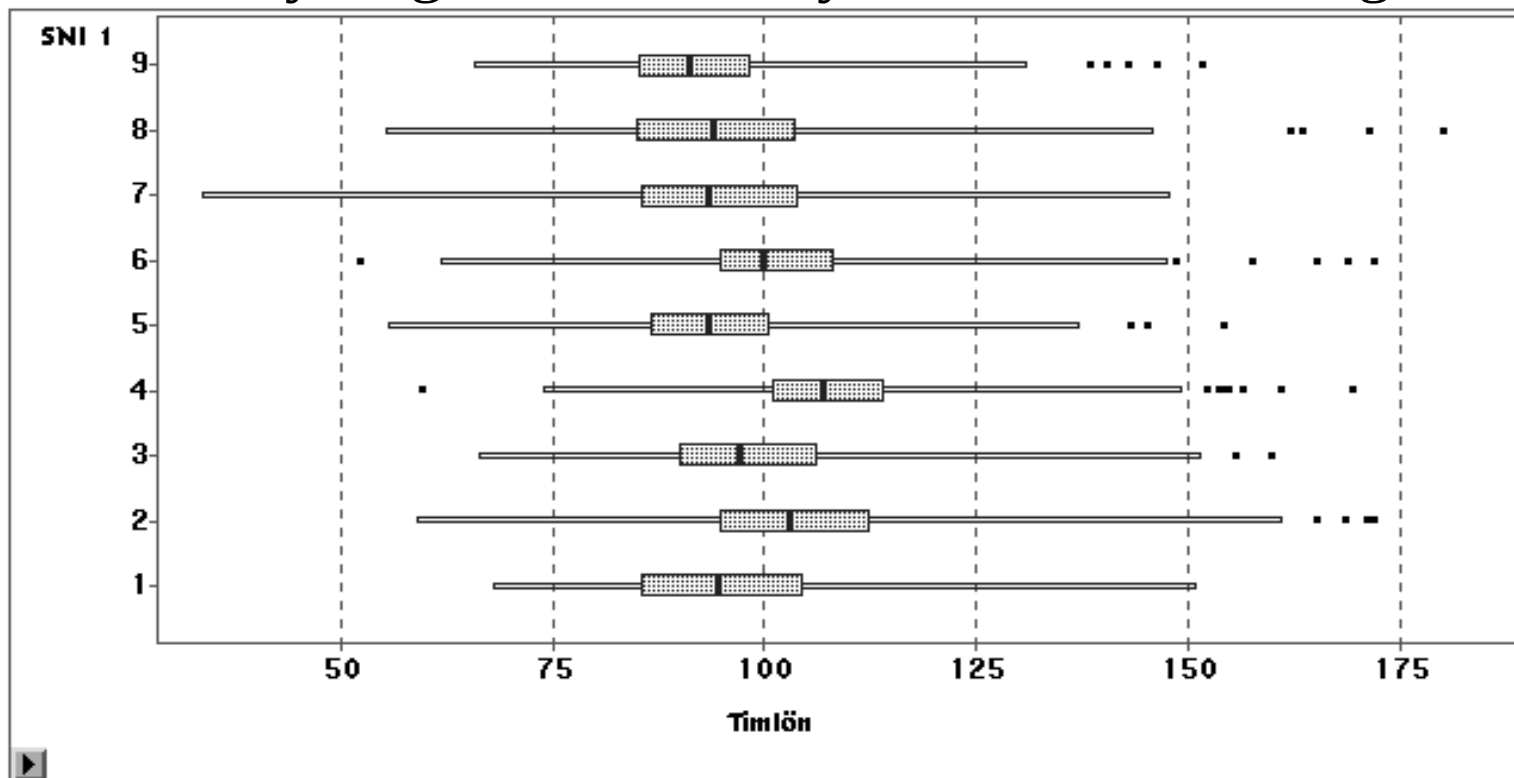
Effektmått

- **Faktisk effekt** = $w (y_{ogr} - y_{gr})$ för objekt k mäter effekten på skattningen av domän - totalen T om y_{ogr} behålls istället för att utreda och söka y_{gr} för objekt k .
- **Potentiell effekt** = $w (y_{ogr} - y_{pred})$ är en proxy för faktiskt effekt att använda i praktiken eftersom ju y_{gr} blir känd först efter utredning. y_{pred} kan vara objektets värde föregående produktionsomgång eller prognos från tidsserieanalys på objekts- eller gruppnivå.
- **Förväntad effekt** är (per domän, variabel, objekt) produkten av misstankemått och potentiell effekt

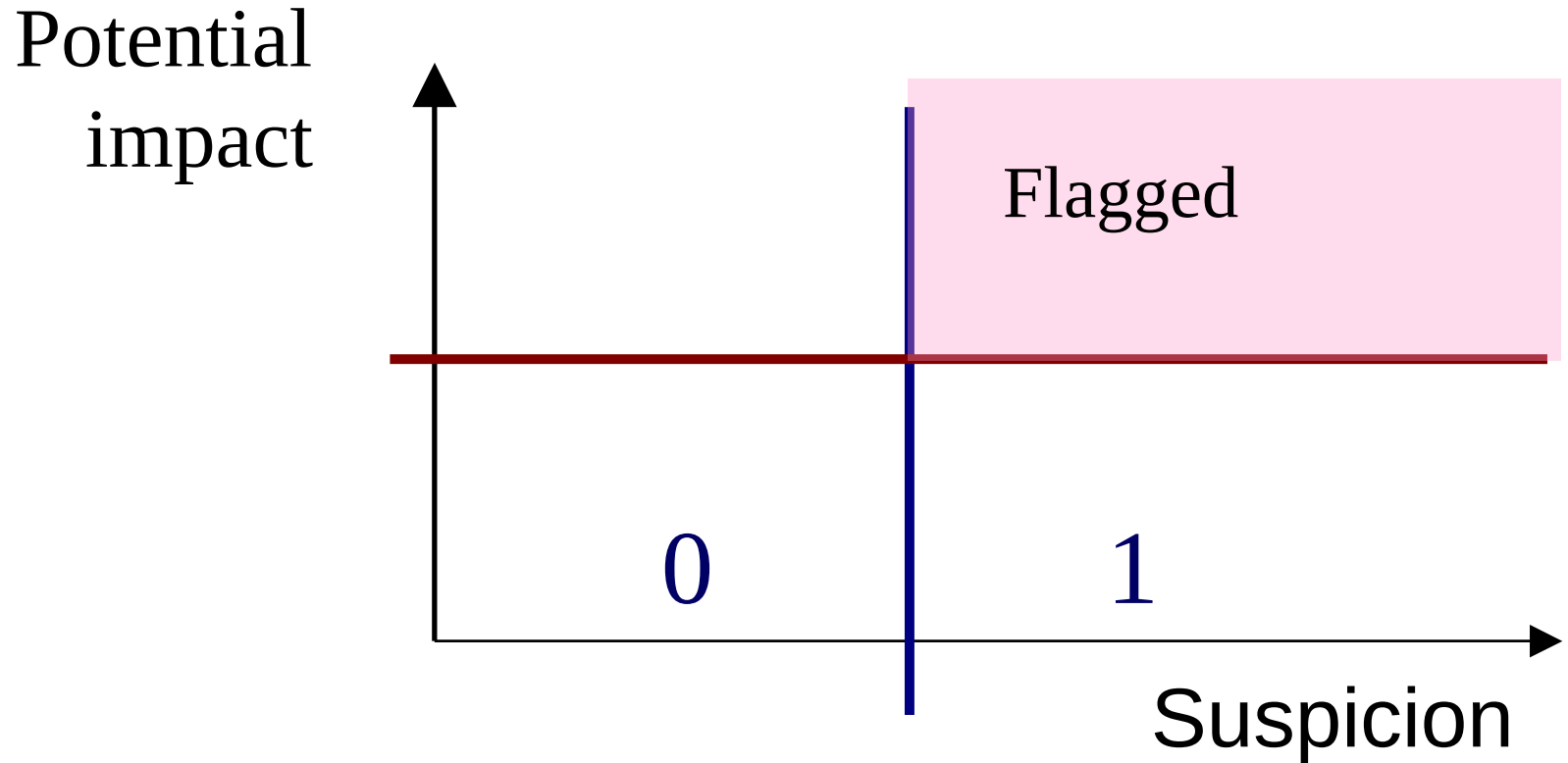
Misstanke: Traditionella kontroller

Att hitta acceptansgränser: Data från tidigare produktionsomgångar

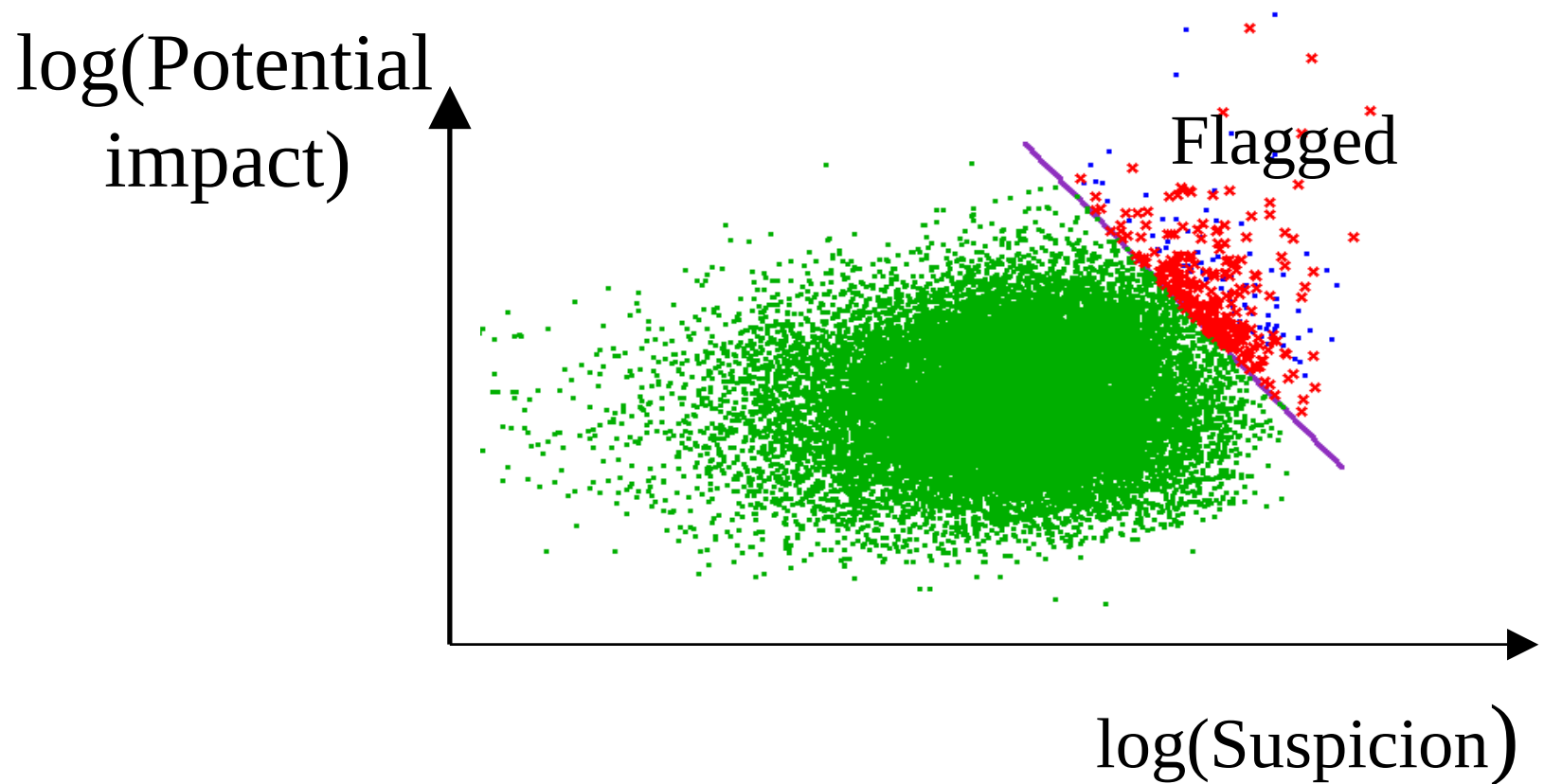
Hourly wage distributed by SNI code at one-digit level.



Flagging suspected errors



Flagging suspected errors



Suspicion

Homogeneous groups

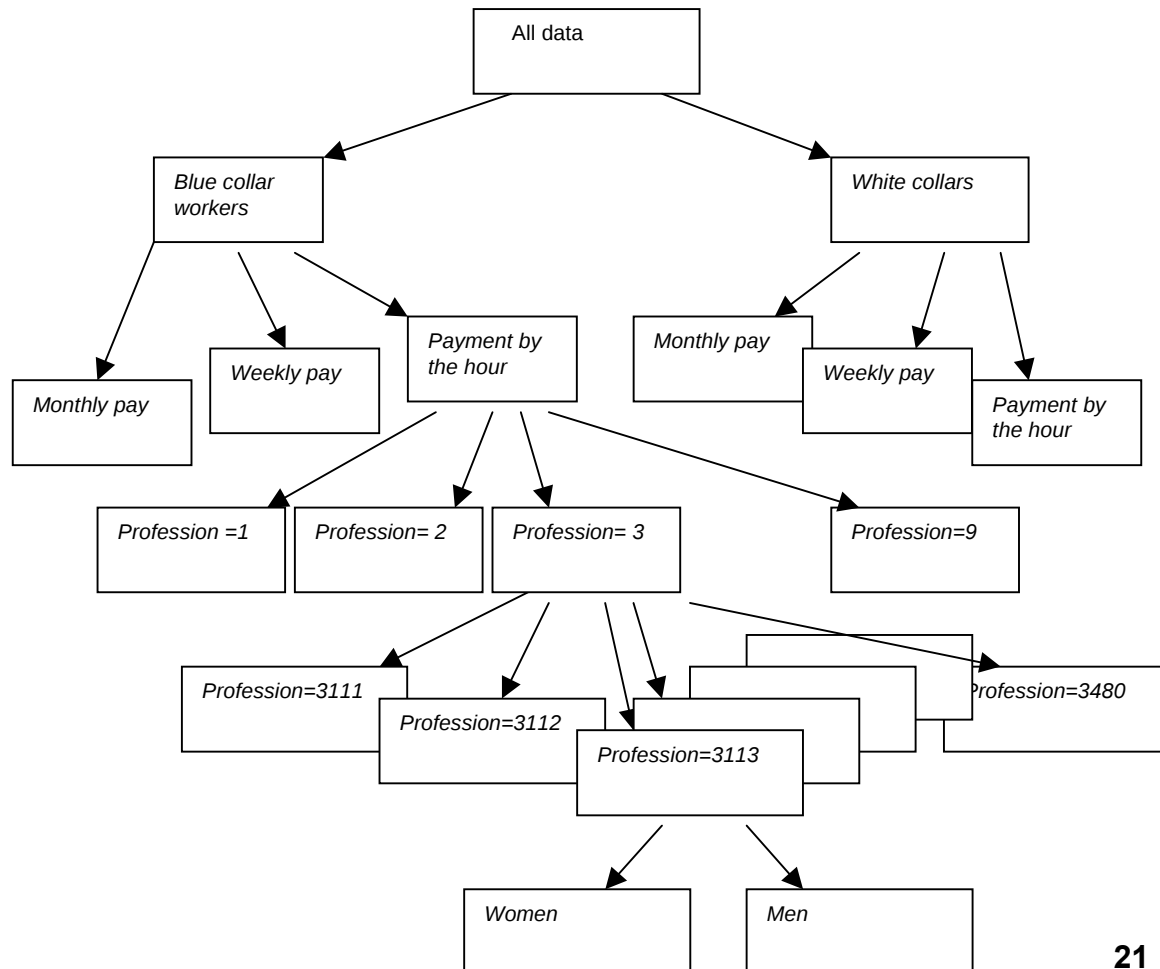
Predicted values

- Time series

- Previous value
- Forecast

- Cross section

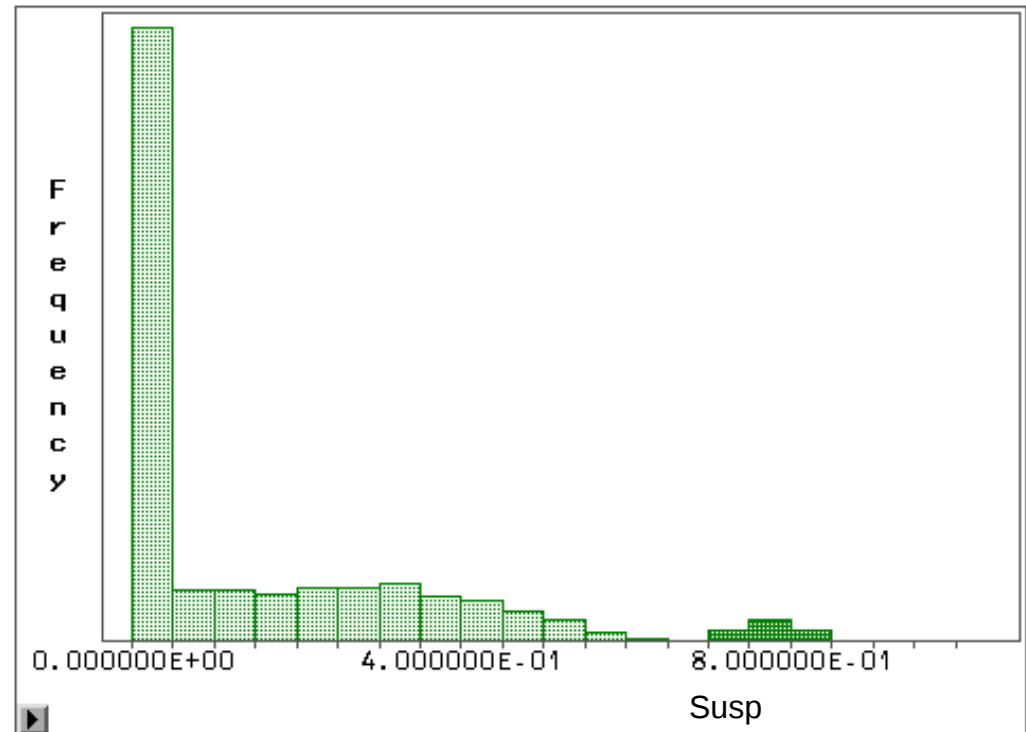
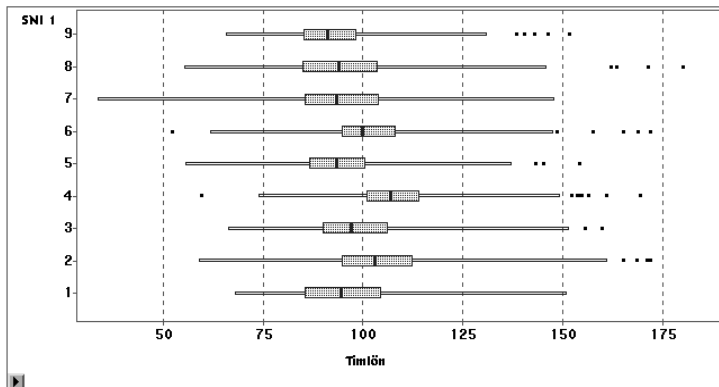
- Mean/standard error
- Median/quartile



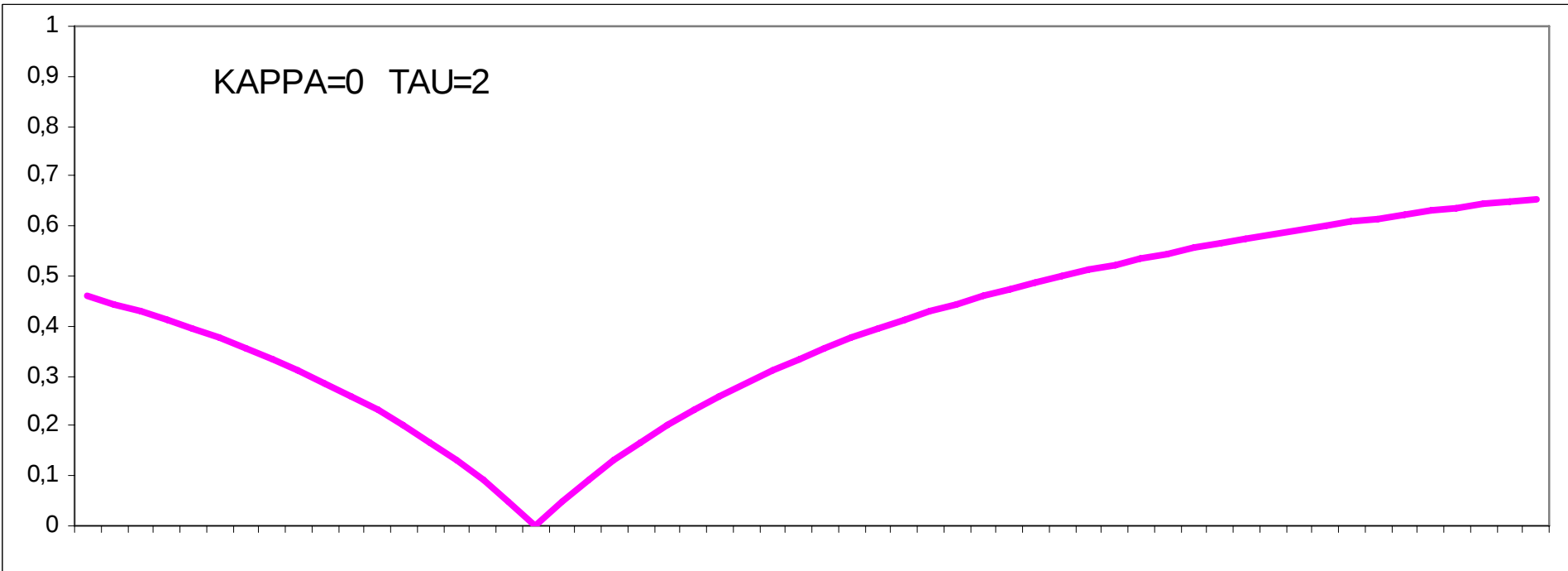
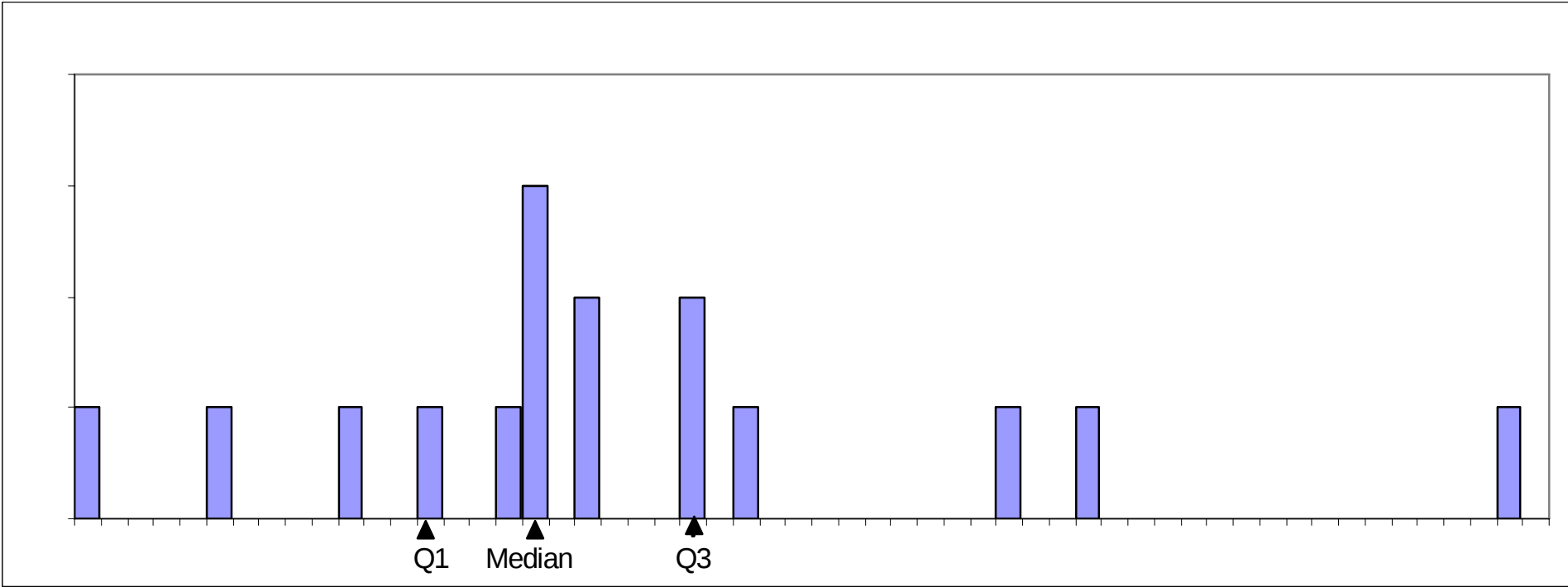
Misstanke

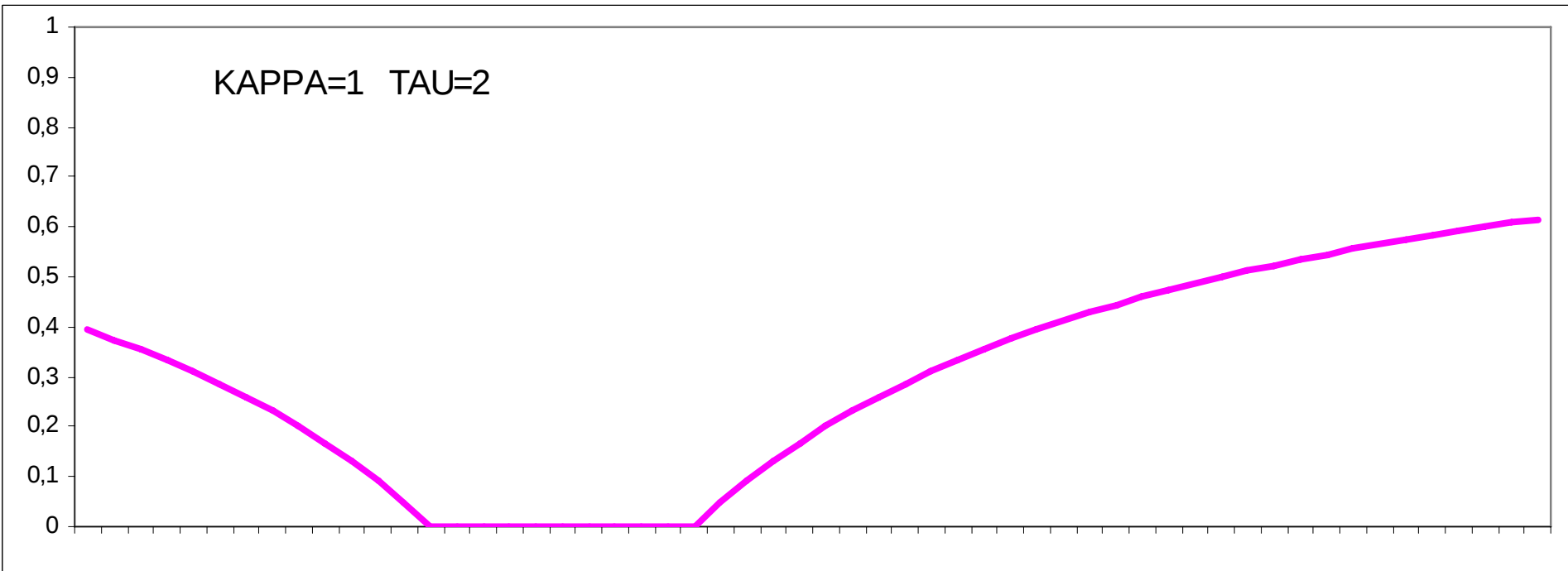
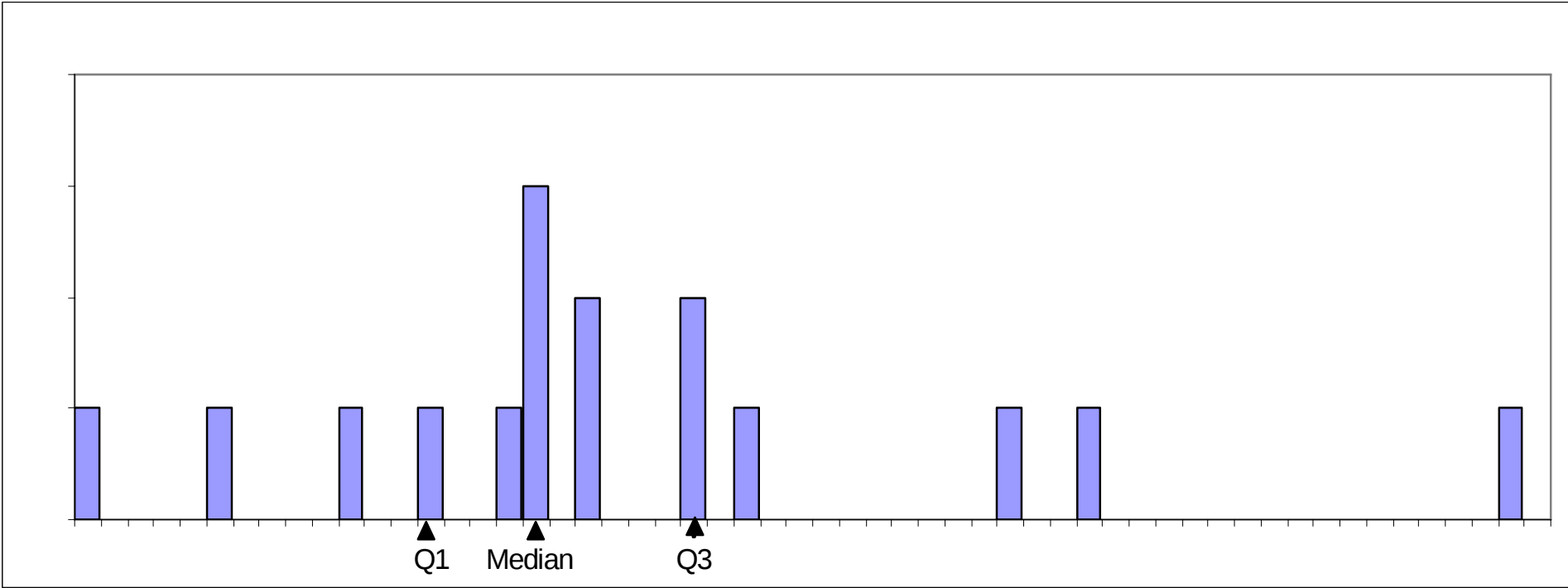
$$R = \begin{cases} (\tilde{z}_{j,k,l} - \text{KAU} \cdot (\tilde{z}_{j,k,l} - \tilde{z}_{j,k,l}^L) - z_{j,k,l}) / (\tilde{z}_{j,k,l}^U - \tilde{z}_{j,k,l}^L) & \text{if } z_{j,k,l} < \tilde{z}_{j,k,l} - \text{KAU} \cdot (\tilde{z}_{j,k,l} - \tilde{z}_{j,k,l}^L) \\ 0 & \text{if } \tilde{z}_{j,k,l} -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

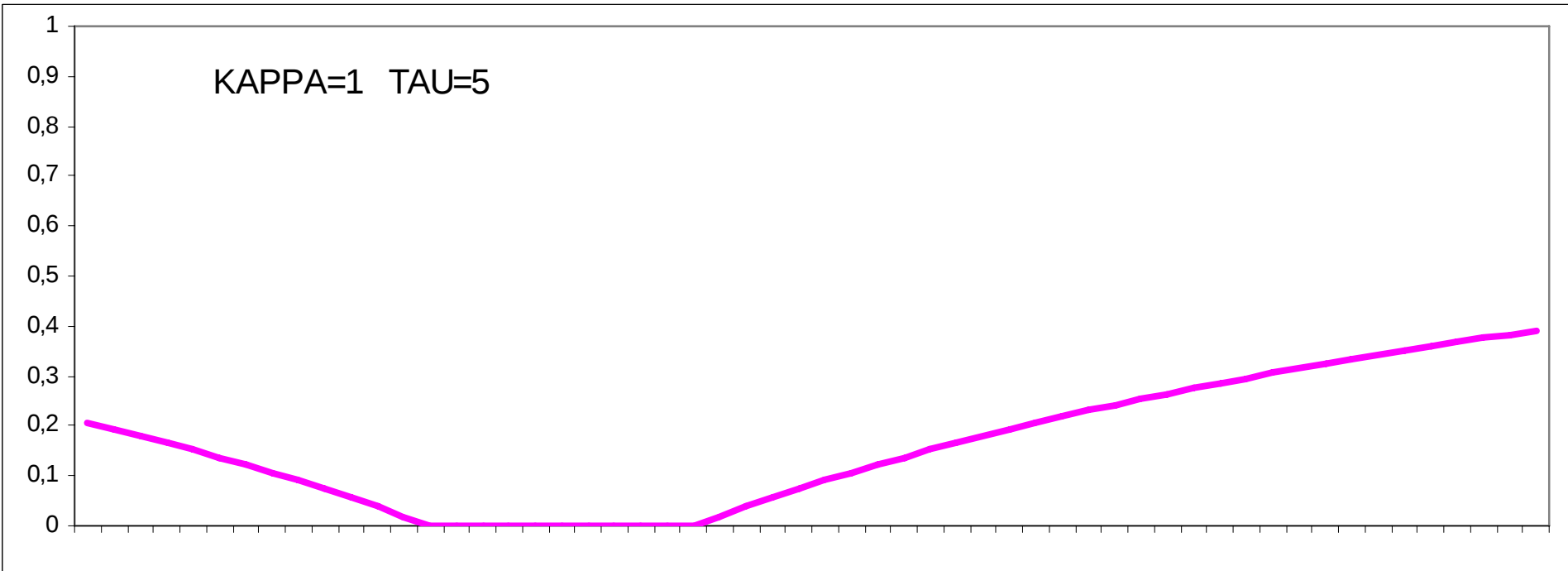
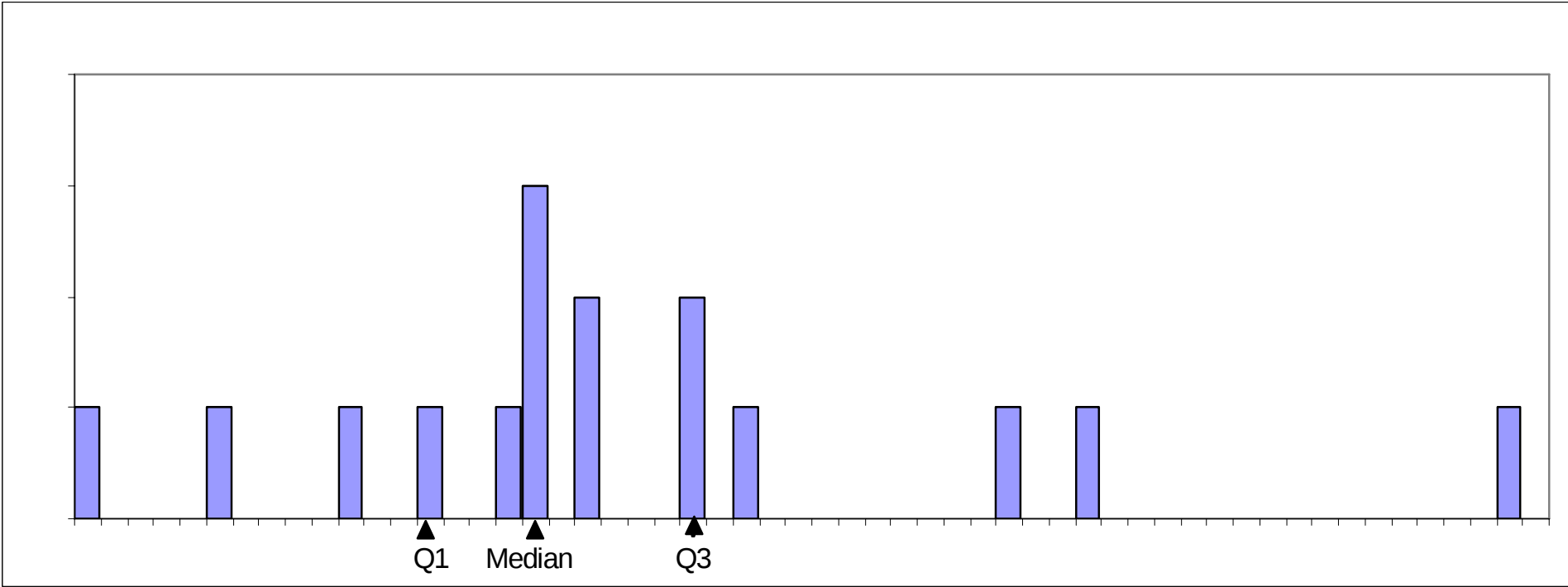
$$\text{Susp} = R / (\text{TAU} + R)$$

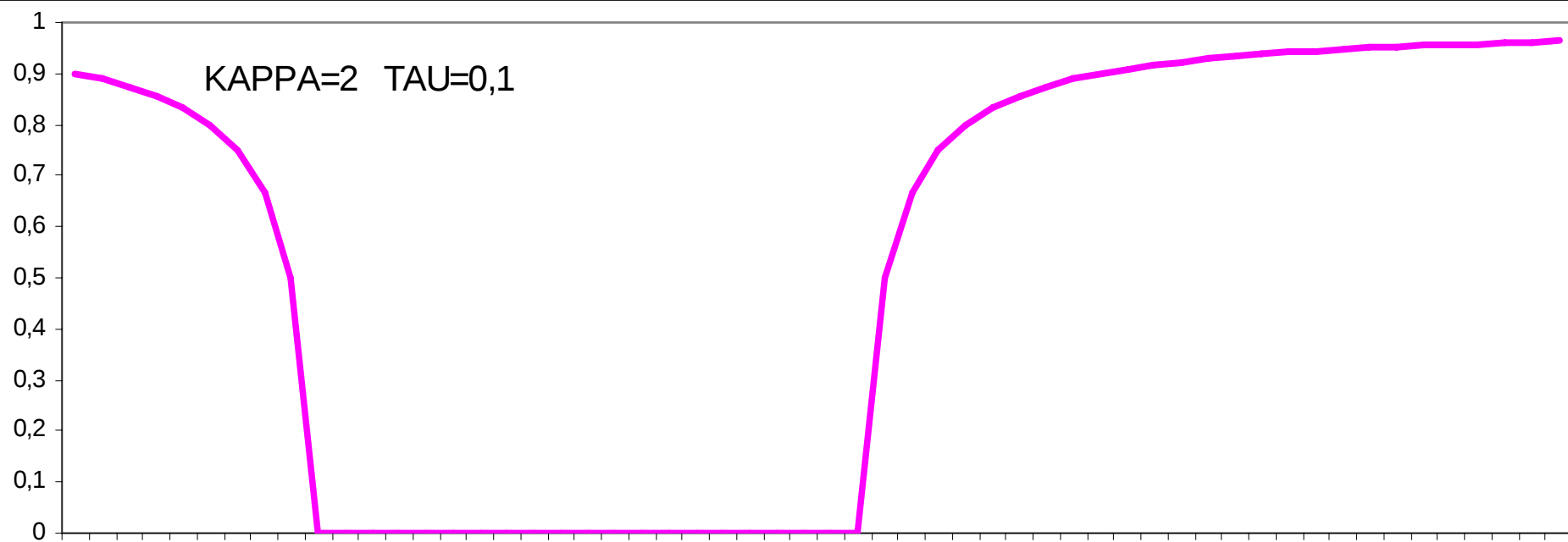
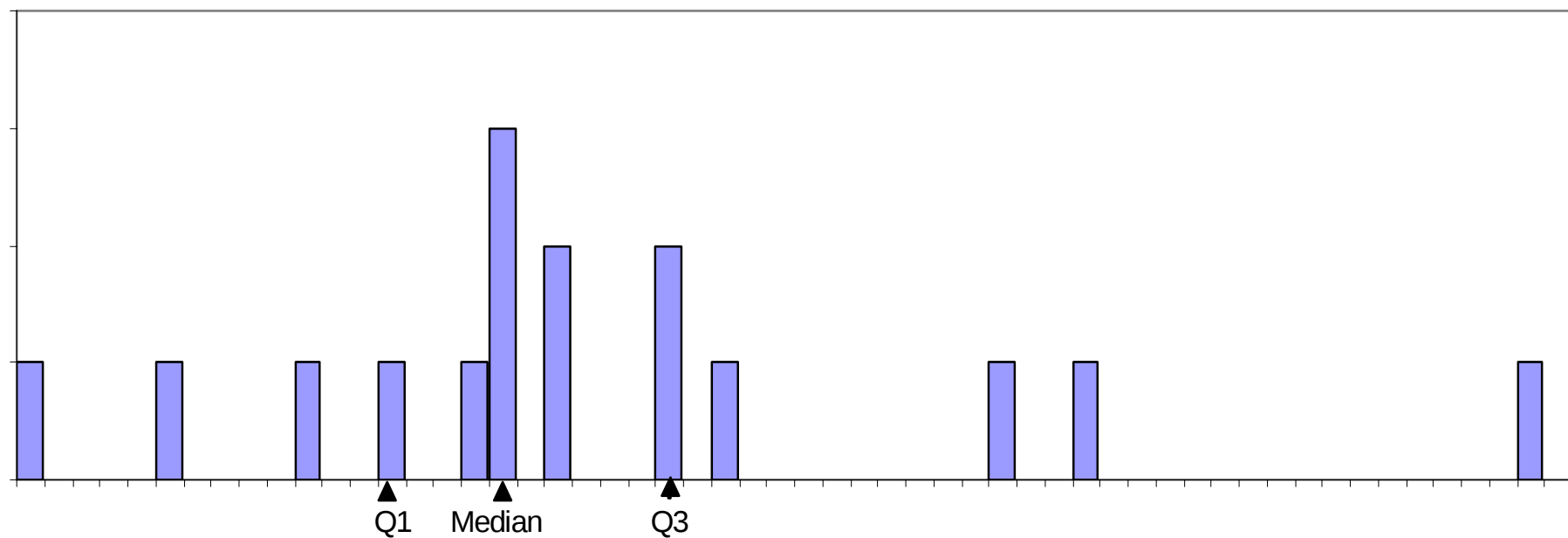


Homogena grupper









Score functions

- 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)
- $LScore_{d,j,k,l} = Susp_{j,k,l} \times (\text{Potential impact}) \times CELLO_{d(c),j}$

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

Score functions

- 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.

$$\text{GScore}_k = \left(\sum_1 \left(\max \{ 0, \text{LScore}_{k,l} - \text{BETA} \} \right)^{\text{LAMBDA}} \right)^{1/\text{LAMBDA}}$$

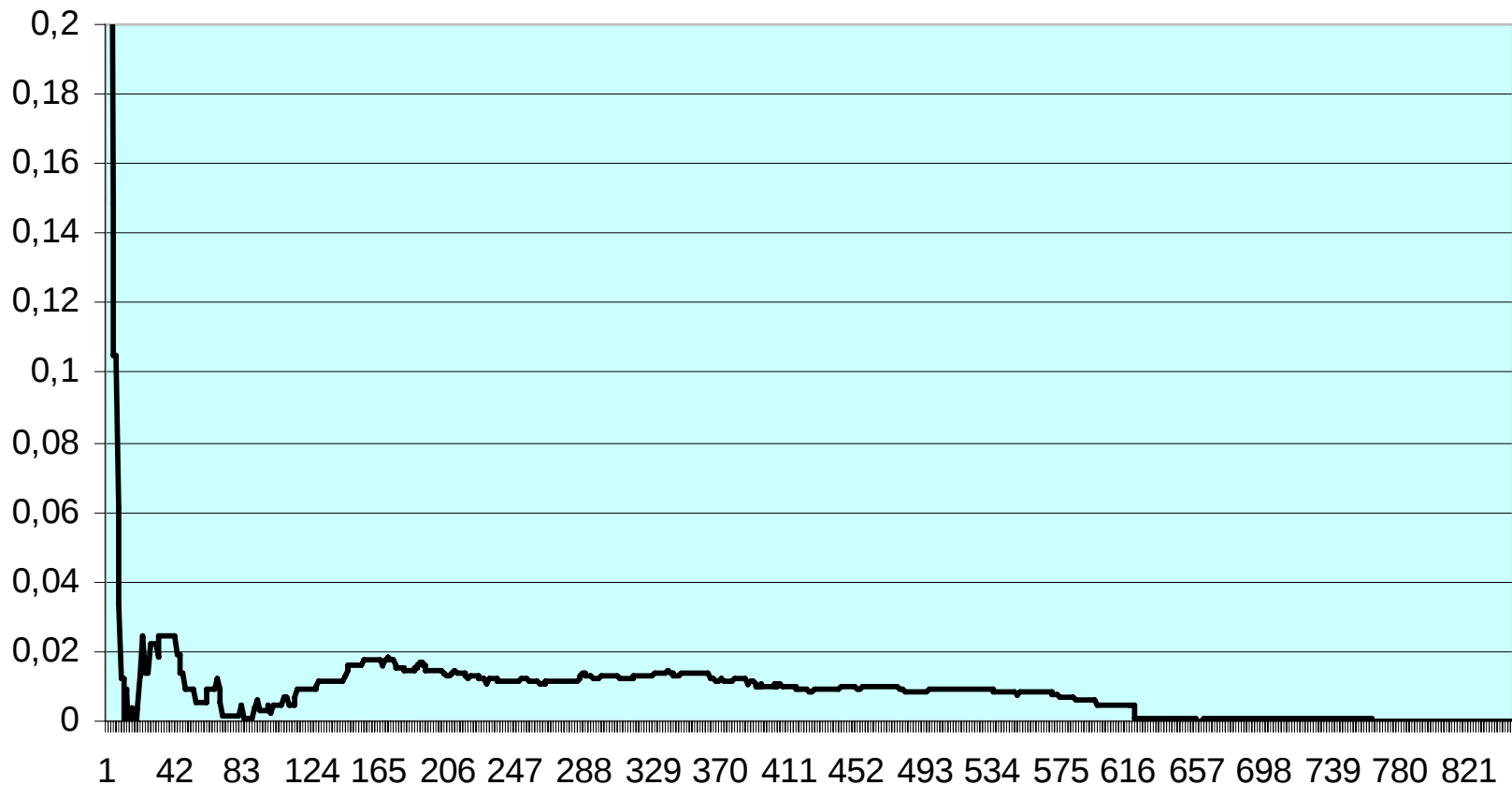
Utvärdering

Relativ pseudobias är ett mått på statistikens kvalitet till följd av ofullständig granskning

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

Felsignalering

Pseudobias for PPI relative to the overall price index. Observation units ordered in descending order of impact.



Felsignalering - metodfrågor

- Uppenbara fel
 - Klassificeringsvariabler
 - Mätvariabler
- Otillåtet värde, validitetsfel
- Värde saknas men måste finnas, partiellt bortfall
- Datastruktur- eller modellfel (totalen = summan av komponenterna)
- Motsägelser
- Förtroende (uppgiftslämnare & kund)
- Nya och gamla uppgiftslämnare
- Redan granskat
 - Elektroniska blanketter
 - Skanning
- Analysmaterial och metoder för att beräkna förväntade värden m.m.
- Hur konstruera homogena grupper
- Prioriteringar; variabler, redovisningsgrupper (ur kundperspektiv)
- Score function
- Bästa metod att fastställa tröskelvärde
- Urval under tröskeln
 - Inferens
 - Analysmaterial

Data editing process

