THE "JACKKNIFE" METHOD

Confidentiality Protection For Complex Statistical Analyses

Jobst Heitzig
Federal Statistical Office
Germany

The goal: Remote Access

- **Web service for statistical analyses of microdata**
- **All kinds of analyses**
  - estimation, (non-)parametric tests, (non-)linear regressions, ...
  - univariate and multivariate, classical and robust, ...

Obvious problem: Confidentiality

- **User gets results, but no micro-data**
  - simple form known as “Statistical Databases”
- **Measure 1: Level of confidentiality protection**
- **Measure 2: Quality of results**
  - no unneccessarily large errors, no bias, consistency of estimation

Why not use anonymised microdata?

- **Recall briefly these disadvantages:**
  - delay
  - guaranteed protection only against specific disclosure strategies
  - some (or even many) individual values remain unprotected
  - low quality of results
    - e.g.: subsampling 70% → relative std. errors increase by 20%
      e.g.: recoding 9 classes to 3 → factor 4 for rel. std. err. of χ^2
    - even for unproblematic analyses (e.g.: population median)
  - biased results for complex analyses
  - no results for some analyses
  - production cost

Why not simply judge by *N*?

- **What can happen anyway when, say, *N*>5?**
- **Example 1:**
  - ***N*=6**, two variables *X* and *Y*
  - *We cannot safely publish mean, variance, skewness, kurtosis, and covariance!*
    - If the "snooper" is one of the six and knows the *X* of his target, he can calculate the corresponding *Y* !
- **Example 2:**
  - ***N*=20**, two variables *X* and *Y*, Pearson correlation 0.99
  - *We should perhaps not publish mean, variance, and covariance!*
    - Anyone who knows an *X* gets the corresponding *Y* by linear regression with a sure precision of at most $\pm 0.6\, s_Y$

→ Using many observations does not imply safety!

But what else can we do?

- **Approach: Coarsen all results more or less**
  - instead of completely suppressing some results
  - e.g., publish intervals instead of precise results
- **Goal: Remove only as much information as necessary**
  - publish approximate results for problematic questions or small N
  - better results for unproblematic (e.g. robust) analyses and large N

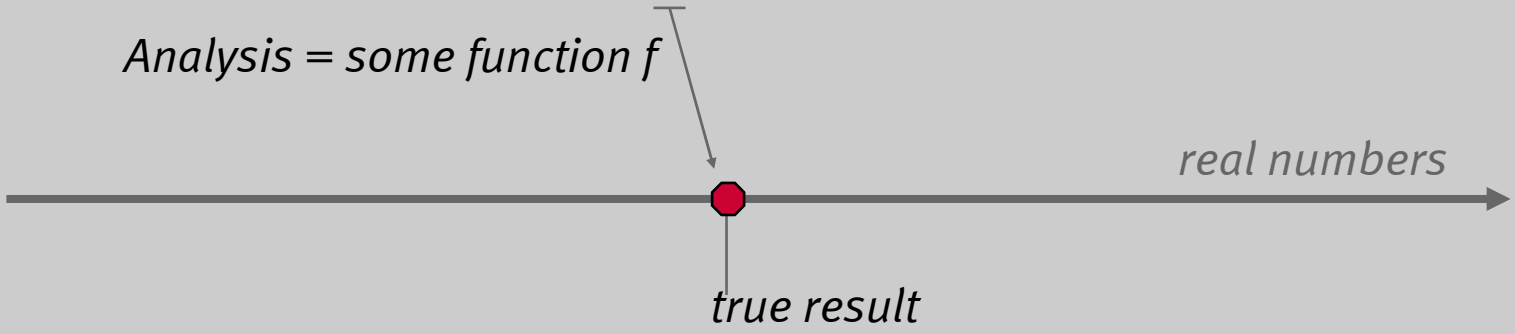
- **Idea (naïve but useful):**
  Replace the snooper's target value with a random value!
  $\rightarrow$ imprecision should typically be $\propto 1/N$
  (standard errors are typically larger: $\propto 1/\sqrt{N}$)

  Only, which is the target value?

...the „jackknife“ method of protection

Without protection: Use original micro-data file

a	b	c	d	e
01	0	03179	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101



...the „jackknife“ method of protection

Instead: Use modified files (modified at only one position each)!

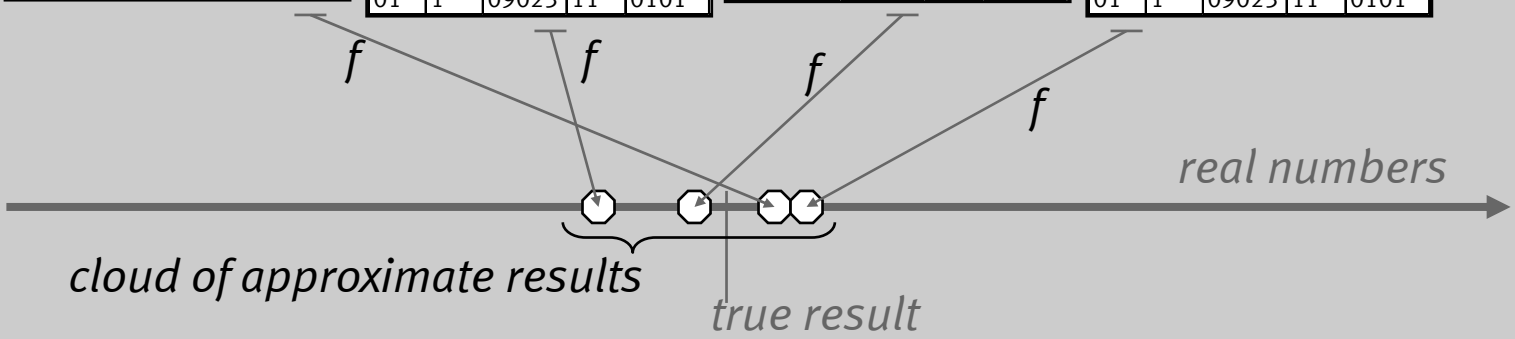
a	b	c	d	e
*	0	03179	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	*	03179	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	0	03179	07	0202
*	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	0	*	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

...



...the „jackknife“ method of protection

Instead: Use modified files (modified at only one position each)!

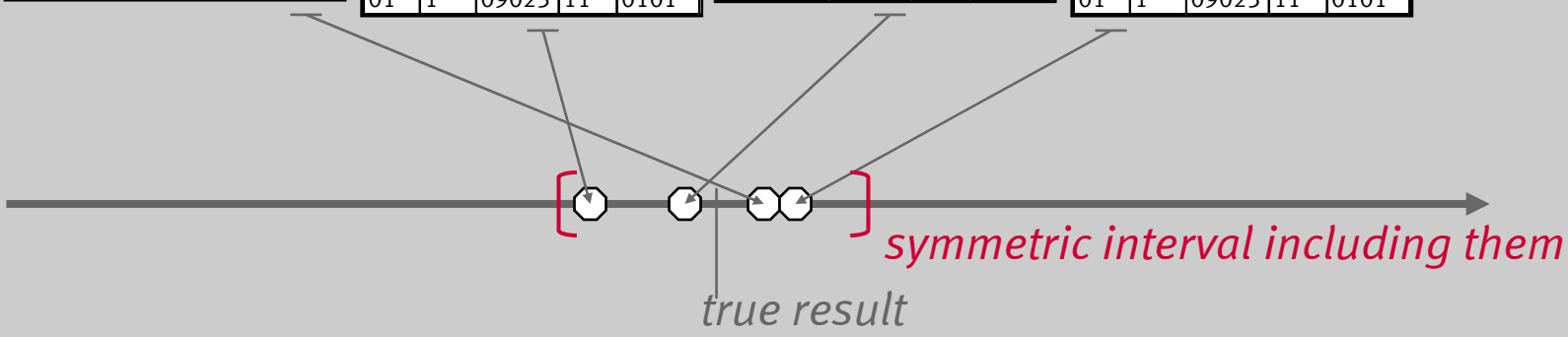
a	b	c	d	e
*	0	03179	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	*	03179	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	0	03179	07	0202
*	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	0	*	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

...



...the „jackknife“ method of protection

Instead: Use modified files (modified at only one position each)!

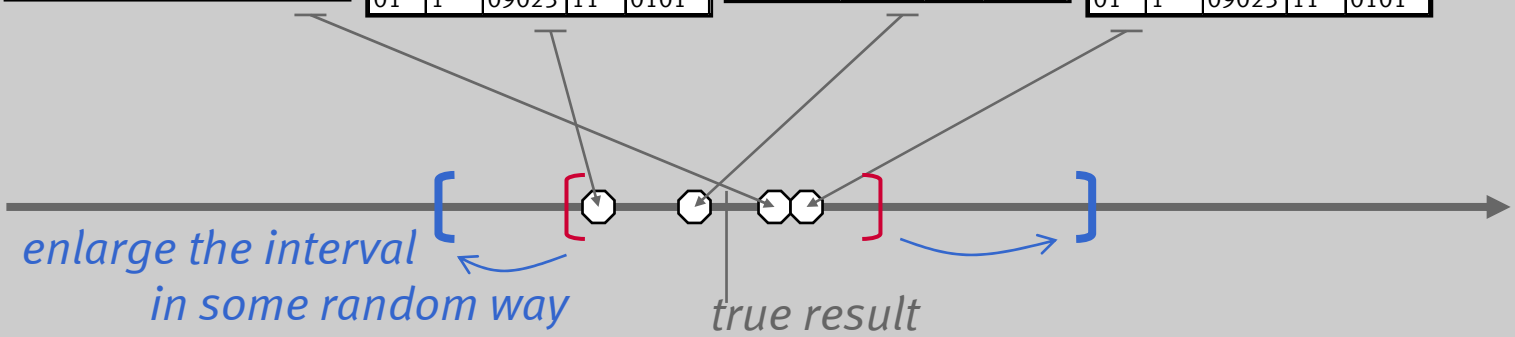
a	b	c	d	e
*	0	03179	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	*	03179	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	0	03179	07	0202
*	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	0	*	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

...



...the „jackknife“ method of protection

Instead: Use modified files (modified at only one position each)!

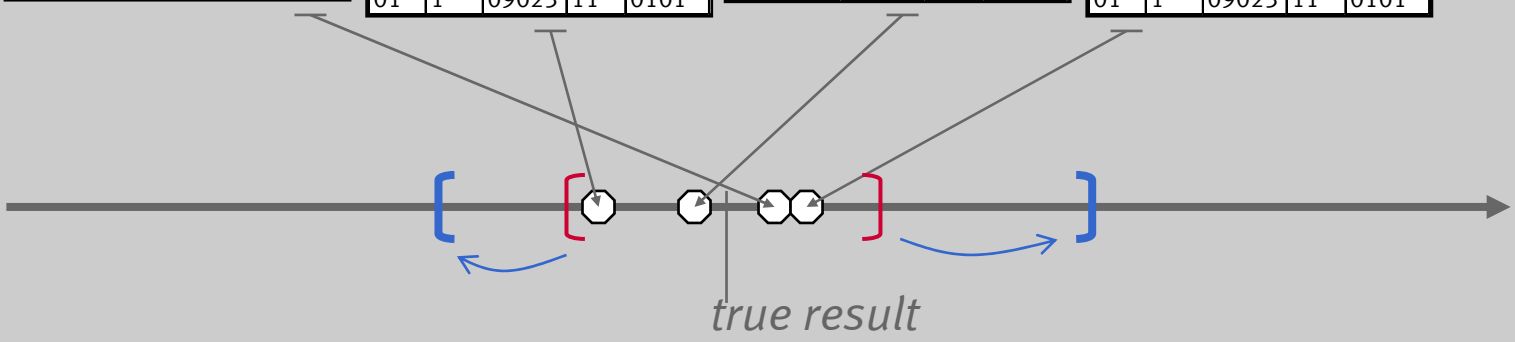
a	b	c	d	e
*	0	03179	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	*	03179	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	0	03179	07	0202
*	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

a	b	c	d	e
01	0	*	07	0202
01	0	10390	00	0202
01	0	67230	06	0101
01	0	67668	13	0101
01	1	02318	06	0101
01	1	02539	03	0101
01	1	05293	28	0101
01	1	06875	16	0202
01	1	09023	11	0101

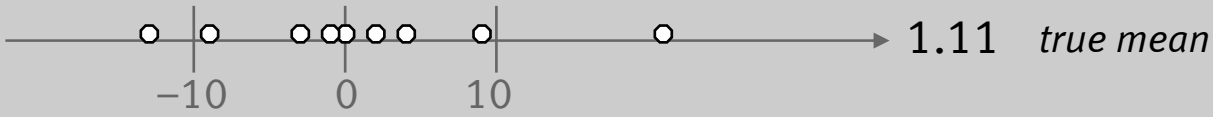
...



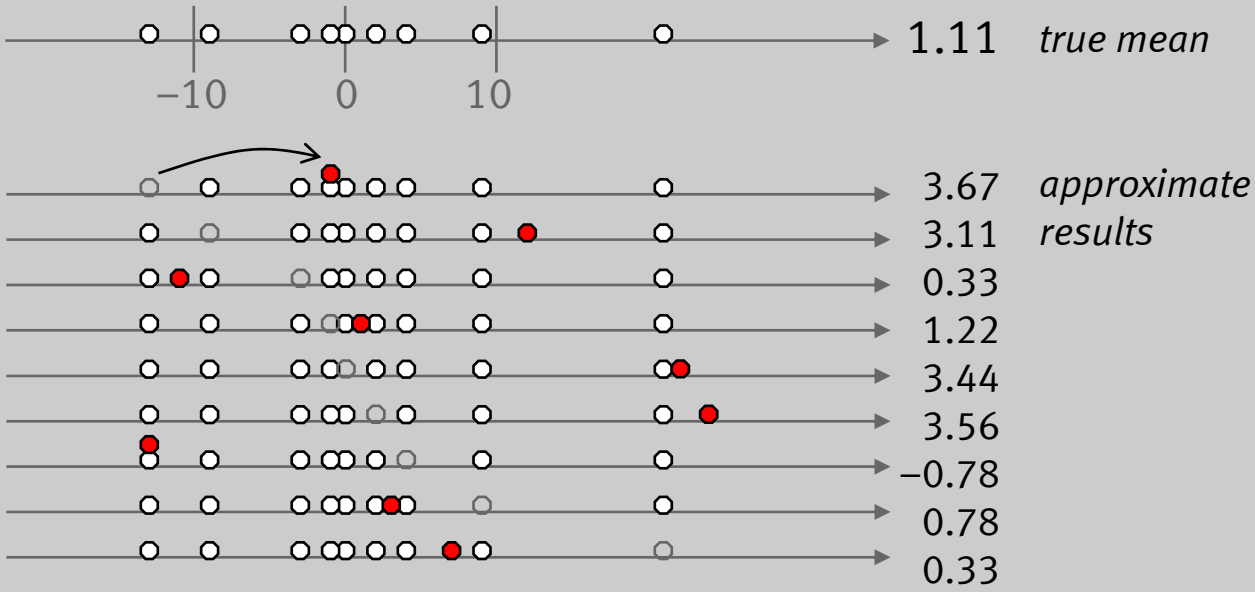
Publish the enlarged interval! (or its centre and width)

fortunately: *interval width* = o(*standard error*)

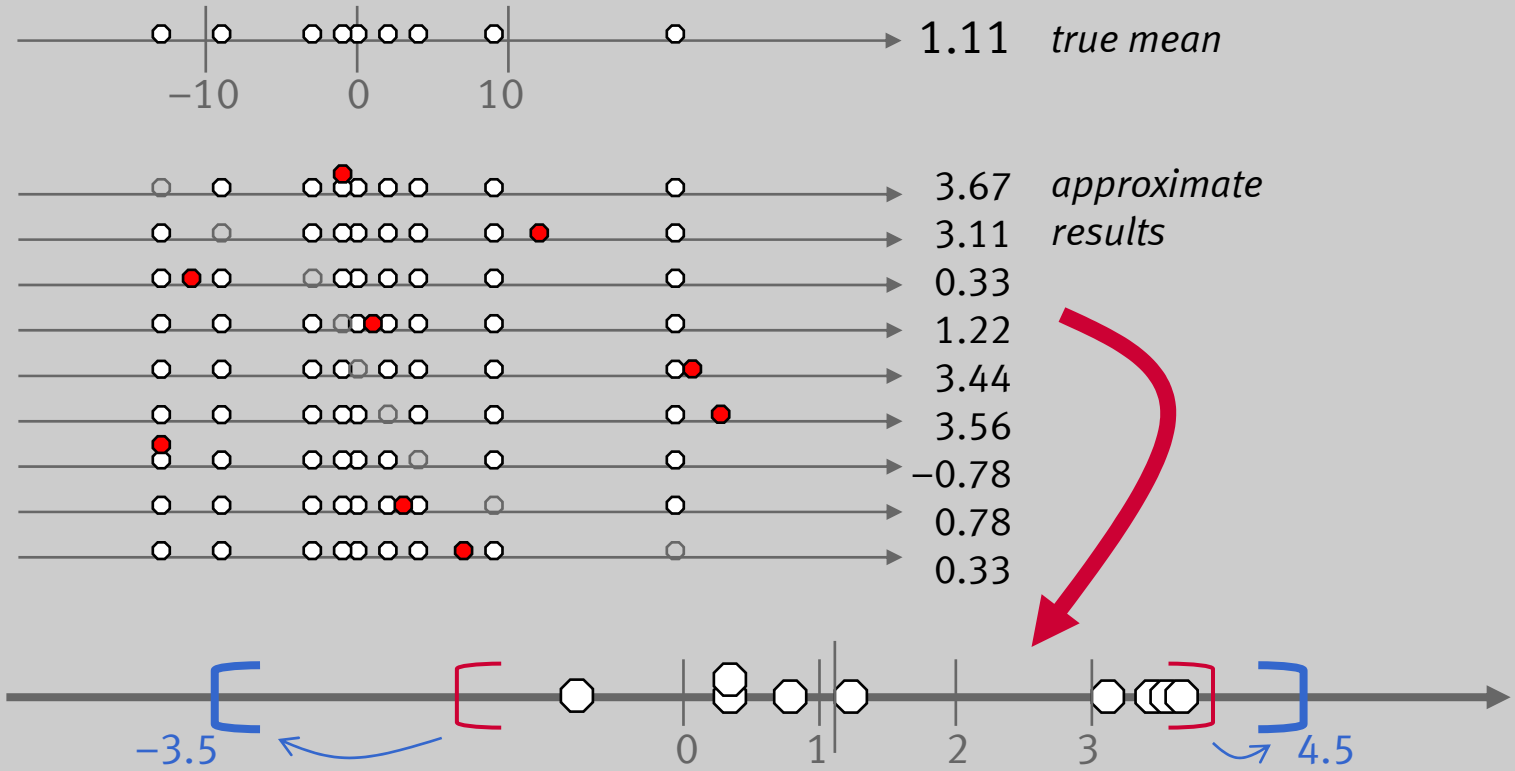
Nonrobust example: sample mean



Nonrobust example: sample mean



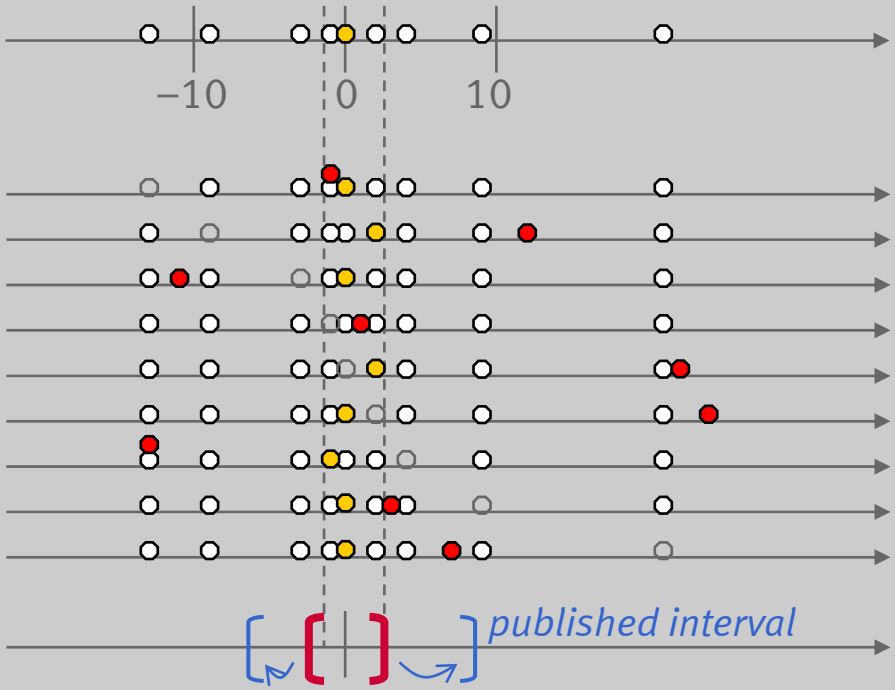
Nonrobust example: sample mean



Publish: *mean* ∈ [−3.5, 4.5] or *mean* = 0.5 ± 4

(*N* = 9, *standard error* = 3.31)

Robust example: sample median



0 *true median*

0 All approximate results
2 are between the two
0 neighboured values
1 of the true median
2
0
−1
0
0

Red interval can be computed without actually repeating the analysis with modified files!

Very robust → individual values have little influence
→ fast, high quality results

More complex example: χ^2 test

	Y=a	Y=b	Y=c	Σ
X=a	1	56	95	152
X=b	28	412	0	440
X=c	0	64	72	136
Σ	29	532	167	728

$$\chi^2 = 339.3$$

More complex example: χ^2 test

	Y=a	Y=b	Y=c	Σ
X=a	1	56	95	152
X=b	28	412	0	440
X=c	0	64	72	136
Σ	29	532	167	728

true $\chi^2 = 339.3$

	Y=a	Y=b	Y=c	Σ
X=a	0	57	95	152
X=b	28	412	0	440
X=c	0	64	72	136
Σ	28	533	167	728

$\chi^2 = 340.6$

	Y=a	Y=b	Y=c	Σ
X=a	0	56	96	152
X=b	28	412	0	440
X=c	0	64	72	136
Σ	28	532	168	728

$\chi^2 = 343.6$

	Y=a	Y=b	Y=c	Σ
X=a	0	56	95	151
X=b	29	412	0	441
X=c	0	64	72	136
Σ	29	532	167	728

$\chi^2 = 342.9$

	Y=a	Y=b	Y=c	Σ
X=a	0	56	95	151
X=b	28	413	0	441
X=c	0	64	72	136
Σ	28	533	167	728

$\chi^2 = 342.7$

	Y=a	Y=b	Y=c	Σ
X=a	0	56	95	151
X=b	28	412	1	441
X=c	0	64	72	136
Σ	28	532	168	728

$\chi^2 = 338.8$

	Y=a	Y=b	Y=c	Σ
X=a	0	56	95	151
X=b	28	412	0	440
X=c	1	64	72	137
Σ	29	532	167	728

$\chi^2 = 339.7$

	Y=a	Y=b	Y=c	Σ
X=a	0	56	95	151
X=b	28	412	0	440
X=c	0	65	72	137
Σ	28	533	167	728

$\chi^2 = 341.2$

	Y=a	Y=b	Y=c	Σ
X=a	0	56	95	151
X=b	28	412	0	440
X=c	0	64	73	137
Σ	28	532	168	728

$\chi^2 = 343.1$

	Y=a	Y=b	Y=c	Σ
X=a	2	55	95	152
X=b	28	412	0	440
X=c	0	64	72	136
Σ	30	531	167	728

$\chi^2 = 338.4$

	Y=a	Y=b	Y=c	Σ
X=a	1	55	96	152
X=b	28	412	0	440
X=c	0	64	72	136
Σ	29	531	168	728

$\chi^2 = 342.3$

	Y=a	Y=b	Y=c	Σ
X=a	1	55	95	151
X=b	29	412	0	441
X=c	0	64	72	136
Σ	30	531	167	728

$\chi^2 = 341.6$

	Y=a	Y=b	Y=c	Σ
X=a	1	55	95	151
X=b	28	413	0	441
X=c	0	64	72	136
Σ	29	532	167	728

$\chi^2 = 341.4$

...

→ interval [334, 344] → enlarge [ ] → publish $\chi^2 = 344 \pm 26$

Advanced example: nonlinear regression

- **Model:** $Y = h\,(X_1, \ldots, X_m\,;\,\theta_1, \ldots, \theta_k\,) + error$
- **Method: Newton with least squares loss function**
- **Result = fitted parameter vector** $\theta = (\theta_1, \ldots, \theta_k)$

How is θ affected by replacing a single value of X_i or Y?

Advanced example: nonlinear regression

- **Model:** $Y = h\,(X_1, \ldots, X_m\,;\,\theta_1, \ldots, \theta_k\,) + error$
- **Method: Newton with least squares loss function**
- **Result = fitted parameter vector** $\theta = (\theta_1, \ldots, \theta_k)$

How is θ affected by replacing a single value of X_i or Y?

- recompute the gradient of the loss function, $g_{\text{new}} = \frac{\partial}{\partial\theta} loss(X_{\text{new}};\theta)$
- multiply with the inverse of the Hessian of the loss function, $H = \frac{\partial^2}{\partial\theta^2} loss$:

$$\theta_{\text{new}} - \theta \approx H^{-1} g_{\text{new}} \quad \text{(approximately)}$$

(Proof: θ is implicitly defined by $\frac{\partial}{\partial\theta} loss(X;\theta) = 0$, Theorem on implicit functions)

Advanced example: nonlinear regression

- **Model:** $Y = h\,(X_1, \ldots, X_m\,;\,\theta_1, \ldots, \theta_k\,) + error$
- **Method: Newton with least squares loss function**
- **Result = fitted parameter vector** $\theta = (\theta_1, \ldots, \theta_k)$

How is θ affected by replacing a single value of X_i or Y?

- recompute the gradient of the loss function, $g_{\text{new}} = \frac{\partial}{\partial\theta} loss(X_{\text{new}};\theta)$
- multiply with the inverse of the Hessian of the loss function, $H = \frac{\partial^2}{\partial\theta^2} loss$:

$$\theta_{\text{new}} - \theta \approx H^{-1} g_{\text{new}} \quad \text{(approximately)}$$

fast
O(N)
O(1/N)

(Proof: θ is implicitly defined by $\frac{\partial}{\partial\theta} loss(X;\theta) = 0$,
Theorem on implicit functions)

Summary of jackknife protection

- **Good level of protection**
  - Even with complete additional knowledge of all other values, it should be impossible to infer a single target value
- **High quality of results**
  - Relative difference of published and true results $O(\ln N/N)$ vs. standard error $O(1/\sqrt{N})$
  - preserves unbiasedness, consistency, asymptotic normality and variance (but not UMVU- or BLUE-properties)
- **Acceptable performance**
  - $O(N)$ time for moment-based statistics and (non-)linear least-squares regression
  - $O(N \ln N)$ time for quantile-based statistics
  - $O(C^2R^2)$ time for $C\times R$-table statistics

Current status of jackknife protection

- **Prototypes available for evaluation:**
  - classical and robust univariate descriptives and tests (as in SAS Proc Mean / Proc Univariate and some more)
  - two-way frequency tables with table statistics and tests (as in SAS Proc Freq)
  - (non-)linear least squares regression (as in SAS Proc NLin)
- **Evaluation of practical quality of results is in progress**
  - comparison with results from anonymised business-data files
- **To do:**
  - further prototypes: correlations, forecasting, plots, ANOVA, ...
  - thorough proof of level of protection
  - comparison with anonymised household-data files
  - integration into a remote access facility

Thank you for your attention!

Jobst Heitzig, PhD

Statistisches Bundesamt
IT User Service / Statistical and Geo-Information Systems
Gustav-Stresemann-Ring 11
65189 Wiesbaden, Germany

jobst.heitzig@destatis.de