

Partially Identified Prevalence Estimation under Misclassification using the Kappa-Coefficient

Anne Kunz

Munich Cancer Registry

IBE, Biometry and Bioinformatics, University of Munich (LMU)

anne.kunz@gmx.net

Thomas Augustin

Department of Statistics, University of Munich (LMU)

thomas@stat.uni-muenchen.de

Helmut Küchenhoff

Department of Statistics, University of Munich (LMU)

kuechenhoff@stat.uni-muenchen.de

Anne Kunz, Thomas Augustin, Helmut Küchenhoff, LMU München

1. Misclassification

- Interest in

$$Y_i = \begin{cases} 1 & \text{diseased} \\ 0 & \text{not diseased} \end{cases}$$

$p := \mathbb{P}(Y = 1)$ *prevalence*

- Often only available

$$Y_i^* = \begin{cases} 1 & \text{test positive} \\ 0 & \text{test negative} \end{cases}$$

$p^* := \mathbb{P}(Y^* = 1)$ *naive prevalence*

- $Y_i^* \longleftrightarrow Y_i$ misclassification



The Signal-Tandmobiel[®] Study

- 6 years longitudinal oral health study (1996 - 2001)
- 4468 children in Flanders (Belgium)
- Annual examinations
- Additional calibration studies 1996, 1998, 2000
- Presence/absence of caries

$$\bullet Y_i^* = \begin{cases} 1 & \text{caries observed} \\ 0 & \text{no caries observed} \end{cases} \iff Y_i = \begin{cases} 1 & \text{caries} \\ 0 & \text{no caries} \end{cases}$$

2. Misclassification Bias

	$Y = 1$	$Y = 0$
$Y^* = 1$	$\mathbb{P}(Y^* = 1 Y = 1)$ sensitivity	$\mathbb{P}(Y^* = 1 Y = 0)$ false positive
$Y^* = 0$	$\mathbb{P}(Y^* = 0 Y = 1)$ false negative	$\mathbb{P}(Y^* = 0 Y = 0)$ specificity

Assume throughout reasonable quality of the test: $sens + spec > 1$. (1)

Prevalence estimation based on misclassified data may be severely biased.

$$\begin{aligned} p^* &= p \cdot sens + (1 - p) \cdot (1 - spec) = & (2) \\ &= p \cdot sens + 1 - spec - p = \\ &= p(sens + spec - 1) + (1 - spec) \end{aligned}$$

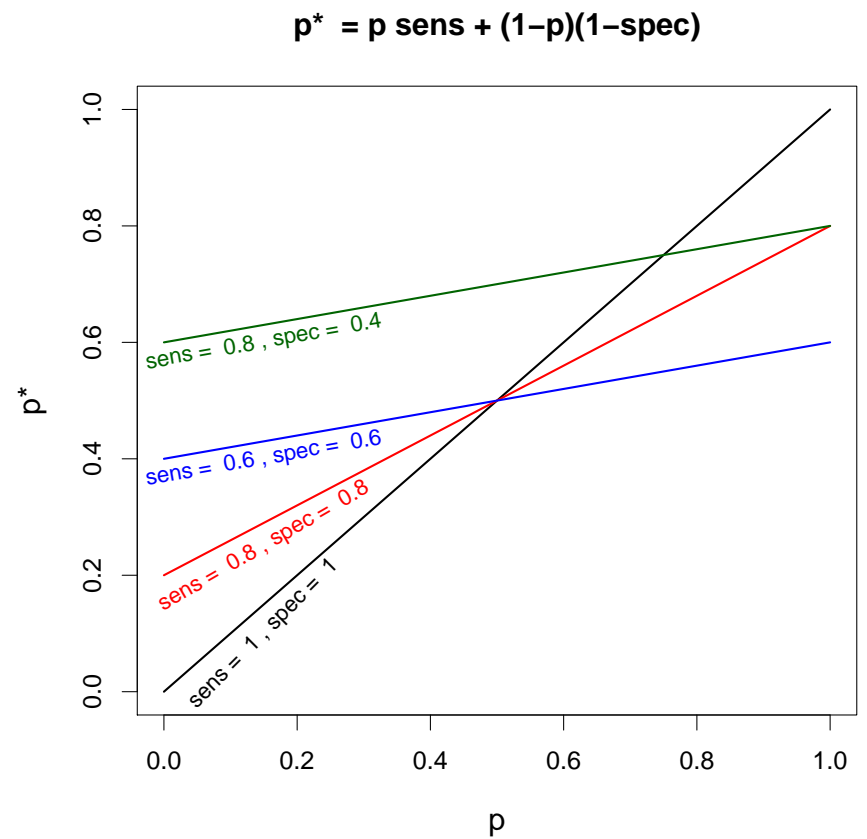
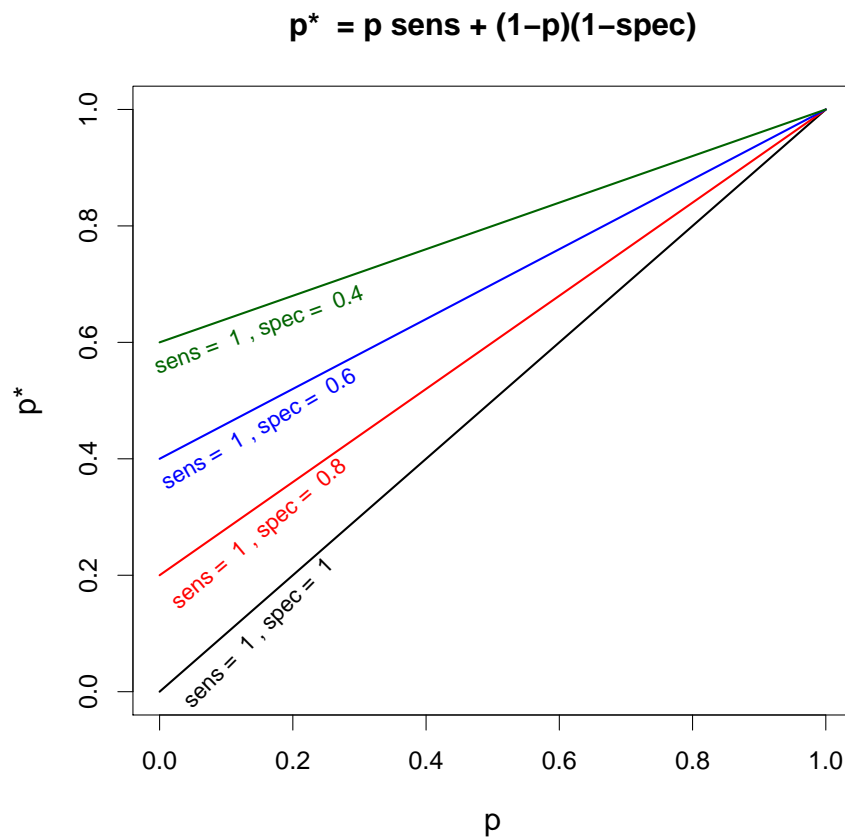


Figure 1: Illustration of misclassification bias (deviation from the angle bisector).

3. Correcting for Misclassification

3.1 The Extreme Cases

- If *sens* and *spec* are precisely known, (2) yields (together with (1)) an unbiased corrected prevalence estimator

$$\hat{p} = \frac{\hat{p}^* + spec - 1}{sens + spec - 1}. \quad (3)$$

- If there is complete ignorance on *sens* and *spec* then the corrected prevalence estimator is vacuous:

$$\hat{p} = [0; 1].$$

3.2 Use Additional Knowledge: kappa Coefficient

- Quite often two replicated measurements (\rightarrow two raters)
- In particular in medicine, common characterization of the quality of measurements by kappa, a coefficient of inter-rater agreement

$Y^{*(2)} \ Y^{*(1)}$	1	0
1	p_{11}	p_{10}
0	p_{01}	p_{00}

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (4)$$

with

$$p_{jk} = P(Y^{*(1)} = j, Y^{*(2)} = k), \quad p_o = p_{00} + p_{11}$$
$$p_e = (p_{00} + p_{01}) \cdot (p_{00} + p_{10}) + (p_{10} + p_{11}) \cdot (p_{01} + p_{11})$$

Classification of κ according to Landis & Koch

Landis and Koch (1977) proposed a widely used classification of the kappa-statistic (see Table 1) to "maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics [...]. Although these divisions are clearly arbitrary, they do provide useful 'benchmarks' for the discussion".

Table 1: Classification of κ according to Landis & Koch (and alternative common terminology)

kappa-statistic	strength of agreement
≤ 0.00	poor
0.01 - 0.20	slight (insufficient)
0.21 - 0.40	fair (satisfactory)
0.41 - 0.60	moderate (sufficient)
0.61 - 0.80	substantial (good)
0.81 - 1.00	almost perfect (excellent)

Under the assumptions

(A1) Independent conditional distributions $Y_1^*|Y$ and $Y_2^*|Y$ for both replicates

(A2) Equal sensitivity and specificity for both replicates

the following equation can be deduced:

$$\kappa = \frac{p(1-p)(sens + spec - 1)^2}{(spec - p(sens + spec - 1)) \cdot (1 - spec + p(sens + spec - 1))} \quad (5)$$

Together with (see (2) above)

$$p^* = p \cdot sens + (1 - p) \cdot (1 - spec)$$

this does not yield a unique solution.

3.3 Point Identification through Additional Constraints

Traditional way to proceed: add additional assumptions leading to point identification: $sens = spec$, or more general by $\frac{sens}{spec} =: \gamma$ known.

Then the system described by (4) and (5) indeed has a unique solution, leading to

$$p(\gamma, \hat{p}^*, \hat{\kappa}) = \frac{(1 - p^*) \cdot \gamma - p^* - \sqrt{w}}{(p^* - 1) \cdot \gamma^2 + (1 - \sqrt{w}) \cdot \gamma - p^* - \sqrt{w}} \quad (6)$$

with $w = \sqrt{(\hat{p}^* - 1)^2 \cdot \gamma^2 - 2 \cdot \hat{p}^* \cdot (\hat{p}^* - 1) \cdot (2 \cdot \hat{\kappa} - 1) \cdot \gamma + (\hat{p}^*)^2}$

3.4 Partial Identification – Basic Ideas

- Point identification only possible under non-testable assumptions
- Manski's (2003) law of decreasing credibility: The credibility of inference decreases with the strength of the assumptions maintained.
- Look at empirical evidence alone, consider *all* models compatible with the data.
- Parallel development in econometrics *partial identification* (initiated by Manski) and in biometrics systematic *sensitivity analysis* (e.g. Vansteelandt et al. (2006, Stat. Sinica))

3.5 Identification Regions for Prevalence Estimations

Let $\kappa > 0$ and (A1), (A2) and (1) be satisfied. Then the identification regions $I(z||p^*, \kappa) = \{z \text{ satisfying (2) and (5) for given } p^*, \kappa\}$, for $z \in \{p, \textit{sens}, \textit{spec}\}$ are

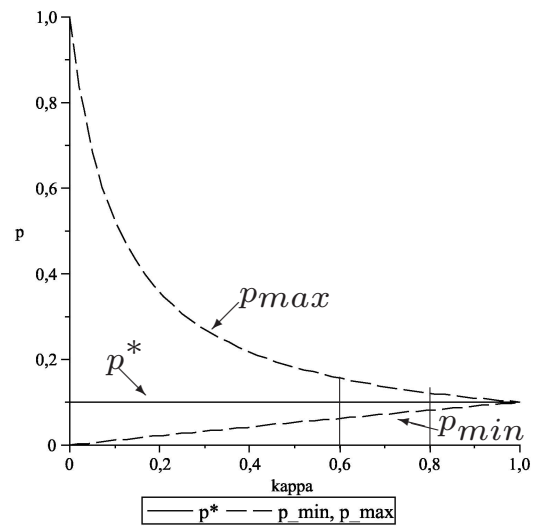
$$I(p||p^*, \kappa) := \left[\frac{p^*}{p^* + \kappa^{-1}(1 - p^*)} ; \frac{p^*}{p^* + \kappa(1 - p^*)} \right] \quad (7)$$

$$I(\textit{sens}||p^*, \kappa) := [p^* + \kappa(1 - p^*) ; 1] \quad (8)$$

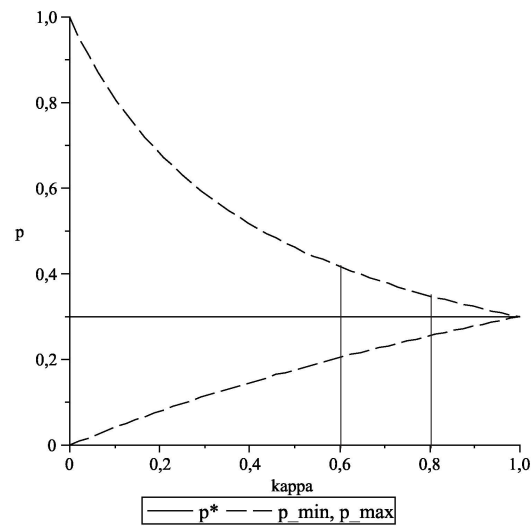
$$I(\textit{spec}||p^*, \kappa) := [1 - p^* + p^* \kappa ; 1] \quad (9)$$

$$\kappa \rightarrow 0 \quad : \quad I_p(p^*, \kappa) = [0, 1]$$

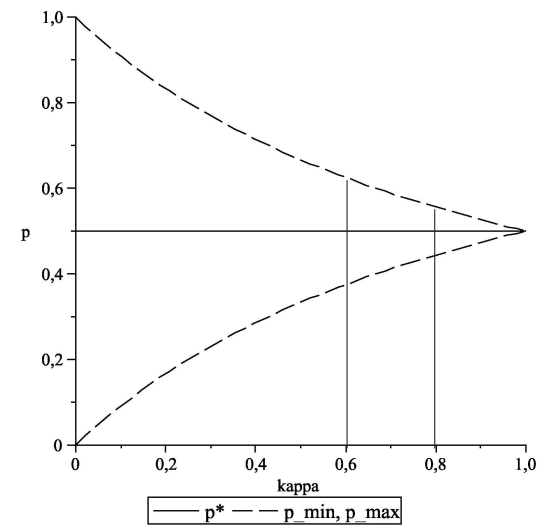
$$\kappa \rightarrow 1 \quad : \quad I_p(p^*, \kappa) = p^*, \quad \textit{sens} = \textit{spec} = 1$$



(a) $p^* = 0.1$



(b) $p^* = 0.3$



(c) $p^* = 0.5$

Figure 2: Identification regions for prevalence p

4. Confidence Intervals for Partially Identified Prevalence Estimation

- Identification regions reflect lack of knowledge („ignorance“), but do not take into account sample variability („uncertainty“).
- Construction of confidence regions (Imbens & Manski (2004, *Econometrica*), Stoye (2009, *Econometrica*); similarly: Vansteelandt, Goetghebeur, Kenword, Molenberghs (2006, *Stat. Sinica*) used here)
- *Identification parameter* $\gamma := \frac{sens}{spec}$; given γ all parameters would be identified (cp. section 3.3)¹

¹Vansteelandt et. al. (2006) use the term *sensitivity parameter*, which, however, is not used here to avoid terminological conflict with the sensitivity *sens*.

- γ can be shown to vary between γ_{min} and γ_{max} , where

$$[\gamma_{min}, \gamma_{max}] = \left[\kappa + p^* - \kappa \cdot p^*; \frac{1}{\kappa \cdot p^* - p^* + 1} \right]. \quad (10)$$

- Construct confidence interval $[L(\hat{p}^*, \hat{\kappa}); U(\hat{p}^*, \hat{\kappa})]$ such that

$$\inf_{\gamma \in [\hat{\gamma}_{min}, \hat{\gamma}_{max}]} Pr_{\gamma}(p \in [L(\hat{p}^*, \hat{\kappa}); U(\hat{p}^*, \hat{\kappa})]) \geq 1 - \alpha \quad (11)$$

- Use suitable confidence intervals $[L(\hat{p}^*, \hat{\kappa}, \gamma); U(\hat{p}^*, \hat{\kappa}, \gamma)]$ given γ and take

$$[L(\hat{p}^*, \hat{\kappa}); U(\hat{p}^*, \hat{\kappa})] := \bigcup_{\gamma \in [\hat{\gamma}_{min}, \hat{\gamma}_{max}]} [L(\hat{p}^*, \hat{\kappa}, \gamma); U(\hat{p}^*, \hat{\kappa}, \gamma)] \quad (12)$$

Consider for fixed γ the point estimator for p derived from (6)

$$\hat{p}(\gamma, \hat{p}^*, \hat{\kappa}) = \frac{(1 - \hat{p}^*) \cdot \gamma - \hat{p}^* - \sqrt{\hat{w}}}{(\hat{p}^* - 1) \cdot \gamma^2 + (1 - \sqrt{\hat{w}}) \cdot \gamma - \hat{p}^* - \sqrt{\hat{w}}} \quad (13)$$

with $\hat{w} = \sqrt{(\hat{p}^* - 1)^2 \cdot \gamma^2 - 2 \cdot \hat{p}^* \cdot (\hat{p}^* - 1) \cdot (2 \cdot \hat{\kappa} - 1) \cdot \gamma + (\hat{p}^*)^2}$

The asymptotic variance is given by the delta method

$$\text{Var}(\hat{p}(\gamma, \hat{p}^*, \hat{\kappa})) = D_p^T \Sigma D_p \quad (14)$$

with, D_p as the vector of derivatives of $\hat{p}(\gamma, \hat{p}^*, \hat{\kappa})$ with respect to \hat{p}^* and $\hat{\kappa}$, and Σ the covariance matrix of \hat{p}^* and $\hat{\kappa}$.

Since the relationship (13) between γ and p is monotone, the choice of the confidence intervals in (12) can be improved. If all the confidence intervals given γ are small compared to the uncertainty region, an asymptotic confidence interval is given by

$$\left[\hat{p}(\widehat{\gamma}_{max}) - z_{1-\alpha} \cdot \sqrt{\widehat{Var}(\hat{p}(\widehat{\gamma}_{max}))}; \hat{p}(\widehat{\gamma}_{min}) + z_{1-\alpha} \cdot \sqrt{\widehat{Var}(\hat{p}(\widehat{\gamma}_{min}))} \right] \quad (15)$$

$$sens + spec > 1$$

Since $\widehat{\gamma}_{min}$, $\widehat{\gamma}_{max}$ estimate γ_{min} and γ_{max} consistently, this is an asymptotic level $(1 - \alpha)$ –confidence interval.

5. Results from the Signal-Tandmobiell[®] Study

Table 2: Signal-Tandmobiell[®] study: Estimation of p^* per year

<i>year</i>	<i>n</i>	\hat{p}^*	$se(\hat{p}^*)$
1996 (age 6)	3378	0.118	0.006
1998 (age 8)	3657	0.280	0.007
2000 (age 10)	3415	0.380	0.008

- For illustration interpret validation measurement as a replicate,
- additionally satisfying (A1) and (A2)
- Further assumption: treat validation sample as a random sample

Table 3: Signal-Tandmobiel[®] study: Estimation of κ per year

<i>year</i>	<i>n</i>	$\hat{\kappa}$	<i>se</i> ($\hat{\kappa}$)
1996	120	0.575	0.084
1998	157	0.602	0.066
2000	148	0.746	0.057

Table 4: Signal-Tandmobiel[®] study: Estimated identification regions for p , *sens* and *spec*

<i>year</i>	\hat{p}^*	$\hat{\kappa}$	\hat{p}		<i>sens</i>		<i>spec</i>	
			<i>min</i>	<i>max</i>	<i>min</i>	<i>max</i>	<i>min</i>	<i>max</i>
1996	0.118	0.575	0.072	0.189	0.625	1.000	0.950	1.000
1998	0.280	0.602	0.190	0.393	0.714	1.000	0.889	1.000
2000	0.380	0.746	0.314	0.451	0.843	1.000	0.903	1.000

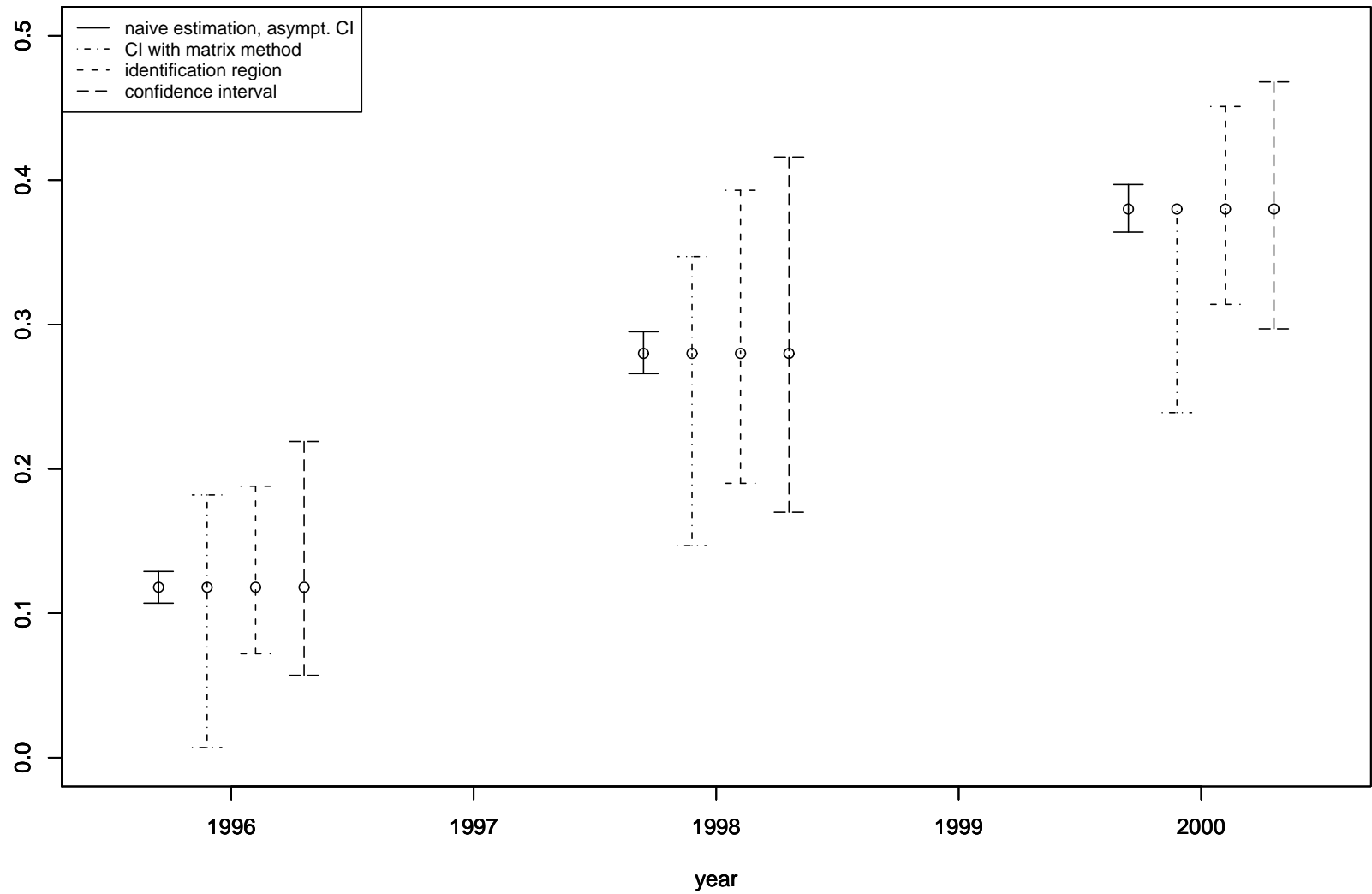


Figure 3: Signal-Tandmobiel[®]: confidence limits for \hat{p}

6. Conclusion

- Prevalence estimation based on kappa is an instance where the idea of partial identification (econometrics) and systematic sensitivity analysis (biometrics) allows for reliable but still informative conclusions.
- Introducing successively additional assumptions would lead to smaller, but less credible regions.
- Optimize confidence interval (finite sample correction)!