AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 1, 57–66

Asymmetric Loss Functions and Sample Size Determination: A Bayesian Approach

Hans Peter Stüger
Institute of Applied Statistics, Joanneum Research, Austria

Abstract: In designing monitoring systems for public health tasks it can be important to give different weights to the cases of under- and overestimation of a binomial parameter. We show how asymmetric loss functions can be used for this aim. Bayesian interval-based approaches can be combined with these loss functions and with prior knowledge about diagnostic classification errors to determine optimal sample sizes.

Zusammenfassung: Beim Design von Monitoringsystemen kann es wichtig sein, der Über- bzw. Unterschätzung eines Binomialparameters unterschiedliches Gewicht zu geben. Wir zeigen wie asymmetrische Verlustfunktionen dafür genutzt werden können. Weiters wird erläutert, wie bayesianische intervallbasierte Ansätze mit diesen Verlustfunktionen sowie mit Priorwissen über diagnostische Fehlklassifikationen kombiniert werden können, um optimale Stichprobenumfänge zu bestimmen.

Keywords: Bayesian Sample Size Determination, Decision Theory, Sensitivity, Specificity.

1 Introduction

Statistical methods have reached quite a high level of importance in many application fields e.g. public health or social-economic problems. Collecting data to get robust answers affords substantial monetary funds. Therefore the decision-makers in politics and administration (but also the tax payers!) are interested in data collection systems with high efficiency. If the sampling process is repeated periodically it shall be called a monitoring system or equivalently a surveillance system (Toma, 1999). In this paper the focus is on monitoring of a binary variable e.g. the health status, antibiotics resistance etc. The proportion of affected subjects is measured with diagnostic tests.

The following aspects are relevant for an efficient design of a monitoring system:

- The sample size should be as small as possible to keep sampling costs low.
- All available prior knowledge shall be included to increase the efficiency of the system.
- The bias due to imperfect diagnostic tests has to be considered.

The paper starts with a short summary on Bayesian parameter estimation and credibility intervals. In the next section a decision theoretic approach is discussed, which is followed by a description of interval-based Bayesian approaches for sample size determination. In a last step the characteristics of imperfect diagnostic tests (sensitivity, specificity) are introduced and built in the complete model. Finally some results for different prior settings are presented. For numerical calculations and graphics SAS 9.1.3 has been used.

2 Bayesian Credibility Intervals

We assume a population model where the target variable has two outcomes and is homogeneously distributed

$$X_i \overset{iid}{\sim} \text{Bernoulli}(\theta)\,.$$

The parameter of interest θ is estimated from an experiment of sample size n with x positive outcomes. Thus, for large populations x follows a binomial distribution

$$X \sim \text{Binomial}(n,\theta)\,, \quad \text{i.e. } P(X=x) = \binom{n}{x}\theta^x(1-\theta)^{(n-x)}\,.$$

If the sample is the only source of information the ML-estimator of θ is $\hat{\theta} = x/n$ (Greene, 2000). In classical statistics confidence intervals for this estimator can be constructed. From a Bayesian point of view one tries to reach (at least) two aims. First, the uncertainty about the parameter θ shall be quantified and, secondly, prior knowledge about θ has to be considered (Gelman et al., 2004, p.11). The uncertainty is quantified by a distribution function. This formal description does not necessarily imply that the parameter itself has a variability but there is uncertainty in our knowledge about it. If we use the available information *before* sampling (e.g. former investigations, expert opinions) we can define a prior (density) function $\pi(\theta)$.

The prior knowledge can be quite poor. For instance, the only information could be that it is a binomial parameter within the bounds $[0,1]$. The sampling results enter in the form of the function $f(x|\theta)$ which is mathematically identical to the likelihood function $l(\theta|x)$. The prior function is amalgamated with the likelihood in a kind of updating mechanism. The resulting posterior function can be written in the well-known form (Gill, 2002, p.66)

$$f(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{m(x)}\,.$$

The function $m(x)$ has the role of an integrating constant and is given by

$$m(x) = \int \pi(\theta)f(x|\theta)d\theta\,.$$

It can be used to calculate probabilities of sampling results based on the prior function and is therefore also called the *prior predictive function* (Gill, 2002, p.66). The Bayesian pendant for the confidence interval is the so-called credibility interval (CI) (Robert, 2001, p.260). It has a length d and a certain probability mass, also called coverage cov. There are several methods for constructing credibility intervals (Gelman et al., 2004, p.37). One way is to define so-called highest posterior density (HPD) intervals, which can easily be obtained by intersecting a line parallel to the x-axis with the density function but are not easy to compute. In case of the binomial parameter θ, which will henceforth be called p, the CI has the bounds $[p_u, p_o]$. This leads to the following definitions, based on a posterior function[1],

$$d = p_o - p_u\,, \qquad 0 \le d \le 1$$

[1]For probability density or distribution functions (PDF) a lower case letter (e.g. f) will be used, for the cumulative distribution function (CDF) the upper case letter (e.g. F).

and

$$cov = \int_{p_u}^{p_o} f(p|x)dp = F_{post}(p_o) - F_{post}(p_u)\,, \qquad 0 \leq cov \leq 1\,.$$

Coverage cov and length d have a positive interrelation. For a given prior or posterior function higher coverage implies a wider interval i.e. higher d. In Figure 1 this is illustrated by four cases with different prior functions and sample results:

- case 1: prior Beta(1,1), sample ($n = 5$, $x = 2$),
- case 2: prior Beta(1,1), sample ($n = 50$, $x = 20$),
- case 3: prior Beta(2,10), sample ($n = 5$, $x = 2$),
- case 4: prior Beta(2,10), sample ($n = 50$, $x = 20$).

Case 1 has the least information, therefore its slope is lower than that of the other cases.

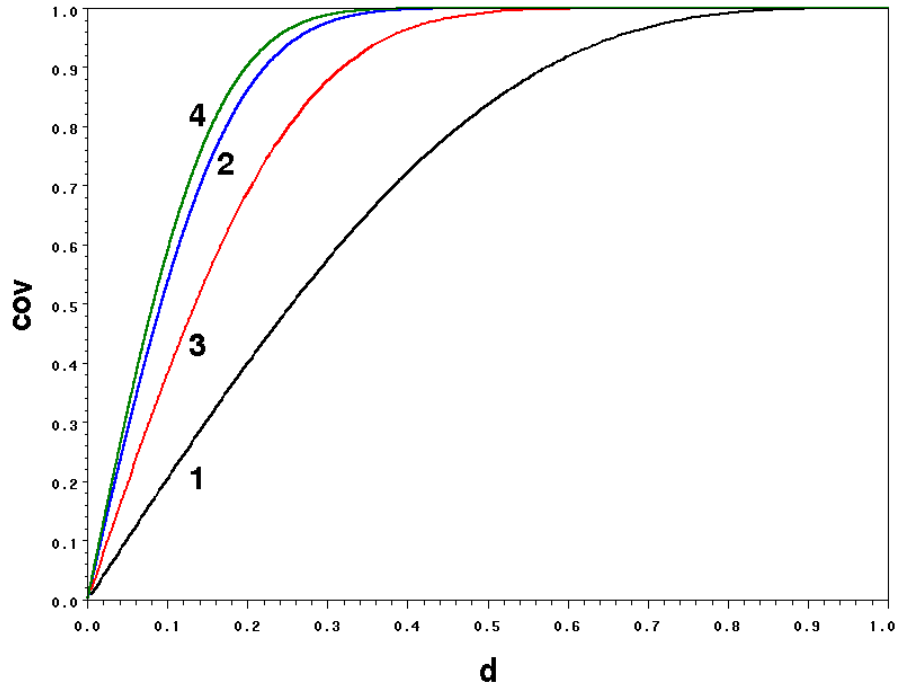

Figure 1: Coverage and length for different prior-sample-scenarios.

3 Loss Functions

As mentioned in the introduction, the results of a monitoring or surveillance system will have implications. High prevalence may afford expensive health programs, costs of additional sampling etc. This could be based on national laws or on recommendations of international organizations (e.g. WHO). In case of low prevalence these costs might be reduced. Due to imperfect information three different situations can occur (Marinell and Steckel-Berger, 2001, p.374):

Correct estimation: The true value p equals the estimated parameter $\hat{p}$, resp. lies within the interval $CI_p\ [p_u, p_o]$.

Underestimation: The true value p is *larger* than the estimated parameter $\hat{p}$, resp. *larger* than the upper bound p_o.

Overestimation: The true value p is *smaller* than the estimated parameter $\hat{p}$, resp. *smaller* than the lower bound p_u.

Both cases, over- and underestimation, may lead to wrong actions. Overestimation might require unnecessary additional programs (eradication programs, additional sampling costs). On the other hand underestimation may have negative long-term consequences if for instance infected animals or food is detected in other countries. This might lead to import bans, negative worldwide publicity and therefor negative economic impacts. There might as well be positive consequences, which can be considered by just looking at the net effects. Since the decisions are dependent on the realization of the parameter, loss-functions can be defined (Berger, 1985, p.8; Robert, 2001, p.52). The equations for the point estimate are (Marinell and Steckel-Berger, 2001, p.374)

$$s(\hat{p}, p) = \begin{cases} 0\,, & \text{if } \hat{p} = p\,, \\ s_u(\hat{p} - p)^r\,, & \text{if } \hat{p} > p\,, \\ s_o(p - \hat{p})^r\,, & \text{if } \hat{p} < p\,. \end{cases}$$

Depending on the exponent r the loss function is called constant ($r = 0$), linear ($r = 1$) or quadratic ($r = 2$) (French and Insua, 2000, p.150). If the loss coefficients s_u and s_o are not equal the loss function is asymmetric. In case of $s_u > s_o$ the possibility of overestimation is judged higher than that of underestimation. The loss functions for an interval estimate are quite similar (Marinell and Steckel-Berger, 2001, p.385)

$$s(p_u, p_o, p) = \begin{cases} 0\,, & \text{if } p_u \le p \le p_o\,, \\ s_u(p_u - p)^r\,, & \text{if } p_u > p\,, \\ s_o(p - p_o)^r\,, & \text{if } p_o < p\,. \end{cases}$$

In a decision theoretic approach one tries to find optimality conditions to minimize the *expected loss* (Raiffa and Schlaifer, 1961). These so-called *Bayes rules* for the optimal point estimate are formulated in the following way (for a detailed description see Marinell and Steckel-Berger, 2001, p.375ff)

$$\begin{aligned} \text{constant loss function:} &\quad s_u f_p(\hat{p} - \varepsilon) = s_o f_p(\hat{p} + \varepsilon)\,, \quad \text{where } 0 < \varepsilon < 1/2 \\ \text{linear loss function:} &\quad s_u F_p(\hat{p}) = s_o(1 - F_p)(\hat{p}) \\ \text{quadratic loss function:} &\quad s_u L_0^{\hat{p}}(p) = s_o L_{\hat{p}}^1(p)\,. \end{aligned}$$

$L_0^{\hat{p}}(p)$ and $L_{\hat{p}}^1(p)$ are the lower and upper linear partial moments. For a continuous random variable X partial moments are defined as (Marinell and Steckel-Berger, 2001, S. 254)

$$\begin{aligned} \text{lower (left) partial moment:} &\qquad m_r^u(c) = \int_{-\infty}^{c} (c - x)^r f(x) dx\,, \\ \text{upper (right) partial moment:} &\qquad m_r^o(c) = \int_{c}^{\infty} (x - c)^r f(x) dx\,. \end{aligned}$$

For a linear partial moment $r = 1$. The equivalent conditions for the interval estimate are

$$\begin{aligned} \text{constant loss function:} \quad & f(p_u) = (s_o/s_u) f(p_o)\,, \\ \text{linear loss function:} \quad & F(p_u) = (s_o/s_u)(1 - F(p_o))\,, \\ \text{quadratic loss function:} \quad & L^u_{p_u} = (s_o/s_u) L^o_{p_o}\,. \end{aligned}$$

Figure 2 illustrates how to find the loss optimal interval in case of a linear loss function. The starting point is the distribution function $F(p_u)$ that is shifted by the specified length d along the p-axis. The intersection of this new curve $F'(p_u)$ with the curve $1 - F(p_o)$ gives the optimal interval in case of a symmetric loss function (case 1, $s_o/s_u = 1$). If an asymmetric loss function is applied, the optimal interval is shifted upward because $F'(p_u)$ has to be intersected with $s_o/s_u(1 - F(p_o)$ (case 2). If $s_u > s_o$ there would of course be a downward shift. In case 1 the interval [0.475, 0.525] has a coverage of 16.6%, in the asymmetric case 2 the interval [0.519, 0.569] has a coverage of 15.6%. The effect on the change of coverage strongly depends on the skewness of the basic distribution function. With right-skewed functions an increase of the loss ratio s_o/s_u reduces the coverage.

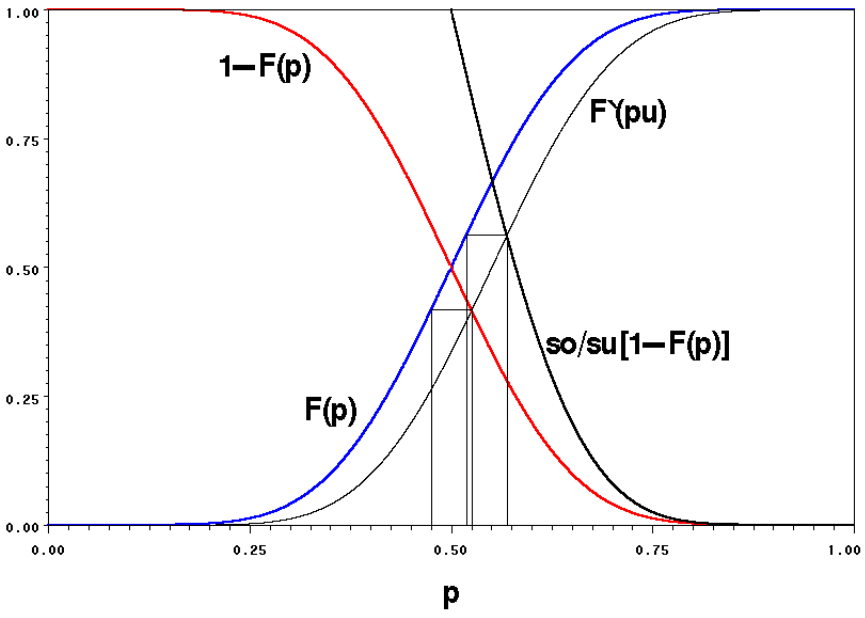

Figure 2: Interval estimate for linear loss function ($d = 0.05$, $s_o/s_u = 2$).

4 Sample Size Determination

4.1 Interval-based Approaches

Note, that it is also possible to specify a certain coverage and then look at the resulting interval length. For a given coverage the length d can only be reduced by increasing the sample size. For the problem of the optimal sample size different solutions within the

Bayesian framework have been developed. One way is to evaluate the additional cost or utility of a sampling unit (expected value of sample information, see Marinell and Steckel-Berger, 2001; Raiffa and Schlaifer, 1961). The practical disadvantage of this approach is that it requires a cost- or utility-function that can be compared with the losses s_o and s_u. Another method is more in the classical direction of calculating sample sizes for specified confidence levels and tolerances, the so-called *interval-based approaches* (Joseph and Wolfson, 1997). Their common basis is the cov-d-relationship described in Figure 1. This relationship is defined for a prior function and a certain sample outcome $(n; x)$. For a sample of size n there are $n + 1$ possible results for x, thus $n + 1$ possible cov-d-curves. Joseph et al. (1995) suggested the *average coverage criteria (ACC)* and the *average length criteria (ALC)*. Both are based on marginal probabilities of the sample results that can be derived from the prior predictive function

$$\text{ACC:} \qquad \overline{cov}(n) = \sum_{x=0}^{n} cov(x)m(x)\,,$$

$$\text{ALC:} \qquad \overline{d}(n) = \sum_{x=0}^{n} d(x)m(x)\,.$$

An example is given in Table 1. To determine the optimal sample size based on one of the criteria it is necessary to calculate $\overline{cov}$ and $\overline{d}$ for several values of n, which can be depicted by ACC- and ALC-curves.

Table 1: Calculation of ACC and ALC ($n = 5$, Prior: Beta(2,10))

x	0	1	2	3	4	5	
cov $(d = 0.1)$	0.485	0.437	0.385	0.353	0.333	0.322	$\overline{cov} = 0.446$
d $(cov = 0.95)$	0.264	0.326	0.374	0.407	0.433	0.447	$\overline{d} = 0.310$
$m(x)$ (%)	45.99	32.66	15.06	5.02	1.14	0.14	

4.2 Integrative approach

The concept of asymmetric loss functions and interval-based sample size determination are now integrated. Based on a prior function and a loss function (loss relation, type) a coverage resp. the length d has to be specified. Then for a certain sample size n the distinct optimal intervals are numerically determined and finally the average coverage (average length) is calculated. This has to be repeated for several sample sizes. For the sake of illustration ACC- and ALC-curves for various loss relations (0.2, 1, 2, 5) are shown in Figure 3 and 4. Table 2 provides the threshold values of n, where an accepted average coverage of 0.95 (average length of 0.1) for different loss relations s_o/s_u is reached.

4.3 Diagnostic Errors

Measurements are done by diagnostic tests, which are characterized by their sensitivity (SE) and specificity (SP). Sensitivity is the probability of getting a positive result for a true

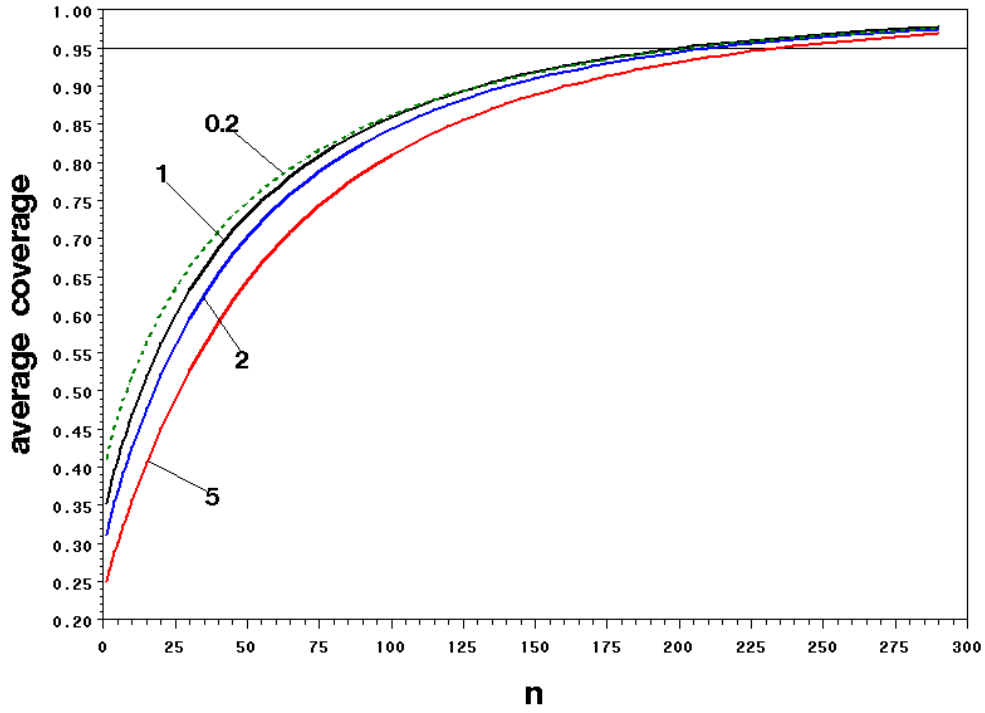

Figure 3: ACC-curve for various loss relations (quadratic, $d = 0.1$, prior: Beta(2, 10)).

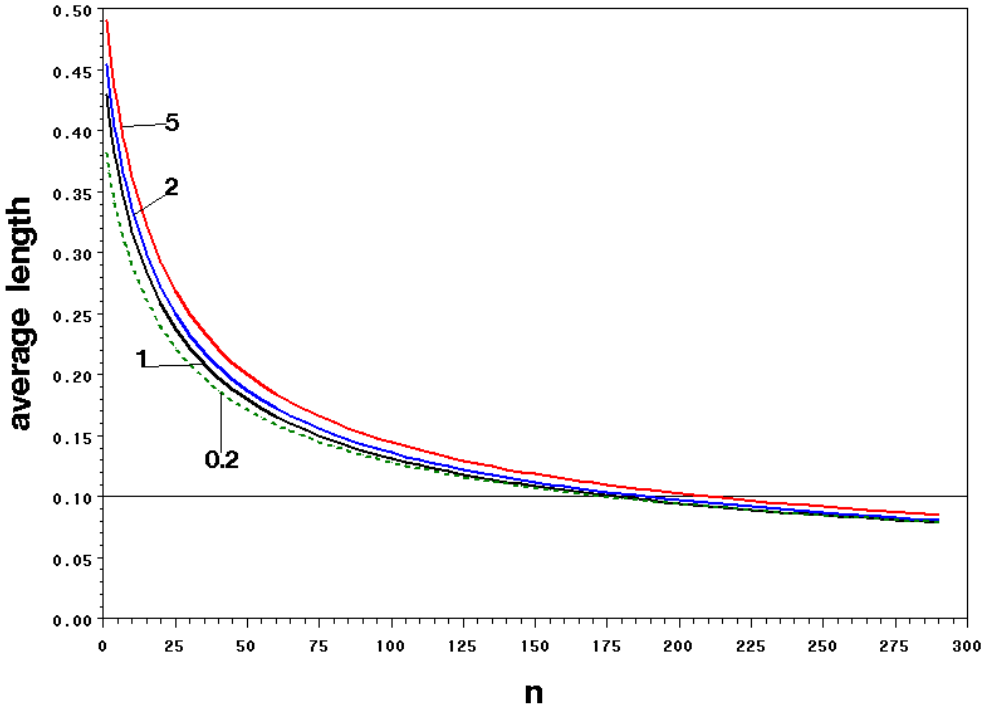

Figure 4: ALC-curve for various loss relations (quadratic, $cov = 0.95$, prior: Beta(2, 10)).

positive subject, specificity the probability of getting a negative result for a true negative subject. Hence the probability p_f of getting test-positive outcomes as a biased measure

Table 2: Optimal sample sizes.

s_o/s_u	0.2	1	2	5
$cov = 0.95 \Longrightarrow \overline{d} = 0.1$	180	178	181	204
$d = 0.1 \Longrightarrow \overline{cov} = 0.95$	205	201	211	235

for the true prevalence p is

$$p_f = p\,\mathrm{SE} + (1-p)(1-\mathrm{SP})\,.$$

This transformation has a deep impact on the distribution function of p. It can be shown that the transformed density function $f(p_f)$ has a lower variance than the unbiased function $f(p)$ (Stüger, 2004). The biased sample results give a picture of false certainty. Therefore the equation for the posterior function has to be corrected in the following way (Rahme et al., 2000)

$$f(p|x, \mathrm{SE}, \mathrm{SP}) \propto f(p)[p\,\mathrm{SE} + (1-p)(1-\mathrm{SP})]^x[(p(1-\mathrm{SE}) + (1-p)\mathrm{SP}]^{(n-x)}\,.$$

It is natural within the Bayesian framework also to include uncertainty about the test characteristics themselves by prior functions for SE and SP. The next equation shows the joint posterior function for p, SE and SP

$$f(p|x, \mathrm{SE}, \mathrm{SP}) \propto f(p)f(\mathrm{SE})f(\mathrm{SP})[p\,\mathrm{SE}+(1-p)(1-\mathrm{SP})]^x[(p(1-\mathrm{SE})+(1-p)\mathrm{SP}]^{(n-x)}\,.$$

To obtain the posterior marginal distribution of p the test characteristics SE and SP have to be integrated out

$$f(p|x) = \int\int f(p, \mathrm{SE}, \mathrm{SP}|x)d\mathrm{SE}d\mathrm{SP}\,.$$

This allows to correct the posterior function for the bias of misclassification and then to proceed in the aforementioned way. Figure 5 shows the effect of specificity. Decreasing values of SP (here from 1 to 0.8) afford a higher sample size to get same amount of credibility.

Figure 6 illustrates the effect of incorporating uncertainty about SE and SP. In case (2) uniform prior distributions for both parameters have been used. In case (3) PERT-functions have been applied, which are quite usual in the context of Bayesian veterinary epidemiology. These are modified Beta-functions with a given minimum, mode and maximum (Audige and Beckett, 1999). The picture illustrates the dramatic effect of additional uncertainty in prior knowledge, which affords quite a lot more of sample information.

5 Final Remarks

We showed how prior knowledge about prevalence but also characteristics of diagnostic tests are used for sample size determination in sample based monitoring systems in conjunction with asymmetric loss functions. The combination of loss functions and Bayesian interval-based approaches allows to include the relative evaluation of negative effects in case of over- or underestimation. The approach described can also account for the misclassification due to an imperfect diagnostic test which leads to increased sample sizes.

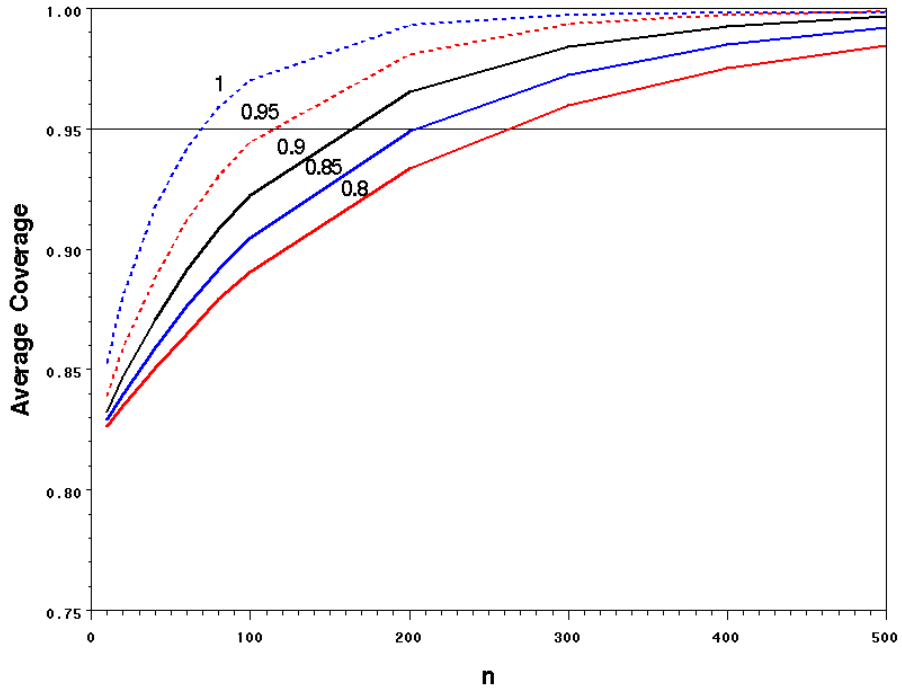

Figure 5: ACC-curves for various constant values of SP; $d_0 = 0.1$, prior: Beta(5, 50), quadratic loss ($s_o/s_u = 2$), SE $= 1$.

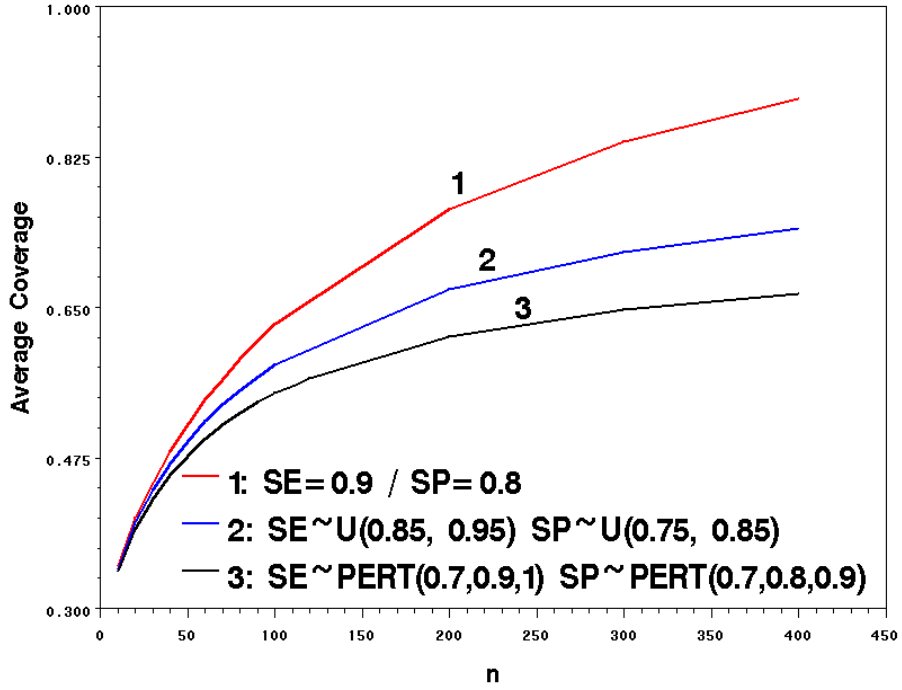

Figure 6: ACC-curves with test uncertainties; prior: Beta(2, 10), $d_0 : 0.1$, constant loss, ($sr = 2$).

References

Audige, L., and Beckett, S. (1999). A quantitative assessment of the validity of animal-health surveys using stochastic modelling. *Preventive Veterinary Medicine*, *38*, 259-276.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2. ed.). New York: Springer.

French, S., and Insua, D. R. (2000). *Statistical Decision Theory*. London: Arnold.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis* (2. ed.). Boca Raton: Chapman&Hall.

Gill, J. (2002). *Bayesian Methods. A Social and Behavioral Sciences Approach*. Boca Raton: Chapman&Hall.

Greene, W. H. (2000). *Econometric Analysis* (4. ed.). London: Prentice-Hall.

Joseph, L., and Wolfson, D. B. (1997). Interval-based versus decision theoretic for the choice of sample size. *The Statistician*, *46*(2), 145-149.

Joseph, L., Wolfson, D. B., and Berger, R. D. (1995). Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician*, *44*(2), 143-154.

Marinell, G., and Steckel-Berger, G. (2001). *Einführung in die Bayes-Statistik: Optimaler Stichprobenumfang* (3. ed.). München: R. Oldenbourg Verlag.

Rahme, E., Joseph, L., and Gyorkos, T. W. (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics*, *49*, 119-128.

Raiffa, H., and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University Graduate School of Business Administration.

Robert, C. P. (2001). *The Bayesian Choice: from decision-theoretic foundations to computational implementation* (2. ed.). New York: Springer.

Stüger, H. P. (2004). *Bayes-Methoden für Monitoringsysteme*. Unpublished doctoral dissertation, University of Graz.

Toma, B. (1999). *Dictionary of Veterinary Epidemiology*. Iowa State University Press.

Author's address:

Hans Peter Stüger
Institute of Applied Statistics
JOANNEUM RESEARCH
Steyrergasse 25a
A-8010 Graz

E-mail: hans-peter.stueger@joanneum.at