Kredit und Kapital, 41. Jahrgang, Heft 2, Seiten 217–238
Abhandlungen

Assessing the Estimation Uncertainty of Default Probabilities

By Jochen Lawrenz, Innsbruck

I. Introduction

Credit risk management is a central task for commercial banks. Within a quantitative approach, the key variable in quantifying credit risk is the probability of default (*PD*), which one may assign to a specific obligor or to a certain rating category. The focus on the *PD* may be either as considering it to be an output variable or an input variable. The former is related to the extensive research on credit risk modelling, where highly sophisticated models are formulated to predict the *PD* of individual obligers or entire portfolios. The latter is more related to the allocation of economic and regulatory capital and risk-adjusted pricing. When treating the *PD* as an input, it may be model-based or estimated from actual default history. If it is estimated from historical data two natural questions arise: (i) How is the *PD* estimated?, and (ii) How good is this estimate?[1]

The issue of assessing the reliability of *PD* estimates has only recently received heightened attention in the academic literature.[2] Not at least due to the revised framework of the Basel Capital Accord (Basel II).[3] Under Basel II, banks may choose to apply an internal ratings based (IRB) approach to assess the regulatory capital requirement needed to cover their credit risk. Basically, the (foundation) IRB approach consists of a risk-weight function which is defined as a function of the *PD*. However, little is said about how to estimate this *PD* and how to deal with estimation uncertainty.[4]

---

[1] Obviously, the issue of estimation precision is also present in the model-based *PD*, since any model requires some input data, but we will not focus on this more involved problem here.

[2] See *Christensen* et al. (2004), *Hanson/Schuermann* (2006), *Stein* (2003), *Pluto/Tasche* (2005).

[3] *BCBS* (2006).

[4] *BSBS* (2006), Sec. 451 tells us that: "In general, estimates of PDs, LGDs, and EADs are likely to involve unpredictable errors. In order to avoid over-optimism,

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

This issue is especially relevant for banks with small portfolios, as it is the case for the majority of regionally active banks. Note, that exactly for those banks the standard approach, which relies on external ratings, is hardly applicable since their portfolio will only contain few if any obligers with a rating from a major rating agency. But even if we consider banks with a reasonable large portfolio, the distribution across rating categories is far from uniform, so that some categories – notably the low-risk categories – contain only a small number of data points. BBA et al. (2004) for example report, that among the seven largest UK banks, 48 % of corporate assets will suffer from insufficient default data to give a statistically significant estimate. For other asset classes, like sovereigns and banks, percentage figures are even worse, being 90 % and 62 % respectively.

In this paper, we provide an assessment of the *PD* estimation uncertainty for different credit portfolio structure and size. The usual way to quantify the "likely range of errors" is to calculate statistical confidence intervals, i. e. upper and lower limits, that cover the true unknown parameter with a certain given probability. The length of the confidence interval depends on the actual estimate, the chosen confidence probability and most notably on the sample size. And as intuition suggests, the interval will be shorter for larger samples. Therefore, for a given *PD* estimate of a certain rating category and given the number of observations in that category, one can show how reliable this estimate is in terms of calculating the length of the confidence interval. As the PD estimation uncertainty is especially important in the context of Basel II, we give an economic taste for the estimation uncertainty by calculating the regulatory capital requirement at the upper and lower limits.[5]

Furthermore, it is important to recognize that confidence intervals are not unique, i. e. there is not one single way of constructing interval estimators, but several alternatives are available. Thereby, the most common approach consists of inverting a test statistic. In particular, the so-called Wald interval is obtained by inverting a test statistic based on the normal distribution. Besides test statistic inversion, there is the Bayesian approach, which assumes a prior distribution for the parameter and then uses the observed data to obtain the updated (posterior) distribution. The

a bank must add to its estimates a margin of conservatism that is related to the likely range of errors." But what exactly "margin of conservatism" and "likely range of errors" means, is left open to interpretation.

[5] See also *Rösch* (2005) for the impact of estimation error on regulatory capital.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

Bayesian approach necessitates the choice of the prior distribution and thus reflects the experimenter's beliefs.[6] Within this approach, we discuss the so-called Jeffrey's interval.

Discussing the impact of interval choice is not only an theoretical issue, but is also of practical importance. For example, OeNB (2004), a publication by the Austrian central bank (OeNB), which gives an outline on the validation of rating systems recommends to use the Wald interval, unless the sample size is small, in which case the Clopper-Pearson interval should be used. Our contribution should help to clarify if this is an adequate recommendation.

Our contribution adds to the growing literature on *PD* estimation. Stein (2003) gives a simple but instructive example about the accuracy of probability estimates and especially poses the question of how large a data set needs to be in order to give reliable estimates. However, he largely relies on the normal approximation to the binomial distribution, which provides doubtful results as will be discussed later. Höse/Huschens (2003) and Huschens (2006) also use the asymptotic property of the binomial distribution but consider correlated default events in the form of the one-factor model underlying the Basel IRB approach. They show that confidence intervals in the correlated case are significantly wider. Pluto/Tasche (2005) address the issue of *PD* estimation if there is no or few default events in the data set. They suggest a most prudent estimation principle, which they interpret as estimating *PD*s by upper confidence bounds. A more qualitative approach of how to deal with low default portfolios is also discussed in BBA et al. (2004).

Focusing more on the issue of how to estimate *PD*s (or in general migration probabilities), Lando/Skodeberg (2002) and Jafry/Schuermann (2004) compare confidence intervals derived from the cohort against the duration method. While the cohort method counts the number of migrations at the end of the observation period (usually one year), the duration method uses continuously observed rating actions and thus also takes into account if a firm that was initially in rating category i was temporarily in category j before ending up in category k at the end of the year. With the duration method, Lando/Skodeberg (2002) show that one obtains non-zero migration probabilities for cases where actually no such migration has been observed in the data sample. This is reasonable, since only because the default of a high-rated company has not been observed

[6] See e.g. *Casella/Berger* (2002), p. 437.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

does not mean that it has probability zero that this event can occur.[7] More importantly, Lando/Skodeberg (2002) and Jafry/Schuermann (2004) show that confidence intervals for duration estimates are usually tighter than those for the cohort estimate, since the former contain more information. However for the duration estimate, no analytical confidence intervals are known and can only be obtained via bootstrapping methods. Hanson/Schuermann (2006) provide a systematic comparison of confidence intervals across different estimation methods and interval. Interestingly, with a rather large sample, they show that interval lengths can be substantial, making it impossible to distinguish statistically between notch-level *PD*s in the investment grade categories. Similar results were derived by Christensen et al. (2004) although their focus is on addressing the issue of rating momentum, i.e. the observation that a recently downgraded firm has a higher risk of being further downgraded than firms being in the same rating category for a longer period of time.

Besides the financial application, the construction of appropriate confidence intervals for a binomial proportion is the topic of recent research in statistical science. Brown et al. (2001, 2002) or Agresti/Coull (1998) have shown that the standard confidence interval based upon the asymptotic property of the binomial distribution performs poor not only for small sample size.

The remainder of the article is organized as follows: Section II discusses the theoretical foundations of confidence intervals. Section III applies four alternative intervals to exemplary representative credit portfolios, and section IV concludes.

II. Theoretical Underpinnings of Confidence Intervals

The industry standard for the estimation of the *PD* is the cohort method. As mentioned in the previous section, the cohort method consists of counting the number of firms that were in rating category i at the beginning of the year and ended up in default (or any non-default category) at the end of the year relative to the total number of firms initially in category i. Denoting in the following the *PD* of a firm in rating category i

[7] Actually, this is easily seen by multiplying a one-year migration matrix with itself, giving the two-year matrix. If the one-year matrix contains a zero *PD* for the highest grade, but a non-zero probability of migrating into a lower category which itself has a positive *PD*, then the two-year matrix will show up a non-zero *PD* for the highest grade also.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

by p^i, and the default frequency calculated from a sample size of n by f_n^i,[8] then according to the cohort method, we have

$$f_n^i = \frac{n_{i,def}}{n_i}, \tag{1}$$

where $n_{i,def}$ is the number of firms migrating from i to default, and n_i is the total number of firms initially in category i.

Usually, the default frequency f is taken to be the *PD* estimate, i. e.:

$$\hat{p}^i = f^i.$$

This is also consistent with the requirements in BCBS (2006): "PD estimates must be a long-run average of one-year default rates for borrowers in the grade [...]".[9] One may come up with different methods to provide *PD* estimates such as the duration method previously mentioned, but since our aim is to provide an assessment of the *PD* estimate uncertainty in the context of standard methods used in practice and for typical portfolio structure and size, we focus on the cohort method.

Once the *PD* estimate in terms of the default frequency is given, we can ask how good this estimate is given the data underlying the estimation, i. e. we can ask for the "likely range of errors" for this estimate in terms of calculating a confidence interval. In general, the construction of a confidence interval consists of finding lower and upper bounds L and U, such that the probability that the interval (L, U) covers the unknown true parameter, i. e. $L < p < U$, or $p \in (L, U)$ equals some predefined confidence level. Denoting the confidence level by α, we seek to determine U and L, such that:

$$Prob\{L < p < U\} = 1 - \alpha, \tag{2}$$

where α is frequently taken to be 5 %. It is important to note that in this formulation the random quantity is the interval and not the parameter.[10] Obviously, L and U will depend on the estimate $\hat{p}$, the sample size n, the confidence level α and distributional assumptions.

[8] If reference to the rating category or sample size is not necessary, we simply drop the sub- and superscripts.

[9] *BCBS* (2006), Sec. 447.

[10] More formally, we can interpret the confidence interval as a set of parameter values for which the hypothesis that it contains the true parameter cannot be rejected at a given confidence level.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

In the present case, since default either occurs or does not occur, we know that the default frequency follows a binomial distribution with some (unknown) parameter p. The first approach to come up with a confidence interval, which is also the standard approach in most textbooks,[11] is to make use of the fact that the binomial distribution can be asymptotically approximated by the normal distribution. The Central-Limit Theorem tells us, that as $n \to \infty$:

$$\frac{f_n - p}{\sqrt{Var(f_n)}} \sim \mathcal{N}(0,1). \tag{3}$$

Therefore, as a corollary,[12] we have:

$$\lim_{n \to \infty} Prob\left\{-z < \frac{f_n - p}{\sigma(f_n)} < z\right\} = \int_{-z}^{z} \phi(s)\, ds \tag{4}$$

$$= \Phi(z) - \Phi(-z), \tag{5}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution function of the standard normal distribution respectively. Because of symmetry of the normal distribution, we have $\Phi(z) - \Phi(-z) = 2\Phi(z) - 1$.

If we want the left hand side of (4) to be equal to $1 - \alpha$ as in (2), we find z to be $z = \Phi^{-1}(1 - \alpha/2)$, i.e. the $(1 - \alpha/2)$-quantile of the standard normal distribution. Therefore, for the remainder we formally define: $z_\alpha \equiv \Phi^{-1}(1 - \alpha/2)$.

Plugging in z_α in (4) and rearranging terms gives

$$\lim_{n \to \infty} Prob\{f_n - z_\alpha \sigma(f_n) < p < f_n + z_\alpha \sigma(f_n)\} = 1 - \alpha.$$

The estimated standard deviation (standard error) of the estimator f_n is easily found to be

$$\hat{\sigma}(f_n) = \sqrt{\frac{f_n(1 - f_n)}{n}}.$$

Considering that $\hat{\sigma}(f_n)$ is a consistent estimator for the standard deviation $\sigma(f_n) = \sqrt{\frac{p(1-p)}{n}}$, we end up with the following interval, for which we can be sure to cover the true parameter with an asymptotical confidence level of $1 - \alpha$:

[11] See e.g. *Mood* et al. (1974) or *Davison* (2003).

[12] Known as the de Moivre-Laplace Theorem. See e.g. *Capinski/Kopp* (1999).

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

$$f_n \pm z_\alpha \sqrt{\frac{f_n(1-f_n)}{n}}. \tag{6}$$

This is called the *Wald confidence interval*. For further use, we abbreviate the lower and upper bound of the Wald interval as L_W and U_W.

It is important to note, that the above reasoning guarantees that the Wald interval covers the true parameter with probability $1-\alpha$ only asymptotically, i.e. as $n \to \infty$. The actual coverage probability, $Cpr = Prob\{p \in (L_W, U_W)\}$, may deviate significantly from its nominal value if n is small. This is why most textbooks give the recommendation to use the Wald interval only if np is larger than 5, or similar conditions. But in fact, Brown et al. (2001, 2002) show that the actual coverage probability of the standard Wald interval may not only be poor for small n, but can still deviate substantially from its nominal value for large n and for p near 0 and 1.

Figure 1 plots the value of the actual coverage probability *Cpr*. In the left panel, *Cpr* is shown for increasing n and a fixed $p = 0.005$. The right panel holds fixed a constant $n = 100$ and shows *Cpr* for p from 0 to 1. The dashed line indicates the nominal confidence level, which is $\alpha = 0.05$. It is easily recognized, that the actual coverage probability can be quite significantly lower than its nominal value. In particular for small n and p near 0, *Cpr* may be only around 75 %. The behavior of the coverage probability is quite erratic, and still for a sample as large as $n = 1018$, where the above-mentioned rule of thumb is satisfied, we can calculate *Cpr* to be only 88.94 %.

To overcome the poor performance of the Wald interval, in particular for small sample size, several alternative formulations for confidence

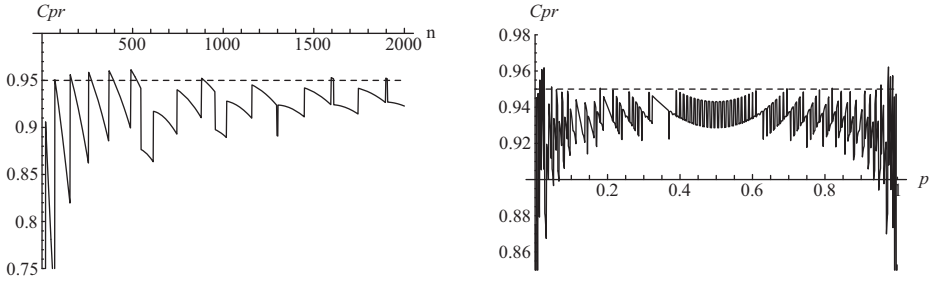


Figure 1: Coverage Probability Cpr of the Wald Interval

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

intervals have been put forward. A natural alternative is to avoid using the asymptotic property and construct a confidence interval from the finite sample binomial distribution. The idea is to find the endpoints of the interval by the following reasoning. The upper limit U is the highest value of p_0 such that the probability of observing not more than n_{def} defaults is just $\alpha/2$, i.e. it is the solution in p_0 of the following equation:

$$\sum_{x=0}^{n_{def}} \binom{n}{x} p_0^x (1-p_0)^{n-x} = \alpha/2. \tag{7}$$

Correspondingly, the lower endpoint of the interval L is found as the smallest solution in p_0, such that the probability of observing more than n_{def} is $\alpha/2$:

$$\sum_{x=n_{def}}^{n} \binom{n}{x} p_0^x (1-p_0)^{n-x} = \alpha/2. \tag{8}$$

The interval obtained in this way is known as the *Clopper-Pearson "exact" confidence interval,* and we will denote its endpoints by L_{CP} and U_{CP}.

Although the Clopper-Pearson interval is called "exact", its coverage probability is known to be too conservative. The construction of the interval guarantees that the coverage probability is *at least* the nominal value. However, for small n and p near 0 and 1 the actual *Cpr* is substantially higher, approaching nearly 100 %.[13]

Therefore, Agresti/Coull (1998) argue that "approximate is better than 'exact'", and propose a confidence interval that relies on the same form as the standard Wald interval, i.e. based upon approximation, but uses a different mid point of the interval. Recall, that the default frequency is $\hat{p} = f_n = \frac{n_{def}}{n}$. Now, define

$$\tilde{n}_{def} = n_{def} + \frac{z_\alpha^2}{2}, \qquad \tilde{n} = n + z_\alpha^2, \qquad \tilde{f}_n = \frac{\tilde{n}_{def}}{\tilde{n}}.$$

For the case that $\alpha = 0.05$, $z_\alpha = 1.96$ which is approximately 2. This means, that we add 4 observations to our sample, out of which 2 are de-

[13] *Brown* et al. (2001) therefore call the Clopper-Peason interval "wastefully conservative" (p. 113). However, it may serve as an alternative (conservative) benchmark. See e.g. *Christensen* et al. (2004) and *Hanson/Schuermann* (2006). See also *Agresti/Coull* (1998) or *Davison* (2003), p. 346.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

faults. Agresti/Coull (1998) summarize their suggestion in the succinct statement: "Add two successes and two failures and then use the Wald formula."[14]

The Agresti-Coull interval is therefore found by

$$\tilde{f}_n \pm z_\alpha \sqrt{\frac{\tilde{f}_n(1-\tilde{f}_n)}{\tilde{n}}}, \tag{9}$$

and we abbreviate the lower and upper endpoints by L_{AC} and U_{AC}.

Note, that the midpoint of this interval is a weighted average of f_n and $1/2$, as can be seen by simple transformations

$$\begin{aligned}\tilde{f}_n &= \frac{n_{def} + z_\alpha^2/2}{n + z_\alpha} = \frac{n_{def}}{n + z_\alpha} + \frac{z_\alpha^2/2}{n + z_\alpha} = \frac{n_{def}}{n}\frac{n}{n + z_\alpha} + \frac{z_\alpha^2/2}{n + z_\alpha} \\ &= f_n\left(\frac{n}{n + z_\alpha}\right) + \frac{1}{2}\left(\frac{z_\alpha^2}{n + z_\alpha}\right).\end{aligned}$$

This formulation can be justified by using standard errors of the null hypothesis value instead of its maximum likelihood estimator.[15]

While the Agresti-Coull interval significantly improves the performance of either the Wald and the Clopper-Pearson interval, it has still deficiencies when n is small. Brown et al. (2001) therefore propose to use Agresti-Coull only for $n > 40$, while they recommend the so called (equal-tailed) Jeffreys interval for small $n < 40$.

Unlike the Wald and Agresti-Coull interval which use maximum likelihood principles, the Jeffreys interval is obtained by applying inference based on Bayes' theorem. Bayesian inference requires to specify some prior density for the random variable. In the case of a binomial proportion, the standard (conjugate) prior[16] is the Beta distribution. For a binomial random variable Y that has a prior beta distribution with parameter a and b, $Beta(a, b)$, the posterior distribution is $Beta(Y + a, n - Y + b)$. In the present context, the (non-informative) prior distribution is $Beta(1/2, 1/2)$. The posterior distribution is accordingly $Beta(n_{def} + 1/2, n - n_{def} + 1/2)$.

[14] *Agresti/Coull* (1998), p. 122.
[15] For more on this, see *Agresti/Coull* (1998), p. 120.
[16] See *Davison* (2003) for the notion of "conjugate priors".

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

Thus, the endpoints of the (equal-tailed) Jeffreys interval are defined as

$$L_J = \mathcal{B}(\alpha/2;\, n_{def} + 1/2,\, n - n_{def} + 1/2) \tag{10}$$

$$U_J = \mathcal{B}(1 - \alpha/2;\, n_{def} + 1/2,\, n - n_{def} + 1/2), \tag{11}$$

where $\mathcal{B}(q; a, b)$ denotes the q-quantile of the Beta distribution with parameter a and b.

Note, that the Bayesian approach underlying the Jeffreys interval provides a slightly different interpretation about the confidence interval. While we took care to talk about the interval that *covers* the true unknown parameter with some probability in the previous three cases, we can say that p *lies inside* the interval with some probability in the (Bayesian) case of the Jeffreys interval. This is because p is not a parameter but a random variable.[17]

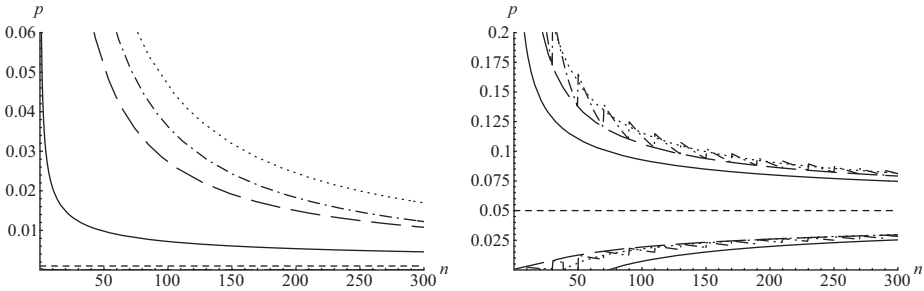
Other intervals are available, but we will confine ourselves to the four above-mentioned alternatives: (i) The standard Wald interval, which is known to have a low coverage probability, (ii) the "exact" Clopper-Pearson interval, which is too conservative, (iii) the Agresti-Coull interval which is recommended unless $n < 40$ and (iv) the Jeffreys interval, which provides adequate results also for smaller n.

Figure 2 shows the four intervals at work.

The left panel shows the upper bounds for increasing sample size n for an estimate of $\hat{p} = 0.001$, i.e. 0,1 % or 10 basispoints (bp) indicated as the horizontal dashed line at the bottom. The confidence level is chosen as 95 %, i.e. $\alpha = 0.05$. The solid line nearest to the estimate is the upper limit of the Wald interval (U_W), the dashed line is the Jeffreys interval (U_J), the dash-dotted line corresponds to the upper endpoint of the Clopper-Pearson interval (U_{CP}), and finally the dotted line is the Agresti-Coull upper limit (U_{AC}). While the lower limits for the Clopper-Pearson and the Jeffreys up to the sample size of 300 are defined to be 0, the lower endpoints of the Wald and Agresti-Coull interval are negative. Obviously, the *PD* cannot be negative, and the calculated negative values are a consequence of the normal approximation to the binomial proportion, which is especially inadequate in this case. Note, that the Clopper-Pearson and the Agresti-Coull interval are farthest from the estimate,

[17] For more details, see *Casella/Berger* (2002), p. 435 f.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/



Solid line: Wald interval; Dashed line: Jeffreys interval; Dot-dashed line: Clopper-Pearson interval; Dotted line: Agresti-Coull interval

Figure 2: Comparison of the Four Confidence Intervals

which is consistent with the fact that their coverage probability tend to be significantly higher than their nominal value.

The right panel shows the confidence intervals for an estimate of $\hat{p} = 0.05$, i. e. 5 %, again for $\alpha = 0.05$. In this case, we see both, the upper and lower limits. Interestingly, the upper endpoint of the Wald interval is nearest to the estimate, while the lower endpoint is farthest. The Clopper-Pearson interval, due to the discreteness of the binomial distribution displays jumps.

For both values of the estimate, however, the interval length can be substantial and differences between the intervals matter especially for small estimates. Table 1 reports numerical values for $\hat{p} = 10$ bp and $\hat{p} = 500$ bp for a sample size of $n = 100$ and $n = 300$.[18]

Table 1

Upper and Lower Endpoints of the Confidence Intervals

(in bp)	n	L_W	U_W	L_J	U_J	L_{CP}	U_{CP}	L_{AC}	U_{AC}
$\hat{p} = 10$	100	0*	71.9	0	273	0	362	0*	465
(0.001 %)	300	0*	45.7	0	107	0	122	0*	169
$\hat{p} = 500$	100	72.8	927	193	1,061	164	1,128	177	1,155
(0.05 %)	300	253	746	295	790	282	811	297	816

[18] Figures indicated with a star are set to zero (although their computed value is negative) since a binomial fraction cannot assume values smaller than zero.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

III. Confidence Intervals for Different Portfolios

The main focus of this article is to provide an assessment of the *PD* estimation uncertainty encountered in typical credit portfolios of different size and structure. Therefore, we will apply the results of the previous section to three different portfolios with three different sample sizes, corresponding to a small, medium and large portfolio.

For the first portfolio, we use the data reported in Hanson/Schuermann (2006) on the rating history of Standard & Poor's for U.S. obligers. Their sample consists of 50.611 firm-years of data, mainly from large corporations and extends over the period 1982 to 2002.

The second portfolio relies on data taken from the Austrian Central Bank (Österreichische Nationalbank, OeNB).[19] Although it is only a data example for the validation of rating systems, we will label it the OeNB portfolio.

The last portfolio pretends to represent a credit portfolio consisting of small and medium-sized enterprises (SME). The corresponding data is taken from Schwaiger (2002), who constructed a rating system of 12 classes based on 11,610 Austrian firms with a yearly turnover between € 1 and 50 Million.[20]

The three portfolios vary in the number of rating classes (7 for S&P, 10 for OeNB and 12 for SME) as well as in the distribution of obligers across rating categories. (See Table 2 and Figure 5 in the Appendix.)

We will consider the three portfolios in three different sizes. The small portfolio is assumed to consist of $n = 3.000$ firm-years of data, while the medium and the large portfolio contain $n = 10.000$ and $n = 25.000$ firm years respectively. If we presume that we have 3 years of rating history, this corresponds roughly to 1.000, 3.300, 8.300 obligers in the corresponding portfolios. By varying the portfolio size, we hold fixed the *PD*s in each rating category, as well as the distribution of obligers across categories, to reflect the different structure of the portfolios.

We apply the four confidence intervals introduced in the previous section to the resulting 9 cases. Figure 3 reports the results. Note, that for the best category in each portfolio there were no defaults and thus the cohort estimate is 0, and we don't calculate a confidence interval in this

[19] See *OeNB* (2004).
[20] For more information on the SME portfolio, see *Schwaiger* (2002).

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

case.[21] Note, that the vertical axes in figure 3 are in log-scale. For each rating category,[22] the four different confidence intervals are plotted, where the left solid bar is the Wald interval, the left dashed bar is the Jeffreys interval, the right solid bar is the Clopper-Pearson interval, and the right dashed bar is the Agresti-Coull interval.

The first observation from figure 3 is that the length of the confidence interval for "real-world" credit portfolios can be substantial. For example, considering the Jeffreys interval, the length for the A rating in the small, medium and large S&P portfolio is 44.5 bp, 20.4 bp and 12.2 bp respectively. This might sound innocent, but recall that the *PD* estimate for A is itself 6.2 bp, so if we express the interval length in terms of the *PD* estimate, we obtain 717 %, 329 % and 196 %.

The same is true for class 5 of the OeNB portfolio, where the Jeffreys interval has a length of 235.1 bp, 125.9 bp and 79.2 bp respectively, which is significantly higher in absolute terms, but translates in smaller relative length figures in terms of the *PD* estimate, which is 122.6 bp, of 191 %, 102 % and 64 % respectively.

Comparing the four different confidence intervals with respect to their lengths, we observe significant differences especially in the good rating categories. In the extreme case of the small S&P portfolio, the interval length for the AA category is 13.04 bp for the Wald interval, while it is 109.78 bp for the Agresti-Coull interval, which is more than eight times as large. Even for the large S&P portfolio the Agresti-Coull interval is still three times larger (5.49 bp versus 15.44 bp) than the Wald interval for category AA.

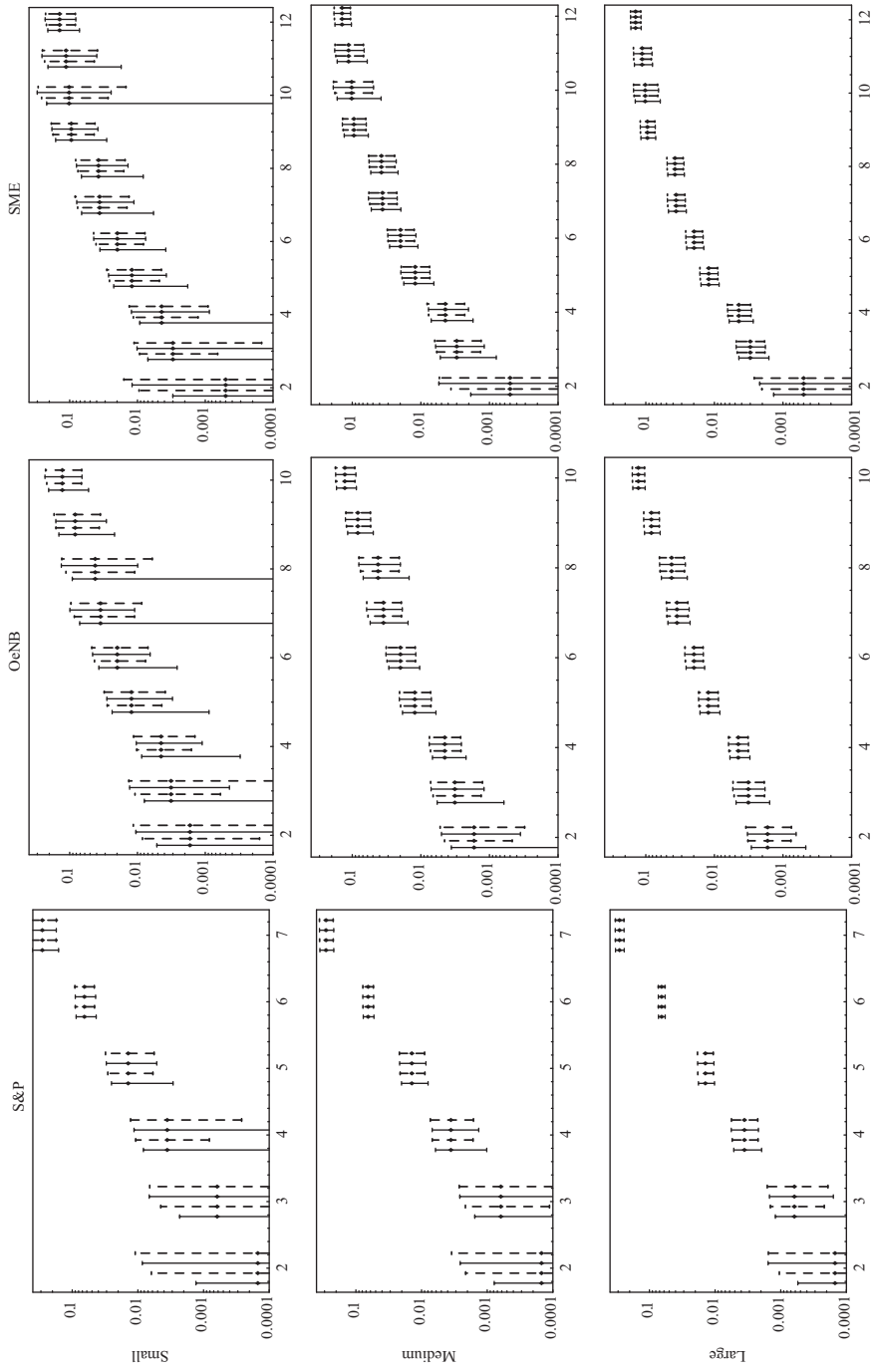
The second pattern we observe is that the upper limit of the Wald interval U_W is always lower than the upper limits of the other intervals, while the lower limit of the Wald interval L_W is usually the lowest one. This is most obviously observed in the small S&P and OeNB portfolios. This refers to the fact, that length alone is not the only criteria to evaluate the performance of a confidence interval.

The poor coverage probability of the Wald interval mentioned above is also partly due to the fact that it is "downward" biased. As we will argue

[21] For reasons of comparability, we do not report confidence intervals, although it would be possible to calculate interval estimates, as *Pluto/Tasche* (2005) have shown.

[22] For the S&P portfolio, 2 corresponds to AA, 3 corresponds to A, ... and 7 corresponds to CCC.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/



Vertical axis: *PD* values, log-scale; Horizontal axis: Rating categories.
For each rating category, four confidence intervals are drawn: Wald (left solid), Jeffreys (left dashed), Clopper-Pearson (right solid) and Agresti-Coull interval (right dashed).

Figure 3: Confidence Intervals for PD Estimates

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

in the next section, this bias has especially adverse implications in terms of capital requirements.

The final remarkable observation we can draw from figure 3 refers to the possibility of distinguishing between different rating categories. Consider the small S&P portfolio in the upper left panel. The *PD* estimate for category BBB is 35.7 bp. Except for the Wald interval, all upper limits of the confidence intervals for category A and AA are higher than this estimate. This means, the BBB category is not statistically significantly different from the A and AA category in the small portfolio case. For the medium and large portfolio, the BBB estimate no longer lies within the interval of the A category, but we still cannot differentiate significantly the AA category from the A category. This is in line with the results of Hanson/Schuermann (2006) who also find that notch-level rating categories are statistically indistinguishable in the investment grade ratings.[23] The same observation is found in the OeNB and SME portfolio, where in the small OeNB portfolio, we cannot distinguish category 4 from categories 2, 3 and 5, since the estimate lies within their confidence intervals. For the small SME portfolio, f. ex. categories 9 through 12 are indistinguishable from each other.

To provide an economic assessment of the uncertainty of the *PD* estimation, we now demonstrate how the above derived confidence intervals translate into risk-weights according to the Basel II IRB approach. This means, we take the *PD* value at the upper and lower limits of the various confidence intervals and calculate the corresponding risk-weight by applying the IRB risk-weight function as it is defined in BCBS (2006). In this way, we obtain a quasi confidence interval for the capital requirement the bank must hold.[24] As before, we apply this to nine representative portfolios. Figure 4 reports the results.

All nine panels show the deviation of the risk-weight at the upper and lower limit from the risk-weight of the point estimate, i. e. the ordinate shows the differences $RW(U_{\cdot}) - RW(\hat{\boldsymbol{PD}})$ and $RW(L_{\cdot}) - RW(\hat{\boldsymbol{PD}})$, where $RW(\cdot)$ denotes the IRB risk-weight function.[25] For better comparison, all

[23] *Hanson/Schuermann* (2006) consider an even greater data sample, and conclude that "ratings are indistinguishable *even with 22 years of data*." p. 16.

[24] Note, that we call this a *quasi* confidence interval. This is meant to stress, that in general it need not be true that for some function f: $(0,1) \to \mathbf{R}$, the interval $(f(L), f(U))$ covers $f(p)$ with the same probability as (L, U) covers p. However, if f is monotone, the two coincide. After all, the quasi confidence interval presented here is mainly meant to provide an economic flavor for the estimation uncertainty.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

nine panels show deviations in the same range of −70 to +80 risk-weight (percentage) points.

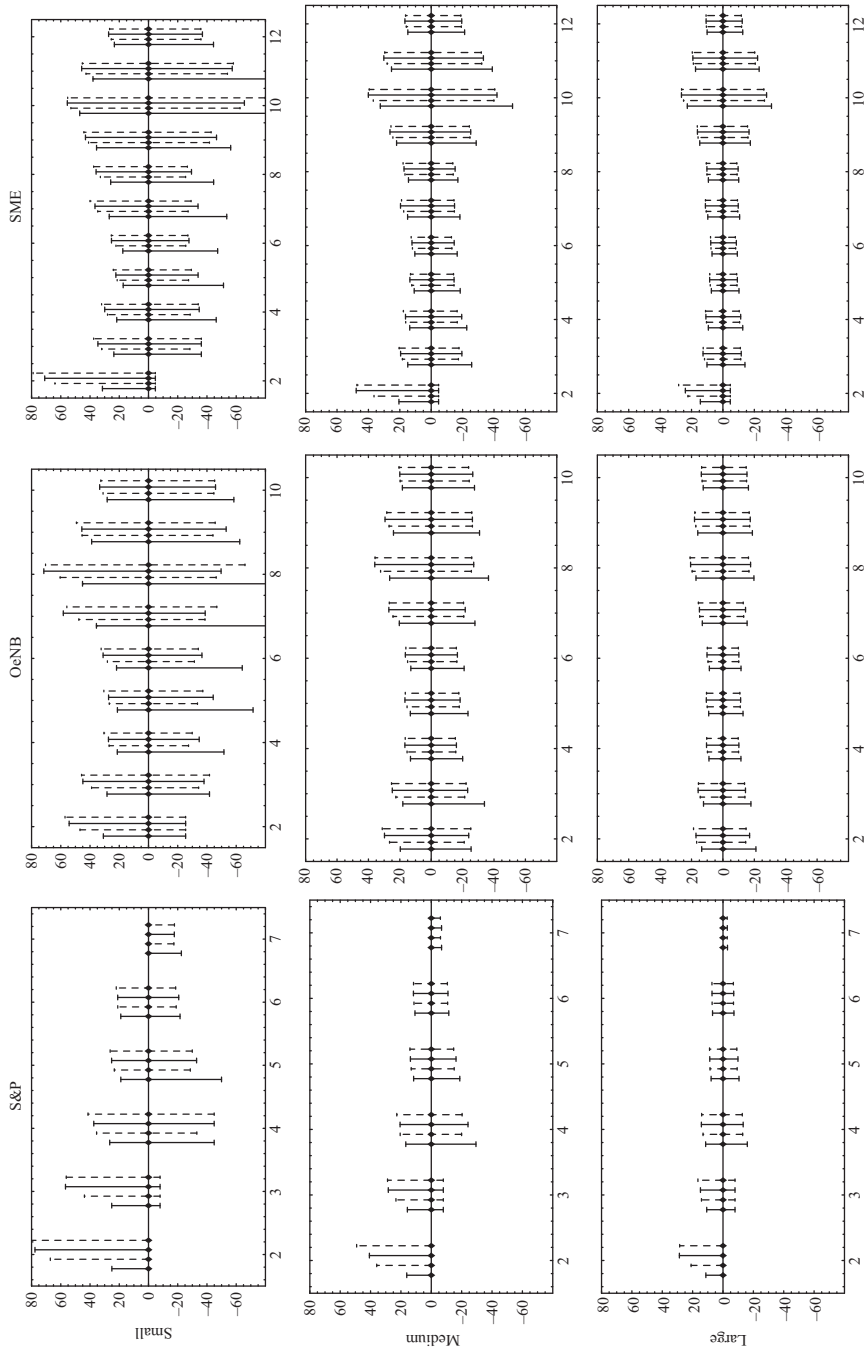
The first observation, we can draw from figure 4 refers to the absolute magnitude of the risk-weight confidence interval. Consider again, the single A rating category of the S&P portfolio. Note, that this category has the highest percentage share of all obligers. The point estimate is 6.2 bp, which translates into a IRB risk-weight of 22.35 %. The deviations in the small portfolio are substantial. The risk-weights calculated at the upper limit of the Clopper-Pearson and Agresti-Coull intervals are the highest, being 79.2 and 78.6 respectively, which is a deviation of +56.8 and +56.2. Thus, for an exposure of 1,000,000 of an A obligor, the risk-weighted asset at the upper limit of the confidence interval is roughly 790,000 instead of 223,500, which translates into a capital requirement of 63,200 instead of 17,880. This is roughly 3.5 times more! Risk-weights at the upper limit of the Jeffreys and Wald interval are lower, having a deviation of +66.2 and +47.5, giving a capital requirement which is 295 % and 212 % higher than that at the point estimate. For the negative deviations, there is a floor at –8, since BCBS (2006) requires that for corporate exposures, risk-weights can not be lower than 14.4, corresponding to a minimum *PD* of 3 bp. The impact of the floor is also responsible that there is no negative deviation for category AA, since the estimated *PD* of 1.49 bp is already below this minimum. The estimation error thus only works in one direction and the risk-weight confidence interval appears highly asymmetric for the high investment grades. The asymmetry is reversed for the lowest category, CCC. For this rating class, we only observe negative deviations. This seems puzzling, especially if we go back to figure 3 where we do have a symmetric confidence interval. The explanation is given in terms of the shape (and a curiosity) of the IRB risk-weight function. In general, the IRB risk-weight function is concave, thus deforming a symmetric *PD* interval to an asymmetric risk-weight interval where we have larger negative than positive deviations. The curiosity is, that the IRB function starts to *decline* for very high *PD* values,[26] giving lower risk-weights at both the lower *and* upper limit of the confidence interval.

[25] We use the foundation IRB-approach, which implies a given $LGD = 0.45$ (*BCBS* (2006), Sec. 287.) and a maturity adjustment with an effective maturity of 2.5 years. (*BCBS* (2006), Sec. 318.)

Note, that in the case of the SME portfolio we use the SME adjustment of the IRB function by assuming $S = 30$, representing an average SME firm.

[26] By analyzing the risk-weight function as it is defined in *BCBS* (2006), we find a maximum for $PD = 0.2962$.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/



Vertical axis: Deviations from the risk weight; Horizontal axis: Rating categories.

For each rating category, four confidence intervals are drawn: Wald (left solid), Jeffreys (left dashed), Clopper-Pearson (right solid) and Agresti-Coull interval (right dashed).

Figure 4: Confidence Intervals for Risk-weights

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

The second striking observation from figure 4 is the large difference between the four intervals for many cases. Consider category AA in the medium S&P portfolio, where we have 1,429 observations. The positive deviation with the Wald interval is +16.2, while the deviation according to the Agresti-Coull interval is +49.2 which is three times larger. The Jeffreys and Clopper-Pearson upper limits lie in between, being 251% and 221% larger.

The Wald interval not only displays the lowest positive risk-weight deviations but also the highest negative deviations, as can be seen most apparently in the small OeNB portfolio for rating categories 5 and 6. As mentioned previously, the Wald interval is generally downward biased, in the sense that the lower endpoint is lower than those of the other intervals. This feature is further aggravated through the concavity of the risk-weight function. Therefore, using the standard Wald interval in this context seems especially inappropriate.

IV. Conclusion

When using estimated probabilities of default for any risk management purposes and especially for the assessment of capital requirements, it is important to have some measure of the reliability of this estimate. In terms of BCBS (2006), a bank needs to evaluate the "likely range of errors" involved in the estimation. This can be done by calculating confidence intervals for the point estimate. A commonly used interval is the standard Wald interval. However, as we have shown in the first section of this article, there is an active theoretical debate in the statistics literature of how to appropriately construct a confidence interval for a binomial proportion. Thereby, there seems to be agreement that the Wald interval performs quite poorly, not only for small sample size. Alternative intervals have been put forward, out of which we have discussed, the "exact" Clopper-Pearson, the Agresti-Coull and the Jeffreys interval. Depending on the sample size, the upper and lower endpoints can be significantly different between these intervals.

To assess the economic significance of the estimation error, we have applied the four intervals to nine exemplary credit portfolios which differ in their structure and size. Although we cannot claim that these are "real-world" portfolios, we consider them as being representative and able to reproduce "real-world" conditions. The impact of the estimation error is shown as the confidence interval for the *PD* estimate on the one

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

hand, and in order to give a better economic intuition in terms of a quasi confidence interval for the risk-weights according to the Basel II IRB approach on the other hand.

Two main conclusions emerge from the analysis: (i) The absolute magnitude of the estimation error in terms of the length of the confidence interval can be large. (ii) Which interval is used matters and can make a significant difference.

The first point is mainly driven by the number of observations available for the portfolio, but due to the distribution of obligers across rating categories it is still relevant for high rating classes for rather large portfolios. This is reflected in the result, that to some extent one cannot statistically distinguish investment grade rating classes. The amount of estimation uncertainty we found in our analysis adds to the impression that the Basel II IRB approach insinuates an accuracy that is not justified, and adds to the finding of Bank/Lawrenz (2003), who have argued that there is model-inherent uncertainty in the IRB approach, which makes results doubtful. It has to be stressed at this point, that our results may even be regarded as a lower bound on the uncertainty, in that they rely on the independence assumption of obligers. For correlated default behavior, confidence intervals are even larger.[27]

While the first point seems to have found repercussion in literature and in practice, the second point is rarely addressed. As our results demonstrate, the widely used standard Wald interval is the worst choice especially for small portfolios and for good rating classes. The Clopper-Pearson interval, which is presented as exact alternative, has also drawbacks. Based on our analysis, we would recommend to use the Jeffreys interval.

To conclude, our results demonstrate that the estimation uncertainty is of a significant magnitude and should be addressed more carefully.

[27] See *Höse/Huschens* (2003) on this.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

Appendix: Characteristics of the Rating Systems

Table 2

Portfolio Characteristics (*f* in bp)

S&P portfolio

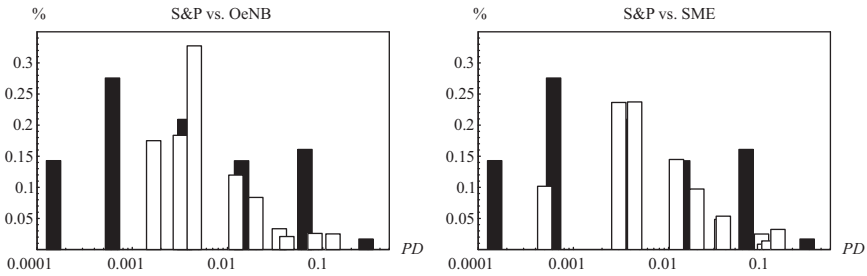
i	AAA (1)	AA (2)	A (3)	BBB (4)	BB (5)	B (6)	CCC (7)
n_i	2417	6690	12907	9794	6681	7533	792
$n_{i,def}$	0	1	8	35	94	491	226
f_n^i	0	1.5	6.2	35.7	141	652	2854

OeNB portfolio

i	1	2	3	4	5	6	7	8	9	10
n_i	50	1788	1876	3345	1223	856	342	214	265	257
$n_{i,def}$	0	3	6	15	15	17	12	9	22	33
f_n^i	0	17	32	45	123	199	351	421	830	1284

SME portfolio

i	1	2	3	4	5	6	7	8	9	10	11	12
n_i	5	1181	2745	2755	1681	1130	563	623	290	100	161	376
$n_{i,def}$	0	1	8	12	20	22	20	23	27	10	18	53
f_n^i	0	5	30	44	121	198	362	376	947	1020	1132	1409



Solid black bars represent the S&P portfolio. Open bars in the left panel represent the OeNB portfolio. Open bars in the right panel represent the SME portfolio.

The abscissa is in log-scale and represents *PD* values, the ordinate represents the percentage share of all obligers in the different rating categories.

Figure 5: Portfolio Structure

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

References

Agresti, A. and *Coull,* B. (1998): Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician, 52(2), 119–126. – *Bank,* M. and *Lawrenz,* J. (2003): Why simple, when it can be difficult? Some remarks on the Basel IRB approach. Kredit und Kapital, 36(4), 534–556. – *BBA, LIBA,* and *ISDA* (2004): The IRB approach for low default portfolios (LDPs) – Recommendations of the joint BBA, LIBA, ISDA Industry Working Group. working paper, Joint British Bankers' Association, London Investment Banking Association, International Swaps and Derivatives Association Industry Working Group. – *BCBS* (2006): International convergence of capital measurement and capital standards. A revised framework. Comprehensive version. Basel Committee on Banking Supervision. – *Brown,* L. D., *Cai,* T., and *DasGupta,* A. (2001): Interval estimation for a binomial proportion. Statistical Science, 16(2), 101–133. – *Brown,* L. D., *Cai,* T., and *DasGupta,* A. (2002): Confidence intervals for a binomial proportion and asymptotic expansions. The Annals of Statistics, 30(2), 160–201. – *Capinski,* M. and *Kopp,* E. (1999): Measure, Integral and Probability. Springer, London and Berlin. – *Casella,* G. and *Berger,* R. L. (2002): Statistical Inference. Duxbury, Pacific Grove, 2 edition. – *Christensen,* J. H., *Hansen,* E., and *Lando,* D. (2004): Confidence sets for continuous-time rating transition probabilities. Journal of Banking & Finance, 28, 2575–2602. – *Davison,* A. (2003): Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. – *Hanson,* S. and *Schuermann,* T. (2006): Confidence intervals for probabilities of default. Journal of Banking & Finance, 30(8), 2281–2301. – *Höse,* S. and *Huschens,* S. (2003): Sind interne Ratingsysteme im Rahmen von Basel II evaluierbar? Zur Schätzung von Ausfallwahrscheinlichkeiten durch Ausfallquoten. Zeitschrift für Betriebswirtschaft, 73(2), 139–168. – *Huschens,* S. (2006): Backtesting von Ausfallwahrscheinlichkeiten. In: Burkhardt, T., Knabe, A., Lohmann, K., and Walther, U., editors, Risikomanagement aus Bankenperspektive. Grundlagen, mathematische Konzepte und Anwendungsfehler. Berliner Wissenschafts-Verlag. – *Jafry,* Y. and *Schuermann,* T. (2004): Measurement, estimation and comparison of credit migration matrices. Journal of Banking & Finance, 28, 2603–2639. – *Lando,* D. and *Skodeberg,* T. M. (2002): Analyzing rating transitions and rating drift with continuous observations. Journal of Banking & Finance, 26, 423–444. – *Mood,* A. M., *Graybill,* F. A., and *Boes,* D. C. (1974): Introduction to the Theory of Statistics. McGraw-Hill series in probability and statistics. McGraw-Hill, 3 edition. – *OeNB* (2004): Ratingmodelle und -validierung. Oesterreichische Nationalbank. – *Pluto,* K. and *Tasche,* D. (2005): Estimating probabilities of default for low default portfolios. working paper, arXiv:cond-mat/0411699. – *Rösch,* D. (2005): Regulatory banking capital, estimation error, and systemic risk in ratings based capital rules. working paper, University of Regensburg. – *Schwaiger,* W. S. (2002): Auswirkungen von Basel II auf den österreichischen Mittelstand nach Branchen und Bundesländern. Bankarchiv, 50(06). – *Stein,* R. M. (2003): Are the probabilities right? Technical report, Moody's KMV.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

Summary

Assessing the Estimation Uncertainty of Default Probabilities

The probability of default (*PD*) is one of the key variables in credit risk management. By using *PD* estimates as input to pricing and capital requirement calculations, one should be concerned of how good these estimates are. Confidence intervals are thereby a convenient way to assess the range that covers the true, but unknown parameter with a certain confidence probability. In this paper, we discuss the issues occurring in the construction of confidence intervals for a binomial proportion, and assess the magnitude of estimation uncertainty for exemplary but representative credit portfolios. To give an economic meaning to the range of errors, we translate the *PD* confidence interval into a risk-weight confidence interval by applying the Basel II IRB approach.

The two main conclusions are: (i) The magnitude of estimation uncertainty can be substantial and is economically relevant. (ii) The choice of confidence interval matters and differences between intervals can be large. (JEL G21, C80)

Zusammenfassung

Eine Schätzung der Schätzungenauigkeit von Ausfallwahrscheinlichkeiten

Die Ausfallwahrscheinlichkeit (Probability of Default, *PD*) ist eine der zentralen Variablen im Kreditrisikomanagement. Um *PD*-Schätzungen als Einflussgrößen für Preisbildung und die Berechnung von Eigenmittelhinterlegung heranzuziehen, sollte man die Frage nach der Güte dieser Schätzungen stellen. Konfidenzintervalle sind dabei eine geeignete Möglichkeit, jenen Bereich zu bestimmen, der den tatsächlichen aber unbekannten Wert des Parameters mit einer bestimmten Wahrscheinlichkeit überdeckt. In diesem Artikel besprechen wir Aspekte, die bei der Bildung von Konfidenzintervallen für binomial verteilte Größen zu beachten sind, und geben eine Einschätzung der Schätzungenauigkeit für exemplarische, repräsentative Kreditportfolios. Um den Schätzfehlern eine ökonomische Bedeutung zu verleihen, übersetzen wir die Resultate der Konfidenzintervalle in Risikogewichte, indem wir den IRB-Ansatz der Neuen Basler Eigenkapitalvereinbarung anwenden.

Die zwei zentralen Resultate sind: (i) Das Ausmaß der Schätzungenauigkeit ist substanziell und ökonomisch relevant. (ii) Die Wahl des Konfidenzintervalls ist von Bedeutung, da Unterschiede zwischen verschiedenen Alternativen bedeutend sind.

DOI https://doi.org/10.3790/kuk.41.2.217 | Generated on 2025-10-30 15:46:04
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/