

Article

A Maximum-Entropy Method to Estimate Discrete Distributions from Samples Ensuring Nonzero Probabilities

Paul Darscheid ¹, Anneli Guthke ²  and Uwe Ehret ^{1,*} 

¹ Institute of Water Resources and River Basin Management, Karlsruhe Institute of Technology—KIT, 76131 Karlsruhe, Germany; p_dar.scheid@t-online.de

² Institute for Modelling Hydraulic and Environmental Systems (IWS), University of Stuttgart, 70569 Stuttgart, Germany; anneli.guthke@iws.uni-stuttgart.de

* Correspondence: uwe.ehret@kit.edu; Tel.: +41-721-608-41933

Received: 18 July 2018; Accepted: 13 August 2018; Published: 13 August 2018



Abstract: When constructing discrete (binned) distributions from samples of a data set, applications exist where it is desirable to assure that all bins of the sample distribution have nonzero probability. For example, if the sample distribution is part of a predictive model for which we require returning a response for the entire codomain, or if we use Kullback–Leibler divergence to measure the (dis-)agreement of the sample distribution and the original distribution of the variable, which, in the described case, is inconveniently infinite. Several sample-based distribution estimators exist which assure nonzero bin probability, such as adding one counter to each zero-probability bin of the sample histogram, adding a small probability to the sample pdf, smoothing methods such as Kernel-density smoothing, or Bayesian approaches based on the Dirichlet and Multinomial distribution. Here, we suggest and test an approach based on the Clopper–Pearson method, which makes use of the binominal distribution. Based on the sample distribution, confidence intervals for bin-occupation probability are calculated. The mean of each confidence interval is a strictly positive estimator of the true bin-occupation probability and is convergent with increasing sample size. For small samples, it converges towards a uniform distribution, i.e., the method effectively applies a maximum entropy approach. We apply this nonzero method and four alternative sample-based distribution estimators to a range of typical distributions (uniform, Dirac, normal, multimodal, and irregular) and measure the effect with Kullback–Leibler divergence. While the performance of each method strongly depends on the distribution type it is applied to, on average, and especially for small sample sizes, the nonzero, the simple “add one counter”, and the Bayesian Dirichlet-multinomial model show very similar behavior and perform best. We conclude that, when estimating distributions without an a priori idea of their shape, applying one of these methods is favorable.

Keywords: histogram; sample; discrete distribution; empty bin; zero probability; Clopper–Pearson; maximum entropy approach

1. Introduction

Suppose a scientist, having gathered extensive data at one site, wants to know whether the same effort is required at each new site, or whether already a smaller data set would have provided essentially the same information. Or imagine an operational weather forecaster working with ensembles of forecasts. Working with ensemble forecasts usually involves handling considerable amounts of data, and the forecaster might be interested to know whether working with a subset of the ensemble is sufficient to capture the essential characteristics of the ensemble. If what the scientist and the forecaster are interested in is expressed by a discrete distribution derived from the data (e.g., the distribution

of vegetation classes at a site, or the distribution of forecasted rainfall), then the representativeness of a subset of the data can be evaluated by measuring the (dis-)agreement of a distribution based on a randomly drawn sample (“sample distribution”) and the distribution based on the full data set (“full distribution”). One popular measure for this purpose is the Kullback–Leibler divergence [1]. Depending on the particular interest of the user, potential advantages of this measure are that it is nonparametric, which avoids parameter choices influencing the result, and that it measures general agreement of the distributions instead of focusing on particular aspects, e.g., particular moments.

For the use cases described above, if the sample distribution is derived from the sample data via the bin-counting (BC) method, which is the most common and probably most intuitive approach, a situation can occur where a particular bin in the sample distribution has zero probability but the corresponding bin in the full distribution has not. From the way the sample distribution was constructed, we know that this disagreement is not due to a fundamental disagreement of the two distributions, but rather that this is a combined effect of sampling variability and limited sample size. However, if we measure the (dis-)agreement of the two distributions via Kullback–Leibler divergence, with the full distribution as the reference, divergence for that bin is infinite, and consequently so is total divergence. This is impractical, as an otherwise possibly good agreement can be overshadowed by a single zero probability. A similar situation occurs if a distribution constructed from a limited data set (e.g., three months of air-temperature measurements) contains zero-probability bins, but from physical considerations we know that values falling into these zero-probability bins can and will occur if we extend the data set by taking more measurements.

Assuring nonzero (NZ) probabilities when estimating distributions is a requirement found in many fields of engineering and sciences [2–4]. If we stick to BC, this can be achieved either by adjusting the binning to avoid zero probabilities [5–9], or by replacing zero probabilities with suitable alternatives. Often-used approaches to do so are (i) assigning a single count to each empty bin of the sample histogram, (ii) assigning a (typically small) preselected probability mass to each zero probability bin in the sample pdf and renormalizing the pdf afterwards, (iii) spreading probability mass within the pdf by smoothing operations such as Kernel-density smoothing (KDS) [10] (an extensive overview on this topic can be found in Reference [11]), and (iv) assigning a NZ guaranteeing prior in a Bayesian (BAY) framework. Whatever method we apply, desirable properties we may ask for are introducing as little unjustified side information as possible (e.g., assumptions on the shape of the full distribution) and, like the BC estimator, convergence towards the full distribution for large samples.

In this context, the aim of this paper is to present a new method of calculating the sample distribution estimate, which meets the mentioned requirements, and to compare it to existing methods. It is related to and draws from approaches to estimate confidence intervals of discrete distributions based on limited samples [12–17]. In the remainder of the text, we first introduce the “NZ” method and discuss its properties. Then we apply the NZ method and four alternatives to a range of typical distributions, from which we draw samples of different sizes. We use Kullback–Leibler divergence to measure the agreement of the full and the sample distributions. We discuss the characteristics of each method and their relative performance with a focus on small sample sizes and draw conclusions on the applicability of each method.

2. The NZ Method

2.1. Method Description

For a variable with discrete distribution p with K bins, and a limited data sample S , thereof of size n , we derive a NZ estimator \hat{p} for p based on S as follows: For the occurrence probability of each bin β_k ($k = 1, \dots, K$), we calculate a BC estimator q_k and its confidence interval $CI_{p,k} = [p_{k,lower}; p_{k,upper}]$ on a chosen confidence level (e.g., 95%).

Based on the fact that the occurrence probability of a given bin from n -repeated trials follows a binomial distribution with parameters n and p_k , there exist several ways to determine a confidence

interval for this situation [18]. Several of these methods approximate the binomial distribution with a normal distribution, which is only reasonable for large n , or use other assumptions. To avoid any of these limitations and to keep the methods especially useful for cases of small n (here the probability of observing zero probability bins is the highest), we calculate $CI_{p,k}$ using the conservative yet exact Clopper–Pearson method [19]. It applies a maximum-likelihood approach to estimate p_k given the sample S of size n . The required conditions for the method to apply are:

- there are only two possible outcomes of each trial,
- the probability of success for each trial is constant, and
- all trials are independent.

In our case, this is assured by distinguishing the two outcomes “the trial falls within the bin or not”, keeping the sample constant and random sampling.

In practice, there are two convenient ways to compute the confidence Interval $CI_{p,k}$. One way is to look it up, for example, in the original paper by Clopper and Pearson [19], where they present graphics of confidence intervals for different sample sizes n , different numbers of observations x , and different confidence levels. The second option is to compute the intervals using the Matlab function $[~, CI] = \text{binofit}(x, n, \text{alpha})$ (similar functions exist for R or python) with $1 - \text{alpha}$ defining the confidence level. This function uses a relation between the binomial and the Beta-distribution, for more details see e.g., Reference [20] (Section 7.3.4) and Appendix A.

For each $k = 1, \dots, K$, the NZ estimate \hat{p}_k is then calculated as the normalized mean value m_k of the confidence interval $CI_{p,k}$ according to Equation (1). Normalization with the sum of all m_k for $k = 1, \dots, K$ is required to assure that the total sum of probabilities in \hat{p} equals 1. For this reason, the normalized values of \hat{p}_k can differ a little from the mean of the confidence intervals.

$$\hat{p}_k := \frac{m_k}{\sum_k m_k}, \text{ with } m_k = \frac{p_{k,lower} + p_{k,upper}}{2} \quad (1)$$

Two text files with Matlab code (Version 2017b, MathWorks Inc., Natick, MA, USA) of the NZ method and an example application are available as Supplementary Material.

2.2. Properties

There are four properties of the NZ estimate \hat{p}_k that are important for our application:

1. **Maximum Entropy by default:** For an increasing number of zero probability bins in q , \hat{p} converges towards a uniform distribution. For any zero probability bin β_k we get $q_k = 0$, assign the same confidence interval, and, hence, the same NZ estimate. Consequently, estimating p on a size-zero sample results in a uniform distribution \hat{p} with $\hat{p}_k = 1/K$ for all $k = 1, \dots, K$, which is a maximum-entropy (or minimum-assumption) estimate. For small samples, the NZ estimate is close to a uniform distribution.
2. **Positivity:** As probabilities are restricted to the interval $[0, 1]$, and it always holds $p_{k,upper} > p_{k,lower}$, the mean value of the confidence interval $CI_{p,k}$ is strictly positive. This also applies to the normalized mean. This is the main property we were seeking to be guaranteed by \hat{p}_k .
3. **Convergence:** Since q_k is a consistent estimator (Reference [21], Section 5.2), it converges in probability towards p_k for growing sample size n . Moreover, the ranges of the confidence intervals $CI_{p,k}$ approach zero with increasing sample size n (Reference [19], Figures 4 and 5) and hence, the estimates \hat{p}_k converge towards p_k .
4. **As described above,** due to the normalization in the method, the NZ estimate does not exactly equal the mean of the confidence interval. However, the interval’s mean tends towards p_k with growing n and, hence, the normalizing sum in the denominator tends towards one. Consequently, for growing sample size n , the effect of the normalization is of less and less influence.

2.3. Illustration of Properties

An illustration of the NZ method and its properties is shown in Figure 1. The first plot, Figure 1a, shows a discrete distribution, constructed for demonstration purposes such that it covers a range of different bin probabilities. Possible outcomes are the six integer values $\{1, 2, \dots, 6\}$, where $p(1) = 0.51$ and all further probabilities are half of the previous, such that $p(6) = 0.015$. Figure 1b shows a random sample of size one taken from the distribution; here, the sample took the value “1”. The BC estimator q for the distribution p for outcomes $\{1, \dots, 6\}$ is shown with blue bars. Obviously, we encounter the problem of zero-probability bins here. In the same plot, the confidence intervals for the bin-occupation probability based on the Clopper–Pearson method on 95% confidence level are shown in green. Due to the small sample size, the confidence intervals are almost the same for all outcomes, and so is the NZ estimate for bin-occupation probability shown in red. Altogether, the NZ estimate is close to a uniform distribution, which is the maximum entropy estimate, except that the bin-occupation probability for the observed outcome “1” is slightly higher than for the others: The NZ estimate of the distribution is $\hat{p} = (0.1737, 0.1653, 0.1653, 0.1653, 0.1653, 0.1653)$. We can also see that the positivity requirement for bin occupation probability is met.

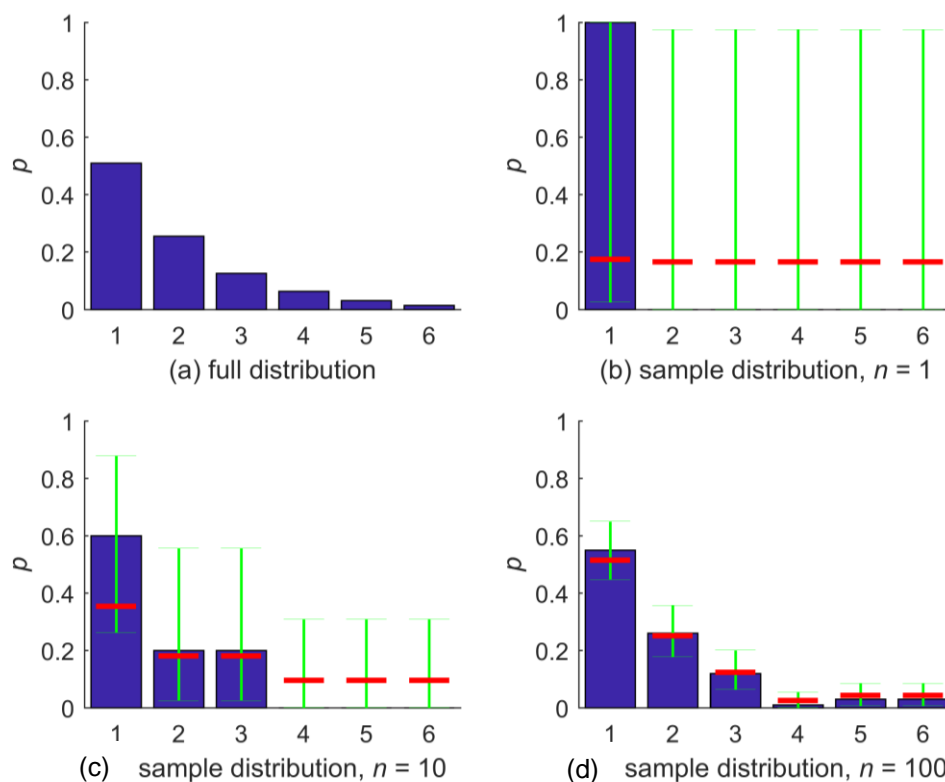


Figure 1. (a) Full distribution and (b–d) samples drawn thereof for different sample sizes n shown as blue bars. Green bars are the sample-based confidence intervals on 95% confidence level for bin-occupation probability based on the Clopper–Pearson method, and the red bar is the nonzero estimate for bin-occupation probability.

In Figure 1c,d, BC and NZ estimates of the bin-occupation probability are shown for random samples of size 10 and 100, respectively. For sample size 10, the BC method still yields three zero-probability bins, which are filled by the NZ method. The NZ estimates for this sample still gravitate towards a uniform distribution (red bars) but, due to the increased sample size, to a lesser degree than before. For sample size 100, both the BC and the NZ distribution estimate of bin-occupation probability closely agree with the full distribution, which illustrates the convergence behavior of the NZ method. Compared to the size-10 sample, the Clopper–Pearson confidence intervals for the

bin-occupation probabilities have narrowed considerably, and, as a result, the NZ estimates are close to those from BC.

3. Comparison to Alternative Distribution Estimators

3.1. Test Setup

How does the NZ method compare to established distribution estimators that also assure NZ bin-occupation probabilities? We address this question by applying various estimation methods to several types of distributions. In the following, we will explain the experimental setup, the evaluation method, the estimation methods, and the distributions used.

We start by taking samples S of size n by i.i.d. picking (random sampling with replacement) from each distribution p . Each estimation method we want to test applies this sample to construct a NZ distribution estimate \hat{p} . The (dis-)agreement of the full distribution with each estimate is measured with the Kullback–Leibler divergence as shown in Equation (2).

$$D_{KL}(p||q) = \sum_{\beta \in X} p(\beta) \log_2 \frac{p(\beta)}{q(\beta)} \quad (2)$$

with D_{KL} : Kullback–Leibler divergence [bit]; p : reference distribution; q : distribution estimate; X : set taking discrete values β_k (“bins”) for $k = 1, \dots, K$.

Note that, for our application, the full distribution of the variable is the reference p , since the observations actually occur according to this distribution; the distribution estimate q is derived from the sample and is our assumption about the variable. We chose Kullback–Leibler divergence as it conveniently measures, in a single number, the overall agreement of two distributions, instead of focusing on particular aspects, e.g., particular moments. Kullback–Leibler divergence is also zero if and only if the two distributions are identical, while, for instance, two distributions with identical mean and variance can still differ in higher moments.

We tested sample sizes from $n = 1$ to 150, increasing n in steps of one. We found an upper limit of 150 to be sufficient for two reasons: Firstly, the problem of zero-probability bins due to the combined effect of sampling variability and limited sample size mainly occurs for small sample sizes; secondly because, for large samples, the distribution estimates by the tested methods quickly become indistinguishable. To eliminate effects of sampling variability, we repeated the sampling for each sample size 1000 times, calculated Kullback–Leibler divergence for each and then took the average. As a result, we get mean Kullback–Leibler divergence as a function of sample size, separately for each estimation method and test distribution.

The six test distributions are shown in Figure 2. We selected them to cover a wide range of shapes. Please note that two of the distributions, Figure 2b,f, actually contain bins with zero p . It may seem that, in such a case, the application of a distribution estimator assuring NZ p 's is inappropriate; however, in our targeted scenarios (e.g., comparison of two distributions via Kullback–Leibler divergence), it is the zero p 's due to limited sample size that we need to avoid, while we accept the adverse effect of falsely correcting true zeros. If the existence and location of true-zero bins were known a priori, this knowledge could be easily incorporated in the distribution estimators discussed here to only produce actual NZ p 's.

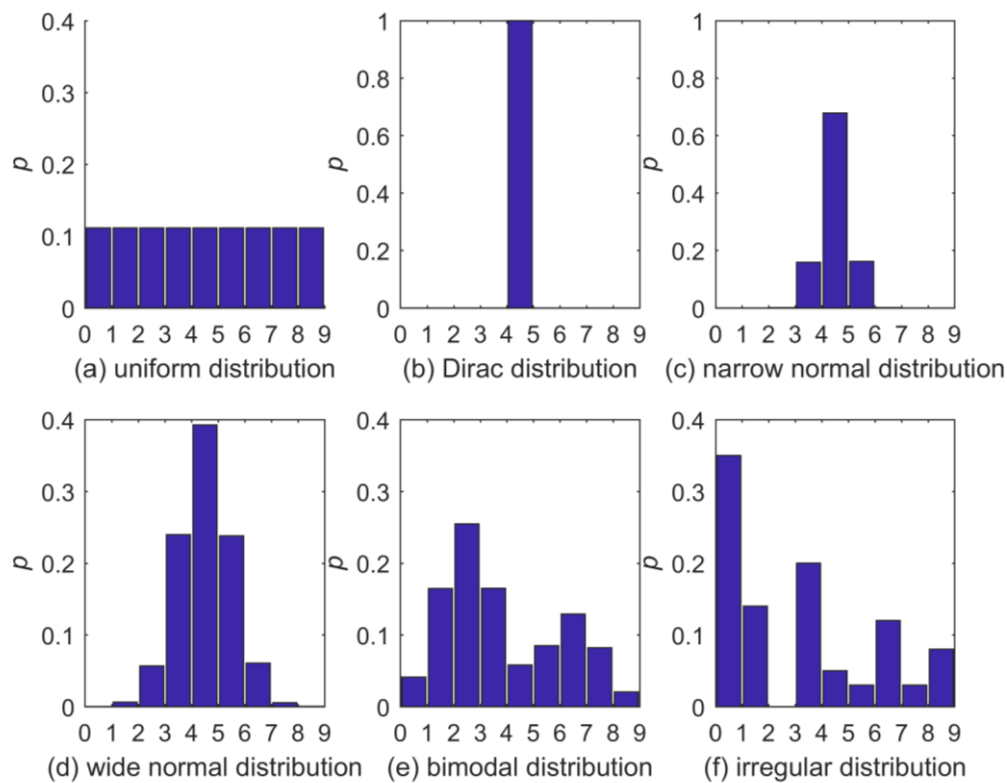


Figure 2. Test distributions: (a) Uniform, (b) Dirac, (c) narrow normal, (d) wide normal, (e) bimodal and (f) irregular. Possible outcomes are divided in nine bins of uniform width. Note that for (b,c), the y-axis limit is 1.0, but for all others it is 0.4.

Finally, we selected a range of existing distribution estimators to compare to the NZ method:

1. BC: The full probability distribution is estimated by the normalized BC frequencies of the sample taken from the full data set. This method is just added for completeness, and as it does not guarantee NZ bin probabilities its divergences are often infinite, especially for small sample sizes.
2. Add one (AO): With a sample taken from the full distribution, a histogram is constructed. Any empty bin in the histogram is additionally filled with one counter before converting it to a pdf by normalization. The impact of each added counter is therefore dependent on sample size.
3. BAY: This approach to NZ bin-probability estimation places a Dirichlet prior on the distribution of bin probabilities and updates to a posterior distribution in the light of the given sample via a multinomial-likelihood function [22]. We use a flat uniform prior (with the Dirichlet distribution parameter alpha taking a constant value of one over all bins) as a maximum-entropy approach, which can be interpreted as a prior count of one per bin. Since the Dirichlet distribution is a conjugate prior to the multinomial-likelihood function, the posterior again is a Dirichlet distribution with analytically known updated parameters. We take the posterior mean probabilities as distribution estimate and, for our choice of prior, they correspond to the observed bin counts increased by the prior count of one. Hence, BAY is very similar to AO with the difference that a count of one is added to all bins instead of only to empty bins; like for AO, the impact of the added counters is dependent on sample size. Like the NZ method, BAY is by default a strictly positive and convergent maximum-entropy estimator (see Section 2.2).
4. Add p (AP): With a sample taken from the full distribution, a histogram is constructed and normalized to yield a pdf. Afterwards, each zero-probability bin is filled with a small probability mass (here: 0.0001) and the entire pdf is then renormalized. Unlike in the “AO” procedure, the impact of each probability mass added is therefore virtually independent of n .

5. KDS: We used the Matlab Kernel density function `ksdensity` as implemented in Matlab R2017b with a normal kernel function, support limited to $[0, 9.001]$, which is the range of the test distributions, and an iterative adjustment of the bandwidth: Starting from an initially very low value of 0.05, the bandwidth (and with it the degree of smoothing across bins) was increased in 0.001 increments until each bin had NZ probability. We adopted this scheme to avoid unnecessarily strong smoothing while at the same time guaranteeing NZ bin probabilities.
6. NZ: We applied the NZ method as described in Section 2.1.

3.2. Results and Discussion

The results of all tests, separately for each test distribution and estimation method are shown in Figure 3. We will discuss them first individually for each distribution and later summarize the results.

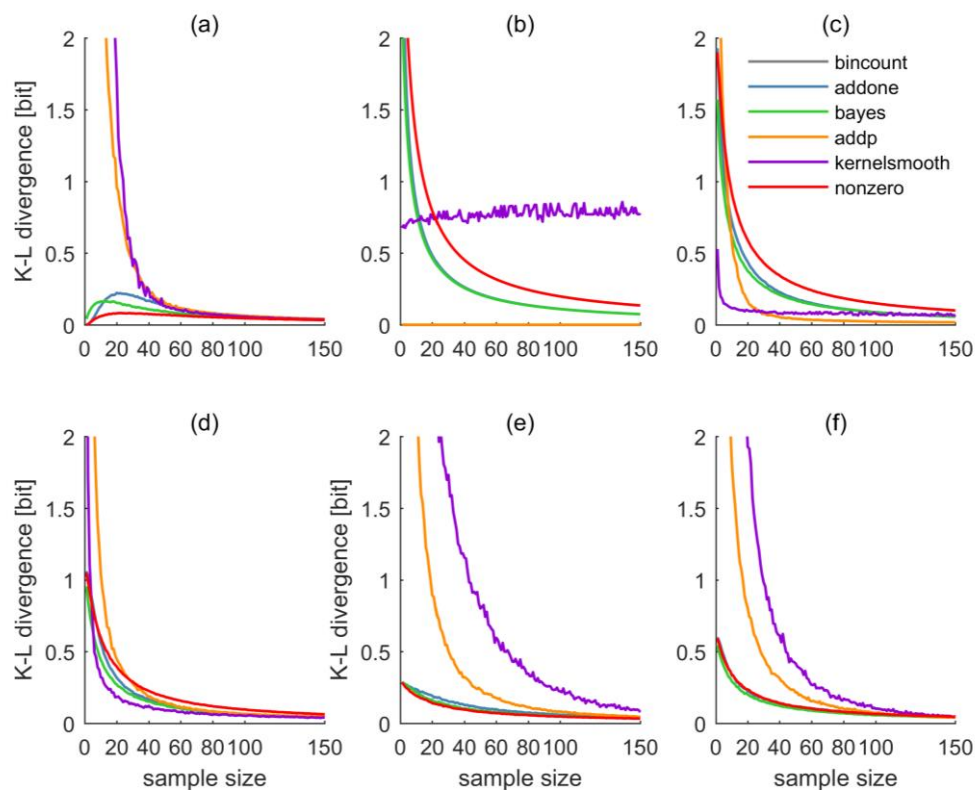


Figure 3. (a) Kullback–Leibler divergences of test distributions uniform, (b) Dirac, (c) narrow normal, (d) wide normal, (e) bimodal, and (f) irregular and size- n samples thereof. Sample-based distribution estimates are based on bin counting (grey), “Add one counter” (blue), “Bayesian” (green), “Add probability” (orange), “Kernel-density smoothing” (violet), and the “nonzero method” (red). In all plots except (b), the “bincount” line is invisible as its divergence is infinite, and in plot (b) it is invisible as it is zero and almost completely overshadowed by the “addp” line. In plots (b,f), the “addone” line is almost completely overshadowed by the “bayes” line. For better visibility, all y -axes are limited to a maximum divergence of 2 bit, although this limit is sometimes clearly exceeded for small sample sizes.

For the uniform distribution as shown in Figure 2a, the corresponding Kullback–Leibler divergences are shown in Figure 3a. For small sample sizes up to approximately 40, both AP and KDS show very large divergences, AO, BAY, and NZ perform considerably better, with a slight advantage of NZ. This order clearly reflects the methods’ different estimation strategies, and how capable they are to reproduce a uniform distribution: For small sample sizes, both AP and KDS will maintain “spiky” distribution estimates, while AO, BAY, and NZ gravitate towards uniform distribution or maximum-entropy estimates. For larger sample sizes, beyond 80, the performance differences among

the methods quickly vanish. For the small sample sizes as shown in the figure, the BC approach was still frequently afflicted with zero-probability bins, resulting in infinite divergence.

Quite expectedly, the relative performance of the estimators for the Dirac distribution (Figures 2b and 3b) is almost opposite from the uniform distribution. BC shows zero and AP almost-zero divergence for all sample sizes. The reason is that even a very small sample from a Dirac distribution yields a perfect estimate of the full distribution, and both methods do not interfere much with this estimate (in fact, BC not at all). AO and BAY show almost identical performance, NZ is similar but slightly worse. All of them show high divergences for small samples and a gradual decrease with sample size. The reason lies in the methods' tendency towards a uniform spreading of probabilities, which is clearly unfavorable if the true distribution is a Dirac. Interestingly, the KDS estimator performs constantly poorly over the entire range of sample sizes, which can be explained by its tendency of locally distributing probability mass around the BC estimate. In particular, as the kernel function was chosen to be normal, the observed divergence of about 0.8 bit corresponds to the divergence of a Dirac and a normal distribution extending over the nine bins covering the codomain.

For the narrow normal distribution as shown in Figures 2c and 3c, obviously the normal kernel of KDS is of advantage, such that, for small sample sizes, divergence is smaller than for any other estimator. The performance of AP varies greatly with sample size: For small samples it is poor, for sample sizes beyond about thirty it scores best. AO and BY are almost identical, NZ is similar to them but shows worse performance; altogether it is the worst estimator. Beyond sample sizes of about 80, all methods perform almost equally well, except for BC, whose divergence is infinite due to the occasional occurrence of zero-probability bins.

For the wide normal distribution as shown in Figures 2d and 3d, KDS remains the best estimator except for very small sample sizes. AO and BAY are similar and perform better than the NZ method. AP performs worst for sample sizes smaller about thirty; for larger samples, NZ performs worst.

For the bimodal distribution as shown in Figures 2e and 3e things look differently: Both KDS and AP show poor performance even for large sample sizes; AO, BAY, and NZ are almost indistinguishable and they perform well even for small sample sizes.

Finally, results for the application to the irregular distribution as shown in Figure 2f are shown in Figure 3f. As this distribution shows no pattern in the distribution of probabilities across the value domain, any approach assuming a particular shape of pattern (like KDS) will have difficulties, at least for small sample sizes. This is clearly reflected in the large divergences of KDS. Interestingly, AP also struggles to reproduce the irregular distribution, but not because of the absence of a probability pattern across the value domain, but because filling a bin that has zero probability due to chance with always the same small probability mass, irrespective of the sample size, here is less effective than filling it with an adaptive probability mass as done by AO. AO, BAY, and NZ, again, perform almost equally well and better than the other methods (BC again has infinite divergences).

4. Summary and Conclusions

We started by describing use cases that involve estimation of discrete distributions with the additional requirement that all bins of the estimated distribution should have NZ probabilities. As the standard BC approach does not guarantee this, we proposed an alternative approach based on the Clopper–Pearson method, which makes use of the binominal distribution. Based on the BC-distribution estimate, confidence intervals for bin-occupation probability are calculated. The mean of each confidence interval is a strictly positive estimator of the true bin-occupation probability and is convergent with increasing sample size. For small samples, it converges towards a uniform distribution, i.e., the method effectively applies a maximum-entropy approach. We compared the capability of this “NZ” method to estimate different distributions (uniform, Dirac, narrow normal, wide normal, bimodal, and irregular) based on i.i.d. samples thereof of different sizes. For comparison, we applied four alternative estimators guaranteeing NZ bin probabilities (adding one counter to each empty bin of the sample histogram, a BAY approach applying a Dirichlet prior, and a multinomial likelihood

function, adding a small probability to the sample pdf, and KDS). We measured the agreement of the distributions and their respective estimates via Kullback–Leibler divergence. The most obvious result is that the relative performance of the estimators strongly depends on whether their estimation strategy matches the shape of the test distribution or not. So if the latter is known (or can be reasonably guessed) a priori, a case-specific choice should be made. However, if this is not the case, it is reasonable to select an estimator that performs, on average, well across all distributions. For the range of distributions tested here, this could be either the straightforward method of adding one counter to each empty bin of a sample histogram, the BAY method, or the NZ method. As could be expected by their design, the first two show almost identical behavior and performance. The NZ method is similar to them in overall performance and its dependency of performance on sample size, except that it performs better for close-to-uniform distributions and worse for spiky distributions. Each of the three methods (AO, NZ, and BAY) is straightforward to implement and computationally inexpensive, so from a practical viewpoint, there is no preference for one method or the other. The main differences are in the formal background: The “AO” method lacks a formal justification; the NZ method is based on a statistical/frequentist background, while the BAY method applies a BAY perspective. Although the NZ and the BAY methods are formulated in different formal frameworks, they are in fact very similar (both are maximum-entropy estimators by construction), and so is their performance. Their main differences are that the NZ method applies the binomial distribution to evaluate each bin separately, while the BAY method applies the multinomial distribution simultaneously to all bins. The second difference is that the NZ method uses the normalized mean of the confidence interval of bin probability as the best estimate of bin probability; the BAY method uses the posterior mean. An advantage of the NZ and the BAY over the AO method is that, in addition to the distribution estimate, they also provide confidence intervals that offer additional avenues of analysis or conditioning. An additional advantage of the BAY method is that it offers adaptability: If a priori estimates of the distribution shape are available, they can be considered via the choice of the Dirichlet distribution parameter alpha. Overall, users may make a choice according to the formal setting they are most comfortable with.

Supplementary Materials: The following are available online at <http://www.mdpi.com/1099-4300/20/8/601/s1>. Two text files with Matlab code (version 2017b) are available as supplementary material. File “f_NonZero_method.m” is a function to compute the NZ estimate of a pdf given its histogram, file “apply_NonZero_method.m” is an application example calling the function. The files are also available on Github at https://github.com/KIT-HYD/NonZero_method (accessed on 13 August 2018).

Author Contributions: P.D. and U.E. jointly developed the NZ method and jointly wrote the paper. A.G. contributed the Dirichlet-Multinomial tests and interpretations.

Funding: This research received no external funding.

Acknowledgments: The third author acknowledges support by Deutsche Forschungsgemeinschaft DFG and Open Access Publishing Fund of Karlsruhe Institute of Technology (KIT). We gratefully acknowledge the fruitful discussions about the NZ method with Manuel Klar, Grey Nearing, Hoshin Gupta, and Wolfgang Nowak. Further we acknowledge comments by an anonymous referee of the first version of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix The Clopper–Pearson Method

Let X be a discrete random variable that may take values in $\chi = \{\beta_1, \dots, \beta_k\}$ and $S_n = S_n(X) = \{x_1, \dots, x_n\}$ a set of realisations of X . For each $\beta_k \in \chi$, we estimate the probability $p(\beta_k)$ of an observation from X to fall into bin β_k by the BC probability

$$q(\beta_k) = \frac{|\{x \in S_n(X) | x = \beta_k\}|}{n}. \quad (\text{A1})$$

The aim is to develop a confidence interval $CI = [p_L, p_U]$ for $p(\beta_k)$ to the confidence level $(1-\alpha) = 95\%$.

We derive a confidence interval for each $p_k := p(\beta_k)$ individually; for the sake of readability, we neglect the index k in the following and only write $p := p_k, q := q_k$. Let $Y = |\{x \in S_n(X) | x = \beta_k\}|$ be the random variable that counts all the observations in S_n which actually fall into bin β_k . This Y has a binominal distribution with parameters n and $p, Y \sim \text{Bin}(n, p)$, i.e., for $y \in \{0, \dots, n\}$

$$\mathbb{P}(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \tag{A2}$$

Based on the result $Y = y$ ("The bin β_k is observed y times in S_n "), we want to give a 95% confidence interval for p around q .

Property A1. Consider a random variable $Y \sim \text{Bin}(n, p)$. Based on the observation of y successes out of the n trials, a confidence interval for p to the confidence level $(1-\alpha)$ is given by $CI = [p_L, p_U]$, such that

$$\sum_{j=y}^n \binom{n}{j} p_L^j (1 - p_L)^{n-j} = \frac{\alpha}{2} \tag{A3}$$

$$\sum_{j=0}^y \binom{n}{j} p_U^j (1 - p_U)^{n-j} = \frac{\alpha}{2} \tag{A4}$$

Proof. These confidence intervals were introduced by Reference [19]. The following proof is based on Reference [20] (Theorem 7.3.4).

For any $p \in (0, 1)$ and corresponding $Y \sim \text{Bin}(n, p)$, take the largest $h_1(p) \in \mathbb{N}$ with $\mathbb{P}[Y \leq h_1(p)] \leq \alpha/2$ and the smallest $h_2(p) \in \mathbb{N}$ with $\mathbb{P}[Y \geq h_2(p)] \leq \alpha/2$. For these values we get

$$\mathbb{P}[h_1(p) < Y < h_2(p)] = 1 - \mathbb{P}[Y \leq h_1(p) \text{ or } Y \geq h_2(p)] = 1 - \alpha \tag{A5}$$

Since the functions $h_1, h_2 : (0, 1) \rightarrow \mathbb{N}$ are monotonically increasing discontinuous step functions, we can define $g_1(y) := \min\{h_1^{-1}(\{y\})\}$ as the minimal p , such that $h_1(p) = y$; analogously, $g_2(y) := \max\{h_2^{-1}(\{y\})\}$ is the maximal p with $h_2(p) = y$. With these definitions, the events $[h_1(p) < Y < h_2(p)]$ and $[g_2(Y) < p < g_1(Y)]$ are equivalent, so we have

$$\mathbb{P}[g_2(Y) < p < g_1(Y)] \geq 1 - \alpha \tag{A6}$$

We have shown that for the observation $Y = y$ and for $p_L := g_2(y), p_U := g_1(y), CI = [p_L, p_U]$ is a $(1-\alpha)$ confidence interval for p . For $Y \sim \text{Bin}(n, p)$ it holds

$$\mathbb{P}[Y \leq h_1(p)] \leq \frac{\alpha}{2} \iff \sum_{j=0}^{h_1(p)} \binom{n}{j} p^j (1 - p)^{n-j} \leq \frac{\alpha}{2} \tag{A7}$$

Hence, the upper limit p_U fulfills the following equation:

$$\sum_{j=0}^{h_1(p_U)} \binom{n}{j} p_U^j (1 - p_U)^{n-j} \leq \frac{\alpha}{2} \tag{A8}$$

With $h_1(p_U) = h_1(g_1(y)) = y$ and since we are looking for the smallest-possible upper bound, p_U has to satisfy (A4). Similarly, we can show that p_L has to satisfy (A3).

Of course, for $y = 0$, we have $p_L = 0$ and easily compute

$$\sum_{j=0}^0 \binom{n}{j} p_U^j (1 - p_U)^{n-j} = \frac{\alpha}{2} \iff (1 - p_U)^n = \frac{\alpha}{2} \iff p_U = 1 - \sqrt[n]{\frac{\alpha}{2}} \tag{A9}$$

In the same way, for $y = n$, we get $CI = [\sqrt[n]{\frac{\alpha}{2}}, 1]$. For all other $y = 1, \dots, n - 1$, there is no such easy closed solution to compute. Reference [19] and others give tables that list the solutions for these equations for some $n \in \mathbb{N}$. For our purposes (i.e., for use in Matlab codes), however, it is more convenient to use the relation between the binomial and the beta distribution to compute the confidence interval limits p_L and p_U [23,24]. \square

Property A2 (see e.g., Reference [25], Equation (3.37), Chapter 3). For a random variable $Y \sim \text{Bin}(n, p)$ and $y \in \{1, \dots, n - 1\}$, it holds

$$\mathbb{P}(Y \geq y) = I_p(y, n - y + 1), \tag{A10}$$

$$\mathbb{P}(Y \leq y) = 1 - I_p(y + 1, n - y), \tag{A11}$$

where $I_x(a, b) = \frac{1}{B(a, b)} \int_0^x t^{a-1} (1 - t)^{b-1} dt$ is the regularized incomplete beta function with the beta function $B(a, b) = \int_0^1 t^{a-1} (1 - t)^{b-1} dt$.

Proof. First of all, with the property

$$I_{1-x}(a, b) = 1 - I_x(b, a) \tag{A12}$$

for $a, b \in \mathbb{R}_{>0}$, $x \in (0, 1)$ (e.g., (olver), Equation (8.17.4)), it holds

$$\mathbb{P}(Y \leq y) = 1 - I_p(y + 1, n - y) = I_{1-p}(n - y, y + 1). \tag{A13}$$

Moreover, for the beta function with $a, b \in \mathbb{N}$, we have

$$B(a, b) = \frac{(a - 1)!(b - 1)!}{(a + b - 1)!} \tag{A14}$$

(e.g., Reference [26], Equation (5.12.1)). Using this, Equation (A13) can be derived by integration by parts:

$$\begin{aligned} I_{1-p}(n - y, y + 1) &= \frac{n!}{(n - y - 1)!y!} \int_0^{1-p} t^{n-y-1} (1 - t)^y dt \\ &= \frac{n!}{(n - y - 1)!y!} \left(\left[\frac{1}{n - y} t^{n-y} (1 - t)^y \right]_0^{1-p} + \int_0^{1-p} \frac{1}{n - y} t^{n-y} y (1 - t)^{y-1} dt \right) \\ &= \frac{n!}{(n - y)!y!} p^y (1 - p)^{n-y} + \frac{n!}{(n - y)!(y - 1)!} \int_0^{1-p} t^{n-y} (1 - t)^{y-1} dt \\ &= \binom{n}{y} p^y (1 - p)^{n-y} + I_{1-p}(n - (y - 1), (y - 1) + 1). \end{aligned} \tag{A15}$$

For $y = 0$, the claim simplifies to

$$I_{1-p}(n, 1) = \frac{n!}{(n - 1)!} \int_0^{1-p} t^{n-1} dt = n \left[\frac{1}{n} t^n \right]_0^{1-p} = (1 - p)^n = \mathbb{P}(Y \leq 0) \tag{A16}$$

so Equation (A11) follows by induction. From there, Equation (A10) follows directly

$$\begin{aligned} \mathbb{P}(Y \geq y) = 1 - \mathbb{P}(Y < y) &= 1 - \mathbb{P}(Y \leq y - 1) = 1 - [1 - I_p((y - 1) + 1, n - (y - 1))] \\ &= I_p(y, n - y + 1). \end{aligned} \tag{A17}$$

The beta distribution $\text{Beta}(a, b)$, $a, b \in \mathbb{R}_{>0}$, is defined by its probability density function (for $x \in (0, 1)$)

$$f(x; a, b) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} \quad (\text{A18})$$

and on the interval $(0, 1)$, the cumulative distribution function is

$$F(x; a, b) = I_x(a, b). \quad (\text{A19})$$

Hence, from all the considerations above results the following theorem, which we use to compute the Clopper–Pearson confidence intervals for the binominal distribution parameter p . \square

Theorem A1. Consider the random variable X with values in $\chi = \{\mathfrak{B}_1, \dots, \mathfrak{B}_k\}$. If bin \mathfrak{B}_k is observed $y = |\{x \in S_n(X) | x = \mathfrak{B}_k\}|$ times in a data set $S_n(X)$ of length n , the $(1 - \alpha)$ confidence interval for the probability of bin \mathfrak{B}_k is $CI = [p_L, p_U]$, where

(a) for $y = 1, \dots, n - 1$:

- (i) p_L is the $\alpha/2$ quantile of the beta distribution $\text{Beta}(y, n - y + 1)$,
- (ii) p_U is the $1 - \alpha/2$ quantile of the beta distribution $\text{Beta}(y + 1, n - y)$;

(b) for $y = 0$:

- (i) $p_L = 0$,
- (ii) $p_U = 1 - \sqrt[n]{\frac{\alpha}{2}}$;

(c) for $y = n$:

- (i) $p_L = \sqrt[n]{\frac{\alpha}{2}}$,
- (ii) $p_U = 1$.

References

1. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
2. Beck, J.L.; Au, S.-K. Bayesian Updating of Structural Models and Reliability using Markov Chain Monte Carlo Simulation. *J. Eng. Mech.* **2002**, *128*, 380–391. [[CrossRef](#)]
3. Au, S.K.; Beck, J.L. Important sampling in high dimensions. *Struct. Saf.* **2003**, *25*, 139–163. [[CrossRef](#)]
4. Kavetski, D.; Kuczera, G.; Franks, S.W. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resour. Res.* **2006**, *42*. [[CrossRef](#)]
5. Pechlivanidis, I.G.; Jackson, B.; McMillan, H.; Gupta, H.V. Robust informational entropy-based descriptors of flow in catchment hydrology. *Hydrol. Sci. J.* **2016**, *61*, 1–18. [[CrossRef](#)]
6. Knuth, K.H. Optimal Data-Based Binning for Histograms. *arXiv*, 2013.
7. Fraser, A.M.; Swinney, H.L. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A* **1986**, *33*, 1134–1140. [[CrossRef](#)]
8. Darbellay, G.A.; Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Inf. Theory* **1999**, *45*, 1315–1321. [[CrossRef](#)]
9. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [[CrossRef](#)] [[PubMed](#)]
10. Blower, G.; Kelsall, J.E. Nonlinear Kernel Density Estimation for Binned Data: Convergence in Entropy. *Bernoulli* **2002**, *8*, 423–449.
11. Simonoff, J.S. *Smoothing Methods in Statistics*; Springer: Berlin/Heidelberg, Germany, 1996.
12. Good, I.J. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* **1953**, *40*, 237–264. [[CrossRef](#)]

13. Blaker, H. Confidence curves and improved exact confidence intervals for discrete distributions. *Can. J. Stat.* **2000**, *28*, 783–798. [[CrossRef](#)]
14. Agresti, A.; Min, Y. On Small-Sample Confidence Intervals for Parameters in Discrete Distributions. *Biometrics* **2001**, *57*, 963–971. [[CrossRef](#)] [[PubMed](#)]
15. Vollset, S.E. Confidence intervals for a binomial proportion. *Stat. Med.* **1993**, *12*, 809–824. [[CrossRef](#)] [[PubMed](#)]
16. Escobar, M.D.; West, M. Bayesian Density Estimation and Inference Using Mixtures. *J. Am. Stat. Assoc.* **1995**, *90*, 577–588. [[CrossRef](#)]
17. Argiento, R.; Guglielmi, A.; Pievatolo, A. Bayesian density estimation and model selection using nonparametric hierarchical mixtures. *Comput. Stat. Data Anal.* **2010**, *54*, 816–832. [[CrossRef](#)]
18. Leemis, L.M.; Trivedi, K.S. A Comparison of Approximate Interval Estimators for the Bernoulli Parameter. *Am. Stat.* **1996**, *50*, 63–68.
19. Clopper, C.J.; Pearson, E.S. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* **1934**, *26*, 404–413. [[CrossRef](#)]
20. Larson, H.J. *Introduction to Probability Theory and Statistical Inference*; Wiley: Hoboken, NJ, USA, 1982.
21. Bickel, P.J.; Doksum, K.A. *Mathematical Statistics*; CRC Press: Boca Raton, FL, USA, 2015; Volume 1.
22. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
23. Brown, L.D.; Cai, T.T.; DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **2001**, *16*, 101–133.
24. Blyth, C.R. Approximate binomial confidence limits. *J. Am. Stat. Assoc.* **1986**, *81*, 843–855. [[CrossRef](#)]
25. Johnson, N.L.; Kemp, A.W.; Kotz, S. *Univariate Discrete Distributions*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2005.
26. Olver, F.W.; Lozier, D.W.; Boisvert, R.F.; Clark, C.W. *NIST Handbook of Mathematical Functions*, 1st ed.; Cambridge University Press: New York, NY, USA, 2010.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).