

A purely statistical bias in efficiency re-weightings - investigations in PIDCalib



Internal Note

Issue:	1
Revision:	0
Reference:	LHCb-XX-XXXX
Created:	Friday 20 th June, 2014
Last modified:	Friday 20 th June, 2014
Prepared by:	Stephen Ogilvy

Abstract

A purely statistical bias in efficiency re-weighting techniques is identified and demonstrated. This effect is common in all efficiency re-weighting procedures in which some calibration dataset is used to construct binned efficiencies. The magnitude of the bias is dependent on the magnitude of the efficiency of the selection being corrected for, on the statistical error from the binomial error on the bin efficiencies, and on the number of bins in a given schema. The effect of the bias is always to underestimate average re-weighted efficiencies. This bias was investigated for likely re-weighting procedures in PIDCalib, and while it was judged to generally be within the typical binomial error on individual bin efficiencies, this does not necessarily mean it is generally within the final error on the re-weighted efficiency. Depending on how analysts propagate PIDCalib bin efficiency errors, this effect can become significant.

Document Status Sheet

1. Document Title: A purely statistical bias in efficiency re-weightings - investigations in PIDCalib			
2. Document Reference Number: LHCb-XX-XXXX			
3. Issue	4. Revision	5. Date	6. Reason for change
Draft	1	June 20, 2014	First version.

Contents

1	Introduction	2
2	A statistical bias	2
3	Toy studies	4
3.1	$l = 10^5, \epsilon = 0.1$ - an exaggerated example	4
3.2	$l = 10^8, \epsilon = 0.002$ - approximation of $K \rightarrow \mu$ mis-ID in tight but realistic cuts	6
4	Conclusion	6

List of Figures

1	Exaggerated example - bin ϵ distributions	5
2	Exaggerated example - ϵ bias vs number of bins	5
3	Practical example - bin ϵ distributions	6
4	Practical example - ϵ bias vs number of bins	7

List of Tables

1	Exaggerated example - re-weighting biases	5
2	Practical example - re-weighting biases	6

1 Introduction

Efficiency re-weighting procedures are commonly conducted on LHCb. Arguably the most prevalent is the `PIDCalib` technique. In this procedure, so-called “golden modes” are reconstructed and selected through decay kinematics alone, without the use of PID discrimination. The tracks from these modes are then used to construct local efficiencies by binning in variables on which the PID response is dependent (usually some combination of p , p_T , η , $nTracks$ etc.). This is necessary as the kinematics of decay modes of interest are different to the calibration samples.

The systematic uncertainties in this procedure come from both increased binomial uncertainties due to the division of the dataset into bins with smaller populations than the whole, and from variations in PID cut efficiency across single kinematic bins. The second is well known to result in biases in local regions where the distributions of calibration and reference samples have very different kinematics. It is the intention of this document to demonstrate that the inflated binomial uncertainties, an effect which is known to result in an increased binomial uncertainty on the propagated reference sample efficiency, can also manifest in biases on the re-weighted PID efficiencies of reference samples. This bias is entirely statistical in nature, and is common to all instances of `PIDCalib`.

The severity of these effects are typically not large given the standard `PIDCalib` applications. Typical hadronic PID calibrations involve PID cut efficiencies of order 10^{-1} , with millions of calibration tracks. As such, the binomial errors on calibration bins are low enough that these biases become negligible. The possibility was considered that in the case of mis-ID measurements with tight PID cuts, as is the case with for example the $B_s^0 \rightarrow \mu^+ \mu^-$ analysis, where the $K \rightarrow \mu$ mis-ID is estimated at 10^{-3} , the effect may be significant. This was investigated with toy methods and the expected bias is shown to be within the individual bin binomial errors on efficiency. It should be stressed, however, that this is no guarantee that the bias is within the final error on the re-weighted PID efficiency - this is treated differently by different analyses and may become significant.

2 A statistical bias

To illustrate this bias it is easiest to consider binnings in variables on which the efficiency is not dependent, which unfolds biases from efficiency variation over individual efficiency bins. The considerations nonetheless apply equally well to those variables on which the efficiency does depend.

Consider a general calibration dataset of N_t tracks. We apply some PID selection ϵ to this population and are left with N_p tracks surviving the selection:

$$N_p = N_t \epsilon. \quad (1)$$

In the unrealistic case whereby the track kinematics are the same in the calibration and reference samples, this ϵ can be used to correct for the efficiency of a reference sample, with a known number of events passing PID selection k , to extract the number of reference tracks before selection, l - this l can be referred to as the adjusted track yield.

Realistically, the distributions of kinematics are different in the reference and calibration samples. The PID response depends on these, so we must use a re-weighting process to extract an efficiency for the reference sample. In these cases, the calibration dataset is binned in the kinematics, where local efficiencies in the kinematic space are derived. The efficiency for bin j is then given as:

$$\epsilon_j = \frac{k_j}{l_j}. \quad (2)$$

Now, in this particular example we consider a binning in a variable, x , which is uncorrelated with the PID efficiency. We bin the calibration data in x , apply some arbitrary PID selection and construct local efficiencies. Without efficiency variation in x , each bin efficiency will be roughly the same but of course statistical fluctuations will result in a distribution of bin efficiencies. Assuming a large enough number of bins, and a large enough number of calibration tracks populating each bin, the distribution of bin efficiencies will be Gaussian-distributed about the effective efficiency for the integrated dataset. The width of the Gaussian will be roughly the average binomial error on the bin efficiencies.

Now say we take a sample of reference data, which has had the same PID selection applied to it. We assume the distribution in x is the same for the calibration and reference sample. We also make the assumption that the track kinematics are the same in the calibration and reference samples. Again, this is not realistic but the effects are also true for cases whereby the kinematic distributions disagree!

We then use the bin efficiencies to assign efficiencies to the reference tracks, extracting the adjusted reference track yield l by summing over the bins:

$$l = \sum_j^j \frac{N_j}{\epsilon_j} \quad (3)$$

where N_j is the number of reference tracks in bin j . The ratio of the number of tracks passing PID selection over the adjusted track yield gives the re-weighted efficiency $\bar{\epsilon}$:

$$\bar{\epsilon} = \frac{k}{l} \quad (4)$$

In this example, x is independent of the PID response. Now suppose x is uniformly distributed in both the calibration and reference sample, say between zero and one. By binning in x , we attain the aforementioned Gaussian distribution of bin efficiencies about the integrated effective efficiency. So there are approximately an equal amount of bins with efficiency above the effective value as there are below the effective value. Therefore we expect approximately equal numbers of tracks to be assigned per-track efficiencies in bins which have efficiency fluctuations above the effective value, and equal events assigned those bin efficiencies with values below the effective value.

To illustrate the source of the bias, consider two reference tracks. If we average the calibration dataset over x , these events are both assigned the effective efficiency and the adjusted track yield of the two tracks is

$$l = \frac{2}{\epsilon} \quad (5)$$

Now consider again the binning in x and our two reference tracks. It is representative of the whole re-weighting to consider that one of these is a track which lies in a calibration bin with a statistical fluctuation in efficiency some value y above the effective efficiency, and one is a track falling in a calibration bin with a statistical fluctuation in efficiency some value y below the effective efficiency. Now, the tracks are assigned the adjusted track yield:

$$\begin{aligned} l &= \frac{1}{\epsilon + y} + \frac{1}{\epsilon - y} \\ &= \frac{(\epsilon - y) + (\epsilon + y)}{(\epsilon + y)(\epsilon - y)} \\ &= \frac{2\epsilon}{\epsilon^2 - y^2} \\ &= \frac{2}{\epsilon - \frac{y^2}{\epsilon}} \end{aligned} \quad (6)$$

In the case of zero statistical fluctuations this reduces to the phase-space averaged adjusted yield. For non-zero y , this always results in an adjusted track yield which is larger, and therefore a reference $\bar{\epsilon}$ which is always biased downwards. In any given bin schema over x there are equal bins fluctuating up and down in efficiency, and assuming equal distribution of the reference sample across x , equal numbers of tracks will be assigned higher fluctuated efficiencies as low fluctuated efficiencies. As can be seen, the downward fluctuations dominate over the upward. This results in a genuine downward bias in the re-weighted efficiencies which is purely statistical in nature.

As can also be seen, the magnitude of this bias is determined by the factor $\frac{y^2}{\epsilon}$. Therefore, the larger the statistical fluctuations, the larger the bias. Therefore the higher the binning granularity of a given track dataset, the greater the severity of this bias. Also, the higher the ratio of the average binomial uncertainty over the average efficiency,

$$\frac{\bar{\sigma}_\epsilon}{\bar{\epsilon}}, \quad (7)$$

the higher the manifested bias becomes. This ratio in typical PIDCalib analyses is tiny and so most do not see this effect, but it is possible that low mis-ID evaluations will be sensitive to it.

3 Toy studies

This is investigated using toy MC, by generating a simulated population of N_t “candidates” - this is our calibration sample. Each candidate is randomly assigned a value in a control variable, x , which is uniformly distributed between 0 and 1. A random selection is applied to this dataset, with a true efficiency ϵ_{true} . As such, the efficiency of the selection is independent of the control variable x . We calculate the effective efficiency of the selection by counting the number of candidates surviving the selection. We divide the sample into j equal bins in x , and derive an effective efficiency for each bin, ϵ_j .

We generate a second sample of l candidates with the same uniform distribution in x - this is our reference sample. We use the binned efficiencies constructed using the calibration data to assign per-event efficiencies to the reference sample. We then calculate the adjusted yield for the reference sample, and use this to extract the re-weighted efficiency $\bar{\epsilon}$. This is, of course, the standard `PIDCaLib` technique. With k_j as the number of reference events in bin j , the adjusted yield l is then:

$$l = \sum_j \frac{k_j}{\epsilon_j} \quad (8)$$

and the re-weighted efficiency is then

$$\bar{\epsilon} = \frac{k}{l}. \quad (9)$$

This is carried out for a number of binning schemas for a given calibration dataset, and we track the change in re-weighted efficiency.

3.1 $l = 10^5$, $\epsilon = 0.1$ - an exaggerated example

In this example we have low calibration statistics, so the effects are exaggerated for purposes of illustration. The selection efficiency is 0.1, the number of generated calibration events is 100000, the number of reference events is 100000 (although this is ultimately irrelevant), the extracted efficiencies are shown in Table 1, along with the typical binomial error for the bin schema, and the evaluation of the bias from the phase-space averaged efficiency. Naturally, as the number of bins increases, the fractional binomial efficiency error increases, and the width of the Gaussian distribution of bin efficiencies becomes larger. This leads to larger biases being introduced. Examples of the distributions of bin efficiencies for different numbers of bins are given in Figure 1. The magnitude of the bias in extracted efficiency is plotted against the number of bins in the re-weighting in Figure 2. The efficiency is being clearly biased downward with increasing granularity. While it is at all times within the typical statistical error on individual bins, this is not indicative of the final statistical error on a re-weighted efficiency. For this to be the case all bins would have to fluctuate in the same direction - which is not possible. As such the bias may be non-negligible.

Table 1: The extracted efficiencies for a variety of binning schemas. In this example, $\epsilon_{\text{true}} = 0.1$, $k = 100000$, $N_t = 100000$, $\epsilon_{\text{ave}} = 0.0994 \pm 0.0009$.

N Bins	ϵ_{RW} [%]	$ \epsilon_{\text{RW}} - \epsilon_{\text{ave}} $ [%]	Average bin σ_ϵ [%]	Fractional bias on ϵ [%]
10	9.92	0.01	0.30	0.15
25	9.91	0.03	0.47	0.26
50	9.89	0.05	0.67	0.46
100	9.84	0.10	0.95	0.97
200	9.73	0.21	1.34	2.13
300	9.66	0.28	1.64	2.77
400	9.55	0.39	1.89	3.94
500	9.41	0.53	2.12	5.32
600	9.33	0.61	2.32	6.14
700	9.11	0.83	2.50	8.33
800	9.08	0.86	2.68	8.65
900	8.91	1.02	2.84	10.31
1000	8.80	1.14	2.99	11.45

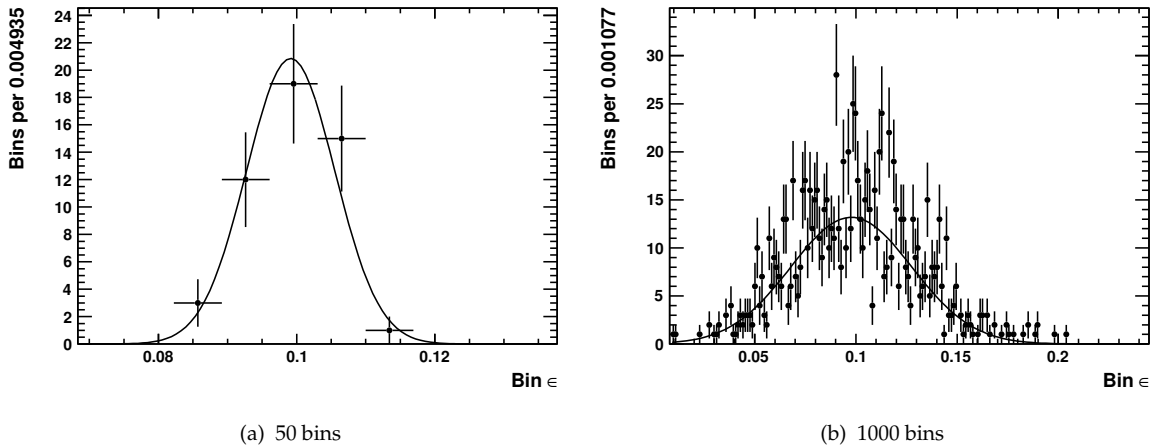


Figure 1: The bin binomial efficiency distributions distributions for 50 bins (a) and 1000 bins (b).

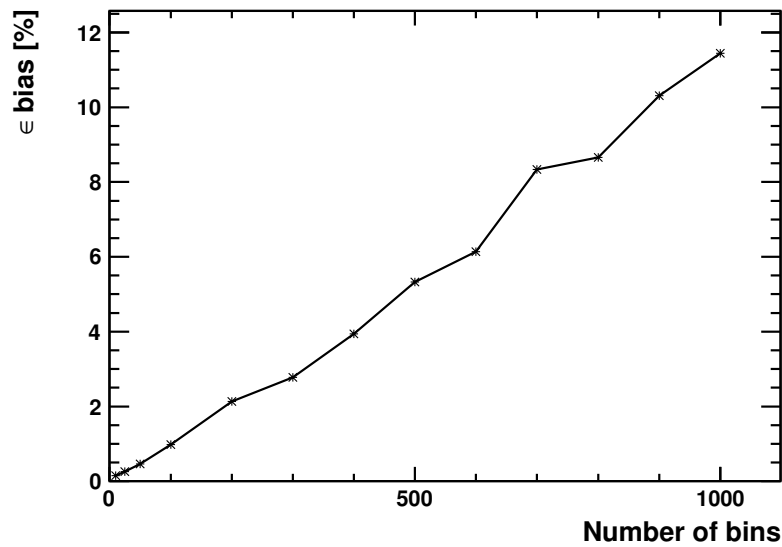


Figure 2: The fractional bias on re-weighted efficiency versus number of bins for the exaggerated example in the text.

3.2 $l = 10^8, \epsilon = 0.002$ - approximation of $K \rightarrow \mu$ mis-ID in tight but realistic cuts

The toy experiment is repeated, but now with a comparable number of calibration candidates as there are calibration tracks for the K species. We take a reasonable estimate for the $K \rightarrow \mu$ mis-ID rate as is evaluated for the $B_s^0 \rightarrow \mu^+ \mu^-$ analysis, around 0.002. As such we can estimate the effects of statistical biases in these mis-ID rates.

The results are shown in Table 2, again with some example efficiency distributions shown in Figure 3 with a plot of the bias against number of bins shown in Figure 4. We observe similar conclusions as before, that in all cases the bias is within the typical binomial error on the bin efficiencies. Again, we stress that this does not guarantee that the bias is smaller than the error on the re-weighted PID efficiency.

Table 2: The extracted efficiencies for a variety of binning schemas. In this example, $\epsilon_{\text{true}} = 0.1, k = 100000, N_t = 100000, \epsilon_{\text{ave}} = 0.0994 \pm 0.0009$.

N Bins	ϵ_{RW}	$ \epsilon_{\text{RW}} - \epsilon_{\text{ave}} [\%]$	Average bin $\sigma_\epsilon [\%]$	Fractional bias on ϵ [%]
10	0.19955	0.00000	0.00141	0.00201
25	0.19952	0.00003	0.00223	0.01576
50	0.19949	0.00006	0.00316	0.03108
100	0.19944	0.00011	0.00446	0.05677
200	0.19937	0.00018	0.00631	0.09012
300	0.19929	0.00026	0.00773	0.13221
400	0.19920	0.00035	0.00893	0.17534
500	0.19906	0.00049	0.00998	0.24695
600	0.19896	0.00059	0.01093	0.29331
700	0.19886	0.00069	0.01181	0.34374
800	0.19877	0.00078	0.01262	0.39152
900	0.19866	0.00089	0.01339	0.44389
1000	0.19859	0.00096	0.01411	0.48129

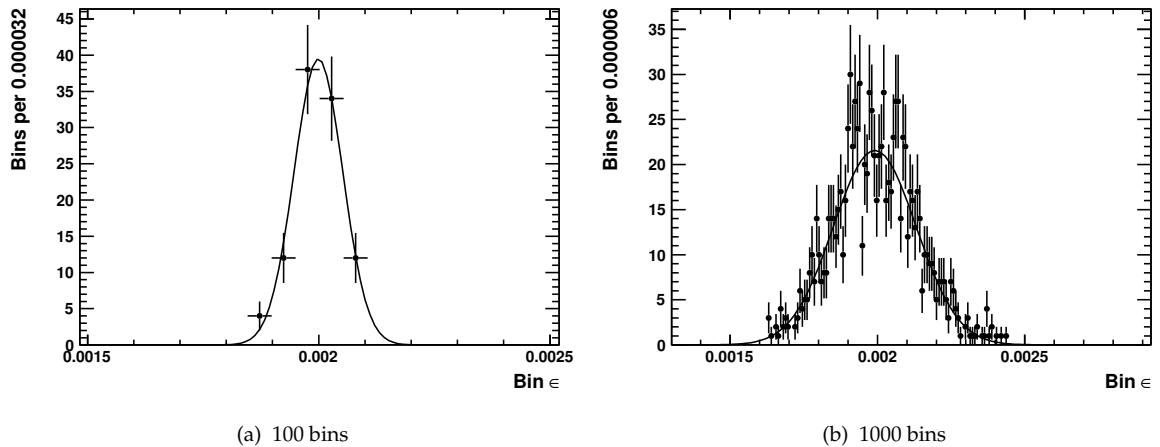


Figure 3: The bin binomial efficiency distributions for 100 bins (a) and 1000 bins (b).

4 Conclusion

A purely statistical bias which arises in efficiency re-weighting procedures was identified and explained. The possible effects of this bias on PIDCalib analyses was investigated using toy MC. It was found that, even in extreme low efficiency cases, as may be found in mis-ID studies, the bias is within the typical binomial uncertainty on bin efficiencies in the schemas. This does not guarantee that the bias is within the error on the final re-weighted efficiency, and as such it is difficult to make conclusions about

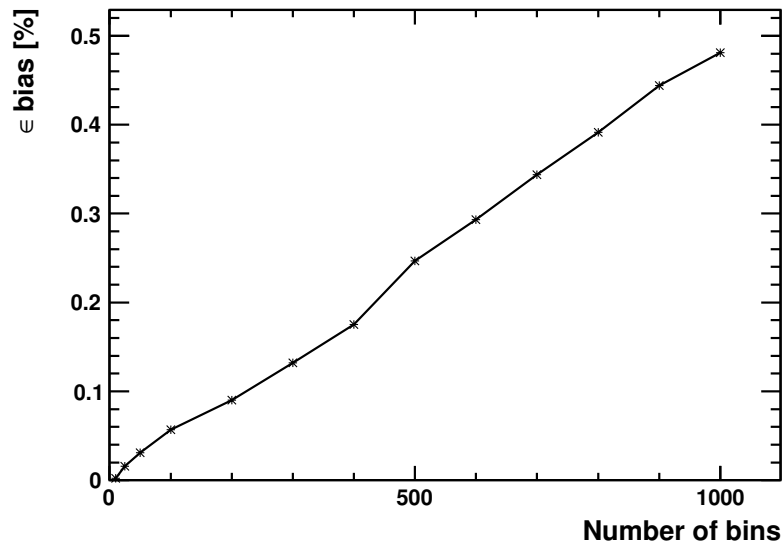


Figure 4: The fractional bias on re-weighted efficiency versus number of bins for the practical example in the text.

this bias. It should nonetheless be recognised that the ultimate effect of such statistical fluctuations is *always* to deflate the re-weighted efficiency, even by such small amounts. Depending on how the analysts handle the `PIDCalib` bin efficiencies, this effect may become significant. It should also be stressed that this is an effect which arises via the re-weighting procedure - it cannot explain why `PIDCalib` appears to overestimate/underestimate the mis-ID efficiencies in individual kinematic bins in some very low efficiency cases.