

Valentin Niess

Clermont Université, Université Blaise Pascal, CNRS/IN2P3

Laboratoire de Physique Corpusculaire,

BP 10448, F-63000 Clermont-Ferrand, France

(Dated: December 8, 2010)

The Banff Challenge 2a poses 2 problems relative to significance of discovery claim (<http://www-cdf.fnal.gov/~trj/bc2probs.pdf>), which are meant to simulate the task of discovering new particle or phenomena in high-energy physics experiments. We tried to apply frequentist hypothesis testing tools to these problems. Following a brief overview of the problematics, we summarise the main steps of the analysis and the results obtained. For problem 1 the test statistic relies on the bracketing of a region potentially rich in signal. The power achieved varies from 12% to 100% for the test bench cases. For problem 2 the test statistic is based on Kolmogorov-Smirnov's one. The power is 76% for 75 signal events.

I. PROBLEM 1: A GAUSSIAN SIGNAL PEAK ON AN EXPONENTIAL BACKGROUND

The observation is a data-set of $\sim 10^3$ events for which one measured the marks, $x \in [0; 1]$, of each event. The events in the data-set are distributed according to 2 statistic laws defining 2 categories: signal or noise. A signal event is Gaussian distributed of known standard deviation $\sigma = 0.03$ but unknown mean $E \in [0; 1]$. A noise event follows an exponential law of parameter $C = 1/\lambda = 10$. Both statistical distributions are truncated to the range $[0; 1]$. The challenge is to build the most powerful method to detect any signal event in the data-set given that the false alarm rate must be smaller than 1%. Secondly, when stating evidence for a signal one should also provide an estimate of its position, E , and its amplitude.

The general analysis strategy proceeds in 2 steps. First we construct a bracketing sub-interval of $[0; 1]$ enriched in signal events. Then we proceed with hypothesis testing for signal or background on the basis of the number of events lying in this sub-interval.

A. Construction of the bracketing sub-interval

Let us assume that the signal distribution is centred at a particular value E . Then, let us build a decision rule that for each event states whereas it is a signal event or a background one. The most powerful test statistic is to select signal events accordingly to the score of the likelihoods ratio of the signal probability density function (pdf) to the background one. Alternatively, taking the logarithm of this score, the test for accepting an event as a signal event writes as:

$$\frac{(x - E)^2}{\sigma^2} - 2Cx \leq T, \quad (1)$$

where T is the threshold value for acceptance. Its optimal value a priori depends on the hypothesis made for E . The acceptance region defined by Eq. 1 resumes to a sub-interval of $[0; 1]$.

1. Centre of the bracketing interval

Let us set $x_0 = E + \sigma^2 C$ and redefine the test threshold value as $T' = \sigma^2 (T + 2CE + \sigma^2 C^2)$. Then Eq. 1 rewrites as:

$$(x - x_0)^2 \leq T'. \quad (2)$$

From the latter equation one sees that the most powerful bracketing interval for the signal is centred somewhat to the right of the most likely signal value, at x_0 , where the background is weaker. Note that x_0 does not depend on the test threshold T' which is a strong result. The width 2Δ of the bracketing interval however depends on the choice of the threshold value, T' . Overall, the performances of the decision rule depend only on the choices made for x_0 and Δ . Therefore, in the following we wont refer anymore to the selection threshold value T' but rather to the half width Δ .

2. Width of the bracketing interval

The centre of the bracketing interval being optimally fixed to x_0 , we are left with picking an appropriate value for the interval's half width Δ . That for, let us write as N_e the total number of events in the data-set, N_s the number of signal events and N_b the number of background events, with $N_e = N_s + N_b$. Let us further write $p_{s,1}$ and $p_{b,1}$ the probabilities for a signal or background event to fall in the bracketing interval. Provided that the bracketing interval does not overlap with boundaries at 0 or 1, one has:

$$p_{s,1} = \frac{\text{erf}_\sigma(\Delta + C\sigma^2) + \text{erf}_\sigma(\Delta - C\sigma^2)}{2 \text{erf}_\sigma(1)}, \quad (3)$$

$$p_{b,1} = 2 \text{sh}(C\Delta) \frac{e^{-Cx_0}}{1 - e^{-C}}, \quad (4)$$

where we write $\text{erf}_\sigma(x) = \text{erf}(x/(\sqrt{2}\sigma))$.

We expect the number of signal events in the data-set to be small as compared to the number of background

events. Therefore, even for an optimal choice of Δ the bracketing sub-interval will be contaminated by a large number of background events, typically $B = N_b p_{b,1}$. The uncertainty on this contamination is of order \sqrt{B} which is to be compared to the expected number of signal events, given as $S = N_s p_{s,1}$. Therefore, in order to get the most significance for a signal excess in the bracketing interval we want to maximise the ratio S/\sqrt{B} . Since N_b and N_s are fixed, but unknown, we are left with maximising the ratio $p_{s,1}/\sqrt{p_{b,1}}$. From equations 3 one can check that there is an optimal choice Δ_0 for Δ , whatever x_0 . With the numerical values $C = 10$ and $\sigma = 0.03$ one gets $\Delta_0 = 0.0421 = 1.4\sigma$. Note also that for a fixed value of Δ , the ratio S/\sqrt{B} increases exponentially with the product Cx_0 .

In the following all bracketing intervals we use are constructed according to the previous optimality criterion. That is to say, for a signal hypothesis centred on E the corresponding bracketing interval is centred on $x_0 = E + \sigma^2 C$ and it has a half width $\Delta_0 = 1.4\sigma$.

B. Evidence for a Signal

In order to test whether or not the data-set contains any signal we build N_{bin} bracketing intervals regularly centred at marks $x_{0,i}$ with $i \in [1; N_{bin}]$. We write n_i the number of events that fall within the i^{th} bracket. Based on this observation, we perform N_{bin} hypothesis tests assuming a pure background distribution. The tests are done by fixing the false alarm rate to a common value $\alpha_{N_{bin}}$. If any of these tests fails we claim evidence for a signal.

1. Ordering rule and test statistic

When building the hypothesis test we are left with a large number of possibilities for the ordering of the various possible values for the observation n_i . The ordering rule we found to be the most powerful is to rank the values of n_i from the highest, the less in favour of a pure background, to the lowest. That is to say, we accept the background hypothesis $H_0^{(i)}$ for the i^{th} bracket provided that:

$$H_0^{(i)} : n_i \leq n_{i,c}, \quad (5)$$

where $n_{i,c}$ is a threshold value to adjust in order to ensure a Type I error lower than $\alpha_{N_{bin}}$. We tried more complex orderings, based on the most likely signal hypothesis, but they were less powerful while requiring more intensive computations.

2. P-value for a single bracket

Generally speaking, the probability to observe n_i events in the i^{th} bracket writes:

$$p_i(n_i; N_s, N_e) = \sum_{k_s = \sup\{0, n_i - N_b\}}^{\inf\{n_i, N_s\}} C_{N_s}^{k_s} p_{s,1}^{k_s} (1 - p_{s,1})^{N_s - k_s} \times C_{N_b}^{k_b} p_{b,1}^{k_b} (1 - p_{b,1})^{N_b - k_b}, \quad (6)$$

with $k_s + k_b = n_i$ and where we assume N_s events of signal and $N_b = N_e - N_s$ of background. In the case of a pure background hypothesis, $N_s = 0$, the sum in Eq. 6 reduces to a single term with $k_s = 0$ and $k_b = n_i$. The p-value corresponding to our choice of test statistic writes as:

$$P_{b,i}(n_i; N_e) = \sum_{k=n_i}^{+\infty} p_i(k; 0, N_e). \quad (7)$$

Note that this p-value takes discrete values. Therefore it is not possible to ensure exact coverage for a given confidence level $CL = 1 - \alpha_{N_{bin}}$. The definition we use in Eq. 7 is a conservative choice resulting in over-coverage provided that one accepts the background hypothesis $H_0^{(i)}$ for $P_b(n_i) \geq \alpha_{N_{bin}}$.

3. P-value for the background only hypothesis

Repeating the hypothesis tests for the different bracketing intervals, the background only hypothesis, H_0 , is accepted as:

$$H_0 : \inf_{i \in [1; N_{bin}]} P_{b,i}(n_i; N_e) \geq \alpha_{N_{bin}}. \quad (8)$$

If the N_{bin} tests would be independent the CL associated to the test on the N_{bin} brackets would simply be the product of the individual CLs for each bracket, resulting in:

$$1 - \alpha = (1 - \alpha_{N_{bin}})^{N_{bin}}, \quad (9)$$

where $\alpha = 1\%$ is the Type I error for the complete test. However, the tests on the individual brackets are not independent since the values of n_i are correlated from one bracket to another. Nevertheless, it was found that by substituting N_{bin} by an effective dimension N_{eff} eq. 9 is satisfied. The effective dimension writes as:

$$N_{eff} = \frac{\Delta x_0}{\sqrt{2\delta x_0 \Delta}}, \quad (10)$$

with $\Delta x_0 = x_{0, N_{bin}} - x_{0,1}$ the full range spanned by the bracketing intervals centres and δx_0 the constant step between two successive brackets centres. Following, we define the p-value for the background only hypothesis as:

$$P_b(N_e) = 1 - \left(1 - \inf_{i \in [1; N_{bin}]} P_{b,i}(n_i; N_e) \right)^{N_{eff}}. \quad (11)$$

We claim an evidence for a signal provided that: $P_b < \alpha$, where the type I error is ensured to be no more than α . For the brackets stepping we take $\delta x_0 = \sigma/2$ and we span the range $[0; 1]$. Note that varying the brackets step size δx_0 by a factor of two did not change significantly the algorithm performances.

C. Confidence belt for signal parameters

Whenever we find an evidence for a signal we build 68% CL belts for the signal position, E , and the number of signal events N_s . That for we rank the parameter set (E, N_s) according to the log ratio of the signal likelihood to the background one, as:

$$\text{LNR}(E, N_s, N_e) = 2 \ln(p_i(n_i; 0, N_e)) - 2 \ln(p_i(n_i; N_s, N_e)), \quad (12)$$

where the bracket probabilities p_i where given previously in Eq. 6. Low values of LNR are in favour of the signal hypothesis over the background one.

1. Dealing with nuisance parameters

From the latter test statistic the most rigorous way to proceed would be to build 2-dimensional confidence belts on the parameter set (E, N_s) . However, the challenge requires to build individual confidence belts on E and N_s separately. Therefore one has to deal with nuisance parameters issues. That for, we define the following Δ_X test statistics:

$$\Delta_E(E, N_e) = \inf_{N_s} \text{LNR} - \inf_{E, N_s} \text{LNR}, \quad (13)$$

$$\Delta_{N_s}(N_s, N_e) = \inf_E \text{LNR} - \inf_{E, N_s} \text{LNR}. \quad (14)$$

The 68% CL belt on X is the set of parameters that satisfy to $\Delta_X \leq \Delta_{X,0}$, which $\Delta_{X,0}$ a threshold value determined by toy Monte-Carlo in order to ensure the correct coverage. However, although the quantities Δ_E and Δ_{N_s} do not depend on N_s and E respectively, their statistic laws however do. Consequently, the threshold levels $\Delta_{X,0}$ depend on the true values of both parameters E and N_s and furthermore on N_e . In order to deal with the additional nuisance parameter, N_s or E , we use the plugin method for the construction of the confidence belt. That is to say, we assume as the truth the most likely parameter value, the one minimising the LNR.

2. Technical issues

The LNR is a priori not a straightforward function to minimise. It is continuous only by steps with E in $[0; 1]$. By moving the centre of the bracketing interval from 0 to 1, the number of events, n_i , within the bracket

varies by ± 1 each time a new event enters or exits the bracket. At these positions the bracket probabilities, and so the LNR, are discontinuous. There are at most $2N_e$ of these discontinuities. However, one can check that the continuous segments between these points are monotonic. Therefore, the minimum of the LNR, as E , is necessarily reached as the limit from the right or from the left at one of the discontinuities. Therefore, in order to minimise the LNR, as E , it is enough to evaluate the limits at the discontinuities. This takes at most $4N_e$ function evaluations. With typical values of $N_e \simeq 1000$ this method is still quite CPU consuming, but it provides a robust result for the minimisation. The minimisation with N_s is more straightforward. For a fixed E the LNR has a single minimum reached in the neighbourhood of $\mu_s = (n_i - N_e p_{b,1}) / (p_{s,1} - p_{b,1})$. Therefore it is enough to evaluate the LNR at a few integer values of N_s around μ_s . The absolute minimum of the LNR is obtained by combining the two latter properties.

Despite the latter optimisations, the determination by toy Monte-Carlo of the threshold values, $\Delta_{X,0}$, is very CPU consuming. Therefore, we tabulated the results for various values of E , N_s and N_e spanning the ranges $[0; 1]$, $[1; 350]$ and $[500; 1500]$. For the analysis of the BC2 data the threshold values are then interpolated back from the tabulated values.

A last issue is that the confidence belts we obtain for E do not necessarily resume to a single interval. They can be an union of several disjoint intervals. However, the challenge does not allow to produce such results as outcome. Therefore, whenever this happens we define as the confidence belt the smallest interval that contains our results. Consequently we can expect our confidence belts on E to be over-conservative.

3. Converting N_s to a signal rate

The challenge actually requires to provide a confidence belt for a signal rate, D , not for the number of signal events, N_s . But, to our understanding it is possible to estimate only the ratio of the signal to the background rates. The absolute scale seems to be ill defined. In order to nevertheless provide an estimate for D , we apply the following conversion rule:

$$\frac{D}{A} = \frac{N_s}{N_e - N_s}, \quad (15)$$

with $A = 10000$, the quoted average background rate.

D. Results

The algorithms described previously were implemented in C++ as a toolbox class. The full code is available for download from <http://c1rwww.in2p3.fr/lhcb/bc2/bc2prob1.tar.gz>.

TABLE I. Statistics for test bench cases. Are indicated the power, β , and the coverages, α_E and α_{N_s} , for parameters E and N_s .

	0.2	0.5	0.9
E			
D	1010	137	18
N_s	92	14	2
β (%)	100	87	12
α_E (%)	78	79	63
α_{N_s} (%)	64	58	9

1. Test bench cases

We computed the power, β , of the test for a signal evidence for the 3 test bench cases: $(D, E) = (1010, 0.1)$, $(137, 0.5)$ and $(18, 0.9)$. For these 3 cases, we also checked the coverage of the confidence belts on parameters E and D obtained with the plugin method. We assumed $N_e = 1000$ total events and computed N_s according to Eq. 15. The results obtained are summarised in Table I. The power varies from $\sim 100\%$ at $(D, E) = (1010, 0.1)$ down to 12% at $(18, 0.9)$. For the 2 first cases, $(D, E) = (1010, 0.1)$ and $(137, 0.5)$ the confidence belts on E tend to overcover E and undercover N_s . The overcoverage on E was expected since we resumed the belt to a single overcovering interval, whereas it is in fact an union of disjoint intervals. In the last case, $(D, E) = (18, 0.9)$, the confidence belt on N_s is way too optimistic. Note that increasing the signal rate while keeping $E = 0.9$ results in a more accurate confidence belt on N_s .

2. BC2 data analysis

The analysis was ran on the $2 \cdot 10^4$ data-sets provided for the BC2 challenge. The results, in the BC2 format, is available from <http://clrwww.in2p3.fr/lhcb/bc2/bc2prob1.out.gz>.

We found evidence for a signal for 11% of the data sets, which includes $\sim 1\%$ of background events from false alarm. For each signal candidate we built a confidence belt on the parameter values E and D . The results obtained are summarised as a 2D map in Figure 1. The map reads as a superposition of the individual confidence belts for the signal candidates. One sees that candidates are distributed as spots along a curve in the (E, D) parameter space. In particular, two hot spots are visible, at $(E, D) = (0.5, 160)$ and $(0.1, 1300)$. One also sees that the last spot at $(E, D) = (0.9, 40)$ is badly resolved along N_s . This is consistent with the results of Table I.

3. Discussion

The results we obtained seem to indicate that our algorithm is less effective for a signal located at high E than at lower values. The bracketing interval was indeed op-

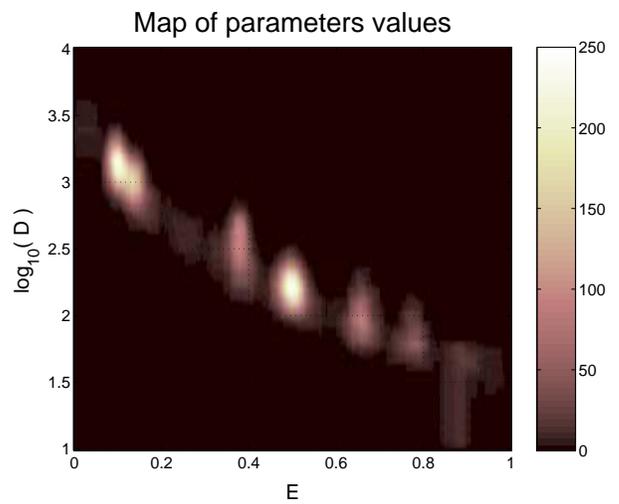


FIG. 1. Map of signal parameter values. The intensity on the 2D map reads as the number of signal candidates for which the individual confidence belts include the point of coordinates (E, D)

timised for S/\sqrt{B} assuming a significant contamination from the noise within the bracket. However, this criterion is no more relevant at high marks values where the background events are scarce. Therefore, we tried to vary the half width Δ of the bracketing interval. It was seen that for the test bench point $(E, D) = (0.9, 18)$ the power of the test only slightly increases with the bracket length, up to 15% for $\Delta = 3\sigma$. However, simultaneously the power then decreases to 83% for signals at $(E, D) = (0.5, 137)$. Therefore, although there might be room for an optimisation of the bracket length Δ varying with E , the improvement one would get seems to be only mild.

II. PROBLEM 2: A MONTE-CARLO PARAMETRISED EXAMPLE

As for problem 1, the observation is a data-set of ~ 1000 events for which one measured the marks, $x \in [0; 1]$. But now the events are distributed according to 3 processes, called signal, background 1 and background 2. Furthermore, we do not have an exact parametrisation of the statistic laws of the 3 processes, but a representative set of 5000 Monte-Carlo events, for each. Again, the challenge is to build the most powerful method to detect any signal event in the data-set for a false alarm rate smaller than 1%, and to provide an estimate of the number of signal events, N_s .

The analysis is based of the Kolmogorov-Smirnov statistic (KSS). First we draw a parametrisation of the signal and noise cumulative density functions (cdf) from the Monte-Carlo samples. Then we compare the data-set empirical distribution functions (edf) to these parametrisation for various hypothesis on the fractions of signal and background events.

TABLE II. Exponents for the parametrisation of the Monte-Carlo distribution functions.

	Backg. 1	Backg. 2	Signal
α_i	0.402	1.008	4.744
D_{N_e} (%)	1.21	0.59	0.86
p - value (%)	46	99	86

A. Parametrisation of the signal and backgrounds statistic laws

Let us assume a set of N_e events with marks x_i , $i \in [1; N_e]$. We write the empirical distribution function of this set as:

$$F_{N_e}(x) = \frac{1}{N_e} \sum_{i=1}^{N_e} H(x - x_i), \quad (16)$$

with H the Heaviside step function, as $H(x) = 0$ for $x < 0$ and $H(x) = 1$ elsewhere. The empirical distribution function converges to the cdf, F , as N_e goes to infinity. A measurement of the agreement between both distributions is given by the Kolmogorov-Smirnov statistic. It writes as:

$$D_{N_e} = \sup_{x \in [0;1]} |F(x) - F_{N_e}(x)|. \quad (17)$$

Numerically, the KSS statistic is easily computed by sorting the marks values x_i in ascending order. The supremum in Eq. 17 is achieved as the limit on the right or on the left at one of the values $x = x_i$ taking the value $|F(x_i) - i|$ or $|F(x_i) - (i - 1)|$. The signal and background cdf were found to be well approximated by simple power law functions, the like:

$$F(x) = x^{\alpha_i}, \quad (18)$$

with exponent values α_i fitted in order to minimise the KSS. The estimates for the exponents, as well as the corresponding KSS p-value are listed in table II. Figure 2 illustrates the agreement between the power law parametrisation and the Monte-Carlo data.

B. Evidence for a signal

1. Test statistic

In order to claim evidence for a signal we compare the KSS for the two hypothesis: $H_0 : \{\text{background only}\}$ and $H_1 : \{\text{signal+background}\}$. Therefore, let us write f_1 , f_2 and f_s the fractions of background 1, background 2 and signal, with $f_1 + f_2 + f_s = 1$. The corresponding cdf writes:

$$F(x; f_1, f_2, f_s) = f_1 x^{\alpha_1} + f_2 x^{\alpha_2} + f_s x^{\alpha_s}, \quad (19)$$

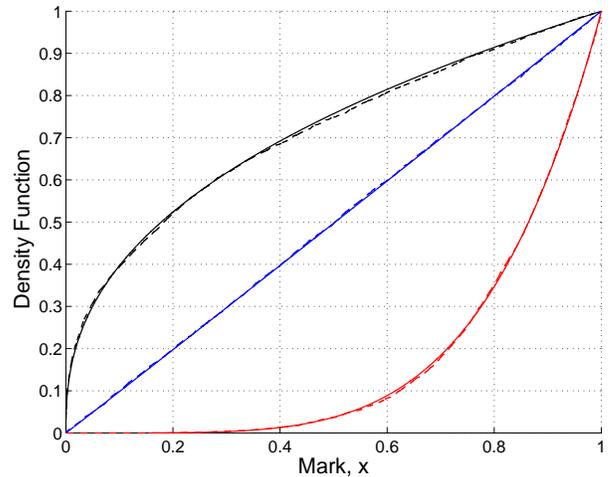


FIG. 2. Processes distribution functions. From the upper to the lower the curves correspond to background 1 (black), background 2 (blue) and signal (red). The solid curve stands for the cdf parametrisation and the dashed one for the Monte-Carlo edf.

where the exponents values α_i were given in Table II. We further define the Δ KSS statistic as:

$$\Delta D_{N_e}(f_s) = \inf_{f_1, f_2} D_{N_e} - \inf_{f_1, f_2, f_s} D_{N_e}, \quad (20)$$

where for the KSS statistic D_{N_e} one assumes the fractions f_i for the cdf and where the minimisation's are subject to the conditions $f_i \in [0; 1]$ and the sum of f_i being equal to 1. Following, the value of $\Delta D_{N_e}(f_s = 0)$ provides a comparison of the hypothesis H_0 and H_1 . Whenever there is a significant improvement in the KSS by assuming H_1 instead of H_0 , that is to say $\Delta D_{N_e}(f_s = 0)$ high enough, we claim evidence for a signal. We tried 2 other test statistics, the KSS pdf and cdf. They were found to be less or as powerful while requiring more intensive computations.

Practically, the minimisation of the KSS with respect to the fractions of signal and background is done numerically with the DMNFB function of the PORT library (<http://www.netlib.org/port/>). We use a penalty term in order to ensure the condition $\sum f_i = 1$. Note that the determination of the absolute minimum, with respect to f_s , suffers from local minima.

2. Test critical value

The number of background events in a set, N_1 and N_2 , are Gaussian distributed with known momentum, ($\mu_1 = 900, \sigma_1 = 90$) and ($\mu_2 = 100, \sigma_2 = 100$), and truncated to positive integer values. Following, significantly high values of N_e -whatever the marks distribution- are against the background only hypothesis. However, we decided not to use this information as an indication against H_0 because we do not want to accept tail background

events even though the distribution shape exhibits no indication for a signal. Therefore, we build p-values given the total number of background events, N_b , with: $N_b = N_1 + N_2 = N_e - N_s$, and where N_s is the assumed number of signal events in the set.

For a given total number of background events, N_b , it is enough to let only N_1 be Gaussian distributed, truncated to $[0; N_b]$, but with parameters values (μ_b, σ_b) given by a weighted mean, as:

$$\mu_b = \frac{\mu_1 \sigma_2^2 + (N_b - \mu_2) \sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \quad (21)$$

$$\sigma_b = \frac{\sigma_1 \sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \quad (22)$$

The number of background 2 events falls from $N_2 = N_e - N_1 - N_s$. Under $H_0 : N_s = 0$, we tabulated by toy Monte-Carlo the critical values $C_{99}(N_e)$ of the Δ KSS for various values of N_e and for a false alarm rate lower than 1%. Following, we claim evidence for a signal, provided that $\Delta D_{N_e}(f_s = 0) > C_{99}(N_e)$.

C. Confidence belt on N_s

The confidence belt on N_s is built from the Δ KSS, as for the test of the background only hypothesis. The value N_s is accepted in the confidence belt provided that the KSS for the hypothesis $f_s = N_s/N_e$ is close enough to the global minimum of D_{N_2} . That it to say: $\Delta D_{N_e}(f_s) \leq C_{68}(N_e, N_s)$. The critical values $C_{68}(N_e, N_s)$ are tabulated by toy Monte-Carlo, given the total number of background events $N_b = N_e - N_s$, in order to ensure a 68% coverage.

D. Results

As for problem 1, the algorithms for the analysis of problem 2 data were implemented in C++ as a toolbox class. The code is available from <http://clrwww.in2p3.fr/lhcb/bc2/bc2prob2.tar.gz>.

1. Test power

From toy Monte-Carlo we computed the power of the test for detecting a signal for various values of the number of events N_s . The background was generated without any assumption on the total number of events. The results obtained are shown on Figure 3. In the limit N_s goes to 0 we recover the 1% false alarm rate. For the test bench case $N_s = 75$ we have a power of $\beta(75) = 76.1 \pm 0.1\%$.

The results for the $2 \cdot 10^4$ data-sets provided for the BC2 challenge can be downloaded from <http://clrwww.in2p3.fr/lhcb/bc2/bc2prob2.out.gz>.

We found evidence for a signal for 9% of the data sets, including $\sim 1\%$ of background events from false alarms. For each signal candidate we built a confidence belt on the parameter value N_s . The frequencies of the obtained values of N_s are shown in Figure 4. About half of the signal candidates are compatible with a number of events in the range 100 – 120. There are almost no candidates compatible with $N_s \leq 30$, whereas the power of the test is still of $\beta(30) = 15.0\%$.

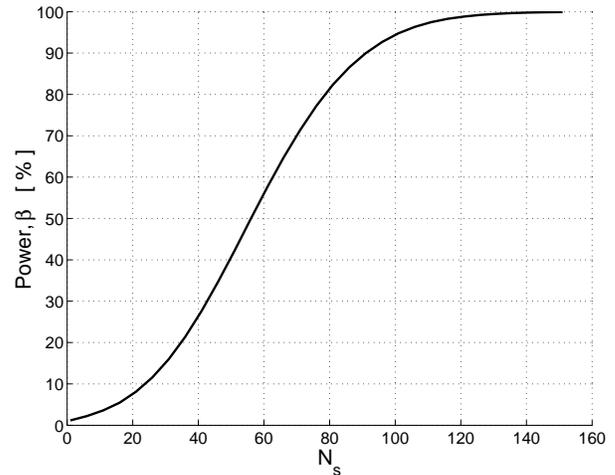


FIG. 3. Power of the test for a signal as a function of the number of signal events N_s .

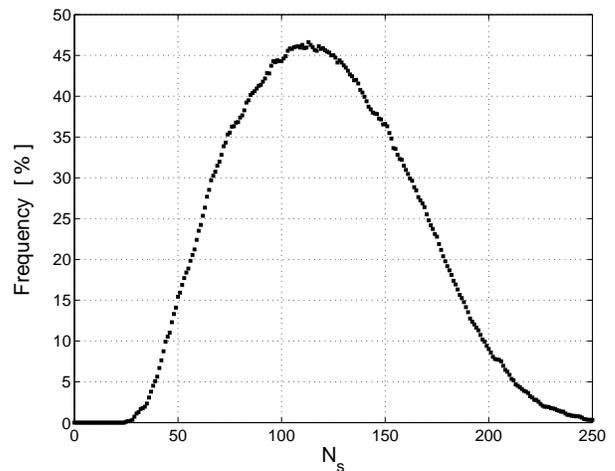


FIG. 4. Frequency of N_s values. The frequency reads as the fraction of signal candidates for which the confidence belt includes the value N_s .