arXiv:1106.6296v1 [physics.data-an] 30 Jun 2011

Calculating a confidence interval on the sum of binned leakage

I. Ruchlin[a,b,1,2,3,*], R.W. Schnee[a]

[a] *Department of Physics, Syracuse University, Syracuse, New York 13244, USA*
[b] *Center for Computational Relativity and Gravitation, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, New York 14623, USA*

Abstract

Calculating the expected number of misclassified outcomes is a standard problem of particular interest for rare-event searches. The Clopper-Pearson method allows calculation of classical confidence intervals on the amount of misclassification if data are all drawn from the same binomial probability distribution. However, data is often better described by breaking it up into several bins, each represented by a different binomial distribution. We describe and provide an algorithm for calculating a classical confidence interval on the expected total number of misclassified events from several bins, based on calibration data with the same probability of misclassification on a bin-by-bin basis. Our method avoids a computationally intensive multidimensional search by introducing a Lagrange multiplier and performing standard root finding. This method has only quadratic time complexity as the number of bins, and produces confidence intervals that are only slightly conservative.

Keywords: confidence interval, likelihood ratio, binomial distribution

1. Introduction

Many real-world processes can assume one of two possible outcomes; each independent trial or observation can be classified as either “success” or “failure,” with the probability of success p. All such trials can be bundled together to form a single experiment with x successes out of a total of n trials. If the experiments are repeated many times, the relative frequency of successes in each experiment follows the binomial distribution (*e.g.* [1]).

For a measurement of x and n, the best estimate $P = x/n$ of the true success probability p can be calculated. Since the ratio $P/(1-P)$ is the best estimate

[*] Corresponding author
Email addresses: `ixr5289@rit.edu` (I. Ruchlin), `rwschnee@phy.syr.edu` (R.W. Schnee)
[1] Current address is at Rochester Institute of Technology.
[2] Phone: +1 585 475 2498
[3] Fax: +1 585 475 7340

of the expected ratio of successes to failures, the best estimate of the number of successes Y of a second experiment that has the same success probability and a known number of failures b is

$$Y = \frac{P}{1-P} b \, . \tag{1}$$

Furthermore, the Clopper-Pearson method [2] provides a (frequentist, or classical) confidence interval $[P_{\text{low}}, P_{\text{high}}]$ with probability content β such that the fraction of experiments with $P_{\text{low}} \le p \le P_{\text{high}}$ is $\ge \beta$ (with the inequality due to the discrete nature of binomial distribution; see *e.g.* [3]). By extension, this method may also be used to calculate the confidence interval $[Y_{\text{low}}, Y_{\text{high}}]$ on the expected number of successes of the second experiment.

Such estimates may be particularly useful for characterizing backgrounds for rare-event searches. A given background event may have some probability p to be misclassified as a signal event. A first, "calibration" experiment may allow estimation of p based on the number of events n and the number x misclassified as signal (the "leakage"). A second, "search" experiment may provide a measurement of the number of correctly identified background events b. If background events in both experiments have the same probability of correct classification, the expected number of misclassified events Y and a confidence interval $[Y_{\text{low}}, Y_{\text{high}}]$ on the expected number may be determined.

Often, however, in order for events in the calibration and search both to have the same probability of misclassification p, events with different characteristics (*e.g.* energy, position, detector, or pixel) must be considered separately, resulting in m separate bins of events for both calibration and search. In the i^{th} calibration bin there are x_i misclassified events out of the total n_i calibration events, resulting in a best estimate $P_i = x_i/n_i$ of the misclassification probability for events in that bin. For the search data, the number of correctly classified events in the i^{th} bin, b_i, is known. If the true misclassification probability, p_i, of an event in the i^{th} bin is the same for both calibration and search data, the best estimate for the total expected number of misclassified events Y is

$$Y = \sum_i^m \frac{P_i}{1-P_i} b_i \equiv f(\vec{P}) \, , \tag{2}$$

where $\vec{P} \equiv \{P_i\}$.

The likelihood $\mathcal{L}$ that x_i events out of the total n_i calibration events in each bin are misclassified is simply the product of the binomial probabilities, with

$$\mathcal{L} \propto \prod_i^m P_i^{x_i} (1-P_i)^{n_i - x_i} \, . \tag{3}$$

The global maximum $\hat{\mathcal{L}}$ of the likelihood is trivially given by the set $\hat{\vec{P}} \equiv \{\hat{P}_i\} = \{x_i/n_i\}$ for all i. Note that it is possible to estimate the expected leakage only if no calibration bin has zero total events (i.e. $n_i \neq 0$ for all i). Substitution of the set $\{\hat{P}_i\}$ into Equation 2 yields the most likely value of the total expected leakage $\hat{Y}(\hat{\vec{P}})$.

Unfortunately, for the case with multiple bins, the Clopper-Pearson method cannot be used to calculate a confidence interval on the total expected leakage. Here we describe a method and provide a practical algorithm for this problem. We use the "Unified Approach" of Feldman and Cousins [4] extended to deal with nuisance variables by means of the profile likelihood [5] without the large-sample approximation used in *e.g.* [1, 6]; here $\vec{P}$ are nuisance variables since they are unknown variables for which we are not setting a confidence interval.

2. Method

For every considered value of Y, we calculate the profile likelihood

$$R_0 = \frac{\mathcal{L}\left(\tilde{\vec{P}}(Y)\right)}{\hat{\mathcal{L}}\left(\hat{\vec{P}}\right)} , \tag{4}$$

where $\tilde{\vec{P}}$ is the combination of P_i, found by a search described in Sect. 2.1, that maximizes the likelihood for the value of Y under test. Asymptotically (and far from physical boundaries), the distribution of $-2\ln(R_0)$ is χ^2-distributed with 1 degree of freedom, but more accurate results may be obtained by determining the expected distribution by Monte Carlo simulation. For each simulated experiment, $\vec{x}$ is randomly determined based on the $\tilde{\vec{P}}$ found above. For each, the best-fit values are found for $\hat{\vec{P}}_{\mathrm{MC}}$ and $\tilde{\vec{P}}_{\mathrm{MC}}$, and then the ratio $R_{\mathrm{MC}} \equiv \mathcal{L}\left(\tilde{\vec{P}}\right) / \hat{\mathcal{L}}\left(\hat{\vec{P}}\right)$ is calculated. If R_0 is larger than $1-\beta$ of the simulated R_{MC} ratios, then Y is included in the confidence interval of probability content β. Since the distributions follow the binomial distribution, the uncertainties $\propto 1/\sqrt{N_{\mathrm{MC}}}$, the inverse of the square root of the number of experiments. Thus, to achieve a relative tolerance t, conduct $N_{\mathrm{MC}} = t^{-2}$ Monte Carlo simulations. A root-finding algorithm hunts for the smallest and largest values of Y that are allowed in order to return the desired confidence interval $[Y_{\mathrm{low}}, Y_{\mathrm{high}}]$.

2.1. Formulation

A multiparameter function minimizer, such as MINUIT [7], could be implemented to hunt for the combination of probabilities P_i that maximize Equation 3 subject to the constraint of Equation 2. However, this method has exponential time complexity, and is not feasible for the analysis of more than a few bins. Instead, the combination of binomial probabilities that maximize Equation 3 subject to the constraint of Equation 2 can be found by introducing a Lagrange multiplier, λ, and solving

$$\frac{\partial}{\partial P_i}\left[\ln(\mathcal{L}(P_i)) + \lambda(f(\vec{P}) - Y)\right] = 0 ,$$

Table 1: Summarized analysis of the behavior of Equation 5.

	Positive Root	Negative Root
$\lambda < 0$	$P_i > 1$	$0 < P_i < x_i/n_i$
$\lambda = 0$	$P_i = 1$	$P_i = x_i/n_i$
$0 < \lambda < c_i$	$\sqrt{x_i/n_i} < P_i < 1$	$x_i/n_i < P_i < \sqrt{x_i/n_i}$

where Y is a given constant. A little algebra yields the solution to this equation for each bin i:

$$P_i = \frac{n_i + x_i - \lambda b_i \pm \sqrt{(\lambda b_i - n_i - x_i)^2 - 4n_i x_i}}{2n_i} \; , \tag{5}$$

while substituting back into Equation 2 yields an equation for the Lagrange multiplier:

$$Y = \sum_i^m b_i \frac{n_i + x_i - \lambda b_i \pm \sqrt{(\lambda b_i - n_i - x_i)^2 - 4n_i x_i}}{n_i - x_i + \lambda b_i \mp \sqrt{(\lambda b_i - n_i - x_i)^2 - 4n_i x_i}} \equiv \sum_i^m b_i Y_i \; . \tag{6}$$

Equation 6 is really 2^m separate equations, depending on the the signs of *each* ± term. One of the 2^m solutions yields the value of λ that gives the most likely combination of binomial probabilities (i.e. $\tilde{\vec{P}}$) for the desired total expected leakage Y. Fortunately, further analysis reveals a significant reduction in the number of viable solutions.

For any bin with $b_i \neq 0$,

$$\lambda > c_i \equiv \frac{n_i + x_i - 2\sqrt{n_i x_i}}{b_i}$$

is unphysical, producing imaginary or negative probabilities. Since λ must be physical for all bins, λ is required to be less than or equal to the smallest c_i, *i.e.* $\lambda \leq \inf\{c_i\} \equiv \lambda_c$. For any bin with $b_i = 0$, $c_i \to \infty$ so that it places no constraint on λ.

Table 1 lists the different limiting values of λ and their corresponding values of P_i from Equation 5. The lower limit on the confidence interval must have $P_i \leq x_i/n_i$ for each bin. Therefore, Table 1 indicates that the solution must use the negative root for each bin, reducing the problem from searching among 2^m solutions to solving a single equation.

It is easiest to understand the viable solutions for the confidence interval's upper bound by first noting that, other than the constraint of Equation 2, each term in

$$\ln(\mathcal{L}) = \sum_i^m \ln(\mathcal{L}_i)$$

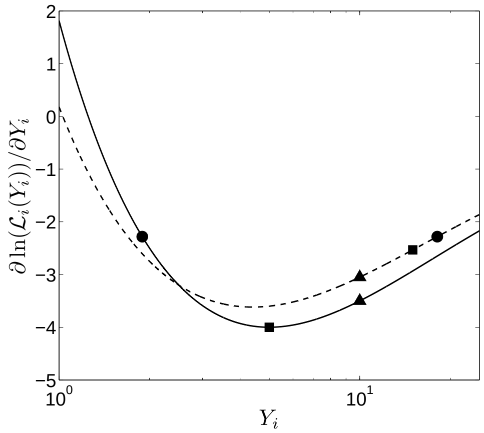

Figure 1: Visualization to demonstrate that the total likelihood is not maximized if more than one bin uses the positive root in Equation 2. Any points to the right of each curve's minimum are given by the positive root, while points at or to the left of the minimum are given by the negative root. Suppose the desired total expected leakage is $Y = 20$, with one possible combination, $Y_1 = Y_2 = 10$, shown as the triangles. Since $\partial \ln(\mathcal{L})/\partial Y_i$ must increase with increasing Y_i for both curves to the right of their minima, a more likely combination also resulting in $Y = 20$ can be found by decreasing the expected leakage in the bin with the more negative value of $\partial \ln(\mathcal{L})/\partial Y_i$ to the minimum while increasing the expected leakage in the other (squares). The maximum of the likelihood may be found by continuing to decrease the expected leakage in the bin with the more negative value of $\partial \ln(\mathcal{L})/\partial Y_i$ until both bins have equal values of $\partial \ln(\mathcal{L})/\partial Y_i$ (circles).

is independent. For each term, $\ln(\mathcal{L}_i)$ decreases monotonically with increasing $P_i > \hat{P}_i$, and there is an inflection point at $P_i = P_i(\lambda = c_i)$, with $\ln(\mathcal{L}_i)$ decreasing ever more slowly for larger P_i.

For any bins i and j, it can be shown that

$$\left.\frac{\partial \ln(\mathcal{L})}{\partial Y_i}\right|_{P_i=\tilde{P}_i} = \left.\frac{\partial \ln(\mathcal{L})}{\partial Y_j}\right|_{P_j=\tilde{P}_j} .$$

This relation is to be expected. If, instead, the left term were larger (smaller) than the right term, a more likely combination with the same total Y could be found by decreasing Y_i (Y_j) and increasing the other by the same amount.

In a similar way, it may be shown that the combination of probabilities $\tilde{\vec{P}}$ that maximize the likelihood for a given total expected leakage Y never includes more than one bin with $P_i > P_i(\lambda = c_i)$, and hence more than one bin using the positive root of Equation 2. Figure 1 helps visualize the reasoning. Consider any two bins i and j that both use the positive roots and hence have $P_i > P_i(\lambda = c_i)$, $P_j > P_j(\lambda = c_j)$. If $\partial \ln(\mathcal{L})/\partial Y_i|_{Y_i=k} \geq \partial \ln(\mathcal{L})/\partial Y_j|_{Y_j=k}$ then $\partial \ln(\mathcal{L})/\partial Y_i|_{Y_i=k+\delta} \geq \partial \ln(\mathcal{L})/\partial Y_j|_{Y_j=k-\delta}$ for $\delta = k - Y_j(c_j)$, since $\partial \ln(\mathcal{L})/\partial Y_i$ becomes larger as Y_i is increased, and $\partial \ln(\mathcal{L})/\partial Y_j$ becomes smaller as Y_j is decreased towards the inflection point. As a result, for a given total leakage, the likelihood is never maximized by having two bins use the positive roots of Equation 2. Since a multi-bin system can be examined in 2-bin pieces, it follows that it is necessary to solve Equation 6 only for $m+1$ cases: the case with all negative roots, and also the m cases with one positive root and $m-1$ negative roots. Thus, the total number of equations to be solved is reduced from 2^m to $m+1$.

A standard root-finding algorithm may be used to hunt for a physical solution to each of these $m+1$ equations. Our implementation uses the Van Wijngaarden-Dekker-Brent Method [8] which combines the secant method, bisection method, and inverse quadratic interpolation to find bracketed roots of a given function—in this case, appropriate λ for a given Y.

2.2. Benchmarking

All tests were conducted with an Intel Core i5 M430 processor and 6 GB of RAM. Since each additional bin increases the length of all of the data storage vectors by one element, the RAM usage increases linearly with the number of bins. The estimated RAM usage is 16.69 ± 0.01 kB per bin. For these tests, five measurements of each quantity were taken and their averages were plotted. The error bars represent the standard deviations of those means.

For m bins, Equation 2 consists of m terms that must be calculated for a given value of λ. To find both the upper and lower limits, $m+1$ equations each with m terms must be solved. Therefore, finding the limits of the confidence interval as a function of the number of bins has time complexity $\mathcal{O}(m^2)$, as shown in Figure 2. The number of Monte Carlo simulations is inversely proportional to the square of the desired relative tolerance. Therefore, as tolerance decreases, the number of basic operations increases polynomially, as shown in Figure 2.

2.3. Coverage Tests

By construction, the confidence interval includes the true expected leakage a fraction β of the time if the true leakage probabilities $\vec{p} = \tilde{\vec{P}}$, *i.e.* the method provides correct "coverage" for the most likely combination of probabilities for each total expected search leakage Y. For other values of the true leakage probabilities, the coverage is not exact. However, since $\tilde{\vec{P}}$ is the combination of nuisance parameters that maximizes the likelihood for the data, it may be expected that the coverage is nearly correct for other combinations of true leakage probabilities, and that there is usually slight over-coverage (*e.g.* a claimed 90% confidence interval may contain the true parameters slightly more than 90% of the time for some combinations of parameters).

To test this algorithm for coverage, several simulated experiments were performed. For each set of tests, values of $\vec{n}$, $\vec{b}$, and $\vec{p}$ were chosen as listed in Table 2 or in the caption to Figure 3. The calibration leakage $\vec{x}$ was simulated from the binomial distribution, given $\vec{n}$ and $\vec{p}$, and the program found the most likely total expected leakage and corresponding 90% confidence interval. For each set of tests, this simulation was repeated 10000 times in order to determine what fraction of the intervals contained the true expected leakage. The results, shown in

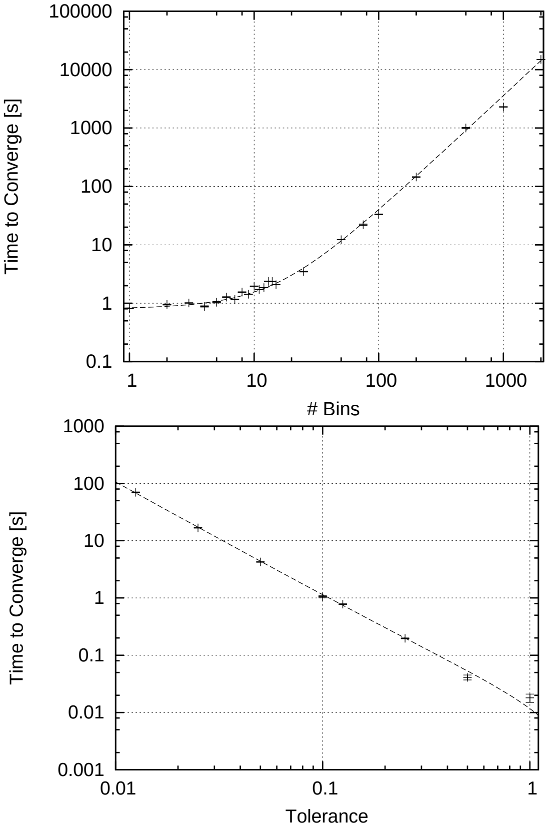

Figure 2: *Top*: Run time as a function of the number of bins m, each with $n_i = 100$ calibration events, $x_i = 5$ leaking calibration events, and $b_i = 10$ correctly tagged events from the search data, for tolerance $t = 0.1$. The data fit time complexity $\mathcal{O}(m^2)$. *Bottom*: Run time as a function of tolerance t, for $m = 5$ bins of data identical to that used in the *top* plot. The data fit time complexity $2^{-\mathcal{O}(\log(t))}$, polynomial time.

Figure 3 and Table 2, suggest that the algorithm produces confidence intervals with approximate coverage and slight conservatism, as expected.

3. Application to Dark Matter Searches

Many collaborations around the world are attempting to make direct detections of dark matter particles (see *e.g.* [9–11]). Dark matter is hypothesized to constitute the majority of the universe's mass, in an effort to explain many astrophysical observations [12]. Since dark matter particles interact very weakly with normal matter, dark matter detectors must be very sensitive, which makes the rejection of background interactions an important task.

The CDMS II collaboration recently published results in which two candidate dark matter events appeared in their signal region. However, with an expected background of 0.9±0.2 events, it was not statistically significant evidence of dark

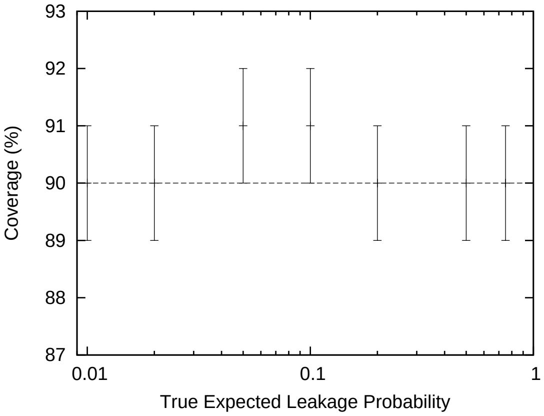

Figure 3: The percentage of simulated 90% confidence intervals that include the true expected leakage given the probability that an event will leak. Each test uses 3 bins, with $n_i = 1000$, $b_i = 10$, and x_i thrown from a binomial distribution given n_i and the true expected leakage.

Table 2: Total number of calibration events $\vec{n}$, expected calibration leakage $\langle\vec{x}\rangle$, number of correctly tagged search events $\vec{b}$, and percentage of the time that the true expected search leakage was in the resulting 90% confidence interval, for the various cases described.

$\vec{n}$	$\langle\vec{x}\rangle$	$\vec{b}$	% In	Description
$(10^3, 10^3)$	$(1, 100)$	$(1, 100)$	93 ± 1	bin with large $\langle x_i\rangle$, b_i
$(10^3, 10^3, 10^3)$	$(500, 5, 5)$	$(10, 10, 10)$	90 ± 1	bin with large $\langle x_i\rangle$
$(10, 10^3, 10^3)$	$(5, 5, 5)$	$(10^4, 10, 10)$	89 ± 1	bin with $n_i \ll b_i$
$(10^5, 10^3, 10^3)$	$(5, 5, 5)$	$(10, 10, 10)$	90 ± 1	bin with $n_i \gg b_i$
$(10^3, 10^3, 10^3)$	$(100, 50, 30)$	$(10, 10, 10)$	91 ± 1	wide range of $\langle\vec{x}\rangle$

matter particle detection [13]. The background estimate was calculated with a Bayesian technique [14] based on three independent "calibration" data sets, two of which are too complicated to be considered here. The third "calibration" set (see Table 3) provided a background estimate only for those detectors not on the top or bottom of a detector stack; this background estimate totaled 0.51 ± 0.27 events. The same data, when analyzed by the method described above, give a total expected leakage of $0.54^{+0.41}_{-0.20}$ events, suggesting the CDMS II estimate may be slightly under-covered. A visual representation of the upper and lower bound solutions are shown in Figure 4.

4. Discussion

The method described here solves the problem of calculating confidence intervals on leakage for several bins. It has quadratic time complexity, so that

Table 3: Data from the final run of the CDMS II experiment. Each detector is taken to be a single bin. Two bins (the end-cap detectors T3Z6 and T4Z6) have been excluded.

Detector Designation	n_i	x_i	b_i
T1Z2	28	0	15
T1Z5	49	0	8
T2Z3	45	0	8
T2Z5	67	2	9
T3Z2	50	0	2
T3Z4	48	0	7
T3Z5	29	0	4
T4Z2	41	0	5
T4Z4	31	0	6
T4Z5	44	1	6
T5Z4	59	0	6
T5Z5	49	1	6

even problems with a tremendous number of bins may be handled. Furthermore, it produces intervals that are only slightly conservative, but exhibit full coverage.

A more rigorous and detailed discussion of the topics presented in this paper has been compiled into a technical note, along with a fully coded version of the algorithm [15].

5. Acknowledgements

This work is supported in part by the National Science Foundation (Grant No. PHY-0855525). We thank Jeff Filippini for his comments.

References

[1] G. Cowan. *Statistical Data Analysis.* Clarendon Press, 1998.

[2] C. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–143, 1934.

[3] F. James. *Statistical Methods in Experimental Physics.* World Scientific, 2^{nd} edition, 2006.

[4] G. J. Feldman and R.D. Cousins. A unified approach to the classical statistical analysis of small signals. *Phys. Rev. D.*, 57:3873–3889, 1998.

[5] G. J. Feldman. Concluding Talk: Physics. In L. Lyons & M. Karagöz Ünel, editor, *Statistical Problems in Particle Physics, Astrophysics and Cosmology*, pages 283–+, 2006.

[6] Wolfgang A. Rolke, Angel M. Lopez, and Jan Conrad. Confidence Intervals with Frequentist Treatment of Statistical and Systematic Uncertainties. *Nucl. Instrum. Meth.*, A551:493–503, 2005.

[7] F. James and M. Roos. Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations. *Comput. Phys. Commun.*, 10:343–367, 1975.

[8] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing.* Golden Cup, 3^{rd} edition, 2007.

[9] N. J. C. Spooner. Direct Dark Matter Searches. *Journal of the Physical Society of Japan*, 76(11):111016–+, November 2007.

[10] Gianfranco Bertone, editor. *Particle Dark Matter: Observations, Models and Searches.* Cambridge University Press, Cambridge, UK, 2010.

[11] R. W. Schnee. Dark matter experiments. In C. Csaki and S. Dodelson, editors, *Physics of the Large and Small: Proceedings of the 2009 Theoretical Advanced Study Institute in Elementary Particle Physics*, Singapore, 2010. World Scientific.

[12] E. Komatsu *et al.* Five-year wilkinson microwave anisotropy probe observations: Cosmological interpretation. *Astrophys. J.*, 180(330), 2009.

[13] The CDMS II Collaboration. Dark Matter Search Results from the CDMS II Experiment. *Science*, 327(5973):1619–1621, 2010.

[14] J. P. Filippini. *A Search for WIMP Dark Matter Using the First Five-Tower Run of the Cryogenic Dark Matter Search,*. PhD thesis, University of California, Berkeley, 2008.

[15] I. Ruchlin. Software and technical note on calculating a confidence interval on the sum of binned leakage. http://cdms.syr.edu/leakage_limit.html.

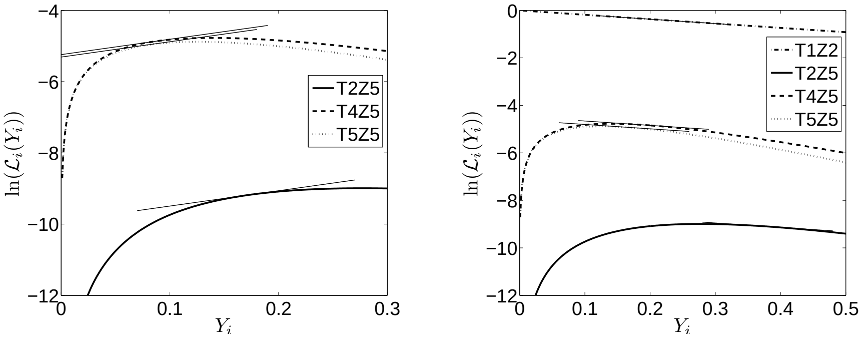

Figure 4: Plot of the likelihood $\mathcal{L}_i$ of the expected leakage Y_i in each bin versus the expected leakage in each bin for the CDMS Run 125 c58. Of the twelve bins under consideration, only the three with $x_i > 0$ (T2Z5: solid, T4Z5: dashes, and T5Z5: dots) have non-zero expected leakage solutions for the lower limit and only one additional bin (T1Z2: dash-dots) has $Y_i > 0$ for the upper limit. The straight lines are tangent to each bin at the points corresponding to the most likely combination of expected leakages, yielding the 68% lower (*left*) and upper (*right*) limits of the confidence interval. For the upper limit, the overall likelihood of the configuration is decreased the least by increasing the expected leakage in the detector with the worst statistics—T1Z2. As expected, the tangent lines are all parallel.