Choosing an estimate of initial activity level for multiple radioactive sources

Christopher Tong,[1] Alex Cardenas, and Richard F. Raubertas

(The authors present this work in their personal capacities, thus no affiliations are listed.)

[1]*Corresponding author email: ctong@mailaps.org*

(Dated: March 18, 2023)

Abstract

For brachytherapy procedures, a small number, n, of radioactive Cs-137 sources are stored together in a lead safe. Their initial activity levels are measured upon delivery from the supplier, and are all found to be within 5% of each other. The physicist "randomly" chooses m sources for an individual therapeutic procedure. To determine dosage, the activity at the time of use can be calculated from knowledge of the initial activity and the decay factor of Cs-137. However, the sources are visually indistinguishable, so the initial activity values for a specific selected subset of sources are unknown, and an estimate must be used. Since the sources have similar activity levels, there is no practical difference among any estimates from within the range of values of the n sources. Nonetheless, some estimate must be chosen for the calculation; the physicist is free to make the choice based on principle. The problem is solved using an optimization approach, either of the mean squared error or the mean absolute error. The former results in the arithmetic mean of the whole set, but the latter results in a new, m-dependent estimator we call the *combinatorial median*. Artificial data sets are used to illustrate the findings.

Keywords: brachytherapy, statistical analysis, Cs-137

I. INTRODUCTION

A hospital orders a set of Cs-137 radioactive sources for use in brachytherapy procedures. These sources have different initial activity levels, and are labeled by the supplier only with nominal activity values (say, 5, 10, 15, and 20 mg-Ra eq). Suppose that a small number of sources, n, is ordered for a given nominal activity. (In the illustrative examples presented below, we will take $n = 5$.) Upon receipt, the initial activity level of each source is measured and their values are all found to be within 5% of each other. All n sources are stored in the same lead safe, and for a given nominal activity they are visually indistinguishable. The sources will be re-used repeatedly over the next, say, fifteen years.

For a particular procedure, the physicist "randomly" chooses a subset of $m \leq n$ sources of a given nominal activity for use on a patient. (Sources from several nominal activities are usually selected; for simplicity, in this paper we consider only sources from one nominal activity used in a procedure. The methods to be described here can be applied separately to sources of each nominal activity selected.) Because of the high-throughput nature of hospital operations, the activities of the selected sources cannot be re-measured prior to each use. Since the sources are visually indistinguishable, the physicist does not know the exact initial activity levels for the "randomly" chosen subset. Therefore, an assumed value (which we will call the *estimate*) must be employed. This estimate of initial activity is used along with the decay factor of Cs-137 to calculate the exposure time needed to deliver the dose prescribed by the physician. Our objective is to find an *estimator*, a mathematical formula for producing such an estimate, for any given subset of sources.

Since the initial activity levels are all within 5% of each other, the error in picking any estimator, within the range of all n values, is physiologically negligible. The resulting error produces a difference in exposure time of at most a few minutes out of perhaps dozens of hours of total exposure. Thus, therapeutically any representative value within the range of n measured values could reasonably be used. Nonetheless, the radiation physicist must make a choice in doing the calculations: should the average of the n values be used, or can one do better? (Note that if the initial activity levels differ among each other by more than 5% , then it is essential to use the values specific to the selected sources.)

We assume that each possible subset of m sources is equally likely to be chosen. Since, as discussed above, there is essentially no practical difference between the alternatives, one

is free to make the choice based on principle. One principle would be to use a conservative estimator; for instance, either the smallest or largest value of the set, depending on whether one believes that too much or too little dosage is the more harmful error. The physicist could make this choice based on the specific circumstances of the case. In this note, however, we focus on the use of a "representative" value rather than the max or min. We will formulate and solve the problem using an optimization approach.

Note that the choice of initial activity value to use for exposure time calculations does *not* fit into the standard framework of statistical inference. In that framework, one uses a known sample of n data points to estimate the unknown mean of a larger population from which the sample was randomly drawn.[1] In our problem, the complete population of n Cs-137 sources at the hospital is known, but in a given therapeutic procedure, an unknown sample of $m \leq n$ sources will be used.

In section II we introduce several artificial data sets that will illustrate various possible distributions of source activity levels. In section III, we present the formal optimization framework and corresponding solutions. A brief section IV concludes the note.

II. ARTIFICIAL DATA SETS

To provide illustrations, we will consider four artificial data sets with $n = 5$ sources in each (see Table I). In all cases, the initial activities differ by less than 5% from each other. Set 1 has a relatively symmetric distribution; set 2 is a little skewed, due to an outlier; and set 3 is heavily skewed. Set 4 has a bimodal structure. For convenience, the initial activity values are listed in ascending order in the table. Two significant digits of precision after the decimal point will be reported throughout the sequel. We also report the estimates of four conventional estimators (arithmetic mean, median, mode, and midrange) in the table. The first three of these are typical location estimators often used as descriptive statistics for data sets[2]. The midrange is the arithmetic average of the smallest and largest values in the data set, and has generally poor statistical behavior. However it is included here since it has sometimes been entertained by physicists trying to solve the stated problem.

The data sets are visualized in Fig. 1a. The plot is reproduced in Fig. 1b but now with the various estimates also indicated. It is convenient to have both plots side-by-side so that the uncluttered data plot can be compared to the plot with the various estimates shown. All

TABLE I: Artificial data sets for illustration. All units are mg-Ra eq.

Data set	Initial activity levels	Mean	Median	Mode	Midrange
Set 1	20.39, 20.40, 20.42, 20.43, 20.45	20.42	20.42	n/a	20.42
Set 2	20.38, 20.40, 20.40, 20.42, 20.49	20.42	20.40	20.40	20.43
Set 3	20.37, 20.37, 20.38, 20.40, 20.57	20.42	20.38	20.37	20.47
Set 4	20.38, 20.39, 20.40, 20.50, 20.50	20.43	20.40	20.50	20.44

calculations and plots were implemented in the R software system[3] (version 2.6.1), an open source statistics system available at `http://www.r-project.org/`. The add-on R package `modest` was used to calculate the mode.

The artificial data sets are somewhat idealized and are designed to illustrate the issues under discussion. Readers are invited to try their own toy examples and evaluate the fitness of various estimators following the principles outlined in this note.

III. THEORY

A. Setup

Let $\{x_1, x_2, \ldots, x_n\}$ denote the initial activity levels of n sources. (Among these n values there may be ties, i.e., sources that have the same initial activity, within measurement error.) These data are the *population* from which a subset (or *sample*) of $m \leq n$ sources will be drawn to treat an individual patient. We assume that the therapeutically important aspect of the sample is the total radiation dose delivered to the patient. That is, given the constraint that source activities are all within 5% of each other, the variation in dose contributed by individual sources is assumed to be of no consequence; only their sum (or equivalently, their mean) is. Such an assumption is, of course, an approximation, since it ignores the spatial distribution of the sources, relative to the locations of the surrounding tissues. Nonetheless, the resulting framework leads to estimators with reasonable properties, without requiring an excursion into the spatial modeling of radiation fields.

Under the above assumption, the quantity of practical interest is the unknown mean of the selected sample of sources, and the question becomes how best to choose an estimator

for that unknown mean when the sample is taken blindly, without knowledge of the specific activities of the sources it includes.

B. Conventional estimators

In conventional statistical inference,[1] the arithmetic sample mean minimizes the *mean square error* (MSE), defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2, \tag{1}$$

where μ is any putative estimator. Similarly, in conventional inference the sample median minimizes the *mean absolute error* (MAE), defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|\,. \tag{2}$$

The remainder of this subsection reviews well-known properties of the mean, median, mode, and midrange.

The sample median is more resistant to extreme "outlier" values, at the tails of the distribution, than the sample mean, which more heavily penalizes errors at the extremes.[4] This is because, as a rank-based statistic, the sample median is influenced only by the ranks of extreme values, rather than the values themselves. However, the sample median can be "jumpy", as best illustrated by a bimodal distribution. Repeated sampling (with odd sample size) from such a distribution will result in the sample median sometimes falling near one mode, and sometimes near the other. The sample mean, as a "center of mass", will instead tend to fall between the two peaks. Incidentally, the initial activity estimation problem is one of the few scenarios where it is appropriate to describe a bimodal or multimodal distribution of data using a *single* estimate; more generally, the locations of each mode or peak should be reported. However, in this problem a single value is used to perform the exposure time calculation for any selected subset of sources.

The sample mode (most frequently occurring value) is not intended to represent the center of a distribution of data, but rather its peak(s). A given data set may actually have more than one mode, or a mode may not even exist (as in set 1). Finally the sample midrange ignores all but the smallest and largest values in the data set, and is thus even more sensitive to "outliers" than the sample mean. The midrange should be avoided, as it tends to be the least "representative" of the full data, among the conventional estimators discussed here.

C. Optimization framework and solutions

Recall that our problem is actually opposite to the one usually encountered in statistical inference. Typically we have a population with an unknown distribution of values. A sample is randomly drawn from the population and the values for the sample are determined. The goal is then to use the known sample values to estimate some characteristic (referred to as a *parameter*) of the unknown population distribution; e.g., the mean or median. Here, on the other hand, we have a known population distribution of activities $\{x_1, x_2, \ldots, x_n\}$, and the goal is to decide what parameter of that distribution should be used to estimate the mean of a random sample with unknown values. Thus the mean, median, mode, and midrange given in Table I are actually *population* statistics, rather than sample statistics. The corresponding sample statistics will never be known for a given blind sample.

The number of possible samples of size m from a population of size n is the binomial coefficient, ${}_nC_m = \frac{n!}{m!(n-m)!}$; all samples are equally likely under random sampling. Let $\bar{x}_i^*$ denote the mean activity for the i-th possible sample, for $i = 1, 2, \ldots, {}_nC_m$. Let θ be some parameter (to be determined) calculated from the population distribution $\{x_1, x_2, \ldots, x_n\}$. Then if θ is used to estimate $\bar{x}_i^*$, the mean square error (MSE) will be $\frac{1}{{}_nC_m} \sum_{i=1}^{{}_nC_m} (\bar{x}_i^* - \theta)^2$ and the mean absolute error (MAE) will be $\frac{1}{{}_nC_m} \sum_{i=1}^{{}_nC_m} |\bar{x}_i^* - \theta|$. To minimize MSE we should take $\theta = \text{mean}(\bar{x}_1^*, \ldots, \bar{x}_{{}_nC_m}^*)$, the average of all possible sample means. This turns out to be just the average of $\{x_1, x_2, \ldots, x_n\}$, which is simply the mean of the entire population.[5]

To minimize the MAE we should take $\theta = \text{median}(\bar{x}_1^*, \ldots, \bar{x}_{{}_nC_m}^*)$, the median of all possible sample means. We call this the *combinatorial median* or c-median. In general the c-median is not equal to the median of $\{x_1, x_2, \ldots, x_n\}$, although in the special case where the samples are single sources ($m = 1$), $\bar{x}_i^*$ is just x_i, so the c-median is the population median. At the other extreme where all n sources are sampled ($m = n$), ${}_nC_m = 1$ and $\bar{x}_1^* = \text{mean}(x_1, \ldots, x_n)$; i.e., the c-median is the population mean. For intermediate sample sizes, $1 < m < n$, the c-median can be calculated for small n simply by enumerating all subsets of size m, calculating the mean for each subset, and taking the median of the means. The following fragment of R code shows the calculations for data set 2 when $m = 2$:

```
set2 <- c(20.38, 20.40, 20.40, 20.42, 20.49)

set2.cmedian <- median(combn(set2,m=2,FUN=mean))
```

```
print(set2.cmedian)
```

The result is 20.41.

D. Performance on artificial data sets

Table II compares the performance of the c-median, population mean, and population median for each of the artificial data sets described in Section II, and for various values of m. Instead of the MSE we report its square root, the *root-mean-square error* (RMSE), for each estimator, so that the figure of merit has the same dimensions as the activity level. Both the RMSE and MAE for a given m are taken over all subsets of size m. Thus, even for m-independent estimators (the population mean and median) their RMSEs and MAEs are m-dependent. Note also that the $m = 1$ and $m = 5$ cases of the c-median correspond to the population median and population mean, respectively. The $m = 5$ case trivially has zero MSE and MAE for the population mean and c-median.

In data set 1, all estimators have the same value regardless of m, and perform equally well. In the remaining data sets, the mean always has the lowest RMSE, and the median has the highest; the c-median's RMSE is usually in between. Also, the c-median always has the lowest MAE. In data sets 2 and 4, the median has the highest MAE; in data set 3, the mean has the highest MAE for $m = 1$ and 2 and the median has the highest MAE for $m = 3, 4, 5$.

The c-median estimators are also plotted in Fig. 2. For $m = 2$, the c-median behaves as a compromise between the population mean and median. For instance, in data sets 2 and 3, the c-median avoids being pulled away from the bulk of the data by the outliers, exhibiting outlier resistance (not unlike that of the population median). On the other hand, in data set 4 where the population median is trapped in one of the two clusters, the c-median falls in between the two clusters, similar to the population mean. For $m = 3$ and 4, the c-median behaves more like the population mean, as expected. Here, the chance that one of the outliers will be included in a subset is higher. Consequently, the sample means should be larger.

TABLE II: Combinatorial medians, their RMSEs, and MAEs, for samples of size m. RMSEs and MAEs of population mean and median are also shown for comparison. All units are mg-Ra eq.

Data set	m	c-median	RMSE, c-median	RMSE, mean	RMSE, median	MAE, c-median	MAE, mean	MAE, median
Set 1	1	20.42	0.02	0.02	0.02	0.02	0.02	0.02
	2	20.42	0.01	0.01	0.01	0.01	0.01	0.01
	3	20.42	0.01	0.01	0.01	0.01	0.01	0.01
	4	20.42	0.01	0.01	0.01	0.01	0.01	0.01
	5	20.42	0	0	0	0	0	0
Set 2	1	20.40	0.04	0.04	0.04	0.03	0.03	0.03
	2	20.41	0.02	0.02	0.03	0.02	0.02	0.02
	3	20.42	0.02	0.02	0.02	0.01	0.01	0.02
	4	20.42	0.01	0.01	0.02	0.01	0.01	0.02
	5	20.42	0	0	0.02	0	0	0.02
Set 3	1	20.38	0.09	0.08	0.09	0.05	0.06	0.05
	2	20.39	0.05	0.05	0.06	0.04	0.05	0.04
	3	20.44	0.04	0.03	0.05	0.03	0.03	0.04
	4	20.43	0.02	0.02	0.04	0.01	0.01	0.04
	5	20.42	0	0	0.04	0	0	0.04
Set 4	1	20.40	0.06	0.05	0.06	0.05	0.05	0.05
	2	20.44	0.03	0.03	0.05	0.02	0.03	0.04
	3	20.43	0.02	0.02	0.04	0.02	0.02	0.04
	4	20.44	0.01	0.01	0.04	0.01	0.01	0.03
	5	20.43	0	0	0.03	0	0	0.03

IV. CONCLUSION

When the goal is to estimate the mean (or total) activity of the selected sources, Table II suggests that the population median should *not* be used: the combinatorial median is as good as or better than the median with respect to both MSE and MAE. The choice between

population mean and combinatorial median depends on whether one prefers to minimize MSE or MAE, respectively. Note that MSE more aggressively penalizes large errors relative to small errors, compared to MAE.

Acknowledgments

We would like to thank A. Liaw, J. Niemkiewicz, S. Oliveira, and A. Palmiotti, for helpful discussions. The original version of this manuscript was prepared in 2008, and has never been published. The views expressed do not necessarily reflect the views, policies, or opinions of the authors' employers.

---

[1] G. Casella and R. L. Berger, *Statistical Inference* (Wadsworth, Belmont, CA, 1990).

[2] D. Zwillinger and S. Kokoska, *CRC Standard Probability and Statistics Tables and Formulae* (Chapman & Hall/CRC, Boca Raton, 2000).

[3] R. Ihaka and R. Gentleman, "R: a language for data analysis and graphics," *J. Comp. Graph. Stat.* **5**, 299–314 (1996).

[4] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust Statistics: Theory and Methods (with R)*, second edition (Wiley, Hoboken, 2019).

[5] W. G. Cochran, *Sampling Techniques*, third edition (Wiley, New York, 1977), p. 22.

Fig. 1(a): Artificial data sets for initial activities of five radioactive sources

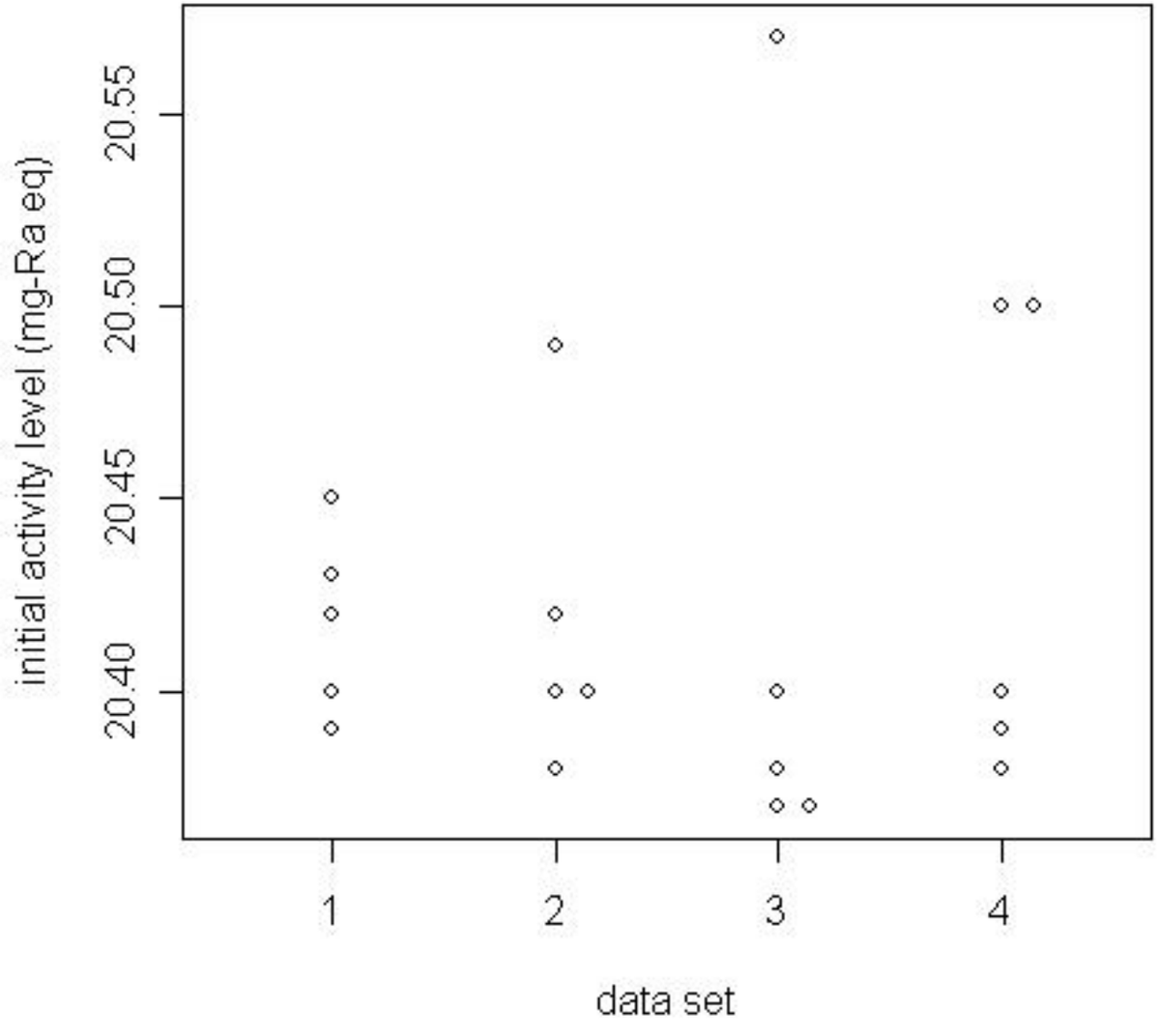

Fig. 1(b): Artificial data sets for initial activities of five radioactive sources, with estimators

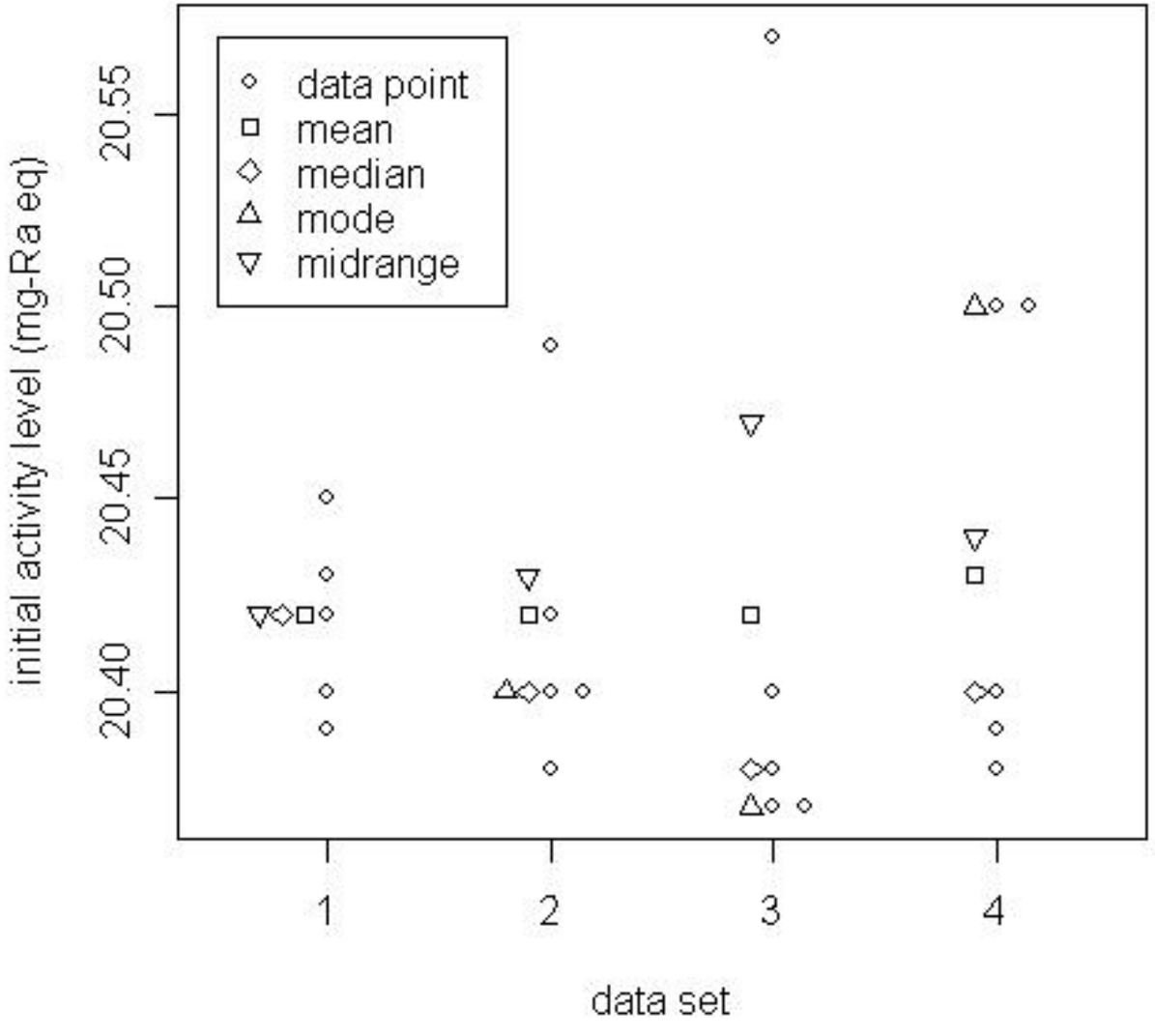

Fig. 2: Artificial data sets for initial activities of five radioactive sources, with c-medians