1 Things People Do with Censored Data that Are Just Wrong

Censored observations are low-level concentrations of organic or inorganic chemicals with values known only to be somewhere between zero and the laboratory's detection/reporting limits. The chemical signal on the measuring instrument is small in relation to the process noise. Measurements are considered too imprecise to report as a single number, so the value is commonly reported as being less than an analytical threshold, for example, "<1." Long considered second-class data, censored observations complicate the familiar computations of descriptive statistics, of testing differences among groups, and of correlation coefficients and regression equations.

Statisticians use the term "censored data" for observations that are not quantified, but are known only to exceed or to be less than a threshold value. Values known only to be below a threshold (less-thans) are left-censored data. Values known only to exceed a threshold (greater-thans) are right-censored data. Values known only to be within an interval (between 2 and 5) are interval-censored data. Techniques for computing statistics for censored data have long been employed in medical and industrial studies, where the length of time is measured until an event occurs, such as the recurrence of a disease or failure of a manufactured part. For some observations the event may not have occurred by the time the experiment ends. For these, the time is known only to be greater than the experiment's length, a right-censored "greater-than" value. Methods for incorporating censored data when computing descriptive statistics, testing hypotheses, and performing correlation and regression are all commonly used in medical and industrial statistics, without substituting arbitrary values. These methods go by the names of "survival analysis" (Klein and Moeschberger, 2003) and "reliability analysis" (Meeker and Escobar, 1998). There is no reason why these same methods should also not be used in the environmental sciences, but until recently their use has been relatively rare. Environmental scientists have not often been trained in survival analysis methods.

The worst practice when dealing with censored observations is to exclude or delete them. This produces a strong bias in all subsequent measures of location or hypothesis tests. After excluding the 80% of observations that are left-censored nondetects, for example, the mean of the top 20% of concentrations is reported. This provides almost no insight into the original data. Excluding censored observations removes the

Statistics for Censored Environmental Data Using Minitab® and R, Second Edition.
Dennis R. Helsel.

primary information contained in them—the proportion of data in each group that lies below the reporting limit(s). And while better than deleting censored observations, fabricating artificial values as if these had been measured provides its own inaccuracies. Fabrication (substitution) adds an invasive signal to the data that was not previously there, potentially obscuring the information present in the measured observations.

Studies 25 years ago found substitution to be a poor method for computing descriptive statistics (Gilliom and Helsel, 1986). Numerous subsequent articles (see Chapter 6) have reinforced that opinion. Justifications for using one-half the reporting limit usually point back to Hornung and Reed (1990), who only considered estimation of the mean, and assumed that data below the single reporting limit follow a uniform distribution. Estimating the mean is not the primary issue. Any substitution of a constant fraction times the reporting limits will distort estimates of the standard deviation, and therefore all (parametric) hypothesis tests using that statistic. This is illustrated in a later section using simulations. Also, justifications for substitution rarely consider the common occurrence of changing reporting limits. Reporting limits change over time due to methods changes, change between samples due to changing interferences, amounts of sample submitted, and other causes. Substituting values that are tied to changing reporting limits introduces an external (exotic) signal into the data that was not present in the media sampled. Substituted values using a fraction anywhere between 0 and 0.99 times the detection limit are equivalently arbitrary, easy, and wrong.

There have been voices objecting to substitution. In 1967, a US Geological Survey report by Miesch (1967) stated that substituting a constant for censored observations created unnecessary errors, instead recommending Cohen's Maximum Likelihood procedure. Cohen's procedure was published in the statistical literature in the late 1950s and early 1960s (Cohen, 1957, 1961), so its movement into an applied field by 1967 is a credit indeed to Miesch. Two other early environmental pioneers of methods for censored data are Millard and Deverel (1988) and Farewell (1989). Millard and Deverel (1988) pioneered the use of two-group survival analysis methods in environmental work, testing for differences in metals concentrations in the groundwaters of two aquifers. Many censored values were present, at multiple reporting limits. They found differences in zinc concentrations between the two aquifers using a survival analysis method called a score test (see Chapter 9). Had they substituted one-half the reporting limit for zinc concentrations and run a *t*-test, they would not have found those differences. Farewell (1989) suggested using nonparametric survival analysis techniques for estimating descriptive statistics, hypothesis testing, and regression for censored water quality data. Many of his suggestions have been expanded in the pages of this book. Since that time, a guide to the use of censored data techniques for environmental studies was published by Akritas (1994) as a chapter in volume 12 of the *Handbook of Statistics*. In an applied setting, She (1997) computed descriptive statistics of organics concentrations in sediments using a survival analysis method called Kaplan–Meier. Means, medians, and other statistics were computed without substitutions, even though 20% of data were observations censored at eight different reporting limits.

Guidance documents have evolved over the years when recommending methods to deal with censored observations. In 1991 the *Technical Support Document for Water-Quality Based Toxics Control* (USEPA, 1991) recommended use of the delta-lognormal (also called Aitchison's or DLOG) method when computing means for censored data. Gilliom and Helsel (1986) had previously shown that the delta-lognormal method was essentially the same as substituting zeros for censored observations, and so its estimated mean was consistently biased low. Hinton (1993) found that the delta-lognormal method was biased low and had a larger bias than either Cohen's MLE or the parametric ROS procedure (see Chapter 6 for more information on the latter). The 1998 *Guidance for data quality assessment: Practical methods for data analysis* recommended substitution when there were fewer than 15% censored observations, otherwise using Cohen's method (USEPA, 1998a). Cohen's method, an approximate MLE method using a lookup table valid for only one reporting limit, may have been innovative when proposed by Miesch in 1967, but by 1998 there were better methods available. Minnesota's *Data Analysis Protocol for the Ground Water Monitoring and Assessment Program* presented an early adoption of some of the better, simpler methods for censored data (Minnesota Pollution Control Agency, 1999). In 2002, substitution of the reporting limit was still recommended in the *Development Document for theProposed Effluent Limitations Guidelines and Standards for the Meat and Poultry Products Industry Point Source Category* (USEPA, 2002c). States have forged their own way at times—in 2005 the California Ocean Plan recommended use of robust ROS when computing a mean and upper confidence limit on the mean (UCL95) for determining reasonable potential (California EPA, 2005, Appendix VI). More recently, the *2009 Stormwater BMP Monitoring Manual* (Geosyntec Consultants and Wright Water Engineers, 2009) states "It is strongly recommended that simple substitution is avoided," and instead recommends methods found in this book for estimating summary statistics. And the 2009 *Unified Guidance* on statistical methods for groundwater quality at RCRA facilities (USEPA, 2009) recommended the use of survival analysis methods, although they unfortunately allowed substitution for estimation and hypothesis testing when the proportion of censored observations was below 15%.

1.1 WHY NOT SUBSTITUTE—MISSING THE SIGNALS THAT ARE PRESENT IN THE DATA

Statisticians generate simulated data for much the same reasons as chemists prepare standard solutions—so that the starting conditions are exactly known. Statistical methods are then applied to the data, and the similarity of their results to the known, correct values provides a measure of the quality of each method. Fifty pairs of X,Y data were generated by Helsel (2006) with X values uniformly distributed from 0 to 100. The Y values were computed from a regression equation with slope $= 1.5$ and intercept $= 120$. Noise was then randomly added to each Y value so that points did not fall exactly on the straight line. The result is data having a strong linear relation between Y and X with a moderate amount of noise in comparison to that linear signal.

The noise applied to the data represented a "mixed normal" distribution, two normal distributions where the second had a larger standard deviation than the first. All of the added noise had a mean of zero, so the expected result over many simulations is still a linear relationship between X and Y with a slope $= 1.5$ and intercept $= 120$. Eighty percent of data came from the distribution with the smaller standard deviation, while 20% reflected the second distribution's increased noise level, to generate outliers. The 50 generated values are plotted in Figure 1.1a.

The 50 observations were also assigned to one of the two groups in a way that group differences should be discernible. The first group is mostly of early (low X) data and second of later (high X) data. The mean, standard deviation, correlation coefficient, regression slope of Y versus X, a t-test between the means of the two groups, and its p-value for the 50 generated observations in Figure 1.1a were then all computed and stored. These "benchmark" statistics are the target values to which later estimates are compared. The later estimates are made after censoring the points plotted as squares in Figure 1.1a.

Two reporting limits (at 150 and 300) were then applied to the data, the black dots of Figure 1.1a remaining as uncensored values with unique numbers, and the squares becoming censored observations below one of the two reporting limits. In total, 33 of 50 observations, or 66% of observations, were censored below one of the two reporting limits. This is within the range of amounts of censoring found in many environmental studies. Use of a smaller percent censoring would produce many of the same effects as found here, though not as obvious or as strong. All of the data between 150 and the higher reporting limit of 300 were censored as <300. In order to mimic laboratory results with two reporting limits, data below 150 were randomly selected and some assigned <150 while others became <300.

1.1.1 Results

Figure 1.1b–g illustrate the results of estimating a statistic or running a hypothesis test after substituting numbers for censored observations by multiplying the reporting limit value by a fraction between 0 and 1. Estimated values for each statistic are plotted on the Y-axes, with the fraction of the reporting limit used in substitution on the X-axes. A fraction of 0.5 on the X axis corresponds to substituting a value of 75 for all <150s, and 150 for all <300s, for example. On each plot is also shown the value for that statistic before censoring, as a "benchmark" horizontal line. The same information is presented in tabular form in Table 1.1.

Estimates of the mean of Y are presented in Figure 1.1b. The mean Y before censoring equals 198.1. Afterwards, substitution across the range between 0 and the detection limits (DL) produces a mean Y that can fall anywhere between 72 and 258. For this data set, substituting data using a fraction somewhere around 0.7 DL appears to mimic the uncensored mean. But for another data set with different characteristics, another fraction might be "best." And 0.7 is not the "best" for these data to duplicate the uncensored standard deviation, as shown in Figure 1.1c. Something larger or smaller, closer to 0.5 or 0.9 would work better for that statistic, for this set of data. Performance will also differ depending on the proportion of data censored, as

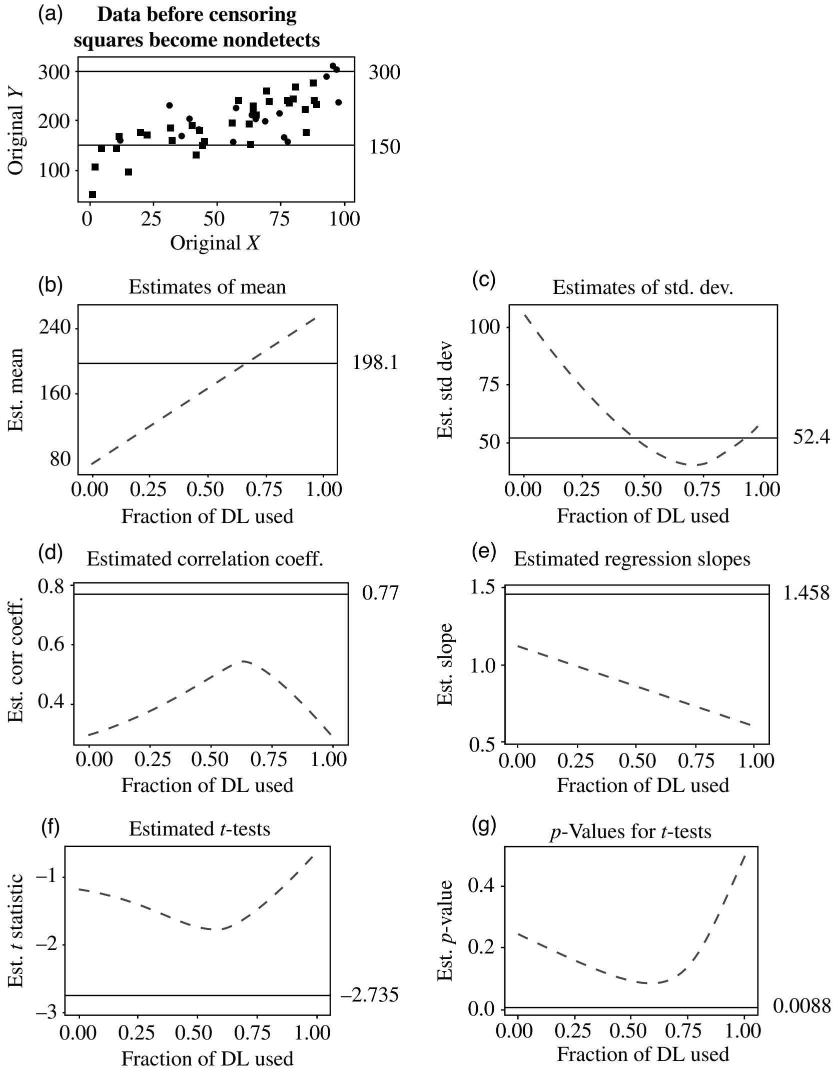

FIGURE 1.1 (a) Data used. Horizontal lines are reporting limits. (b–g) Estimated values for statistics of censored data (Y) as a function of the fraction of the detection limit (X) used to substitute values for each nondetect. As an example, 0.5 corresponds to substitution of one-half the detection limit for all censored values. Horizontal lines are at target values of each statistic obtained using uncensored values.

discussed later. Results for the median (not shown) were similar to those for the mean, and results for the interquartile range (not shown) were similar to those for the standard deviation. The arbitrary nature of the choice of fraction, combined with its large effect on the result, makes the choice of a single fraction an uncomfortable one. As shown later, it is also an unnecessary one.

TABLE 1.1 Statistics and Test Results Before and After Censoring

Procedure	Before Censoring	Range Using Substitution	Using MLE
Mean	198.1	72–258	191.3
Standard deviation	52.4	41–106	54.0
Correlation coefficient	0.77	0.29–0.54	0.55
Regression slope	1.46	0.62–1.12	1.46
t Statistic	−2.74	−1.8 to −0.68	−1.81
p-value for *t*	0.009	0.08–0.50	0.07

Data in the middle two columns are also shown in Figure 1.1. The right column reports the results of MLE tests expressly designed to work with censored data, without requiring substitution for censored observations.

Substitution results in poor estimates of correlation coefficients (Figure 1.1d) and regression slopes (Figure 1.1e), much further away from their respective uncensored values than was true for descriptive statistics. The closest match for the correlation coefficient appears to be near 0.7, while for the regression slope, substituting 0 would be best! With data having other characteristics, the "best" fraction will differ. Because substituted values at a given reporting limit produce a horizontal line, correlation coefficients and regression slopes are particularly suspect when values are substituted for censored observations, especially if the statistics are found to be insignificant.

The generated data were split into two groups. In the first group were data with *X* values of 0–40 and 60–70, while the second group contained those with *X* values from 40 to 60 and then 70 and above. For the most part, values in the first group plotted on the left half of Figure 1.1a, and the second group plotted primarily on the right half. Because the slope change is large relative to the noise, mean *Y* values for the two groups are significantly different. Before the data were censored, the two-sided *t*-statistic to test equality of the mean *Y* values was −2.74, with a *p*-value of 0.009. This is a small *p*-value, so before censoring the means for the two groups are determined to be different.

Figure 1.1f and g, and Table 1.1 report the results of two-group *t*-tests following substitution of values for censored observations. The *t*-statistics never reach as large a negative value as for the uncensored data, and the *p*-values are therefore never as significant. At no time do the *p*-values go below 0.05, the traditional cutoff for statistical significance. Results of *t*-tests after using substitution, if found to be insignificant, should not be relied on. Much of the power of the test has been lost, as substitution is a poor method for recovering the information contained in censored observations. Figure 1.1f and g show a strong drop-off in performance when the best choice of substituted fraction, which in practice is always unknown, is not chosen.

Clearly, no single fraction of the reporting limit, when used as substitution for a nondetect, does an adequate job of reproducing more than one of these statistics. This exercise should not be used to pick 0.7 or some other fraction as "best"; different fractions may do a better job for data with different characteristics. The process of substituting a fraction of the reporting limits has repeatedly been shown to produce poor results in simulation studies (Gilliom and Helsel, 1986; Singh and Nocerino, 2002; and many others—see Chapter 6). As demonstrated by the long list of research findings and this simple exercise, substitution of a fraction of the reporting limit for censored observations should rarely be considered acceptable in a quantitative analysis. There are better methods available.

When substitution might be acceptable? Research scientists tend to use chemical analyses with relatively high precision and low reporting limits. These chemical analyses are often performed by only one operator and piece of equipment, and reporting limits stay fairly constant. Research data sets may include hundreds of data points, and in comparison our 50 observations appears small. For large data sets with a censoring percentage below 60% censored observations, the consequences of substitution should be less severe than those presented here. In contrast, scientists collecting data for regulatory purposes rarely have as many as 50 observations in any one group; sizes near 20 are much more common. Reporting limits in monitoring studies can be relatively high compared to ambient levels, so that 60% or greater censored observations is not unusual. Multiple reporting limits arise from several common causes, all of which are generally unrelated to concentrations of the analyte(s) of interest. These include using data from multiple laboratories, varying dilutions, and varying sample characteristics such as dissolved solids concentrations or amounts of lipids present. Resulting data like that of She (1997) with 8 different reporting limits out of 11 censored observations is quite typical. In this situation, the cautions given here must be taken very seriously, and results based on substitution severely scrutinized before publication. Reviewers should suggest that the better methods available from survival analysis be used instead.

Is there a censoring percentage below which the use of substitution can be tolerated? The short answer is “who knows?” The US Environmental Protection Agency (USEPA) has recommended substitution of one-half the reporting limit when censoring percentages are below 15% (USEPA, 1998a). This appears to be based on opinion rather than any published article. Even in this case, answers obtained with substitution will have more error than those using better methods (see Chapter 6). Will the increase in error with substitution be small enough to be offset by the cost of learning to use better, widely available methods of survival analysis? Answering that question depends on the quality of result needed, but substitution methods should be considered at best “semiquantitative,” to be used only when approximate answers are required. Their current frequency of use in research publications is certainly excessive, in light of the availability of methods designed expressly for analysis of censored data.

1.1.2 Statistical Methods Designed for Censored Data

Methods designed specifically for handling censored data are standard procedures in medical and industrial studies. Results for the current data using one of these methods, maximum likelihood estimation (MLE), are reported in the right-hand column of Table 1.1. The method assumes that data have a particular shape (or distribution), which in Table 1.1 was a normal distribution, the familiar bell-shaped curve.

The right-hand column of Table 1.1 shows that a method designed for censored data produces values for each statistic as good or better than the best of the estimates produced by substitution. MLE accomplishes this without substituting arbitrary values for censored observations. Instead, it fits a distribution to the data that matches both the values for uncensored observations, and the proportion of observations falling below each reporting limit. The information contained in censored observations is efficiently captured by the proportion of data falling below each reporting limit. The specific procedures used, such as the likelihood *r* correlation coefficient, are described in subsequent chapters. Table 1.1 shows that for two-group tests, correlation coefficients and regression slopes, true differences and nonzero slopes can be missed when substitution is used for censored observations.

1.2 WHY NOT SUBSTITUTE?—FINDING SIGNALS THAT ARE NOT THERE

Comparing two groups of data, one a possibly contaminated test group and the other a control group, is a basic design in environmental science. Trace metal concentrations in the bodies of mayflies in pristine streams could be contrasted to those in streams with industrial outfalls. Particulates in the atmosphere are compared inside and outside of a national park. Cadmium concentrations in soils are tested upwind and downwind of an old smelter site. Blood lead levels in children are contrasted between homes with old and peeling paint to those in homes with lead-free paint. Are concentrations in the test group higher than in the control group?

The classic approach for this design is the two-sample *t*-test. If data distributions do not follow a normal distribution, the nonparametric Mann–Whitney (also called Wilcoxon rank-sum) test is used instead. With either test, a roadblock looms in the data shown in Table 1.2—there are values below detection limits; several detection limits.

Substitution for the Table 1.2 data produces the data of Table 1.3, and a Mann–Whitney test *p*-value of 0.015. The equivalence of the groups is rejected, and the test group is declared higher than the control group. Expensive remediation actions might be mandated for conditions that have caused the elevated concentra-

TABLE 1.2 Contaminant Concentrations with Multiple Reporting Limits in a Test and a Control Group

Control Group		Test Group	
<1	<1	<2	<5
<1	<1	<2	<5
<1	<1	3.3	<5
<1	4.1	3.4	<5
1.0	7.0	<2	4.7
1.8	7.5	12.2	<5
2.2	15.4	<5	22.5
<2		6.6	

TABLE 1.3 Contaminant Concentrations in a Test and a Control Group After Substituting One-Half the Reporting Limit for Censored Observations

Control Group		Test Group	
0.5	0.5	1.0	2.5
0.5	0.5	1.0	2.5
0.5	0.5	3.3	2.5
0.5	4.1	3.4	2.5
1.0	7.0	1.0	4.7
1.8	7.5	12.2	2.5
2.2	15.4	2.5	22.5
1.0		6.6	

tions in the test group. Soil is removed. Industrial equipment is modified. Wells are abandoned. People are given new medications.

Now let us pull back a curtain. These data were not field data, but were computer generated. By generating data, the true situation is known. All of the data in Table 1.2 came from the same distribution—there is actually NO difference in their mean or median levels (see Figure 1.2). For the original uncensored data, the Mann–Whitney test produced a one-sided *p*-value of 0.43, stating that there is no evidence for difference between the two groups. Any reasonable method for analyzing the data with censored observations should also find no difference in the two groups. For example, in Chapter 9 a Wilcoxon score test is presented, a nonparametric test to compare two groups of data with multiple thresholds. No substitution is involved, and the test produces a *p*-value of 0.47 for the censored Table 1.2 data. No difference. No contamination. No remediation. But following substitution, a difference was declared.

The examples in these two sections have demonstrated that substitution for censored observations can lead to "finding" either false differences that are not there, or false no-differences when data are truly not equivalent. Substitution implants

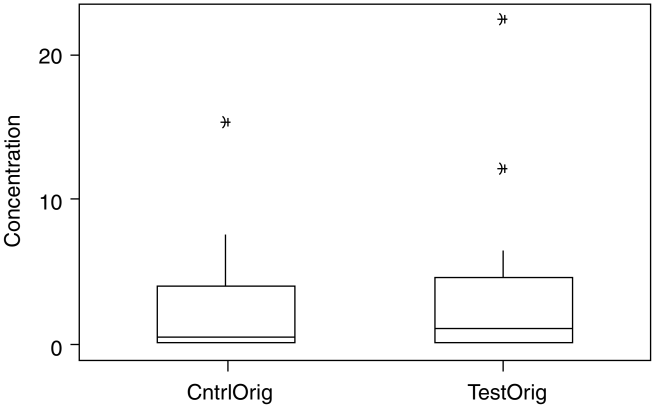

FIGURE 1.2 Boxplots for data of Table 1.2 prior to setting artificial reporting limits. Mann–Whitney test *p*-value (uncensored data) = 0.43.

an invasive pattern into the data that may be quite different than the pattern of the data itself. Substitution is not neutral.

1.3 SO WHY NOT SUBSTITUTE?

The only conclusion possible based on these two simulations is that substitution of values tied to the reporting limit, still the most commonly used method in environmental studies today, is NOT a reasonable method for interpreting censored data. The first simulation demonstrated that an invasive pattern not present in the original data was implanted by substitution, hiding signals that are really there. Causes of contamination are missed, and human or ecosystem health is needlessly endangered. The second simulation shows that the invasive pattern of substitution can introduce a signal that is not there in the data. Expensive cleanup measures may be implemented where none are needed. Substituting values as "real data" that are a function of the process used by the laboratory, are a function of time, or of the dilution of the samples, or of interferences in some samples but not others, or of the mass of material submitted to the laboratory, can easily impose an artificial, invasive pattern that originally was not there. The result is not just an incorrect conclusion by a hypothesis test. In the real world, contamination goes unnoticed. Remediation goes undone. Public health is unknowingly threatened.

There are better ways.

1.4 OTHER COMMON MISUSES OF CENSORED DATA

In addition to the two previous misuses of censored data:

(1) deleting/ignoring nondetects and computing the mean of what's left, or
(2) substituting a fraction of the reporting limit for censored observations, these two flawed approaches to evaluating censored data are fairly common:
   - substituting a value for the variance, standard deviation, or coefficient of variation (CV)
   - interpreting changes in the percent of detections while the reporting limit is changing.

There are methods for estimating the variability of censored data (see Chapter 6), and measures of location such as mean and median. Unknowingly, people have instead fabricated a number that seems "reasonable" to them. Fabricated values have made their way into some environmental regulations, where 0.6 for the CV (the ratio of the standard deviation to the mean) is currently popular. Douglas Adams would no doubt have chosen 0.42. These guessed values could be very far off, with unwarranted consequences either to human or ecological health, or to the cost of monitoring programs. The three methods in Chapter 6—MLE, Kaplan–Meier, and ROS—will each estimate the mean and standard deviation, and so the CV, for censored environmental data. There is little reason to guess a value.

Scientists also draw conclusions based on the percent of detected values, as that statistic changes between groups or through time. We will recommend the practice later in this book. However, this analysis is suspect when the definition of “detection” changes—the reporting limit changes—between groups or through time. Envision two sets of identical concentrations where the first was measured 10 years ago, and the second measured this year. They are exactly the same concentrations. There has been no physical or chemical change. The early data were censored with a mix of two reporting limits, at 1 and 10 μg/L:

```
<1 <1 <1 3 5 7 9 <10 <10 <10 <10 <10
```

while this year’s data were measured with better instruments. Now the only reporting limit is at 1 μg/L:

```
<1 <1 <1 3 5 7 9 <1 2 2 3 5
```

The analyst then computes that there were only 33% detects 10 years ago, but now there are 67% detects of this dangerous chemical. The percentage has dramatically increased, and something must be done to correct it! As you can see, this change is entirely due to the change in the mix of reporting limits used in the two groups. Comparing percent detections between groups, over space or over time only makes sense when the mix of reporting limits is constant.

Government agencies have routinely reported percent detections of pesticides and other organics in drinking water supplies, surface waters, or ground waters by compiling existing data from multiple sources. Detection limits for each chemical usually varies by source of data and over time. Maps of percent detections purport to give a regional picture of where water quality is better or worse. Decreased detection rates are cited as evidence for improving quality. Yet with the definition of “detection” changing, a change in the proportion of data sources or amounts of recent versus early data at each site can severely skew the resulting statistics. Rather than summarizing the “percent detections,” statements about “the percent of concentrations above 1 μg/L” or another well-defined threshold are much more easily interpreted. In the midst of moving detection thresholds, statements such as “Data was closely checked and it was confirmed that the detection limit changes did not affect the trend [in percent detections] significantly” (Ontario Ministry of the Environment, 2010) are hard for a reader to evaluate or believe.

Instead of computing the percent detections above a moving target, this book recommends either doing so only after recensoring all data to the highest reporting limit in the data set, a simple procedure but which may lose information, or instead using survival analysis methods that correctly account for differing reporting limits. If the metric reported and discussed is the percent of detected observations, inspect the definition of “detection” to certify that the reporting limit has not changed as in the small example above. If it has, it and not the underlying concentrations may be the cause of any shift in the percent of detections observed.