

# Appendix B

---

## *Fitting Models to Percentile Data*

---

The Exposure Factors Handbook (EFH) (U.S. EPA, 1997a) often uses percentiles to summarize data for an exposure factor. Let  $x$  denote the random variable of interest, that is,  $x$ =daily tap water consumption or  $x$ =daily inhalation rate. Theoretically, the 100 $p$ th percentile of a continuous distribution with cumulative distribution function (CDF)  $F(x)$  is the value  $x_p$  for which  $F(x_p)=p$ . That is, the 100 $p$ th percentile is the value  $x_p$  for the variable of interest that places 100 $p$ % of the probability below  $x_p$ .

A precise definition for empirical percentiles is rather involved because of finite sample size complications. If the sample size is large enough, think of the 100 $p$ th percentile simply as the smallest data value ( $x_p$ ) with at least 100 $p$ % of the sample below it. It can be estimated from the linearly interpolated empirical distribution function (EDF) by reading over from  $p$  on the vertical axis to the graph of the linearized EDF, then dropping straight down to the horizontal axis to obtain  $x_p$ .

The EDF contains all the information in the sample. Ideally, raw data would be available, and we could calculate and work with the EDF. However, raw data often is unavailable because the published literature rarely provides it. Even if raw data are available, it is not practical to include all data points for large samples in the EFH. A summary of percentiles such as those corresponding to  $p=0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95,$  and  $0.99$  contains much of the information in the original data and can be used as a basis for estimation of the distribution and testing goodness-of-fit (GOF).

A variety of methods for fitting distributions to percentile data can be identified. Four are discussed, and three of them are illustrated with a drinking water example from the EFH.

The problem of estimating distributions for exposure factors seems complicated enough by the fact that more than a dozen families of theoretical probability distributions may be needed in a toolkit for fitting environmental data. The most credible and widely used fitting method is maximum likelihood (ML) estimation. Why not simply use ML estimation? Because it may not be the best method. Some evidence of this is shown in the treatment of the tap water consumption data in Section 3.

## B.1 Four Methods of Fitting Parametric Models to Percentile Data

Serfling (1980) provides procedures for statistical inference for quantiles based on a large sample.

We concentrate here on three methods that have better small sample properties, which basically select an estimated distribution by attempting to make the fitted probabilities  $F(x_p)$  close to the nominal values of 0.01, 0.05, 0.10, etc. Graphically, the data are summarized as a plot of the nine points with  $x_p$  plotted on the horizontal axis and  $p$  plotted on the vertical axis. The goal is to find a theoretical model that passes close to the nine data points. The three methods are obtained by using different notions of closeness and are referred to as weighted least squares (WLS), minimum chi-square (MCS), and ML approaches.

EXAMPLE: Calculation of WLS, MCS, and ML measures for the tap water consumption data of older adults.

This example is from Table 3-7 of the EFH. The empirical quantile values  $x_p$  have the property that 100

% of the sample are below them. The values of  $x_p$  and  $p$  are in columns 3 and 4 of Table B-1. The quantile values  $x_p$  in Table B-1 are those from Table 3-7 divided by 100. This rescaling improves the performance of iterative search methods used to fit the curves.

The results in Table B-1 are from fitting a gamma distribution. The notes for Table B-1 indicate how the various columns are calculated. Column 5 contains the estimated or fitted probabilities  $F(x_p)$ . The goal of fitting is to choose  $F$  to make these  $F(x_p)$  values close to the target  $p$ 's. This gamma distribution was chosen to minimize a weighted sum of squares of errors (WSE) whose individual terms are

$$n*[F(x_p)-p]*[F(x_p)-p]/[p*(1-p)].$$

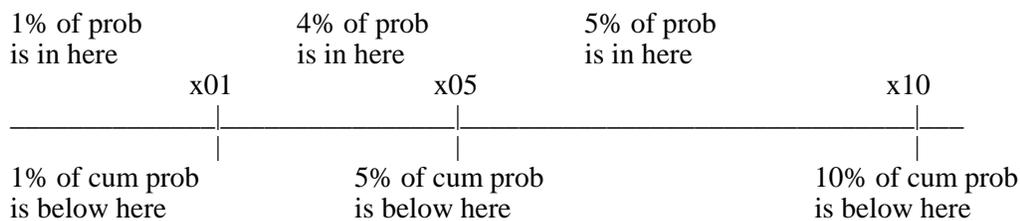
These terms are given in column 6 of Table B-1, labeled "Wtd Sqd Err (WSE)." For example, the WSE term corresponding to  $p=0.50$  is

$$2541*[ (.5 - .4942)*(.5 - .4942) ] / [ (.5)*(.5) ] = .345.$$

The column total 13.57 is the minimized WSE. That is, F was chosen as the gamma distribution, which minimizes the sum of these nine WSE terms.

By comparison with the defining formula for the Anderson-Darling (AD) statistic (Law and Kelton, 1991), it can be seen that this WSE measure is the AD discrepancy limited to the nine available quantiles. Intuitively, if a parametric distribution that agrees closely with the data at the available quantiles is selected, good agreement with respect to any aspect of the distribution, such as the mean, should be obtained.

The chi-square and log-likelihood values for this particular fitted model also are calculated on the right-hand side of Table B-1. Unlike the WSE/AD measure, the chi-square and likelihood measures focus on individual rather than cumulative probabilities associated with intervals. This distinction is illustrated in the diagram below.



Thus, column 7 of Table B-1 for nominal probability mass (labeled "Nom Prob Mass pm") contains successive differences between the nominal cumulative probability values. Similarly, column 8 for estimated probability mass (labeled "Estd Prob Mass pm^") contains successive differences between the gamma estimated cumulative probability values  $F(x_p)$ . The observed and expected numbers (O and E) of sample points in each interval are the products of the sample sizes times these nominal and estimated individual probabilities. That is, column 9 is the product of column 2 times column 7, and column 10 is the product of column 2 times column 8. The chi-square values in column 11 are calculated as  $(O-E) \cdot (O-E) / O$ . The first chi-square value is  $(25.41-9.57) \cdot (25.41-9.57) / 25.41 = 9.874$ . The log-likelihood values are the natural logarithms of pm^ raised to the O power, that is,  $O \cdot \log(\text{pm}^)$ .

The sum of the chi-square and log-likelihood values for the fitted gamma distribution are 17.60 and -4870. To obtain the MCS and ML solutions, the gamma parameters would be selected to minimize the chi-square or maximize the likelihood, rather than to minimize the WSE measure.

