# Appendix A

# Statistical Approach

February 2007

*[This page intentionally left blank.]*

## A.1 Background

The AP-42 emissions factors examined are based on the average, or the mean, of the supporting emissions data. As any statistic the emissions factors are subject to several sources of variation such as measurement error and non-representativeness of the available data. See Frey and Li (2003) for more detail with respect to sources of variability and uncertainty in the AP-42 emissions factors. The objective of this task focuses on developing uncertainty ratios for a range of probability levels. Even though this approach did not directly address all variability and uncertainty issues, it accounts for some fraction of the sampling error by adjusting for the number of tests used to produce the AP-42 Emissions Factors. Simulation techniques used in this approach allows to account for the uncertainty due to measurement error, since several simulations of the data are used to calculated different values of the EF.

This approach focuses on the basic idea that the emissions factor is a statistic, usually an average, and that the uncertainty associated with emissions factors used to represent emissions from a single or limited number of sources may be expressed/explained as a population parameter, which will be referred to as the Target Statistic; examples of Target Statistics are the 5th percentile, the median, and the 99th percentile. This statement leads to the following equation:

$$\mathrm{EF}_{\text{target statistic}} = EF_{\text{uncertainty ratio}} \times (EF) \qquad (1)$$

where EF denotes the emissions factor based on *n* tests, and $\mathrm{EF}_{\text{uncertainty ratio}}$ denotes the uncertainty ratio value for an emissions factor based on *n* tests, Solving Equation (1) for $\mathrm{EF}_{\text{uncertainty ratio}}$ results in the following equation:

$$\frac{\mathrm{EF}_{\text{target statistic}}}{\mathrm{EF}} = \mathrm{EF}_{\text{uncertainty ratio}} \qquad (2)$$

If values for the Target Statistics were known, then Equation (2) can be used to estimate the uncertainty ratio.

## A.2 Methodology

A first step in characterizing the uncertainty ratio consisted of applying exploratory data analysis techniques to obtain measures of skewness, centrality, and spread of the data.

A second step required the specification of a parametric probability distribution for the data. Parametric probability distributions are determined by a finite number of parameters which can be estimated as functions of the data. Parametric distributions make it possible to obtain interpolations (prediction within the limits of the data) and extrapolations (prediction outside the limits of the data), which will allow the generation of many values that otherwise would not be observed.

Goodness of fit tests are used to assess how well a model fits the data. One of the most popular goodness of fit tests is the Komogorov-Smirnov test. The Kolgomorov-Smirnov Goodness of fit test (KS test) is used to assess the fit of the parametric distribution to the data; in other words, the KS-test is used to decide if a sample comes from a population with a specific distribution. The KS test has the advantage of making no assumptions about the distribution of data. The KS test is based on the empirical distribution function (ECDF). Given *n* ordered data points $Y_1, \ldots, Y_n$, the ECDF is defined as $E_n = \frac{n(i)}{n}$, where *n*(*i*) is the number of points less than $Y_i$, and the $Y_i$ values are ordered from smallest to the largest value. Lillifors (1969, 1967) and Pierce (1982) showed that when the parameters of the distribution are estimated from the sample, the KS test provides non-correct p-values. Corrections for the KS tests are available for the normal and exponential distribution (Stephens, 1976,1970, 1974; Dallal and Wilkinson, 1986; Iman 1982 and Finkelsen and Schafer, 1971) but not for the Gamma distribution. Cheng and Stephens (1986) proposed a goodness of fit test based on the Moran's statistic. The proposed goodness of fit test has the same asymptotic distribution when the parameters are estimated from the sample as when the parameters are known. The test is based on the spacing of the data and provides reliable statistics for small sample sizes. The Moran's statistic has the form:

$M = \sum_{i=1}^{M} \log\left(X_i - X_{i-1}\right)$ where $X_i = F(Y_i)$, $Y_1, \ldots, Y_n$ are the ordered data points, *F* is the ECDF defined above, and *M*=*n*+1.

Once a parametric distribution was determined for the data, the parameters of the distribution were estimated using the `ms` function in `Splus`. The function `ms` maximizes the likelihood function using the Newton-Raphson algorithm. The `ms` function requires the user to provide initial values for the parameters. The Newton-Raphson algorithm is an iterative procedure that can be used to calculate maximum likelihood estimators (MLEs), which are the maximum of the likelihood function. Based on the initial values, the Newton-Raphson algorithm will search for a maximum using information from the first and second derivatives, which must be provided by the user. Initial values for the parameters were obtained using the maximum likelihood and Taylor approximations or the method of moments approach.

The Weibull, Gamma, and Log-normal distributions were considered. The distribution, likelihood, gradient, Hessian and initial values calculations are shown in Section A.3.

After the parameters of the probability distribution that best fit the data were obtained, a numerical method known as the Monte Carlo approach was used to generate 10,000 possible outcomes from the selected parametric distribution. The 10,000 simulations are referred to from now on as the hypothetical distribution. From the hypothetical distribution it is possible to obtain the mean, which is estimated by the AP-42 emissions factor, and any population parameter or Target Statistic, such as the 1st percentile, 5th percentile, median, and 99th percentile.

It was of interest to obtain $EF_{\text{uncertainty ratio}}$ for AP-42 emissions factors based on the following number of tests, (*n*), 1, 3, 5, 10, 15, 20, and 25. For each specific number of tests (*n*), 10,000 samples of size equal to *n* were drawn with replacement, and the mean was calculated for each sample.

The 10,000 means of size *n* produced a distribution of emissions factors based on *n* tests. Figure A-1 shows in the first row the distribution of 10,000 emissions factors (means) based on 3 tests.

The next step towards the characterization of the $EF_{\text{uncertainty ratio}}$ consisted of substituting each of the 10,000 emissions factor values based on a specific number of tests in Equation (2). This step resulted in a collection of distributions of $EF_{\text{uncertainty ratio}}$, one for each Target Statistic of interest and specified number of tests, *n*. The second row of Figure A-1 shows three $EF_{\text{uncertainty ratio } 3}$ (Uncertainty ratio based on *n* = 3 tests) distributions corresponding to the following Target Statistics for carbon monoxide from Wood Residue Combustion: mean, 10th percentile, and 90th percentile, respectively. The three $EF_{\text{uncertainty ratio } 3}$ distributions are highly skewed, suggesting the mean of the $EF_{\text{uncertainty ratio}}$ is affected by the extreme values.

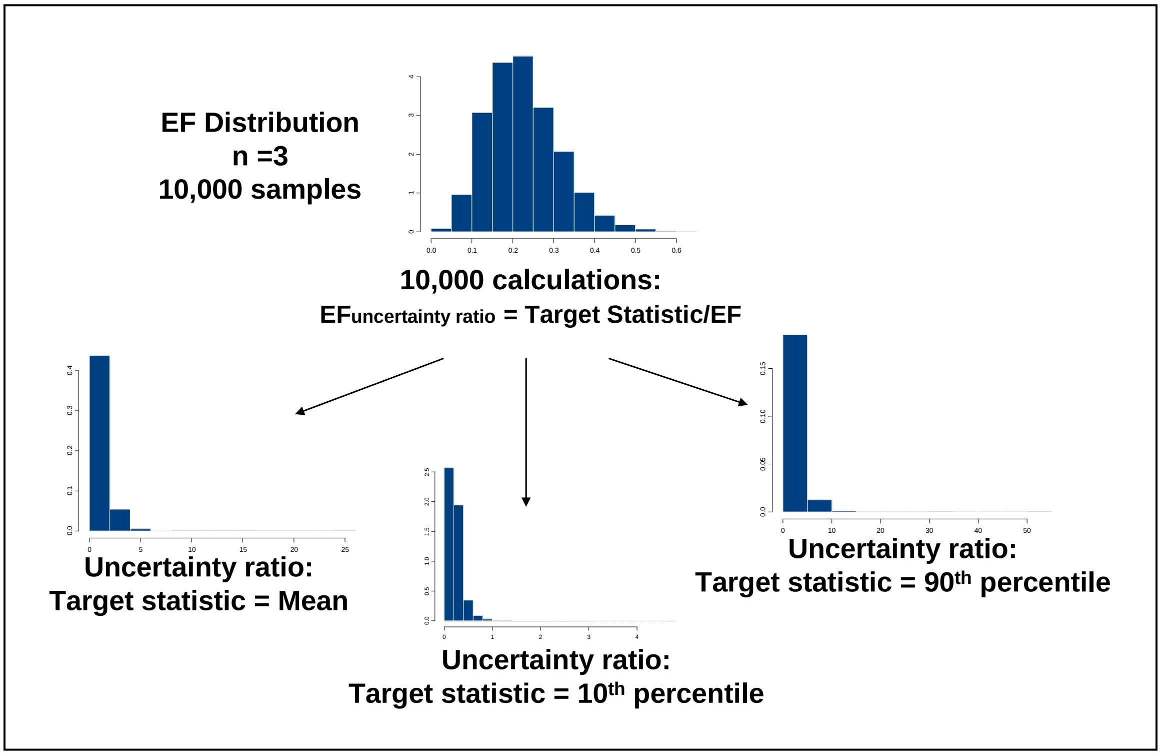

**Figure A-1. Distribution of 10,000 Emissions Factors Based on Three Tests and $EF_{\text{uncertainty ratio}}$ for Selected Target Statistics for Carbon Monoxide.**

By selecting the Median of the distribution of $EF_{\text{uncertainty ratio}}$ to adjust the Target Statistic equal to the Mean, it is safe to conclude that 50 percent of all possible adjusted emissions factors are less than or equal to the value Median($EF_{\text{uncertainty ratio}}$)*EF; while the remaining 50 percent of all possible adjusted emissions factors are greater than or equal to the value Median($EF_{\text{uncertainty ratio}}$)*EF. Furthermore, the values 95th percentile($EF_{\text{uncertainty ratio}}$)*EF and 5th percentile($EF_{\text{uncertainty ratio}}$)*EF can be interpreted as Monte Carlo Upper and Lower Confidence Limits for the ($EF_{\text{uncertainty ratio}}$)*EF, which is estimating the Target Statistic. These Monte Carlo Upper and Lower Confidence Limits provide upper and lower bounds for the EF based on *n* tests.

## A.3 Density, Log-likelihood, Gradient, Hessian and Initial Values for Probability Distributions Considered

### *A.3.1 Weibull Distribution*

1. **Density:** $f(x) = \left(\frac{\gamma}{\beta}\right)\left(\frac{x}{\beta}\right)^{\gamma-1} e^{-\left(x/\beta\right)^{\gamma}}$

2. **Log-likelihood:**

$$\log\left(f(x)\right) = \log\left(\left(\frac{\gamma}{\beta}\right)\left(\frac{x}{\beta}\right)^{\gamma-1} e^{-\left(x/\beta\right)^{\gamma}}\right) = \log(\gamma) - \log(\beta) + (\gamma-1)\log(x) - (\gamma-1)\log(\beta) - \frac{x^{\gamma}}{\beta^{\gamma}}$$

3. **Partial Derivatives (gradient):**

$$\frac{\partial l}{\partial \beta} = \frac{\gamma}{\beta}\left(-1 + \frac{x^{\gamma}}{\beta^{\gamma}}\right)$$

$$\frac{\partial l}{\partial \gamma} = \frac{1}{\gamma} + \log\left(\frac{x}{\beta}\right) - \left(\frac{x}{\beta}\right)^{\gamma} \ln\left(\frac{x}{\beta}\right)$$

4. **Hessian matrix components:**

$$\frac{\partial^2 l}{\partial \beta^2} = \frac{\gamma}{\beta^2}\left(1 - (\gamma+1)\left(\frac{x}{\beta}\right)^{\gamma}\right)$$

$$\frac{\partial^2 l}{\partial \gamma^2} = -\frac{1}{\gamma^2} - \left(\frac{x}{\beta}\right)^{\gamma} \ln\left(\frac{x}{\beta}\right)^{2}$$

$$\frac{\partial^2 l}{\partial \gamma \partial \beta} = \frac{1}{\beta}\left(-1 + \frac{x^{\gamma}}{\beta^{\gamma}}\right) + \frac{\gamma}{\beta}\frac{x^{\gamma}}{\beta^{\gamma}} \ln\left(\frac{x}{\beta}\right)$$

5. **Initial values for the NR-approach:** The initial values were obtained using a method of moments approach. For the Weibull, the non-central moments are defined as:

$$\mu_1 = \beta\,\Gamma\left(1+\frac{1}{\gamma}\right)$$

$$\mu_2 = \beta^2\,\Gamma\left(1+\frac{2}{\gamma}\right)$$

which lead to the following equation system:

$$\bar{x} = \beta\,\Gamma\left(1+\frac{1}{\gamma}\right) \Rightarrow \frac{\bar{x}}{\Gamma\left(1+\frac{1}{\gamma}\right)} = \beta$$

$$s^2 = \beta^2\,\Gamma\left(1+\frac{2}{\gamma}\right) - \left(\beta\Gamma\left(1+\frac{1}{\gamma}\right)\right)^2 \Rightarrow s^2 = \left(\frac{\bar{x}}{\Gamma\left(1+\frac{1}{\gamma}\right)}\right)^2 \Gamma\left(1+\frac{2}{\gamma}\right) - \bar{x}^2$$

$$\Rightarrow s^2 = \bar{x}^2\left(\frac{\Gamma\left(1+\frac{2}{\gamma}\right)}{\Gamma\left(1+\frac{1}{\gamma}\right)^2} - 1\right) \Rightarrow \frac{s^2}{\bar{x}^2} = \frac{\Gamma\left(1+\frac{2}{\gamma}\right)}{\Gamma\left(1+\frac{1}{\gamma}\right)^2} - 1\,.$$

Using the following Taylor approximation (Abramowitz et al., 1968):

$$z^{b-1} = \frac{\Gamma(z+a)}{\Gamma(z+b)} \approx 1 + \frac{(a-b)(a+b-1)}{2z}$$

the following approximations are obtained

$$\frac{\Gamma\left(1+\frac{2}{\gamma}\right)}{\Gamma\left(1+\frac{1}{\gamma}\right)} \approx 1 + \frac{\left(\frac{2}{\gamma}-\frac{1}{\gamma}\right)\left(\frac{2}{\gamma}+\frac{1}{\gamma}-1\right)}{2} = 1 + \frac{1}{2\gamma}\left(\frac{3-\gamma}{\gamma}\right) = \frac{2\gamma^2+3-\gamma}{2\gamma^2}$$

and

$$\frac{1}{\Gamma\left(1+\frac{1}{\gamma}\right)} \approx 1 + \frac{\left(0-\frac{1}{\gamma}\right)\left(0+\frac{1}{\gamma}-1\right)}{2} = 1 - \frac{1}{2}\frac{1}{\gamma}\left(\frac{1}{\gamma}-1\right) = \frac{2\gamma^2-1+\gamma}{2\gamma^2}$$

then

$$\frac{s^2}{\overline{x}^2} = \frac{\dfrac{2\gamma^2+3-\gamma}{2\gamma^2}}{\dfrac{2\gamma^2-1+\gamma}{2\gamma^2}} - 1 = \frac{4-2\gamma}{2\gamma^2-1+\gamma}$$

$$\Rightarrow \frac{s^2}{\overline{x}^2} = \frac{4-2\gamma}{2\gamma^2-1+\gamma} \Rightarrow \left(\frac{s^2}{\overline{x}^2}\right)(2\gamma^2-1+\gamma)-(4-2\gamma)=0$$

$$\Rightarrow \gamma^2\left(\frac{2s^2}{\overline{x}^2}\right)+\gamma\left(\frac{s^2}{\overline{x}^2}+2\right)+\left(-\frac{s^2}{\overline{x}^2}-4\right)=0$$

$$\gamma = \frac{-\left(\dfrac{s^2}{\overline{x}^2}+2\right) \pm \sqrt{\left(\dfrac{s^2}{\overline{x}^2}+2\right)^2 - 4\left(\dfrac{2s^2}{\overline{x}^2}\right)\left(-\dfrac{s^2}{\overline{x}^2}-4\right)}}{2\left(\dfrac{2s^2}{\overline{x}^2}\right)}$$

the positive values from the above equation is selected as an initial value for gamma

$$\gamma = \frac{-\left(\dfrac{s^2}{\overline{x}^2}+2\right) + \sqrt{\left(\dfrac{s^2}{\overline{x}^2}+2\right)^2 + 4\left(\dfrac{2s^2}{\overline{x}^2}\right)\left(\dfrac{s^2}{\overline{x}^2}+4\right)}}{2\left(\dfrac{2s^2}{\overline{x}^2}\right)}$$

### *A.3.2 Gamma Distribution*

1. **Density**: $f(x) = \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$
2. **Log-likelihood:**

$$\log\left(f(x)\right) = \log\left(\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\right) = \alpha \log(\lambda) - \log(\Gamma(\alpha)) + (\alpha-1)\log(x) - \lambda x$$

3. **Partial Derivatives (Gradient):**

$$\frac{\partial l}{\partial \alpha} = \log(\lambda) - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \log(x)$$

Using the asymptotic approximation for $\dfrac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \psi(\alpha) = \log(\alpha) - \dfrac{1}{2\alpha}$

$$\frac{\partial l}{\partial \alpha} = \log\left(\frac{\lambda x}{\alpha}\right) + \frac{1}{2\alpha}$$

$$\frac{\partial l}{\partial \lambda} = \frac{\alpha}{\lambda} - x$$

**Hessian matrix components:**

$$\frac{\partial^2 l}{\partial \alpha^2} = -\frac{1}{\alpha} - \frac{1}{2\alpha^2}$$

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{\alpha}{\lambda^2}$$

$$\frac{\partial^2 l}{\partial \lambda \partial \alpha} = \frac{1}{\lambda}$$

4. **Initial values for the NR-approach:**

$$\alpha = \frac{\bar{x}}{\beta},\ \beta = \frac{s^2}{\bar{x}}$$

### *A.3.4 Lognormal Distribution*

1. **Density:** $f(x) = \dfrac{\exp\left(-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right)}{\sqrt{2\pi x^2\sigma^2}}$

2. **Log-likelihood:**

$$\ln(f(x)) = \ln\left(\frac{\exp\left(-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right)}{\sqrt{2\pi x^2\sigma^2}}\right) = -\frac{1}{2}\frac{(\ln(x)-\mu)^2}{\sigma^2} - \frac{1}{2}\ln(2\pi x^2) - \frac{1}{2}\ln(\sigma^2)$$

3. **Partial Derivatives (Gradient):**

$$\frac{\partial l}{\partial \mu} = \frac{\ln(x)-\mu}{\sigma^2}$$

$$\frac{\partial l}{\partial\sigma^2} = \frac{\left(\ln(x)-\mu\right)^2}{2\left(\sigma^2\right)^2} - \frac{1}{2\sigma^2}$$

**4. Hessian matrix components:**

$$\frac{\partial^2 l}{\partial\alpha^2} = -\frac{1}{\alpha} - \frac{1}{2\alpha^2}$$
$$\frac{\partial^2 l}{\partial\mu^2} = -\frac{1}{\sigma^2}$$
$$\frac{\partial^2 l}{\partial\left(\sigma^2\right)^2} = -\frac{\left(\ln(x)-\mu\right)^2}{\left(\sigma^2\right)^3} + \frac{1}{2\left(\sigma^2\right)^2}$$

$$\frac{\partial^2 l}{\partial\mu\partial\sigma^2} = -\frac{\ln(x)-\mu}{\left(\sigma^2\right)^2}$$

**5. Initial values for the NR-approach:**

The first and second moments of the lognormal are:

$$\mu_1' = \exp\left(\mu + \frac{1}{2}\sigma^2\right) = \bar{x} \Rightarrow \mu + \frac{1}{2}\sigma^2 = \log(\bar{x})$$

$$\mu_2' = \exp\left(2\mu + 2\sigma^2\right)$$

$$s^2 = \exp\left(2\mu + 2\sigma^2\right) - \bar{x}^2$$
$$\frac{s^2}{\bar{x}^2} = \frac{\exp\left(2\mu + 2\sigma^2\right) - \bar{x}^2}{\bar{x}^2} = \frac{\exp\left(2\mu + 2\sigma^2\right)}{\exp\left(2\mu + \sigma^2\right)} - 1$$
$$\frac{s^2}{\bar{x}^2} + 1 = \exp\left(\sigma^2\right) \Rightarrow \sigma^2 = \log\left(\frac{s^2}{\bar{x}^2} + 1\right)$$
$$\Rightarrow \mu = \log(\bar{x}) - \frac{1}{2}\log\left(\frac{s^2}{\bar{x}^2} + 1\right)$$

## A.4 References

Abramowitz, Milton and Segun, Irene A. 1968. Handbook of Mathematical Functions. Dover Publications Inc., p.257

Babu, G. Jogesh, and Rao, C.R. (2004). Goodness-of-fit Tests when Parameters are Estimated. Sankhya: The Indian Journal of Statistics. 66(1):63-74.

Chandra, M.; Singpurwalla, N. D. and Stephens, M. A. 1981. Kolmogorov Statistics for Tets of Fit for the Extreme Value and Weibull Distributions. Journal of the American Statistical Association. 76(375):729-731

Frey H. Christopher and Li Song. 2003. Methods for quantifying Variability and Uncertainty in AP-42 Emission Factors: Case Studies for Natural Gas-Fueled Engines. *Journal of the Air and Waste Management Association*. 53:1436-1447

Lilliefors, Hubert W. 1967, On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association. 62(318):399-402