Distribution fitting for Database Manuscript 1: report

March 29, 2018

Distribution-fitting Methods: Hurdle distributions

It is common in empirical studies of biomass (or, more commonly, abundance) for there to be excessive density at zero (Welsh et al. 1996, Lecomte et al. (2013)) relative to the density functions commonly estimated for such data. Often the non-zero distribution is skewed to the right, implying a distribution such as the log-normal or the gamma distribution is more appropriate than the normal distribution (Lecomte et al. 2013). One method to contend with excessive density at zero is to estimate two models for the data, one that predicts the probability of observing a zero, and a second that models the distribution of non-zero values (Welsh et al. 1996, Lachenbruch (2002)). It can be shown that the maximum likelihood estimate for the two-part model can be obtained by finding maximum likelihood estimates for each part individually (Welsh et al. 1996, Duan et al. (1983)). Such a two-stage (two-part) estimation procedure has been called by many names, but we will use the nomenclature of a "hurdle model," since it's very descriptive. Qualitatively, the hurdle to be crossed is having a non-zero fuel loading, and once that hurdle is crossed (x>0) a continuous distribution is estimated for the data. The density function for the jth fuel type in the kth EVT group ($f_{kj}(x)$) can be written as (Lachenbruch 2002):

$$f_{kj}(x,d) == [\pi_{kj}^{1-d}((1-\pi_{kj})h_{kj})^d]$$

where h(x) is the estimated continuous distribution function (in this case, gamma or lognormal) for x>0, d = 1 if the the fuel type is non-zero, 0 if the fuel type is 0, and π is the probability of observing a zero. For this distribution, E(x) = $(1-\pi)$ E(h(x)) $ E(h(x)))

We will estimate and compare lognormal and gamma distributions for each fuel type and EVT group, where:

lognormal probability density function (pdf):

$$h(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0$$

$$E(x) = e^{\mu+\frac{\sigma^2}{2}}$$

$$V(x) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$

gamma pdf:

$$h(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, x > 0$$

$$E(X) = \alpha\beta$$

$$V(X) = \alpha\beta^2$$

For the fuel loading database estimation then occurs in two steps. Let n_{kj} be the total number of entries in the database for a particular fuel type (j) in a particular EVT group (k), and y_{kji} be the i^{th} fuel loading value for fuel type j in EVT group k. Then:

1. Estimate $\hat{\pi}_{kj} = \frac{\sum I(y_{kij}=0)}{n_{kj}}$, where I is an indicator function that takes a value of 1 if the entry has a value of 0, 0 otherwise.
2. For the remaining non-zero entries (x), use the fitdistr function in the R fitdistrplus package to find the maximum likelihood estimates of distribution parameters for the lognormal and gamma distributions.

Note that there are 30 total fuel types, and **_?___** total EVT groups. In general it is best-practice to inspect distribution fits graphically as part of an assessment of the distribution fit, but this is untenable with so many individual distributions that will be estimated. We will use several goodness of fit quantities to evaluate the distribution fits, with graphical spot-checks of distributions for which the goodness of fit values are not satisfactory (see below).

For initial distribution fitting we decided on a minimum of 30 non-zero entries required for a distribution to be estimated. This balanced our ability to estimate more distributions with the uncertainty in estimating distributions for small sample sizes (where with 95% confidence n = 30 is expected to obtain an estimated distribution with cumulative distribution function at most 0.25 away from the true cumulative distribution (Massart 1990). (See supplementary material for error analysis for distribution fitting)

Assessment of distribution estimates

Some of these details in supplementary?

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test is used for the null hypothesis that a given data set follows a specified theoretical distribution. In general it is designed for situations where the full theoretical distribution is specified *a priori*, and performs poorly if distribution parameter values estimated from the data are used to specify the distribution for the KS test (Lilliefors 1967). We used a Monte-Carlo procedure to estimate the p-value for the estimated distribution against the data, where a smaller p-value indicates that the observed data is statistically different than the estimated distribution (following (Lilliefors 1967)).

Monte Carlo Method: calculate KS statistic for observed distribution relative to "theoretical" distribution at estimated parameter values. Then, for 5000 MC replicates, take n (n=number of observed values in original distribution fit) random draws from "theoretical" distribution at estimated parameter values. For each of these, estimate the same theoretical distribution, then perform KS test of random to theoretical distribution at estimated parameter values. This will generate 1000 KS values when the null hypothesis is true, thus a "null" distribution. The p-value is then calculated as:

$$1 - \frac{\sum_{i=1}^{n_{mc}} I(d_{obs} > d_i)}{n_{mc} + 1}$$

where n_{mc} is the number of simulated statistics in the null distribution, d_i is the ith simulated statisic, and d_{obs} is the observed statistic. I is an indicator function that takes a value of 1 if the observed statistic is less than the simulated, 0 otherwise. The sum tallies the number of simulated statistics are smaller than the observed statistic. Note that we divide by n_{mc}+1 because we have n_{mc}+1 total statistics (including d_{obs}). We can then evaluate, against some α value, which if any distributions are "fail to reject" (ftr).

Interpretation of the KS test, for an application like this, suffers from two issues related to sample size. At low sample sizes the test has less statistical power to reject the null hypothesis, such that with low sample size a fail to reject result (ftr) does not necessarily provide evidence in favor of the estimated distribution (the null hypothesis). A large sample size presents the opposite problem–as sample size increases, the effect size necessary to reject the null hypothesis decreases. At large sample sizes this means that although the observed data are statistically different than the estimated distribution, the difference may not be of practical

significance. We use equivlance tests to aid our interpreration of the goodness of fit between observed data and estimated distributions.

Equivalence test: TOST

Robinson and Froese (2004) recommend an equivalence test to compare empirical data to model predictions using a two-one-sided t-test (TOST). In equivalence testing a maximum allowable error (or error tolerance) is defined, and the null hypothesis is that the observed distribution is outside of the error tolerance relative to the theoretical distribution. If the observed distribution is seen to be within the maximum error (or error tolerance), then the null hypothesis is rejected and the observed data is judged to be "equivalent" to the theoretical distribution (within the error tolerance). Here we use TOST to assess adequate matching between our observed empirical cumulative distribution of fuel type and the theoretical cdf associated with each candidate distribution. Let $x_{(i)}$ be the ith quantile of the empirical data distribution, and $\hat{x}_{(i)}$ be the ith quantile of the theoretical distribution. Then the difference between the observed and theoretical (x_{d_i}) is:

$$x_{d_i} = x_{(i)} - \hat{x}_{(i)}$$

We then calculate $\bar{x}_d$ as the mean distance between observed and theoretical, and use TOST to determine statistically if the observed and theoretical distributions differ by more than a specified error tolerance (ϵ). This requires an error tolerance to be specified, which for our application would be a relatively arbitrarily defined threshold.

Prichard et al (2014) use a similar equivalence procedure to evaluate the uncertainty of the fits of observed fuel consumption relative to those predicted by empirical consumption equations. For their analysis, rather than choosing a single arbitrary error threshold, they repeated the equivalence test with increasing ϵ until the first ϵ at which the equivalence test null hypothesis was rejected. This then defined the bound of uncertainty for that fuel type. We adapt their approach here, repeating the equivalence test for increasing error thresholds between observed and theoretical distributions for distributions estimated both with zeroes (and an offset), and distributions estimated for only values > 0. We then compare the minimum ϵ that rejects the null hypothesis to assess the uncertainty in the distribution estimates.

For assessing distribution estimates, a fuel type that both has a low KS p-value and a high equivalence test threshold (*could, should?*) be flagged for further evaluation.

Uncertainty in distribution estimates

Finally, we use a bootstrap procedure to estimate a standard deviation for estimated distribution parameter values, and to generate a 95% confidence interval for each distribution parameter value. The bootstrap estimates are generated using the bootdist function in the fitdistrplus package. In general, for a bootstrap, the observed data are resampled with replacement and the distribution parameters estimated for each resampling. This is repeated (**5000**) times to generate a distribution of parameter values. From this distribution a standard deviation can be calculated, and a 95% confidence interval as the 0.025 and 0.975 quantiles of the boostrap distribution.

EVT groups and fuel types

For the purposes of demonstrating comparisons of distributions among fuel types and EVT groups we present here distributions for EVT groups: (. . .) and the following fuel types: (. . .). We chose these comparisons because (. . .). Results for all other EVT groups and fuel types are given in supplementary material (S for graphics, S for Excel file of master tables).

Results

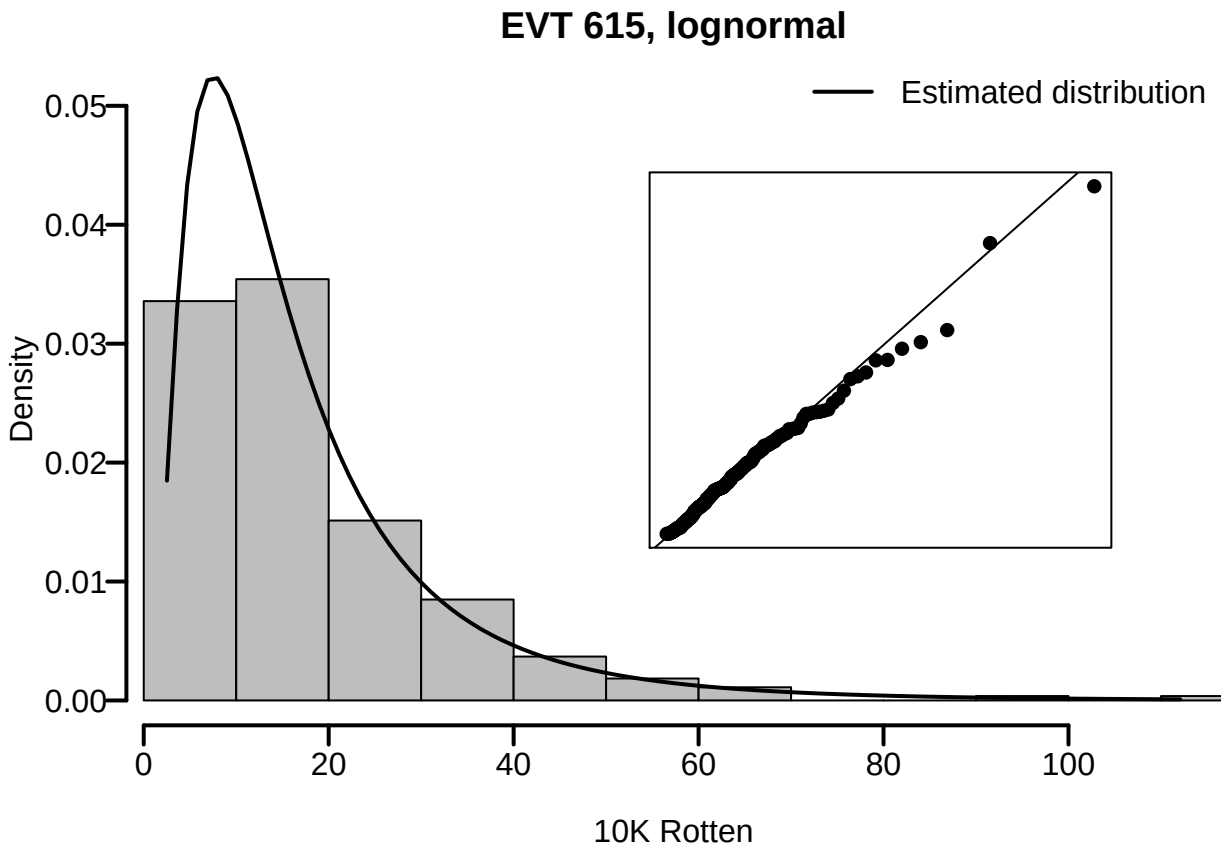
1) Summary of EVTGroups and number of observations

2) Identified data gaps

3) Distribution-fitting

Here I'm not sure how to summarize the distribution-fitting. I think we should probably pick a subset of EVT groups and fuel types to illustrate here, give the fit statistics and distribution graphs. We can choose based on diverse vegetation types, and/or representative fuel types (with more or less zeroes). I give an example below for EVT 615 (Douglas-fir-Western Hemlock Forest and Woodland)

Example fit distributions that performs well on all uncertainty measures (***Note: in a lot of the "good" fits, the tail really falls off the QQ plot, even though the fit stats are good and the other graphs look good***:



Same graph with axes on QQ plot. A little busier, but more informative.

```
## Warning in hist.default(mydata, plot = FALSE, ...): argument '...' is not
## made use of
```

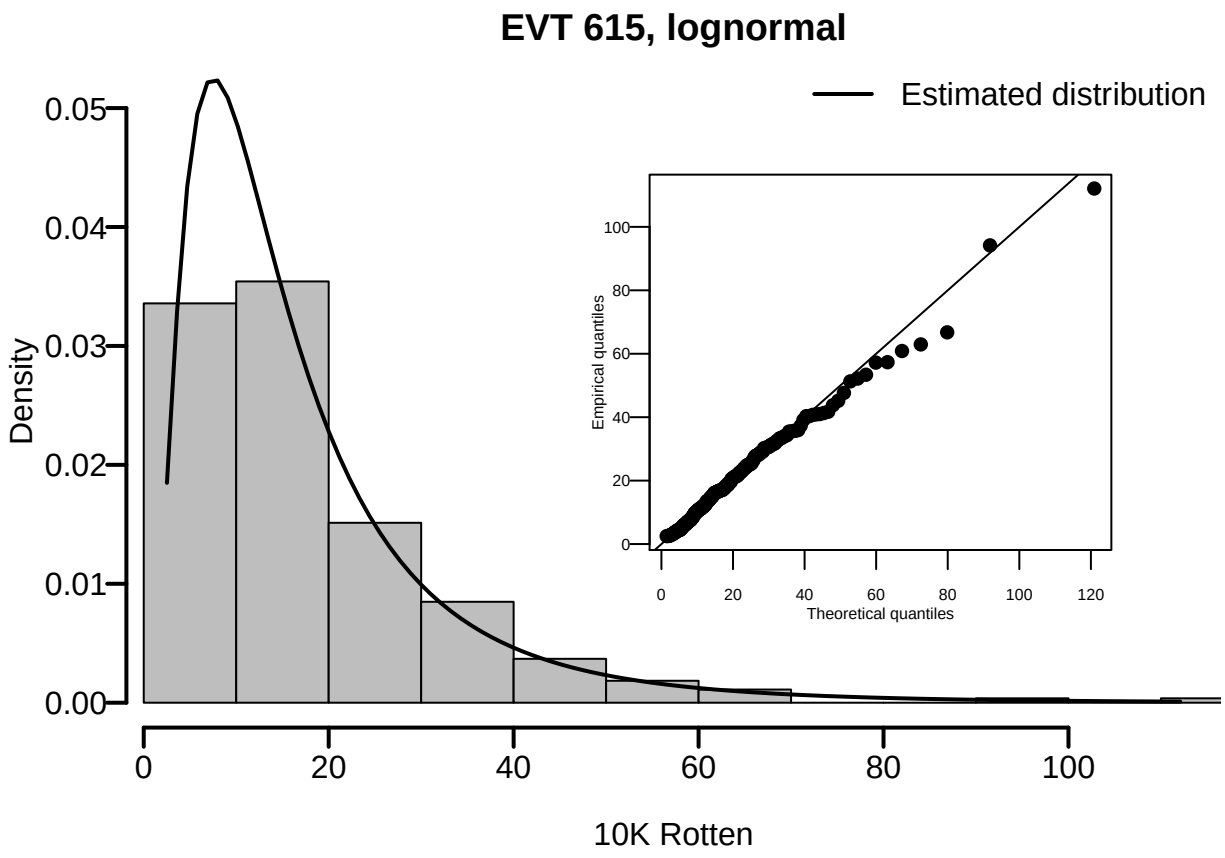


Figure: Example "Good fit".

```
## Warning in hist.default(mydata, plot = FALSE, ...): argument '...' is not
## made use of
```

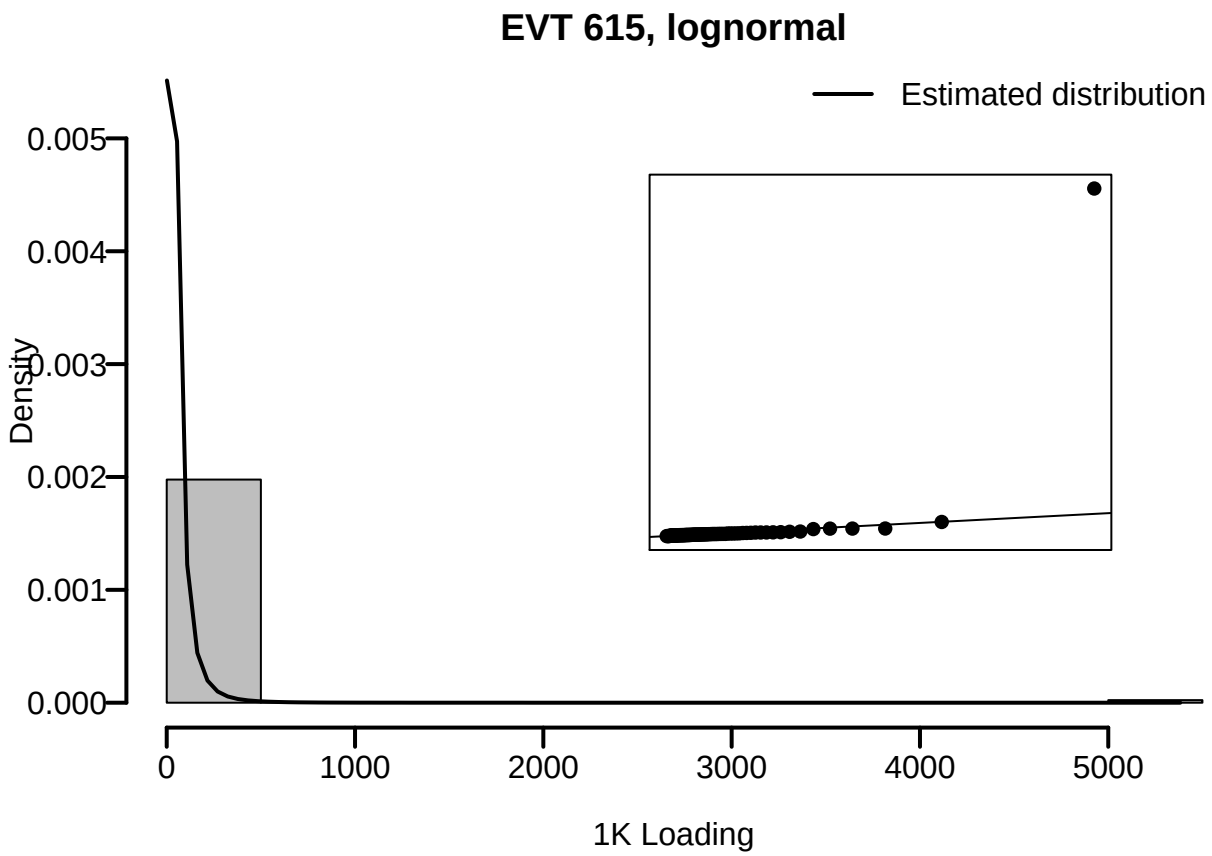


Figure: example “bad”" fit, obviously driven by outlier (so what to do?)

```
## Warning in hist.default(mydata, plot = FALSE, ...): argument '...' is not
## made use of
```

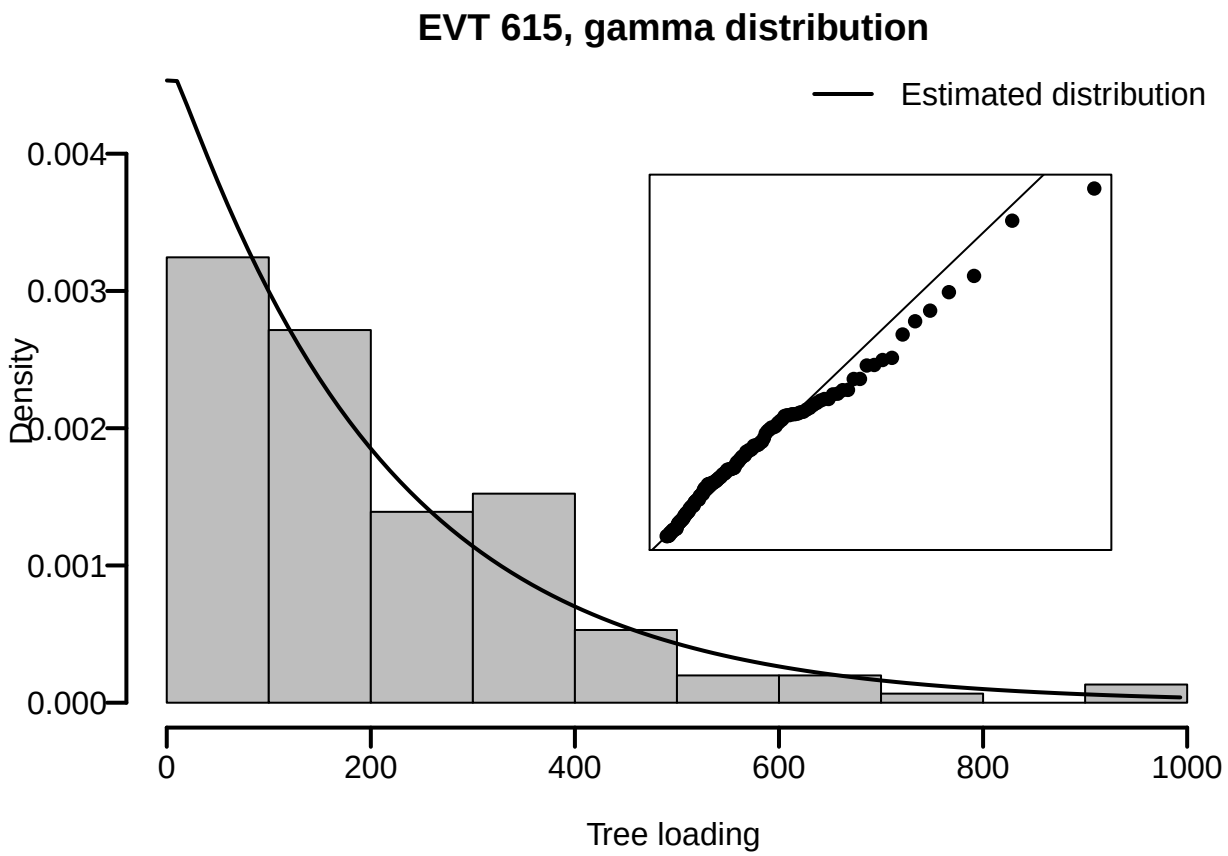


Figure: Example "bad" fit. Both the QQ plot and the cdf plot show deviations in the tail, with a substantial number of observations that fall off the fit.

Note: I have some graphical assessment of the epsilon value and KS p-value in how they relate to each other and to the number of non-zero observations.

Recommended use

For a particulate application, refer to the appropriate Master Table and fit graph to assess quality distribution fit for fuel types of interest. Consider bootstrap error values on estimated distribution parameters and visual correspondence between observed and estimated distributions. Make a judgement for appropriateness of estimated distribution for application.

Decisions to be made:

1. What is our threshold for an unacceptable fit?

It is what it is. It's an untidy distribution, or it's missing data–id's a gap. Good, not good based on threshold (or e.g., 1-5 rating). Might indicate extreme variance or poor data.

2. What to do with the unacceptable fits?
3. What to do with extreme outliers? Estimate distributions without them? So, keep in the database but fit without extreme outliers > Q3+X*IQR
4. Which results to give in the main paper, and which in supplementary?

Show distributions of (e.g., shrubs) across different regions. Inset QQ plot on histogram. Pick fuel type. For emissions applications: Course wood total (cwd_loading_Mgha), duff, shrub some regions, fine wood, herbs

Region comparison for each–pick a modal EVT Group by region, which regions, same regions for each

Global carbon, might want more aggregated/coarser aggregations (clean up)

For Monday meeting: modal EVT by region, Can estimate distributions at different levels of aggregation

Distribution-fitting power analysis

(Massart 1990) (Dvoretzky-Kiefer-Wolfowit Inequality):

$$P(\sqrt{n}\sup_x|\hat{F}_n(x) - F(x)| > \lambda) \leq 2e^{-2\lambda^2}$$

Rearranged as:

$$P(\sup_x|\hat{F}_n(x) - F(x)| > \frac{\lambda}{\sqrt{n}}) \leq 2e^{-2\lambda^2}$$

where $\hat{F}_n$ is the empirical cumulative distribution function for a sample of size n taken from distribution F.

For 1-α confidence, the rhs evaluates to α, assuming a 2-sided confidence interval:

$$\alpha = 2e^{-2\lambda^2}$$

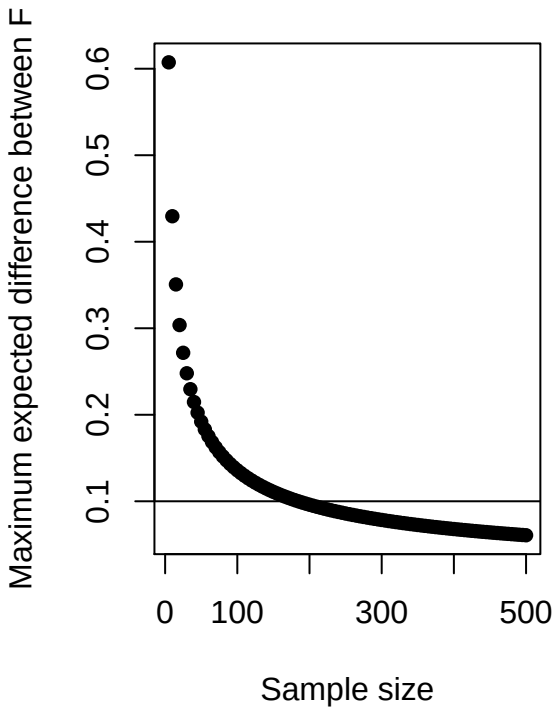
For any value of α we can thereby solve for λ:

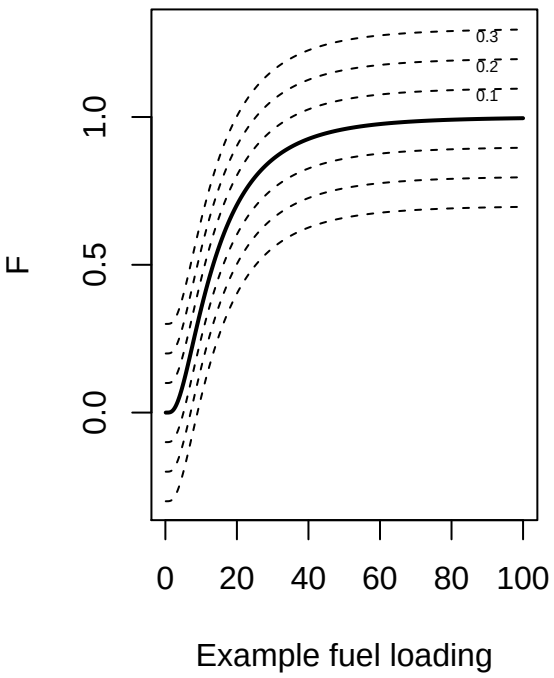
$$\lambda = \sqrt{-\frac{1}{2}\ln(\frac{\alpha}{2})}$$

Can then take the maximum expected error at the given level of confidence as:

$$\frac{\lambda}{\sqrt{n}}$$

Here are some examples for different sample sizes with 95% confidence, and a visualization of some of those different error magnitudes against a lognormal distribution:


Maximum expected difference between F
0.1
0.2
0.3
0.4
0.5
0.6
0
100
300
500
Sample size


F
0.0
0.5
1.0
0.3
0.2
0.1
0
20
40
60
80
100
Example fuel loading

page break

References

```
citation("equivalence")
```

```
##
## To cite package 'equivalence' in publications use:
##
##   Andrew Robinson (2016). equivalence: Provides Tests and Graphics
##   for Assessing Tests of Equivalence. R package version 0.7.2.
##   https://CRAN.R-project.org/package=equivalence
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {equivalence: Provides Tests and Graphics for Assessing Tests of Equivalence},
##     author = {Andrew Robinson},
##     year = {2016},
##     note = {R package version 0.7.2},
##     url = {https://CRAN.R-project.org/package=equivalence},
##   }
##
## ATTENTION: This citation information has been auto-generated from
## the package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

```
citation("fitdistrplus")
```

```
##
## To cite fitdistrplus in publications use:
##
##   Marie Laure Delignette-Muller, Christophe Dutang (2015).
##   fitdistrplus: An R Package for Fitting Distributions. Journal of
##   Statistical Software, 64(4), 1-34. URL
##   http://www.jstatsoft.org/v64/i04/.
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {{fitdistrplus}: An {R} Package for Fitting Distributions},
##     author = {Marie Laure Delignette-Muller and Christophe Dutang},
##     journal = {Journal of Statistical Software},
##     year = {2015},
##     volume = {64},
##     number = {4},
##     pages = {1--34},
##     url = {http://www.jstatsoft.org/v64/i04/},
##   }
##
## Please cite both the package and R when using them for data
## analysis. See also 'citation()' for citing R.
```

```
citation("base")
```

```
##
## To cite R in publications use:
```

```
##
##   R Core Team (2017). R: A language and environment for
##   statistical computing. R Foundation for Statistical Computing,
##   Vienna, Austria. URL https://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2017},
##     url = {https://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating R, please
## cite it when using it for data analysis. See also
## 'citation("pkgname")' for citing R packages.
```


Duan, Naihua, Willard G. Manning, Carl N. Morris, and Joseph P. Newhouse. 1983. "A Comparison of Alternative Models for the Demand for Medical Care." *Journal of Business & Economic Statistics* 1 (2): 115–26. doi:10.1080/07350015.1983.10509330.

Lachenbruch, Peter A. 2002. "Analysis of data with excess zeros." *Statistical Methods in Medical Research* 11 (4): 297–302. doi:10.1191/0962280202sm289ra.

Lecomte, Jean Baptiste, Hugues P. Benoît, Sophie Ancelet, Marie Pierre Etienne, Liliane Bel, and Eric Parent. 2013. "Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume." *Methods in Ecology and Evolution* 4 (12): 1159–66. doi:10.1111/2041-210X.12122.

Lilliefors, H W. 1967. "On the Kolmogorov-Smirnov test for normality with mean and variance unknown." *American Statistical Journal* June (318): 399–402.

Massart, P. 1990. "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality." *The Annals of Probability* 18 (3): 1269–83.

Prichard, S J, E C Karau, R D Ottmar, M C Kennedy, J B Cronan, C S Wright, and R E Keane. 2014. "Evaluation of the CONSUME and FOFEM fuel consumption models in pine and mixed hardwood forests of the eastern United States." *Canadian Journal of Forest Research-Revue Canadienne de Recherche Forestiere* 44 (April): 784–95. doi:DOI 10.1139/cjfr-2013-0499.

Robinson, Andrew P., and Robert E. Froese. 2004. "Model validation using equivalence tests." *Ecological Modelling* 176 (3-4): 349–58. doi:10.1016/j.ecolmodel.2004.01.013.

Welsh, A.H., R.B. Cunningham, C.F. Donnelly, and D.B. Lindenmayer. 1996. "Modelling the abundance of rare species: statistical models for counts with extra zeros." *Ecological Modelling* 88 (1-3): 297–308. doi:10.1016/0304-3800(95)00113-1.