Minimize risk. Maximize potential.

Deep Dive into Input Distribution Selection with @RISK

Part I: Distribution Fitting of Univariate Data

Dr. Steve Van Drew

June 19, 2020

Palisade

ICEAA Technology Showcase Webinar Series
Deep Dive: Input Distribution Selection with @RISK

Part I – Distribution Fitting of Univariate Data
Fri, Jun 19th, 10am EDT

Part II – Revisiting the Triangular and PERT "Three Point" Distributions
Fri, Jul 17th, 10am EDT

Part III – Turning Expert Opinion into Defensible Distributions
Fri, Aug 14th, 10am EDT

Palisade

Part I Objectives

- Provide @RISK users with:
  - guidance on settings to use in the @RISK Fit Distributions to Data dialog box
  - insight necessary for narrowing down the often-overwhelming candidate list of probability distributions available in @RISK for modeling inputs
    - six goodness of fit tests performed when fitting univariate data
    - some judgmental aspects of distribution selection that involve both art and science

Palisade

Input Distribution Selection

- Every risk analyst should expect questions from their peers, senior technical folks, and occasionally even management like, *Why did you use a {distribution} to model {area of uncertainty/risk}?*
- Some answers I've heard (and probably even used) … *Because:*
  - that's what the analyst before me was using
  - that's the distribution our industry/profession always uses to model {area of uncertainty/risk}
  - {distribution} is the easiest to explain to management
  - I plotted the data and it sort of looked like a {distribution}
  - that's what the computer/software told me to use

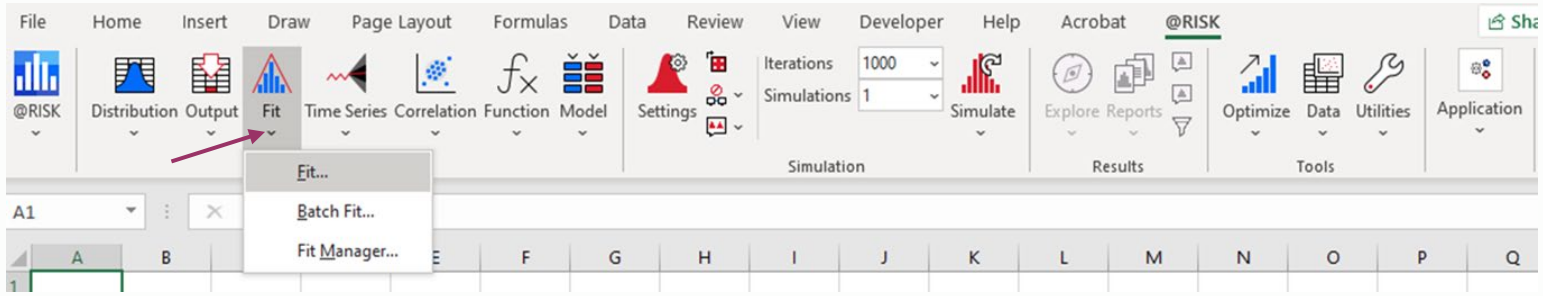
Palisade

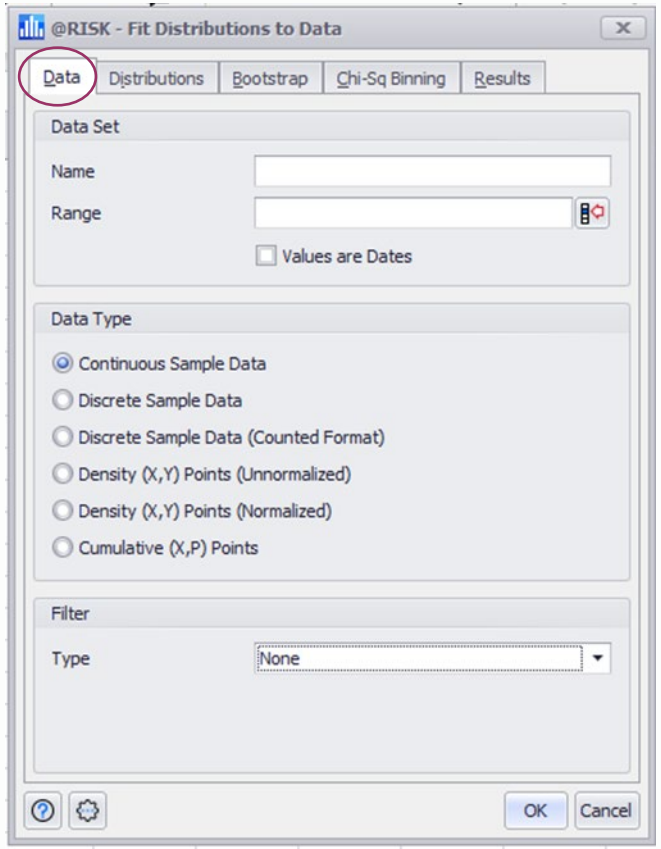
A Better Answer

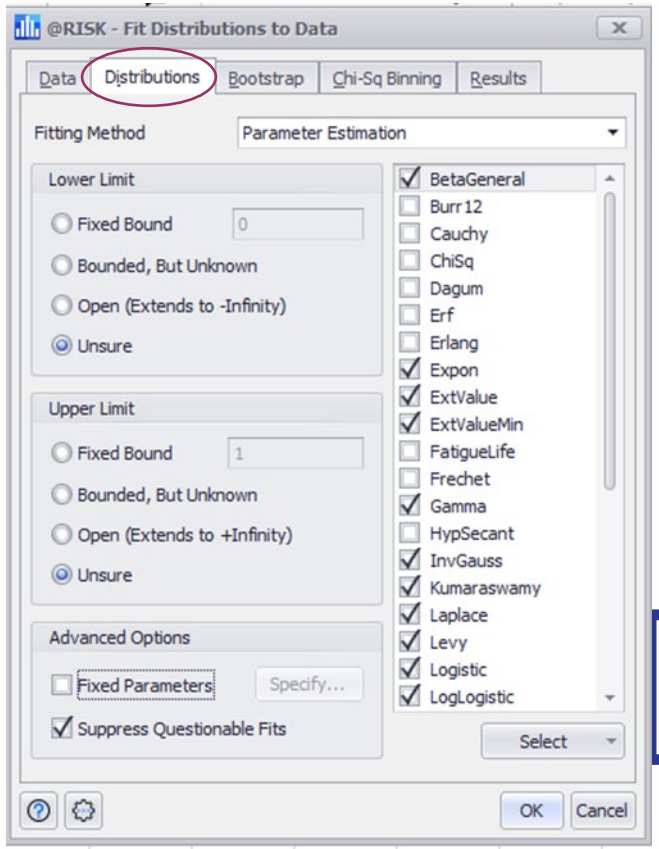
- Because after doing exploratory data analysis and ensuring the data was stable and had no frivolous outliers, I used @RISK to fit distributions to the data, and after factoring in:
  - results from all six goodness-of-fit “tests”
  - “traditional” distribution(s) used to model {area of uncertainty/risk}
  - characteristics of the data compared to the candidate distributions
  - making appropriate modifications such as truncation
- I selected {distribution}

Palisade

Distribution Fitting in @RISK 8.0

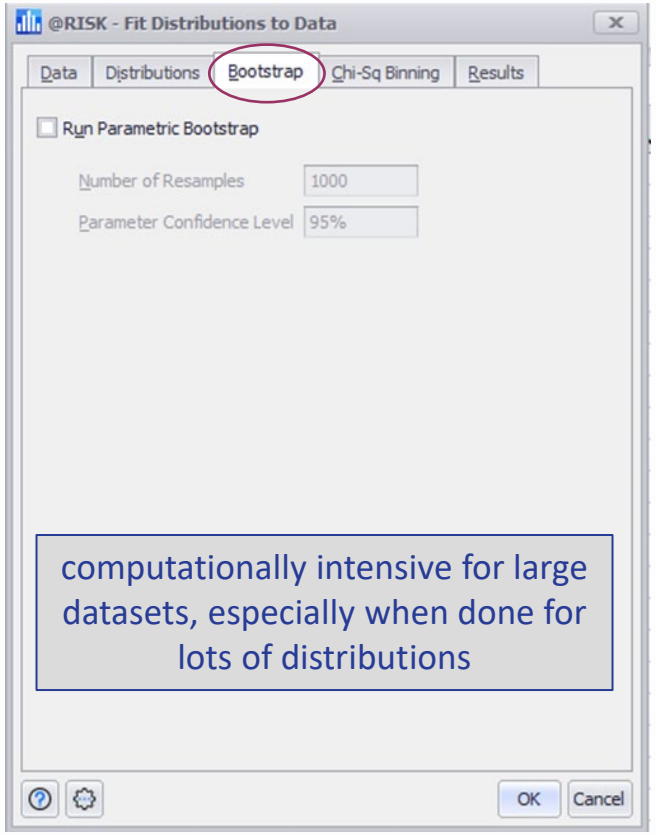




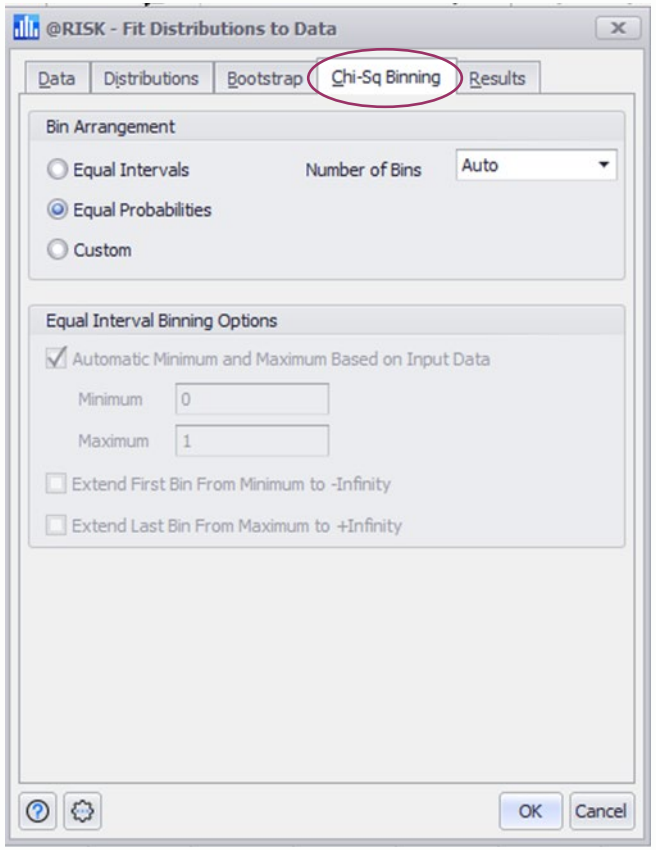


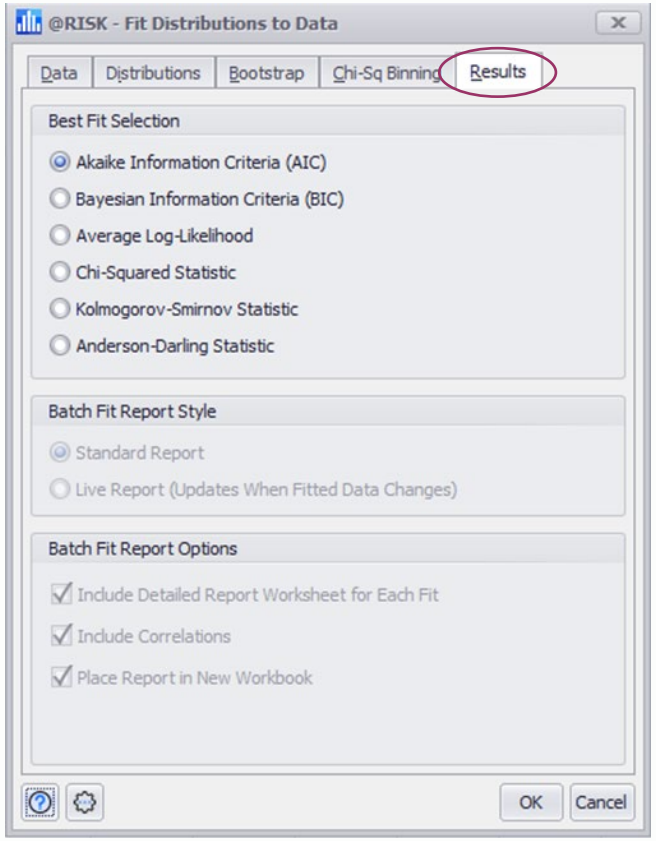
Palisade

Distribution Fitting in @RISK 8.0 (cont.)



computationally intensive for large datasets, especially when done for lots of distributions





Palisade

Goodness-of-Fit Tests

- Chi-square
  - formal comparison of histogram of data with density or mass function of hypothesized distribution
  - sensitive to grouping of data (bin/class/histogram intervals)
  - rules of thumb for number of bins: $\sqrt{n}\,;(4n)^{2/5}\,;1+\log_2 n$
- Kolmogorov-Smirnov
  - compares an empirical distribution function with the distribution function of the hypothesized distribution
  - does not require grouping of data like chi-square test
- Anderson-Darling
  - similar to K-S test but places more weight on comparison in tail(s)

Palisade

Goodness-of-Fit Tests (cont.)

- Bayesian Information Criteria (BIC)
  - calculated from the log-likelihood function
  - takes into account the number of free parameters of the fitted distribution
  - provides only a relative measure of the goodness of a particular fit
- Akaike Information Criteria (AIC)
  - similar to BIC except tends to penalize the number of parameters less strongly
- Average Log-Likelihood
  - also uses the log-likelihood function, but uses the average across the number of samples
  - largest relative measure indicates best fit

Palisade

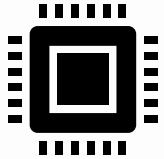
Recommended Approach

- Fit univariate continuous sample data using choices identified on earlier slides, i.e., no filtering, unsure lower and upper limits, select recommended distributions, no parametric bootstrapping, and equi-probable chi-square bins
- Look for distribution fit ranked generally highest by all six test statistics. If no clear winner give preference to AIC result
- Rerun fit for chosen distribution only, with parametric bootstrapping run to obtain p values and parameter confidence intervals

Palisade

Live Demo

Time to failure in days for 50 randomly sampled electronic chips tested at 1.5 times their nominal voltage



Initial thoughts:

- “Time” so continuous and nonnegative values
- Univariate unless other “paired” variable(s) recorded
- “Time to failure” so wouldn’t be surprised if exponential or Weibull
- Outliers anticipated, and relevant … some things just last forever

Palisade

Questions?

https://www.palisade.com/trials.asp

sales@palisade.com

Palisade

Input Distribution Selection with @RISK

- Part II – Revisiting the Triangular and PERT "Three Point" Distributions
  Fri, Jul 17th, 10am EDT

- Part III – Turning Expert Opinion into Defensible Distributions
  Fri, Aug 14th, 10am EDT

Palisade

Back Up Slides

Palisade

Chi-Square Test

H_o: the random variable, X, conforms to the distributional assumption with the parameter(s) given by the parameter estimate(s)

H_1: the random variable X does not conform

$$\chi_o^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

O_i is the observed frequency in the *i* th class interval

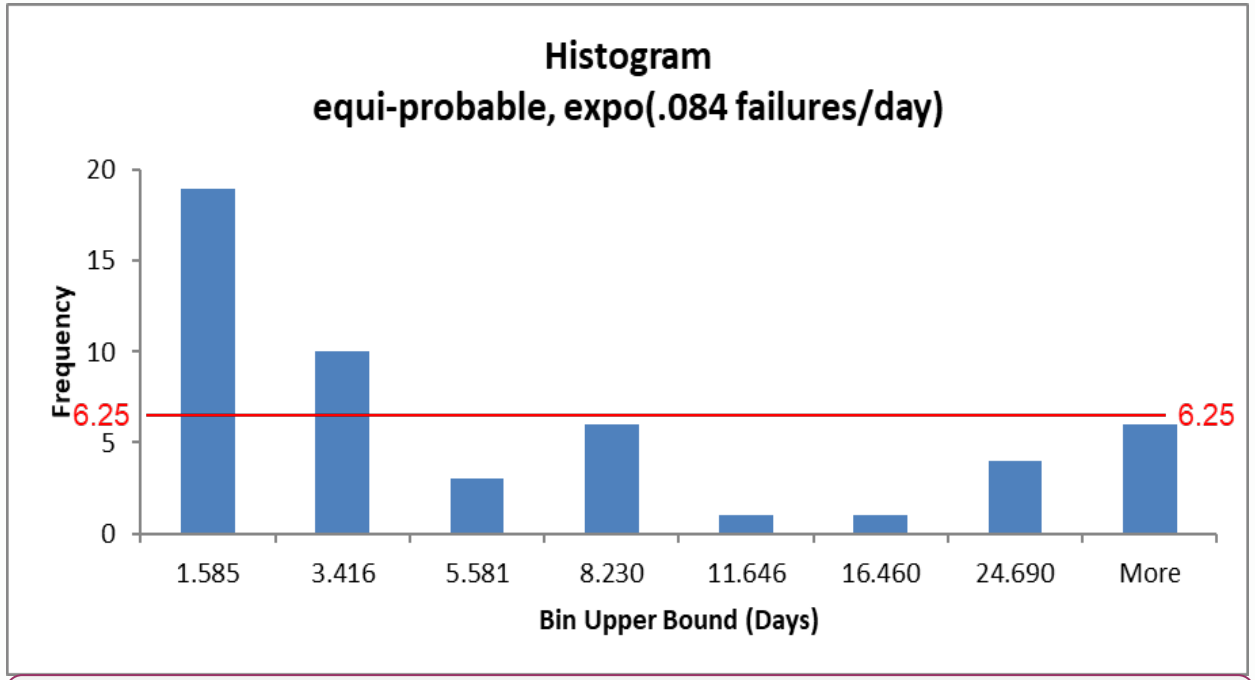
E_i is the expected frequency in the *i* th class interval, computed as $E_i = np_i$

p_i is the theoretical, hypothesized probability associated with the *i* th class interval

Reject H_o if $\chi^2_0 > \chi^2_{\alpha,k-s-1}$ {for $\chi^2_{\alpha,k-s-1}$ use Excel f_x CHIINV(probability,deg_freedom)}

s represents the number of parameters of the hypothesized distribution estimated by sample statistics

Chi-Square Test



note that bins are equi-probable, not equal interval width

Palisade

Kolmogorov-Smirnov Test

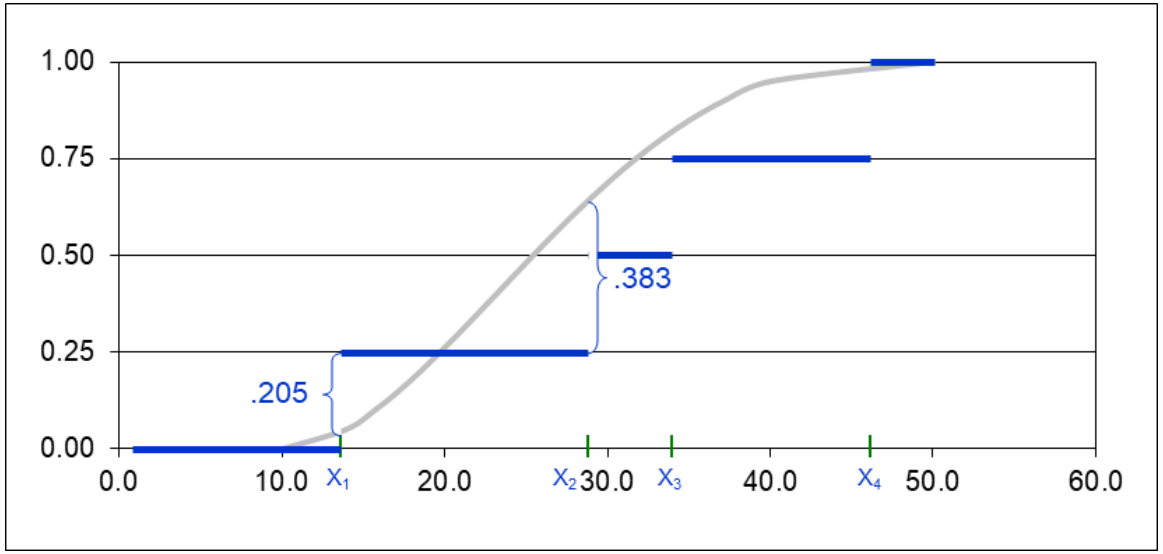
$$D = \max_x \left| F_x(x) - \hat{F}(x) \right|$$

The statistic *D* measures the largest vertical distance between the hypothesized cumulative distribution function $\hat{F}_x(x)$ and the empirical (observed) cumulative distribution function $F_x(x)$ developed from the sample data.

For continuous distributions with all parameters known (not estimated from data), reject the claim that the observed values come from the hypothesized distribution $\hat{F}_x(x)$ if $\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D > c_{1-\alpha}$; otherwise fail to reject.

α	1-α	$c_{1-\alpha}$
0.010	0.990	1.628
0.025	0.975	1.480
0.050	0.950	1.358
0.100	0.900	1.224

Kolmogorov-Smirnov Test



i	xi	$\hat{F}$xi	D+	D-
1	13.7	0.045	0.205	0.045
2	28.6	0.633	-0.133	0.383
3	33.9	0.818	-0.068	0.318
4	46.2	0.997	0.003	0.247

H_0: Beta (2, 3, 10, 50)

Palisade

Anderson-Darling Test

$$A_n^2 = n\int_{-\infty}^{\infty}[F_n(x) - \hat{F}(x)]^2 \psi(x)\hat{f}(x)dx$$

$$\text{where} \quad \psi(x) = \frac{1}{\{\hat{F}(x)[1-\hat{F}(x)]\}}$$

For continuous distributions with all parameters known (not estimated from data) and $n \geq 5$, reject the claim that the observed values come from the hypothesized distribution $\hat{F}_x(x)$ if $A_n^2 \geq a_{n,1-\alpha}$; otherwise fail to reject.

α	1-α	$a_{n,1-\alpha}$
0.010	0.990	3.857
0.025	0.975	3.070
0.050	0.950	2.492
0.100	0.900	1.933

Information Criteria

$AIC = 2k - 2 \ln L$

$BIC = k \ln n - 2 \ln L$

where:

L is the likelihood function

k is the number of parameters estimated

n is the number of sampled points

$L(\theta) = p_\theta(X_1)p_\theta(X_2) \ldots p_\theta(X_n)$

for a discrete distribution with single unknown parameter θ

$L(\theta) = f_\theta(X_1)f_\theta(X_2) \ldots f_\theta(X_n)$

for a continuous distribution with single unknown parameter θ

Palisade