

#### 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



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

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

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

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



### **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



### **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



#### **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}



# Presented for the International Cost Estimating & Analysis Association - www.iceaaonline.com Distribution Fitting in @RISK 8.0

File	Home	Insert	Drav	v Page	Layout	Formul	as D	ata P	Review	View	Develop	er Help	Acrobat	@RIS	c			🖻 Sha
@RISK	Distribution	n Output	Fit	Time Series	Correlation	$f_X$	Model	Setting	s 🚺 🗸	Iterations Simulatior	1000 1	Simulate	Explore Repo	nts 🕅	Optimize	Data	G Utilities	Application
										Circulat	1.0					Teste		
0			E	jit	-			1		Simulat	ion		Results			10015		i. i
A1	•	×	Ē	jit Jatch Fit				1		Simula	ion		Results			10015		1

	December	- or of our ing	<u>I</u> courto					
Data Set								
Name								
Range				<b> </b> ¢				
	Values are Dates							
Data Type								
Ontinuous Sample	e Data							
O Discrete Sample D	ata							
O Discrete Sample D	ata (Counted	Format)						
O Density (X,Y) Points (Unnormalized)								
O Density (X,Y) Poin	ts (Normalized	d)						
Cumulative (X,P) Points								
Filter								
Туре	None			•				

ata Distributions Bootstrap	Chi-Sq Binning	Results	
tting Method Paramet	er Estimation		•
Lower Limit	V Be	taGeneral	
<ul> <li>Fixed Bound</li> <li>Bounded, But Unknown</li> <li>Open (Extends to -Infinity)</li> <li>Unsure</li> </ul>	Bu Ca Ca Da Eri Eri Eri Eri	rr 12 auchy aisq agum f lang pon	
Upper Limit  Fixed Bound  Fixed Bound  Open (Extends to +Infinity)  Unsure  Advanced Options  Fixed Parameters  Spec  Spec  Suppress Questionable Fits	V Ex V Ex Fa V Ga V Ga V In V La V La V Lo	tValue tValueMin tigueLife echet mma vpSecant vGauss maraswamy place vy gistic gLogistic	Ţ



#### **Distribution Fitting in @RISK 8.0 (cont.)**

@RISK - Fit Distributions to D	)ata		2
Data Distributions Bootstrap	Chi-Sq Binning	Results	
Run Parametric Bootstrap			
Number of Resamples	1000		
Parameter Confidence Level	95%		
computationally	, intensive	forla	
compationany	11100110110		rge
datasets, especi	ally when	done	rge for
datasets, especi lots of d	ally when istribution	done s	rge for
datasets, especi lots of d	ally when istribution	done s	rge for
datasets, especi lots of d	ally when istribution	done s	rge for

O Equal I	ntervals	Number of Bins	Auto	-
Equal P	robabilities			
O Custom	1			
Equal Inter	val Binning Opt	ions		
🗸 Automa	tic Minimum and	Maximum Based on Inpu	ut Data	
Minimu	m 0			
Maxim	um 1			
Extend	First Bin From M	finimum to -Infinity		
Extend	Last Bin From M	laximum to +Infinity		





#### **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)



#### **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



#### **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



#### **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



#### **Questions?**



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



sales@palisade.com



#### Input Distribution Selection with @RISK

- Part II Revisiting the Triangular and PERT "Three Point" Distributions Fri, Jul 17<sup>th</sup>, 10am EDT
- Part III Turning Expert Opinion into Defensible Distributions Fri, Aug 14<sup>th</sup>, 10am EDT



### **Back Up Slides**



#### **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**





#### **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	<b>C</b> <sub>1-α</sub>
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	Êxi	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<sub>0</sub>: Beta (2, 3, 10, 50)



#### **Anderson-Darling Test**

$$A_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - \hat{F}(x)]^2 \psi(x) \hat{f}(x) dx$$
  
where  $\psi(x) = \frac{1}{\{\hat{F}(x)[1 - \hat{F}(x)]\}}$ 

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

α	1-α	<b>a</b> <sub>n,1-α</sub>
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 functionk is the number of parameters estimatedn is the number of sampled points

$$\begin{split} \boldsymbol{L}(\theta) &= p_{\theta}(X_1)p_{\theta}(X_2) \dots p_{\theta}(X_n) \\ \text{for a discrete distribution with single unknown parameter } \theta \\ \boldsymbol{L}(\theta) &= f_{\theta}(X_1)f_{\theta}(X_2) \dots f_{\theta}(X_n) \\ \text{for a continuous distribution with single unknown parameter } \theta \end{split}$$

