Monte Carlo, Resampling, And Other Estimation Tricks

Mauricio Aguiar

ti MÉTRICAS, President

IFPUG Immediate Past President

Agenda

- Introduction
- A Simple Example
- Another Example
- An Alternative
- Do It Yourself Monte Carlo

Introduction

Introduction
Estimates

- Estimates are quantitative projections of project characteristics such as:
  - **Product Size**
  - **Effort**
  - **Schedule**
  - **Quality**

Introduction
Uncertainty and Monte Carlo

- There is a certain degree of uncertainty in the input parameters of an estimation model
- We want to know how uncertainty may affect results
- This can be accomplished with the Monte Carlo method

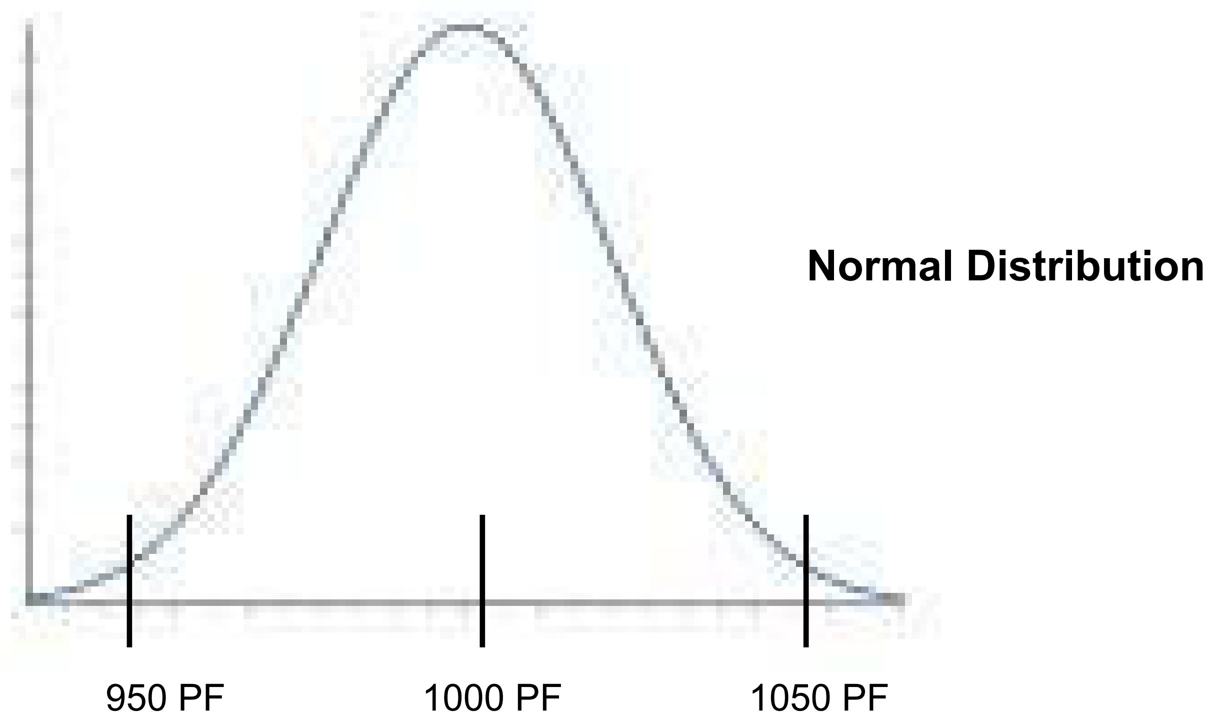
Introduction
Inputs to an Estimation Model

- Size (Function Points, etc.)
- Project and Product Characteristics
- Estimated Effort for Each Activity
- Etc.

Introduction
Modeling Uncertainty

- Allow inputs to vary according to specific statistical distributions

Example: Size



A Simple Example

A Simple Example - I

- Execute unit test and construction for 5 modules (classes, functions, subroutines...)

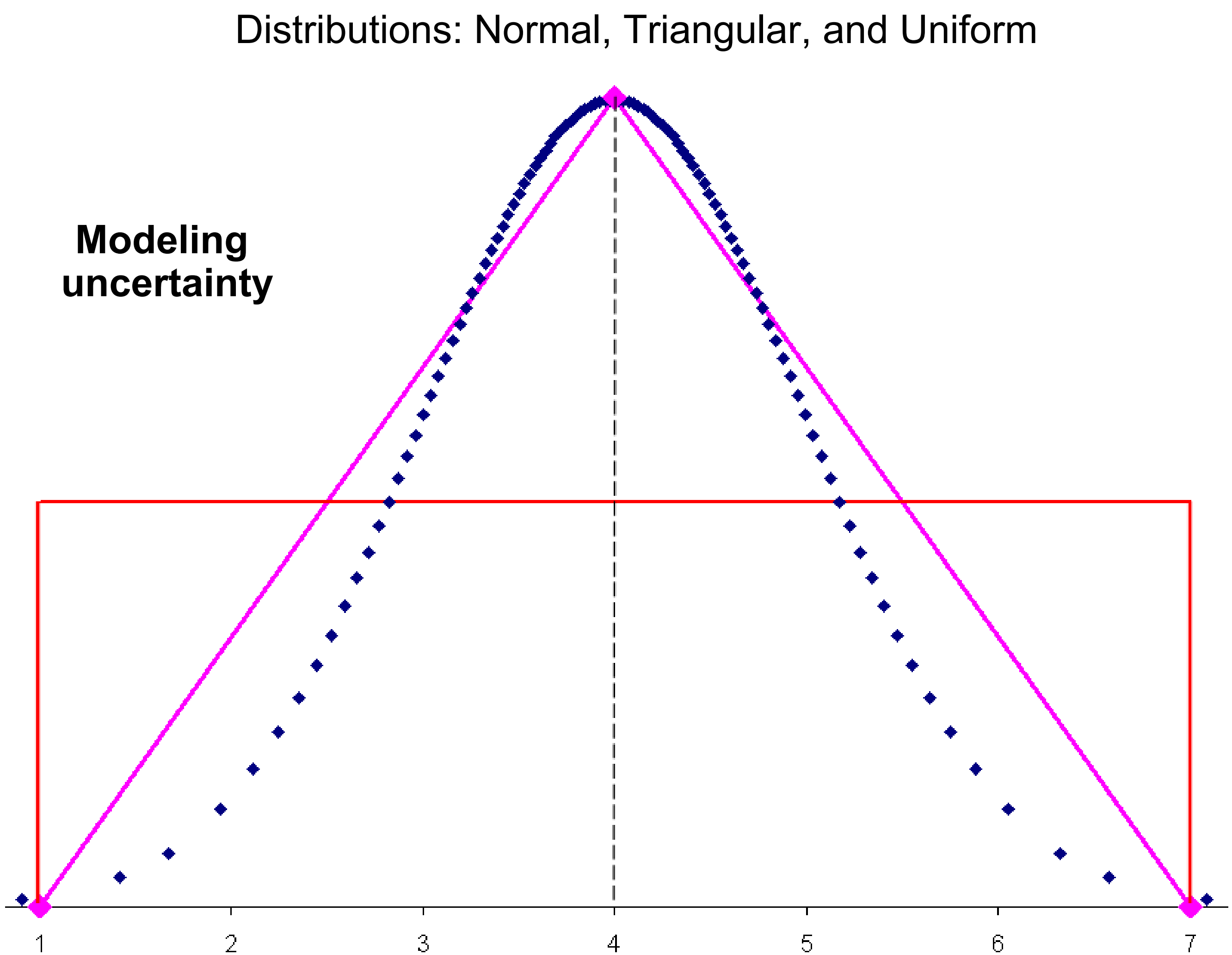
Module	Min (d)	Expected	Max(d)	Programmer	Estimate Quality
A	2	4	10	Arnie	Low
B	4	6	10	Ronnie	Low
C	8	12	16	Andy	Medium
D	3	5	6	Paul	High
E	2	4	6	Norman	Medium
Totals	**19**	**31**	**48**		

- Assume work will be done sequentially

A Simple Example – II
The Problem

- The 100% probability schedule would be **48** days
- A shorter schedule with 90% probability would be better
- What would that schedule be?

A Simple Example – III
Estimate Quality



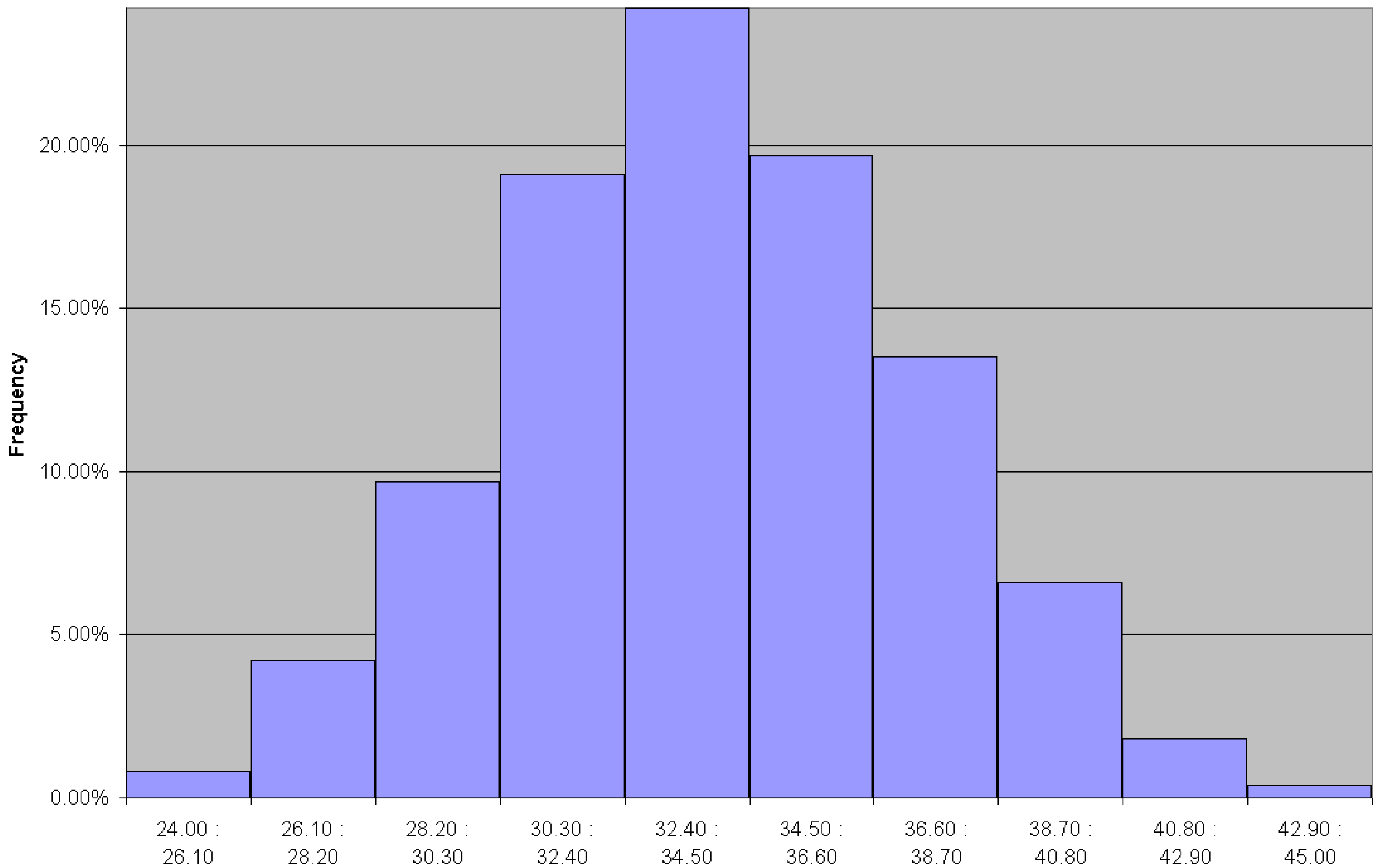
A Simple Example – IV
Monte Carlo

- Simulate the construction and unit test of the 5 modules 10000 times
  - **Let individual effort vary according to the corresponding distribution**
  - **Assess total effort variation**

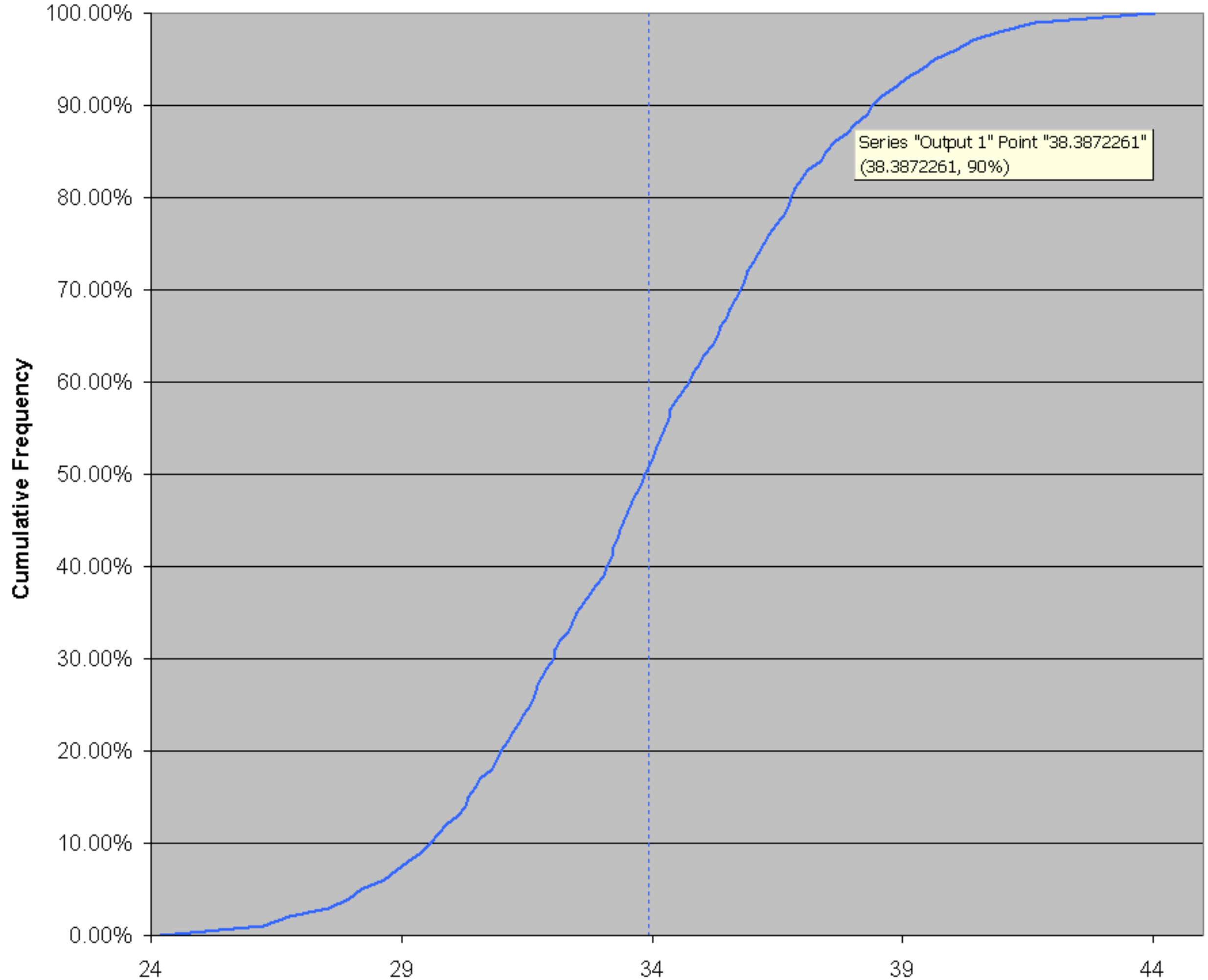
A Simple Example – V
Monte Carlo

Module	Min (d)	Expected	Max(d)	Programmer	Simulated	Distribution
A	2	4	10	Arnie	3.31064	Uniform
B	4	6	10	Ronnie	6.95388	Uniform
C	8	12	16	Andy	11.16812	Triangular
D	3	5	6	Paul	5.81785	Normal
E	2	4	6	Norman	3.68017	Triangular
Totals	**19**	**31**	**48**		**30.93066**	

A Simple Example – VI
Histogram of Simulated Schedule



A Simple Example – VII
Cumulative Frequency of Simulated Scheduled



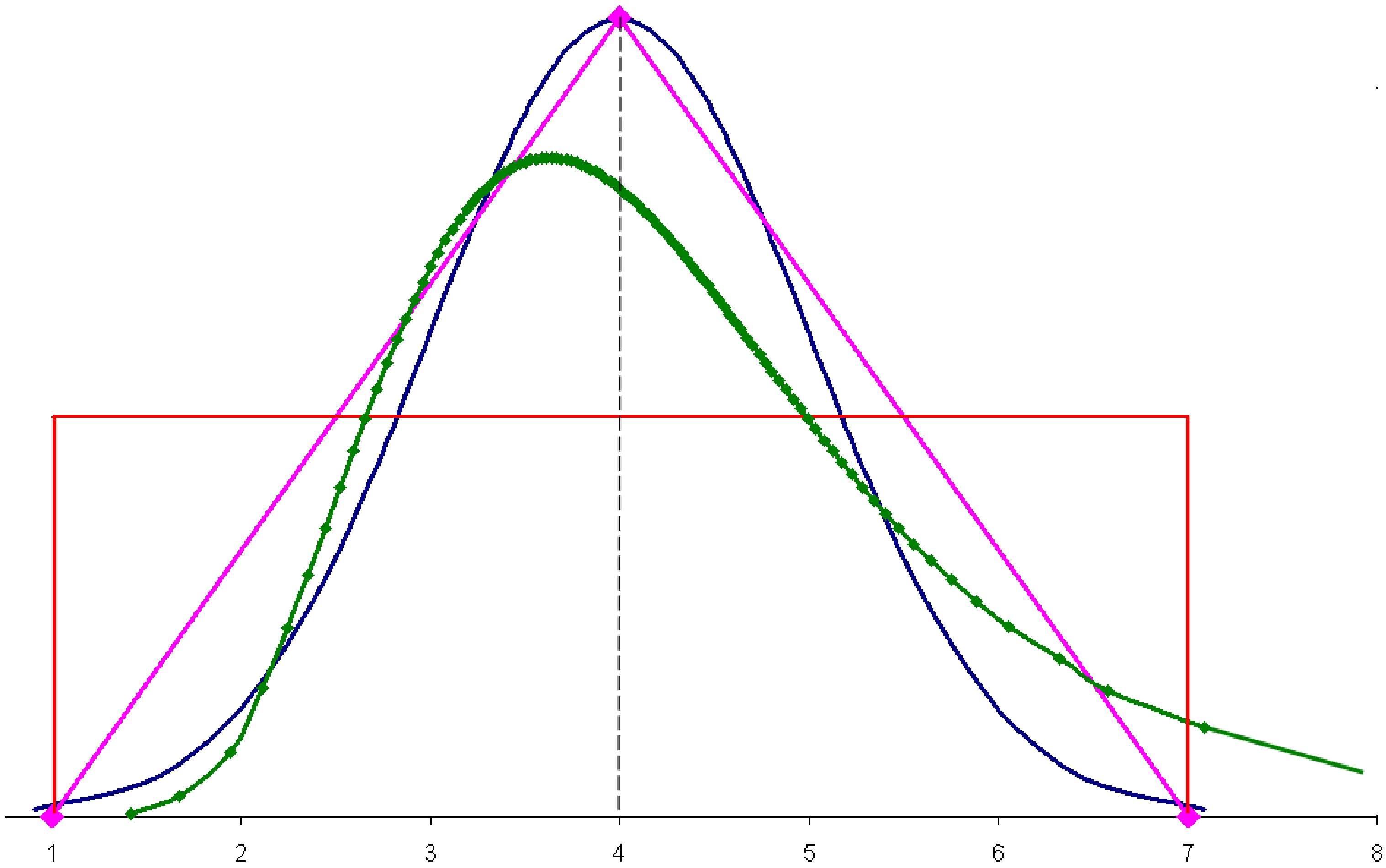
Schedule less or equal to 38 days with 90% probability

A Simple Example – VIII
Questions

- Do the distributions used reflect reality?
- Is a shorter schedule as likely to occur as a longer one?

A Simple Example – IX
Questions

Distributions: Normal, Triangular, Uniform, and Lognormal



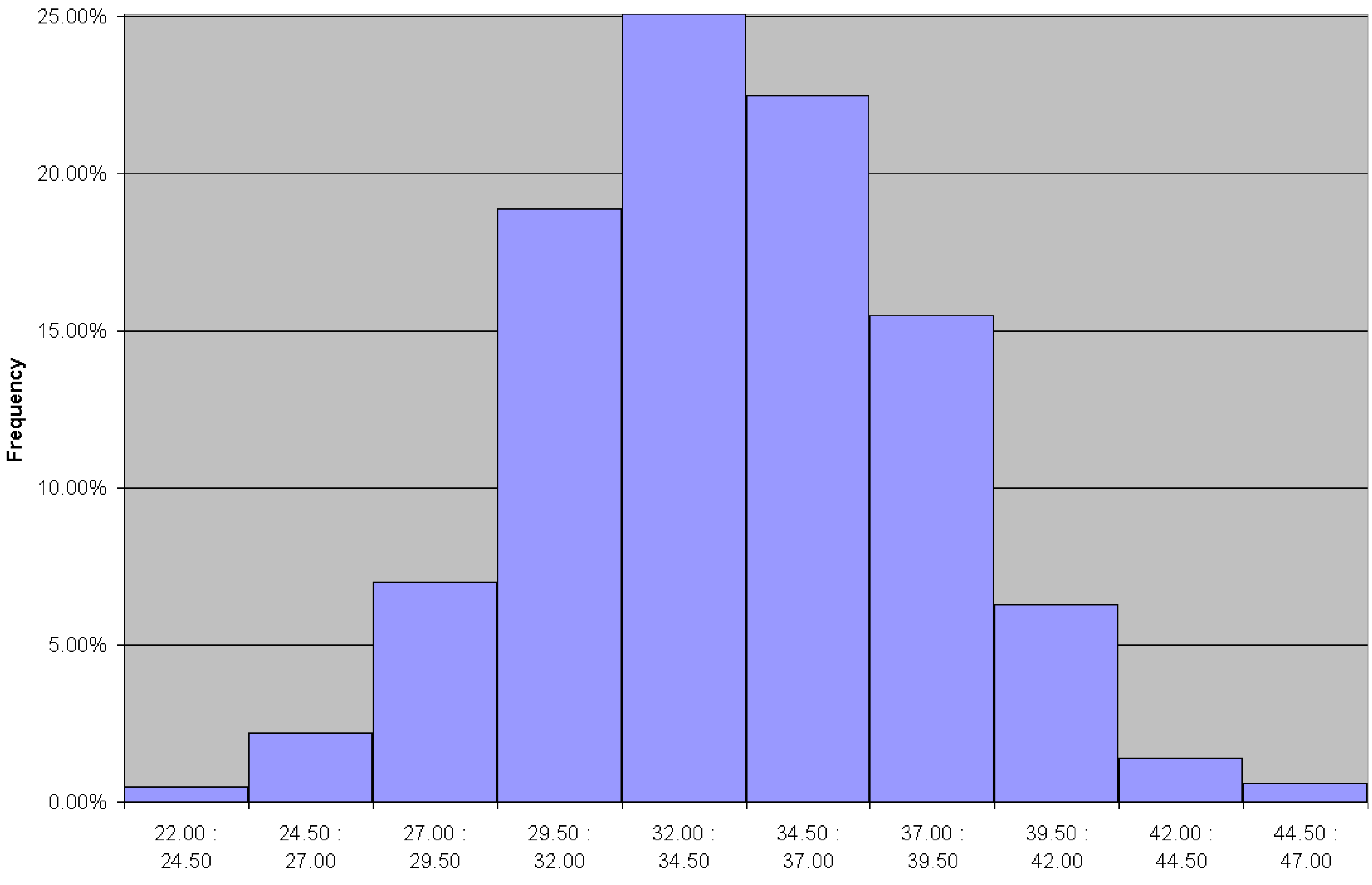
A Simple Example – X
Questions

- Simulate construction and unit test 10000 times substituting the lognormal distribution for the normal distribution

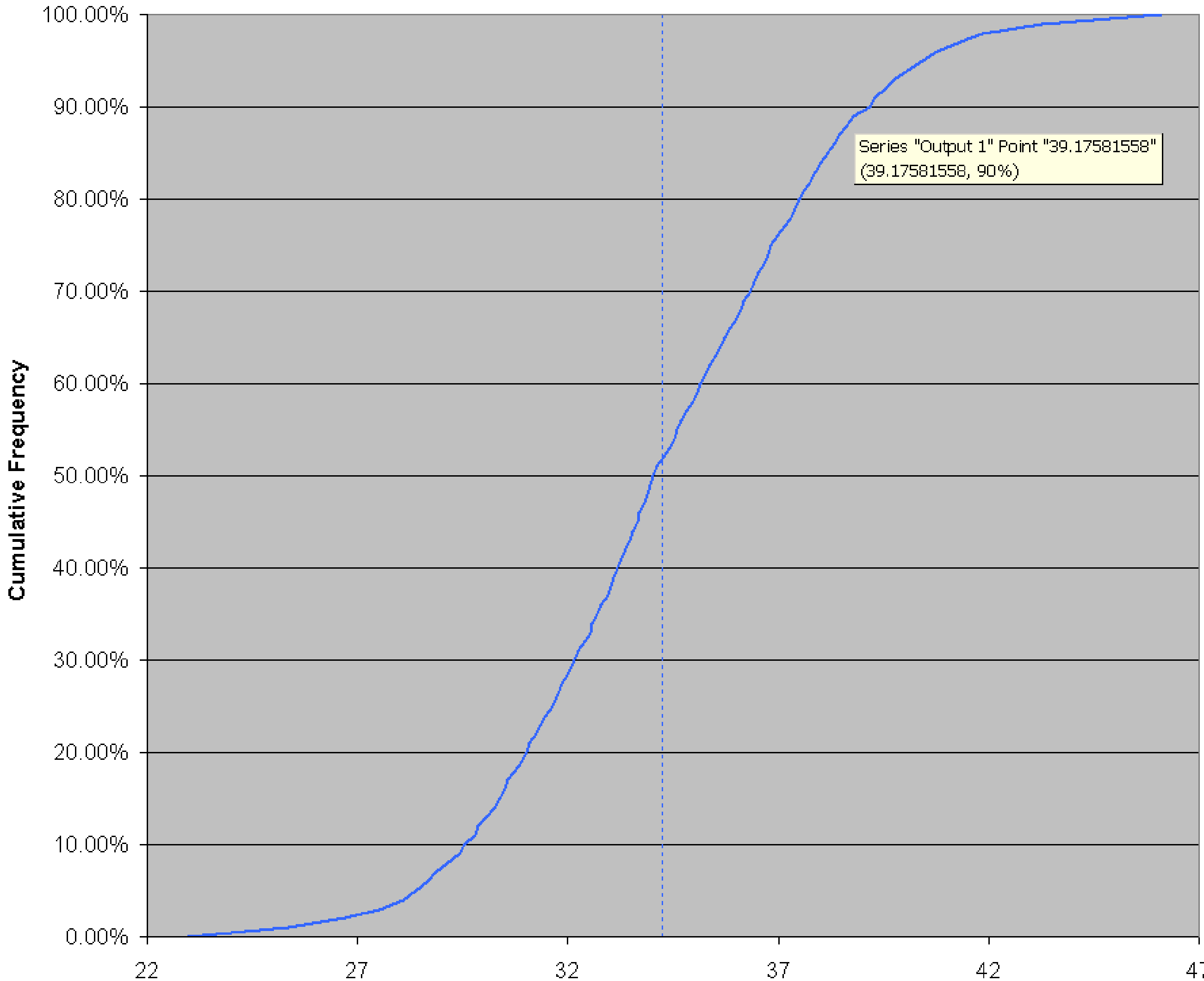
A Simple Example – XI
Questions

Module	Min (d)	Expected	Max(d)	Programmer	Simulated	Distribution
A	2	4	10	Arnie	9.21425	Uniform
B	4	6	10	Ronnie	5.52832	Uniform
C	8	12	16	Andy	9.84395	Triangular
D	3	5	6	Paul	3.47258	**Lognormal**
E	2	4	6	Norman	4.867	Triangular
Totals	**19**	**31**	**48**		**32.9261**	

A Simple Example – XII
Questions



A Simple Example – XIII
Questions



Schedule now is 39 days

Another Example

Another Example – I
Productivity - 9 Projects

Project	Size (FP)	Effort	Productivity
P1	98	1910	19.5
P2	184	2760	15.0
P3	212	2010	9.5
P4	196	1620	8.3
P5	261	1855	7.1
P6	257	1980	7.7
P7	430	7830	18.2
P8	190	1740	9.2
P9	310	4890	15.8
		Average	**12.2**

Productivity in Hours/Function Points

Note: Not real data

Another Example – II
Estimation Error

((Estimated – Actual)/Actual) * 100

Project	Size (FP)	Effort	Productivity	Estimate	% Abs Error
P1	98	1910	19.5	1200	37.2%
P2	184	2760	15.0	2253	18.4%
P3	212	2010	9.5	2596	29.1%
P4	196	1620	8.3	2400	48.1%
P5	261	1855	7.1	3195	72.3%
P6	257	1980	7.7	3147	58.9%
P7	430	7830	18.2	5265	32.8%
P8	190	1740	9.2	2326	33.7%
P9	310	4890	15.8	3795	22.4%
		Average	**12.2**	**MRE**	**39.2%**

Mean Relative Error

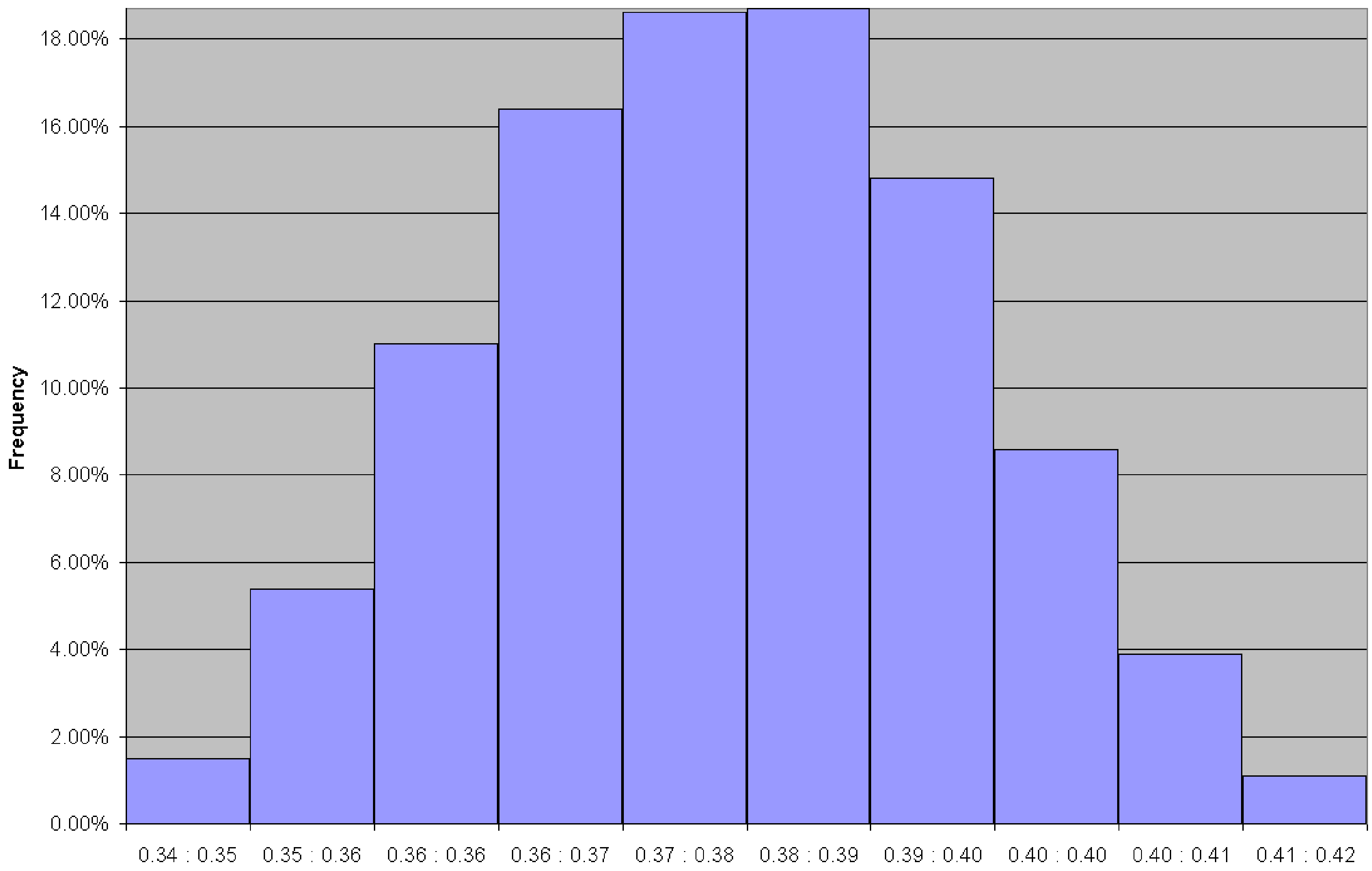
Another Example – III
Effort Data

- The quality of effort data is often low
- How would error in effort data affect the estimate?

Another Example – IV
Effort Data

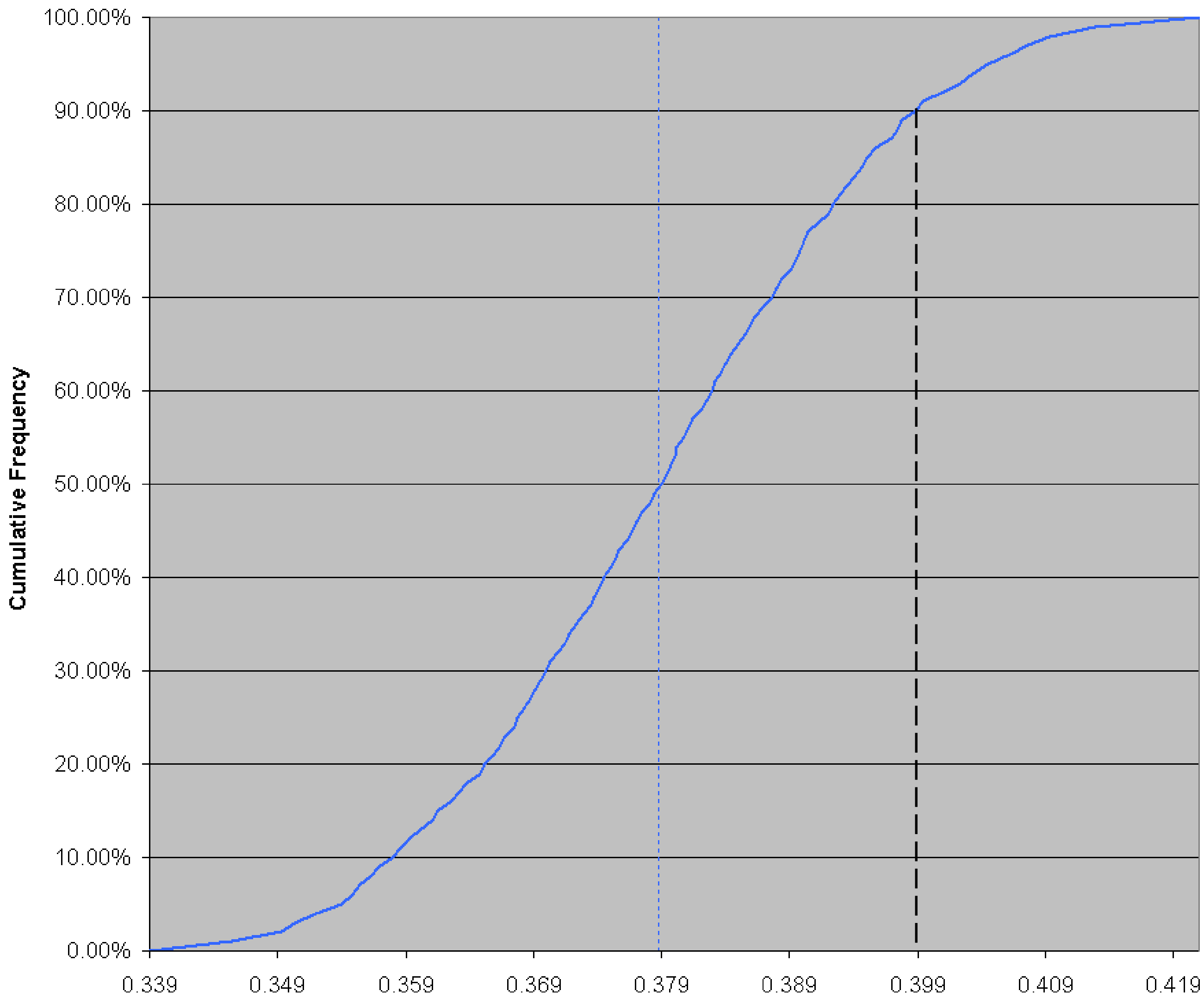
Proj	PF	Quality	Min	Exp	Max	Distribution	Effort	Prod	Est	% Abs Err
P1	98	Low	1600	1910	2100	Uniform	2094	21.4	1238	35.2%
P2	184	Low	2500	2760	3000	Uniform	2621	14.2	2325	15.8%
P3	212	High	1950	2010	2100	Normal	2013	9.5	2679	33.3%
P4	196	Medium	1500	1620	1800	Triangular	1671	8.5	2477	52.9%
P5	261	Low	1500	1855	2200	Uniform	2122	8.1	3298	77.8%
P6	257	Low	1500	1980	2400	Uniform	1681	6.5	3247	64.0%
P7	430	Low	5000	7830	9000	Uniform	8872	20.6	5433	30.6%
P8	190	High	1700	1740	1800	Normal	1707	9.0	2401	38.0%
P9	310	Medium	4600	4890	4900	Triangular	4899	15.8	3917	19.9%
							Average	12.6	MRE	40.8%

Another Example – V
Effort Data



ti MÉTRICAS

Another Example – VI
Effort Data



Less than 40% error with 90% probability.

Maximum 42%.

An Alternative

Resampling - I
The Idea

- Monte Carlo requires assumptions about the distributions
- Resampling is based on replicating a sample a large number of times with replacement
- Statistics based on resampling will approach true values as the number of samples grow
- Resampling is independent of distribution assumptions

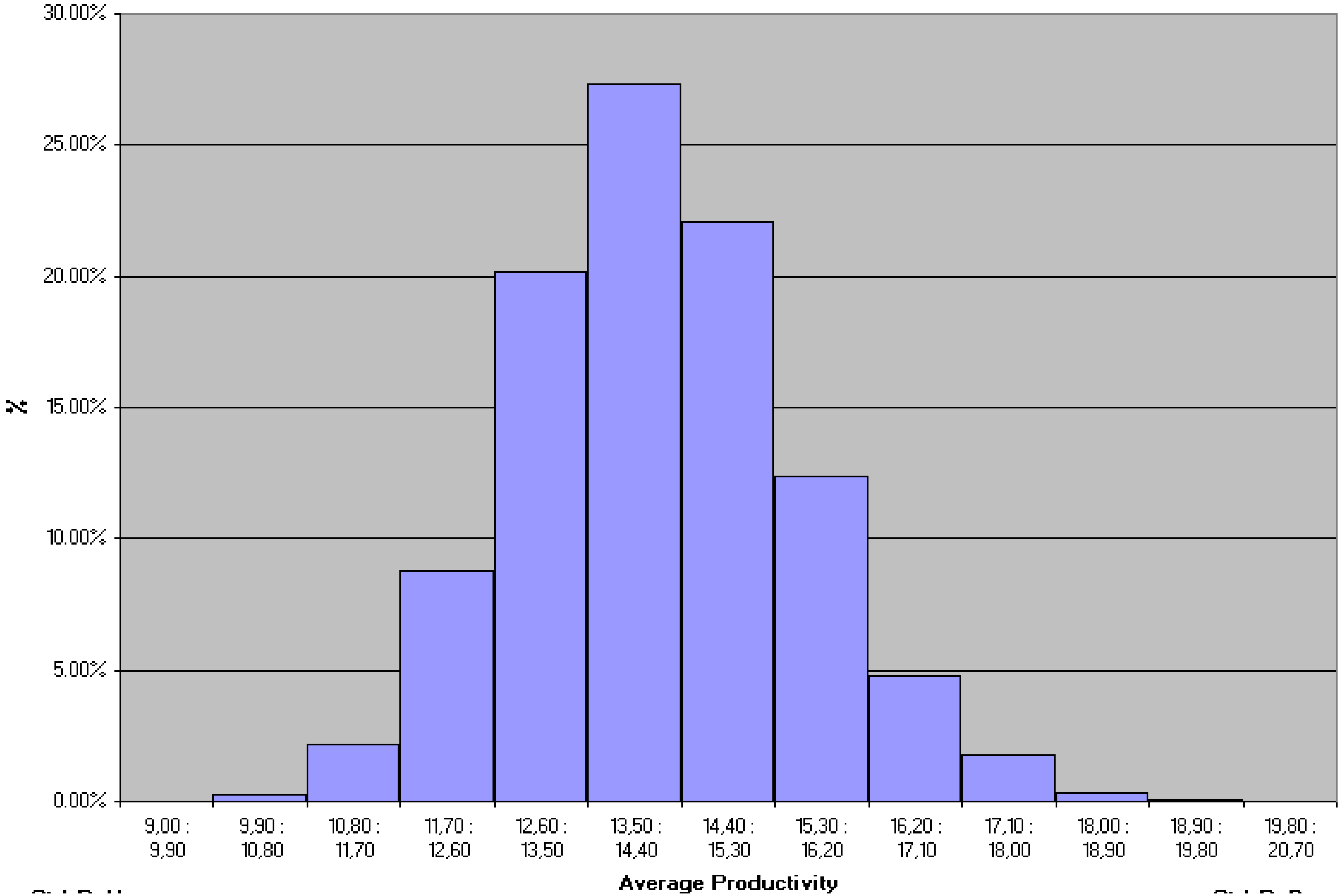
Resampling - II
An Example

- Build a 95% confidence interval for COBOL productivity based on ISBSG V10 data
- The data: 615 projects, average productivity **20.5** H/FP, median productivity **11.9** H/FP

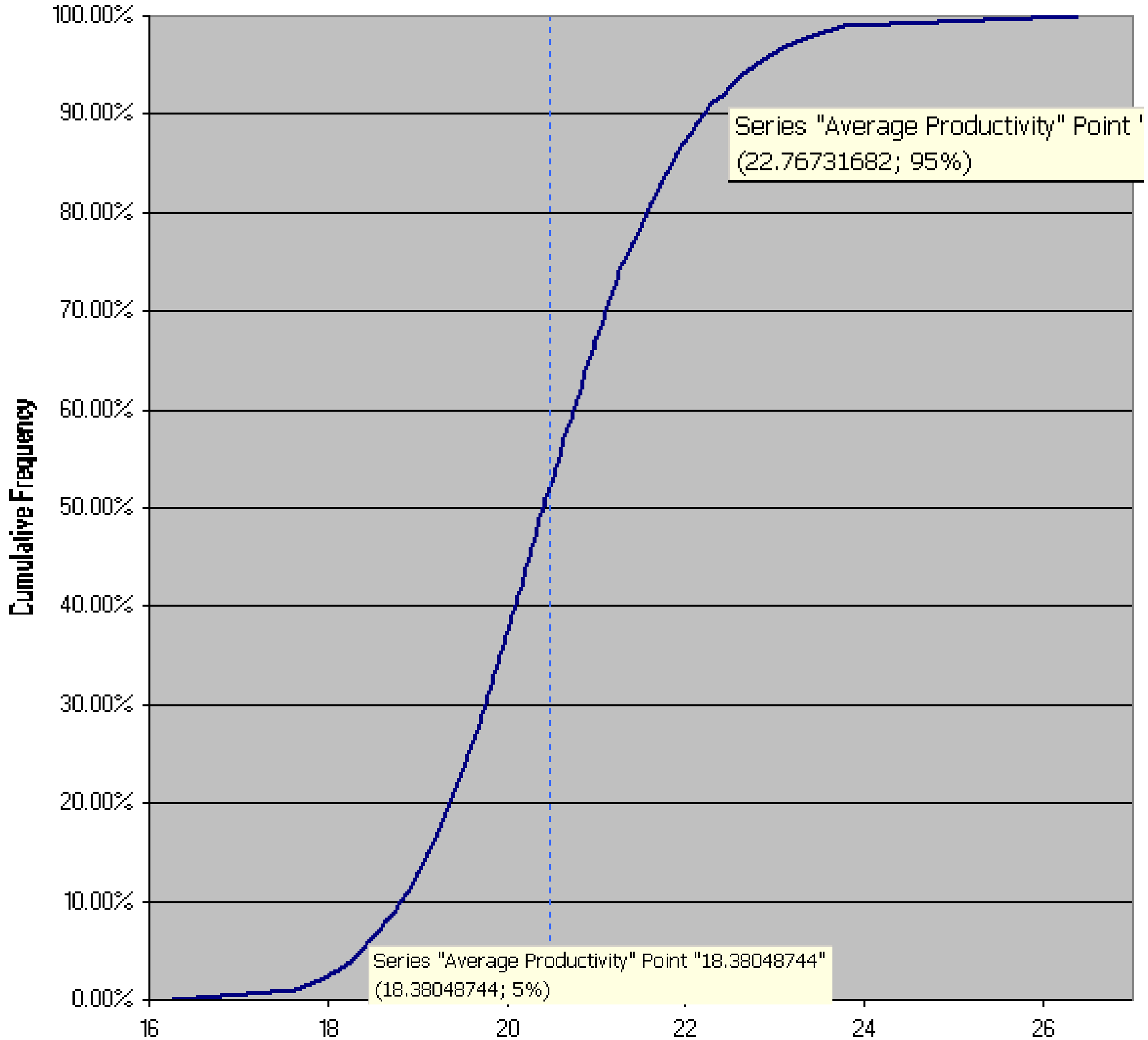
Resampling - III
Sampling Distribution: Mean, Median

	Average Productivity	Median Productivity
Average	**20.47295952**	**11.70256901**
Std Dev	1.32359835	0.76645153
Std Err	0.013235984	0.007664515
Max	26.36926842	14.30000019
Min	16.26861763	9.399999619

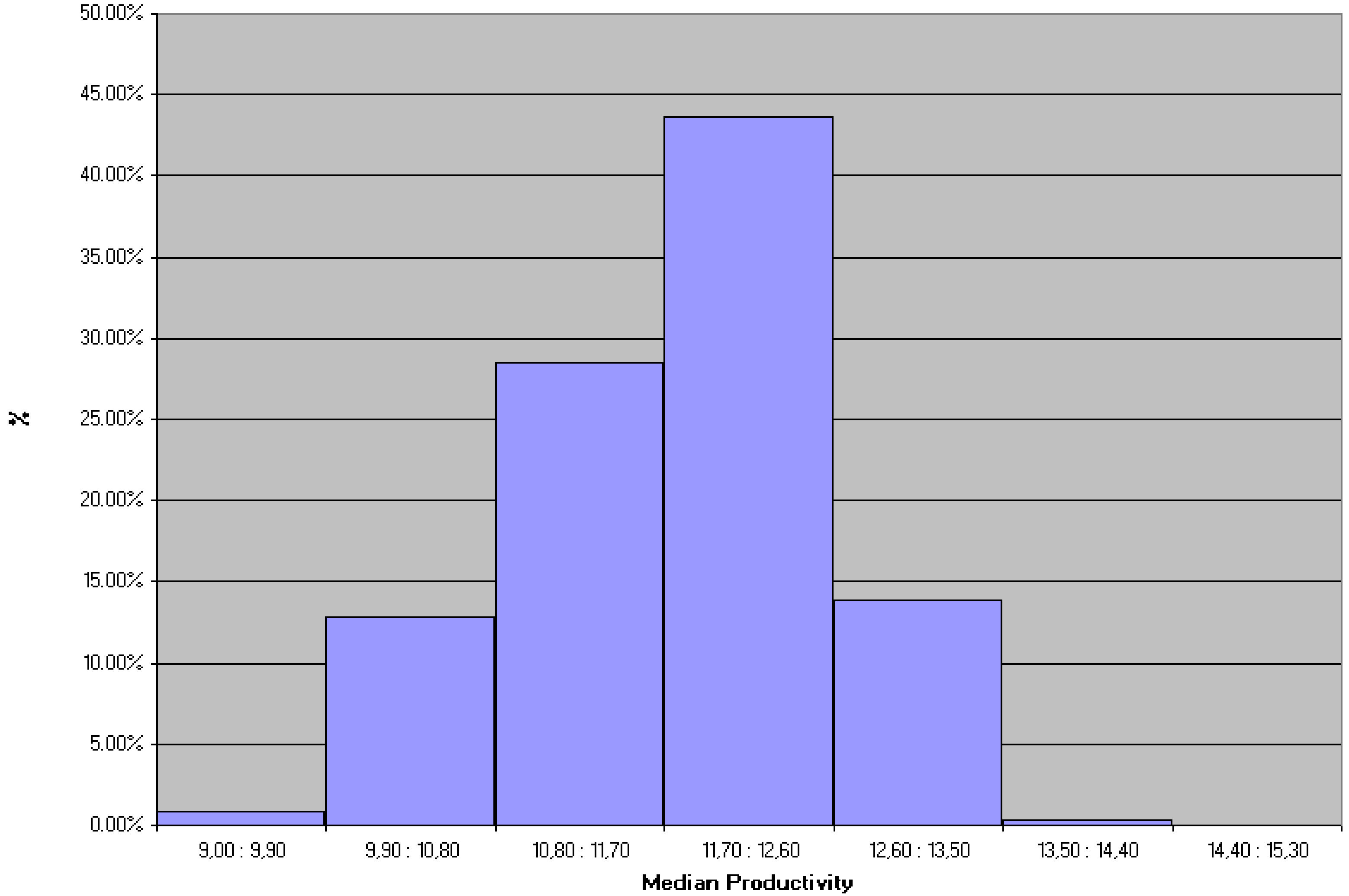
Resampling – IV
Sampling Distribution: Mean



Resampling – V
Sampling Distribution: Mean

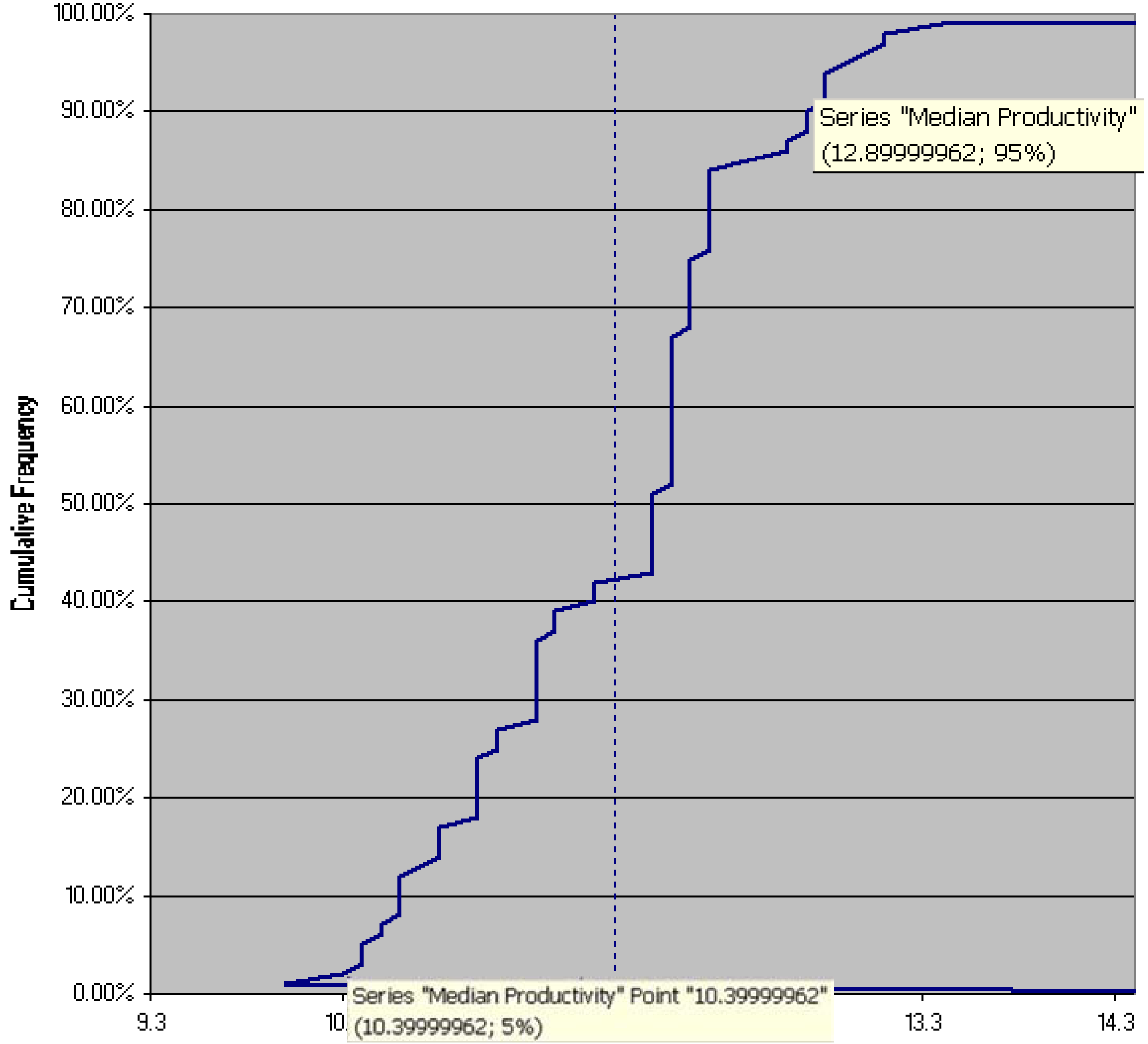


Resampling - VI
Sampling Distribution: Median



tiMÉTRICAS

Resampling - VI
Sampling Distribution: Median



ti MÉTRICAS

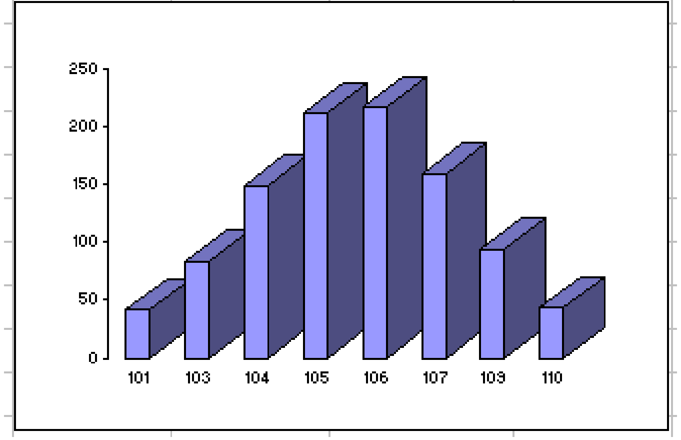
Do It Yourself

Do It Yourself - I
Simulating the Triangular Distribution

	A	B	C	D	E
1	**Simulating the Triangular Distribution**				
2					
3	**Iterations**	1000	Simulate		
4					
5	**Minimum**	100			
6	**Expected**	105			
7	**Maximum**	110			
8					
9	**Uniform**	0.93631			
10	**Triangular**	108.2155			
11					

Do It Yourself - II
Simulating the Triangular Distribution

Minimum	100.2054	Average	105.077
Maximum	109.821	Median	105.063
Range	9.615588		
Freqs	Bins	%	
42	101.4073	4.20%	4.20%
84	102.6093	8.40%	12.60%
149	103.8112	14.90%	27.50%
212	105.0132	21.20%	48.70%
217	106.2151	21.70%	70.40%
159	107.4171	15.90%	86.30%
94	108.619	9.40%	95.70%
43	109.821	4.30%	100.00%
1000		100.00%	



Do It Yourself - III
Simulating the Triangular Distribution

Working with Excel VBA

```
Function TriDist(ByVal prob As Single, ByVal opt As Single, ByVal expect As
'Esta função retorna um valor segundo a distribuição triangular com
'opt = valor mínimo
'expect = valor esperado
'pess = valor máximo

    Dim x, d As Single
    d = pess - opt
    x = (expect - opt) / d
    If prob <= x Then TriDist = opt + (((prob * x) ^ 0.5) * d)
    If prob > x Then TriDist = pess - ((((1 - prob) * (1 - x)) ^ 0.5) * d)

End Function
```

Do It Yourself - IV
Simulating the Triangular Distribution

```
Sub Simular()

    Dim intI, intNumIteracoes As Integer

    intNumIteracoes = Sheets("Main").Range("B3").Value
    If intNumIteracoes <= 0 Then
        MsgBox ("Número de iterações deve ser > 0")
        Exit Sub
    End If
    If intNumIteracoes > 10000 Then
        MsgBox ("Número máximo de iterações = 10000")
        Exit Sub
    End If

    'Resultados: Usar Colunas A, B, C, D, E
    Sheets("Dados").Select
    Sheets("Dados").Range("A27").Select
    ActiveWindow.FreezePanes = True
    Sheets("Dados").Range("A1").Select
    Sheets("Dados").Range("A1:E10000").Clear

    For intI = 1 To intNumIteracoes
        Application.StatusBar = "Processando Iteração... " & intI
        'Distr Triangular
        Sheets("Main").Range("B10").Copy
        Sheets("Dados").Cells(intI, 1).PasteSpecial Paste:=xlValues, _
                                                Operation:=xlNone, _
                                                SkipBlanks:=False, _
                                                Transpose:=False
    Next intI
    ActiveWindow.FreezePanes = False
    Application.StatusBar = False

End Sub
```

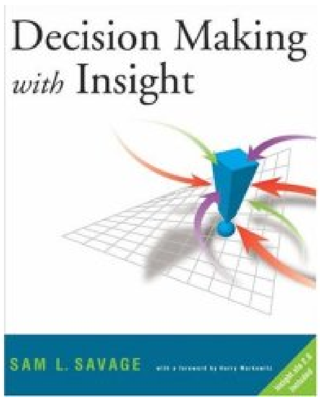
Summary

Summary
Monte Carlo and Resampling

- Monte Carlo helps to identify variation in the results as a function of the uncertainty in the inputs
- The choice of the correct statistical distribution is very important in Monte Carlo
- Resampling will work even when the distribution is unknown
- A lot can be done using Excel & VBA
- There are professional Monte Carlo tools (Crystal Ball, @Risk, XLSim, etc.)

Summary
References

- Savage, Sam L., *Decision Making with Insight*
  - **XLSim 2.0**



Decision Making with Insight (with Insight.xla 2.0 and CD-ROM) (Paperback)
by Sam L. Savage (Author)
(3 customer reviews)

List Price: ~~$73.95~~
Price: **$60.42** & this item ships for **FREE with Super Saver Shipping**. Details
You Save: $13.53 (18%)

- Mooney, C.Z. and Duval, R.D, *Bootstrapping: A Nonparametric Approach to Statistical Inference*, SAGE, 1993
- Davidson, A.C. and Hinkley, D.V., *Bootstrap Methods and Their Application*, Cambridge University Press, 1997

Thank You

Mauricio Aguiar

ti MÉTRICAS

mauricio@metricas.com.br

www.metricas.com.br