Triangular Distributions and Correlations

The simple math behind triangular distributions and correlations in Monte Carlo simulations

Jennifer Lampe Jeffrey Platten

June 9, 2015

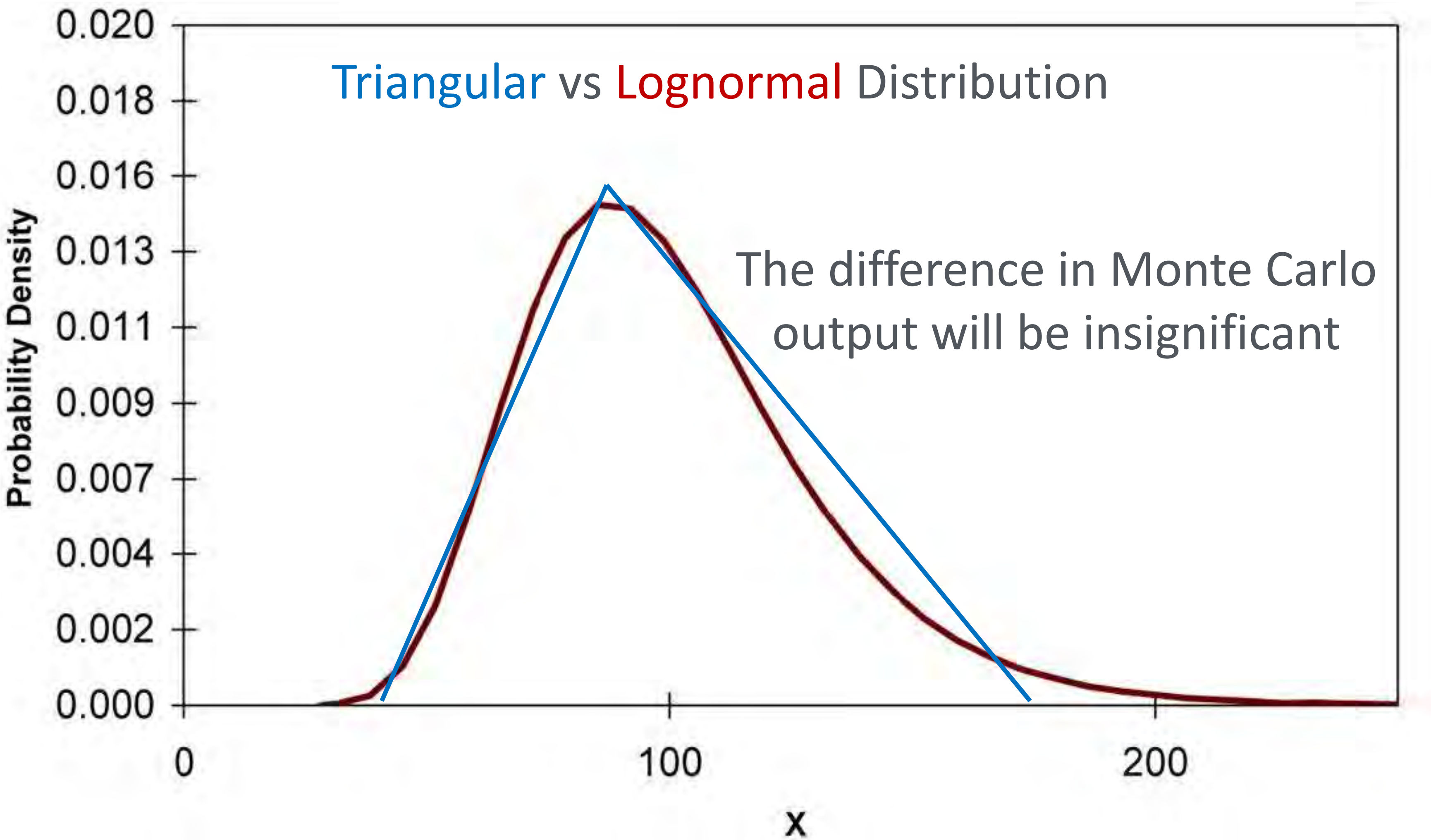
San Diego, CA

Agenda

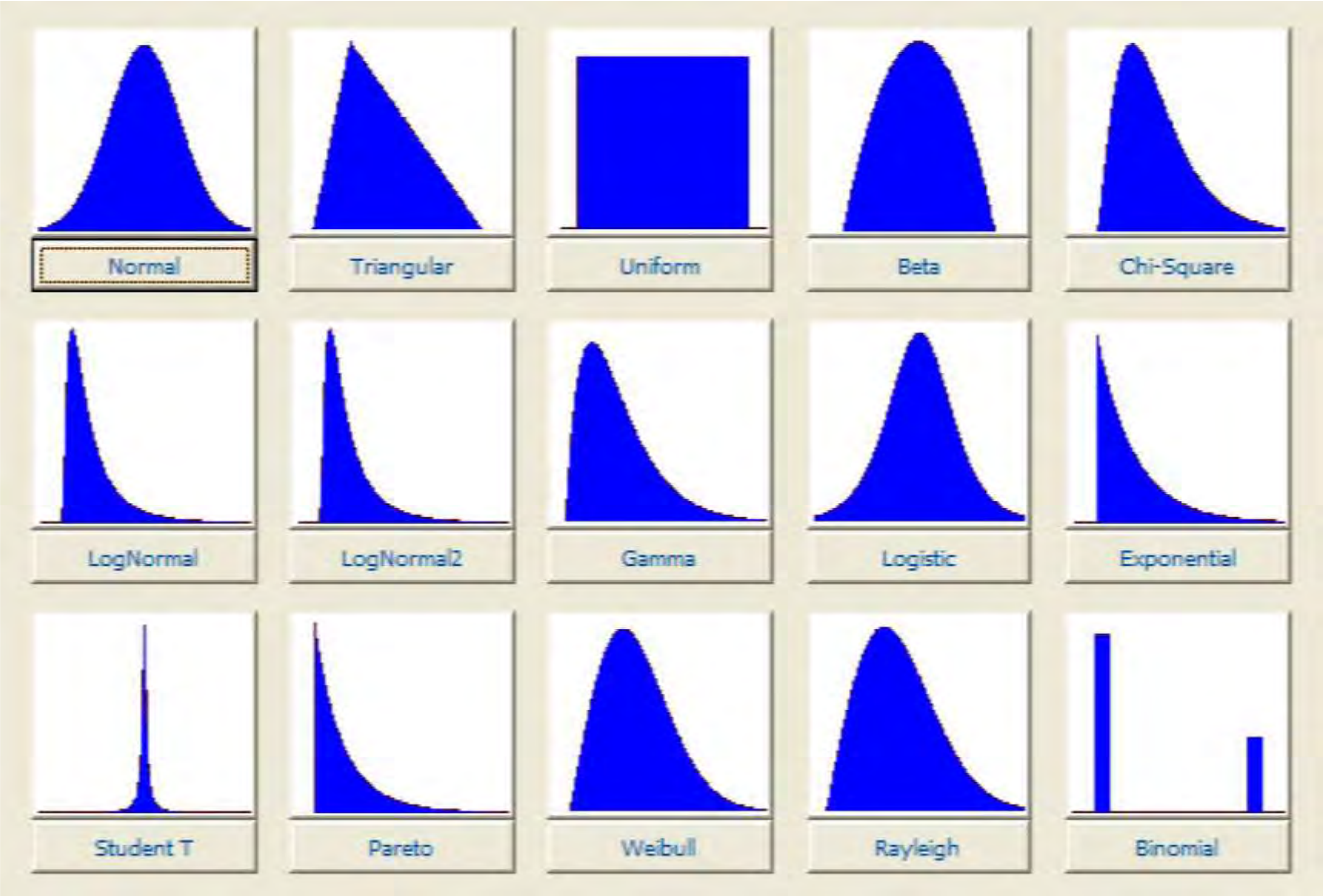
Types of distributions	3
Simple trigonometry	6
The risk adjusted mean	14
Using triangular distributions in Monte Carlo simulations	16
Correlation	21

Why use a triangular distribution ?

- Triangular distributions are often used in estimating cost risks because the math is relatively simple and because it nearly approximates a lognormal distribution

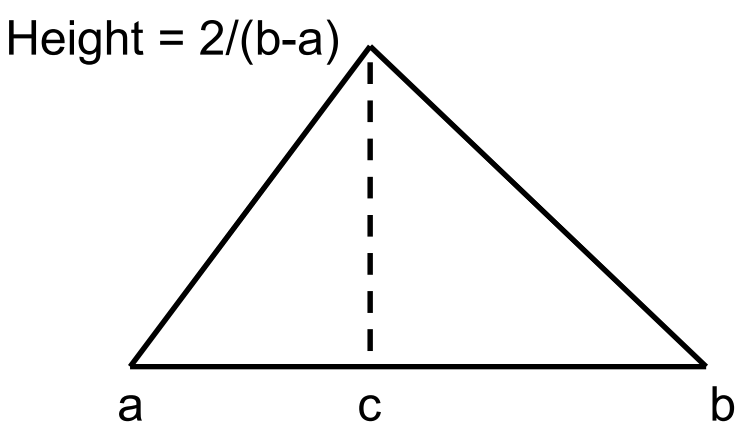


Some (mathematically) possible distributions

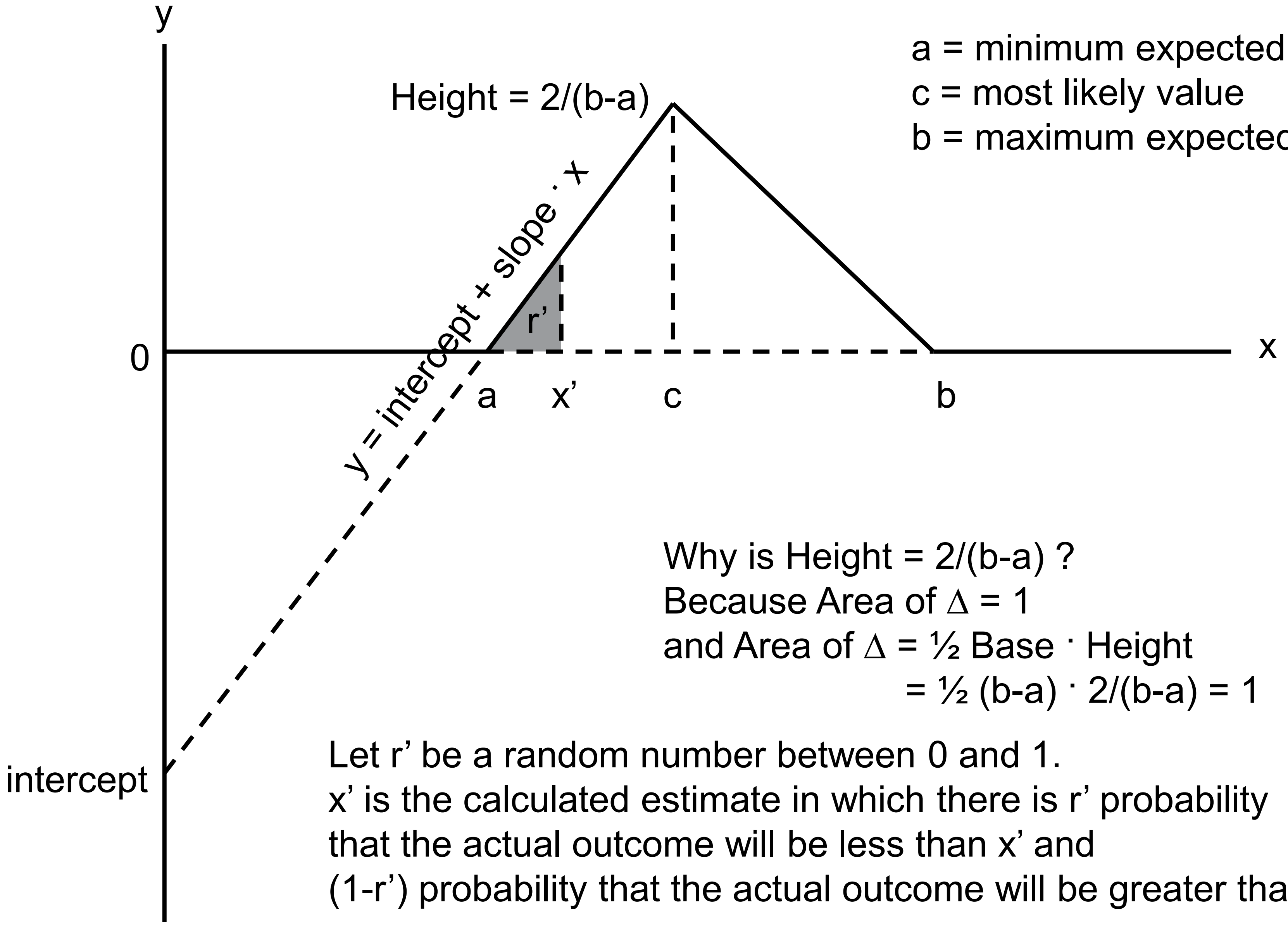


What is a triangular distribution ?

- A triangular area which visually represents mathematically the likelihood of possible outcomes by defining
  - lower limit (a)
  - upper limit (b)
  - mode (c) or most likely
- Implies that the likelihood increases consistently (on a straight line) as the estimate approaches the mode from either side
- The area of the triangle is 1 (or 100%, if a, b and c are percents)
- The base of the triangle is b-a
  - So the height of the triangle is 2/(b-a)



Triangular distribution simple trigonometry

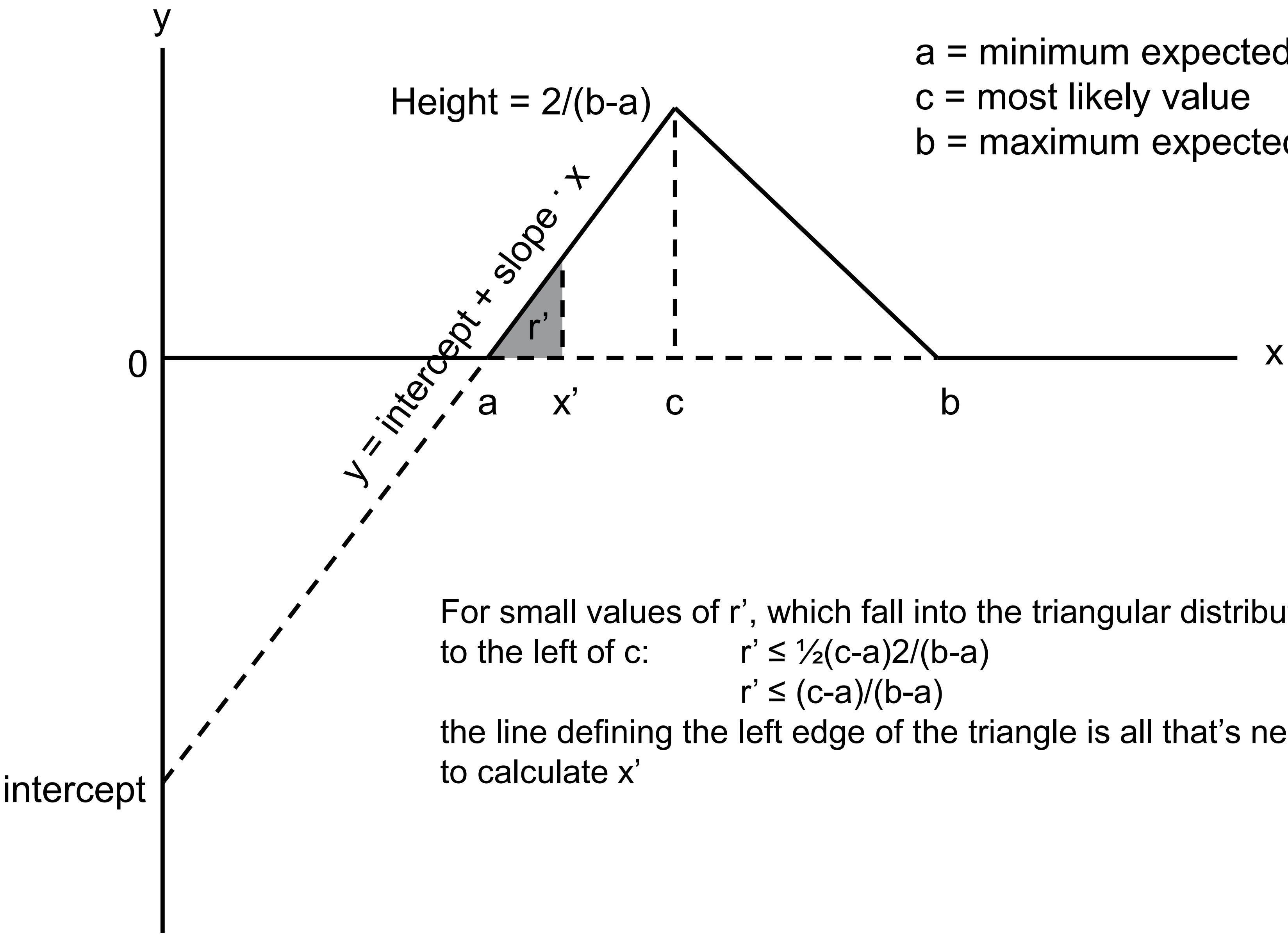


a = minimum expected value
c = most likely value
b = maximum expected value

Why is Height = 2/(b-a) ?
Because Area of Δ = 1
and Area of Δ = ½ Base · Height
= ½ (b-a) · 2/(b-a) = 1

Let r' be a random number between 0 and 1.
x' is the calculated estimate in which there is r' probability that the actual outcome will be less than x' and (1-r') probability that the actual outcome will be greater than x'

Triangular distribution simple trigonometry



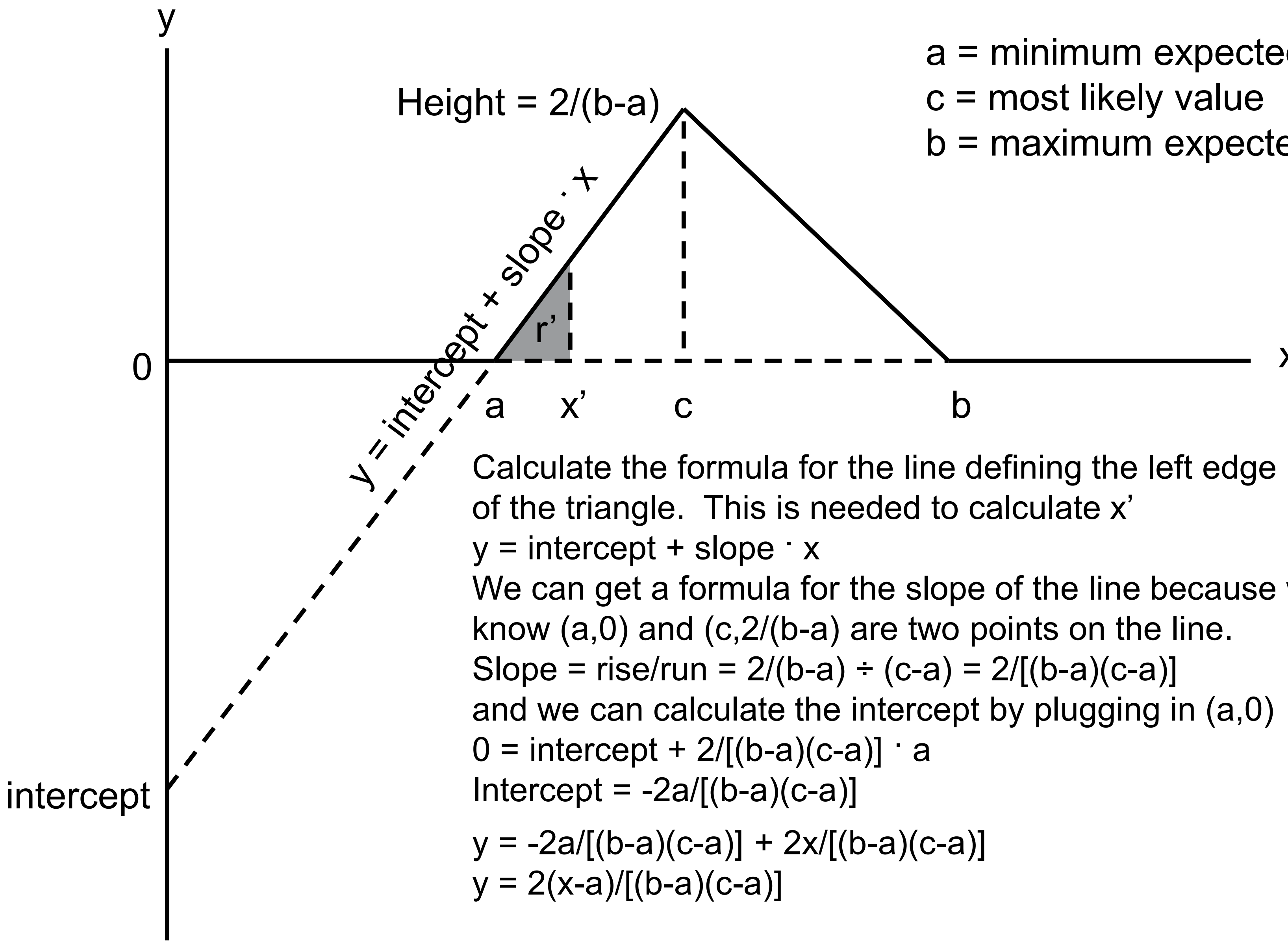
a = minimum expected value
c = most likely value
b = maximum expected value

For small values of r’, which fall into the triangular distribution to the left of c: r’ ≤ ½(c-a)2/(b-a)

r’ ≤ (c-a)/(b-a)

the line defining the left edge of the triangle is all that’s needed to calculate x’

Triangular distribution simple trigonometry



Calculate the formula for the line defining the left edge of the triangle. This is needed to calculate x'

$y = \text{intercept} + \text{slope} \cdot x$

We can get a formula for the slope of the line because we know (a,0) and (c,2/(b-a) are two points on the line.

$\text{Slope} = \text{rise/run} = 2/(b-a) \div (c-a) = 2/[(b-a)(c-a)]$

and we can calculate the intercept by plugging in (a,0)

$0 = \text{intercept} + 2/[(b-a)(c-a)] \cdot a$

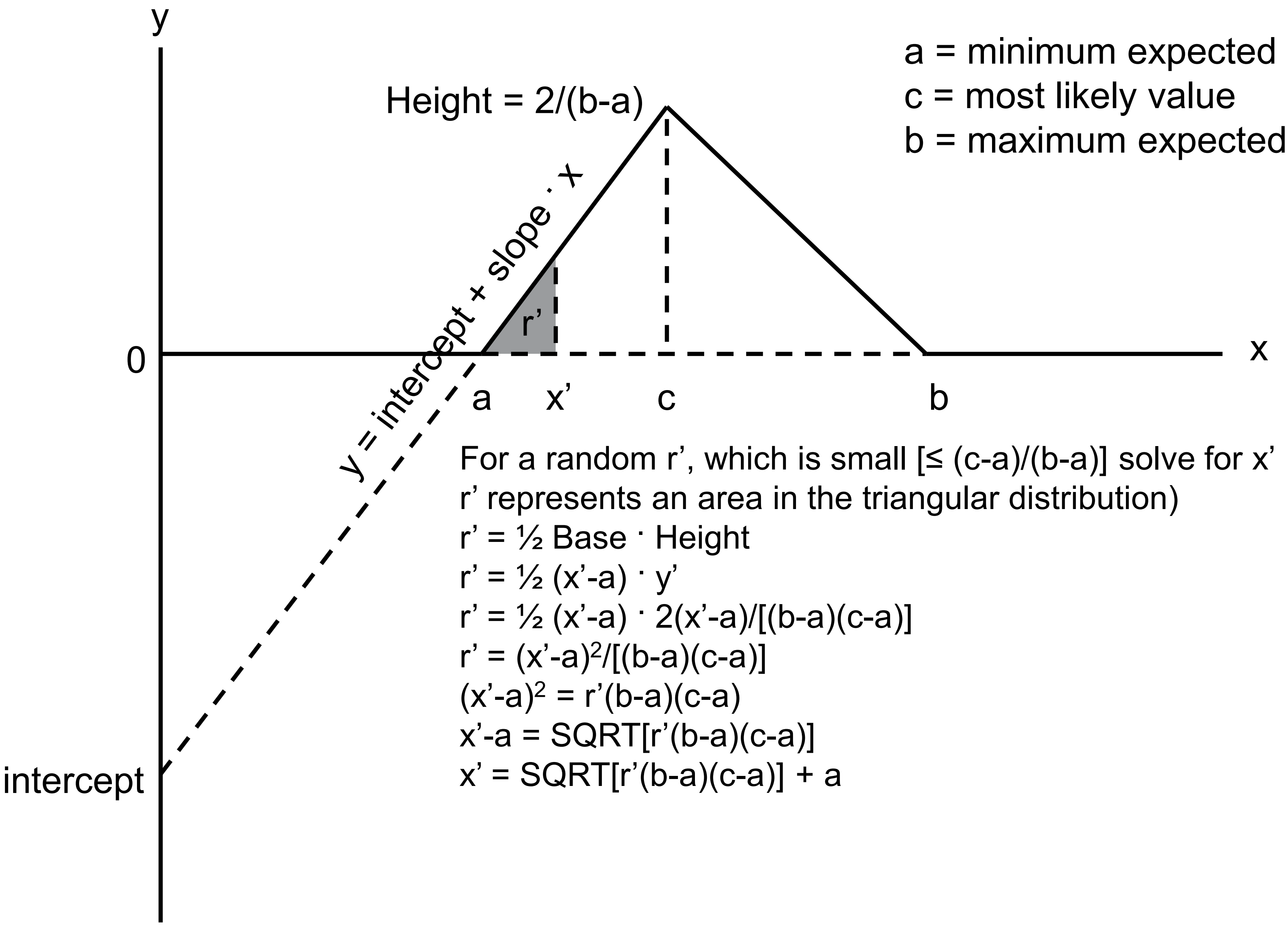
$\text{Intercept} = -2a/[(b-a)(c-a)]$

$y = -2a/[(b-a)(c-a)] + 2x/[(b-a)(c-a)]$

$y = 2(x-a)/[(b-a)(c-a)]$

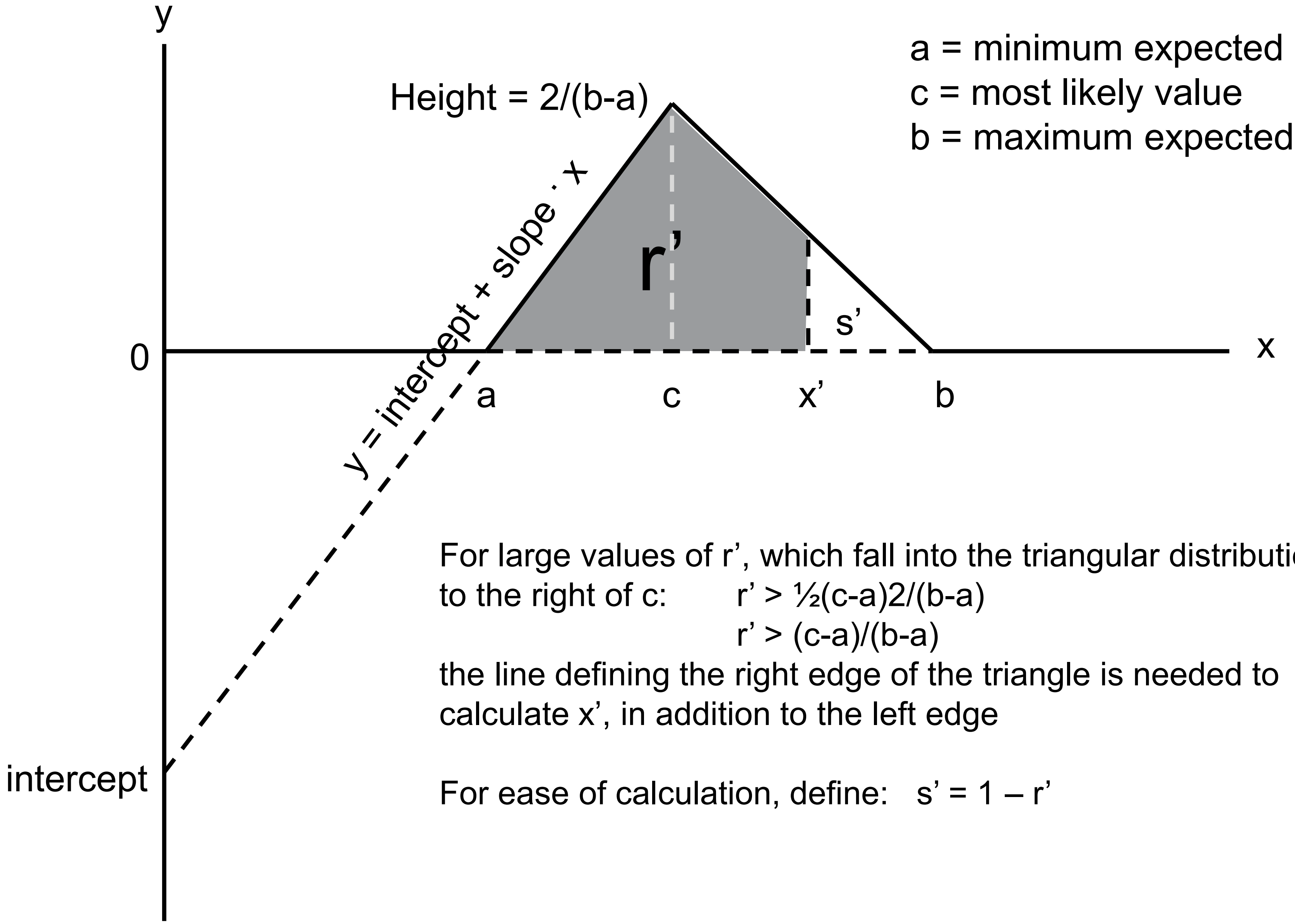
Triangular distribution simple trigonometry

a = minimum expected value
c = most likely value
b = maximum expected value



For a random r’, which is small [≤ (c-a)/(b-a)] solve for x’
r’ represents an area in the triangular distribution)
r’ = ½ Base · Height
r’ = ½ (x’-a) · y’
r’ = ½ (x’-a) · 2(x’-a)/[(b-a)(c-a)]
r’ = $(x'-a)^2$/[(b-a)(c-a)]
$(x'-a)^2$ = r’(b-a)(c-a)
x’-a = SQRT[r’(b-a)(c-a)]
x’ = SQRT[r’(b-a)(c-a)] + a

Triangular distribution simple trigonometry



a = minimum expected value
c = most likely value
b = maximum expected value

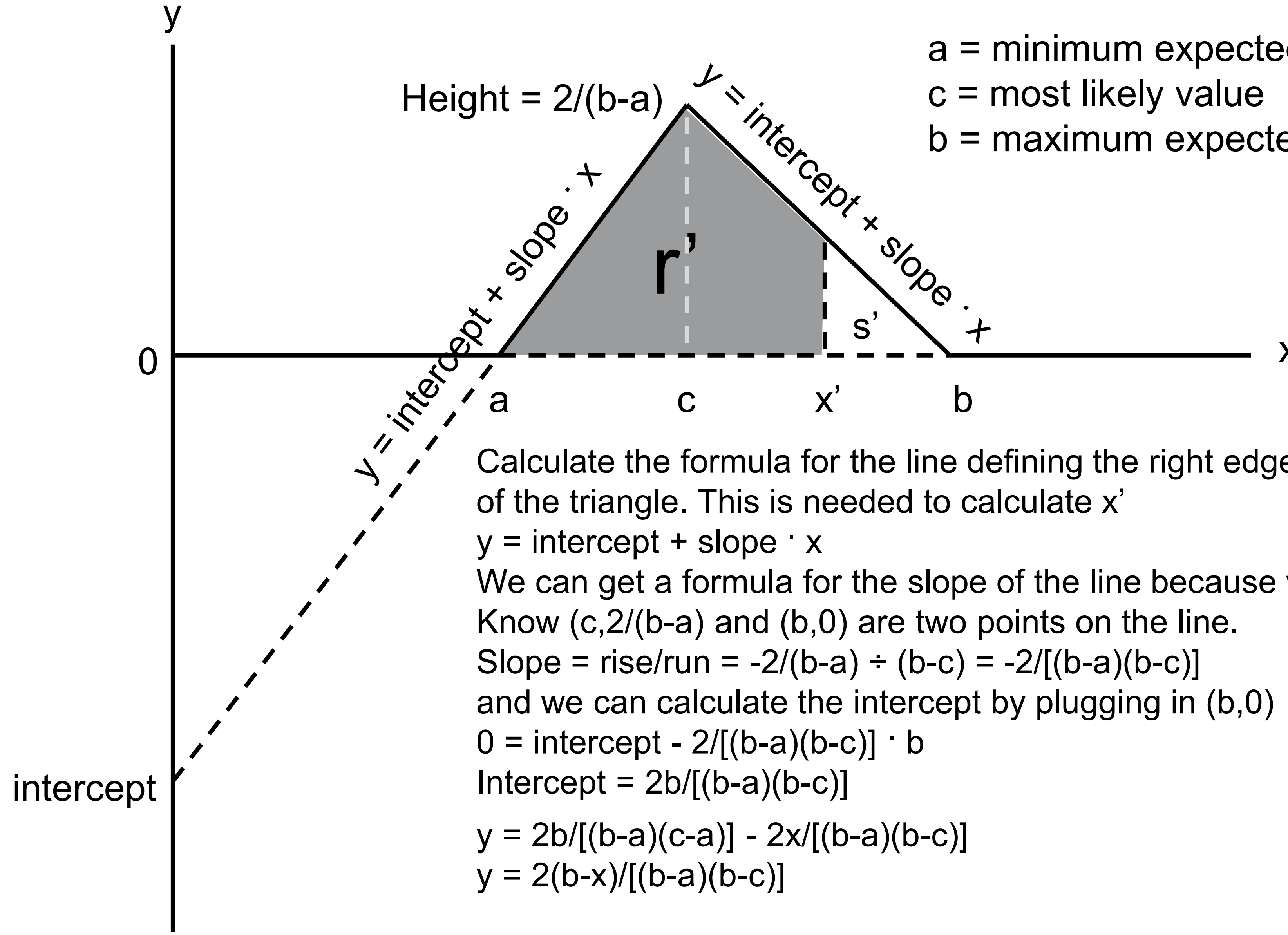
For large values of r', which fall into the triangular distribution to the right of c: r' > ½(c-a)2/(b-a)

r' > (c-a)/(b-a)

the line defining the right edge of the triangle is needed to calculate x', in addition to the left edge

For ease of calculation, define: s' = 1 – r'

Triangular distribution simple trigonometry



a = minimum expected value
c = most likely value
b = maximum expected value

Calculate the formula for the line defining the right edge of the triangle. This is needed to calculate x’

y = intercept + slope · x

We can get a formula for the slope of the line because we Know (c,2/(b-a) and (b,0) are two points on the line.

Slope = rise/run = -2/(b-a) ÷ (b-c) = -2/[(b-a)(b-c)]

and we can calculate the intercept by plugging in (b,0)

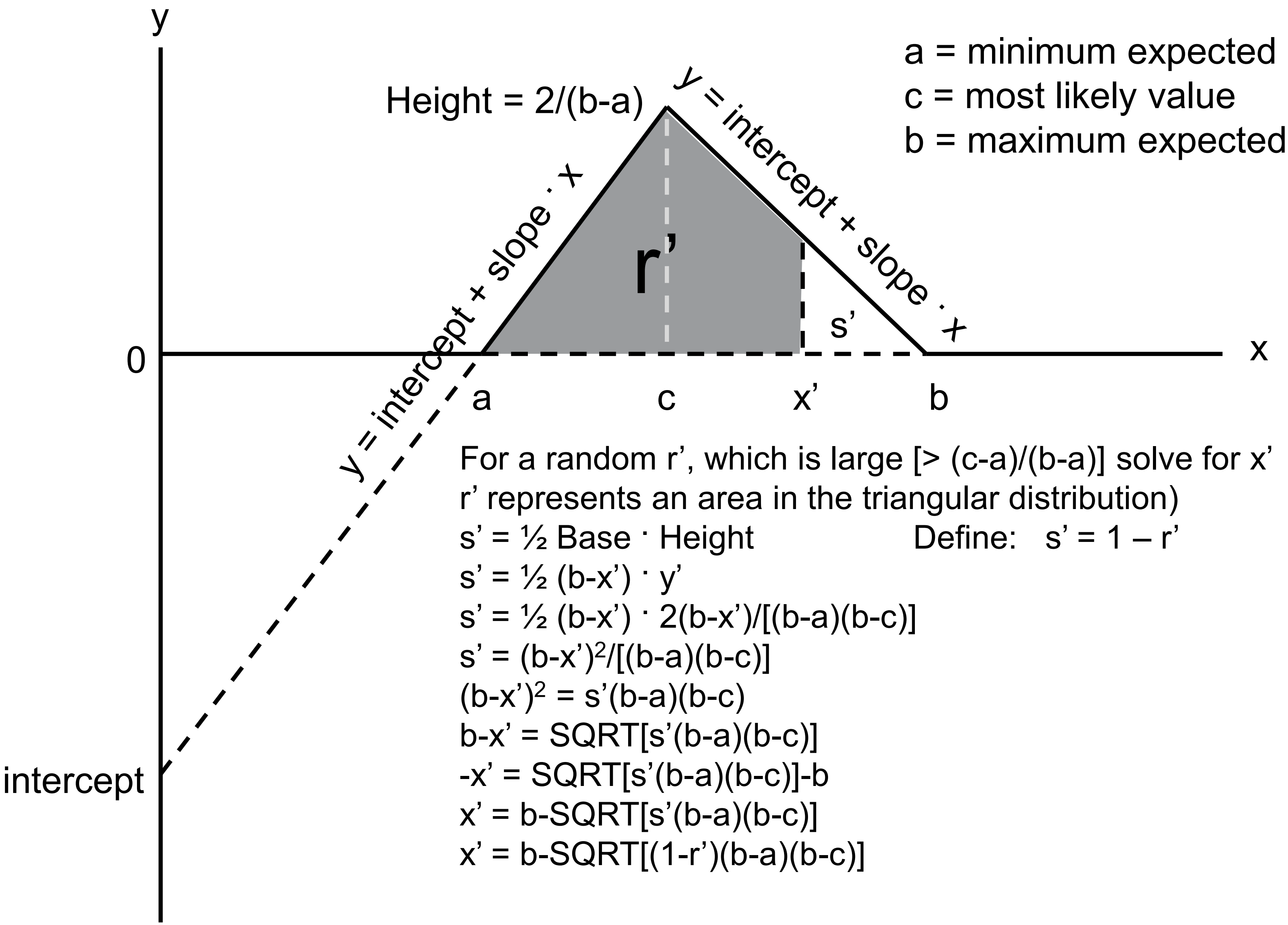
0 = intercept - 2/[(b-a)(b-c)] · b

Intercept = 2b/[(b-a)(b-c)]

y = 2b/[(b-a)(c-a)] - 2x/[(b-a)(b-c)]

y = 2(b-x)/[(b-a)(b-c)]

Triangular distribution simple trigonometry



a = minimum expected value
c = most likely value
b = maximum expected value

For a random r', which is large [> (c-a)/(b-a)] solve for x'
r' represents an area in the triangular distribution)

s' = ½ Base · Height   Define: s' = 1 – r'

s' = ½ (b-x') · y'

s' = ½ (b-x') · 2(b-x')/[(b-a)(b-c)]

s' = $(b-x')^2$/[(b-a)(b-c)]

$(b-x')^2$ = s'(b-a)(b-c)

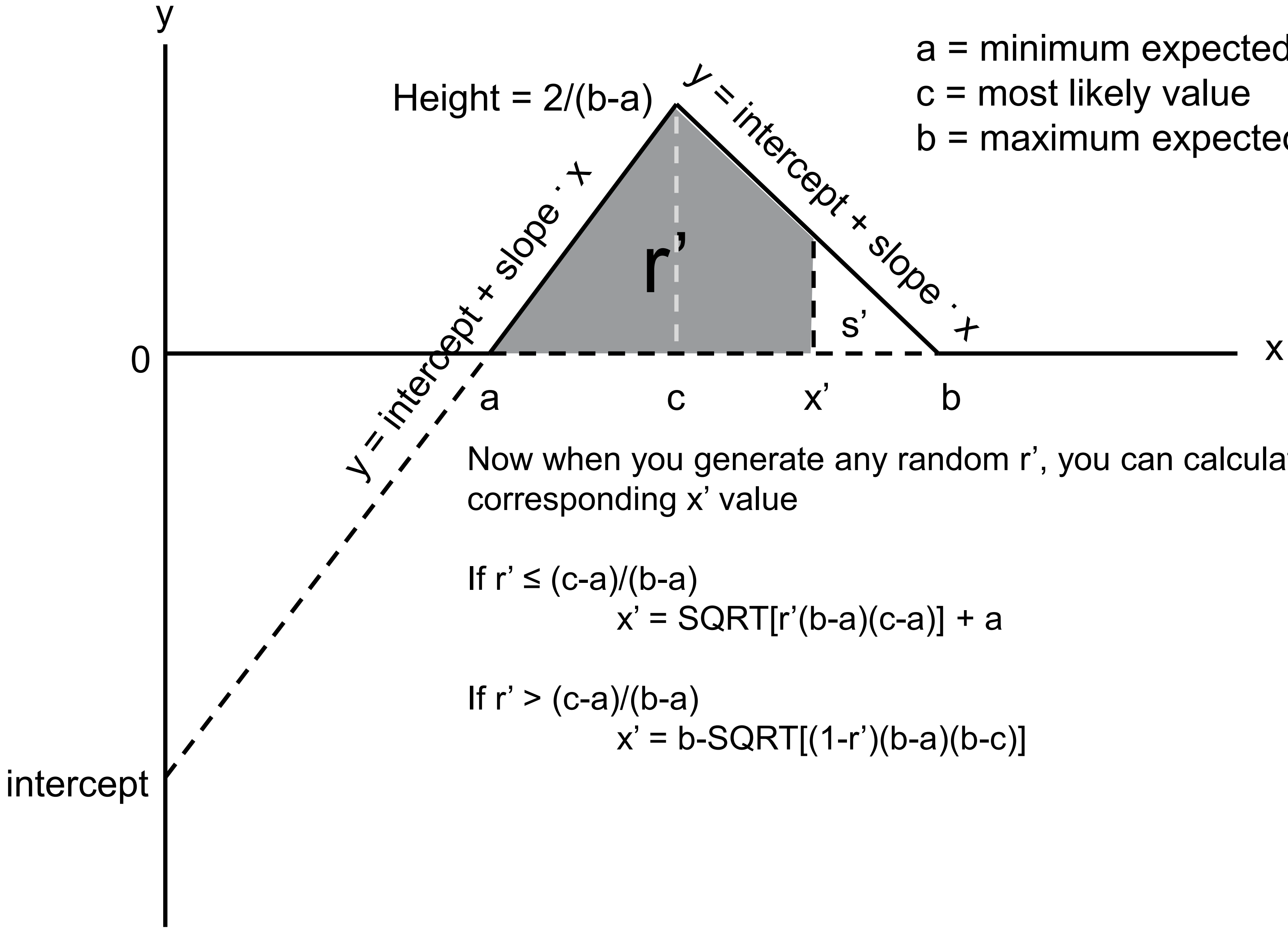
b-x' = SQRT[s'(b-a)(b-c)]

-x' = SQRT[s'(b-a)(b-c)]-b

x' = b-SQRT[s'(b-a)(b-c)]

x' = b-SQRT[(1-r')(b-a)(b-c)]

Triangular distribution simple trigonometry



a = minimum expected value
c = most likely value
b = maximum expected value

Now when you generate any random r’, you can calculate the corresponding x’ value

If r’ ≤ (c-a)/(b-a)

$$x' = \mathrm{SQRT}[r'(b-a)(c-a)] + a$$

If r’ > (c-a)/(b-a)

$$x' = b-\mathrm{SQRT}[(1-r')(b-a)(b-c)]$$

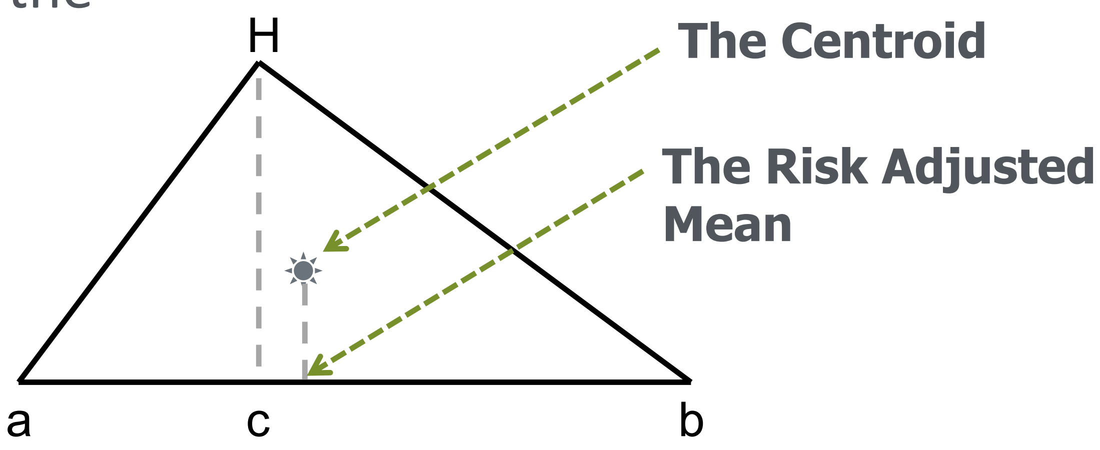
What is the centroid of triangle, if a and b are absolute min and max ?

- A = (a , 0) — a is the absolute minimum
- B = (b , 0) — b is the absolute maximum
- H = (c , 2/(b-a)) — H is the height at the mode c (most likely)
- The Centroid is the center of balance

$$\left[\frac{X_1 + X_2 + X_3}{3} , \frac{Y_1 + Y_2 + Y_3}{3} \right]$$

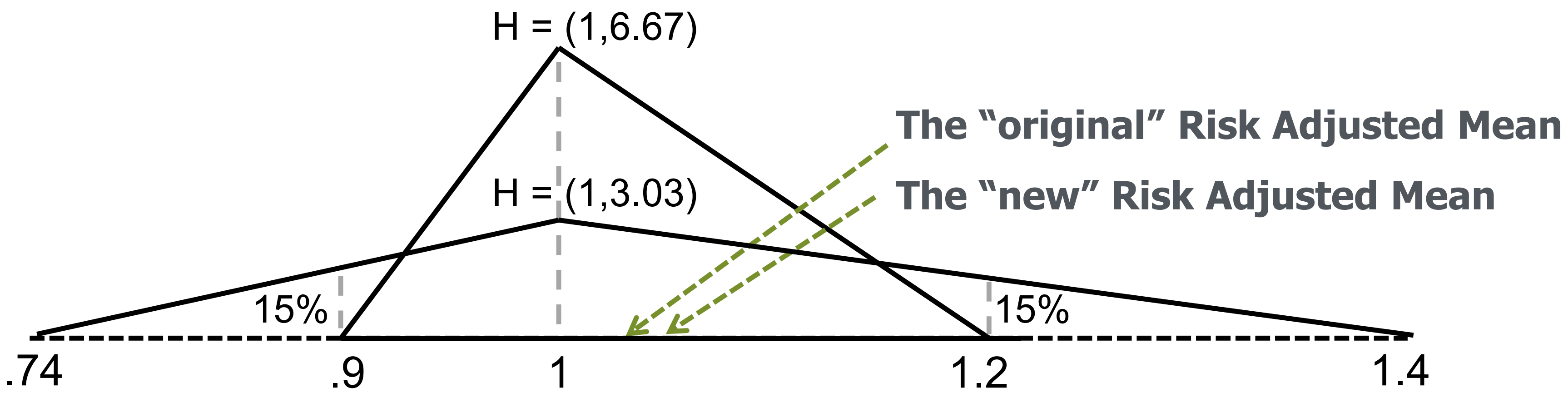
$$= \left[\frac{(a+b+c)}{3} , \frac{2}{3\,(b-a)} \right]$$

- The X-component is the risk adjusted mean



What if a and b are not absolute, but are the 15% and 85% percentiles?

- Example: a = 0.9 = the 15th percentile
  b = 1.2 = the 85th percentile
  "Original" Risk Adjusted Mean = 1.033
- The adjusted distribution is:
  A = (0.74 , 0)
  B = (1.4 , 0)
  H = (1.0 , 3.03) height = 2/0.66
  "New" Risk Adjusted Mean = 1.047
- In many defense acquisition situations, it is customary to apply the 15th and 85th percentiles to the min and max estimates of subject matter experts



Using triangular distributions in Monte Carlo simulations

- Define the triangular parameters (min, most likely, max) for each line (or each WBS or group) of your estimate.
  - Min and Max are typically defined as percentages (of the most likely value)
- Do a Monte Carlo Simulation by running thousands of iterations (applying random r' values) to the defined distribution for each line.
- According to central limit theorem, the resulting distribution of total combined cost estimates will approximate a normal distribution.

Example in Excel

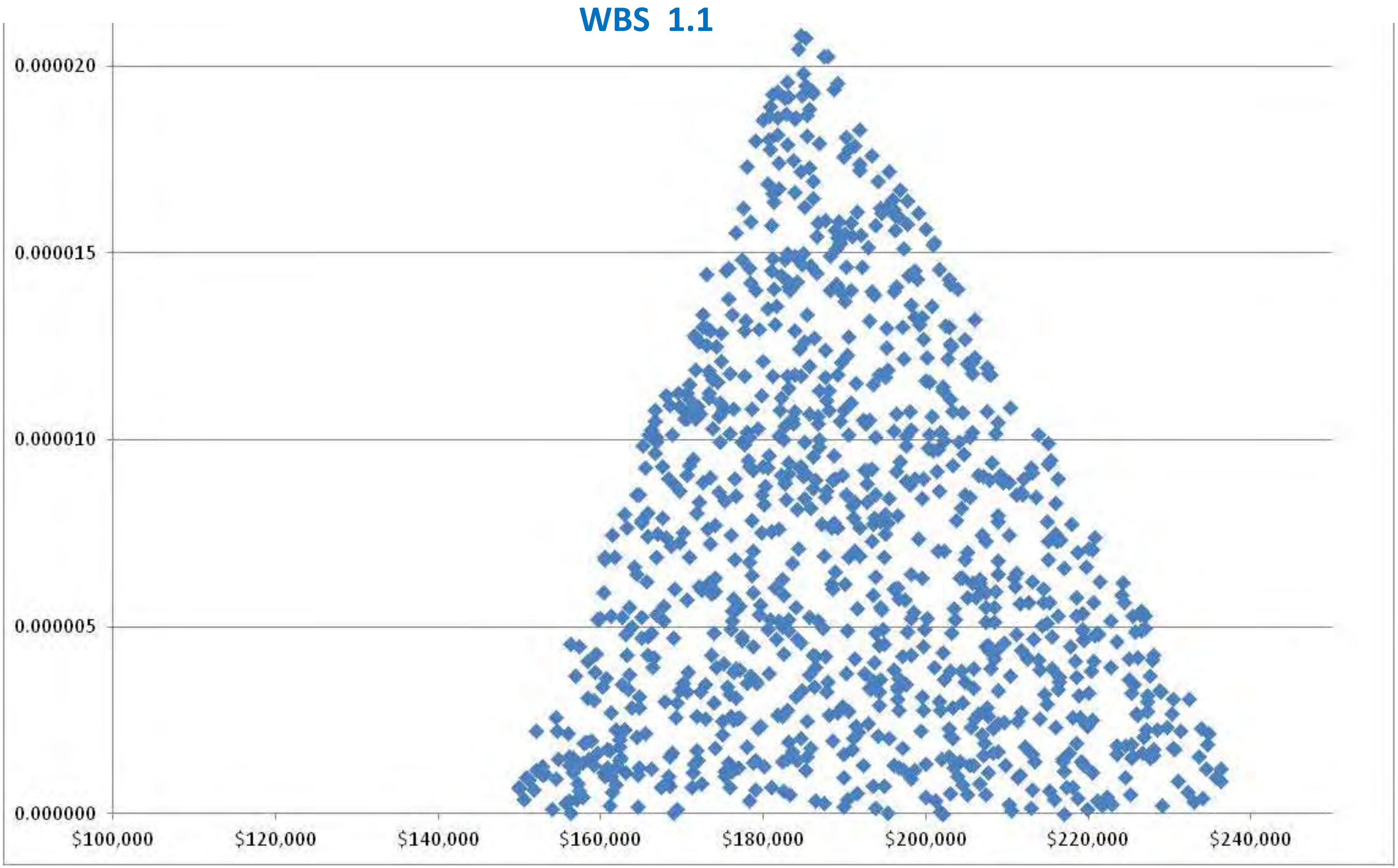
	Monte Carlo Simulation, Triangular Distributions								
		Enter estimates in blue cells only.				H	I	J	L
		c				a	c	b	
		Point Est	Triangular Risk Range			Triangular Cost Range			
	WBS	Most L	Min	ML	Max	Min	Most L	Max	
	0	$907,500				$759,250	$907,500	$1,174,200	
	1	$280,000				$231,400	$280,000	$358,700	
9	1.1	$185,000	80%	100%	130%	$148,000	$185,000	$240,500	RAND()
10	1.2	$74,000	90%	100%	120%	$66,600	$74,000	$88,800	RAND()
11	1.3	$21,000	80%	100%	140%	$16,800	$21,000	$29,400	RAND()
	2	$240,000				$190,900	$240,000	$327,200	
	2.1	$65,000	90%	100%	130%	$58,500	$65,000	$84,500	
	2.2	$99,000	80%	100%	130%	$79,200	$99,000	$128,700	
	2.3	$76,000	70%	100%	150%	$53,200	$76,000	$114,000	
	3	$387,500				$336,950	$387,500	$488,300	
	3.1	$1,500	80%	100%	120%	$1,200	$1,500	$1,800	
	3.2	$153,000	90%	100%	120%	$137,700	$153,000	$183,600	
	3.3	$233,000	85%	100%	130%	$198,050	$233,000	$302,900	

```
=IF(L9<=($I$9-$H$9)/($J$9-$H$9),SQRT(L9*($J$9-$H$9)*($I$9-$H$9))+$H$9,
                    $J$9-SQRT((1-L9)*($J$9-$H$9)*($J$9-$I$9)))
```

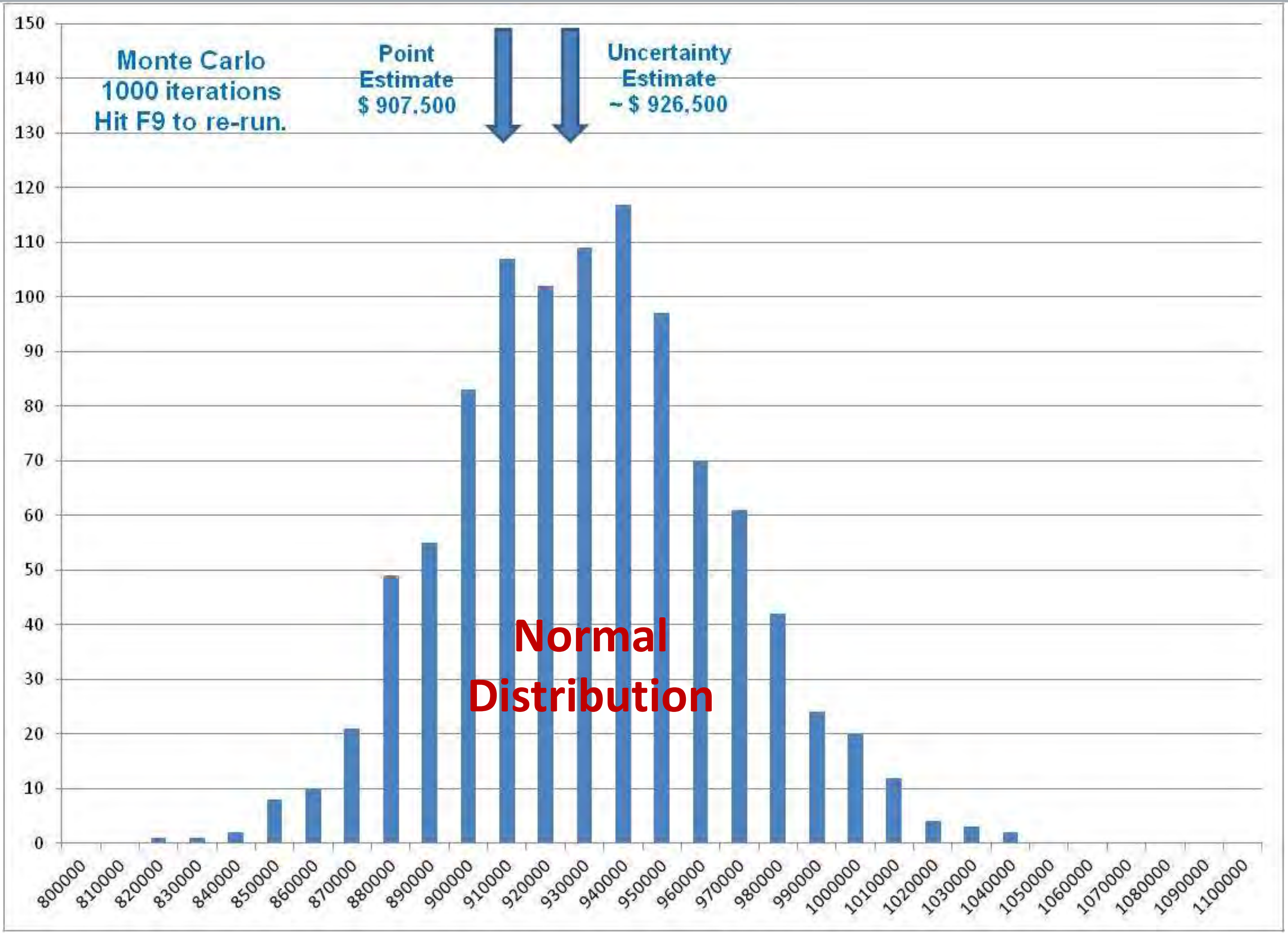
where:

- column L contains a random number, RAND()
- column H contains the minimum value of the triangular distribution, a
- column I contains the most likely value of the triangular distribution, c
- column J contains the maximum value of the triangular distribution, b

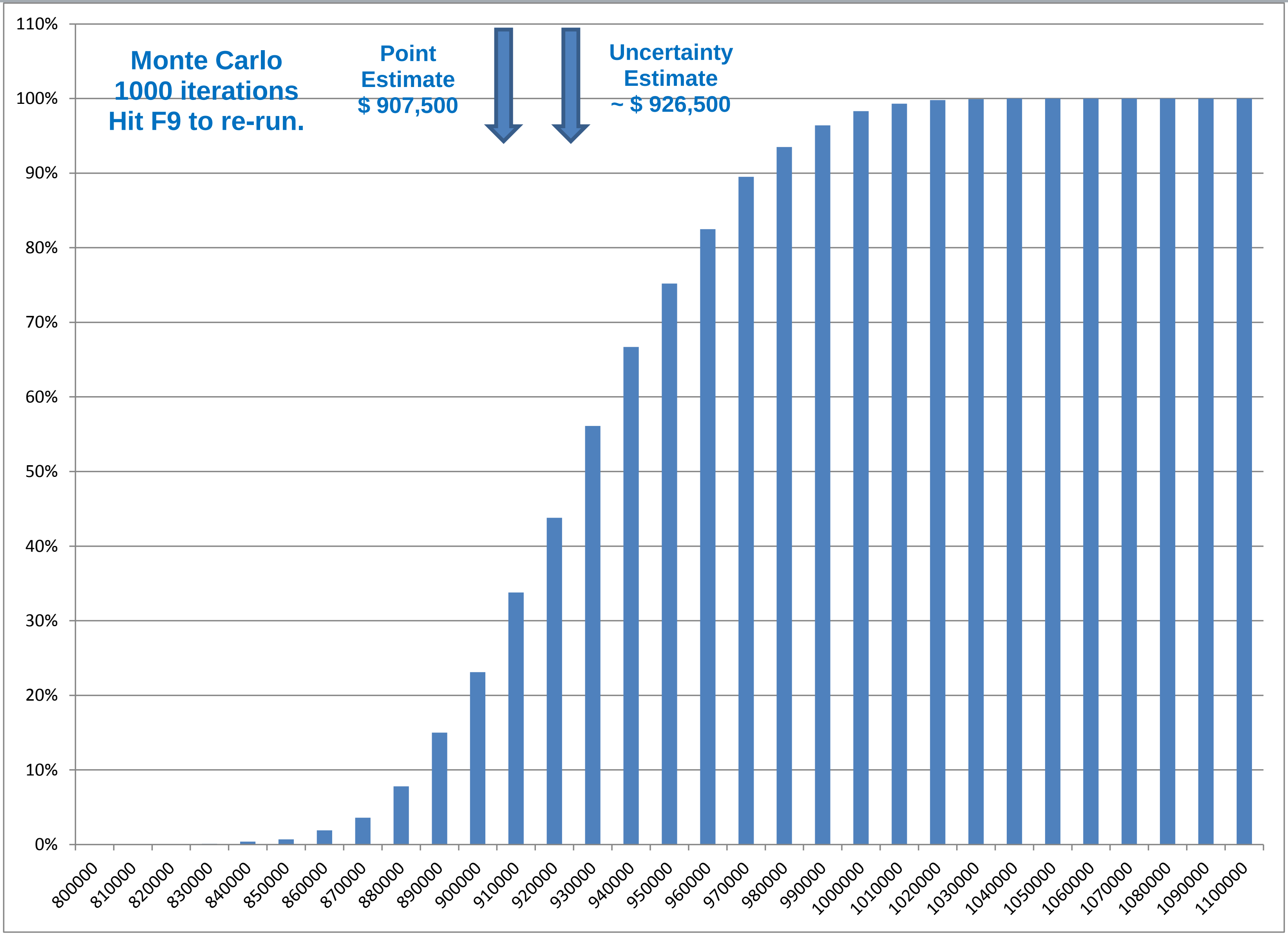
Example: Triangular distribution of one of the cost elements



Example: normal distribution of total combined cost estimates

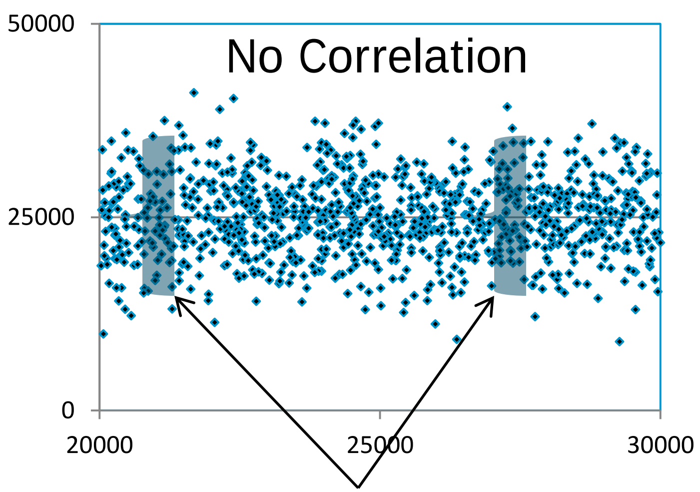


Example: S-curve

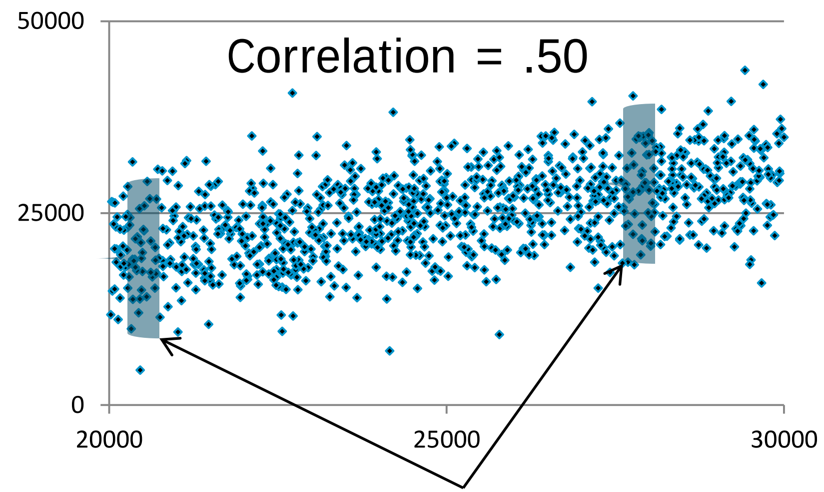


What is Correlation ?

- Correlation is a measure of the relationship between two variables
- There is correlation when a relation between variables exists which tend to vary in a way not expected on the basis of chance alone.
- Due to some underlying relationship between two variables, X and Y, knowing the outcome of one provides additional information as to the likely outcome of the other may be.



Knowing the result of X does not provide any additional information as to what the likely outcome of Y may be.



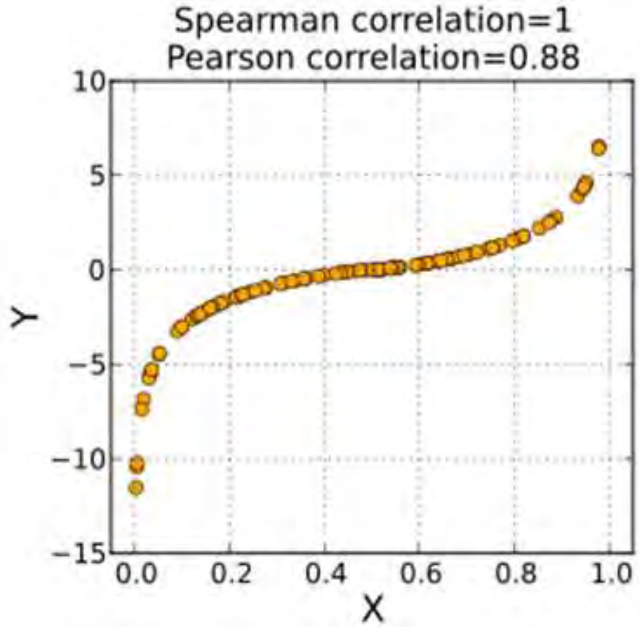
Knowing the result of X provides additional information as to the likely outcome of Y.

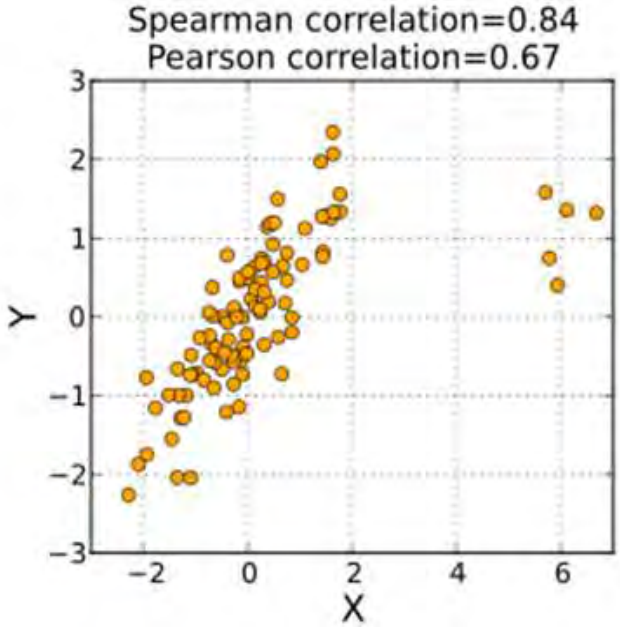
Pearson vs. Spearman's Rank Correlation

- There are various ways to measure correlation
- Pearson Correlation is the most widely used
- It is a measurement of the **linear** relationship between two variables
- Crystal Ball uses Spearman's Rank Correlation
- Spearman's Rank Correlation assumes a **monotonic** relationship rather than a linear relationship

Pearson vs. Spearman's Rank Correlation

- The Image on the left is perfectly monotone, meaning that each successive point is greater than the previous point, but it is not linear. This shows that the Spearman correlation will show perfect correlation while the Pearson correlation will not
- The second image shows another example of the differences between Spearman and Pearson correlation





Calculating Spearman's Rank Correlation with Excel

A	B	C	D
Data 1	Data 2	Rank 1	Rank 2
5	85	10	1
24	41	6	3
13	25	9	6
26	16	5	10
5	13	10	11
1	24	13	7
2	1	12	14
95	8	1	13
46	9	3	12
35	41	4	3
1	32	13	5
20	19	7	9
14	23	8	8
74	76	2	2

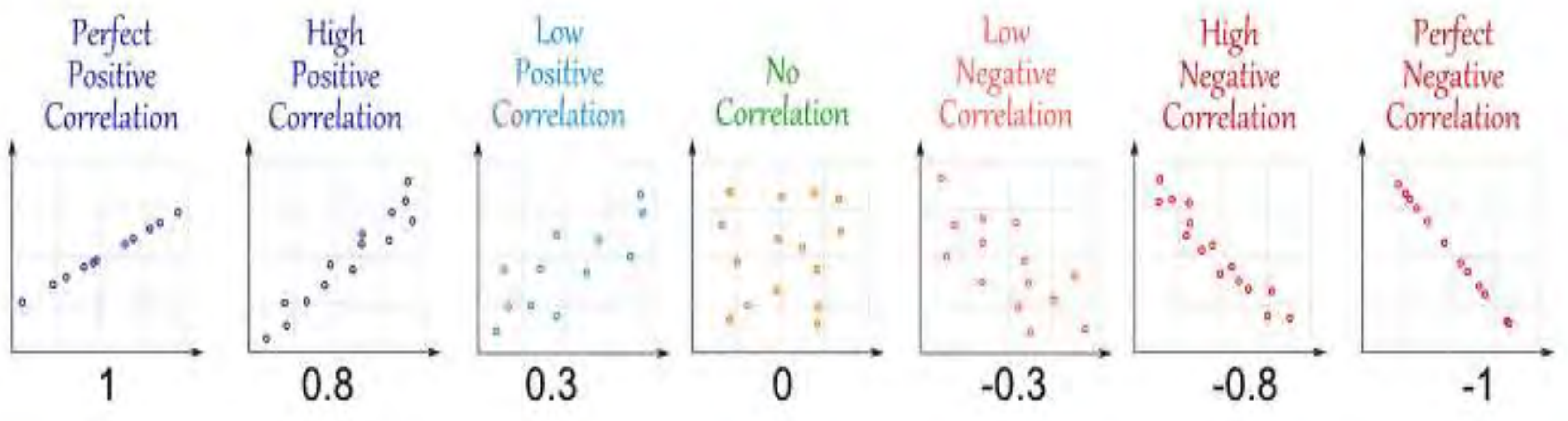
In Excel:

1. Start with two data sets (Columns A and B)
2. Rank each data set using the Rank function in excel. For column, C the formula would be Rank(A2,A2:A15)
3. Using the Correlation function in Excel, you can find the correlation of your data set using Columns C and D. Use the formula Correl(C2:C15,D2:D15)

Correlation
-0.032184496

Evaluating Correlation Results

- In the Excel example, the Correlation was approximately -0.03
- This shows that the two data sets are slightly negatively correlated. This correlation is so small that it is barely recognizable.
- Below shows in photos linear relationships for correlation
- The -0.03 would look as if there were no correlation



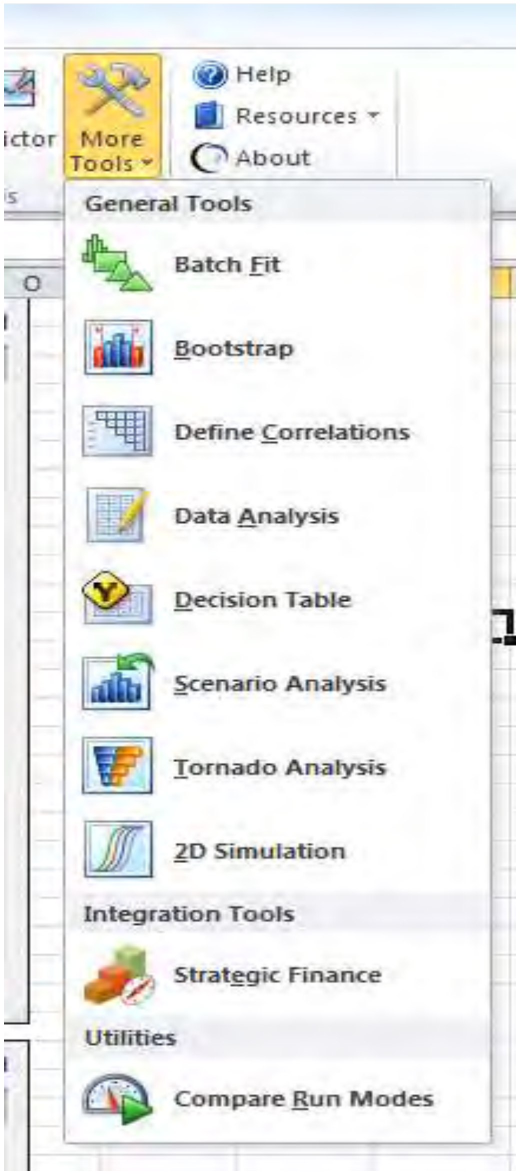
Defining Correlation for a PLCCE model

- When applying correlation to assumption variables in Crystal Ball, there are certain guidelines to consider
- First, only correlate variables that are similar.
- For instance, if you put a risk assumption on separate labor rates, those rates can be correlated. But, if you put risk on labor rates and hardware unit cost, these are probably not correlated.
- The below table are default correlation factors from Page 46 of the Joint Cost Schedule Risk and Uncertainty Handbook

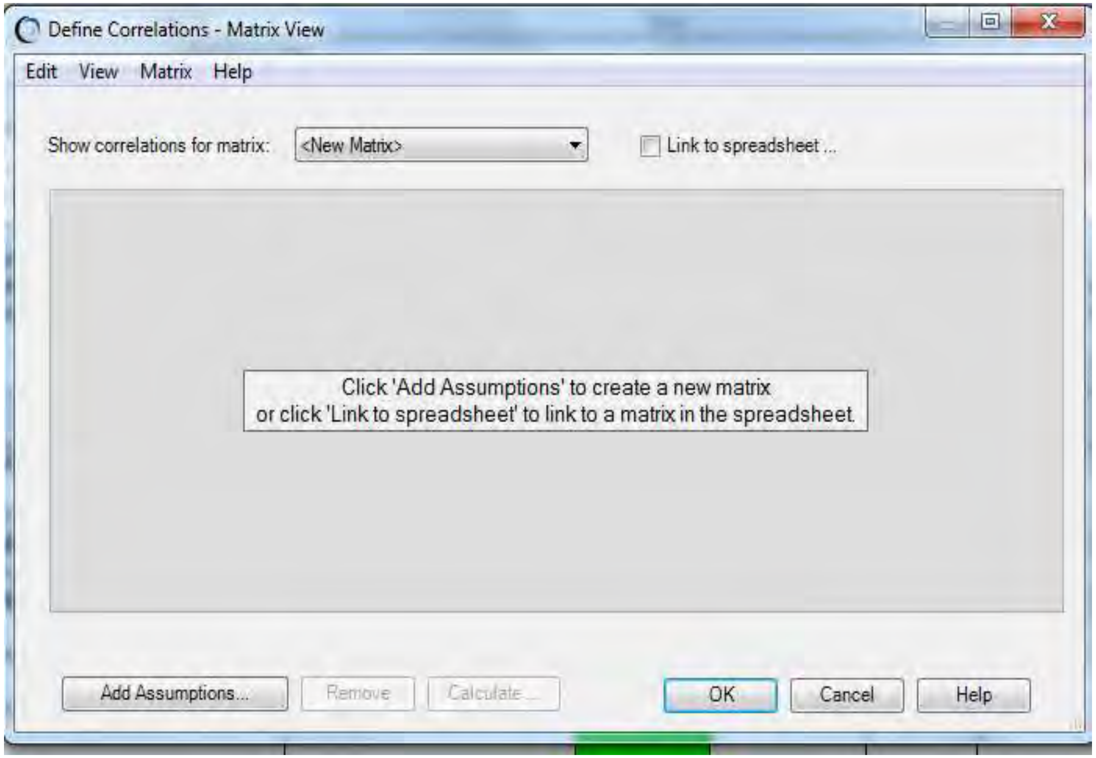
Table 3-1 Default Correlation Factors

Strength	Positive	Negative
None	0.0	0.0
Weak	0.3	-0.3
Medium	0.5	-0.5
Strong	0.9	-0.9
Perfect	1.0	-1.0

Defining Correlation in Crystal Ball



1. More Tools > Define Correlations
2. Add your assumptions (add all assumptions for complete matrix). You can build the correlation matrix in Excel and link it to the Crystal Ball simulation.
3. Add correlation to your assumptions
   - Weak Correlation = 0.3 (use as default)
   - Medium Correlation = 0.5
   - Strong Correlation = 0.9



Example Correlation Matrix

- There are many possible combinations for Crystal Ball assumptions and correlation

  This is only an example of how a correlation matrix may look

- All PLCCEs are unique

	Labor Rate 1	Labor Rate 2	HW Unit Cost 1	HW Unit Cost 2	HW Unit Cost 3	SW License Cost	Reuse Code Growth	New Code Growth
Labor Rate 1	1	0.5						
Labor Rate 2		1						
HW Unit Cost 1			1	0.9	0.9			
HW Unit Cost 2				1	0.9			
HW Unit Cost 3					1			
SW License Cost						1		
Reuse Code Growth							1	0.3
New Code Growth								1

Questions ?

Thank you

For more information, contact

Jennifer Lampe
jennifer.lampe@engilitycorp.com
724-875-9396

Jeffrey Platten
jeffrey.platten@engilitycorp.com
310-892-9150