Chapter 3

Spreadsheet Simulation
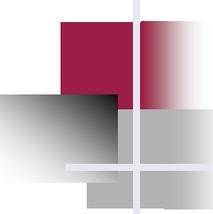
Objectives

- To be able to generate random numbers within a spreadsheet environment.

- To be able to generate random variables with various distributional models within a spreadsheet.

- To be able to perform simple Monte-Carlo simulation in a spreadsheet environment.

- To be able to compute statistical quantities for a Monte-Carlo simulation.

- To be able to determine the sample size for a Monte-Carlo simulation.
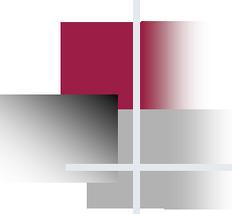
Overview

- Simulation in a spreadsheet environment
- Examples
  - Area estimation via simple Monte Carlo
  - News vendor inventory problem
  - Simulating a cash flow
- Statistical Concepts
  - Estimates and confidence intervals
  - Determining the sample size

Generating Randomness in a Spreadsheet
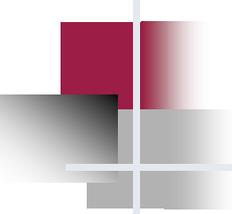
- Implement the ideas and concepts presented in Chapter 2
  - Inverse transform using already defined inverse CDF functions for various distributions
  - Convolution
  - Write VBA code
 - Buy special add-on software for your spreadsheet
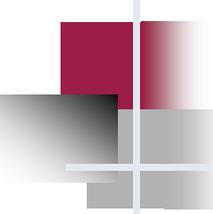
Two Basic Functions

- RAND()
  - Returns a U(0,1) pseudo-random number
  - "Live" function – every time any cell changes in the worksheet (i.e. when worksheet recalculates) a new U(0,1) is produced
- RANDBETWEEN(a,b)
  - Returns a discrete uniform over the range [a,b]
  - "Live" function

Setting Workbook to Manual Calculation

Change options related to formula calculation, performance, and error handling

Calculation options

Workbook Calculation ⓘ
- ⦿ Automatic
- ○ Automatic except for data tables
- ○ Manual
  - ☑ Recalculate workbook before saving

☐ Enable iterative calculation
Maximum Iterations: 100
Maximum Change: 0.001

Working with formulas

- ☐ R1C1 reference style ⓘ
- ☑ Formula AutoComplete ⓘ
- ☑ Use table names in formulas
- ☑ Use GetPivotData functions for PivotTable references

Error Checking

- ☑ Enable background error checking
- Indicate errors using this color: [color picker] [Reset Ignored Errors]

Error checking rules

- ☑ Cells containing formulas that result in an error ⓘ
- ☑ Inconsistent calculated column formula in tables ⓘ
- ☑ Cells containing years represented as 2 digits ⓘ
- ☑ Numbers formatted as text or preceded by an apostrophe ⓘ
- ☑ Formulas inconsistent with other formulas in the region ⓘ
- ☑ Formulas which omit cells in a
- ☑ Unlocked cells containing form
- ☐ Formulas referring to empty ce
- ☑ Data entered in a table is invali

Ribbon

ess Toolbar

ter

[OK]

- Use options to set workbook to perform manual calculation
- Saves time for large workbooks
- Allows you to control when simulation changes/updates
- Use F9 to cause calculation to occur

6

Inverse Transform
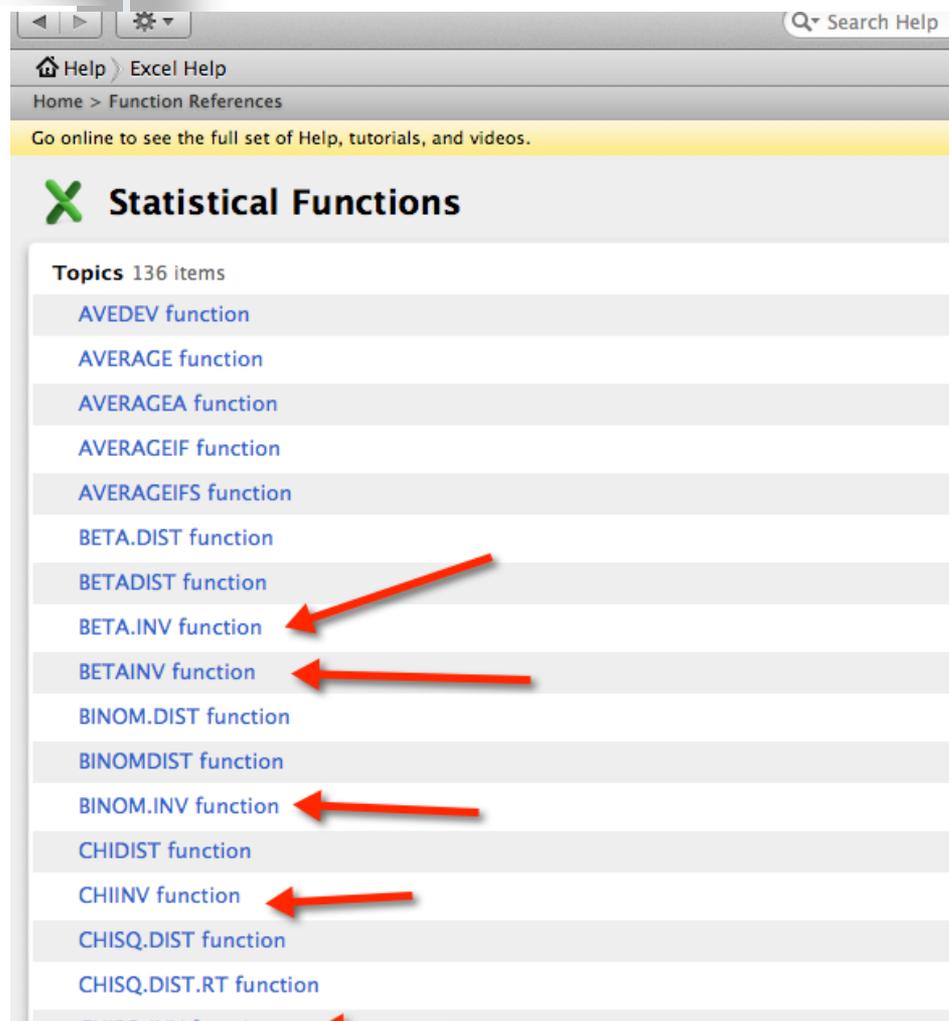
- Use distribution functions denoted with INV or .INV

- Allows you to provide U(0,1) to the function and will return a random variate with appropriate distribution

7

Examples:

- NORM.INV(probability, mean, standard deviation)
  - NORM.INV(RAND(), 3, 1) ~ N(3,1)
- BINOM.INV(#trials, probability of success, p)
  - BINOM.INV(6, 0.5, RAND()) ~ Binomial(n=6,p=0.5)
- U(a,b) ~ a + (b-a)*RAND()
- EXPO(mean) ~ -mean*LN(1-RAND())

SimulatingRandomnessInExcel.xlsm

	A	B	C	D	E
1					Formula
2	Uniform(0,1)			0.845009617	=RAND()
3		min	max		
4	Discrete Uniform(min,max)	3	10	8	=RANDBETWEEN(B5,C5)
5		min	max		
6	Uniform(min,max)	4	20	18.66724518	=B7+(C7-B7)*RAND()
7			p		
8	Bernoulli(p)		0.7	1	=IF(RAND()<=C9,1,0)
9			p		
10	Geometric(p)		0.3	2	=1+INT(LN(1-RAND())/LN(1-C11))
11			mean		
12	Exponential(mean)		5	4.359013352	=-C13*LN(1-RAND())
13		scale	shape		
14	Weibull(scale, shape)	5	3	3.975532185	=B15*POWER(-LN(1-RAND()),1/C15)
15		alpha1	alpha2		
16	Beta(alpha1, alpha2)	5	1.5	0.807041407	=BETAINV(RAND(),B17,C17)
17			DF		
18	ChiSquare		5	14.12091438	=CHIINV(RAND(),C19)
19		scale	shape		
20	Gamma(scale,shape)	2	10	19.06188666	=GAMMAINV(RAND(),C21,B21)
21		mean	std dev		
22		5	3		
23	Lognormal(mean, std dev)	1.455695563	0.554513	5.012218975	=LOGINV(RAND(),B24,C24)
24		mean	std dev		
25	Normal(mean, std dev)	10	2	10.50135199	=NORMINV(RAND(),B25,C25)

Discrete Distributions & VLOOKUP
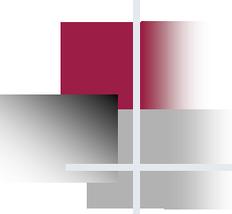
- The VLOOKUP cell function provides an easy way to implement the inverse transform method for discrete distributions.

- Organize data associated with distribution in a special way to take advantage of VLOOKUP functionality

Example

Let X be a discrete random variable, with $X \in \{1, 2, 3, 4\}$, with PMF, CDF, and inverse CDF:

x_i	1	2	3	4
$f(x_i)$	0.4	0.3	0.2	0.1
$F(x_i)$	0.4	0.7	0.9	1.0

$$F^{-1}(u) = \begin{cases} 1 & \text{if } 0.0 \leq u \leq 0.4 \\ 2 & \text{if } 0.4 < u \leq 0.7 \\ 3 & \text{if } 0.7 < u \leq 0.9 \\ 4 & \text{if } 0.9 < u \leq 1.0 \end{cases}$$

Via VLOOKUP

	A	B	C	D	E
A8			f_x =VLOOKUP(RAND(),B3:D6,3)		
	A	B	C	D	E
1					
2	PMF	LR	CDF	X	
3	0.4	0	0.4	1	
4	0.3	0.4	0.7	2	
5	0.2	0.7	0.9	3	
6	0.1	0.9	1	4	
7					
8		2	=VLOOKUP(RAND(),B3:D6,3)		

VLOOKUP(lookup_value, table_array, col_index_num)

- The column PMF is the probability mass function values.
- The column LR is the lower limit on the inverse transform range.
- The column X should contain the possible values for the random variable.

12
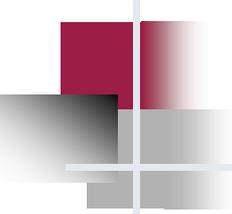
Using Data Tables

- Data tables can be used to repeatedly execute an equation

- If the equation represents a simulation, then data tables can be used to repeatedly execute a simulation to get a random sample of results

What is a Data Table

- A data table is a range of cells that shows how changing one or two variables in your formula will affect the results of those formulas.

- Data tables provide a shortcut for calculating multiple results in one operation and a way to view and compare the results of all the different variations together on your sheet.

- Search on Data Tables in Help to see detailed overview of their setup and use.

Example

	A	B	C	D	E	F	G
1							
2	PMF	LR	CDF	X		Data Table for Replications	
3	0.4	0	0.4	1		Replication	Observation
4	0.3	0.4	0.7	2			2
5	0.2	0.7	0.9	3		1	1
6	0.1	0.9	1	4		2	2
7						3	1
8		2 =VLOOKUP(RAND(),B3:D6,3)				4	1
9						5	3
10		Data Table Example				6	1
11						7	3
12		1) Make a column to count the replications				8	3
13		a) Column F here				9	3
14		2) Make a cell that you want repeated				10	2
15		a) Cell A8 will be repeated here				11	1
16		3) Tie the data table to the cell by setting the cell				12	2
17		above and to the right of the 1st replication equal				13	2
18		to the cell to be repeated				14	1
19		a) Cell G4 here is set to cell A8				15	3
20						16	1
21		4) Select the cells to form the data table				17	1
22		a) Select cells F4:G24				18	1
23		5) Select Data, What-If, Data Table				19	2
24		6) Enter any blank cell in for the Column input cell				20	1
25		a) Here cell F4 is used					
26		b) Leave the Row input cell blank					
27							

15

Data Table Steps

- Make a column to count the replications.
  - Column F is used in this example.
- Make a cell that you want repeated.
  - Cell A8 is used in this example.
- Tie the data table to the cell that needs repeating by setting the cell above and to the right of the 1st replication equal to the cell to be repeated.
  - In this example Cell G4 is set equal to cell A8.
- Select the cells that are to form the data table.
  - In this example, the cells in the range F4:G24 are the required cells.
- From the spreadsheet menu invoke the data table functionality. This is Data ⇒ What- IF ⇒ Data Table.
  - A confusing dialog box will appear asking for a row input cell and a column input cell. Since the table is column oriented, we only need the column input cell.
- Enter any blank cell in for the Column input cell.
  - Cell F4 was chosen as the input cell in this example.
  - Leave the Row input cell blank. Select the OK button.

16

Estimating the Area of a Function

- Suppose we have some function, $g(x)$, defined over the range $a \leq x \leq b$ and we want to evaluate the integral:

$$\theta = \int_{a}^{b} g(x)\, \mathrm{d}x$$

- It turns out that the problem can be translated into estimating the expected value of a well-chosen random variable.

- While a number of different choices for the random variable exist, we will pick one of the simplest for illustrative purposes. Define Y as follows with $X \sim U(a, b)$:

$$Y = (b - a)\, g(X)$$

Why does this work?

$$E[Y] = E[(b-a)g(X)] = (b-a)E[g(X)]$$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$E_X[g(X)] = \int_a^b g(x)f_X(x)\,dx = \int_a^b g(x)\frac{1}{b-a}\,dx$$

Substituting yields,

$$E[Y] = E[(b-a)g(X)] = (b-a)E[g(X)]$$

$$= (b-a)\int_a^b g(x)\frac{1}{b-a}\,dx$$

$$= \int_a^b g(x)\,dx = \theta$$

18

Why does it work?

- From basic statistics, we know that a good estimator for $E[Y]$ is the sample average of observations of Y.
- Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of observations of Y. Let X_i be the ith observation of X. Substituting each X_i yields the ith observation of Y,

$$Y_i = (b - a)\, g(X_i)$$

- Then, the sample average of $Y_1, Y_2, \ldots, Y_n$ is:

$$\bar{Y}(n) = \frac{1}{n}\sum_{i=1}^{n} Y_i = (b - a)\frac{1}{n}\sum_{i=1}^{n}(b - a)\, g(X_i)$$

$$= (b - a)\frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

where $X_i \sim U(a, b)$.
- By simply generating $X_i \sim U(a, b)$, plugging the X_i into the function, $g(x)$, taking the average over the values and multiplying by $(b - a)$, we can estimate the integral.

19

Example

Suppose that we want to estimate the area under $f(x) = x^{\frac{1}{2}}$ over the range from 1 to 4 as illustrated in the figure below. That is, we want to evaluate the integral:

$$\theta = \int_1^4 x^{\frac{1}{2}}\, dx = \frac{14}{3} = 4.6\overline{6}$$

Example

	A	B	C	D	E	F	G	H	I	J
1				True Area =	4.6666667	**Simulating the Area**				
2		Lower Limit a =	1							
3		Upper Limit b =	4			Want to estimate the area under the curve:				
4				Average =	4.7421464					
5						$f(x) = sqrt(x)$				
6	n	U(0,1)	X ~ U(a,b)	f(X)	Y =(b-a)f(X)					
7	1	0.997191473	3.9915744	1.9978925	5.9936775	over the range [1, 4]				
8	2	0.989081344	3.967244	1.9917942	5.9753825					
9	3	0.21772085	1.6531625	1.2857537	3.8572611	The true area is 14/3 = 4.6666 (see cell E1).				
10	4	0.661895436	2.9856863	1.7279139	5.1837416					
11	5	0.404142642	2.2124279	1.4874233	4.4622698	1) column B holds values from rand() to generate U(0,1)				
12	6	0.296216657	1.88865	1.3742816	4.1228449	2) column C generates X ~ U(a, b) using inverse transform				
13	7	0.613861009	2.841583	1.6856996	5.0570987					
14	8	0.37081048	2.1124314	1.4534206	4.3602618	$X = a + (b-a)*U$				
15	9	0.904623825	3.7138715	1.9271408	5.7814223	e.g. row 7: =C2 + (C3-C2)*B7				
16	10	0.99243608	3.9773082	1.994319	5.982957					
17	11	0.452834101	2.3585023	1.5357416	4.6072248	3) column D evaluates f(x) at the generated X				
18	12	0.174478216	1.5234346	1.2342749	3.7028248	e.g. row 7: =SQRT(C7)				
19	13	0.228178088	1.6845343	1.2978961	3.8936883					
20	14	0.035379788	1.1061394	1.0517316	3.1551948	4) column E evaluates Y = (b-a)*f(X)				
21	15	0.507718358	2.5231551	1.5884442	4.7653327					
22	16	0.983199878	3.9495996	1.98736	5.9620799	e.g. row 7: =(C3-C2)*D7				
23	17	0.689793097	3.0693793	1.7519644	5.2558932					
24	18	0.272101333	1.816304	1.3477032	4.0431097	5) cell E4 computes the average of the Y's				
25	19	0.799774928	3.3993248	1.8437258	5.5311774					
26	20	0.034046554	1.1021397	1.0498284	3.1494852	=AVERAGE(E7:E26)				
27						Notice how close cell E4 is to cell E1				

News Vendor Problem

Sly's convenience store sells BBQ wings for 25 cents a piece. They cost 15 cents a piece to make. The BBQ wings that are not sold on a given day are purchased by a local food pantry for 2 cents each. Assuming that Sly decides to make 30 wings a day, what is the expected revenue for the wings provided that the demand distribution is as show in the following table:

d_i	5	10	40	45	50	55	60
$f(d_i)$	0.1	0.2	0.3	0.2	0.1	0.05	0.05
$F(d_i)$	0.1	0.3	0.6	0.8	0.9	0.95	1.0

Setup a simulation to estimate the expected profit for various order quantities.
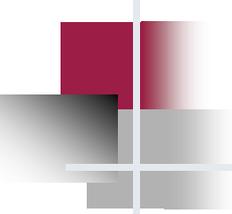
Solution Method

- Understand the "logic" of the problem
- Develop representation of randomness
- Develop formulas/logic that implement relationships between inputs/outputs and randomness
- Implement in spreadsheet
- Execute model

Solution

- Let D be a random variable representing the demand for the period.
- Let $F(d) = \mathrm{P}\{D \le d\}$ be the cumulative distribution function for the demand.
- Define $G(q, D)$ as the profit at the end of the period when q units are ordered at the start of the period with D units of demand.
  - The parameter q is the decision variable associated with this model. Depending on the value of q chosen by Sly, a profit or loss will occur.
- Let c be the purchase cost. This is the cost that Sly pays the supplier for the item.
- Let s be the selling price. This is the amount that Sly charges customers.
- Let u be the salvage value (i.e. the amount per unit that can be received for the item after the planning period).

24

What is the profit?

- Profit = selling price × amount sold + salvage value × amount unsold − unit cost × amount ordered
- How much is sold? This depends on demand, which is random and the amount ordered.
  - If D is the demand for the day and we start with q items, then the most that can be sold is $\min(D, q)$. That is, if D is bigger than q, you can only sell q. If D is smaller than q, you can only sell D.
- How much is unsold? This depends on demand, which is random and the amount ordered.
  - If D is the demand and we start with q, then there are two cases: 1) demand is less than the amount ordered $(D < q)$ or 2) demand is greater than or equal to the amount ordered $(D \geq q)$. In case 1, we will have $q - D$ items left over. In case 2, we will have 0 left over. Thus, the amount left over at the end of the day is $\max(0, q - D)$.
- Final Profit Equation

$$g(q) = \mathrm{E}\left[G(D, q)\right] = s\,\mathrm{E}\left[\min(D, q)\right] + u\,\mathrm{E}\left[\max(0, q - D)\right] - cq$$

25

	A	B	C	D	E	F	G
1	Sly's Convenience Store Example						
2	Inputs			Demand Distribution			
3	cost per unit c =	0.15		PMF	LR	CDF	D
4	price per unit s =	0.25		0.1	0	0.1	5
5	salvage per unit u =	0.02		0.2	0.1	0.3	10
6				0.3	0.3	0.6	40
7	Decision Variable			0.2	0.6	0.8	45
8	Order Quantiy q =	30		0.1	0.8	0.9	50
9				0.05	0.9	0.95	55
10	Outputs			0.05	0.95	1	60
11	Daily Demand	5					
12	Amount Sold	5		u =	0		
13	Amount Left Over	25		D =	5		
14	Sales Revenue	1.25					
15	Salvage Revenue	0.5					
16	Ordering Cost	4.5					
17	Profit	-2.75					
18							
19	Simulation Summary						
20	Average =	1.528					
21							
22	Data Table						
23		-2.75					
24	1	3					
25	2	3					
26	3	-1.6					
27	4	-1.6					
118	95	3					
119	96	-1.6					
120	97	-1.6					
121	98	-2.75					
122	99	-1.6					
123	100	3					

26

	A	B	C	D	E	F	G
1	Sly's Convenience S						
2	Inputs			Demand			
3	cost per unit c =	0.15		PMF	LR	CDF	D
4	price per unit s =	0.25		0.1	0	=D4	5
5	salvage per unit u =	0.02		0.2	=F4	=F4+D5	10
6				0.3	=F5	=F5+D6	40
7	Decision Variable			0.2	=F6	=F6+D7	45
8	Order Quantiy q =	30		0.1	=F7	=F7+D8	50
9				0.05	=F8	=F8+D9	55
10	Outputs			0.05	=F9	=F9+D10	60
11	Daily Demand	=E13					
12	Amount Sold	=MIN(B11,B8)		u =	=RAND()		
13	Amount Left Over	=MAX(0,B8-B11)		D =	=VLOOKUP(E12,E4:G10,3)		
14	Sales Revenue	=B12*B4					
15	Salvage Revenue	=B13*B5					
16	Ordering Cost	=B8*B3					
17	Profit	=B14+B15-B16					
18							
19	Simulation Summary						
20	Average =	=AVERAGE(B24:B123)					
21							
22	Data Table						
23		=B17					
24	1	=TABLE(,A23)					
25	2	=TABLE(,A23)					
26	3	=TABLE(,A23)					
27	4	=TABLE(,A23)					

27

Simulate a Cash Flow

- A firm is trying to decide whether or not it should purchase a new scale to check a package filling line in the plant.

- The scale would allow for better control over the filling operation and result in less overfilling. It is known for certain that the scale costs $800 initially.

- The annual cost has been estimated to be normally distributed with a mean of $100 and a standard deviation of $10.

- The extra savings associated with better control of the filling process has been estimated to be normally distributed with a mean of $300 and a standard deviation of $50.

- The salvage value has been estimated to be uniformly distributed between $90 and $100.

Simulating a Cash Flow

- The useful life of the scale varies according to the amount of usage of the scale. The manufacturing has estimated that the useful life can vary between 4 to 7 years with the chances given below:

years	4	5	6	7
f(years)	0.3	0.4	0.1	0.2
F(years)	0.3	0.7	0.8	1.0

- The interest rate has been varying recently and the firm is unsure of the rate for performing the analysis. To be safe, they have decided that the interest rate should be modeled as a beta random variable over the range from 6 to 9 percent with alpha = 5.0 and beta = 1.5.

- Perform a simulation analysis in order to assess the expected present value of the decision and the chance that the decision has a negative return.
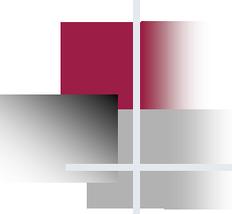
29

Solution Method

- Understand the "logic" of the problem
- Develop representation of randomness
- Develop formulas/logic that implement relationships between inputs/outputs and randomness
- Implement in spreadsheet
- Execute model

Solution

- The spreadsheet function PV will compute the present value of a future value that occurs at the end of a period as follows:

$$P = F\left[\frac{1}{(1+i)^n}\right] = F(P/F, i, n) = PV(i, n, \cdot, -F)$$

- Notice that PV has optional parameters that determine its use.
- The function PV will compute the present value of a series of future values that are assumed to occur at the end of the periods as follows:

$$P = A\left[\frac{(1+i)^n - 1}{i(1+i)^n}\right] = A(P/A, i, n) = PV(i, n, -A)$$
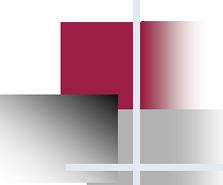
Solution

- Let $S \sim U(90, 100)$ be a random variable that represents the salvage value
- Let $C \sim N(\mu = 100, \sigma = 10)$ be a random variable that represents the annual cost
- Let $Y \sim N(\mu = 300, \sigma = 50)$ be a random variable that represents the annual savings
- Let N be a random variable that represents the useful life
- Let I be a random variable that represents the interest rate
- Present value of cost: $P_1 = C(P/A, I, N)$
- Present value of savings: $P_2 = Y(P/A, I, N)$
- Present value of salvage value: $P_3 = S(P/F, I, N)$
- Net Present Value: $P = -800 - P_1 + P_2 + P_3$

Implementation

	A	B	C	D	E
1	**Inputs**				
2	**Initial Cost**	-$800.00			
3	**Annual Cost**				
4	mean	-$100.00			
5	std dev	$10.00			
6	**Annual Savings**				
7	mean	$300.00			
8	std dev	$50.00			
9	**Salvage Value**				
10	min	$90.00			
11	max	$100.00			
12	**Useful Life**	**PMF**	**LR**	**CDF**	**Life**
13		0.30	0	0.3	
14		0.40	0.3	0.7	
15		0.10	0.7	0.8	
16		0.20	0.8	1	
17	**Interest Rate**				
18	alpha	5.00			
19	beta	1.50			
20	min	6.00%			
21	max	9.00%			

- Organize spreadsheet into inputs and outputs

33

	A	B
22		
23	**Outputs**	
24	Initial Cost	=B2
25	Annual Cost	=NORMINV(RAND(),B4,B5)
26	Annual Savings	=NORMINV(RAND(),B7,B8)
27	Salvage Value	=B10+(B11-B10)*RAND()
28	Useful Life	=VLOOKUP(RAND(),C13:E16,3)
29	Interest Rate	=BETAINV(RAND(),B18,B19,B20,B21)
30	PV of Salvage	=PV(B29,B28,,-B27)
31	PV of Annual Cost	=PV(B29,B28,-B25)
32	PV of Annual Savings	=PV(B29,B28,-B26)
33	Net Present Value	=B24+B31+B32+B30

|	A	B	C
22			
23	**Outputs**		
24	Initial Cost	-$800.00	
25	Annual Cost	-102.069	
26	Annual Savings	183.9268	
27	Salvage Value	$98.14	
28	Useful Life	7	
29	Interest Rate	0.072705	
30	PV of Salvage	$60.05	
31	PV of Annual Cost	($544.93)	
32	PV of Annual Savings	$981.96	
33	Net Present Value	-$302.93	

- Separate output area with individual components for parts of cost.
- Allows analysis of individual components.

34
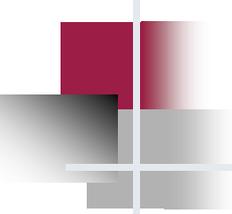
Statistical Concepts

- Simulation has randomness
- Randomness implies that there is sampling error in the output from the simulation
- Treat simulation like experimentation
- Need to quantify uncertainty of results
- Need to determine how much to sample

Estimator versus Estimate

- Represent the random variables in the sample as $X_1, X_2, \ldots, X_n$.
- Let x_i represent the ith observations in a realized sample, $x_1, x_2, \ldots, x_n$, of size n.
- An estimator is a function of a sample that computes a statistic. It is a random variable.
    - The sample average can be used as an estimator of the the expected value

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- An estimate is the number computed from an estimator using the actual data
    - The point estimate for the sample average is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

36

Confidence Intervals

- A confidence interval expresses a degree of certainty associated with a point estimate.
- A specific confidence interval does not imply that the parameter of interest, μ, is inside the interval. The true parameter is either in the interval or it is not within the interval.
- You should think about the confidence level $1 - \alpha$ as an assurance about the procedure used to compute the interval.
  - A confidence interval procedure ensures that if a large number of confidence intervals are computed each based on n samples, then the proportion of the confidence intervals that actually contain the true value of μ should be close to $1 - \alpha$.
  - The value α represents risk that the confidence interval procedure will produce a specific interval that does not contain the true parameter value.

Confidence Interval for the Expected Value

- Under the assumption that the sample size is large enough such that the distribution of $\bar{X}$ is normally distributed then you can form an approximate confidence interval on the point estimate, $\bar{x}$.

- Assuming that the sample variance:

$$s^2(n) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

is a good estimator for $\text{Var}[X] = \sigma^2$, then a $(1-\alpha)100\%$ confidence interval estimate for μ is:

$$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

- $t_{\alpha/2, n-1}$ is the upper $100(1 - \alpha/2)$ percentage point of the Student t-distribution with $n - 1$ degrees of freedom.

- In order to compute $t_{\alpha/2, n-1}$ use the following formula in Excel:

$$t_{\alpha/2, n-1} = T.INV(1 - \alpha/2, n - 1)$$

38

- The quantity, $h = t_{\alpha/2, n-1} \dfrac{s}{\sqrt{n}}$ is called the half-width of the confidence interval.

- You can place a bound, E, on the half-width by picking a sample size that satisfies:

$$h = t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \leq E$$

- You must try different values of n until the condition is satisfied.

- The required sample size can be approximated using the normal distribution. Solving for n yields:

$$n \geq \left(\frac{t_{\alpha/2, n-1} s}{E} \right)^2$$

As n gets large, $t_{\alpha/2, n-1}$ converges to the upper upper $100(1 - \alpha/2)$ percentage point of the standard normal distribution $z_{\alpha/2}$. This yields the following approximation:

$$n \geq \left(\frac{z_{\alpha/2} s}{E} \right)^2$$

39

	A	B	C	D	E	F
1			Sample size determination			
2	alpha	0.05				
3	S	66.083	Run goal seek to find the values of n			
4	bound	20	such that the half-width is less than			
5	n	44.40159	the specified desired bound.			
6	alpha/2	0.025				
7	t-alpha/2	2.016692	1) Specify alpha			
8	half-width	20.00001	2) Specify S (standard deviation)			
9	difference	5.82E-06	3) Specify bound			

4) Specify initial n
5) Run Goal Seek
   a) Tools> Goal Seek
   b) Set cell b9 to Value: 0
   by changing cell b5

Goal Seek

Set cell: B9

To value: 0

By changing cell: B5

Cancel   OK

20

40

Example

Suppose we are required to estimate the output from a spreadsheet simulation so that we are 99% confidence that we are within ± 0.1 of the true population mean. After taking a pilot sample of size $n = 10$, we have estimated $s = 6$. Recommend a suitable sample size.

- We will use the normal approximation. For a 99% confidence interval, we have $\alpha = 0.01$ and $\alpha/2 = 0.005$. Thus, $z_{0.005} = 2.576$. This yields,

$$n \geq \left(\frac{z_{\alpha/2} s}{E} \right)^2 = \left(\frac{2.576 \times 6}{0.1} \right)^2 = 23888.8 = 23889$$

41

Cash Flow Example

- Suppose we are interested in estimating the expected present value of the cash flow problem.

- We desire to be 95% confident that our estimate of the true expected present value is within $\pm$ 10 dollars.

- In addition, we are interested in estimating the probability that the present value is negative. That is, we want to estimate the risk of losing money on the situation.

Estimating a Probability

- To estimate a probability define an indicator variable that has the value 1 if the condition associated with the probability is true and has the value 0 if it is false.
- Let X represent the unknown random present value. Then, we are interested in estimating $E[X]$ and $p = P\{X < 0\}$. To estimate p, define the following indicator variable:

$$Y_i = \begin{cases} 1 & X_i < 0 \\ 0 & X_i \geq 0 \end{cases}$$

- This definition of the indicator variable Y_i allows $\bar{Y}$ to be used to estimate p.

$$\hat{p} = \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{\#\{X_i < 0\}}{n}$$

43

	G	H	I	J	
1	Simulation Statistics	NPV	P(NPV < 0)		
2	Sample Average	85.16036	0.4		
3	Standard Error	40.71579	0.090972		
4	Median	61.3453	0		
5	Standard Deviation	223.0095	0.498273		
6	Sample Variance	49733.26	0.248276		
7	Range	891.265	1		
8	Minimum	-387.753	0		
9	Maximum	503.5121	1		
10	Sum	2554.811	12		
11	Count	30	30		
12	Alpha	0.05	0.05		
13	T-Value(Alpha/2,n-1)	2.04523	2.04523		
14	CI Half-Width	83.27313	0.186058		
15	LL-Confidence Interval	1.887226	0.213942		
16	UL-Confidence Interval	168.4335	0.586058		
17					
18	Simulation Data Table				
19		-$344.64			
20	1	423.6894	0	=IF(H20<0,1,0)	
21	2	-247.635	1		
22	3	-167.293	1		
23	4	83.08767	0		
24	5	439.5662	0		
25	6	245.4963	0		
26	7	163.4001	0		
27	8	-55.7688	1		
28	9	-97.0344	1		
29	10	-207.699	1		

	G	H	I
1	Simulation Statistics	NPV	P(NPV < 0)
2	Sample Average	=AVERAGE(H20:H49)	=AVERAGE(I20:I49)
3	Standard Error	=H5/SQRT(H11)	=I5/SQRT(I11)
4	Median	=MEDIAN(H20:H49)	=MEDIAN(I20:I49)
5	Standard Deviation	=STDEV(H20:H49)	=STDEV(I20:I49)
6	Sample Variance	=VAR(H20:H49)	=VAR(I20:I49)
7	Range	=H9-H8	=I9-I8
8	Minimum	=MIN(H20:H49)	=MIN(I20:I49)
9	Maximum	=MAX(H20:H49)	=MAX(I20:I49)
10	Sum	=SUM(H20:H49)	=SUM(I20:I49)
11	Count	=COUNT(H20:H49)	=COUNT(I20:I49)
12	Alpha	0.05	0.05
13	T-Value(Alpha/2,n-1)	=T.INV(1-H12/2,H11-1)	=T.INV(1-I12/2,I11-1)
14	CI Half-Width	=H13*H3	=I13*I3
15	LL-Confidence Interval	=H2-H14	=I2-I14
16	UL-Confidence Interval	=H2+H14	=I2+I14
17			
18	Simulation Data Table		
19		=B33	
20	1	=TABLE(,G19)	=IF(H20<0,1,0)
21	2	=TABLE(,G19)	=IF(H21<0,1,0)
22	3	=TABLE(,G19)	=IF(H22<0,1,0)
23	4	=TABLE(,G19)	=IF(H23<0,1,0)
24	5	=TABLE(,G19)	=IF(H24<0,1,0)
25	6	=TABLE(,G19)	=IF(H25<0,1,0)
26	7	=TABLE(,G19)	=IF(H26<0,1,0)
27	8	=TABLE(,G19)	=IF(H27<0,1,0)
28	9	=TABLE(,G19)	=IF(H28<0,1,0)
29	10	=TABLE(,G19)	=IF(H29<0,1,0)

45

- The results based on a pilot of $n = 30$ observations
  - average present value is 85.16
  - the standard deviation is quite large at 223
- To meet the specification of a half-width of 10 dollars will require a larger sample. Using the normal approximation, we get the following:

$$n \geq \left(\frac{z_{\alpha/2}s}{E}\right)^2$$

$$\geq \left(\frac{1.96 \times 223}{10}\right)^2 \approx 1910.3$$

	G	H	I
1	Simulation Statistics	NPV	P(NPV < 0)
2	Sample Average	69.12068	0.425955
3	Standard Error	5.546058	0.011315
4	Median	49.18056	0
5	Standard Deviation	242.4458	0.494616
6	Sample Variance	58779.99	0.244645
7	Range	1668.144	1
8	Minimum	-713.756	0
9	Maximum	954.3876	1
10	Sum	132089.6	814
11	Count	1911	1911
12	Alpha	0.05	1.05
13	T-Value(Alpha/2,n-1)	1.961207	-0.06272
14	CI Half-Width	10.87697	-0.00071
15	LL-Confidence Interval	58.24371	0.426665
16	UL-Confidence Interval	79.99765	0.425245

46
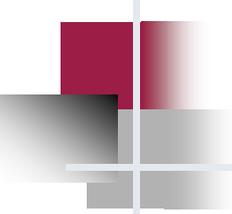
Summary

- Spreadsheets can be used to simulate randomness

- Built in functions facilitate random variate generation

- Carefully set up and organize your spreadsheet

- Randomness implies that we need to use statistics to analyze the results