Cost Estimating: It's a Risky Business

Ian Anderson
The Naval Center for Cost Analysis (NCCA)
16 September 2010, prepared for:
4th Annual Navy/Marine Corps Cost Analysis Symposium

Objective

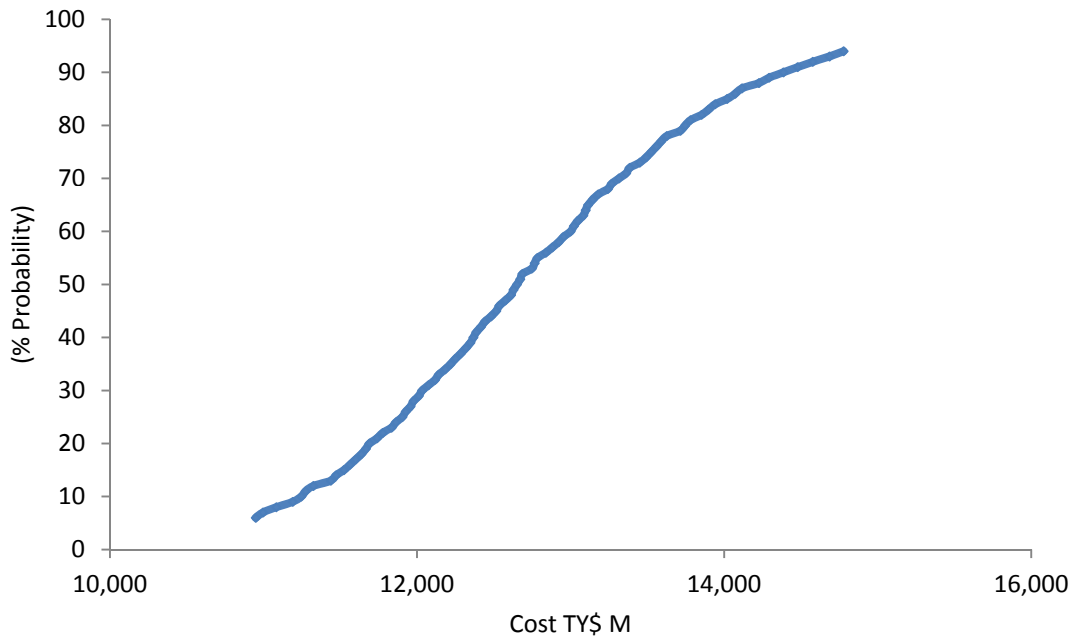
The purpose of this presentation is to describe the process of uncertainty and risk analysis from a mathematical perspective and explain how this relates the field of cost estimation. The presentation will conclude with a real world case study of the Consolidated Afloat Networks and Enterprise Services (CANES) System.

What's the difference between "risk" and "uncertainty"?

- Mathematically, risk is very easy to handle. We simply identify it at as a quantity.
  - Sometimes it's added into the estimate and included in the budget.
  - Sometimes it's simply called "risk" and it's book-kept separately from the budget.
  - Either way, little math is required.
- Uncertainty is much more complex. Most of this brief will focus on uncertainty.

The Cumulative Distribution Function (CDF)

- Regardless of whether we're talking about risk or uncertainty, I've observed that people typically associate these terms with the following illustration.



CDF Myths & Other Characteristics

- A CDF is only used to describe “uncertainty” considerations only. Risk considerations are binary.
- A CDF is NOT an S-Curve and shouldn’t be called one. Sometimes it resembles an S, sometimes it’s a straight line, it all depends on the probability distributions functions (PDFs) chosen to represent the random variable inputs. (I’ll explain what’s meant by PDF’s and random variables in a bit)

CDF Myths & Other Characteristics (…cont'd)

- A few other characteristics to keep in mind
  - A CDF is always increasing
  - Y-Axis shows the probability from zero to 1 (or 0% to 100%)
  - X-Axis ranges are important

CDFs invented by theories in probability and statistics

- Okay, so since it's a requirement that we produce these CDF's, let's talk about how they are generated.
- BUT, to do this we need to understand some basic theoretical concepts in the mathematical field of probability and statistics.
- Without a basic understanding of probability and statistics, these curves are literally useless.

First basic theory of probability

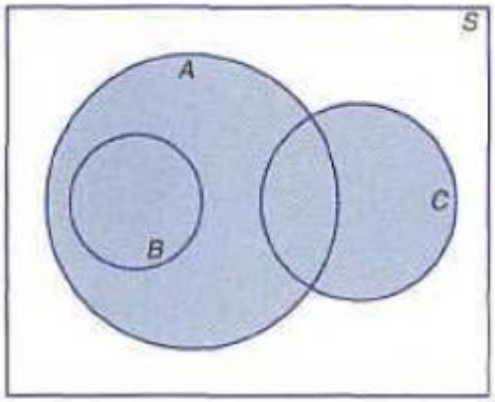
- The first basic underlying theory of probability is that of a repetitive operation or repetitive experiment, that is, the notion of being able to repeat an operation over and over again "under the same conditions." Some examples of repetitive operations are:
  1) Tossing a coin.
  2) Throwing a die.
  3) Selecting a "spade" from a shuffled deck of cards.
  4) Selecting 2 screws "at random" from a box of 100 screws.

Second basic theory of probability

- The second basic underlying theory of probability is the idea of a "sample space." The sample space is the collection of all possible outcomes (or elements) of a repetitive operation. For example it's:
  1) Tossing a coin: {H,T} Simultaneously tossing two coins: {Hh,Tt,Ht,hT}
  2) Throwing a die: {1,2,3,4,5,6}
  3) Selecting a "spade" from a deck of cards: The deck of cards {52 outcomes}
  4) 2 screws from box of 100 screws: {4950}

Third basic theory of probability

- The third basic concept is the idea of an "event." An event is any subset of the sample space.
- Each circle represents an event. A and C together represent an event.



Venn-Diagram

Third basic theory of probability (…cont'd)

- Previous examples applied the "events"
  1) Not defined.
  2) Not defined. If we were throwing a 1 the event would be {1}
  3) {2s, 3s, 4s, 5s, 6s, 7s, 8s, 9s, 10s, Js, Qs, Ks, As} or {13 outcomes}
  4) Not defined. But if we expect defects, then {Nn, Dd, Nd, nD}, where N means non defective and D means defective.

Fourth basic theory of probability

The fourth basic concept is the **probability of an event**. If we perform a repetitive operation over and over again under the same conditions, it's simply the fraction of trials in which a specific event occurs.

If you imagine a blind person throwing a huge supply of darts at the Venn Diagram and you count the number of darts that fall in B, divide that by the total number of darts, you've calculated the probability of an event B (of course it becomes more exact the more darts you throw).

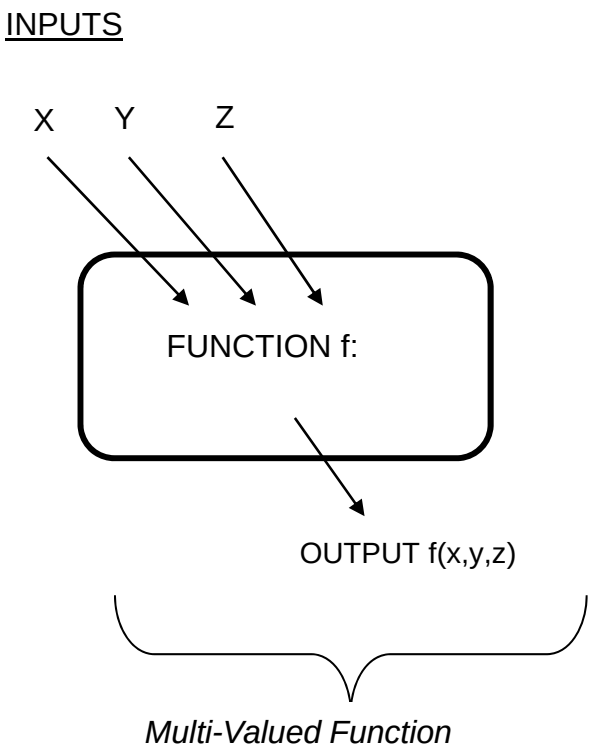
Fourth basic theory of probability (…cont'd)

- Previous examples applied the **probability of an event**
  1) Not defined.
  2) Not defined. If we were throwing a 1 the probability would be 1/6 or 0.16667 or 16.7%
  3) 13/52 or 1/4 or 25%
  4) Not defined. But if we expect defects, then you could calc:
     - Probability of 1 defective screw
     - Probability of both defective screws
     - To do this you need to know something about the "box of screws" or in statistical terms the "population."

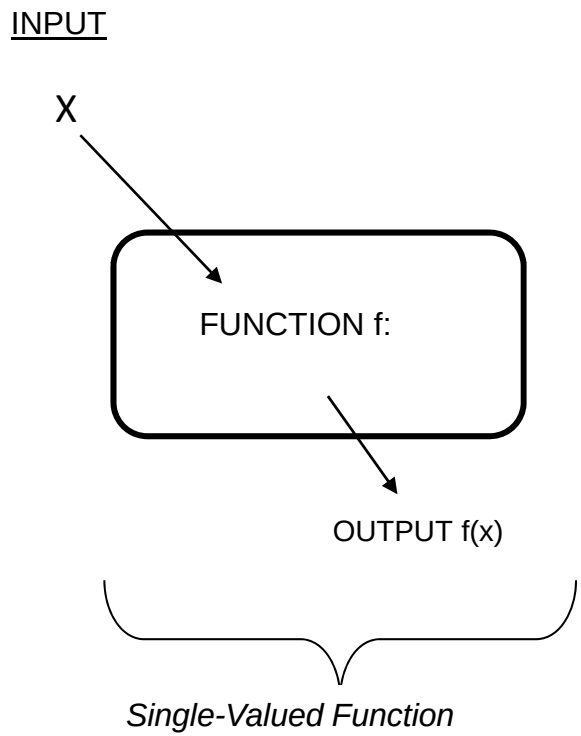
Fifth basic theory of probability

- The fifth basic concept the **random variable (RV).** A RV is a function that associates a number with each element in the sample space.
- What is meant by “function”? A function is the mathematical idea that one quantity completely determines another quantity. A function assigns a *unique* value to each input of a specified type.

Functions

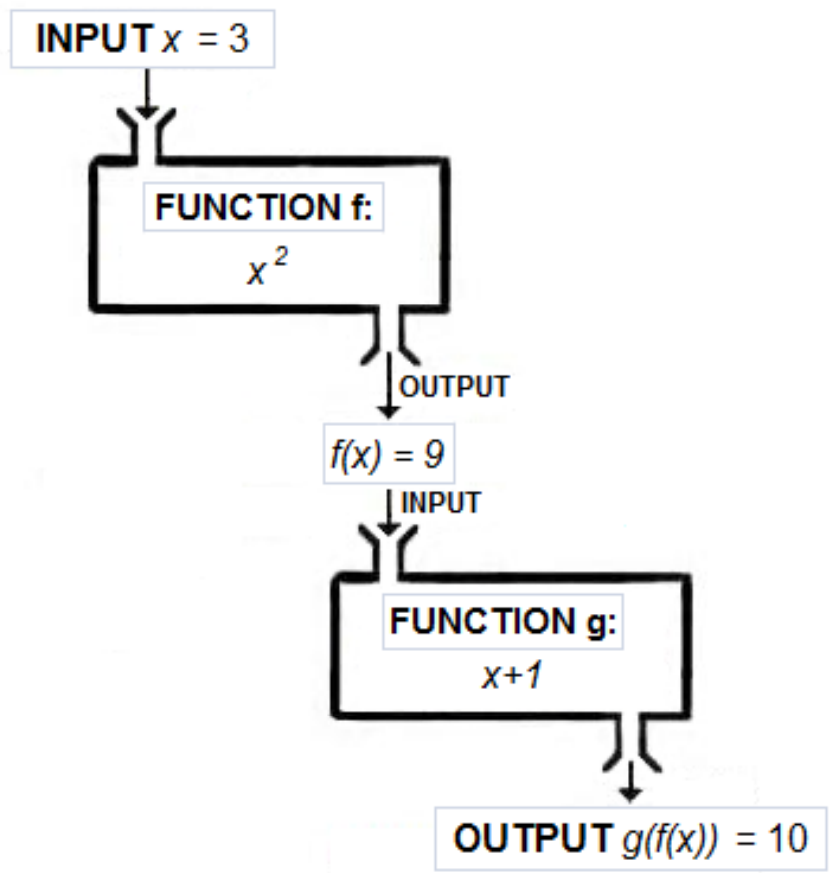


Multi-Valued Function



Single-Valued Function

Functions (…cont'd)



Random Variables

- Characteristics of a RV
  - Single valued function
  - Applies to the outcomes of a “repetitive operation”

Example #1 of a RV

Suppose your testing defective LCD monitors from a batch, before they leave the manufacturing facility. For simplicity, the test is a simple pass/fail. You choose 3 monitors from the batch. This means your sample space is:

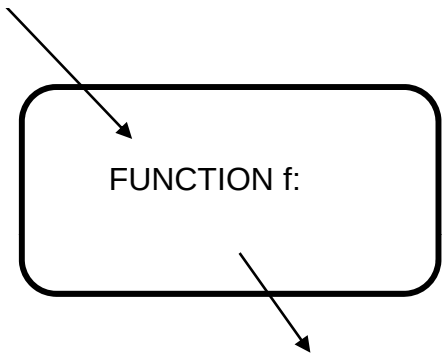
S = {NNN, NND, NDN, DNN, NDD, DND, DDN, DDD}, where D means defective and N means non-defective

Suppose you're interested in the number of defects. Let f(e) represent a RV for the number of defects, where e is each element of the sample space.

The possible outcomes (or outputs) from the RV called f(e) are the numbers 0,1,2, and 3. These outcomes are random quantities determined by what 3 monitors you choose.

Example #1 of a RV (…cont'd)

elements of sample space
S = {NNN, NND, NDN, DNN, NDD, DND, DDN, DDD}



f(e) = 2

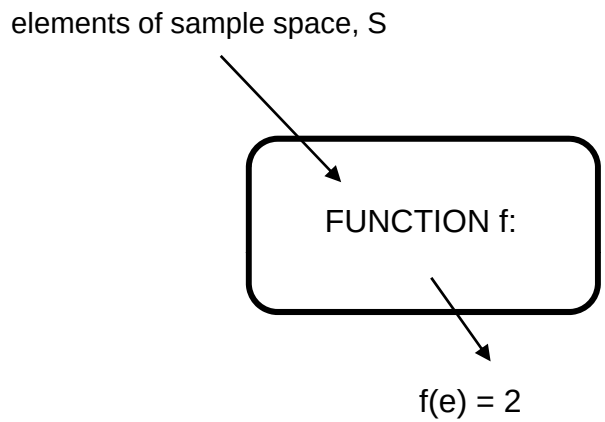
when f(e) = 2 the event, E = {DDN, DND, NDD}
when f(e) = 1 the event, E = {NND, NDN, DNN}
when f(e) = 0 the event, E = {NNN}

Example #2 of a RV

This time suppose you're rolling two dice simultaneously. An example of a RV could be number of dots that appear, denoted f(e). First define the sample space.

$$S = \left\{ \begin{bmatrix} (1,1) & (2,1) & (3,1) & (4,1) & (5,1) & (6,1) \\ (1,2) & (2,2) & (3,2) & (4,2) & (5,2) & (6,2) \\ (1,3) & (2,3) & (3,3) & (4,3) & (5,3) & (6,3) \\ (1,4) & (2,4) & (3,4) & (4,4) & (5,4) & (6,4) \\ (1,5) & (2,5) & (3,5) & (4,5) & (5,5) & (6,5) \\ (1,6) & (2,6) & (3,6) & (4,6) & (5,6) & (6,6) \end{bmatrix} \right\}$$

Example #2 of a RV (…cont'd)



when f(e) = 2 the event, E = { (1,1) }
when f(e) = 3 the event, E = { (2,1) (1,2) }
when f(e) =4 the event, E= { (1,3) (3,1) (2,2)}

Sixth basic theory of probability

The sixth basic concept is that of a **probability distribution function (pdf)**. This is the function that identifies the probability of each output value from a random variable. Going back to the previous example let's calculate all possible output values for the random variable. They are: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

when f(e) = 2 the event, E = {(1,1)}, # of outcomes = 1
when f(e) = 3 the event, E = {(2,1) (1,2) }, # of outcomes = 2
when f(e) =4 the event, E= {(1,3) (3,1) (2,2)}, # of outcomes = 3
...
when f(e)=12 the event, E= {(6,6)}, # of outcomes=1

The total number of outcomes (or elements in the sample space) is 36. Recall the definition of probability to be the fraction of trials in which a specific event occurs. For RV value, f(e), an associated probability, denoted p, can be calculated.

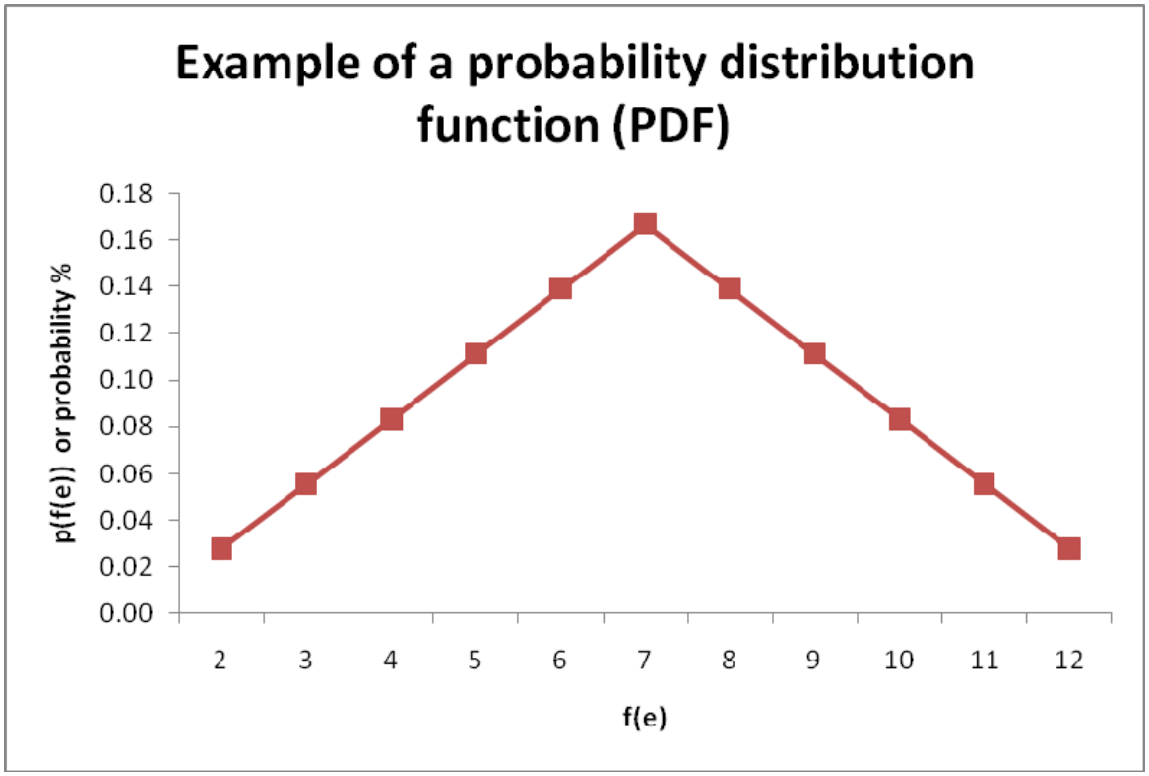
f(e)	2	3	4	5	6	7	8	9	10	11	12
p(f(e))	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Sixth basic theory of probability (…cont'd)

From previous slide:

f(e)	2	3	4	5	6	7	8	9	10	11	12
p(f(e))	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Graphing this data:



The CDF, Revisited

Now that we've defined a PDF we finally have the framework to define and discuss a CDF. Before doing so, let's simplify the notation a bit. f(e) will now be denoted as x and represents the output of the RV function. P refers to the probability function. Recall, the inputs of this function are all elements in the sample space, along with all specific event of interest; it outputs the fraction of occurrences an event occurs when repeated over and over again.

Having said that, let's denote the CDF as F(x'). x' is the new variable of interest, it's the input to the function. It can take on essentially any number.

$F(x') = P(x \leq x') = \sum' p(x_i)$

where $\sum'$ denotes the summation for all values of i for which $x_i \leq x'$

A CDF describes the probability that a RV x with a given PDF will be found at a value less than or equal to x'. Intuitively, it is the "*area so far*" function of the probability distribution.

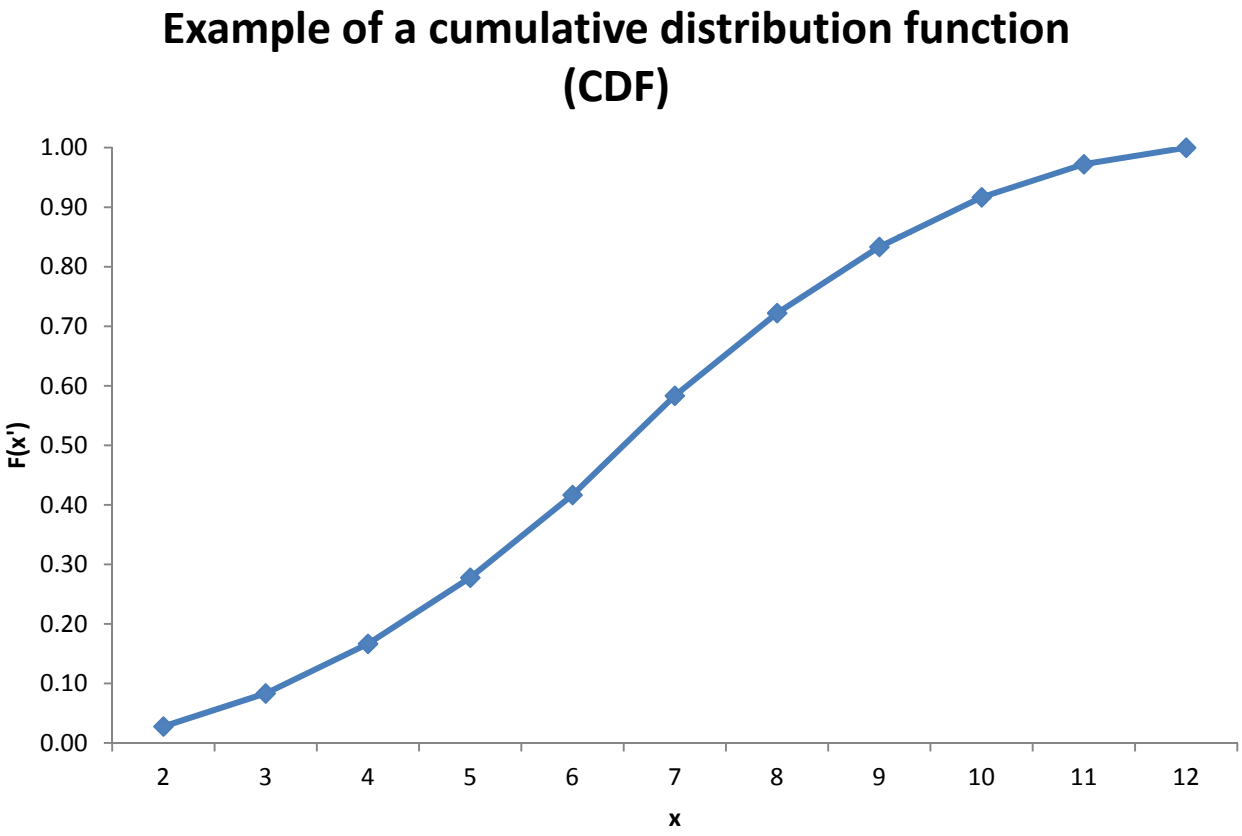
The CDF, Revisited (…cont'd)

Going back to the die example, let's calculate the CDF:

x	2	3	4	5	6	7	8	9	10	11	12
F(x')	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	1
		= 1/36	=1/36	=1/36	=1/36	=1/36	=1/36	=1/36	=1/36	=1/36	=1/36
		+ 2/36	+2/36	+2/36	+2/36	+2/36	+2/36	+2/36	+2/36	+2/36	+2/36
			+3/36	+3/36	+3/36	+3/36	+3/36	+3/36	+3/36	+3/36	+3/36
				+4/36	+4/36	+4/36	+4/36	+4/36	+4/36	+4/36	+4/36
					+5/36	+5/36	+5/36	+5/36	+5/36	+5/36	+5/36
						+6/36	+6/36	+6/36	+6/36	+6/36	+6/36
							+5/36	+5/36	+5/36	+5/36	+5/36
								+4/36	+4/36	+4/36	+4/36
									+3/36	+3/36	+3/36
										+2/36	+2/36
											+1/36

The CDF, Revisited (…cont'd)

Graphing data from previous slide:



The CDF, Revisited (…cont'd)

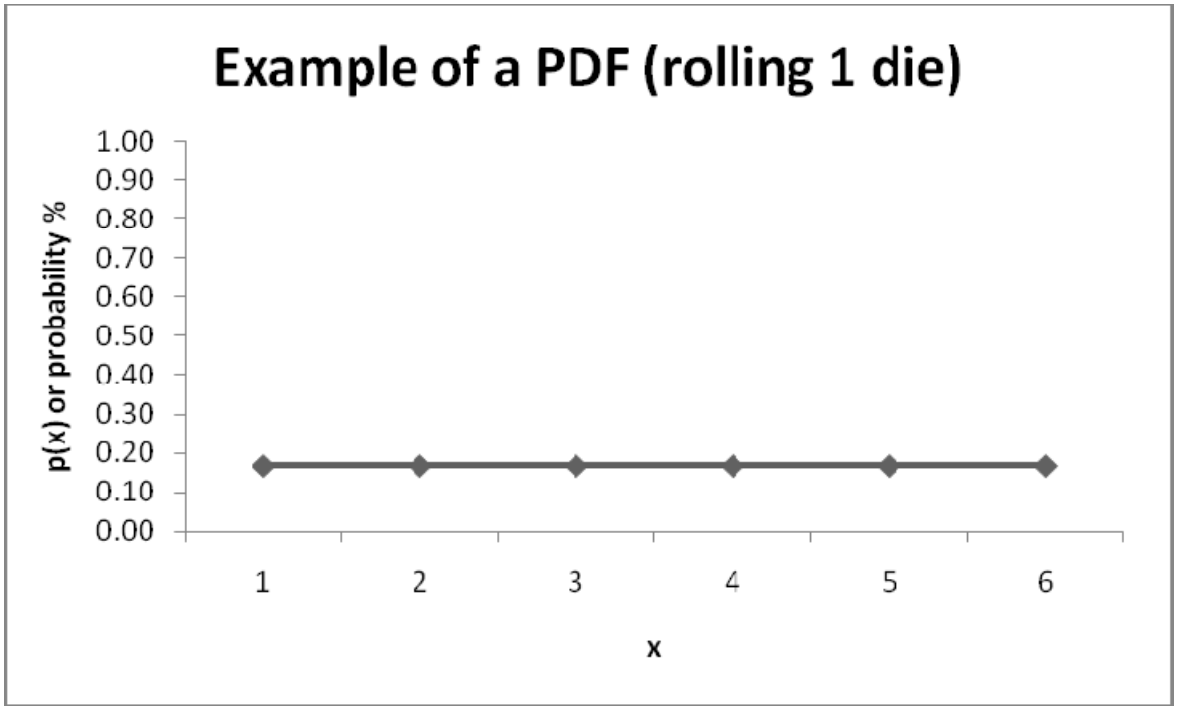
Interestingly enough, in this example the CDF looks like an S! Was a lying to you? Let's simplify the dice example. Suppose instead of throwing two dice, we now decide to only throw one die. And like the previous example, we're still interested in counting the number of dots. So what's the sample space S? It's very simply S = {1,2,3,4,5,6}. We know the RV is the number of dots appearing, just like the last example.

The CDF, Revisited (…cont'd)

So to calculate the CDF, first we need to know the PDF.

x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6

Graphing this data:

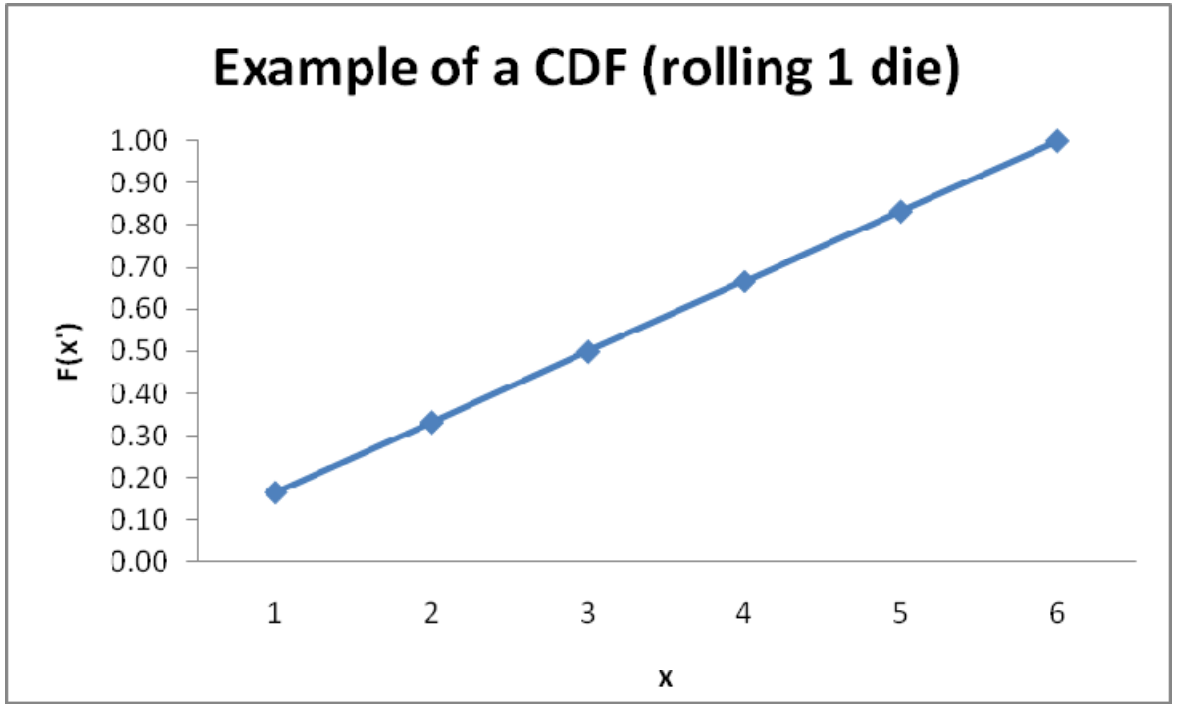


The CDF, Revisited (…cont'd)

Given the PDF, we can now calculate the CDF:

x	1	2	3	4	5	6
F(x')	1/6	2/6	3/6	4/6	5/6	6/6

Graphing this data:



First basic theory of probability applied to the field of **cost estimating**

First off, we need to accept that we're dealing with a repeatable process. Is purchasing an aircraft a repeatable process? Is purchasing a Boeing 737 a repeatable process? Is purchasing UAV a repeatable process? Is purchasing a computer network a repeatable process? What about a single computer, a hard drive, a processor, a circuit card assembly, a breadboard, a capacitor, a resistor, silicon, soldering paste? What about heating and electricity at the manufacturing plant? What about all the tooling and robotics used to mass produce the circuit card assembly?

Yes, you can do all these things many times

But what about if you're doing something new? No. Doing something new is by definition not-repeatable. If it's new, it's never been done before. So we have to careful when we apply probability and statistics to new designs.

Application to CANES

- CANES has very little new capability. It's a consolidation of existing shipboard networks. It was created during the POM-08 budget process as a cost savings initiative. No SW development is expected on CANES. The program is structured to only purchase COTS SW licenses.

First basic theory of probability applied to the field of **cost estimating** (cont'd)

Is the process repeatable "under the same conditions"?

Well, it's also much easier to accept this notion the deeper you dig. For example, the heating costs in the summer at the Boeing 737 plant producing one 737 per month has much more clearly defined conditions than say the price of a commercial aircraft. In many cases, the more specific the component, the easier it is to control the conditions or the environment.

The problem of homogeneous conditions is everywhere. It's especially evident when looking at comparable data sets. Sure it costs $10 million for Northrop Grumman to develop this computer system. But since this is Lockheed Martin, will the system be cheaper or more expensive?

First basic theory of probability applied to the field of **cost estimating** (cont'd)

The other issue related to homogeneous conditions is that of "time." The time variable is a really tricky one. Not only does it have to do with "inflation" but it's also related to changes in technology and worker productivity. Generally, the shorter the time-span, the more accurate the estimate. If we wanted to improve the accuracy of an estimate, we might simply shorten the time-span.

- Certainly, this area merits further discussion, but for the sake of explaining what was done on the CANES estimate, we assume that the TOC to develop and support the CANES system is repeatable "under the same conditions"

Second basic theory of probability applied to the field of **cost estimating**

The second basic theory of probability is that of sample space. For any cost estimate, what's the sample space? It's all the possible values (or elements) for which the system could cost. Rolling of dice, flipping of coins, sampling defects, etc., all have a clearly defined sample space. A cost estimate is much trickier. Should zero be included? What happens if there's is a positive return? What's the maximum possible cost?

For the CANES estimate, the sample space for the TOC was {$10.2 billion – $17.3 billion}. I'll explain how, and why this sample space was chosen in a minute.

Third basic theory of probability applied to the field of **cost estimating**

- The third basic theory of probability is the concept of event. Recall an event is any subset of the sample space. Quite simply, for CANES any range of values, or any point value between $10.2 billion and $17.3 billion qualifies an event.

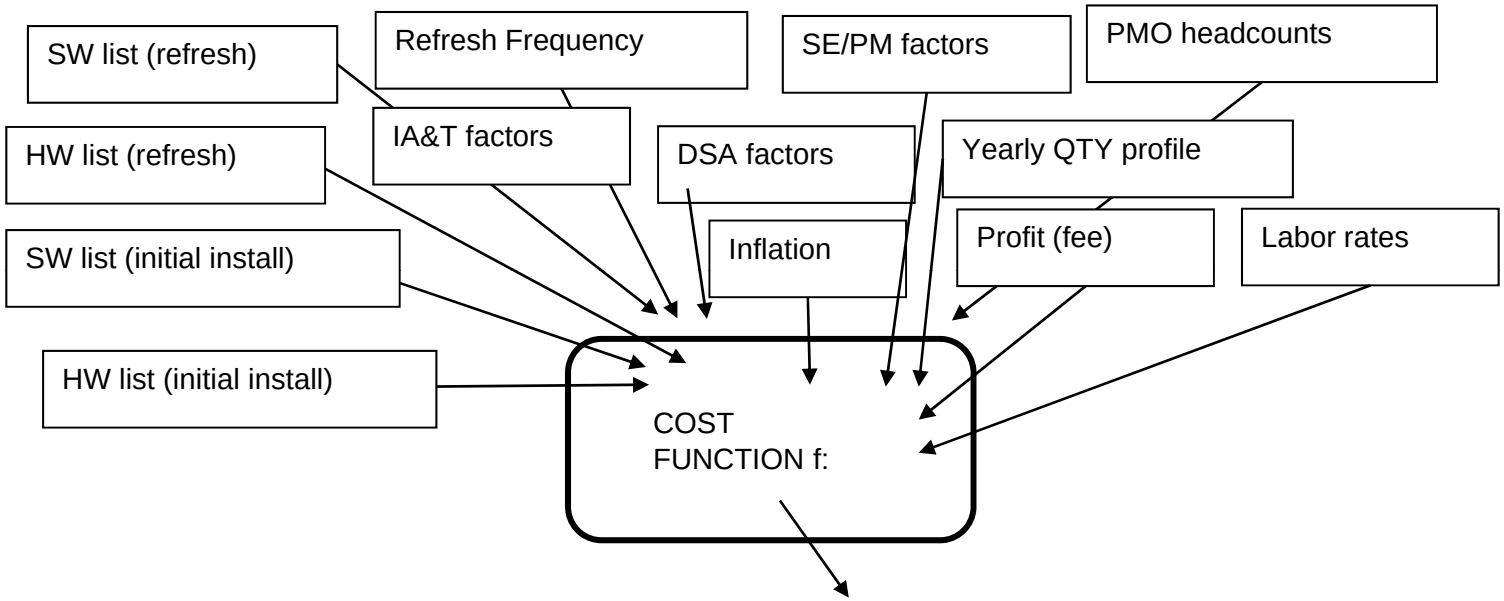
Fourth and Fifth principles applied to CANES

In order to talk about the probability of an event in the sample space, a cost function (or cost model) must first be developed. This is what many analysts refer to as the "base estimate," or "base model." The heart of the CANES estimate is a long list of required HW equipment and required SW applications. Each part is separately priced based on either the:

- commercial market price
- GSA Prices
- Actual costs incurred on predecessor shipboard networks (ISNS, SCI-Net, SubLan, CENTRIX-N)

For an individual ship, most of HW and SW is then planned to be repurchased every 6 years. The assumed lifecycle is from FY10-FY31.

The CANES deterministic model



F(HW list, SW list, Refresh Frequency, IA&T factors, DSA factors, SE/PM factors, inflation, labor rates, yearly QTY profile, PMO headcounts, etc.) = 12.7 billion

CANES model discussion

All the inputs to the cost function (or cost model) exhibit a degree of uncertainty. Not only that, the base model by no means accounts for all the possible areas of uncertainty. These areas are limitless and can include things such as:

- Supply and demand of worldwide COTS electronic equipment
- Labor rate fluctuations in a given geographic market
- Aggressive or conservative testing requirement
- Major changes in requirements
- Unethical employees or managers
- Political decisions
- Human instinct to spend you're entire budget so it won't get cut
- Trade imbalances
- Etc.

CANES model discussion (…cont'd)

Any of the inputs of the base model can be designated a RV. It's the analysts job to choose the most relevant areas of uncertainty. Because of limited time, resources, data, and computing power, great care must be taken in choosing the RVs. The process is subjective and most certainly open to debate. In choosing the RVs, the mathematical field of "sensitivity analysis" is a very useful tool.

Sensitivity Analysis

It's important to note that sensitivity analysis is it's own separate field, and can be done without any knowledge of probability and statistics. It simply answers the question, "What's the largest cost driver?" What happens to the output if we change the values of one or many inputs? It becomes very interesting when you're working on an optimization problem. For example, suppose you may want to minimize the transportation cost (a function of distance traveled) of a school bus route, but, the school district wants to intermix students from different parts of town so your routes are subject to "constraints."

CANES model discussion (…cont'd)

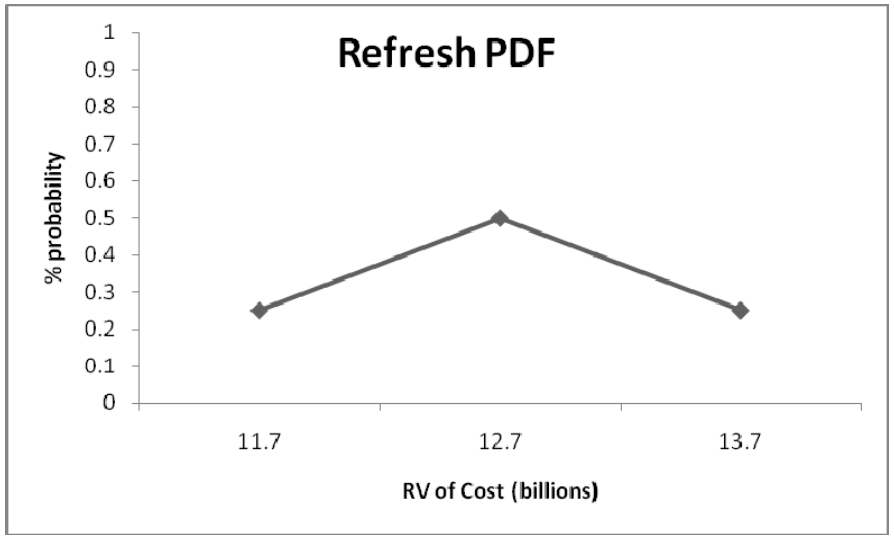
In building the model and playing with all the possible RVs, and in discussion with the program office managers and engineers, we chose the following random variables:

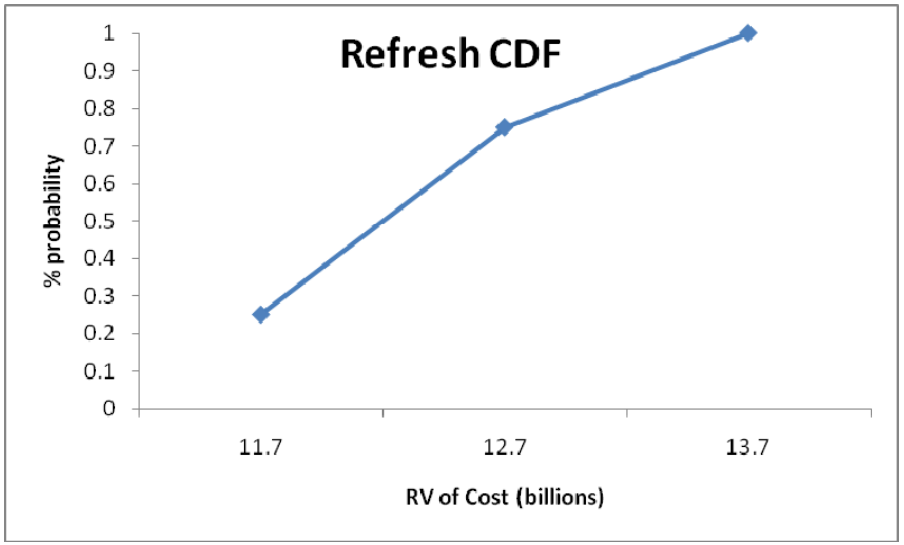
- 6 - Year refresh rate
- ∑ HW material costs
- ∑ SW application licenses costs

CANES Refresh

- The data from the CNO ship availability plan, limited the refresh rate to 5- Years. Discussions with PM's and engineers indicated the max refresh rate was 7- Years.
- Because the base model accounts for all 249 ships, defining a RV for each individual ship proved infeasible. But, using the offset function, in MS Excel, allowed us to create three distinct refresh profiles.
  - Profile 1: 6 - Year refresh rate: Twice as likely to occur than either profile 2 or 3 (50% probability)
  - Profile 2: 5 - Year refresh rate, equal probability as the 7-Year refresh rate (25% probability)
  - Profile 3: 7 - Year refresh rate, by deduction, 25% probabibility)

CANES Refresh PDF & CDF

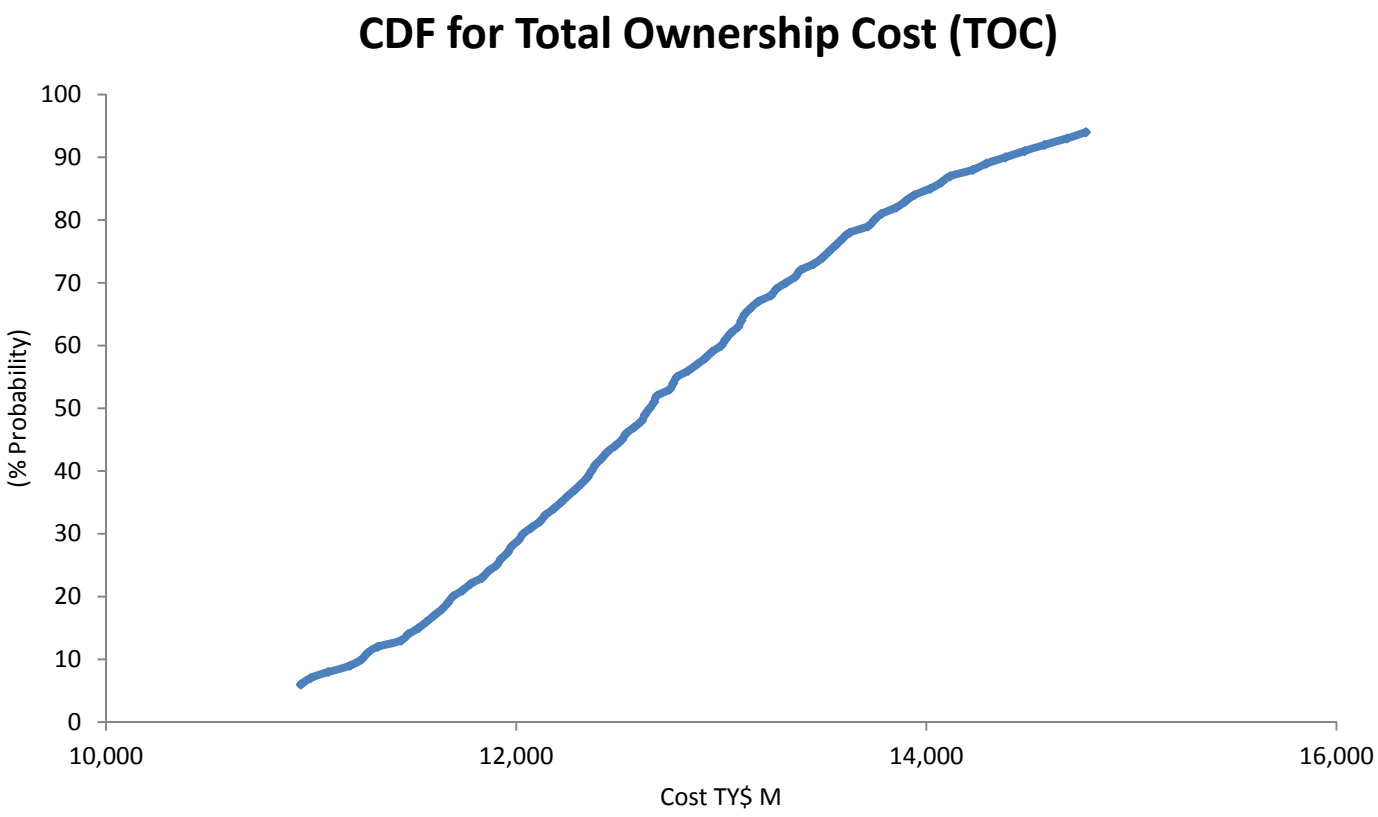




CANES HW & SW uncertainty

- The data from the Bureau of Labor and statistics indicated that the cost of SW application and HW electronics is best characterized by a normal distribution.
- A normal distribution is a specialized PDF. Recall it's output is the probability of an event. The inputs are: the RV of interest, the mean of the elements in the sample space, the standard deviation of all elements in the sample space. Without formally defining mean or standard deviation, I'll just say observed standard deviation from the BLS data was about 16% when sampling commercial sales of different electronics components, adjusted for inflation. The mean is trivial since we were applying this RV to ∑, not individual items on the list.
- In layman's terms, the standard deviation is a measure of dispersion. In this example, if we purchased a LCD display from many different vendors, where the average cost (mean) is $100, then 68% of the time when we purchased the LCD display it would be within $16 ($84-$116), and 95% of the time the cost would be within 2 * standard deviation or $32 ($68 - $132).

Putting it all together



Questions?