SAND2010-4637P

Validation, Sensitivity and Uncertainty Analyses

KRMC Training Program

Module 3: WIPP Performance Assessment

July 16, 2010

Thomas Kirchner
Sandia National Laboratories

Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

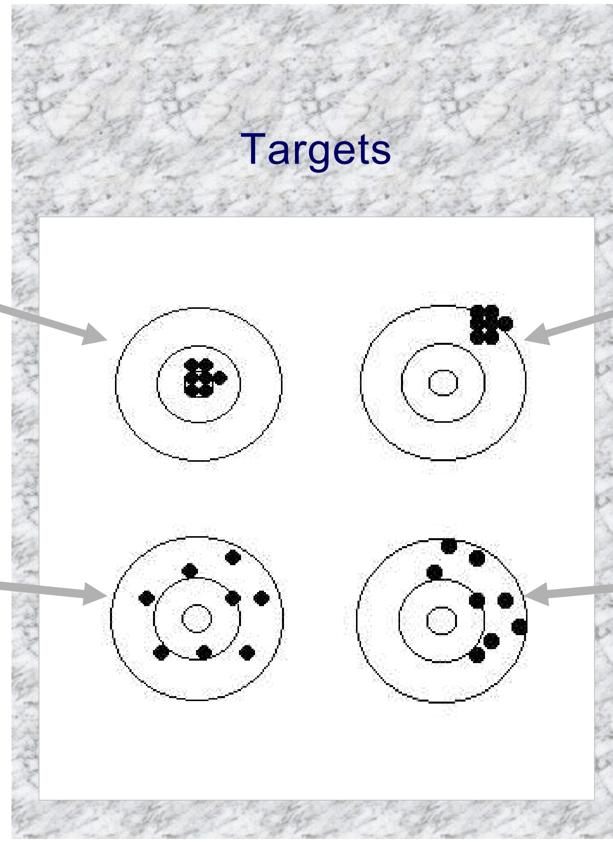
Evaluating models and their reliability

- **Validation** (testing numerical predictions against measured quantities)
- **Uncertainty analysis** (estimating the possible error distribution in numerical estimates)
- **Sensitivity analysis** (determining the relative influence of parameter estimates and their uncertainties on the numerical prediction and its uncertainty)

Accuracy and precision determine model reliability

High accuracy
High precision

Low accuracy
High precision



High accuracy
Low precision

Low accuracy
Low precision

Model validation to establish credibility

- Face validity (peer review)
- Verification (quality of code & calculations)
- Evaluation of model structure, parameter values
- Testing model predictions against independent measurements

Methods of model testing

- **Subjective judgement** (comparing predictions (P) against observations(O): plot P against O values; calculate P/O ratios)
- **Statistical methods**

  -**Paired *t*-test** with H_0: mean of paired differences = 0

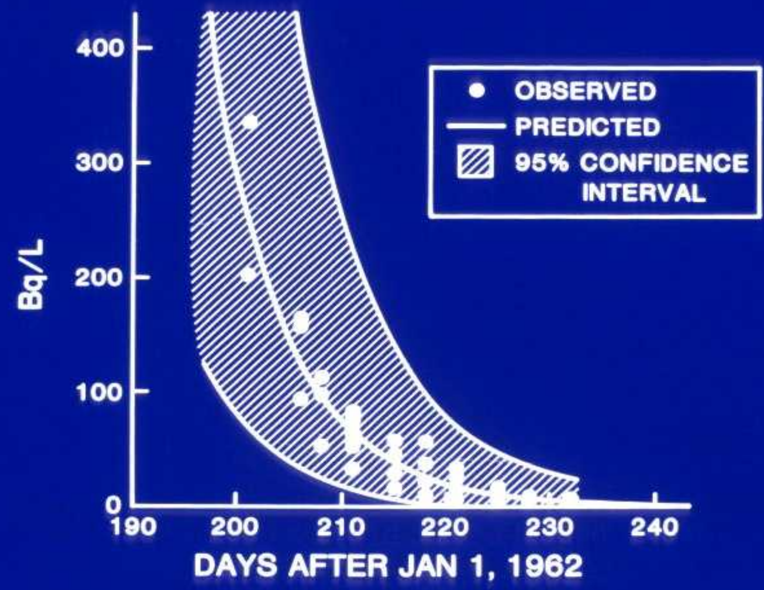
  -**Interval test**: How many observations fall outside a distribution of predictions?

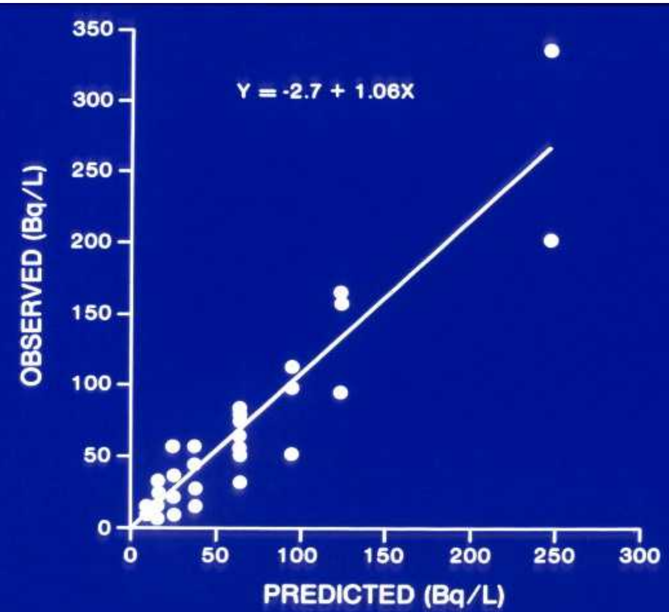
  -**Regression**: Plot O vs. P and calculate slope, intercept & correlation coefficient

  -**Runs test**: Plot (P-O) values against time to test for dynamic errors

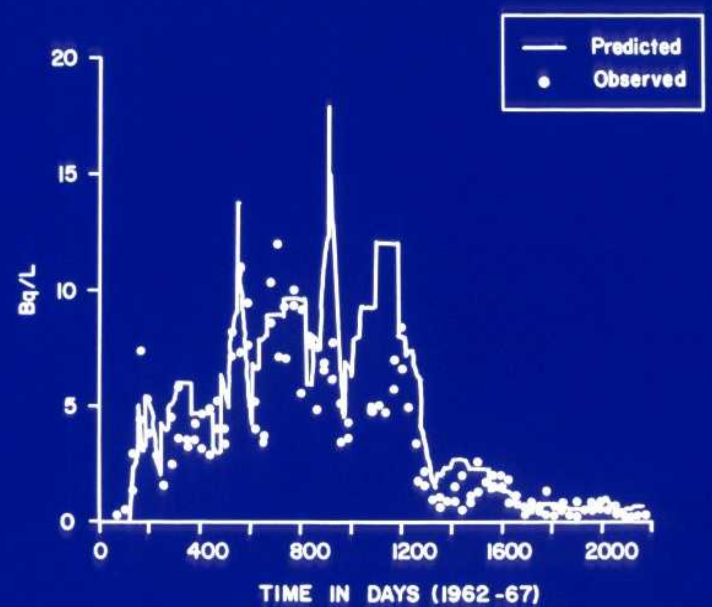
Sandia National Laboratories

Examples of comparing model predictions to observed data





^{131}I in milk



^{131}I in milk

^{90}Sr in milk

Uncertainty Analysis Implies Probabilistic Modeling

If we were “certain” of everything in our calculations, there would be no uncertainty.

But we are rarely certain of everything.

Simple (but typical) Risk Model

Many risk models are simple algebraic expressions:

R = A·B/C + D·E/F

for example

$$C_{air} = S_{release}\frac{\chi}{Q}$$

$$D_{thyroid} = \frac{C_{air}V_{dep}I_{cow}T_{forage/milk}I_{child}D_{child}}{\left(K_{weathering} + K_{decay}\right)D_{forage}}$$

- **Each model input can be either**
  - **Constant**
  - **Variable**

Constants

Unit Conversions

- Kilograms per pound
- Liters per cubic meter
- Days per year

Physical Properties

- Avogadro's number
- Half-life of Pu

Parameters

- Release rate
- Maximum ingestion rate

Constants

Unit Conversions	• Kilograms per pound • Liters per cubic meter • Days per year
Physical Properties	• Avogadro's number • Half-life of Pu
Parameters	• Release rate • Maximum ingestion rate

Are these constants?

Sandia National Laboratories

Variables

- Amount of milk ingested in a day by a child
- Number of days per year spent outside
- Amount of pasture grass consumed by a dairy cow
- Fraction of the diet of a cow that is hay
- Body weight of the individuals in this room

Variable (V) or Constant (C)?

V / C

- ❑ ❑ Body weight of individuals in this room
- ❑ ❑ Average body weight of individuals in this room
- ❑ ❑ Sample average concentration of Pu in soil
- ❑ ❑ Concentration of Pu in soil at different sampling locations in an exposure unit
- ❑ ❑ 95% upper confidence limit for the arithmetic mean concentration
- ❑ ❑ 95th percentile concentration

Variable (V) or Constant (C)?

V	C	
☒	❑	Body weight of individuals in this room
❑	☒	Average body weight of individuals in this room
☒	❑	Sample average concentration of Pu in soil
☒	❑	Concentration of Pu in soil at different sampling locations in an exposure unit
☒	❑	95% upper confidence limit for the sample arithmetic mean concentration
☒	❑	95^{th} percentile concentration from sampled data

Uncertainty

- Variability is an inherent property of a feature; it cannot be reduced by collecting more data.

  Example: rate of ingestion in cows

- Lack of knowledge; it can be reduced (but rarely eliminated) by collecting better data.

  Example: rate of ingestion of milk by a specific child in 1957

Subjective Versus Objective Uncertainty

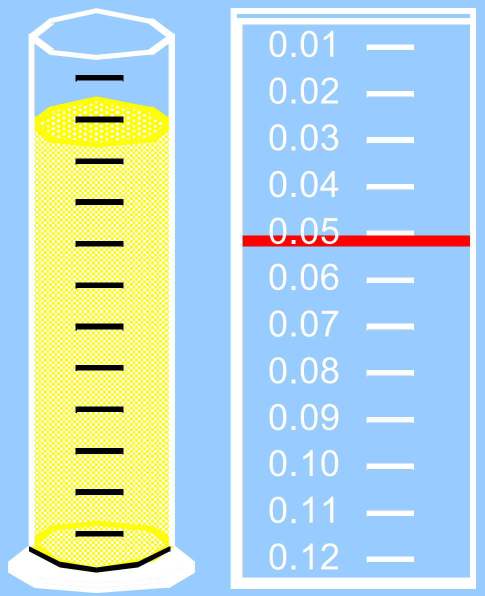
- **Aleatory (Type A) uncertainty is natural variability.**
- **Epistemic (Type B) uncertainty represents lack of knowledge or confidence.**

- **Epistemic uncertainty is sometimes referred to as subjective uncertainty, because subjective methods are often used to quantify the uncertainty – but don't fall into that habit**

Sandia National Laboratories

Is Imprecision Subjective or Objective Uncertainty?

- Precision traditionally has referred to the uncertainty (Type B) that arises when making a measurement, such as reading a level in a graduated cylinder or reading the weight on a scale.
- With today's instruments, the uncertainty on measurements can include both Type A and Type B uncertainty.



Sandia National Laboratories

Random Error versus Systematic Error

- **Random error is stochastic variability in a measurement**

- **Systematic error, or bias, is a consistent difference between the true mean and the estimated mean of measurements**

Sandia National Laboratories

Random Error versus Systematic Error

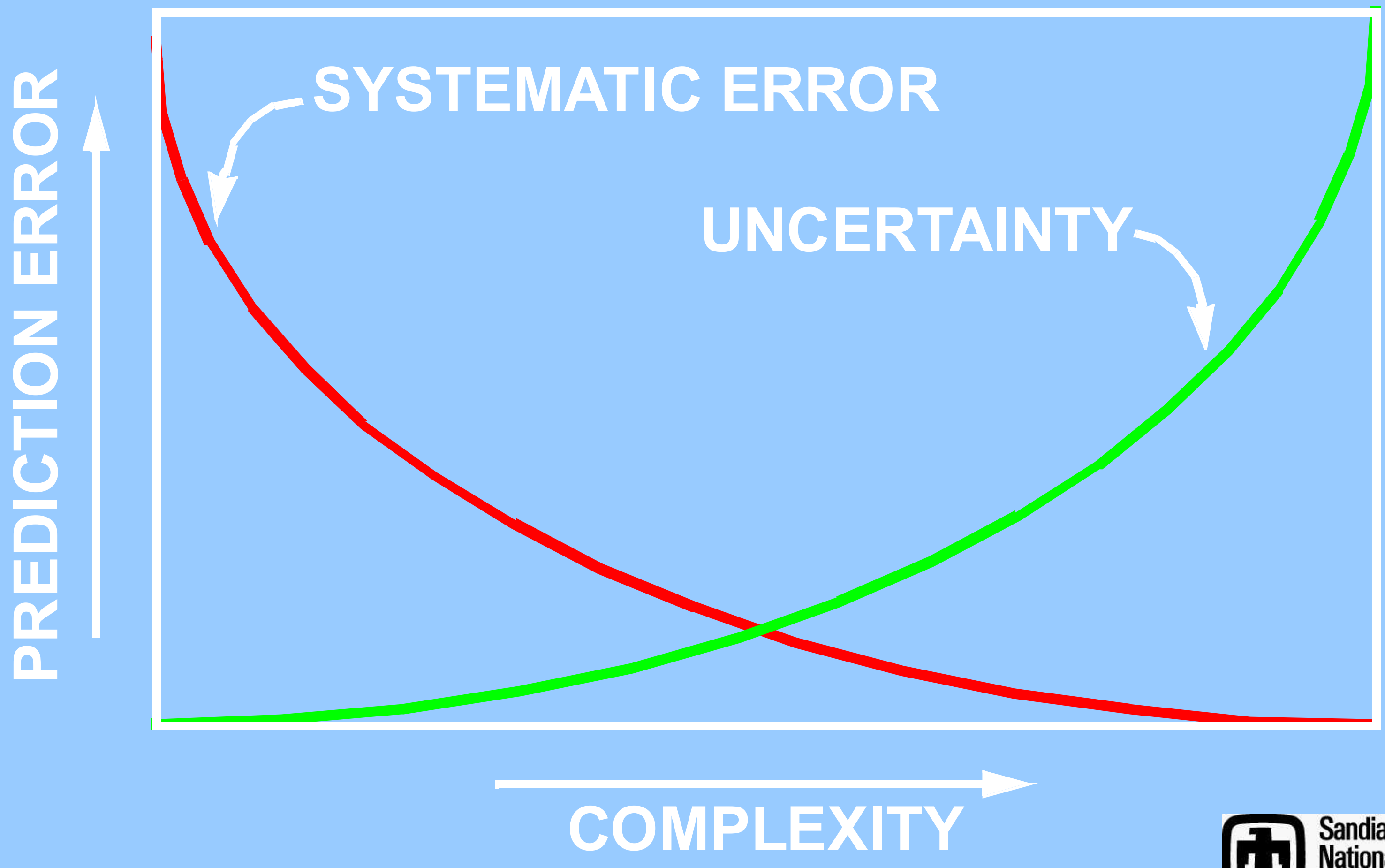
- **Random error can be estimated using replicated (repeated) measurements**

- **Systematic error, or bias, can not be estimated using replicated (repeated) measurements. It can sometimes be identified using other methods, e.g. testing instruments against known standards**

Random Error

Random sampling error gives rise to differences between the true parameters of a distribution and the values estimated from samples.

Random Error versus Systematic Error



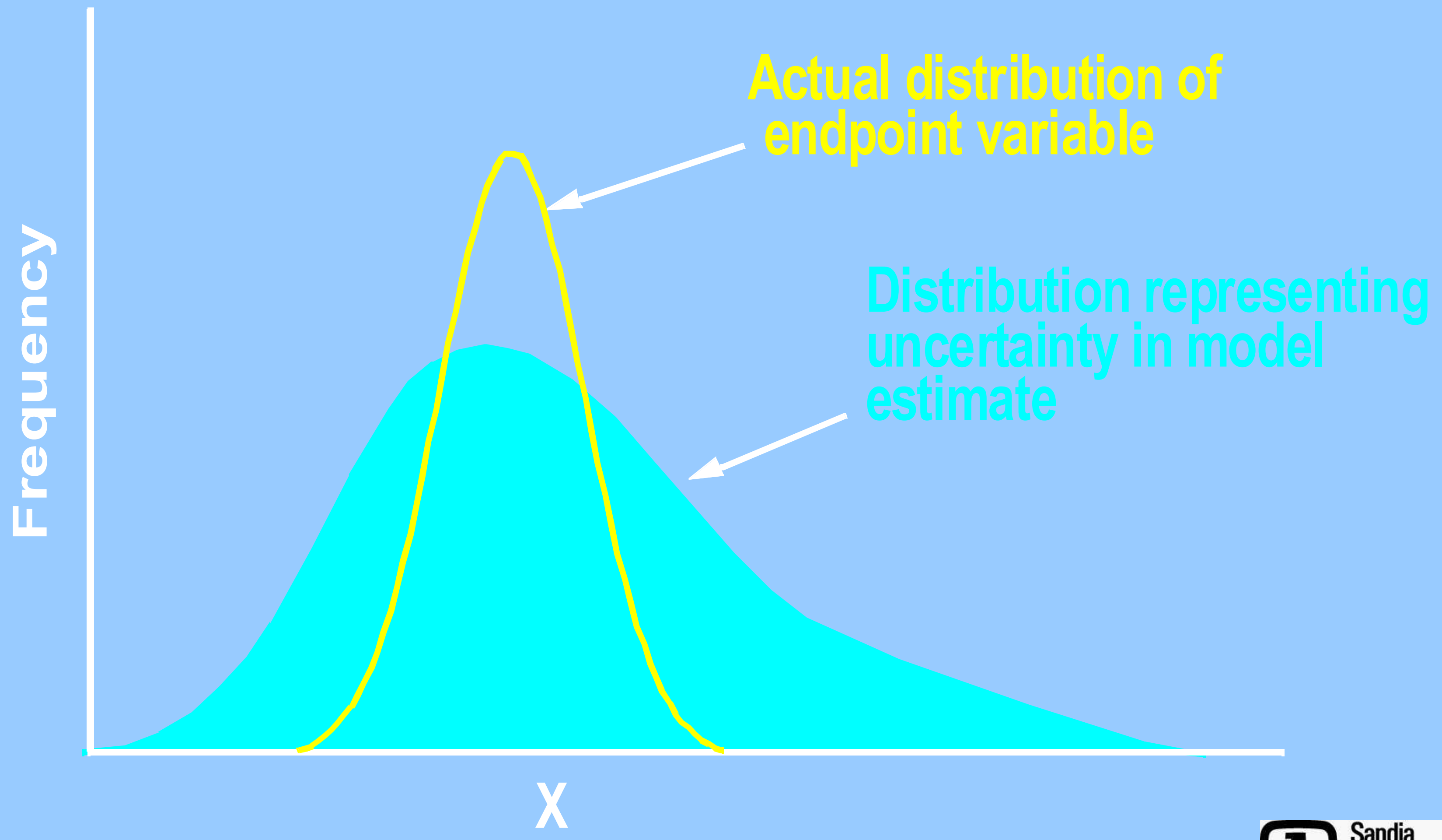
Uncertainty Analysis

- Uncertainty analysis provides a distribution of possible real values, or of possible errors in a prediction
- Important for maintaining honesty and gaining credibility
- Model predictions are of little utility if accuracy and uncertainty are unknown
- Uncertainty can be quantified from numerous actual measurements, or from propagating errors in model inputs and assumptions to obtain a distribution of possible output values
- Error propagation techniques:
  - -analytical
  - -random sampling from distributions (Monte Carlo)

Model Output

- If one or more of the model inputs are RANDOM VARIABLES, then the model output will be a distribution of values.
  - A "random variable" is a function that maps an event into a real number
  - For example, consider tossing a coin. The events are heads or tails. A random variable, X, can be defined to map a head to 1 and a tail to 0.
- Thus, when the risk manager asks "What is the risk?", the answer CANNOT be just some isolated number. Rather, the answer MUST convey information on the nature of the output distribution.

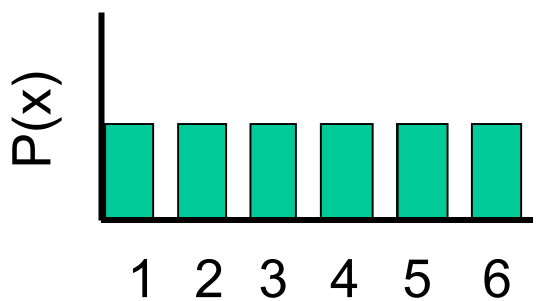
Uncertainty Should Encompass True Results



What is a Distribution?

A distribution represents the probabilities of one or more random variables occurring

For example, the distribution of the values from the roll of 1 die would look like:



Distribution of values for 1 die

P(x) = 1/6 for all x

What is a Distribution?

A distribution represents the probabilities of one or more random variables occurring

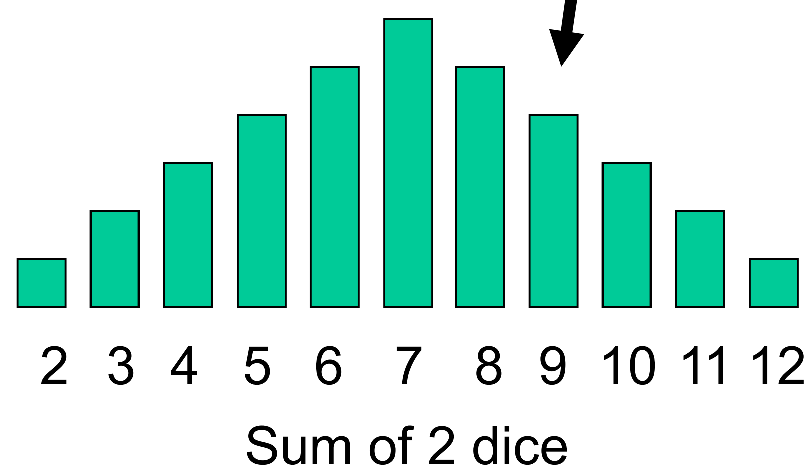
Second Die

First Die

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Permutations

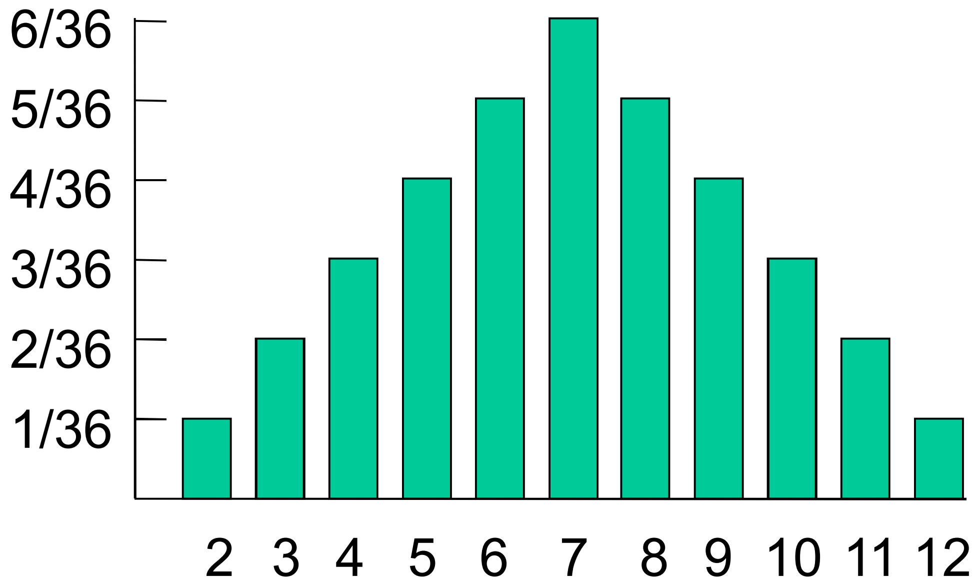
Combinations



Sandia National Laboratories

Probability Density Function

A PDF represents the probabilities of random variables as a function of their values



$$f_X(x) = \begin{cases} P[X = x_j] \; if \; x = x_j, \; j = 1,2,3,\ldots \\ 0 \quad\quad\quad\quad\quad if \; x \neq x_j \end{cases}$$

Sandia National Laboratories

Probability Density Function

The integral of the PDF equals 1

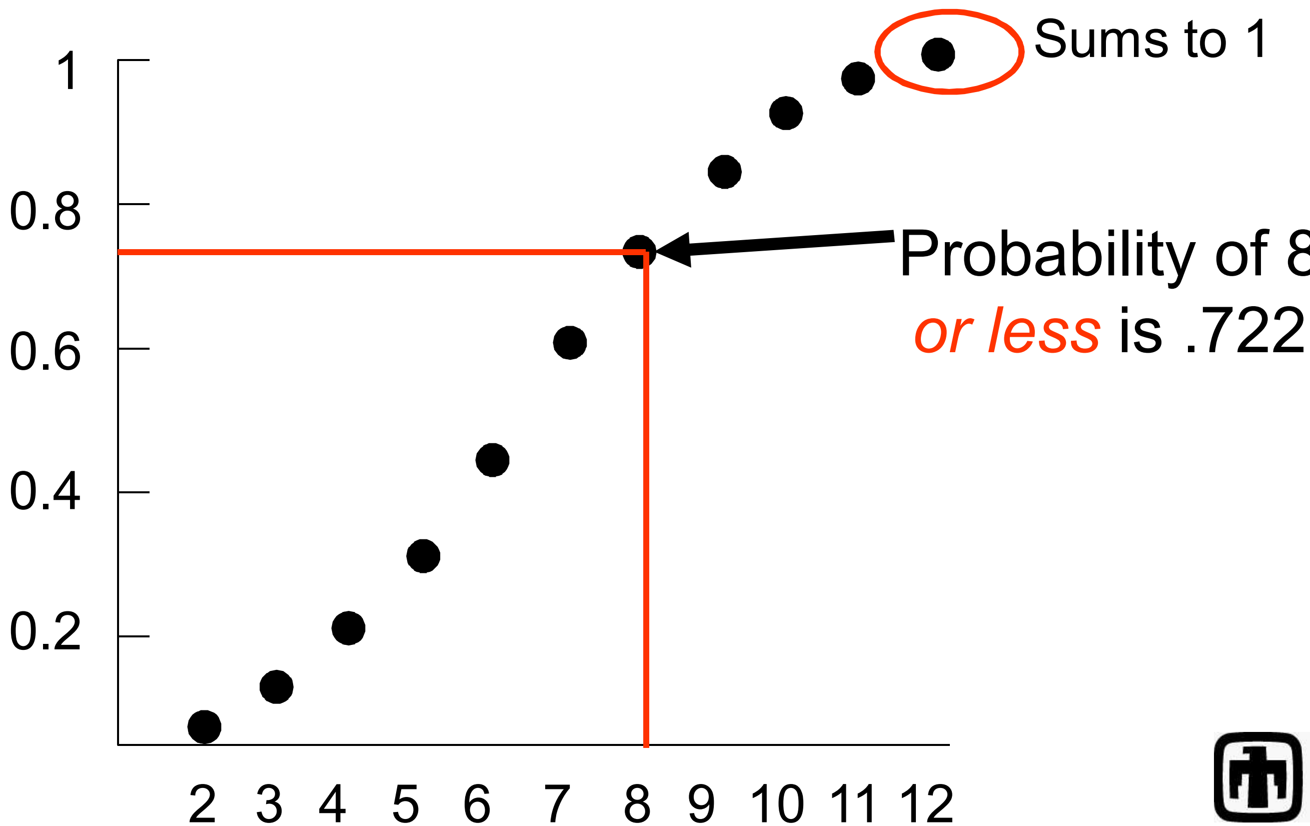
$$\sum_{i} P[X = x_i] = 1$$

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

Sandia National Laboratories

Cumulative Distribution Function

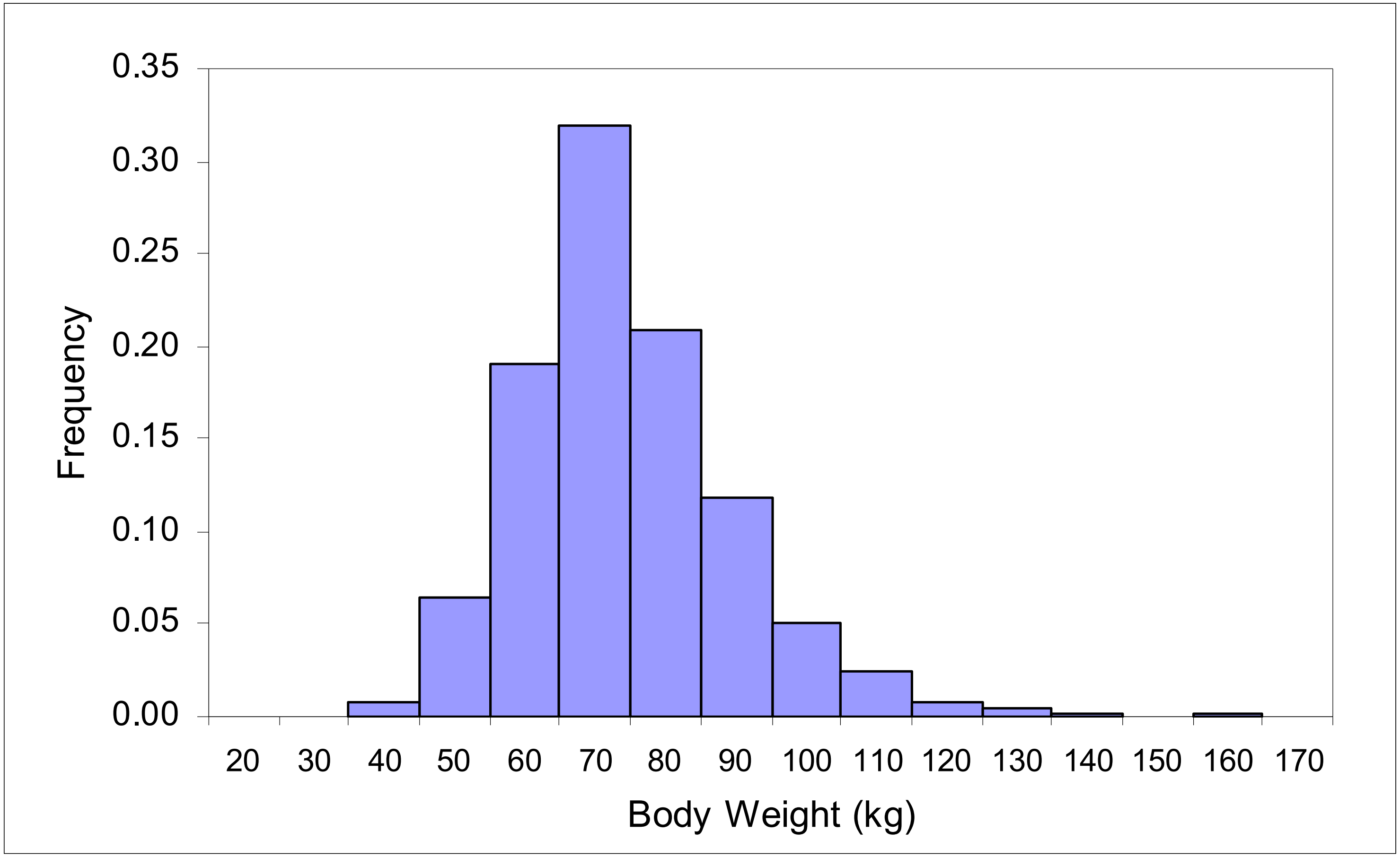
A CDF represents the cumulative probabilities of random variables as a function of their values



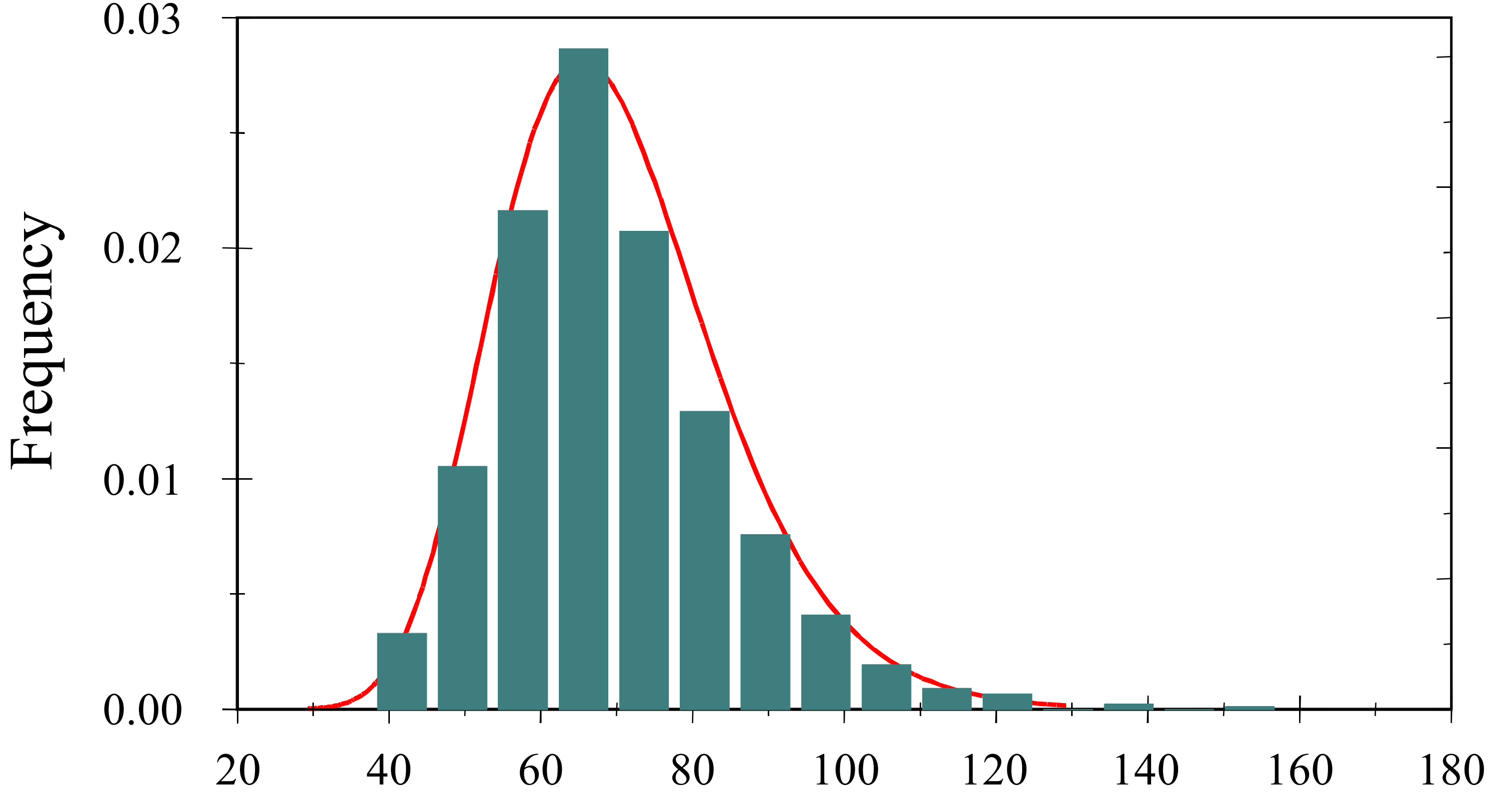
Cumulative Distribution Function

A CDF can also represent the cumulative probabilities of a "continuous" random variable as a function of its values

Body Weight of Men - Histogram



Body Weight of Men – Probability Density Function



Body Weight of Men - Summary Statistics

N	= 1000
Mean	= 69.7 kg
Stdev	= 15.2 kg
Min	= 38 kg
Max	= 157 kg
95th %ile	= 98 kg

Cumulative Distribution Function

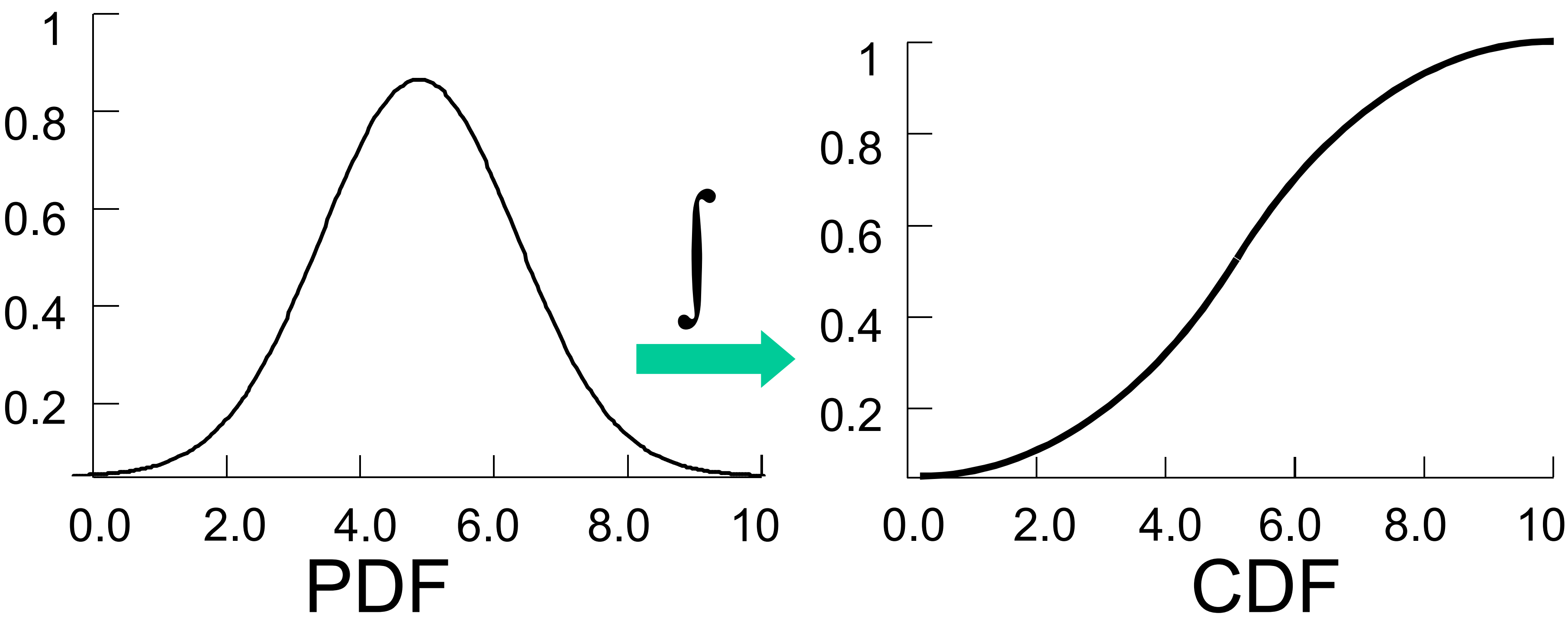
The CDF is the integral of the PDF

$$F_X(x) = \sum_{i;x_j<x} P[X = x_i]$$

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$$

Cumulative Distribution Function

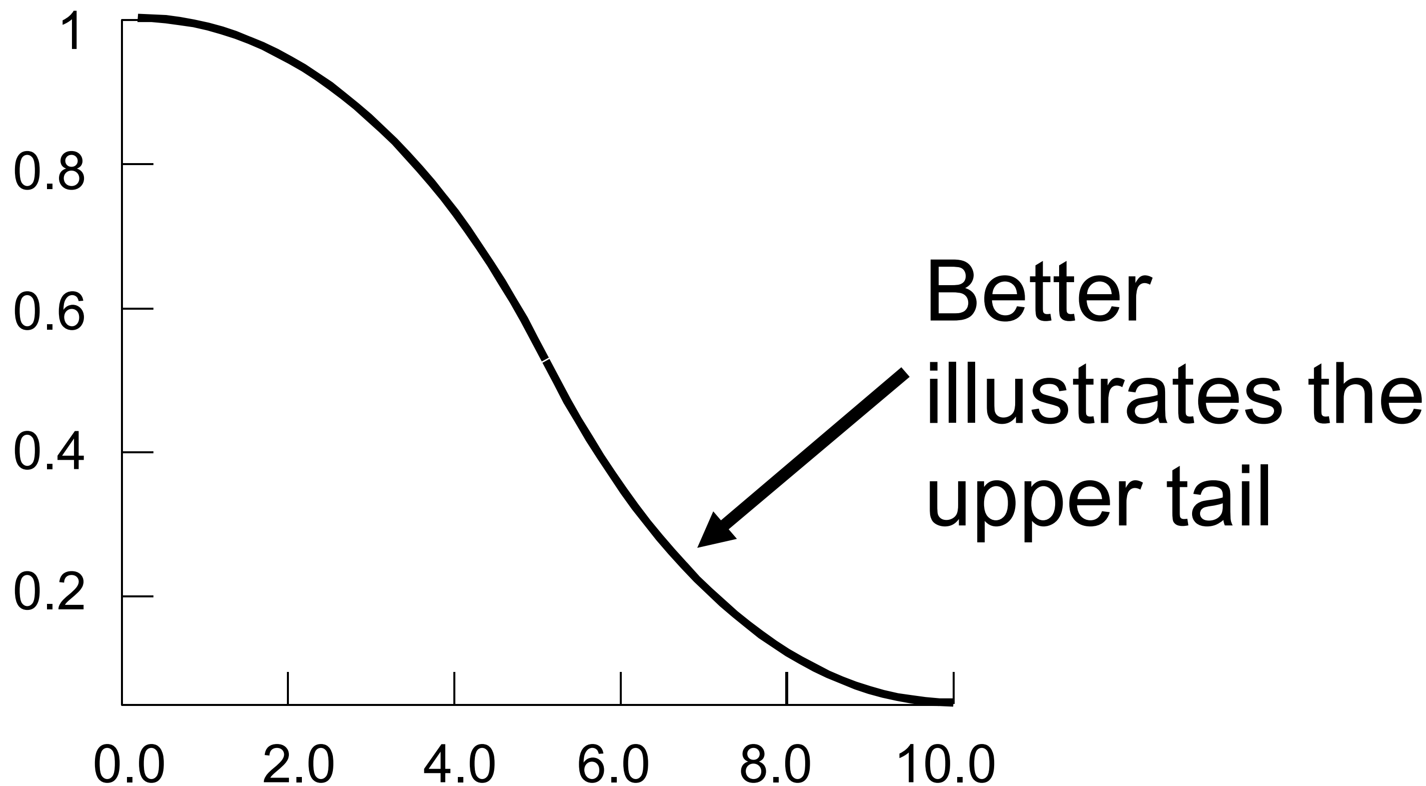
The CDF is the integral of the PDF



Sandia National Laboratories

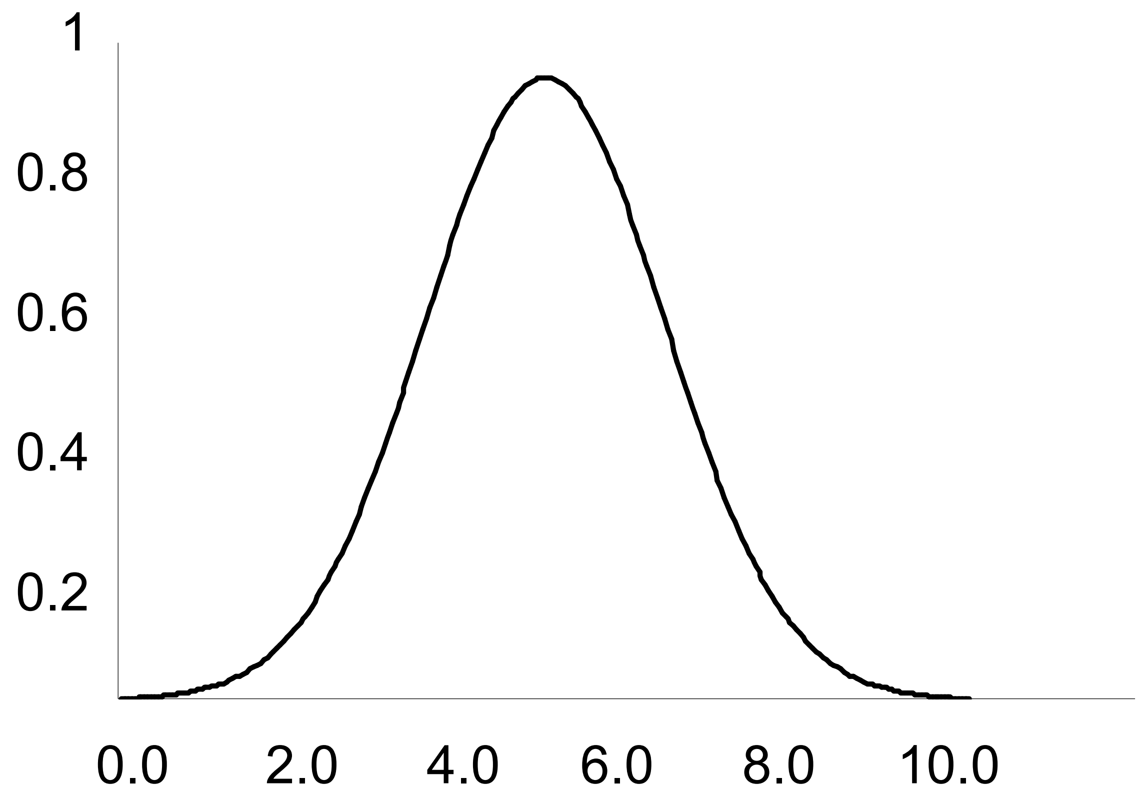
Complimentary Cumulative Distribution Function

A CCDF represents the cumulative probabilities of being greater than some value

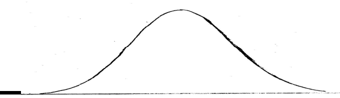
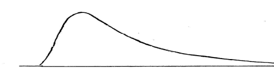
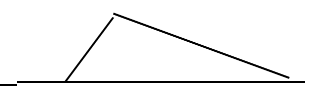
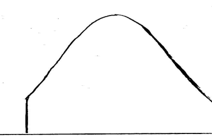
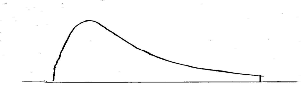
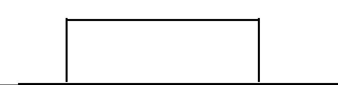


Probability Density Function

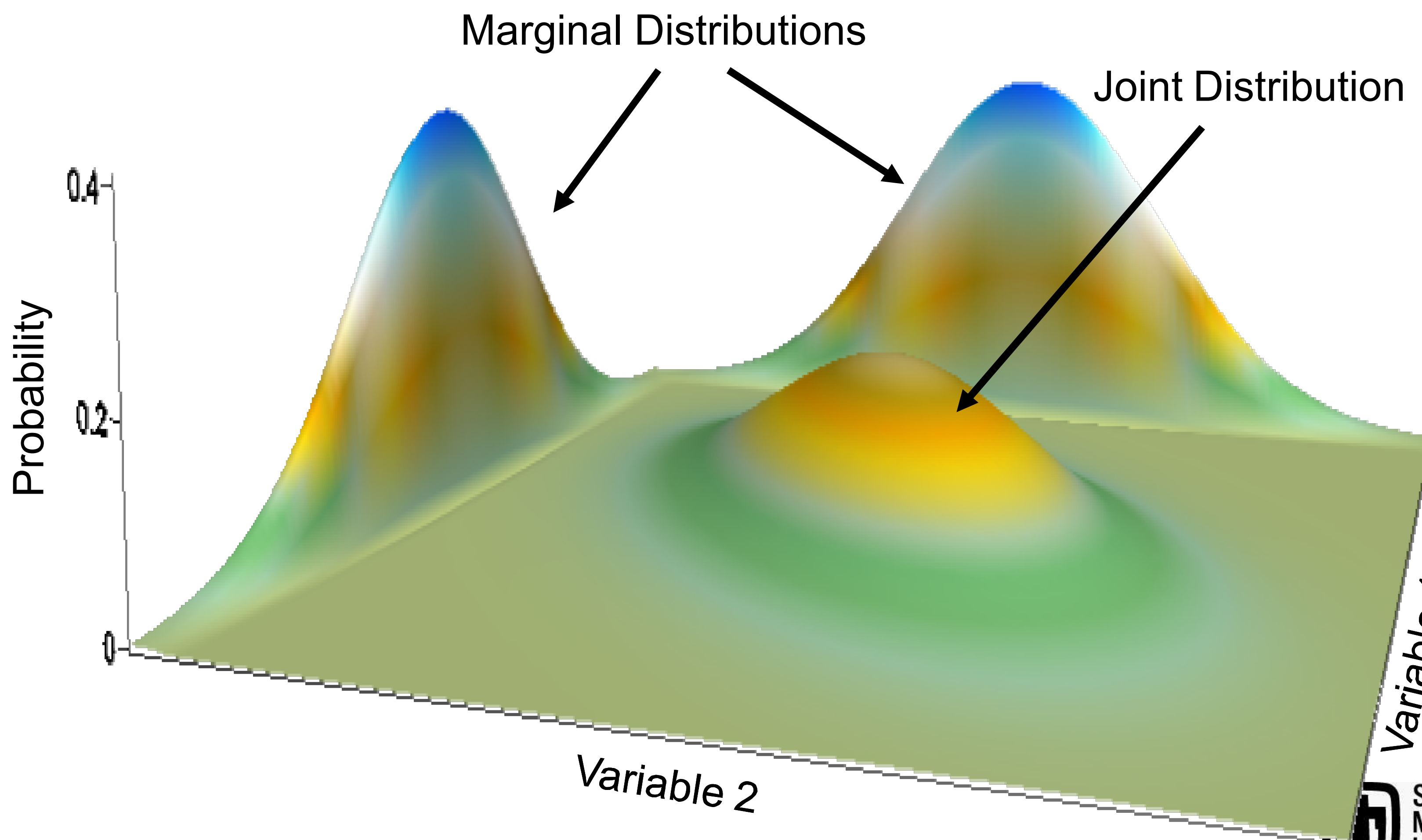
A PDF can also represent the cumulative probabilities of a "continuous" random variable as a function of its values



Distributions Often Used for Uncertainty Analysis

Distribution type	Most likely value (mode)	Descriptors of width	General shape
Normal	Mean (x)	Std dev (s)	
Lognormal	Geometric mean (GM)	Geometric std dev (GSD)	
Triangular	Mode	Min, max	
Truncated normal	Mean (x)	Std dev + min, max	
Truncated lognormal	Geometric mean (GM)	GSD + min, max	
Rectangular	none	Min, max	

Conditional variables can be described as multivariate distributions

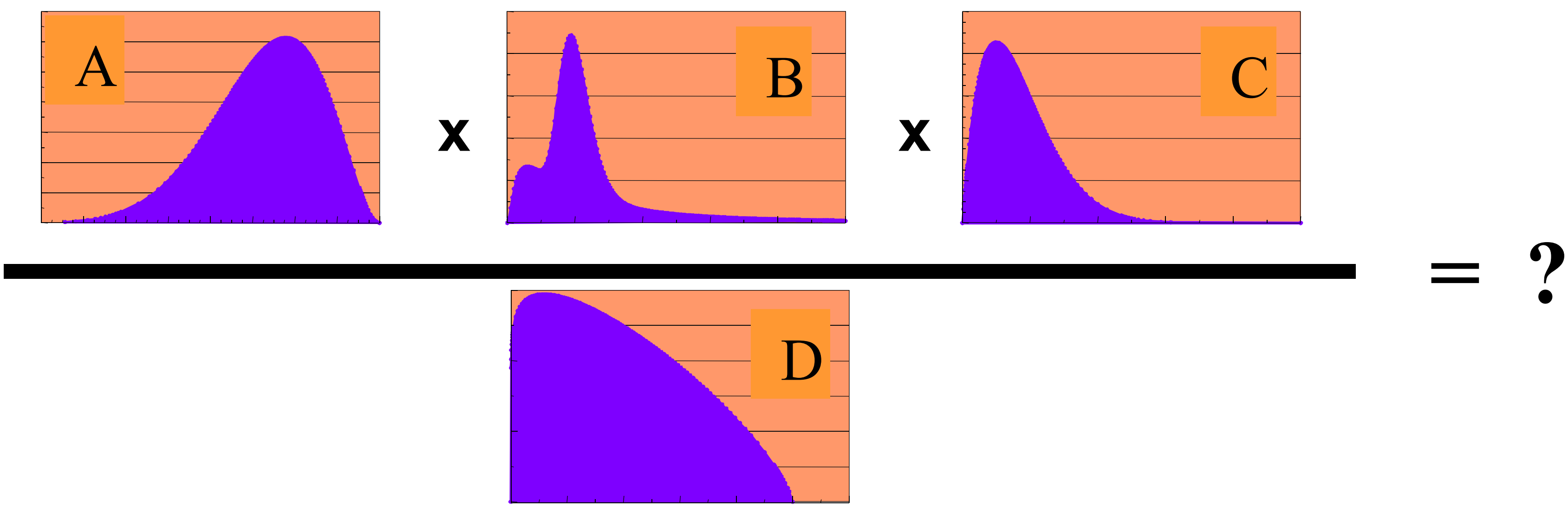


Multivariate Distributions

- Conditional variables can be described as multivariate distributions
- There are a limited number of multivariate distributions
- All have marginal distributions of same type:
  - Normal
  - Lognormal

One Minor Problem:
How do you combine distributions?

$$Y = \frac{A \times B \times C}{D}$$



Sandia National Laboratories

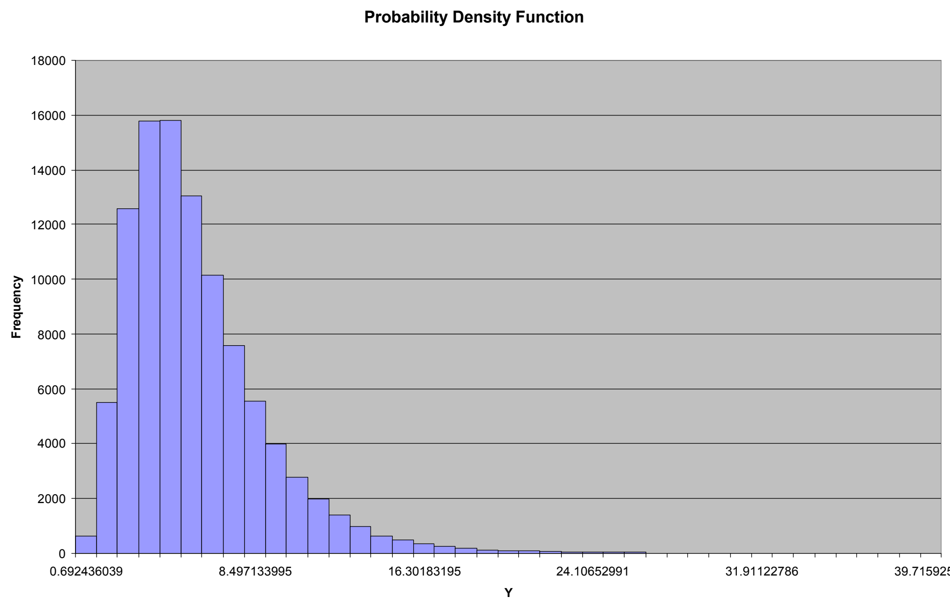
Two Alternatives

- **Analytical Solutions**
  - Useful in special cases, but often not mathematically feasible for complex models
- **Monte Carlo Simulation**
  - Relatively easy to perform with modern programs
  - Can usually be implemented, no matter how complex the model or the distributions

Example

$$Y = \frac{A \times B}{C}$$

A = LN(10,3)
B = LN(5,2)
C = LN(10,2)

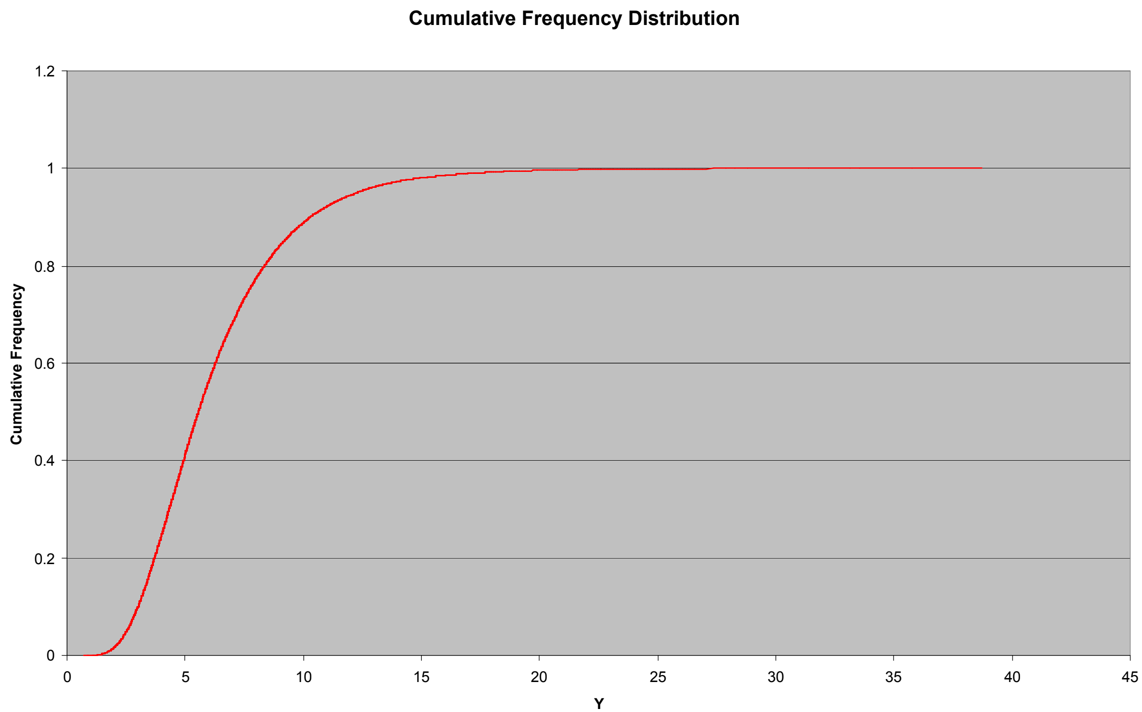


PDF = Probability Density Function

Example

$$Y = \frac{A \times B}{C}$$

A = LN(10,3)
B = LN(5,2)
C = LN(10,2)



CDF = Cumulative Distribution Function

Example

Descriptive Statistics for Y

Mean	6.236E+000	Skewness	1.625E+000
Median	5.550E+000	Kurtosis	4.888E+000
Standard deviation	3.173E+000	Standard error	1.003E-002
Variance	1.006E+001	Geometric mean	5.556E+000
Minimum	6.924E-001	Geometric SD	1.618E+000
Maximum	4.069E+001	N	100000
N positive values	100000		

Example

Percentiles Based on Order Statistics

0.1%: 1.280E+000	25%:4.012E+000	60%:6.275E+000	95%:1.223E+001
1.0%: 1.812E+000	30%:4.318E+000	65%:6.690E+000	97.5%:1.419E+001
2.5%: 2.168E+000	35%:4.625E+000	70%:7.165E+000	99.0%:1.695E+001
5%: 2.514E+000	40%:4.924E+000	75%:7.694E+000	99.9%:2.410E+001
10%: 3.002E+000	45%:5.229E+000	80%:8.347E+000	
15%: 3.372E+000	50%:5.551E+000	85%:9.157E+000	
20%: 3.699E+000	55%:5.895E+000	90%:1.029E+001	

Analytical Methods of Propagation

Variance and Covariance

$$\bar{X} = E[X] = \int_{-\infty}^{\infty} x f_x(x) dx$$

$$Var[X] = E[(x-u)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f_x(x) dx$$

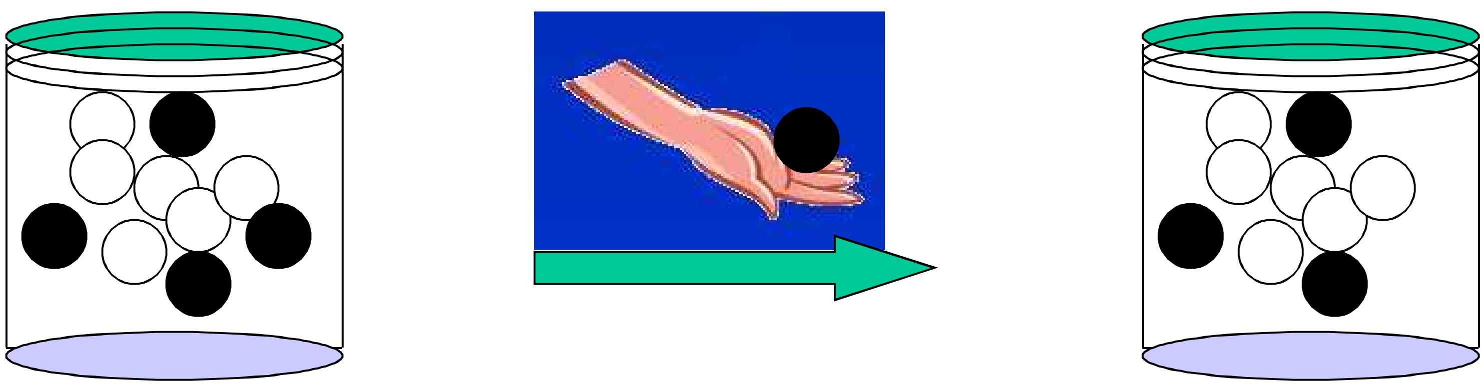
$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

Conditional Probability

- The probability of event A occurring, given that event B has already occurred
- P(A|B)
- Read this as "probability of A given B"

Dependence

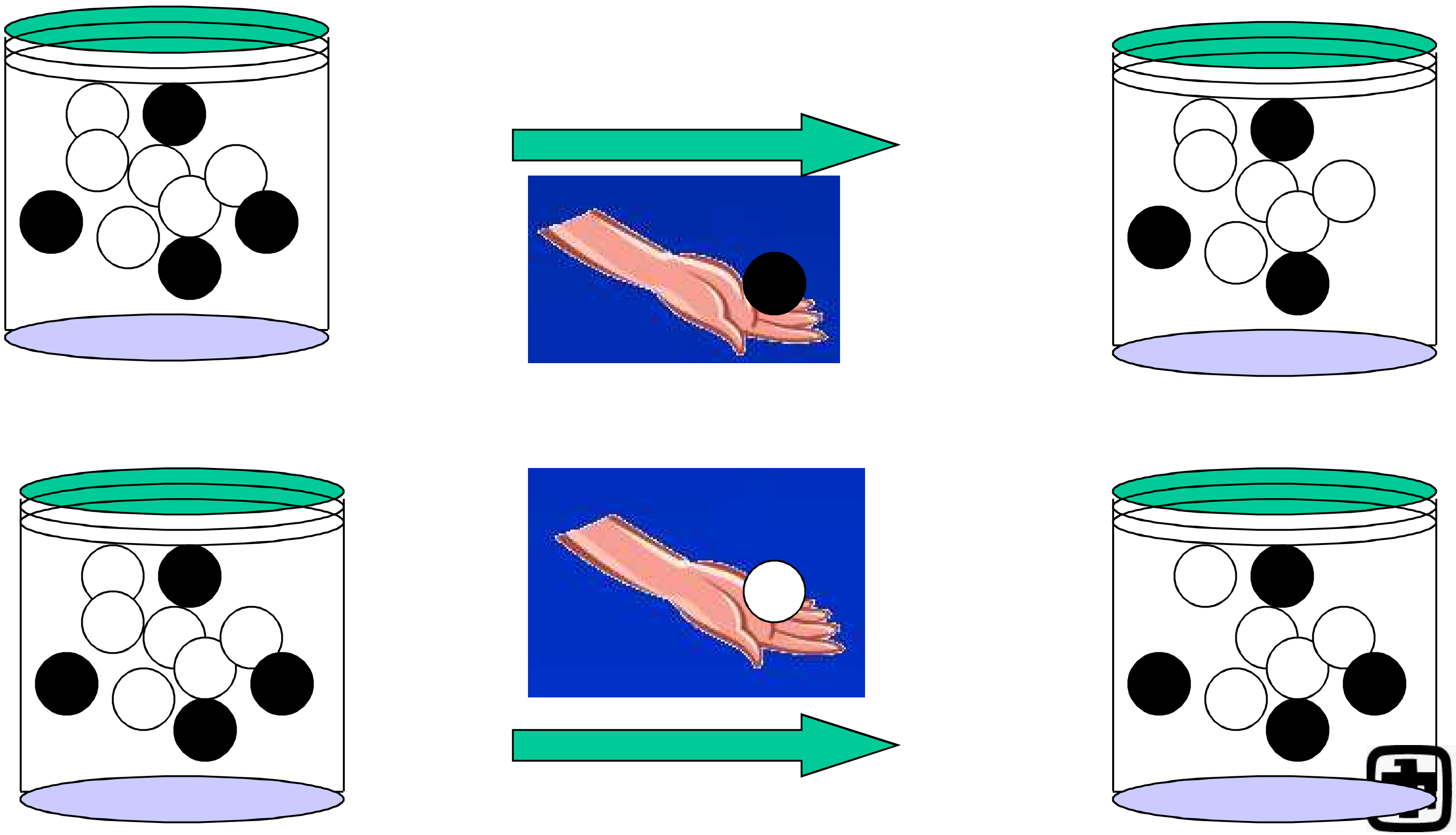
- If the occurrence of A changes the probability that B occurs then A and B are dependent
- Example: A jar contains 4 black balls and 6 white balls.
- Event A = take 1 ball at random, don't return to jar: P(black)=4/10



Dependence

Event B = take another ball,

– P(black) = 3/9 if A was black

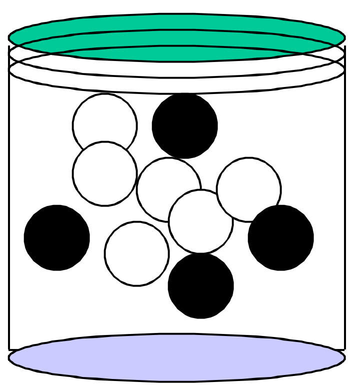


Independence

- **If the occurrence of A does not change the probability that B occurs then A and B are independent**
- **Example: A jar contains 4 black balls and 6 white balls.**

Event A = take 1 ball at random, return to jar: P(black)=4/10

Event B = take another ball, P(black) = 4/10 whether or not A was black or white



Independence and Dependence

- If A and B are independent then
  - P(B|A) = P(B)
- If A and B are dependent then
  - P(B)=P(B|A)P(A)+P(B|not A)P(not A)

  Example of jar sampled without replacement,

  What is the probability of B being black if you don't know the color of A?

  P(B = black) = (3/9)(4/10)+(4/9)(6/10) = 4/10

Sandia National Laboratories

Sum and Difference of Random Variables

- **Mean**

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

- **Variance of Sum**

$$\mathrm{var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{var}[X_i] - 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{cov}(X_i, X_j)$$

- **Variance of Difference**

$$\mathrm{var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{var}[X_i] + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{cov}(X_i, X_j)$$

- **If the variables are *independent*:**

$$\mathrm{var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{var}[X_i]$$

Product of Random Variables

- **Mean**

$$E\left[XY\right]=\mu_X\mu_Y+\mathrm{cov}(X,Y)$$

- **Variance**

$$\mathrm{var}\left[XY\right]=\mu_X^2\,\mathrm{var}\left[Y\right]+\mu_Y^2\,\mathrm{var}\left[X\right]+2\mu_X\mu_Y\,\mathrm{cov}(X,Y)-$$

$$(\mathrm{cov}\left[X,Y\right])^2+E\left[(X-\mu_x)^2(Y-\mu_Y)^2\right]+2\mu_Y E\left[(X-\mu_X)^2(Y-\mu_Y)\right]+$$

$$2\mu_x E\left[(X-\mu_x)(Y-\mu_Y)^2\right]$$

- **If X and Y are *independent*:**

$$E\left[XY\right]=\mu_X\mu_Y$$

$$\mathrm{var}\left[XY\right]=\mu_X^2\,\mathrm{var}\left[Y\right]+\mu_Y^2\,\mathrm{var}\left[X\right]+\sigma_x\sigma_y$$

Sandia National Laboratories

Quotient of Random Variables

- **Mean**

$$E\left[\frac{X}{Y}\right]=\frac{\mu_X}{\mu_Y}-\frac{1}{\mu_Y^2}\operatorname{cov}(X,Y)+\frac{\mu_X}{\mu_Y^3}\operatorname{var}[Y]$$

- **Variance**

$$\operatorname{var}\left[\frac{X}{Y}\right]=\left(\frac{\mu_X}{\mu_Y}\right)^2\left(\frac{\operatorname{var}[X]}{\mu_X^2}+\frac{\operatorname{var}[Y]}{\mu_Y^2}-\frac{2\operatorname{cov}(X,Y)}{\mu_X\mu_Y}\right)$$

- **If X and Y are *independent*:**

$$E\left[\frac{X}{Y}\right]=\frac{\mu_X}{\mu_Y}+\frac{\mu_X}{\mu_Y^3}\operatorname{var}[Y]$$

$$\operatorname{var}\left[\frac{X}{Y}\right]=\left(\frac{\mu_X}{\mu_Y}\right)^2\left(\frac{\operatorname{var}[X]}{\mu_X^2}+\frac{\operatorname{var}[Y]}{\mu_Y^2}\right)$$

Example

Descriptive Statistics for Y

Mean	6.236E+000	Skewness	1.625E+000
Median	5.550E+000	Kurtosis	4.888E+000
Standard deviation	3.173E+000	Standard error	1.003E-002
Variance	1.006E+001	Geometric mean	5.556E+000
Minimum	6.924E-001	Geometric SD	1.618E+000
Maximum	4.069E+001	N	100000
N positive values	100000		

Analytical Solution

$$E[ab/c] = \frac{10 \times 6}{10} + \frac{10 \times 6}{10^3} \times 2^2 = 6.24$$

$$Var[ab] = 10^2 \times 2^2 + 6^2 \times 3^2 + 3 \times 2 = 730$$

$$Var[ab/c] = \left(\frac{6 \times 10}{10}\right)^2 \times \left(\frac{730}{60^2} + \frac{2^2}{10^2}\right) = 8.74$$

Analytical Solutions for Functions

Function	Mean	Variance
$x = e^y$	$\mu_x = e^{\mu_y + \frac{\sigma_y^2}{2}}$	$\sigma_x^2 = e^{2\mu_y + 2\sigma_y^2} - e^{2\mu_y + \sigma_y^2}$
† $x = \ln(y)$	$\mu_x = \ln(\mu_y) - \frac{\sigma_y^2}{2\mu_y^2}$	$\sigma_x^2 = \frac{\sigma_y^2}{\mu_y^2}$
$x = \ln(y)$	$\mu_x = \ln(\mu_y) - \frac{1}{2}\left(\frac{\sigma_y^2}{\mu_y^2} + 1\right)$	$\sigma_x^2 = \ln\left(\frac{\sigma_y^2}{\mu_y^2} + 1\right)$

† **based on assumption of normality**

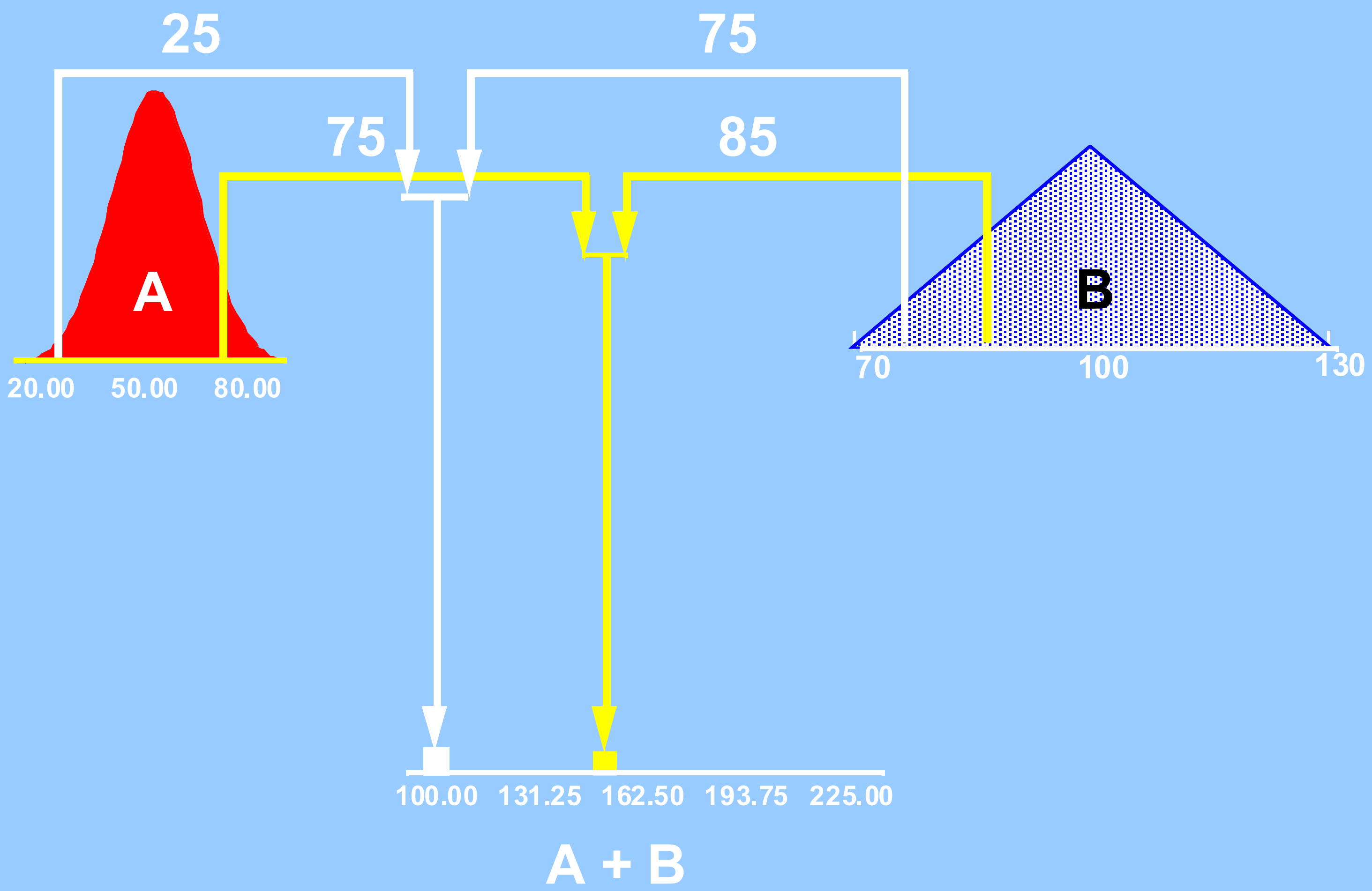
Sandia National Laboratories

Analytical Solutions for Functions

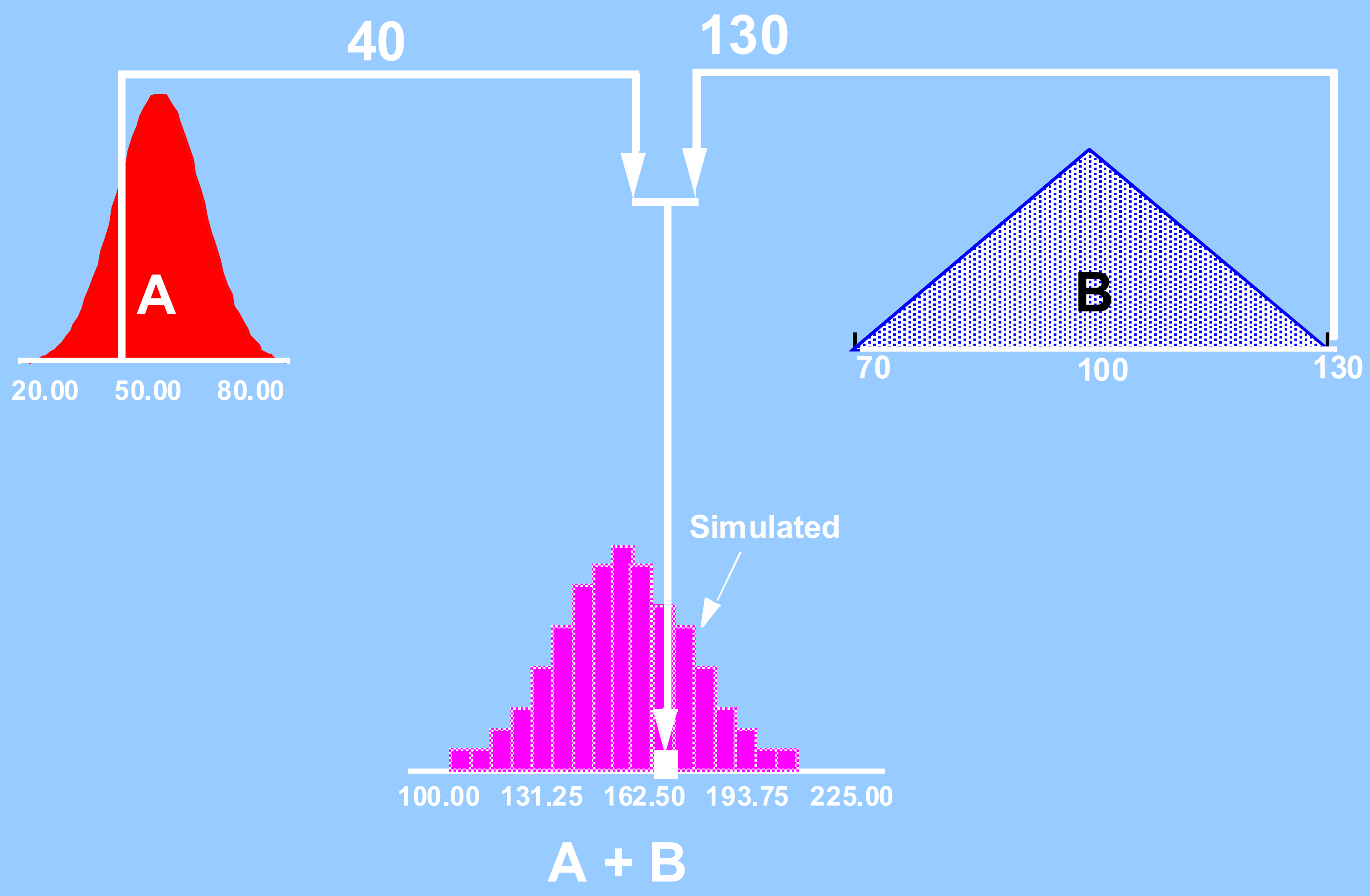
Function	Mean	Variance
$x = \sin(y)$	$\mu_x = \sin(\mu_y)\left(1 - \frac{\sigma_y^2}{2}\right)$	$\sigma_x^2 = \sigma_y^2 \cos(\mu_y)^2$
$x = \cos(y)$	$\mu_x = \cos(\mu_y)\left(1 - \frac{\sigma_y^2}{2}\right)$	$\sigma_x^2 = \sigma_y^2 \sin(\mu_y)^2$
$x = \sqrt{y}$	$\mu_x = \sqrt{\mu_y} - \frac{1}{8}\mu_y^2 \sigma_y$	$\sigma_x^2 = \frac{\sigma_y^2}{4\mu_y}$
$x = c^y$	$\mu_x = c^{\mu_y} + \frac{1}{2}\sigma_y^2 \ln(\mu_y)^2 c^{\mu_y}$	$\sigma_x^2 = \sigma_y^2 c^{2\mu_y} \ln(\mu_y)^2$

Sandia National Laboratories

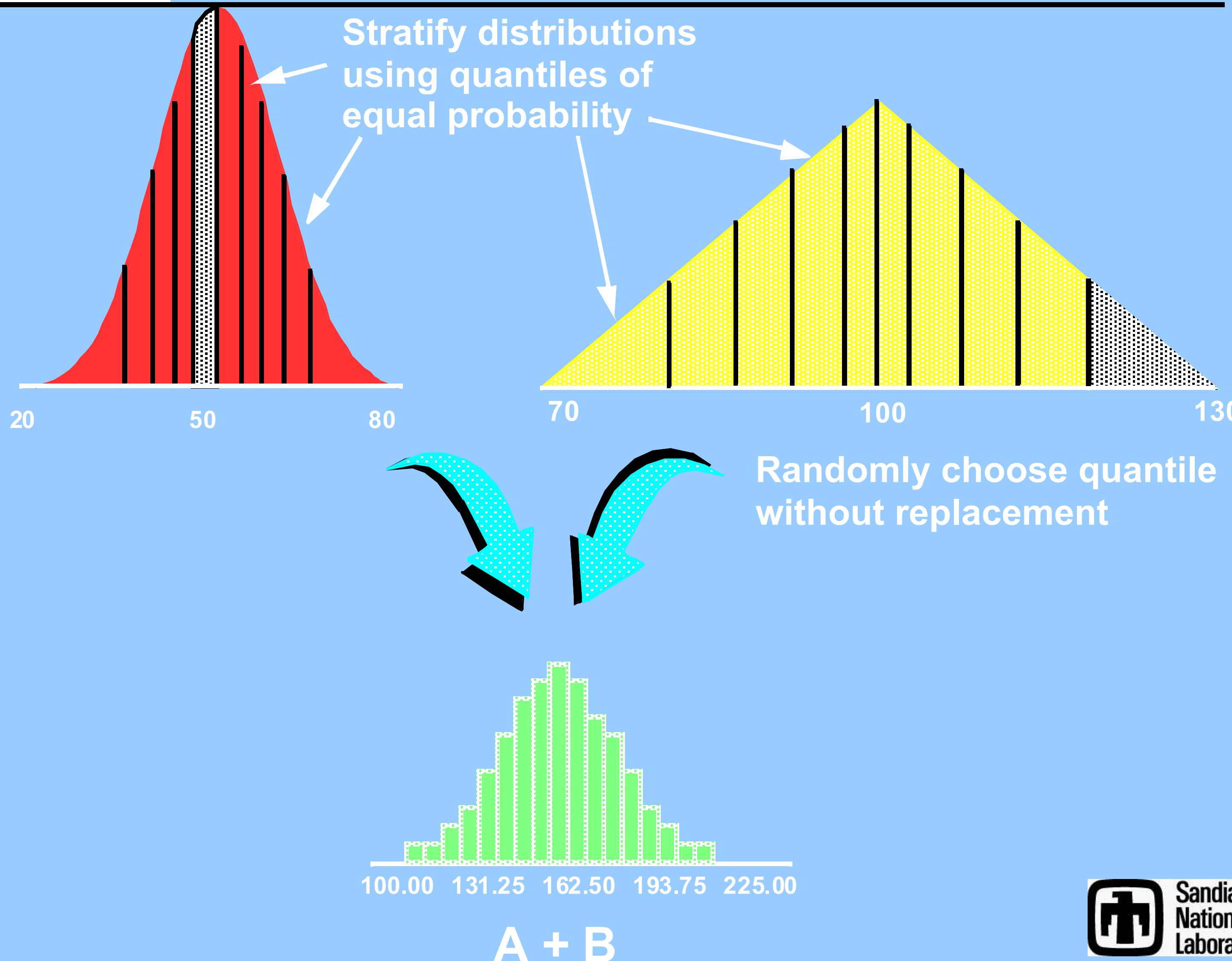
Monte Carlo Sampling



Monte Carlo Sampling



Latin Hypercube Sampling



Sandia National Laboratories

Latin Hypercube Sampling

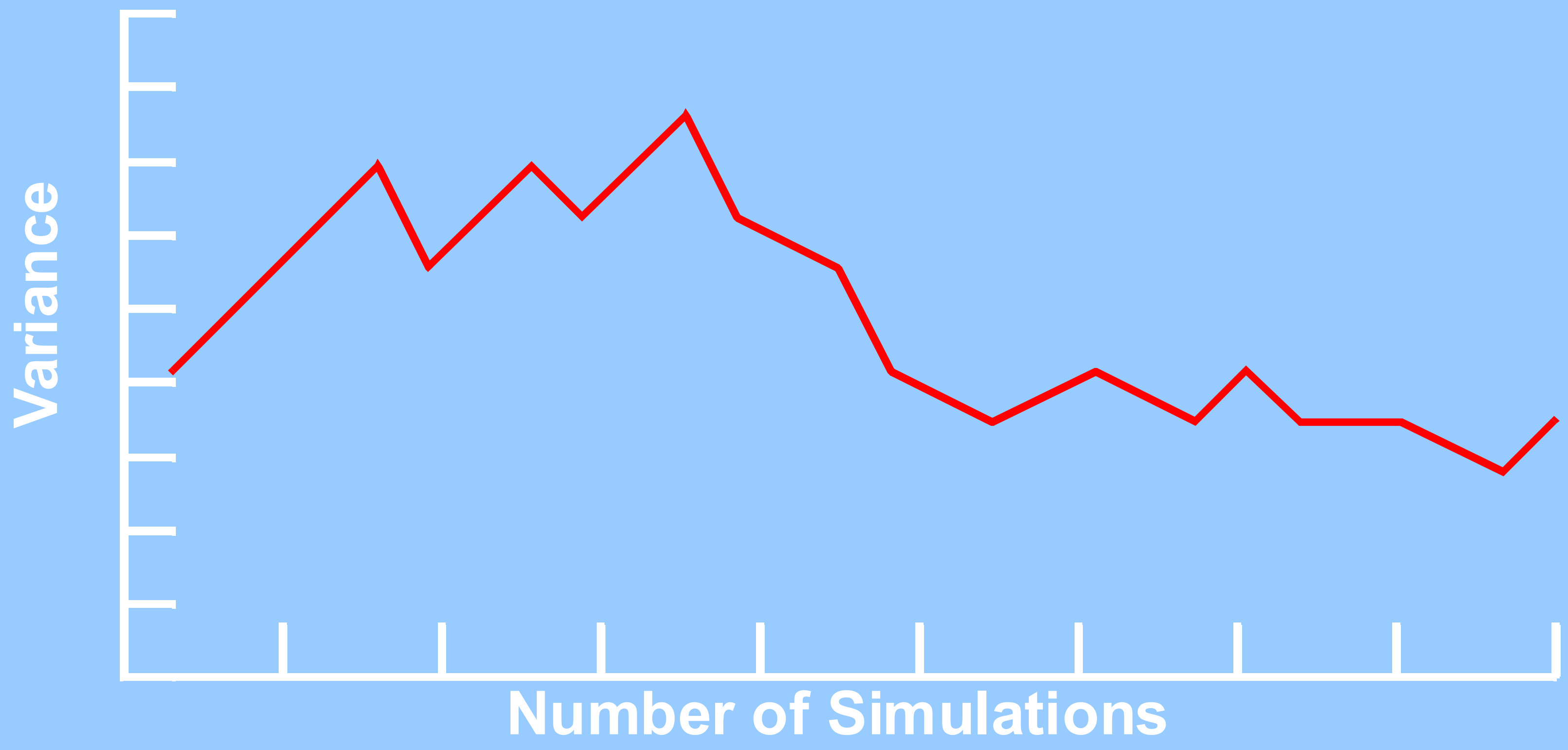
Midpoint sampling

- A variation of LHS sampling uses the midpoint of the interval
- This method can introduce bias into the results
- More importantly, with periodic functions like Sine the method can fail badly if the period of the function falls on the midpoints of the intervals

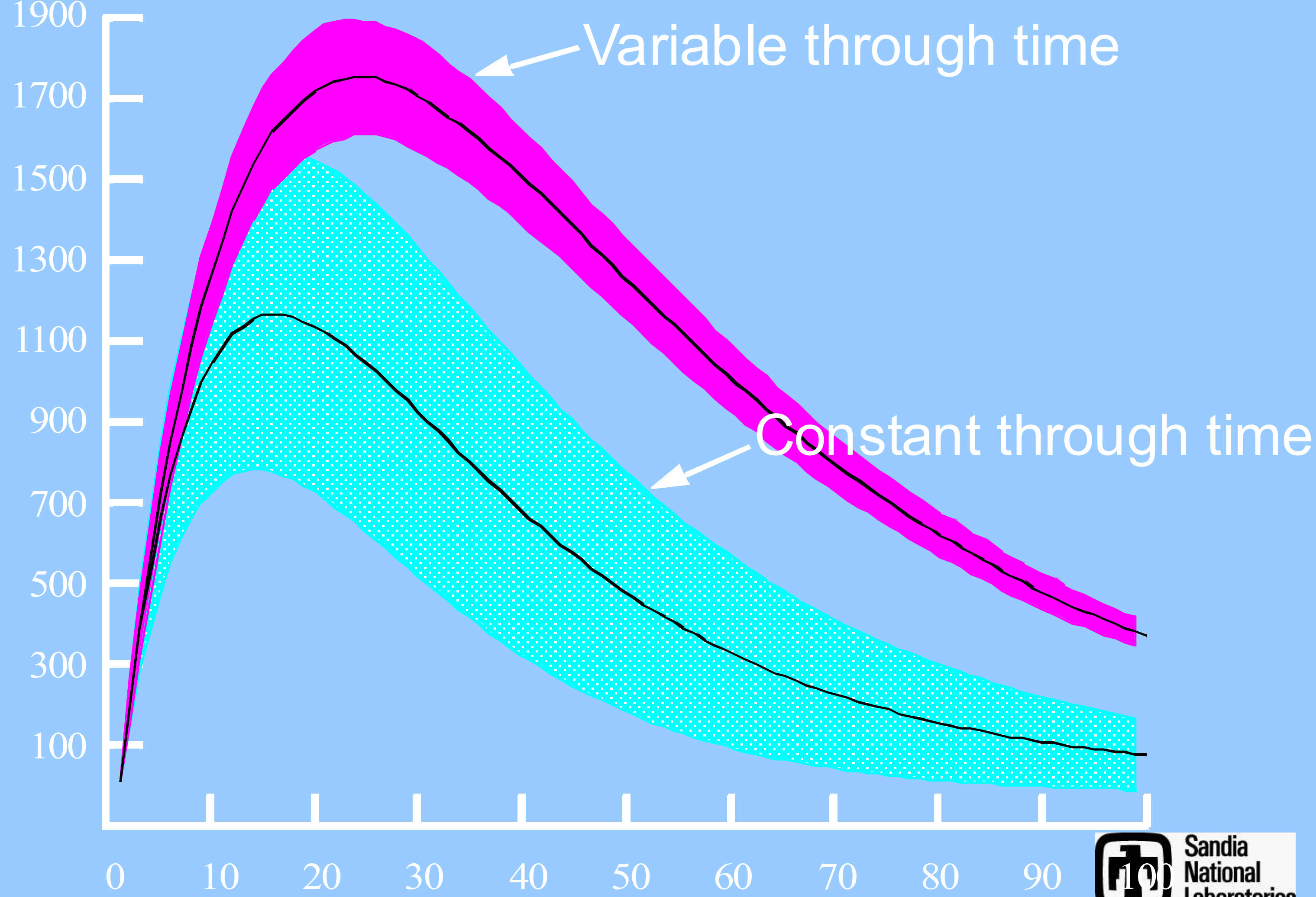
Correlated Sampling

- Correlations can be introduced into LHS sampled data by re-ordering the data
- An algorithm exists which uses the inverse of the correlation matrix to modify the data
- Rank order correlations are created within the data
- The marginal distributions can be of any shape – they don’t have to form a true multivariate distribution

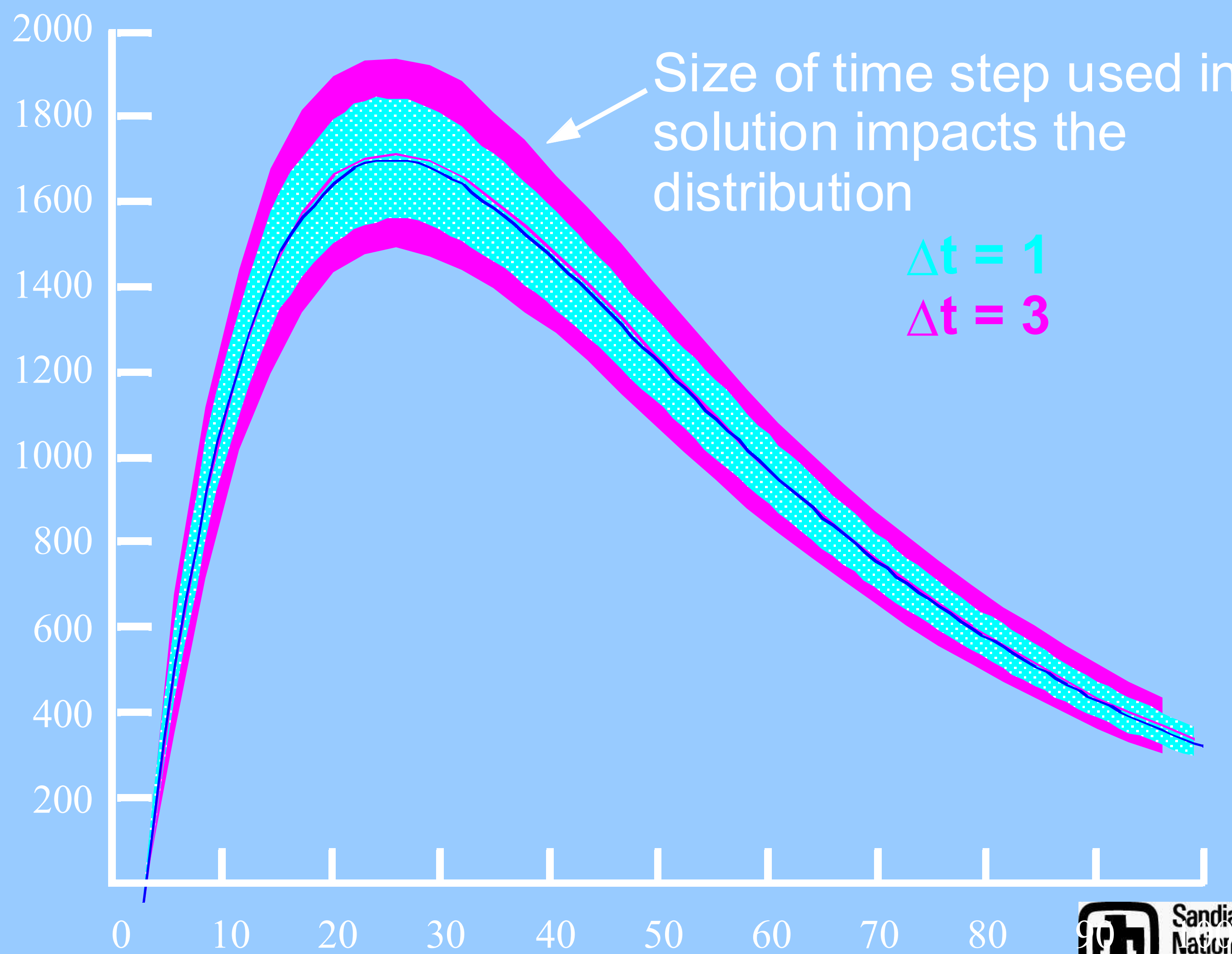
How Many Simulations?



Are Parameters Constant Through Time?



Are Parameters Constant Through Time?



Case Study

The model equations

$$C_{air} = S_{release} \frac{\chi}{Q}$$

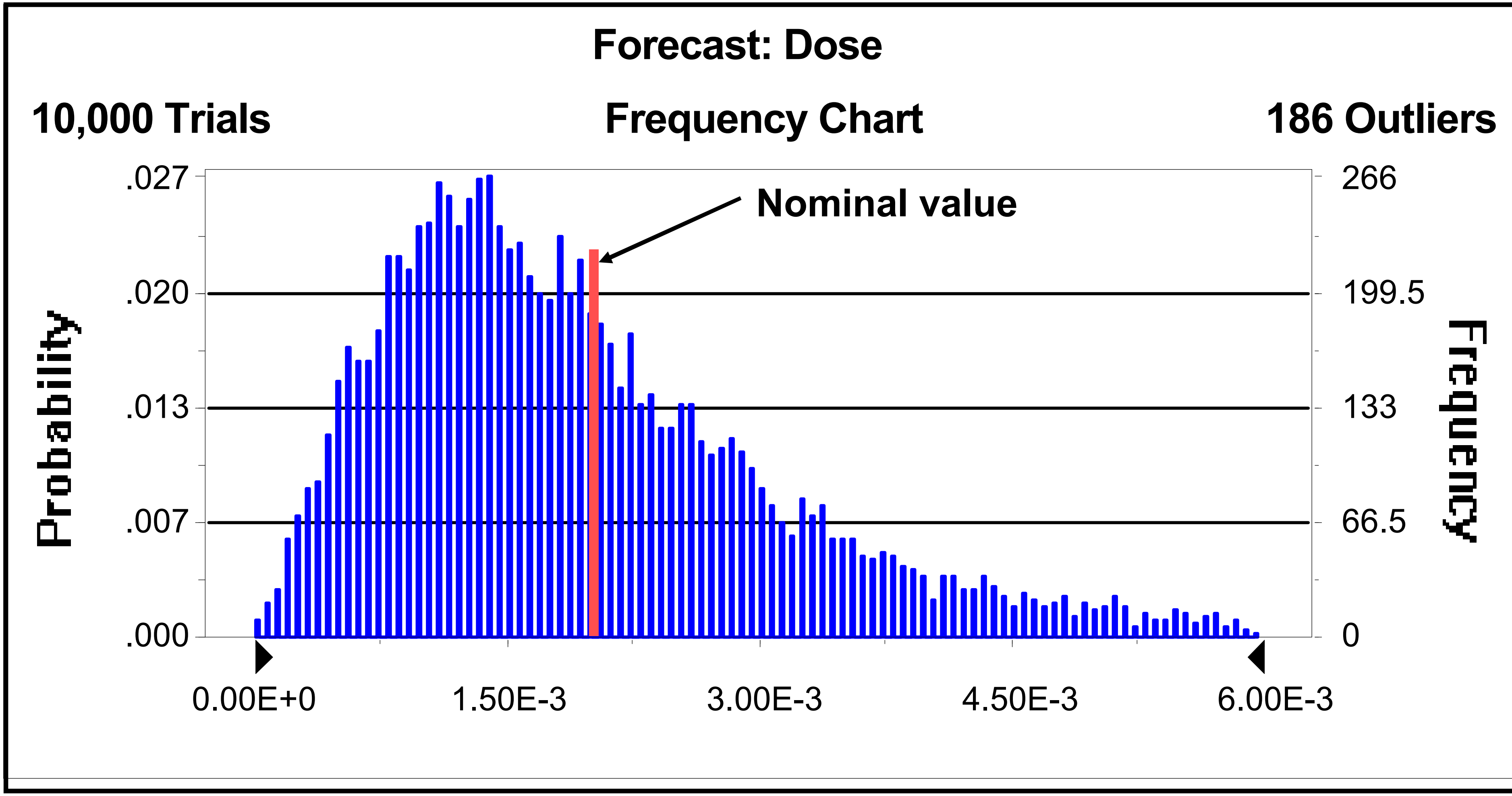
$$D_{thyroid} = \frac{C_{air} V_{dep} I_{cow} T_{forage/milk} I_{child} D_{child}}{\left(K_{weathering} + K_{decay}\right) D_{forage}}$$

Case Study

Uncertainties

Factor	Distribution	Parameters	Units
V_{dep}	Constant	1	m/s
I_{cow}	Normal	Mean = 8, SD = 0.8	kg/d
T_{forage}	Lognormal	GM=0.01, GSD=1.31	d/l
I_{child}	Triangular	Min=0, Mode=0.6, Max=1.2	l/d
$K_{weathering}$	Lognormal	GM=7.65E-2, GSD=1.2	d^{-1}
D_{forage}	Triangular	Min=0.1, Mode=0.3, Max=0.6	kg/m^3
K_{decay}	Constant	0.086	d^{-1}
D_{child}	Constant	3.6E-6	Sv/Bq
$S_{release}$	Constant	5000	Bq/day
X	Constant	88	Bq/m^3
Q	Constant	1000	Bq/day

Case Study

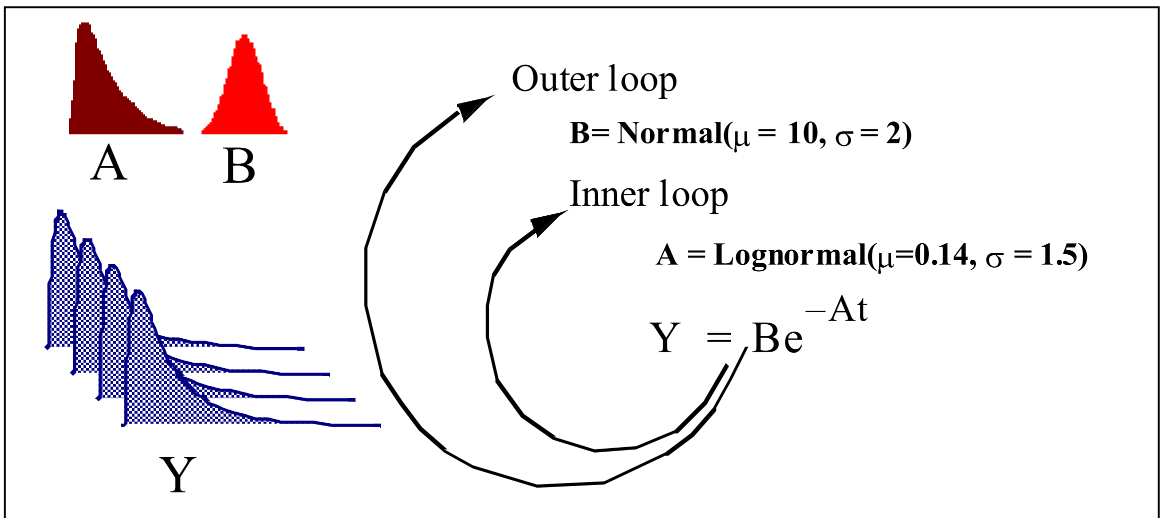


Forecast: Dose
10,000 Trials
Frequency Chart
186 Outliers
Nominal value
Probability
Frequency
.027
.020
.013
.007
.000
266
199.5
133
66.5
0
0.00E+0
1.50E-3
3.00E-3
4.50E-3
6.00E-3

Case Study

Statistics (Sv/y)	
Nominal	2.1E-3
Mean	1.62E-3
SD	1.08E-3
GM	1.31E-3
GSD	2.00
2.5 percentile	2.73E-4
97.5 percentile	4.50E-3

Nested Sampling of Aleatory and Epistemic Uncertainties



Sampling Design- WIPP PA

- Outer loop steps through 100 vectors of input parameters per replicate
  - 56 parameters are sampled for each 2004 PABC vector
- Inner loop simulates 10,000 futures for each vector
- Hence, 1,000,000 simulations of 10,000 year futures are simulated per replicate
- Three replicates are normally run for a PA

Sandia National Laboratories

Choosing Distributions

Objective methods:

- **Traditional statistical methods for fitting distributions to data**
- **Analytical methods for determining distribution**
  - **Example: Counting error is assumed to be Poisson because radioactive decay of atoms is random in time**

Subjective methods:

- **Expert elicitation, meta-analysis of data, other statistical tools for limited data**

Cautions When Using Subjective Methods

Subjective methods are frequently used to assign distributions to parameters in risk assessment models.

It is generally better to assign subjective uncertainty than to pretend that a variable has no uncertainty

How Well Would You Do As An Expert If Asked To Estimate Probabilities? Case I

"The prevalence of breast cancer is 1% (in a specified population). The probability that a mammogram is positive if a woman has breast cancer is 79%, and 9.6% if she does not. What is the probability that a woman who tests positive actually has breast cancer?______%" (Gigerenzer 1994).

Case I: The Answer

The correct answer of about 8% of the women testing positive really have cancer might be more easily found by using counts of people

Out of 1000 women:

Expected to have cancer: $1000 \times 0.01 = 10$

Expected to test positive: $10 * 0.79 \approx 8$

False positives: $990 \times 0.096 = 95$

$$\frac{\text{8 positives with cancer}}{\text{95 positive without cancer + 8 with cancer}} = 7.7\ \%$$

Sandia National Laboratories

Estimating Probabilities: Case II

- "Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations. Which of the these two alternatives is more probable?
  (a) Linda is a bank teller
  (b) Linda is a bank teller and active in the feminist movement."

Sandia National Laboratories

Case II: The Answer

In previous experiments, 80% to 90% of the subjects of the experiment answered (b), regardless of the fact that (b) is the product of the probability of (a) and the probability that Linda is active in the feminist movement and thus cannot be larger than (a) (Gigerenzer 1994).

What Can Bias Estimates of Uncertainty or Risk?

- Base rate bias: Illustrated by Case I.
- Representativeness bias: Unusual patterns in data, such as HHHTTT in a coin toss experiment are thought to have lower probability than “more random” patterns, such as HHTHTT.

Sandia National Laboratories

What Can Bias Estimates of Uncertainty or Risk?

- Motivational bias: People tend to weight their estimates of frequency of events by their perceived importance.

- Availability bias: People tend to weight their estimates of frequency of events by the ease with which they can recall previous occurrences of the event or with which they can imagine the event to occur.

What Can Bias Estimates of Uncertainty or Risk?

- Anchoring and adjustment bias: If a "known" frequency of an event (the "anchor point") is presented when eliciting estimates of frequencies of similar events, people tend to use those estimates to help scale their responses(Kahneman and Tversky 1973).

- Conjunction fallacy: Illustrative above in Case II, causes people judge alternatives by seeking the greatest similarity.

What Can Bias Estimates of Uncertainty or Risk?

- Overconfidence bias: Researchers have claimed to show that scientists generally tend to underestimate the true uncertainties when reporting their results (e.g.,Schlyakhter 1994; Henrion and Fischhoff 1986).

10 Step Elicitation Protocol

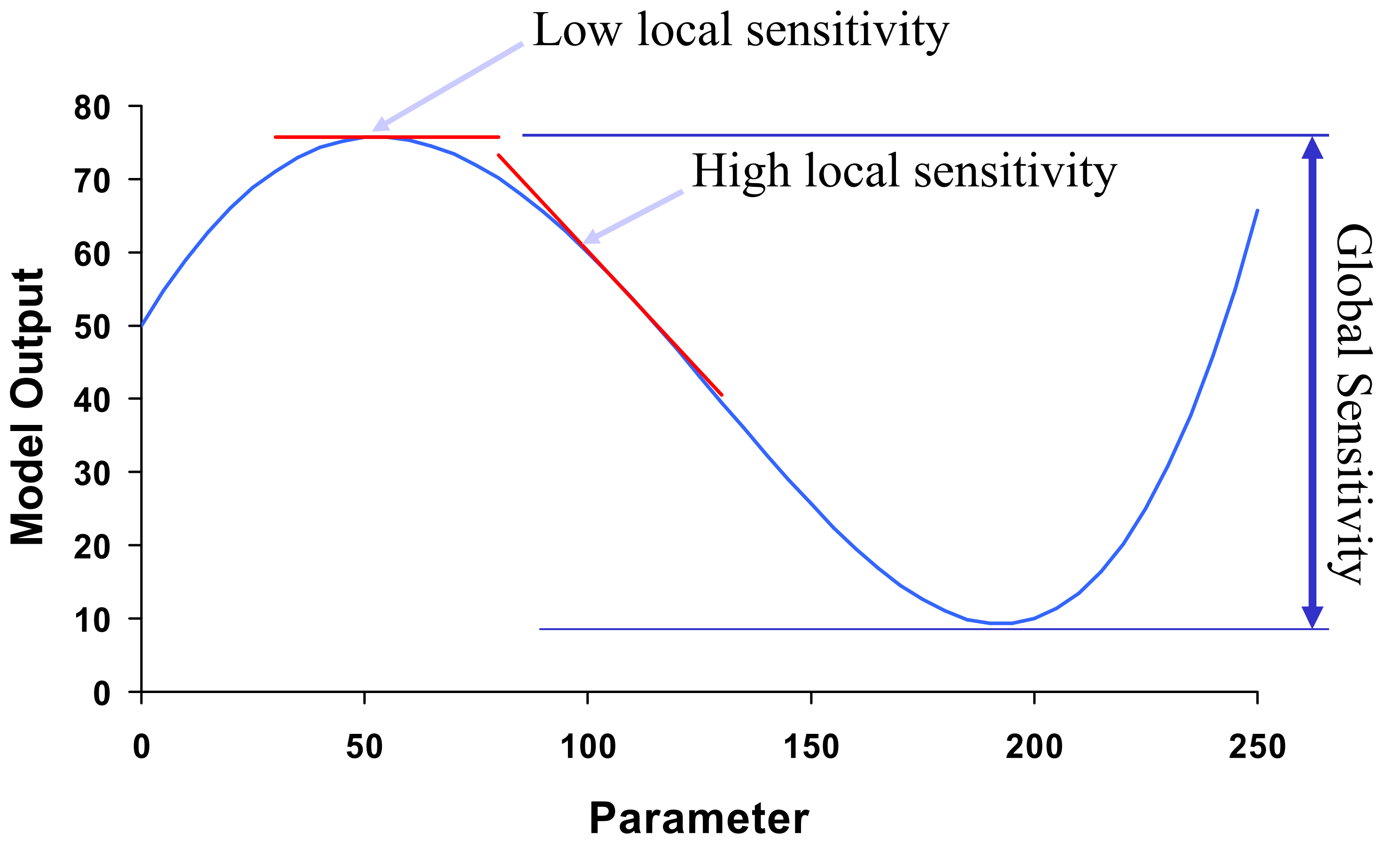
1. selection of issues and associated parameters;
2. selection of experts;
3. preparation of background information;
4. training of experts in the elicitation method;
5. presentation of issues, questions, and background information to experts;

10 Step Elicitation Protocol

6. analysis of issues, questions and background information by experts;
7. discussion of analyses of issues and background information by experts;
8. elicitations of judgment from the experts;
9. consolidation (and possibly aggregation) of the elicitation results; and
10. review and communication of the elicitations.

Sensitivity Analysis

Sensitivity Analysis



Local Sensitivity Analysis

One-at-a-time (OAT) methods

The partial derivatives at the nominal values:

$$S_i' = \frac{\partial G(x_1, x_2, \ldots, x_3)}{\partial x_i} = \frac{\partial Y}{\partial x_i}$$

Sometime normalization is done to get comparable magnitudes

$$S_i = \frac{\partial y / y}{\partial x_i / x_i} = \frac{x_i}{y} \frac{\partial y}{\partial x_i}$$

Local Sensitivity Analysis

One-at-a-time (OAT) methods

The partial derivatives at the nominal values are estimated using 2 points:

$$\frac{x_{i,2} - x_{i,1}}{x_{i,1}} = \frac{(x_{i,1} + x_{i,1}\Delta) - x_{i,1}}{x_{i,1}} = \Delta x$$

$$S_i = \frac{1}{\Delta x}\frac{\Delta y}{y}$$

A sensitivity index can be used to normalize the estimates

Sandia National Laboratories

Local Sensitivity Analysis

One-at-a-time (OAT) methods

Sensitivity scores adjust the sensitivity index by multiplying them by the CV or relative range of the response variable.

- Multiply S_i by $\frac{\sigma}{\mu}$ or $\frac{x_{i,\max} - x_{i,\min}}{\overline{x}}$

Local Sensitivity Analysis

Interactions

Full-factorial sampling design

- If there are *p* parameters each of which is assigned *l* levels then simulation must be done (i.e. all possible combinations)
- *l* is usually small (2 or 3) and values are set near the point of interest (nominal value)

Analysis of Variance can be used to analyze resulting data

Global Sensitivity Analysis

Global Sensitivity Analysis

- Analytical methods
- Regression-based methods
- Analysis of Variance
- Correlation Ratio
- Fourier Amplitude Sensitivity Test

Sensitivity Analysis

Global Sensitivity of a Sum

Remember the propagation formulas?

The variance of a sum = sum of variances (if independent)

$$G_i = \frac{Var[x_i]}{Var\left[\sum_{i=1}^{n} x_i\right]} = \frac{Var[x_i]}{\sum_{i=1}^{n} Var[x_i]}$$

Sensitivity Analysis

Global Sensitivity of a Product

Propagation formula is a bit more complex

For 3 independent variables

$$G_i = \frac{\sigma_i^2}{\mu_1^2\mu_2^2\sigma_3^2 + \mu_1^2\mu_3^2\sigma_2^2 + \mu_2^2\mu_3^2\sigma_1^2 + \mu_1^2\sigma_2^2\sigma_3^2 + \mu_2^2\sigma_1^2\sigma_1^2 + \mu_3^2\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_2^2\sigma_3^2}$$

Regression Methods

Regression methods are popular because

- **They are familiar**

- **They don't require any specialized sampling design, so they can use the results generated for uncertainty analysis**

- **There are methods for identifying and handling non-linearities**

Linear Regression

Linear Regression can be used to relate the values of one variable (the dependent variable) to another (the independent variable)

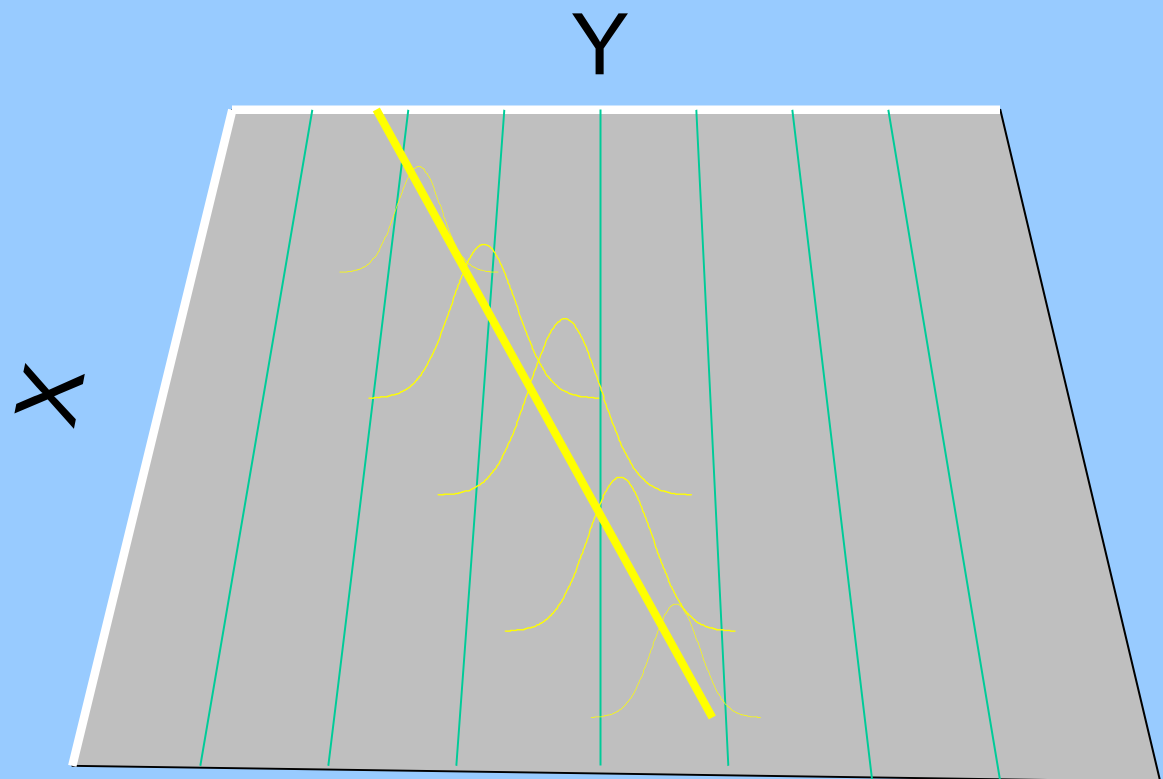
- For example, the concentration of a radionuclide may be a function of soil depth
- Note that the reverse is not true – soil depth is not dependent on radionuclide concentration

Linear Regression

Assumption of Simple Linear Regression

- **The relationship between X and Y is modeled by a straight line**
- **The independent variable (X) is measured without error**
- **For each value of X the Y value is sampled from a normal distribution**
- **The normal Y distributions are equal to each other**

Y-Values Normally Distributed



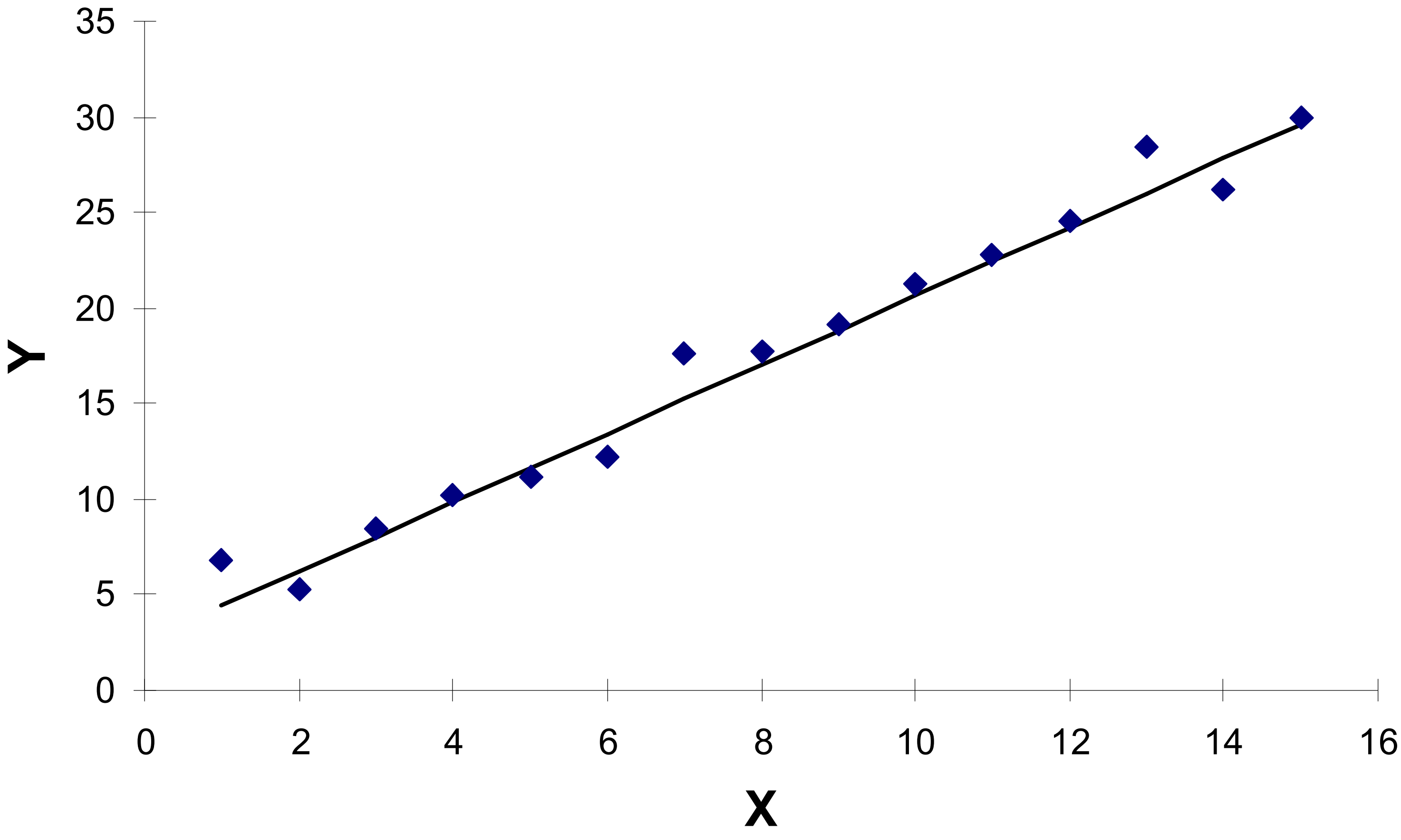
Linear Regression

Assumption of Simple Linear Regression

- The relationship between X and Y is modeled by a straight line
- The independent variable (X) is measured without error
- For each value of X the Y value is sampled from a normal distribution
- The normal Y distributions are equal to each other
- **The Y values are independent of each other**
- **The errors in Y are additive**

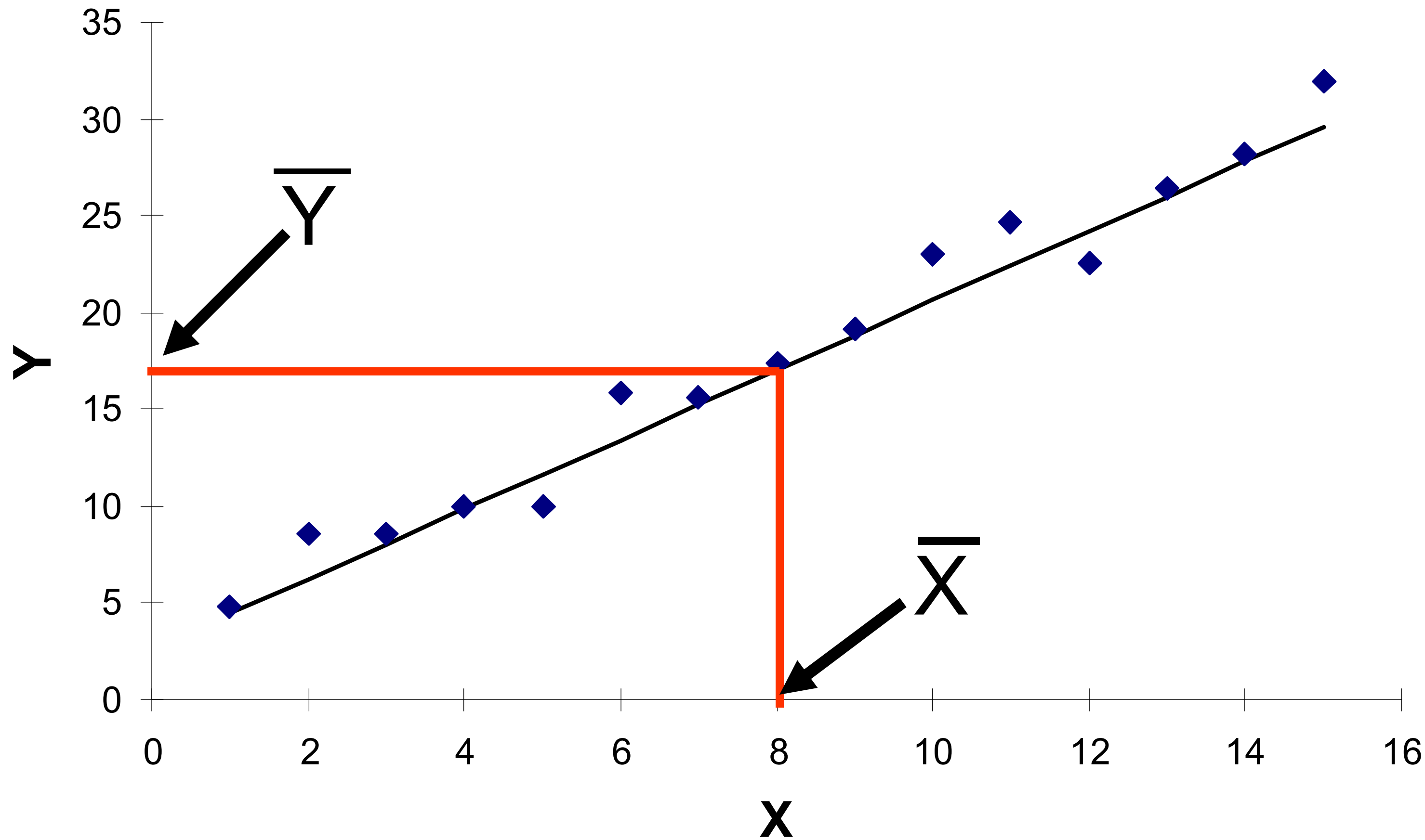
Sandia National Laboratories

Simple Linear Regression



Sandia National Laboratories

Simple Linear Regression



Simple Linear Regression

$$Y = aX + b + \varepsilon$$

$$\text{slope} = a = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\text{intercept} = b = \bar{Y} - a\bar{X}$$

Sandia National Laboratories

Simple Linear Regression

Is the relationship significant?

- **Compute the F-value and test it**
  - **Regression and analysis of variance use the same "model"**
- **Compute the correlation coefficient, r, and test it for significance**
- **Compute the coefficient of determination, r^2**
  - **Shows proportion of variation in Y accounted for by X**

Sandia National Laboratories

Simple Linear Regression

Is the relationship significant?

- **Compute the F-value and test it**
  - **Regression and AOV use the same "model"**
- **Compute the correlation coefficient, r, and test that**
- **Compute the coefficient of determination, r^2**
  - **Shows proportion of variation in Y accounted for by X**

Statistical Significance

Simple Linear Regression

Is the relationship significant?

- Compute the F-value and test it
  - Regression and AOV use the same "model"
- Compute the correlation coefficient, r, and test that

- **Compute the coefficient of determination, r^2**
  - **Shows proportion of variation in Y accounted for by X**

Practical Significance

Simple Linear Regression

$$F = \frac{\text{regression MS}}{\text{residual MS}} = \frac{\text{Sum of Squares Regression/DF Regression}}{\text{Sum of Squares Residual/DF Residual}}$$

$$F = \frac{\sum(\hat{Y}_i - \overline{Y})^2 / 1}{\left(\sum(Y_i - \overline{Y})^2 - \sum(\hat{Y}_i - \overline{Y})^2\right)/(n-2)}$$

Where the predicted Y is

$$\hat{Y}_i = a + bX_i$$

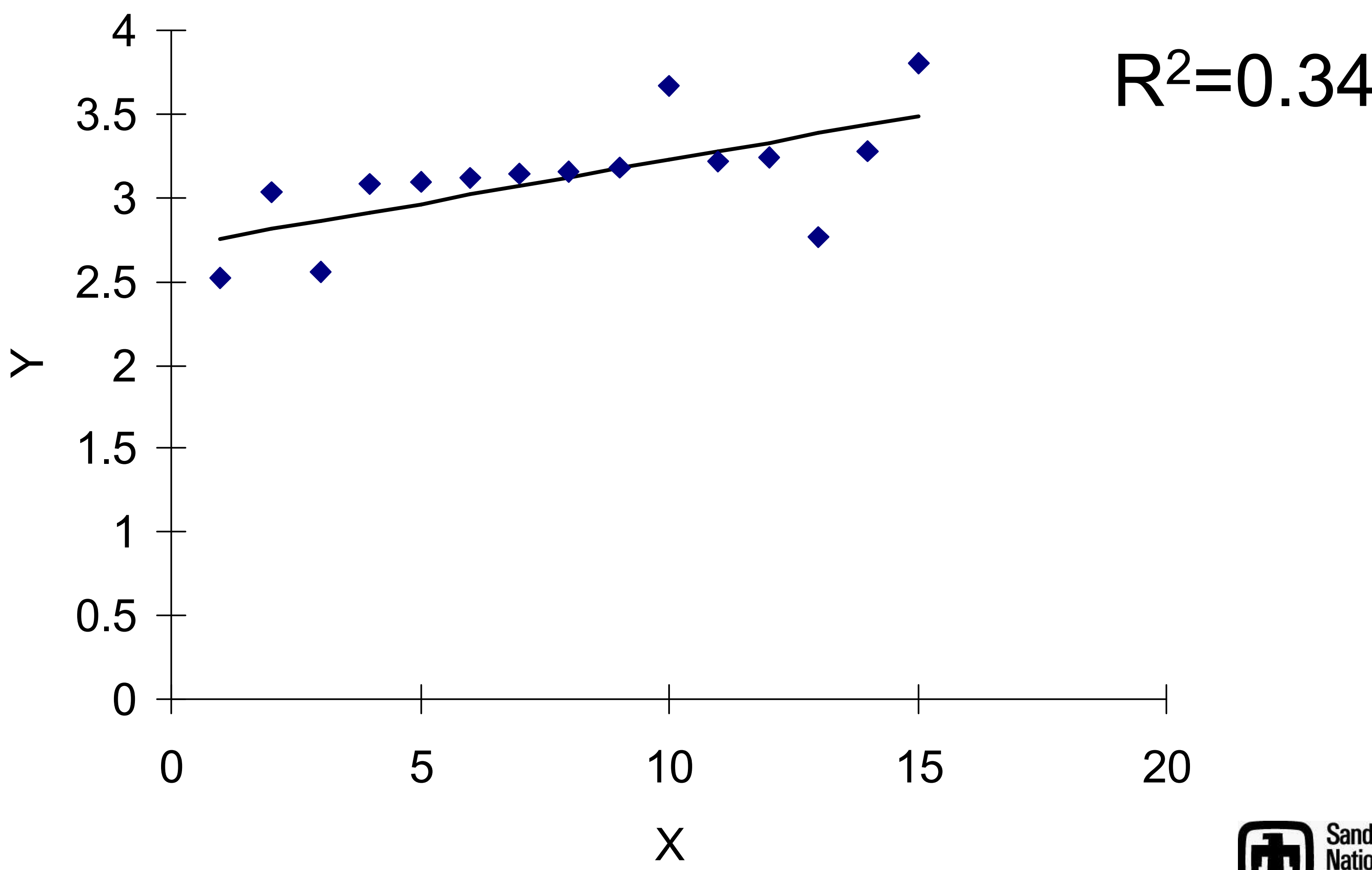
The Coefficient of Determination

The coefficient of determination gives the proportion of the variance in the data accounted for by the linear model

$$r^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}$$

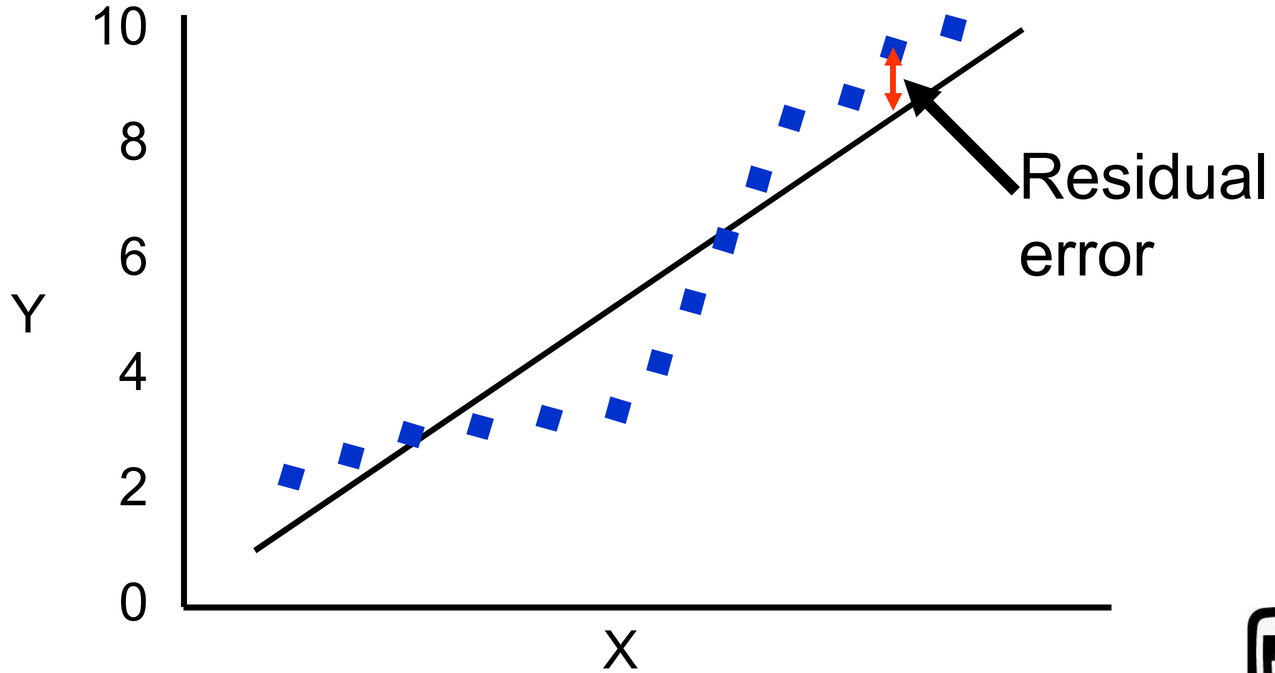
Sandia National Laboratories

Significance Falls as the Slope Approaches 0



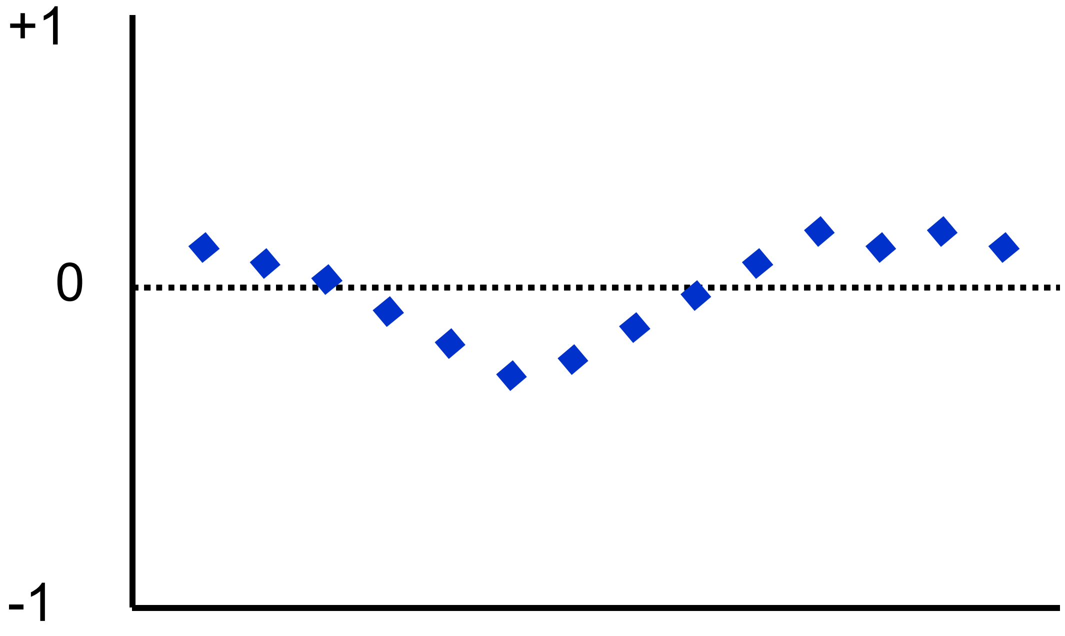
Check the Residuals

Patterns in the residuals can indicate that the model is inadequate



Check the Residuals

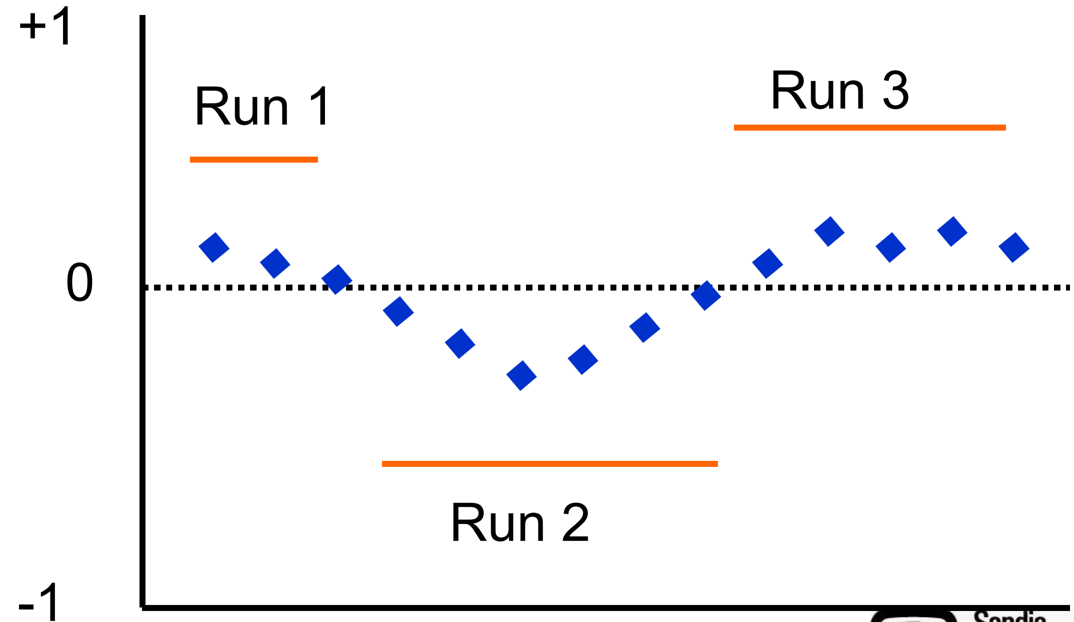
Patterns in the residuals can indicate that the model is inadequate



Runs test

A "Runs test" can be used to see if a pattern of high and low values is significant.

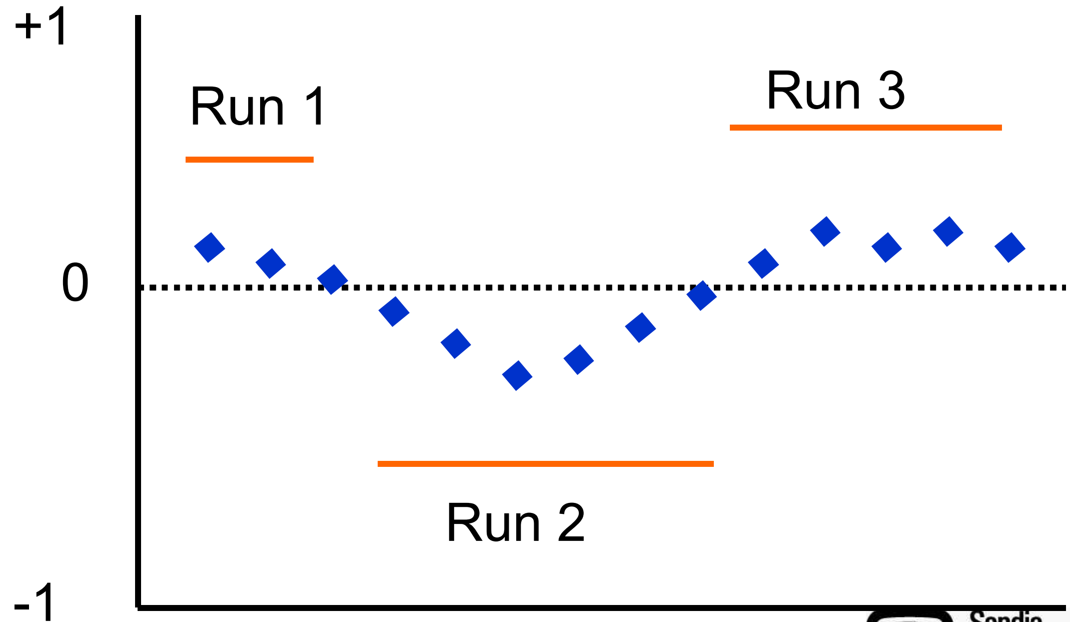
- Error should be random above and below
- Runs indicate linear model is missing dynamics



Runs test

A “Runs test” can be used to see if a pattern of high and low values is significant.

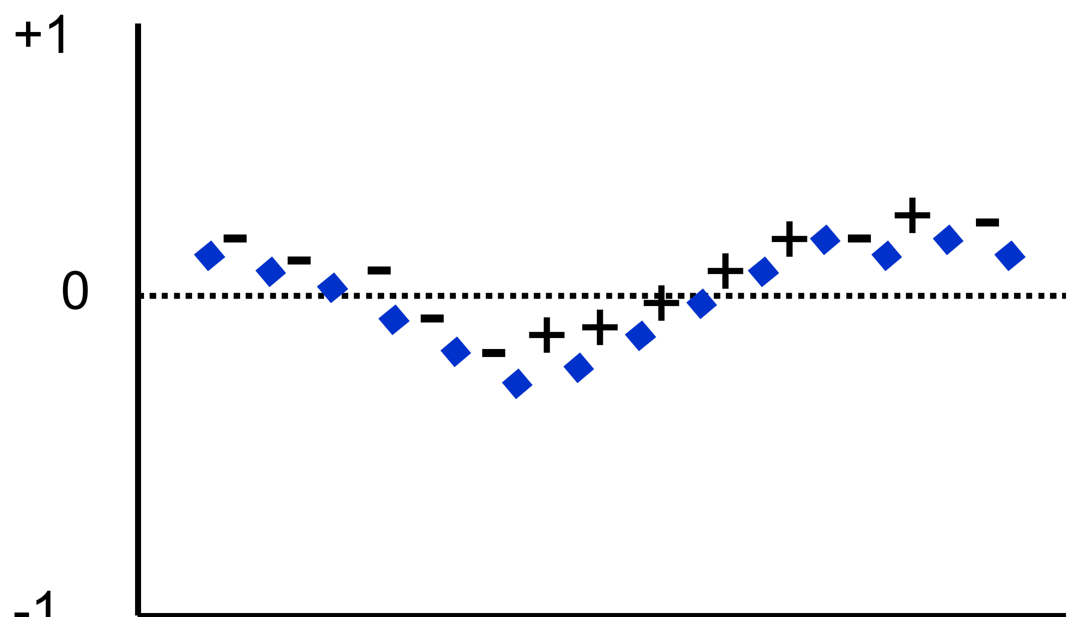
- Count the runs of positive and negative values
- Check against tabulated values for significance



Signs test

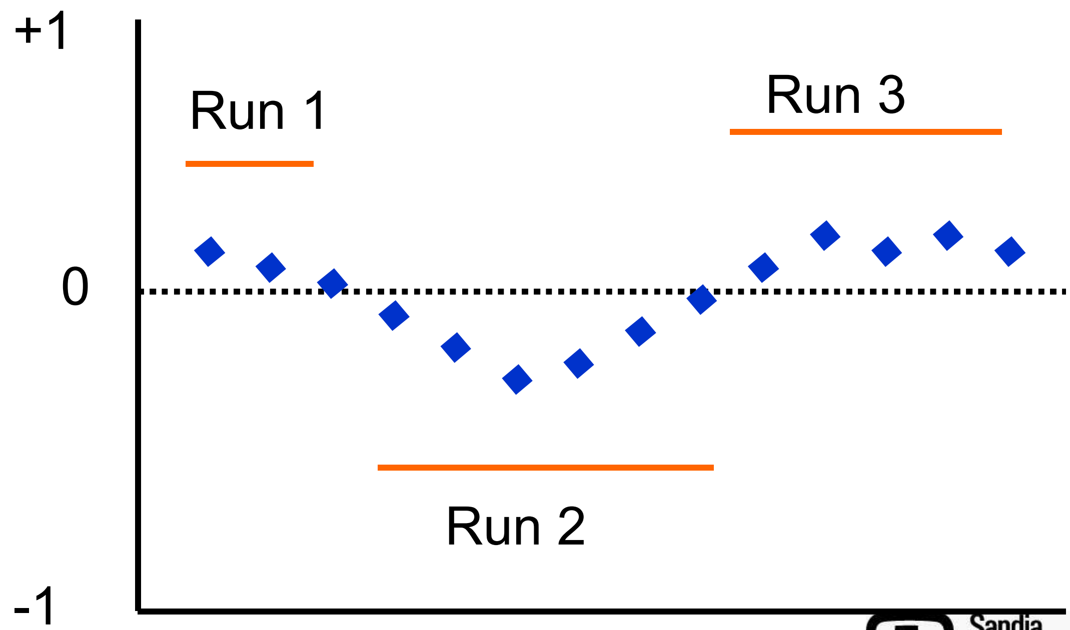
A “Signs test” can be also used to see if a pattern in the residuals is significant.

- Assign + if the change from 1 point to the next is positive, - if negative
- Check counts against tabulated values for significance

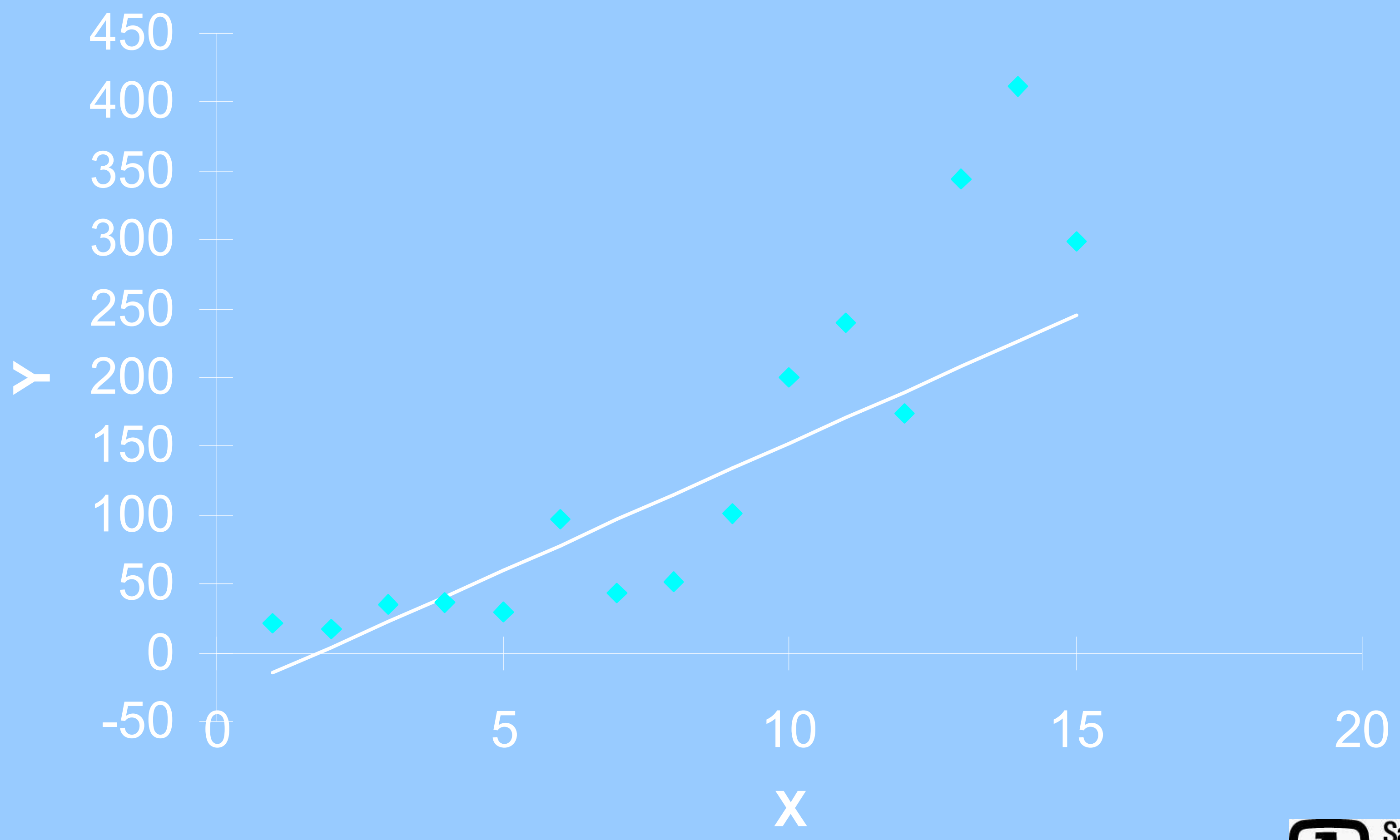


Runs test

Can you guess what distribution is used to figure out the probabilities of runs or signs?



What do you do if the data are not linear?



What do you do if the data are not linear?

Try some transforms

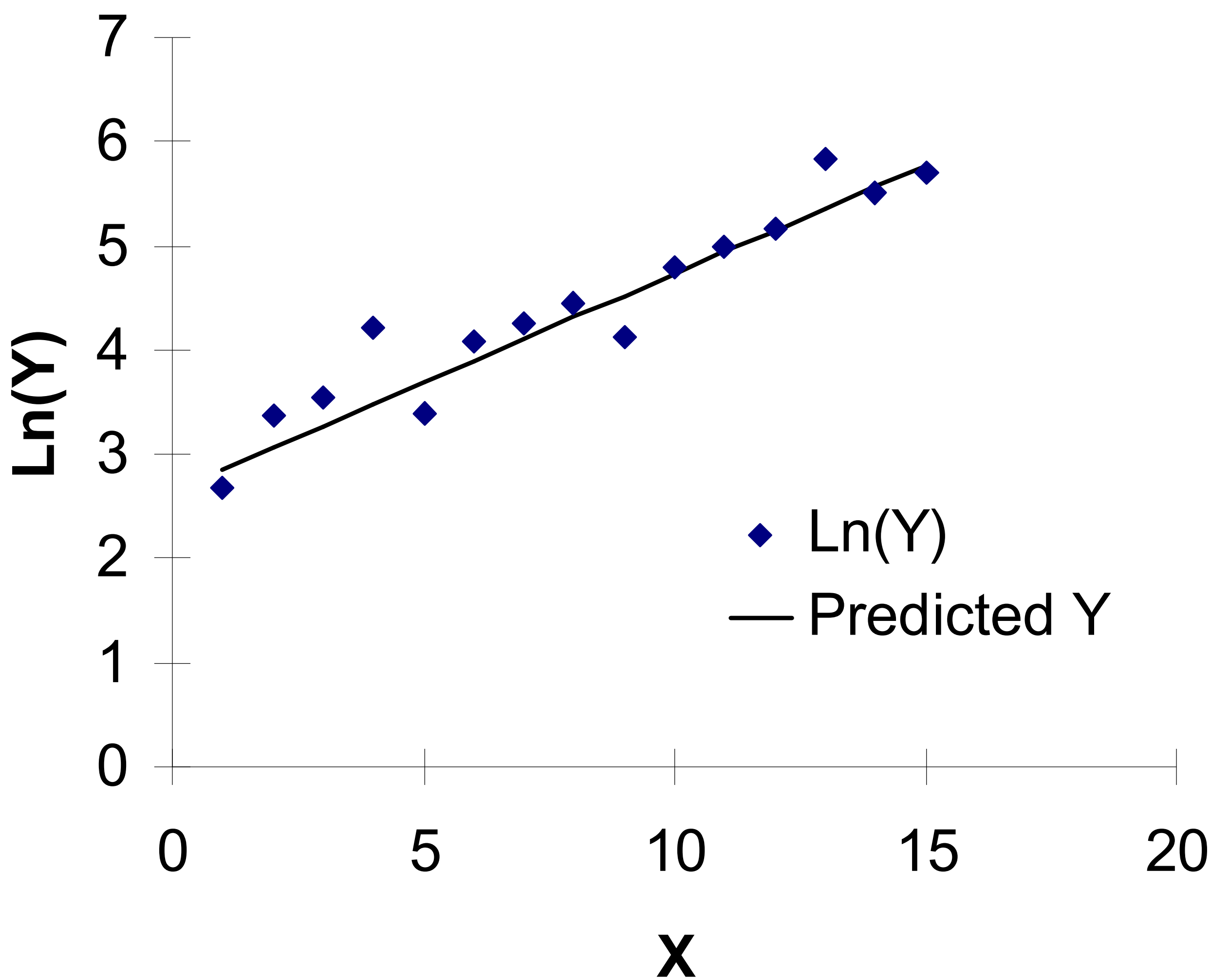
Log(Y)

Log(X)

$\sqrt{Y}$

Etc.

What do you do if the data are not linear?

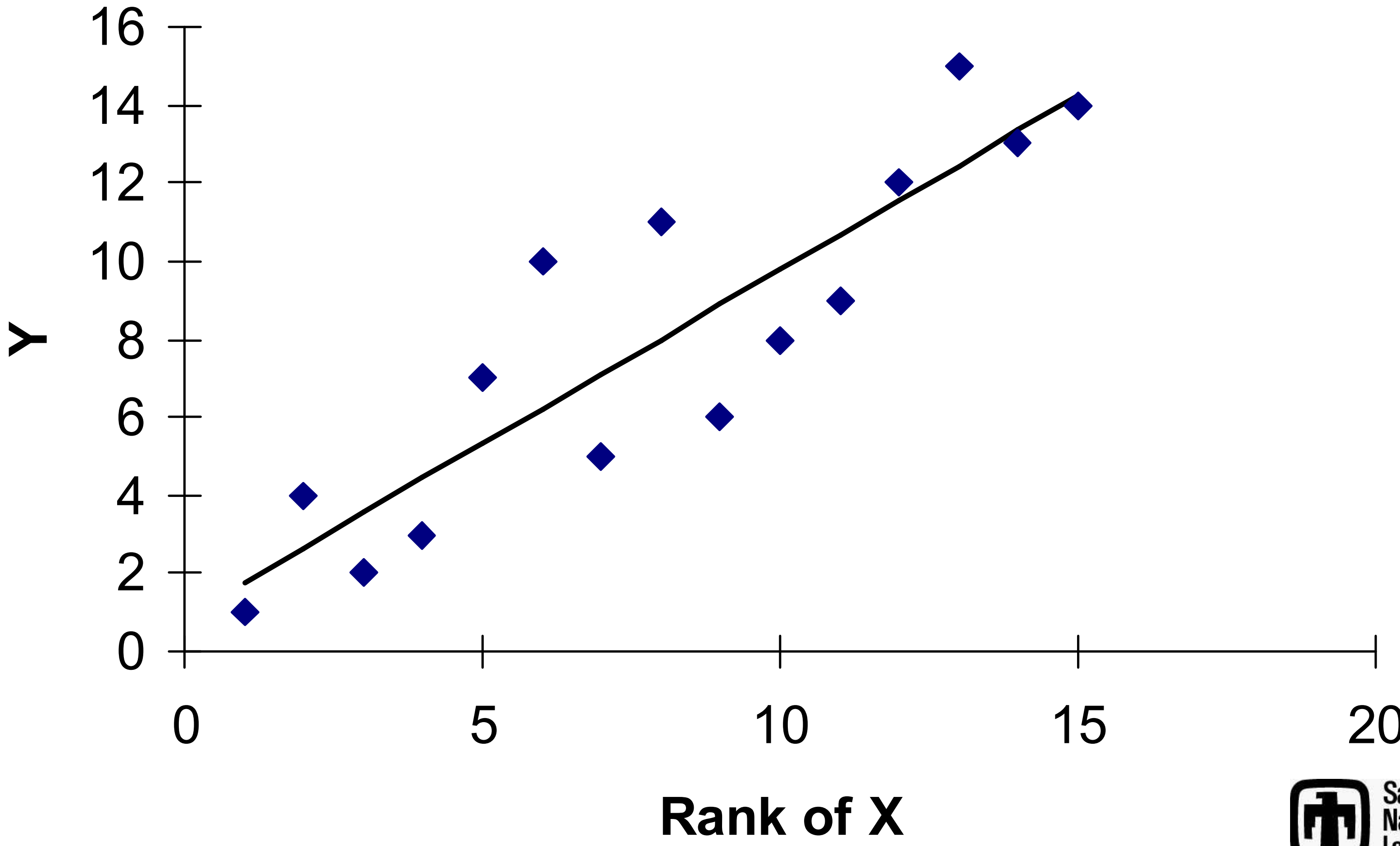


What do you do if the data are not linear?

Rank Regression

- Replace the values with their “rank”
- The rank is the position in the sorted list of values

Ranking Linearizes Non-Linear but Monotonic Data



Sandia National Laboratories

Error in the Independent (X) Variable

Sometimes the independent variable is

- Known to contribute to the variability in Y
- Cannot be measured without error

Regression assumes there is no error in X.

So what is the impact of the error?

Error in the Independent (X) Variable

If the error is "classical" error then the mean is "biased toward 0".

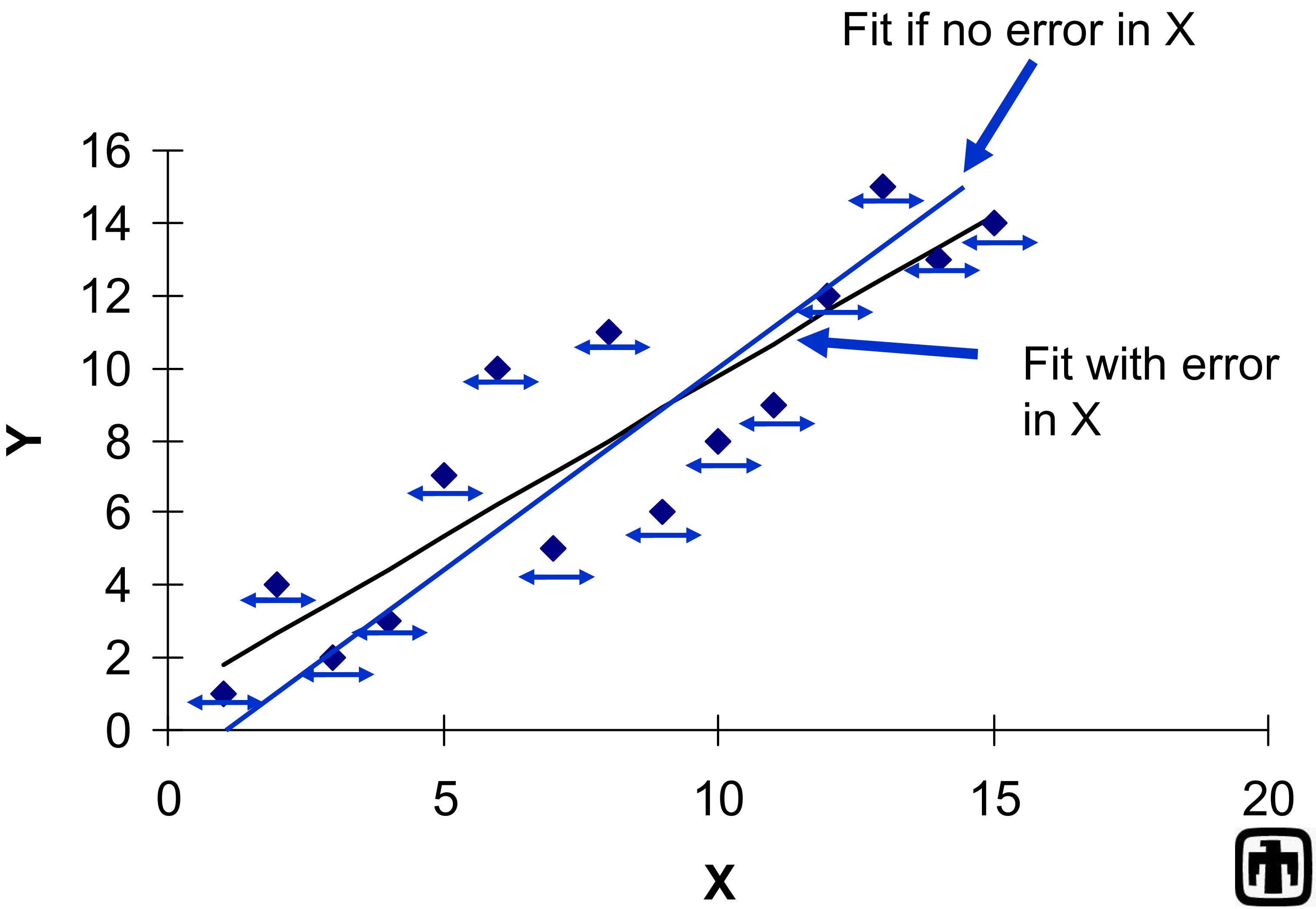
- Classical error occurs if the measurement of X is subject to random error

$$W = X + e$$

- "Biased toward 0" means that the slope is estimated to be more horizontal than the response curve really is.

Sandia National Laboratories

Classical Error in X



Berkson Error

Suppose X cannot be measured directly, but must be inferred from Y. For example:

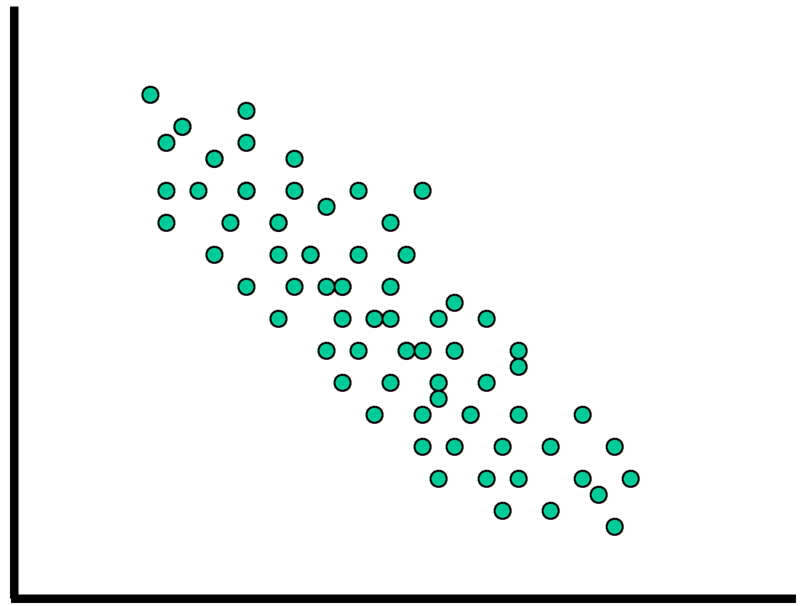
- Badge dose is proportional to exposure but is inferred from optical density
- Temperature is not measured but is inferred from the setting on a heater

In this case the error in X has no impact on the slope of the regression line

Correlation

What if there is no dependency, only a relationship?

Look for correlations in the data



Sandia National Laboratories

Correlation

What if there is no dependency, only a relationship?

- Look for correlations in the data
- Similar assumptions to Regression but the X values can have variability
  - The X values at each Y must be normally distributed

The correlation coefficient, r, is used to test for significance

Rank-Correlation

- Often used in Sensitivity Analysis of Models
- Useful for assigning a relative order of importance
- Always inspect the scatter plots
  - Means little if curve is not monotonic

Regression Methods

Rank regression is popular because it handles many non-linear responses without having to search for transformations

Partial Regression Methods

Partial regression or partial rank regression is also popular because it removes the linear effects of the other variables (model parameters) when fitting each model parameter

Sandia National Laboratories

Simple or Rank Regression?

Large differences between the ordering of model parameters using simple and rank regression can indicate non-linear responses are present

Using UCalc to do Uncertainty and Sensitivity Analyses

The problem: Calculate the dose rate to a deer after feeding for 100 days on ^{137}Cs-contaminated vegetation and do an uncertainty analysis of the result

Parameter values assumed:

Quantity	Symbol	Most likely value	Distribution type	Std Dev or GSD	Min value	Max value
Feeding rate	R	1.2 kg/d	normal	Std dev = 0.2		
Concentr. in veg.	CV	10 Bq/kg	lognormal	GSD =1.5		
Absorbed fraction	A	0.4	triangular		0.3	0.6
Loss rate constant	K	0.058 d^{-1}	triangular		0.043	0.077
Muscle mass	M	28 kg	truncated normal	Std. dev = 4	16	45
Dose factor	DF	8.2E-9 Gy/d per Bq/kg	lognormal	GSD = 1.2		

Sandia National Laboratories

The Deer Model

Rate of change in deer inventory

$$\frac{dq}{dt} = R \text{ kg-d}^{-1} \times CV \text{ Bq-kg}^{-1} \times A - k \text{ d}^{-1} \times q \text{ Bq}$$

Concentration in deer

$$C_m \text{ Bq/kg} = q \text{ Bq} / M \text{ kg}$$

Dose rate

$$DR \text{ Gy-d}^{-1} = C_m \text{ Bq-kg}^{-1} \times DF \text{ Gy-d}^{-1}\text{Bq}^{-1}\text{-kg}$$

Sandia National Laboratories

Algebraic solution (deterministic) using the most probable single-values

$$q = \frac{R \cdot CV \cdot A}{K}\left(1 - e^{-Kt}\right)$$

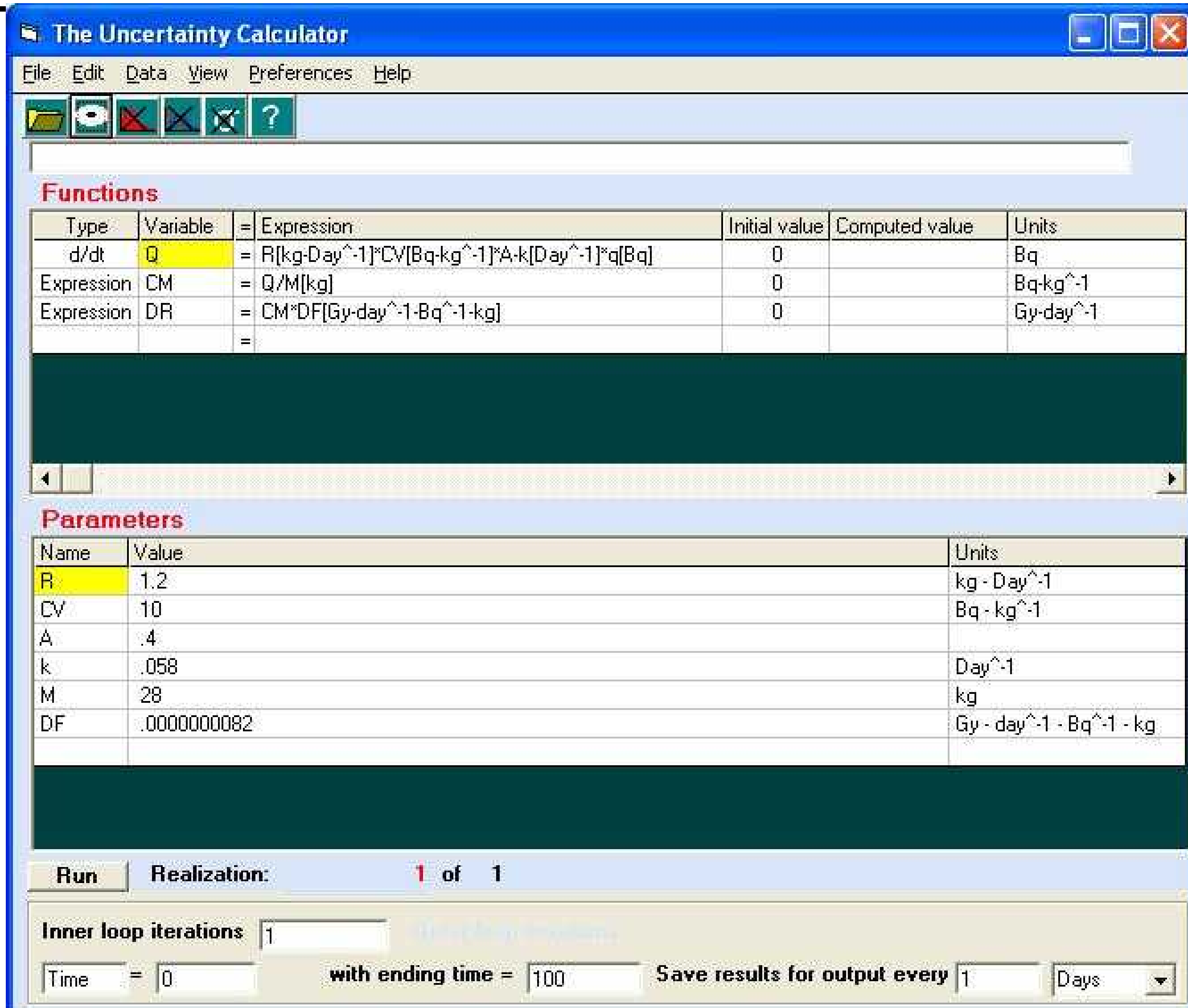
$$q = \frac{1.2 \cdot 10 \cdot 0.4}{0.058}\left(1 - e^{-0.058x100}\right)$$

$$= 82.5Bq$$

$$CM = \frac{82.5}{28} = 2.95Bq/kg$$

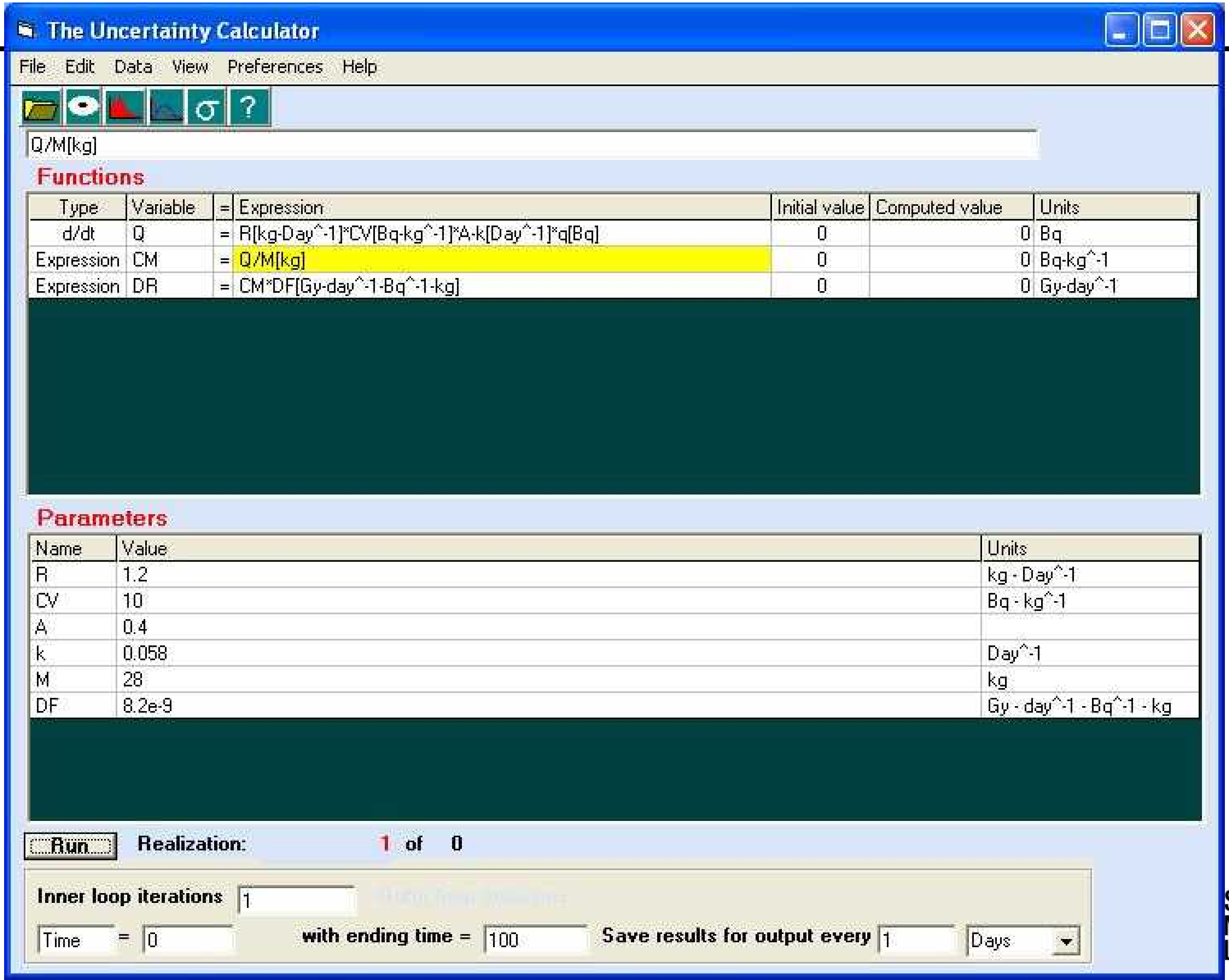
$$DR = 2.95 \cdot 8.2E - 9 = 2.42E - 8Gy/d$$

UCalc screen for deer model



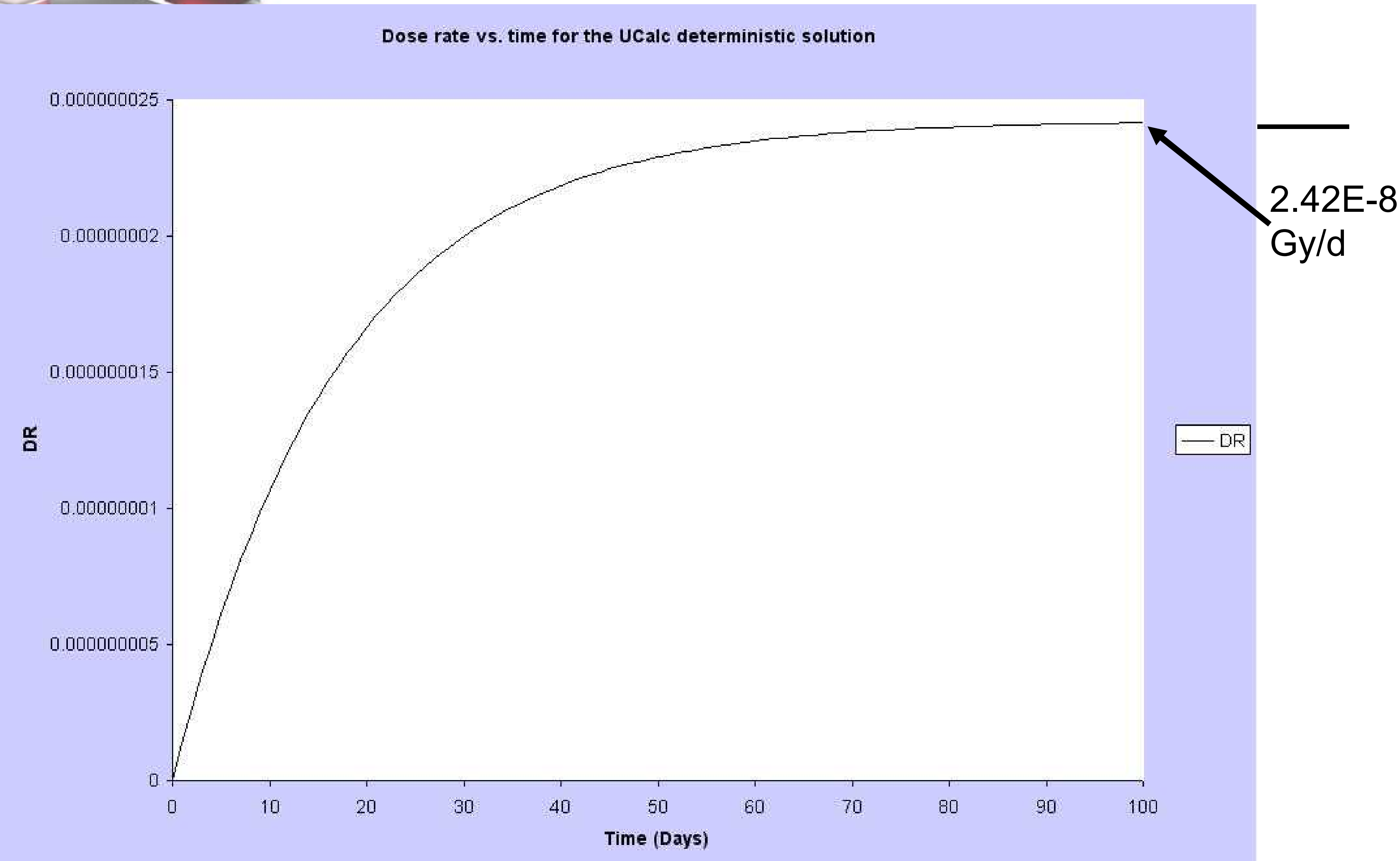
Sandia National Laboratories

UCalc screen for deer model with units checking



Sandia National Laboratories

UCalc solution-deterministic



Now, the question becomes: What is the uncertainty of this prediction?

Input data for the model (UCalc report)

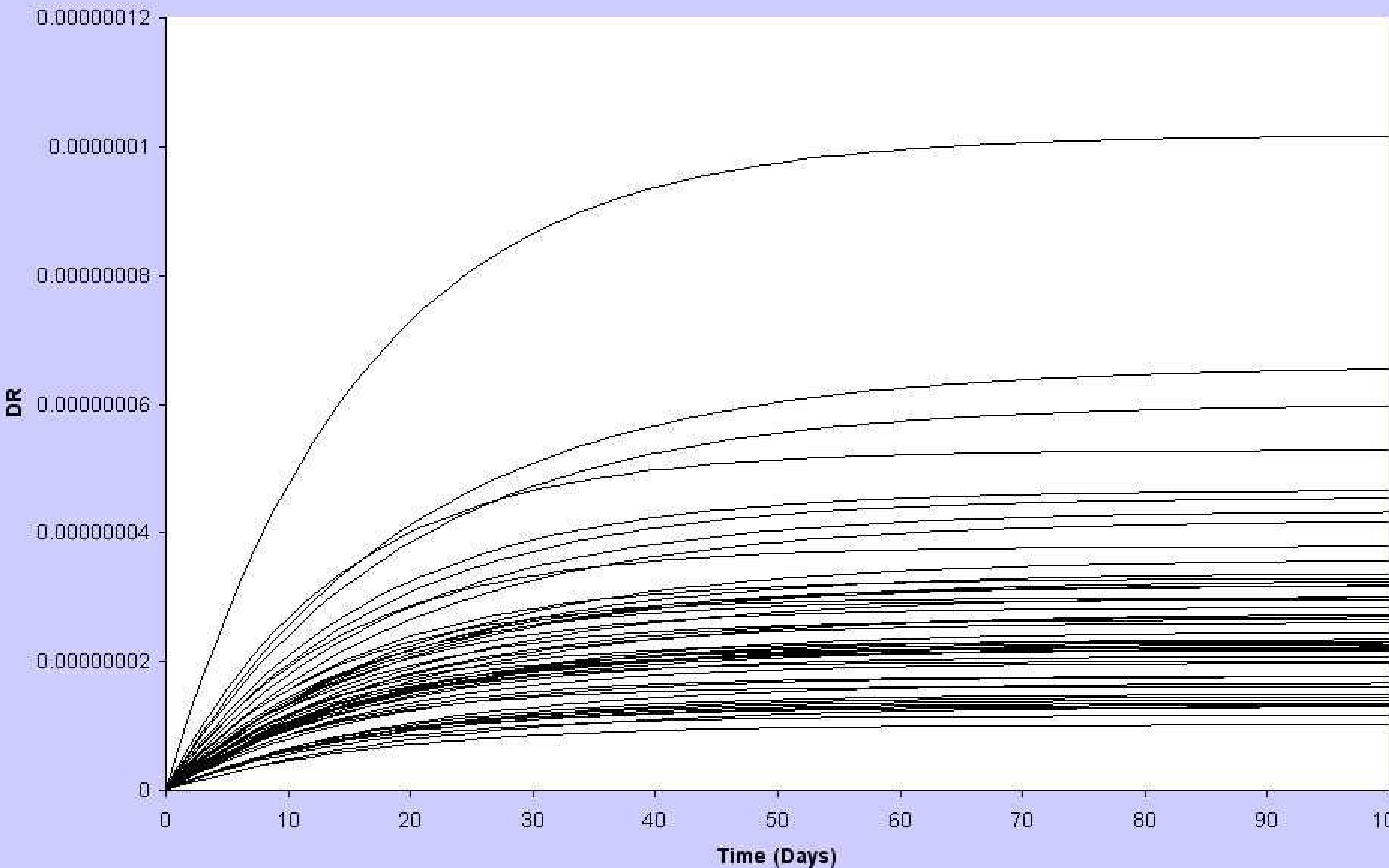
Expressions

d/dt Q = R[kg-Day^-1]*CV[Bq-kg^-1]*A-k[Day^-1]*q[Bq] 0 [Units = Bq]
Expression CM = Q/M[kg] 0 [Units = Bq-kg^-1]
Expression DR = CM*DF[Gy-day^-1-Bq^-1-kg] 0 [Units = Gy-day^-1]

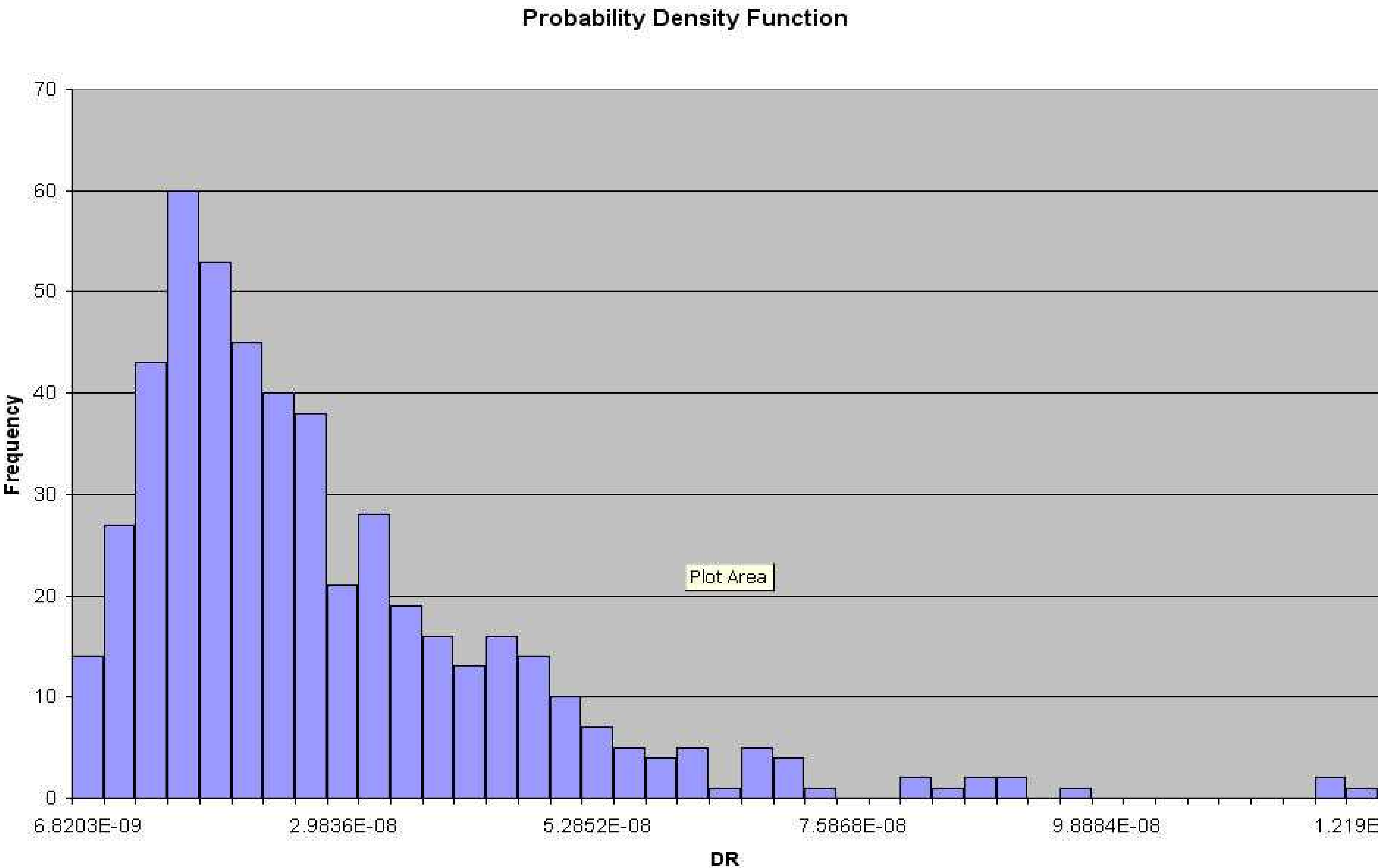
Model Parameters

R 1.2 kg - Day^-1
CV 10 Bq - kg^-1
A .4
k .058 Day^-1
M 28 kg
DF .0000000082 Gy - day^-1 - Bq^-1 - kg
Number of Inner Realizations: 1
Number of Outer Realizations:
Time Step for Saving Results: 1
Time Step for Sampling Time-Varying Parameters: 1
Name of Independent Variable: Time
Starting Value of Independent Variable: 0
Ending Value of Independent Variable: 100
Unit of Time: Days
Random number seed: 31415

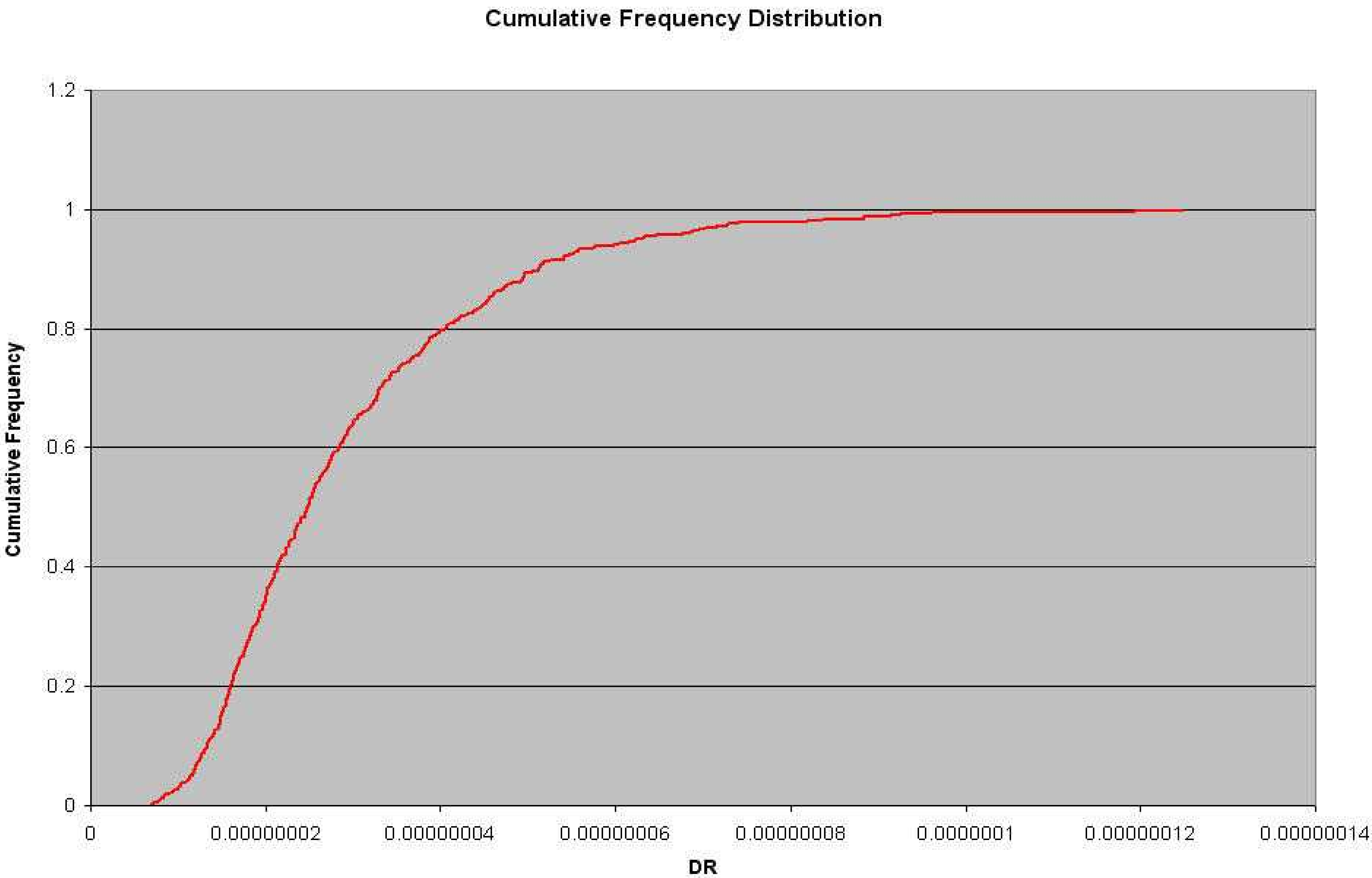
UCalc output with 50 realizations



Frequency histogram for dose rate at time = 100 days (500 runs)



Cumulative distribution for dose rate at time = 100 d (500 runs)



Numerical descriptors of the dose rate output distribution

Mean	2.940E-008	Minumim	6.820E-009	Geometric mean	2.543E-008
Median	2.482E-008	Maximum	1.248E-007	Geometric SD	1.699E+000
Standard deviation	1.744E-008	Skewness	1.883E+000	N	500
Variance	3.043E-016	Kurtosis	5.190E+000	N positive values	500
Standard error	7.802E-010				

Percentiles Based on Order Statistics

1.0%: 7.934E-009	25%: 1.718E-008	55%: 2.622E-008	85%: 4.551E-008
2.5%: 9.381E-009	30%: 1.869E-008	60%: 2.834E-008	90%: 5.130E-008
5%: 1.143E-008	35%: 2.008E-008	65%: 3.046E-008	95%: 6.225E-008
10%: 1.335E-008	40%: 2.143E-008	70%: 3.300E-008	97.5%: 7.292E-008
15%: 1.493E-008	45%: 2.334E-008	75%: 3.663E-008	99.0%: 9.142E-008
20%: 1.612E-008	50%: 2.495E-008	80%: 4.046E-008	

Show DR at Time 100 ☐ Show correlations

UCalc results on the sensitivity analysis through statistical regressions

Mean	2.920E-008	Minumim	3.553E-009	Geometric mean	2.529E-008
Median	2.632E-008	Maximum	1.290E-007	Geometric SD	1.732E+000
Standard deviation	1.609E-008	Skewness	1.486E+000	N	500
Variance	2.589E-016	Kurtosis	3.961E+000	N positive values	500
Standard error	7.196E-010				

Percentiles Based on Order Statistics

1.0%: 6.869E-009	25%: 1.738E-008	55%: 2.806E-008	85%: 4.458E-008
2.5%: 8.158E-009	30%: 1.900E-008	60%: 2.983E-008	90%: 4.984E-008
5%: 1.014E-008	35%: 2.040E-008	65%: 3.217E-008	95%: 5.780E-008
10%: 1.245E-008	40%: 2.300E-008	70%: 3.430E-008	97.5%: 7.488E-008
15%: 1.425E-008	45%: 2.459E-008	75%: 3.681E-008	99.0%: 8.162E-008
20%: 1.602E-008	50%: 2.645E-008	80%: 3.953E-008	

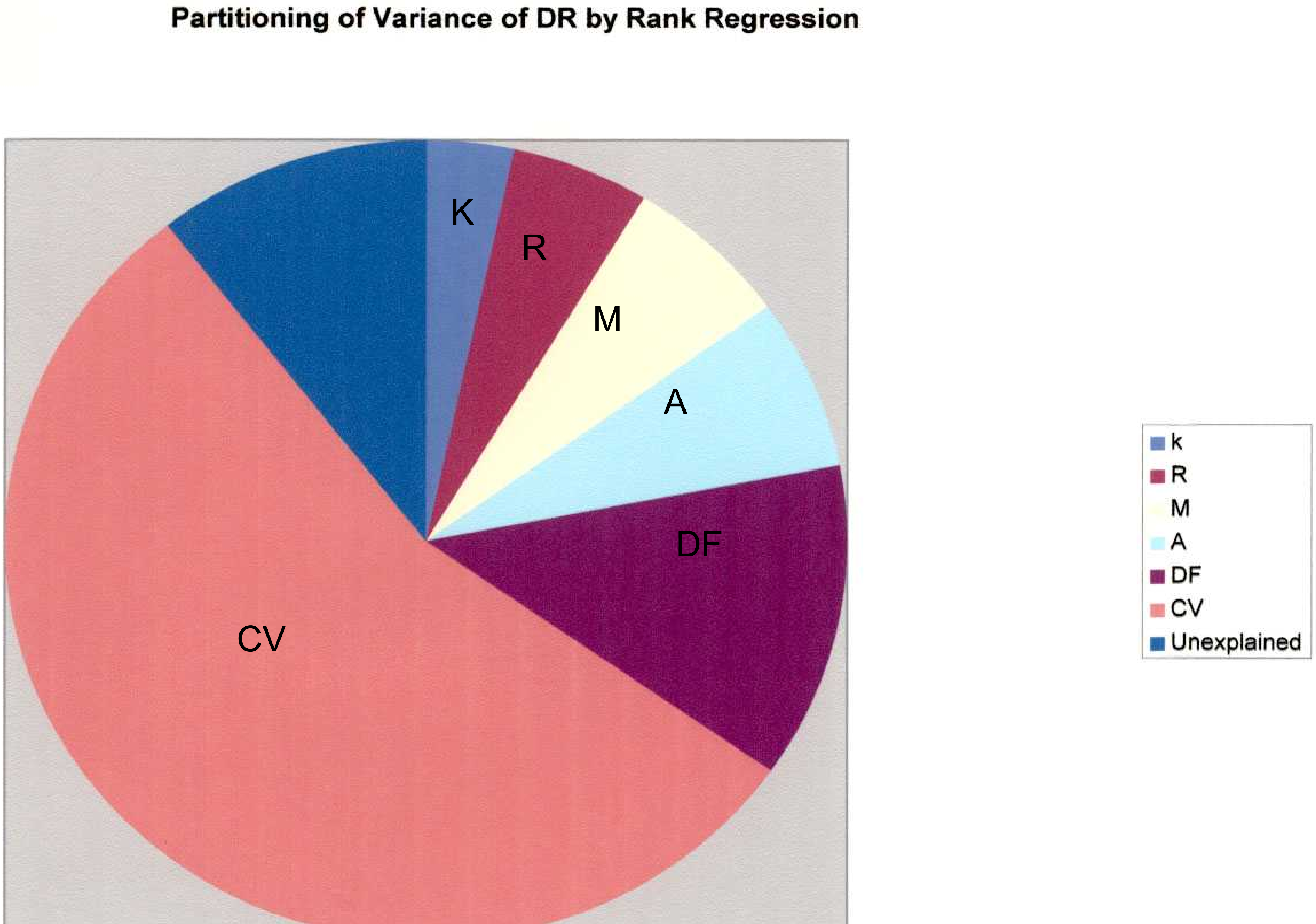
Show DR **at Time** 100 ☑ **Show correlations**

Correlations (r^2)

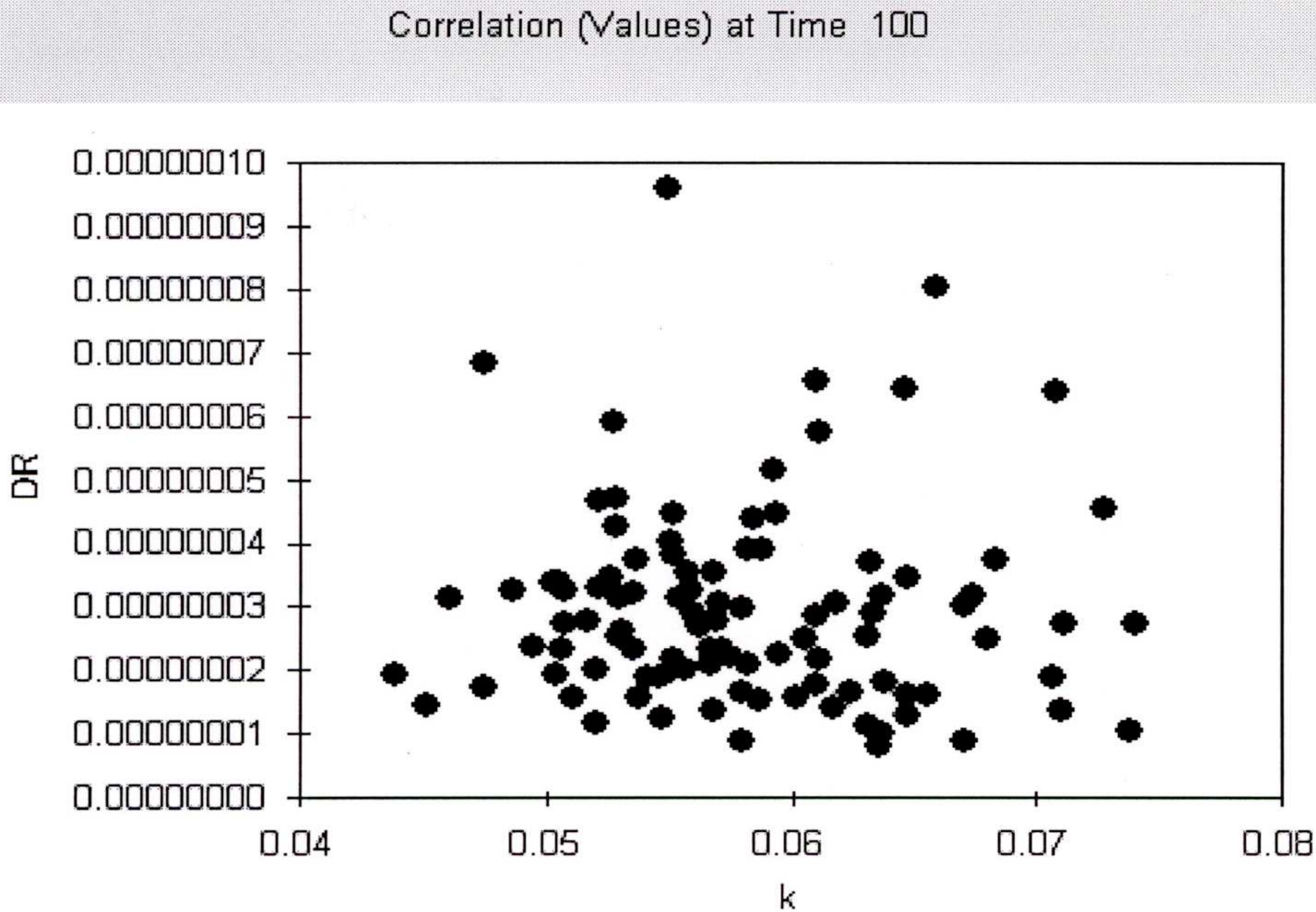
Input variable	Partial	Simple	Partial Ranked	Simple ranked
CV	0.86474	0.54573	0.90892	0.59304
DF	0.56157	0.10167	0.64031	0.10896
M	0.43304	0.07038	0.51034	0.0615
R	0.49921	0.05655	0.60722	0.05307
A	0.41323	0.05493	0.52159	0.05545
k	0.33212	0.03435	0.42231	0.03625
Total		0.86362		0.90827

Chart of Variance **Plot Correlations**

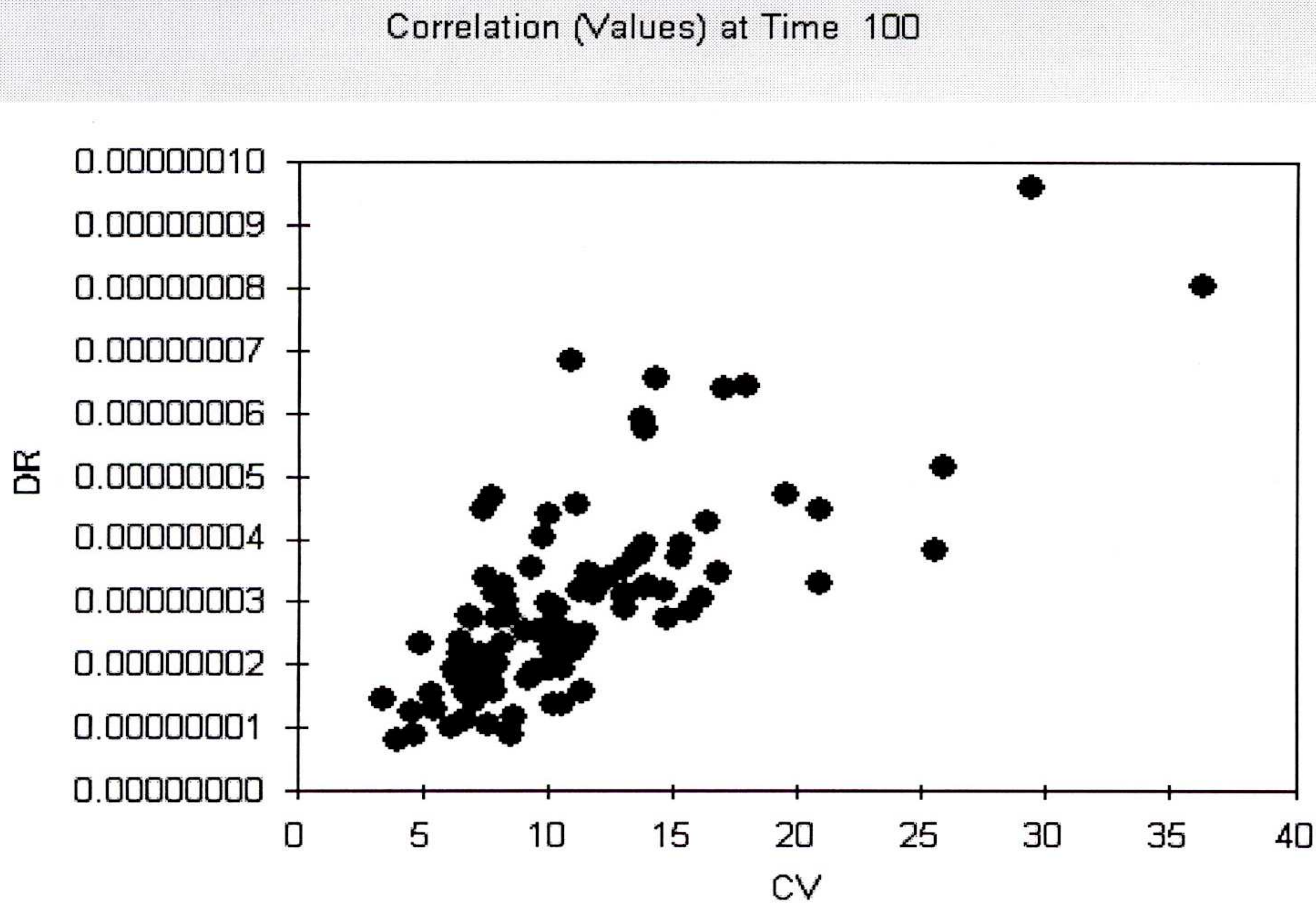
Pie chart showing how variance is partitioned among the input parameters



Dose rate vs. K, the least influential parameter, based on rank regression

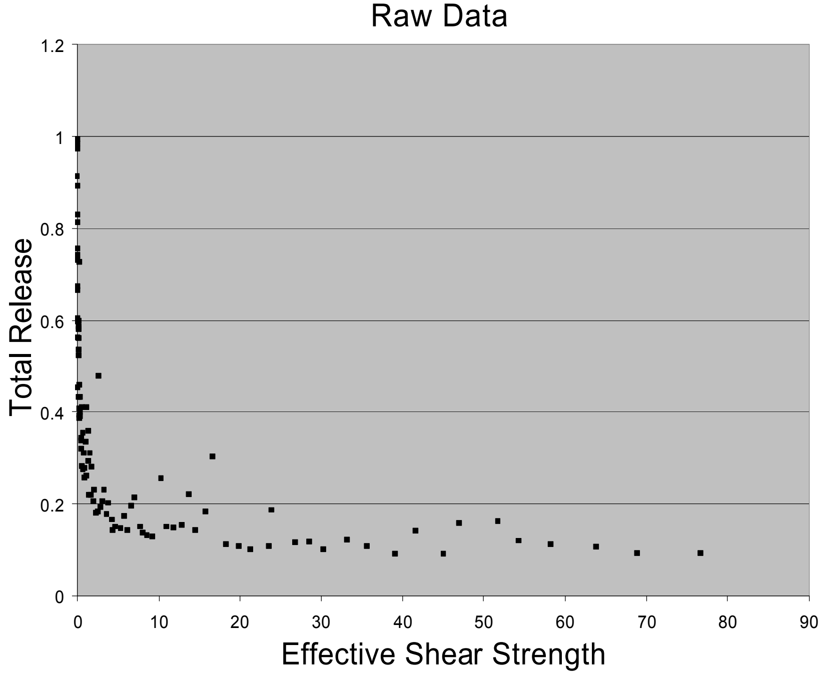


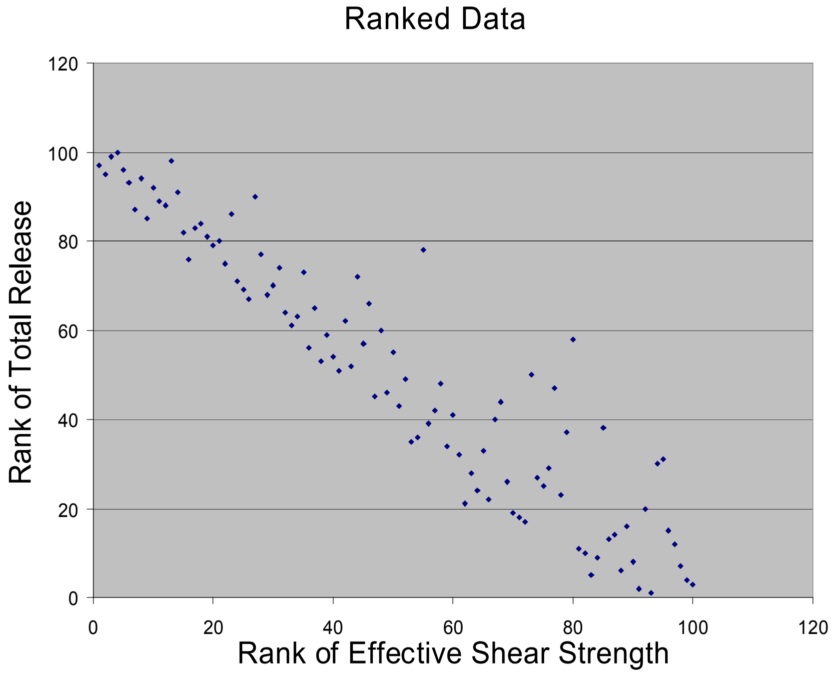
Dose rate vs. CV, the most influential parameter, based on rank regression



Slides below here are extras – to be used if questions arise

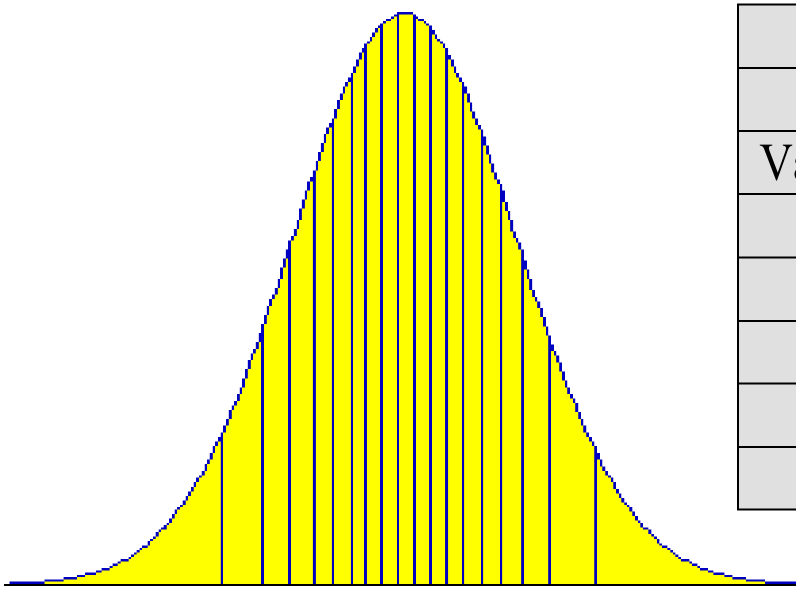
Sensitivity of Total Release to Effective Shear Strength





Latin Hypercube Sampling

- **Latin Hypercube Sampling is used for parameters**
  - **Compared to simple random sampling, LHS reduces the number of vectors required to estimate means, variances, etc.**
  - **Extreme values of input parameters are sampled**



	Simulation Number									
	1	2	3	4	5	6	7	8	9	10
Variable	Quantile Sampled									
A	6	3	9	1	2	7	8	5	10	4
B	8	5	2	9	6	4	10	7	1	3
C	2	7	10	8	4	9	6	1	5	3
D	5	9	4	3	8	10	2	7	6	1
E	7	10	2	5	8	9	6	3	1	4

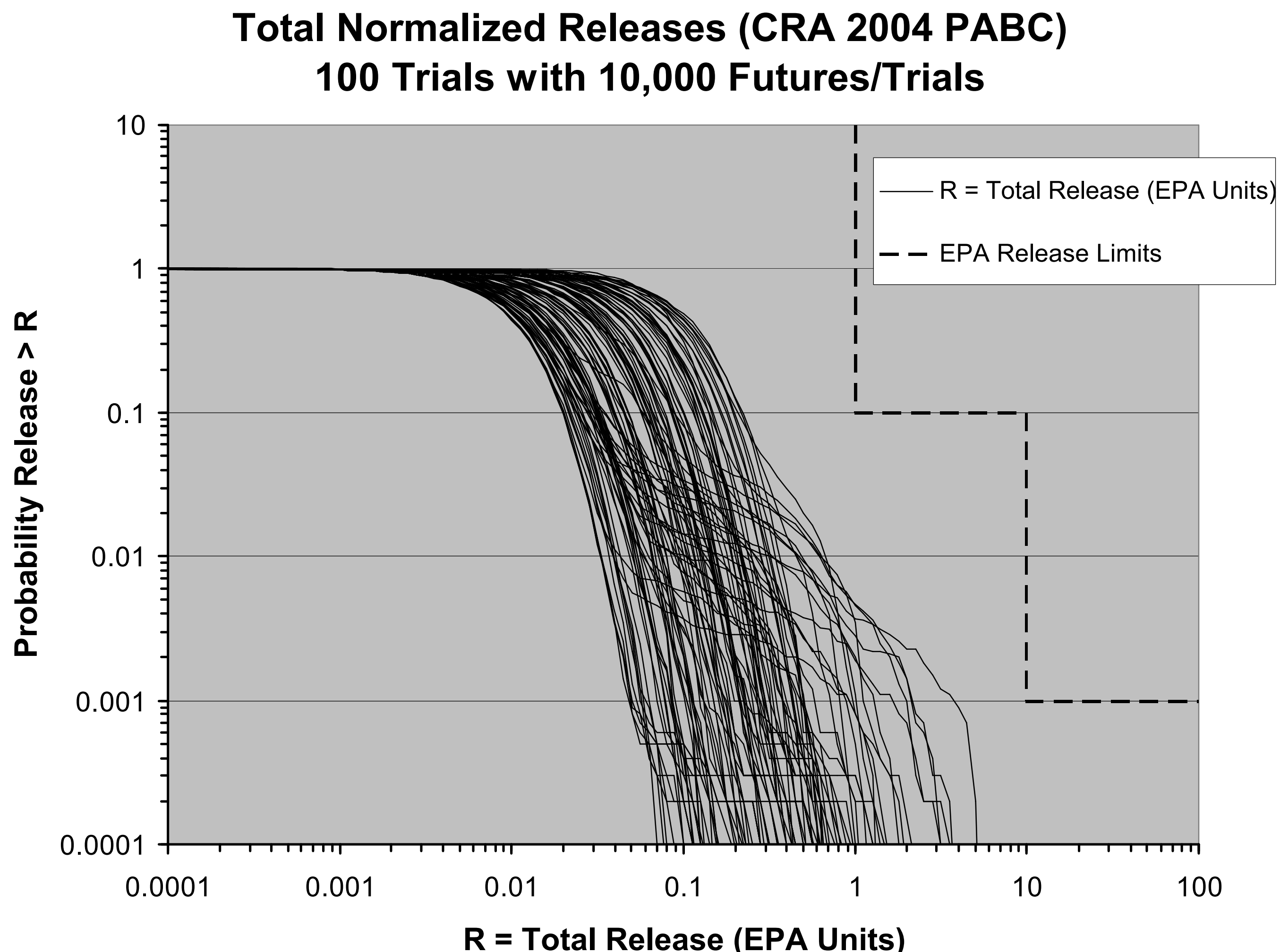
Sandia National Laboratories

Modeling Intrusion Scenarios (Aleatory Uncertainty)

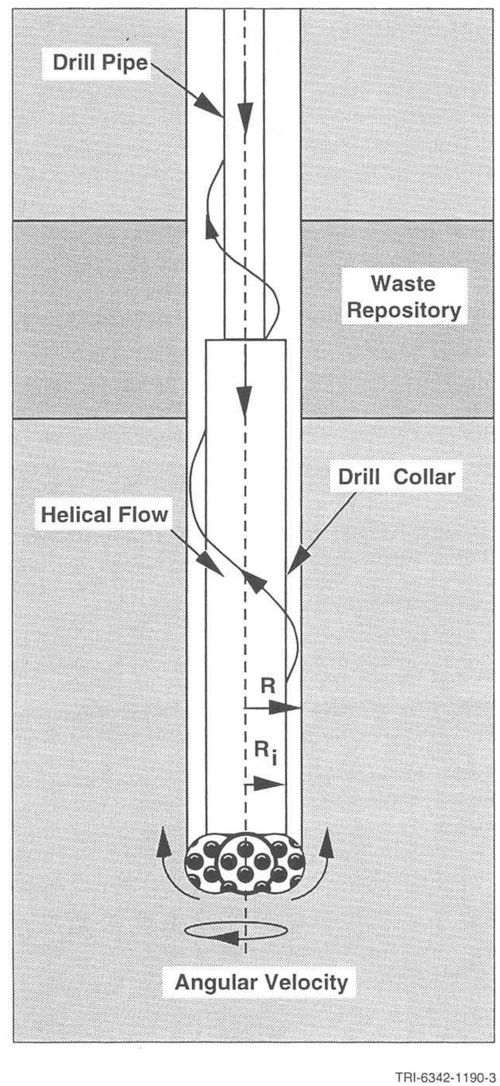
CCDFGF generates 10,000 possible futures for each vector.

- A future is the cumulative release from one possible sequence of events from 0 to 10,000 years.
- Each future consists of a series of randomly occurring drilling intrusions.
- Each set of 10,000 simulations yields one CCDF

“Horsetail” Plot



Shear Strength Impacts Releases



Sensitivity Analysis Methods(cont.)

- **Partial correlation coefficients**
  - Remove the linear effects of variables already included in the model.
  - Used to order the addition of variables to the model.
  - Used to rank the importance of a parameter

Case Study

The model equations

$$C_{air} = S_{release}\frac{\chi}{Q}$$

$$D_{thyroid} = \frac{C_{air} V_{dep} I_{cow} T_{forage/milk} I_{child} D_{child}}{\left(K_{weathering} + K_{decay}\right) D_{forage}}$$

Sensitivity Analysis

Rank correlation (r) and contribution to variance (%Var)

	r	%Var
Milk consumption rate	0.69	51.9
Forage mass	0.51	27.5
Forage/milk transfer coefficient	0.40	18.8
Avg. daily intake rate	0.14	2.1
Weathering rate	0.12	1.6

Analysis of Variance

Analysis of Variance

- **Requires full factorial sampling design**
- **Levels of parameters are set across the range of the distributions, e.g. at the 1st, 20th, 40th, 60th, 80th and 99th percentiles.**

Sensitivity Analysis

Correlation Ratio

$$\frac{Var_{X_i}\left[E\left[Y \mid X_j = x_j\right]\right]}{Var\left[Y\right]} = \frac{VCE}{Var\left[Y\right]}$$

VCE: Variance Correlation Expectation

Where the estimate of the total variance is

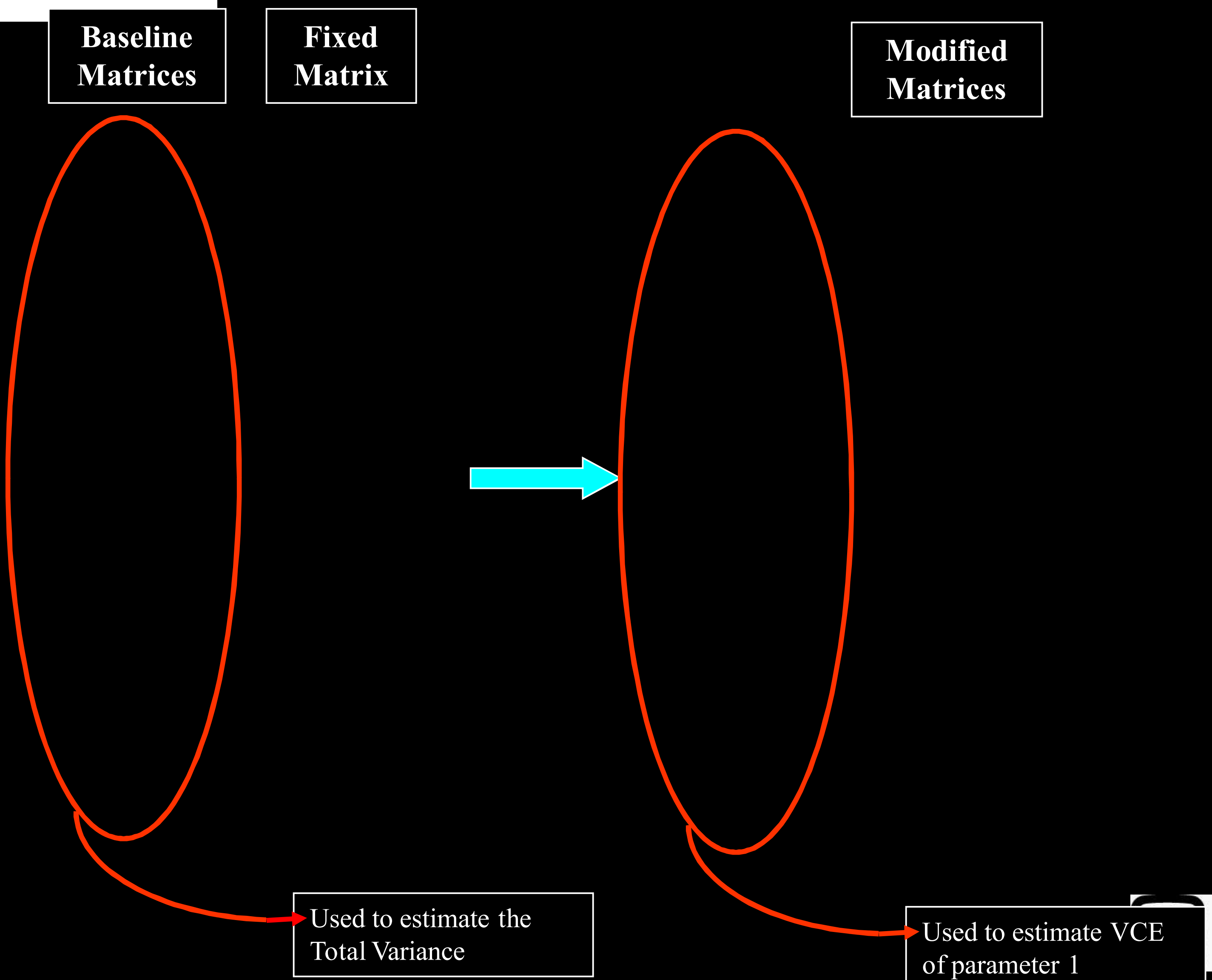
$$\hat{V}ar\left[Y\right] = \frac{1}{mr}\sum_{j=1}^{m}\sum_{l=1}^{r}\left(y_{jl} - \bar{y}\right)^2$$

k= = number of parameters
r = number of baseline LHS matrices
m = samples per parameter

And the grand mean is

$$\bar{y} = \frac{1}{mr}\sum_{j=1}^{m}\sum_{l=1}^{r} y_{jl}$$

Sandia National Laboratories



Sandia National Laboratories

Sensitivity Analysis

Correlation Ratio

$$\hat{V}CE(X_i) = \frac{1}{m}\sum_{j=1}^{m}\left(\bar{y}_j - \bar{y}\right)^2 - \frac{1}{mr^2}\sum_{j=1}^{m}\sum_{l=1}^{r}\left(y_{jl}^{(i)} - \bar{y}_j\right)^2$$

$$\bar{y}_j. = \frac{1}{r}\sum_{l=1}^{r} y_{jl}$$

$$\eta^2 = \frac{\frac{1}{m}\sum_{j=1}^{m}\left(\bar{y}_{j.} - \bar{y}\right)^2 - \frac{1}{mr^2}\sum_{j=1}^{m}\sum_{l=1}^{r}\left(y_{jl}^{(i)} - \bar{y}_j.\right)^2}{\frac{1}{mr}\sum_{j=1}^{m}\sum_{l=1}^{r}\left(y_{jl} - \bar{y}\right)^2} = \frac{r\sum_{j=1}^{m}\left(\bar{y}_{j.} - \bar{y}\right)^2 - \frac{1}{r}\sum_{j=1}^{m}\sum_{l=1}^{r}\left(y_{jl}^{(i)} - \bar{y}_j.\right)^2}{\sum_{j=1}^{m}\sum_{l=1}^{r}\left(y_{jl} - \bar{y}\right)^2}$$

$$\eta_{bias}^2(X_i) = \frac{r\sum_{j=1}^{m}\left(\bar{y}_{j.} - \bar{y}\right)^2}{\sum_{j=1}^{m}\sum_{l=1}^{r}\left(y_{jl} - \bar{y}\right)^2}$$

Sandia National Laboratories

Fourier Amplitude Sensitivity Test (FAST)

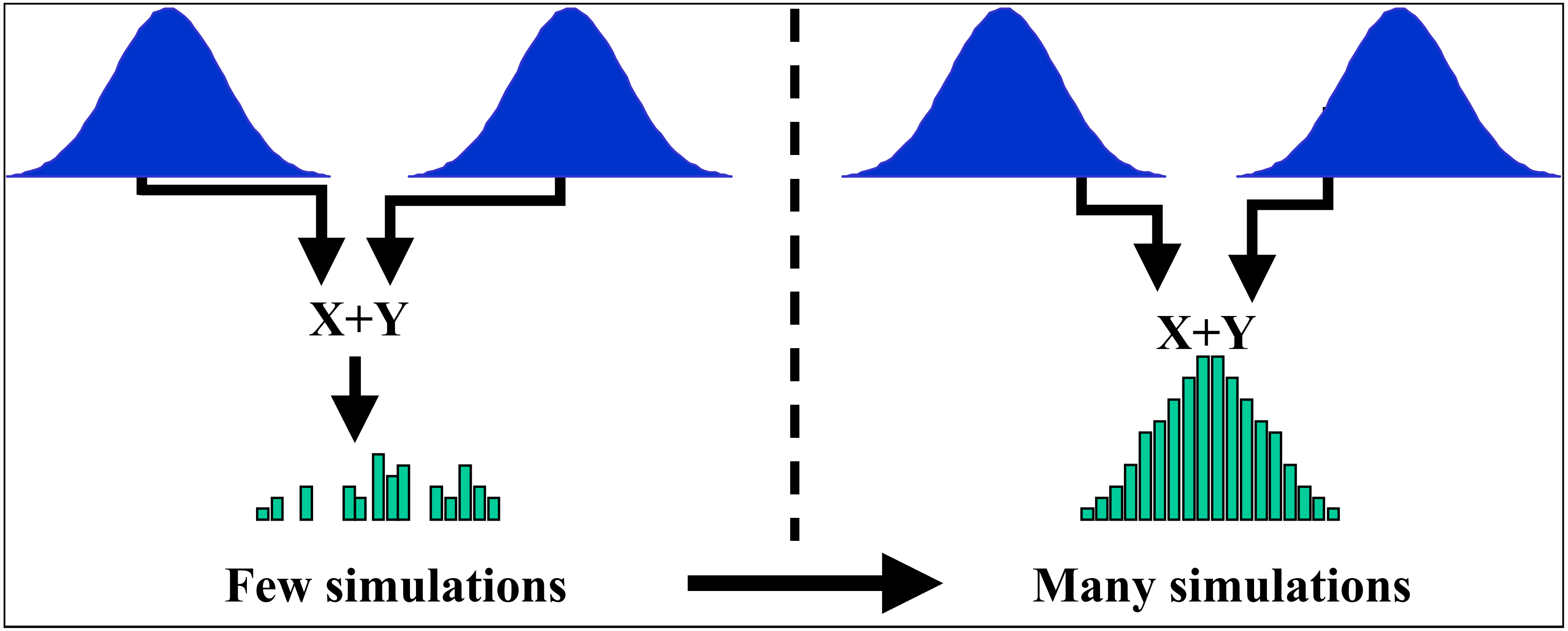
A Fourier series is fit to the model results to convert a multidimensional integral over the model parameters into a 1-dimensional integral

A Fourier decomposition is used to obtain the fractional contributions of the input parameters to the total variance of the model prediction

Uncertainty & Sensitivity Analyses

- Global sensitivity analysis can be conducted using only results from uncertainty analysis
  - No additional simulations required
  - Statistical model used to relate outputs to inputs

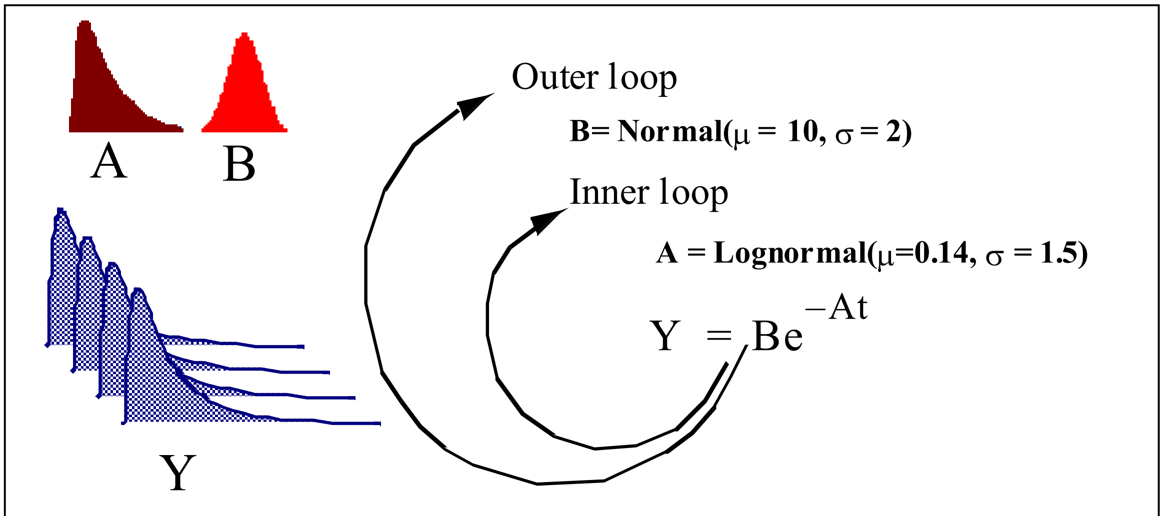
Propagation of Uncertainty



Classification of Uncertainty

- Epistemic (subjective): Arises from a lack of knowledge about parameters assumed to have fixed values within the computational implementation of a PA.
  - Examples: Permeability, Porosity, etc.
- Aleatory (stochastic): Arises because the system can potentially behave in many different ways. The sequence of future events is not known.
  - Example: Timing of future drilling events.

Nested Sampling of Aleatory and Epistemic Uncertainties

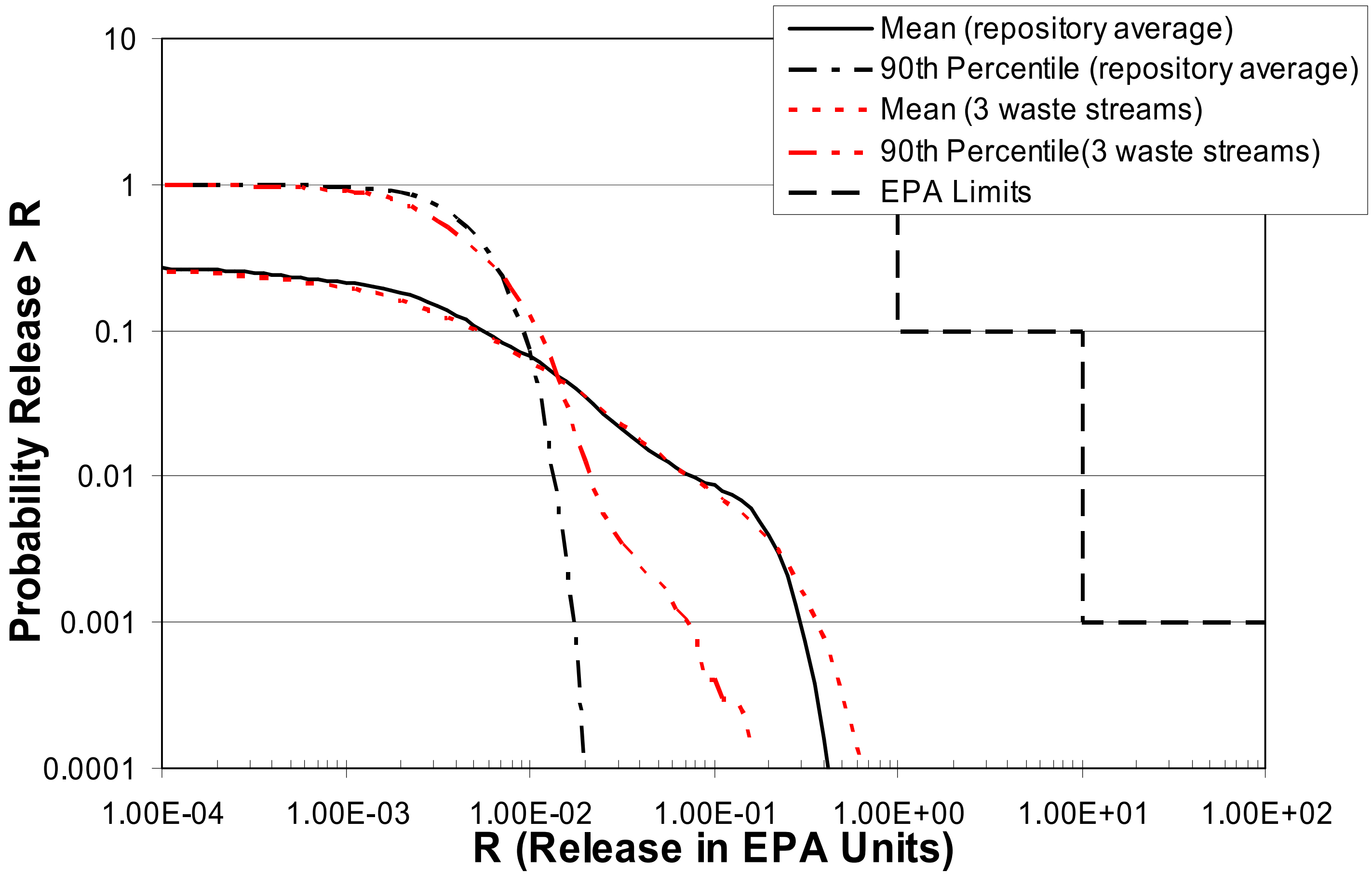


[redacted]

Sampling Design

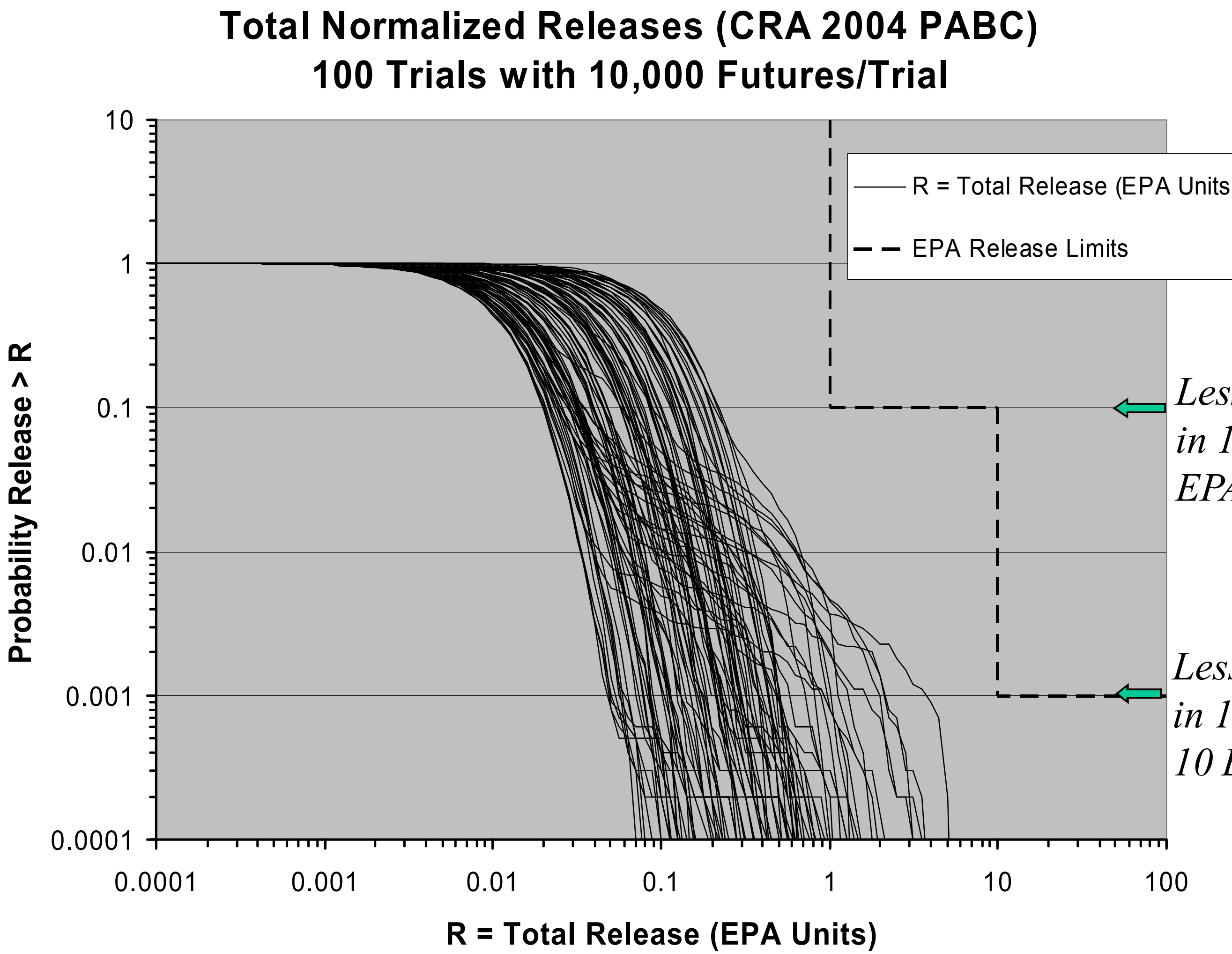
- Outer loop steps through 100 vectors of input parameters per replicate
  - 56 parameters are sampled for each 2004 PABC vector
- Inner loop simulates 10,000 futures for each vector
- Hence, 1,000,000 simulations of 10,000 year futures are simulated per replicate
- Three replicates are normally run for a PA

Sensitivity to Repository Scale Assumptions



Sandia National Laboratories

“Horsetail” Plot of CCDFs (Cumulative Complimentary Distribution Functions)



Less than 1 chance in 10 of exceeding 1 EPA unit

Less than 1 chance in 1000 of exceeding 10 EPA units

Sandia National Laboratories

Sensitivity Analysis Methods

- Stepwise linear multiple regression
  - Evaluates the relative importance of the various sampled parameters on the estimates of potential releases.
  - Constructs a sequence of regression models starting with the most influential input parameter.

Sensitivity Analysis Methods (cont.)

Rank regression

- Replaces the values of the data with their ranks

Benefits:

- Tends to linearize the response curves
- Standardizes the variability in the outputs and parameters by mapping the data into identical ranges.
- Tends to de-emphasize the impact of "outliers".