Raytheon

Customer Success Is Our Mission

Examples of Statistical Methods at CMMI Levels 4 and 5

Jeff N Ricketts, Ph.D.
jnricketts@raytheon.com

November 17, 2008

Agenda

- Overview
- Definitions
- Current state
- General Measurement Issues
- Steps in the Scientific Method
- Example Statistical applications to engineering
- Example Statistical Model
- Summary

Overview

The current practice of using control charts to achieve level 4-5 maturity ratings is inadequate to demonstrate that the organization is identifying sources of variation within the product development process or testing hypotheses. This presentation proposes the application of the scientific method and inferential statistical models to identify, control and eliminate sources of variation in product and system development by identifying independent variables that may be used to predict their effects on subsequent dependent variables. Examples of hypothesis testing and inferential statistical models and their application to this process are provided.

Scientific Method and inferential Statistics Defined

Raytheon

The SCIENTIFIC METHOD is a body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge. It is based on gathering observable, empirical and measurable evidence subject to specific principles of reasoning. The scientific method consists of the collection of data through observation and experimentation, and the formulation and testing of hypotheses. The scientific method is used to explain and predict the causes of variability in natural phenomena.

INFERRENTIAL STATISTICS or statistical induction comprises the use of sample statistics to make inferences concerning relationships within a population. These relationships are expressed in causal terms.

Current State of the Practice

Engineering Measures:

- Staffing
- CPI/SPI
- Productivity
- Defect Density
- Defect Containment
- Problem Report Open and Closure status
- Requirements Volatility

General Measurement Issues

The standard measures commonly in use today all have one thing in common: they are historical vs. predictive

They are all reactive vs. proactive

Some metrics have little relationship to the real questions that need to be answered

Corrective actions are only applied to 0.03% of the observations because 99.7% of the variation is “under control” (3σ)

There are no standard measurement definitions

No one seems to be doing anything about the measures

Observe the Process

Raytheon

The product development process consists of many variables (tools, people, processes, inputs, outputs)

There is a lot of variation in these factors and consequences to the variation:

- stability of requirements
- makeup of peer review teams
- stability of design
- types of tools and technology used
- number of defects identified in peer reviews
- amount of hrs of training per engineer
- maturity of technology
- types of development environments used
- skill sets/mix
- programming language or design methods used

Connect the Dots (Formulate conditional associations)

Raytheon

X seems to happen more often when Y is around

We always seem to do better when we use this product/method/tool/process

Do we really save time by conducting formal peer reviews for reused and ported code?

Are peer reviews even necessary on a product line?

Use cases take a long time to develop. Are they really necessary?

The key is to identify factors that appear to be associated with each other or are reducing/increasing cost and schedule

Formulate Null Hypotheses

If you believe/observe that there is a causal relationship between two variables, the relationship is stated in the form of “no difference”.

e.g. Systems engineers find the same number of defects during peer reviews as software engineers.

e.g. The amount of preparation time one takes for a peer review has no relationship to the number of defects identified.

Measure the Process

Measurements must be consistent, precise and repeatable

Measures are targeted for the type of statistics that will be generated

- Nominal - categorical/dichotomous- systems engineers vs. software engineers
- Ordinal - categorical -low medium high- complexity factors, lift/mod/reuse
- Interval - frequency distributions- 1…n - years of experience
- Ratio - frequency distributions with an absolute zero

Measures by category of data

Nominal	Difference in proportions, Chi square, Lambda, student's t test
Ordinal	Analysis of Variance, Exactness tests, Rank Order correlation, Gamma
Interval	Correlation and regression, Multiple and stepwise regression, path analysis
Ratio	Correlation and regression, multiple and stepwise regression, path analysis

Generate a Sample (test) Statistic

Raytheon

Samples must be representative of the population under study

Samples must be randomly selected (can be simple, stratified, cluster, etc)

Samples cannot be the whole population

Statistics computed must be appropriate for the level of measurement

Test the Hypotheses

What is the observed difference between Group A and Group B?

What is the measure of association between the independent variable (X) and the dependent variable (Y)?

Significance levels tell you if the observed difference is statistically significant

- Given no relationship between what you measured, this is the probability (.05, .01, .001) that you would observe this result in a randomly drawn sample from the target population.

Example One: Categorical Data (Chi^2)

Issue: Who makes a better tester? Systems (because they write the requirements) or software (because they coded the implementation of the requirements)?

A random sample of 458 developers is drawn, half systems engineers and half software engineers. They are provided the same software components, test procedures and tools to integrate and test the code. Who did better?

	Integration test defect yield	Formal test defect yield	Total
Software Engineers	126 (96.8)	99 128.2)	225
Systems Engineers	71 (100.2)	162 (132.8)	233
Total	197	261	458

The product of the marginals is divided by N to obtain expected frequencies if there were no difference. These are then subtracted from the observed frequencies, squared, divided by the expected frequency and summed to obtain a chi square test statistic.

Cell	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
a	126	96.8	29.2	852.64	8.808
b	99	128.2	-29.2	852.64	6.651
c	71	100.2	-29.2	852.64	8.509
d	162	132.8	29.2	852.64	6.42
Total	458	458	0		**30.388**

p>.001

What is the conclusion?

- Software engineers do a better job of finding defects during integration?
- Systems engineers do a better job of finding defects during formal test?
- If there is no difference between software engineers and systems engineers ability to identify defects during integration and formal testing, the probability of drawing a random sample that is distributed this way is less than one in one thousand.

What Else Could be Causing this?

- Are the software engineers less familiar with target hardware environment then SEs?
- Are the SEs less familiar with the development/integration environment than software engineers?
- Did the systems engineers miss identifying defects during integration because they overlooked design issues and focused on requirements?
- Did the software engineers find less defects during formal test because they had already found them during integration?
- Were the systems engineers cranky because they had to do software work?

- Further investigation may be warranted into what types of errors the two groups found, how much time they spent on finding the errors, and how familiar the two groups were with the tool sets.

Example Two: Nominal/Interval Data (ANOVA)

Raytheon

Three vendors are promoting design analysis tools that they say identify inconsistencies, holes, gaps, and other design problems. You decide to do a DAR to determine if one of them is significantly superior to the others. 8 software components are analyzed by each tool with the average defect discovery recorded below.
computed per module.

	Defect Removal Rate			
	Tool A	Tool B	Tool C	Total
	4.3	5.1	12.5	
	2.8	6.2	3.1	
	12.3	1.8	1.6	
	16.3	9.5	6.2	
	5.9	4.1	3.8	
	7.7	3.6	7.1	
	9.1	11.2	11.4	
	10.2	3.3	1.9	
sum	68.6	44.8	47.6	161
mean	8.58	5.6	5.95	6.71

	Sums of Squares	Degrees of Freedom	Estimate of Variance	F
Total	373.538	N - 1 = 23		
Between	42.303	k - 1 =2	21.152	
Within	331.235	N - k = 21	15.773	1.34

AVOVA (F test) is used to compare the variation within each category to the variation between categories.

F is the probability of observing the differences between the categories given there is no difference in the population.

Even thought tool A identified what appears to be a significantly greater number of defects, F is well below the .05 significance level.

Example Three: Interval/Interval Data (Correlation and Regression)

Raytheon

Management thinks that projects are spending too much money on reviewing products prior to the formal peer review meetings and decided to find out how valuable pre-reviews were in the first place. 25 peer reviews were randomly sampled and examined to determine if the number of hours spent reviewing products prior to the peer review meetings were impacting the number of defects identified during the reviews.

Hrs	Defects
1.5	3
6	20
2	5
8	24
3	7
9	30
4.5	10
12	19
20	40
30	50
15	21
9	15
25	50
4	6
22	35
1	0
16	25
40	80
32	60
15	20
6	10
11	17
8	20
3	12
18	30

This technique compares the covariance of hours spent vs defects found to the total variance in defects creating the linear (least squares) equation

$Y = a + bX$, where

Y = defects

a = Y intercept

b = slope of the line

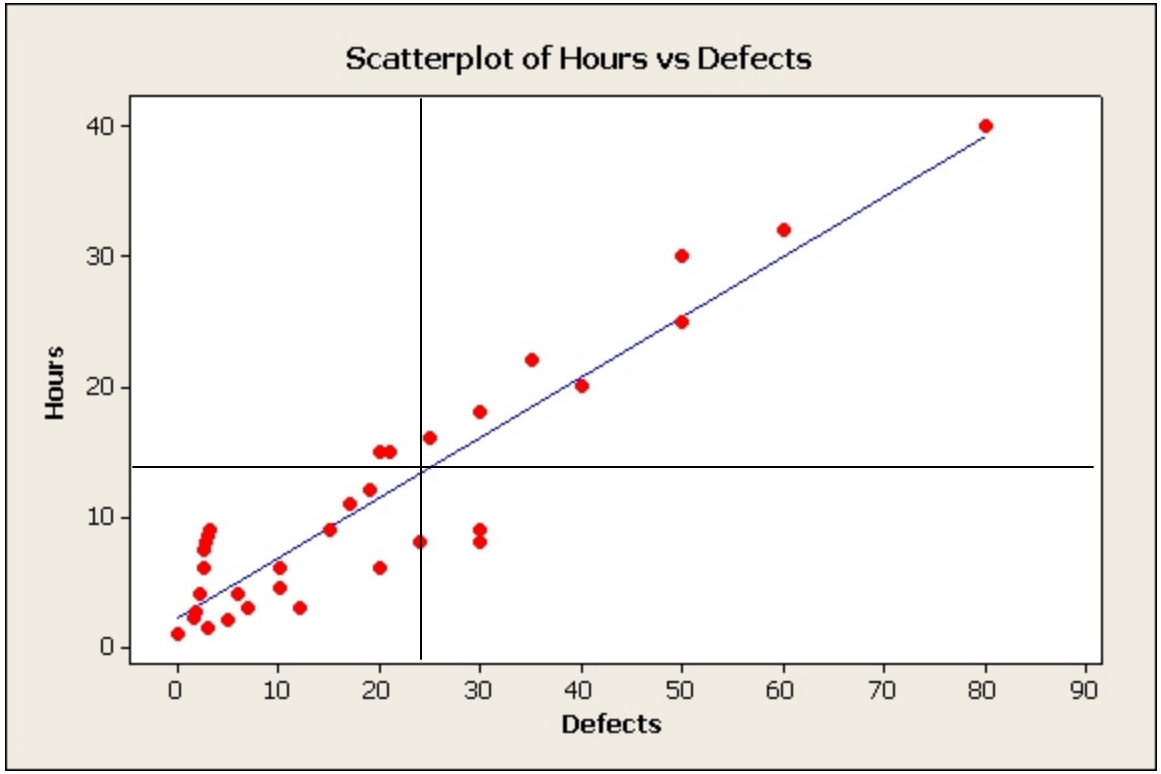
X = hours spent

The variation around this equation is then compared to the original variation around defects found. The percent reduction in variation is said to be “explained”

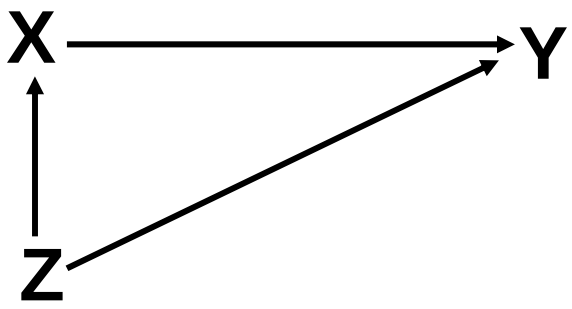
Analysis Results

$R^2 = .86$

Source	DF	SS	F	P<	
Regression	1	2571.8	203.33	0.0000	
Residual	31	392.1			

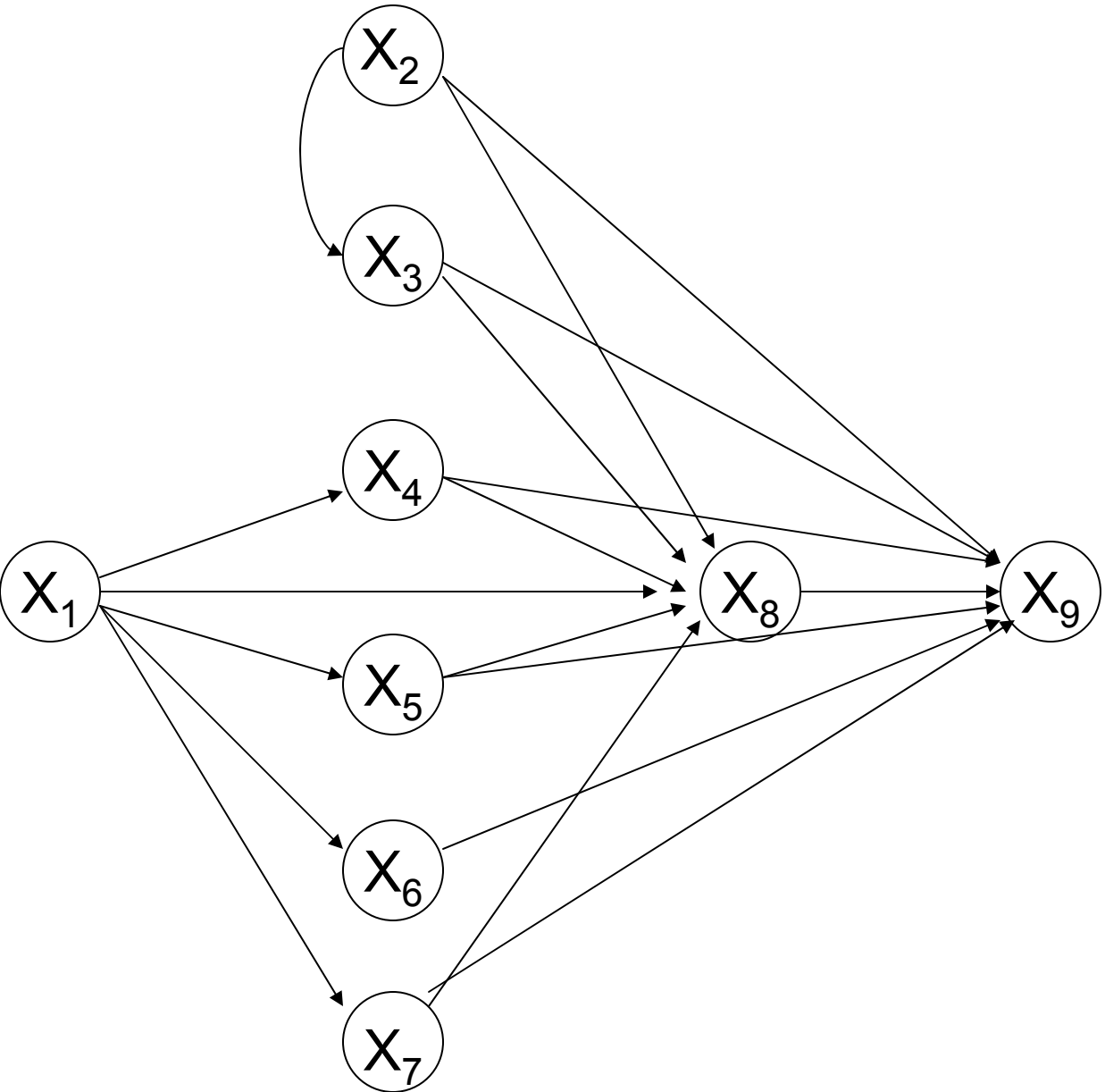


Beware of Spurious Relationships



Changes in X appear to be causing changes in Y when in fact Z is strongly correlated with both X and Y so when Z varies both X and Y vary.

What Causes Variation in Integration SPI/CPI?



X_1 = Training
X_2 = Technology Maturity
X_3 = Team Composition
X_4 = Hrs Spent In Peer Review
X_5 = Type of Review
X_6 = Domain
X_7 = Development Env
X_8 = Peer Review Efficiency
X_9 = IV&V CPI/SPI

Statistical Analysis Tools

1. SPC-PC - Excel macro based tool used for control charts
2. Minitabs - Excel macro based tool used for control charts, analysis of variance and regression analysis
3. MATLAB - Engineering modeling tool with statistical plugin
4. SAS - Powerful engineering based statistical modeling tool
5. SPSS - Powerful social science based statistical modeling tool
6. BMDP - Powerful medical based statistical modeling tool

Summary

- We could be doing a much better job and adding more value to our level 4-5 processes by incorporating the use of the scientific methods and inferential statistical models into our measurement and analysis processes
- The data is there, but being collected inconsistently
- Random samples allow us to create probability distributions, generate sample statistics and to test null hypotheses that will aid us in being able to predict the effect of fine tuning our processes used to build our products and Dispel myths and non truths regarding the value of non-value added tasks.
- Statistically significant results typically warrant further investigation
- Correlation is not necessarily causation

Questions

- For any questions someone might have on today's presentation

- For future questions the presenter contact information is:
  - Jeff N Ricketts, Ph. D.
  - jnricketts@raytheon.com
  - 714.446.4598

Presenter Biography

Dr. Ricketts has over 25 years of experience in software intensive system development in the areas of communications, air defense and air traffic control. He has been involved in the CMM/CMMI since it's inception and has participated in 12 formal appraisals (SCE/SPA/SCAMPI). He recently was part of the Raytheon NCS hardware, software and systems engineering SCAMPI effort that resulted in a level 5 rating for Raytheon's Network Centric Systems five major design centers. He holds a Doctorate degree in social statistics from Washington State University. He currently Serves as a technical director at Raytheon's Fullerton, California system design center.