Building Models from Your PSM Data

Brad Clark, Ph.D.

Software Metrics, Inc.

Objectives

- Share data analysis experiences with real PSM data
- Show how models created from data are based on the average or *mean* of the data and its spread or *standard deviation*
- Show how model performance improves with the removal of *assignable causes of variation*

Using Data to Estimate

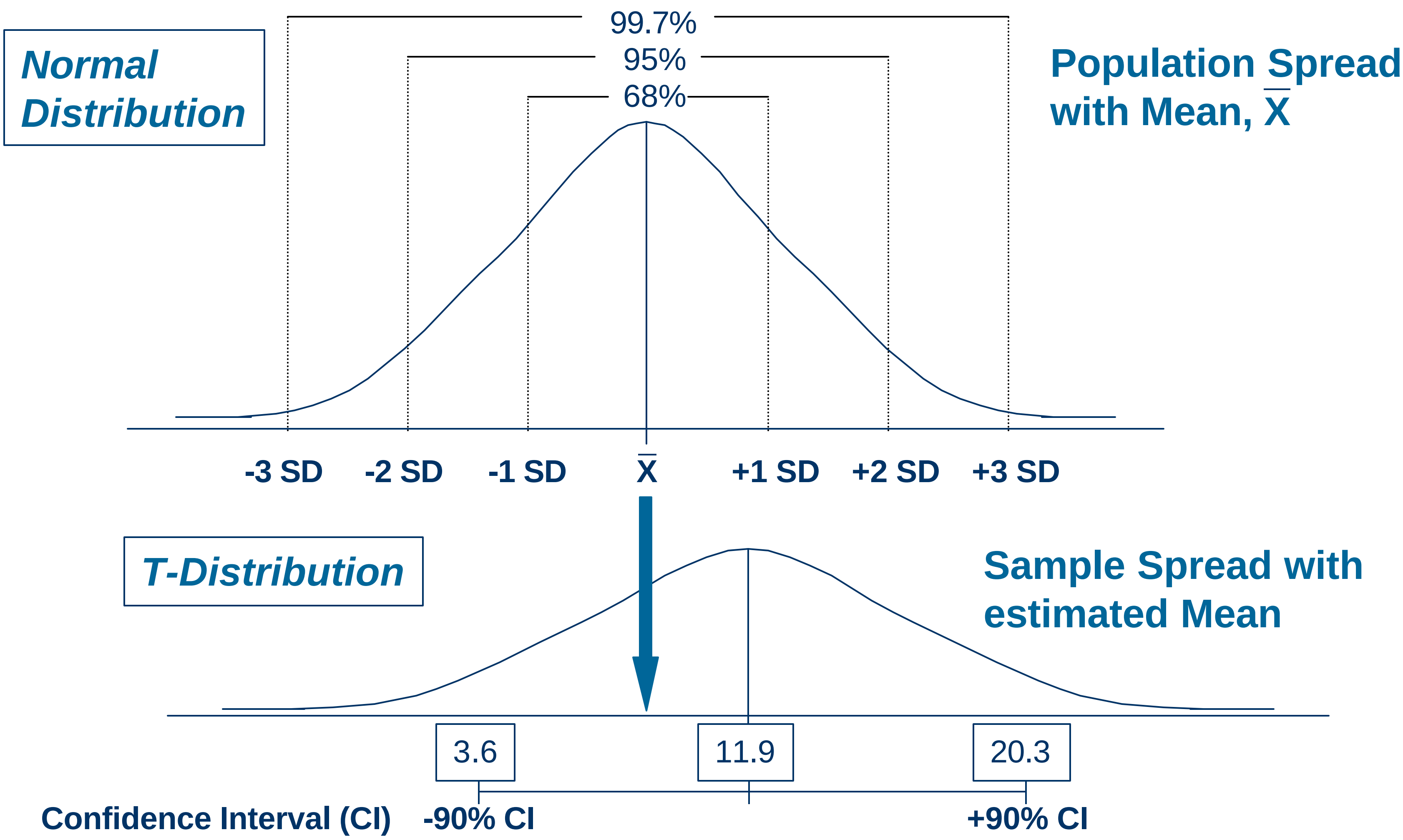
Effort Consumption = 11.9 Person Hours / Function Point

What does this mean?

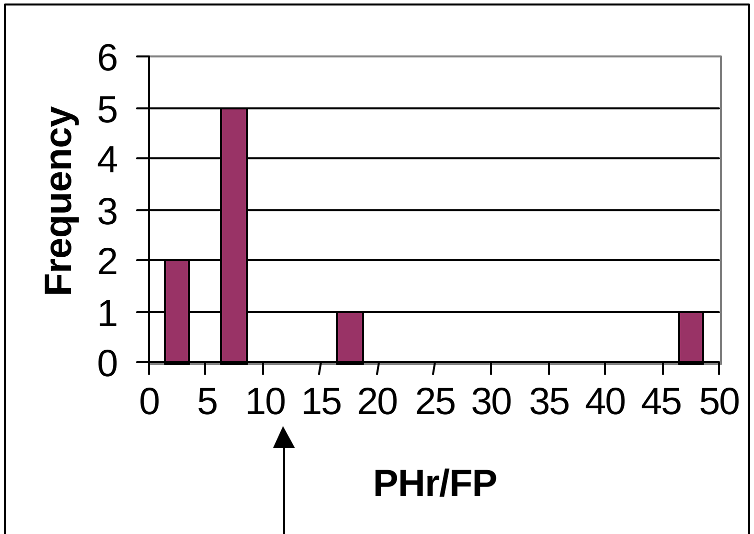
Actual Effort	Estimated Effort	=	Function Points	*	Effort Consumption
165	880.6	=	74	*	11.9
14,080	3,665.2	=	308	*	11.9
3,602	5,057.5	=	425	*	11.9

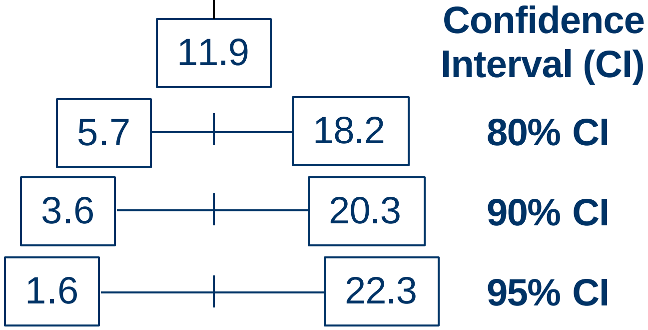
Yikes!

Accuracy and Precision



Data Analysis: PHr/FP



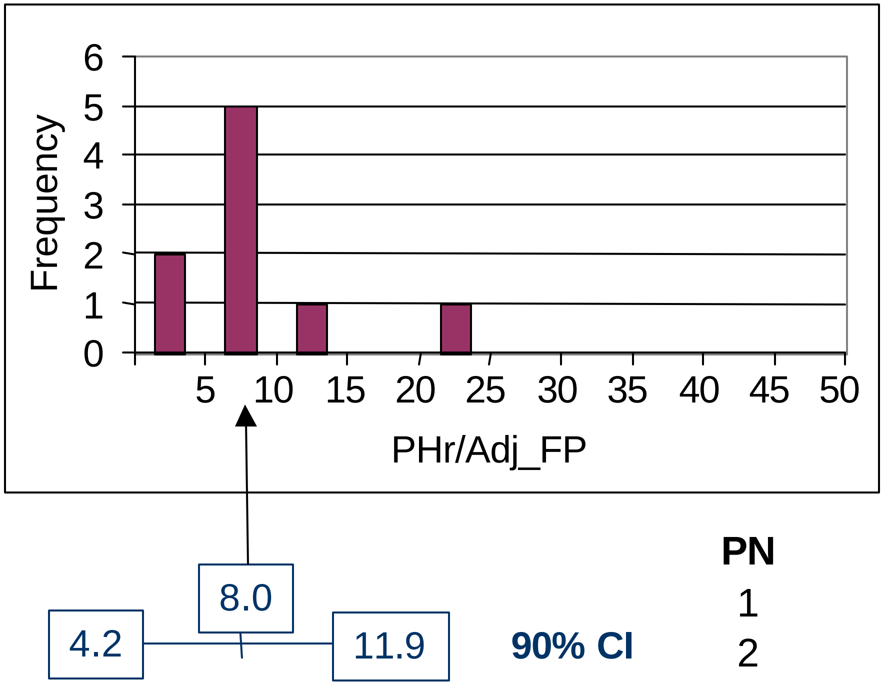


How can the CI be reduced?

PHr/FP	
Mean	11.92
Standard Error	4.48
Median	7.50
Standard Deviation	13.44
Range	43.48
Minimum	2.23
Maximum	45.71
Confidence Level(90.0%)	8.33

PN	FP	PHrs	PHr/FP
1	40	300	7.50
2	931	6,400	6.87
3	425	3,602	8.48
4	181	1,550	8.56
5	308	14,080	45.71
6	163	1,090	6.69
7	74	165	2.23
8	333	1,070	3.21
9	241	4,350	18.05

Data Analysis: PHr/Adj_FP



90% CI

PHr/Adj_FP	
Mean	8.01
Standard Error	2.07
Median	6.62
Standard Deviation	6.21
Range	20.81
Minimum	2.05
Maximum	22.86
Confidence Level(90.0%)	3.85

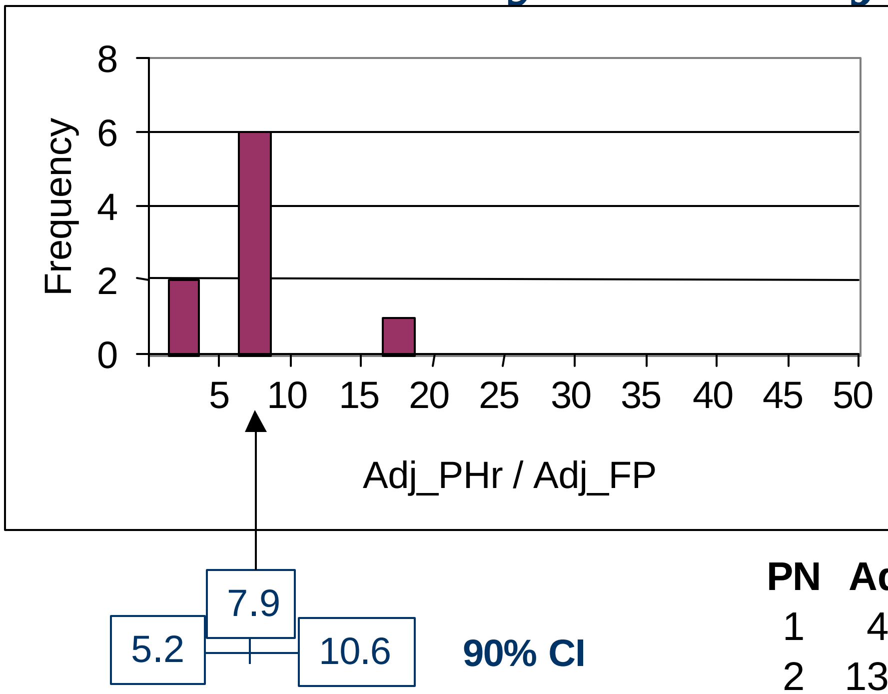
Assignable cause of variation:

Adjust the size with the effects of requirement's volatility (REVL)

Adj FP = FP * (1 + REVL%)

PN	FP	REVL	Adj_FP	PHrs	PHr/Adj_FP
1	40		40.00	300	7.50
2	931	50	1396.50	6,400	4.58
3	425	30	552.50	3,602	6.52
4	181	10	199.10	1,550	7.79
5	308	100	616.00	14,080	22.86
6	163	1	164.63	1,090	6.62
7	74	9	80.66	165	2.05
8	333	10	366.30	1,070	2.92
9	241	60	385.60	4,350	11.28

Data Analysis: Adj_PHr/Adj_FP



90% CI

Adj_PHr/Adj_FP	
Mean	7.87
Standard Error	1.46
Median	8.05
Standard Deviation	4.39
Range	15.19
Minimum	2.53
Maximum	17.72
Confidence Level(90.0%)	2.72

PN	Adj_FP	PHrs	PCON	Adj_PHrs	Adj_PHr / Adj_FP
1	40.00	300	0.90	333.33	8.33
2	1396.50	6,400	0.81	7901.23	5.66
3	552.50	3,602	0.81	4446.91	8.05
4	199.10	1,550	0.81	1913.58	9.61
5	616.00	14,080	1.29	10914.73	17.72
6	164.63	1,090	1.00	1090.00	6.62
7	80.66	165	0.81	203.70	2.53
8	366.30	1,070	0.81	1320.99	3.61
9	385.60	4,350	1.29	3372.09	8.75

Assignable cause of variation: Adjust the effort with the effects of Personnel Continuity (PCON)

Adj_PHr = PHr / PCON

Models Depend on Solid Data

- Models are created from data $\equiv$ Models are only as good as the data used to create them
  - life-cycle phase
  - overtime to get work done
  - experience
  - tools
  - complexity
  - reuse
- Data used to create models must be well specified

PSM Measurement Specifications

- Staff Turnover Specification Guidance
  - Typical Data Items
    - Number of personnel
    - Number of personnel gained (per period)
    - Number of personnel lost (per period)
  - Typical Attributes
    - Experience factor
    - Organization
  - Typical Aggregation Structure
    - Activity
  - Typically Collected for Each
    - Project
  - Count Actuals Based On
    - Financial reporting criteria
    - Organization restructuring or new organizational chart

Staff Turnover

Impact of Personnel Continuity on Effort

This factor captures the turmoil caused by the project losing key, lead personnel. The loss of key personnel leads to extra effort in new people coming to work for the project and having to spend time coming up to speed on what has to be done. The rating scale is in terms of the project's personnel turnover normalized to a year.

Descriptors:	48% per year	24% per year	12% per year	6% per year	3% per year
Rating Levels	Very Low	Low	Nominal	High	Very High
Effort Multipliers	1.29	1.12	1.00	0.90	0.81

Effect on Effort: +15% +12% -11% -11%

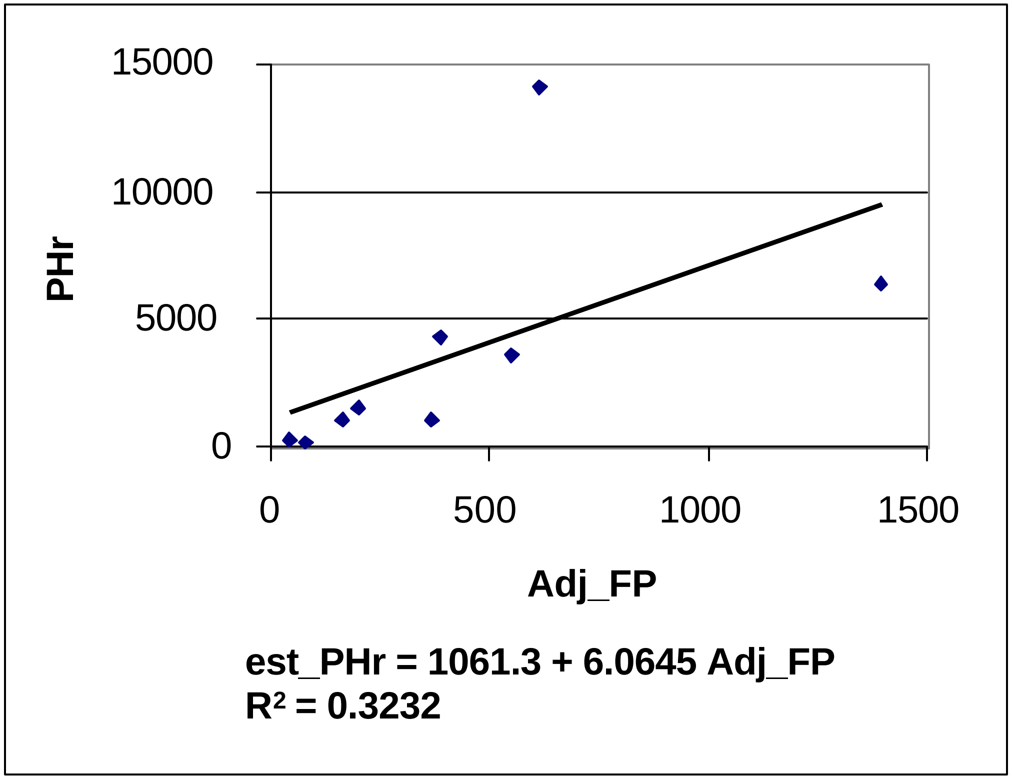
Source: Software Cost Estimation with COCOMO II, Boehm et. al.

One More Model: Linear Regression Analysis

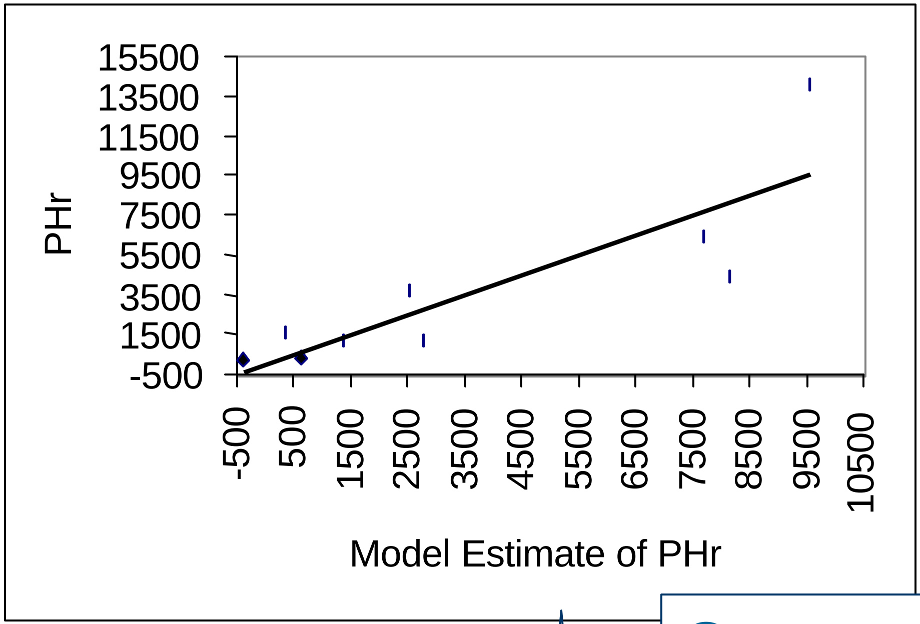
My favorite!

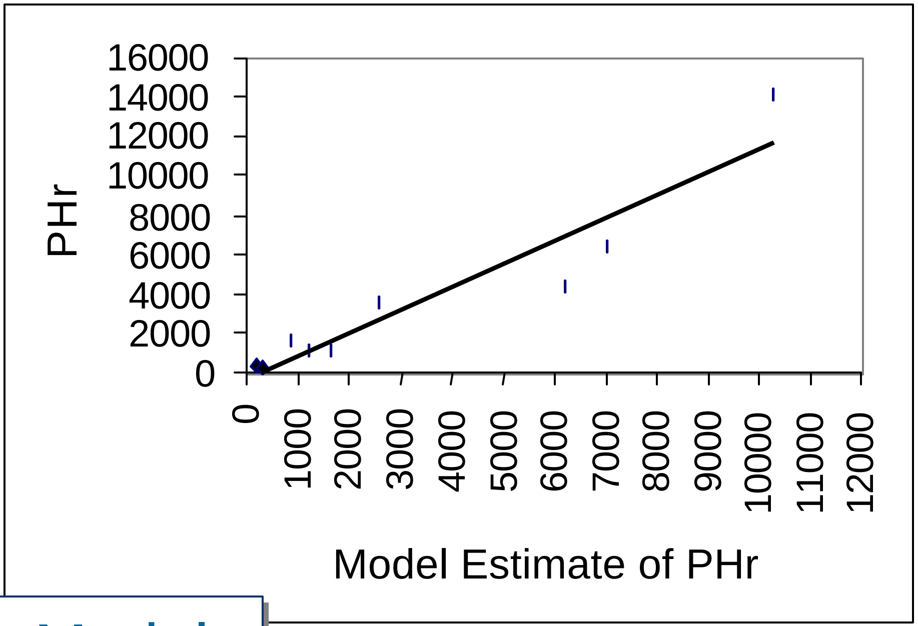
- **Statistical Regression fits a line through points minimizing the least square error between the points and the line**
- **The regression analysis yields a line with a slope, M, and intercept, A:**

  $Y = A + MX$
- **The goodness of fit is given by a statistic called R^2. The closer to 1.0, the better the fit.**



Regression Analysis Example





Compare Models

est_PHr = -12127 + 6.16 Adj_FP + 13879 PCON
Adj. R^2 = 0.64

est_PHr = 4.84* Adj_FP$^{1.08}$ * PCON$^{2.72}$
Adj. R^2 = 0.88

Model Accuracy

PN	PHr/FP	Adj_PHr / Adj_FP	Linear Model	Multiplicative Model	Actual PHrs
1	476.80	314.80	612.26	200.76	300
2	11,097.52	7,326.97	7,703.17	7,026.10	6,400
3	5,066.00	3,344.75	2,512.57	2,570.26	3,602
4	2,157.52	1,424.47	339.16	849.69	1,550
5	3,671.36	2,423.96	9,565.01	10,274.27	14,080
6	1,942.96	1,282.81	2,764.17	1,227.46	1,090
7	882.08	582.38	-389.25	318.92	165
8	3,969.36	2,620.71	1,367.44	1,645.88	1,070
9	2,872.72	1,896.67	8,148.05	6,181.82	4,350
PRED(.30)	0.0	0.55	0.33	0.44	

PRED(L) = X means that the model estimates within L% of the actual values X% of the time

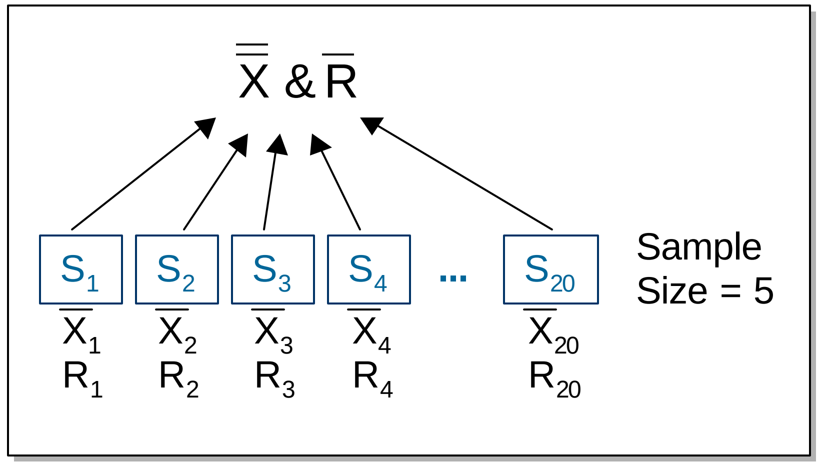
Another Model: Statistical Process Control

- Application of statistics in the area of quality control
  - important in manufactured goods
  - vital to service operations
- Variation in quality
  - *common causes*: no two outputs from any production process are exactly alike
  - *assignable causes*: sporadic changes that can be identified and eliminated or explained
- Use SPC to:
  - identify special causes and correct them
  - not react to common causes over which we have no control

SPC Indicator

- **Samples a *single* variable or attribute that represents process performance.**
- **Uses the Mean of sample-means or the mean of sample Ranges to determine if a process is “in” or “out of control”**

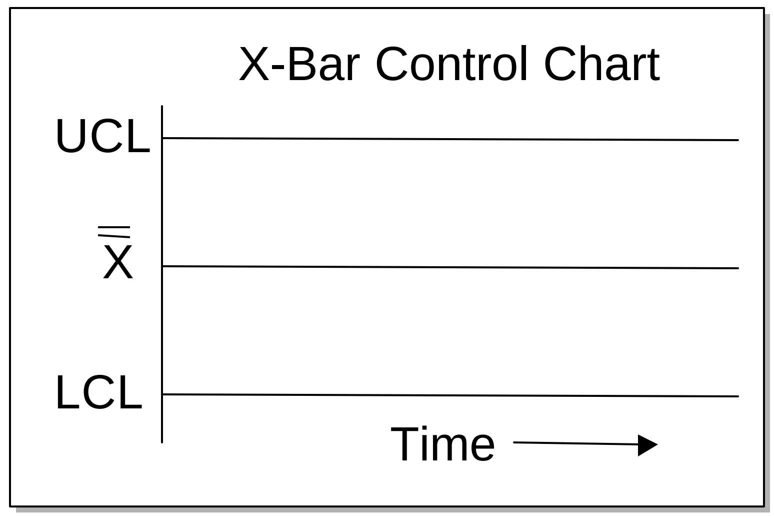
$\bar{\bar{X}}$	Mean of sample-means
R	Range of samples



- **An indicator called a “control chart” is used to show some aspect of process behavior over time**

Upper Control Limits (UCL) = $\bar{\bar{X}} + c\,R$

Lower Control Limits (LCL) = $\bar{\bar{X}} - c\,R$



Control Chart types

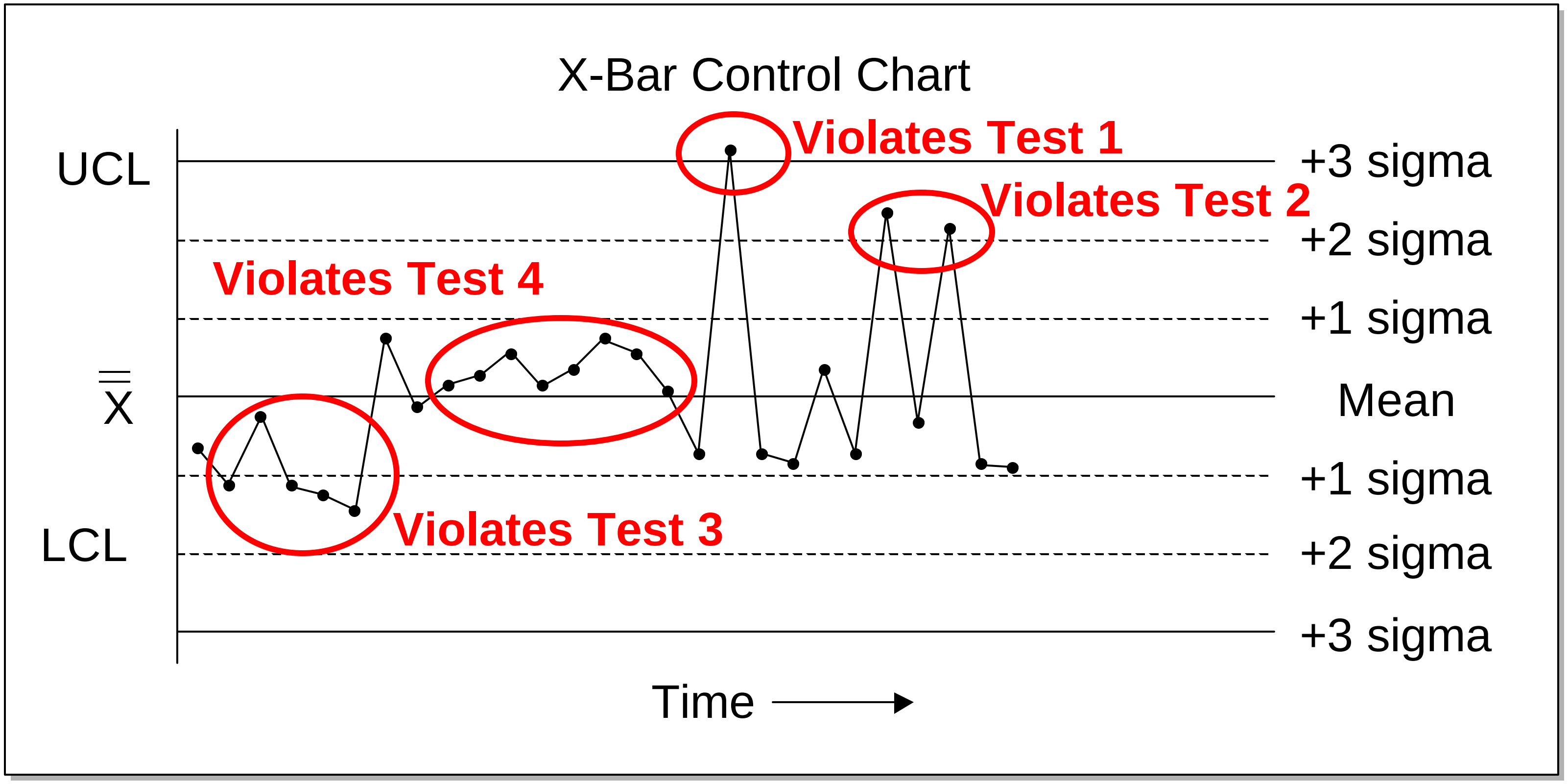
- X-Bar chart
  - Shows the mean of the process performance attributes
- R chart
  - Measures the amount of spread in process performance
- Attributes charts (P, np, c, & u)
  - Based on theoretical models (e.g. Binomial or Poisson distributions) to compute limits about the process mean
- Individual charts (XmR)
  - Show the mean, X, and variation, mR, of single point samples of process variables. Not as sensitive as X-Bar or R charts in detecting assignable causes.

SPC Indicator Analysis

- In-Control Processes
  - variation in output due to common causes, process is stable and predictable within a range around a mean
- Out-Of-Control Processes
  - variation in output due to assignable causes, unpredictable due to change in distribution
- Detecting Out-Of-Control situations
  - Test 1: A single point falls outside the UCL or LCL
  - Test 2: At least 2 out of 3 successive values fall on the same side of, and more than 2 sigma away from, the centerline
  - Test 3: At least 4 out of 5 successive values fall on the same side of, and more than 1 sigma away from, the centerline
  - Test 4: At least 8 successive values fall on the same side of the centerline

Source: Measuring the Software Process, Statistical Process Control for Software Process Improvement by Florac and Carleton

Indicator Ex. of Unstable Process



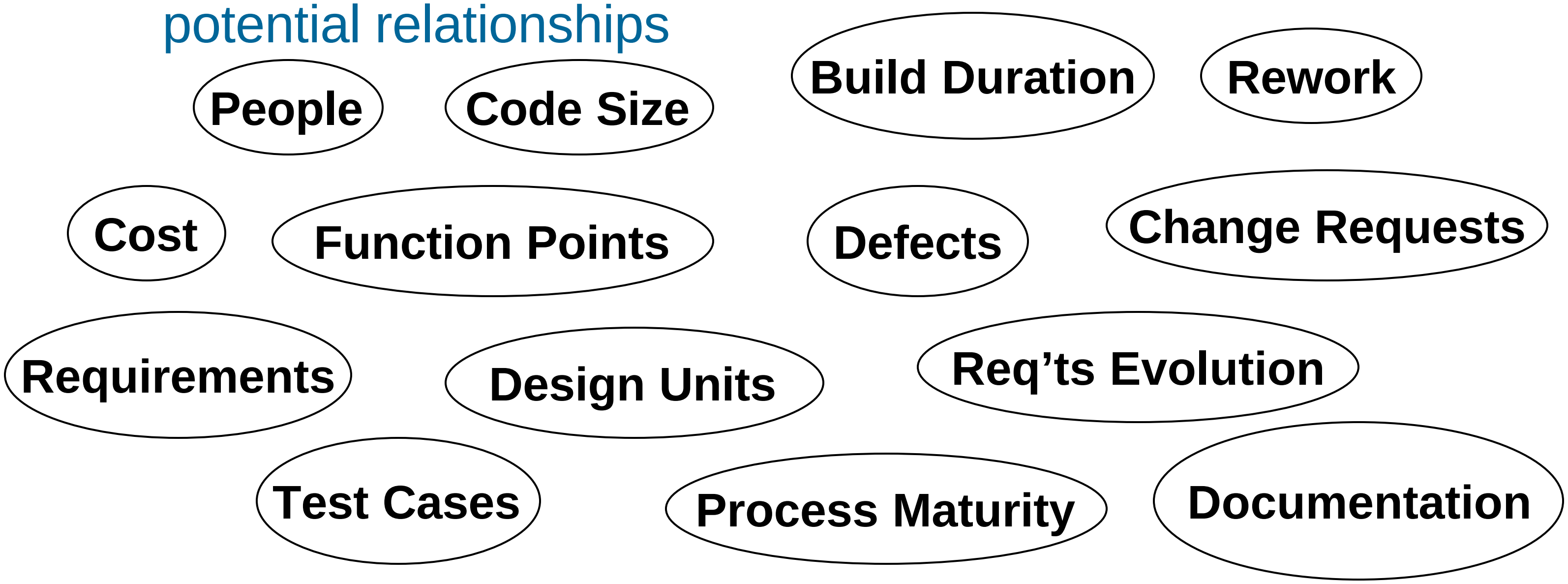
Source: Measuring the Software Process, Statistical Process Control for Software Process Improvement by Florac and Carleton

Can SPC be Used for Software Processes?

- Statistical Process Control is used on the same manufacturing process occurring many times concurrently
- Data requirement is for a large number of samples
  - 20 to 25 samples of size 5 = 100 to 150 samples total
- Items of interest in software processes have many assignable causes of variation - these must be eliminated, e.g.
  - Experience
  - Product complexity
  - Personnel turnover

Summary -1

- Statistical models describe “what is” and not “what should be”
  - They are simplified representations of reality
  - They formalizes a relationship
  - In Software and Systems Engineering, there are many potential relationships



Summary -2

- Good models depend on lots of good data
- Data attributes can be useful in reducing variation in the data
- Thinking about what causes variation is a good way to pick attributes to collect
- The *mean* is a model that describes data “on average”
- The *standard deviation* is a model that describes distances “in general”

Further Information

- Measuring the Software Process, Statistical Process Control for Software Process Improvement, by William Florac and Anita Carleton, Addison-Wesley, 1999
- Software Cost Estimation with COCOMO II by Barry Boehm, Chris Abts, Winsor Brown, Sunita Chulani, Brad Clark, Ellis Horowitz, Ray Madachy, Donald Reifer, and Bert Steece, Prentice Hall PTR, 2000.
- Statistics, Data Analysis, and Decision Making, by James Evans and David Olson, Prentice-Hall, 1999
- Statistical Analysis Simplified, by Glen Hoffherr and Robert Reid, McGraw-Hill, 1997

Contact Information

Brad Clark

Software Metrics, Inc.

Washington, D.C. area

(703) 754-0115

Brad@Software-Metrics.com

http://www.software-metrics.com