Scientific Test and Analysis Techniques (STAT)

ITEA MDO
El Paso, TX
July 19, 2022

Participant Guide

Mark Kiemele, Ph.D.
Air Academy Associates
12295 Oracle Blvd, Ste 340
Colorado Springs, CO 80921
Phone: 719-531-0777
email: aaa@airacad.com

22-ITEASTAT-7A

www.airacad.com

Introductions

- Name
- Position (where are you from and what do you do)
- Experience in test design and data analysis
- Expectations

Session Guidelines and Information

- Materials
- Cell phones
- Start and stop times
- Breaks
- Active participation – questions are good / discussion is great / involvement is wonderful
- Importance of applying right away

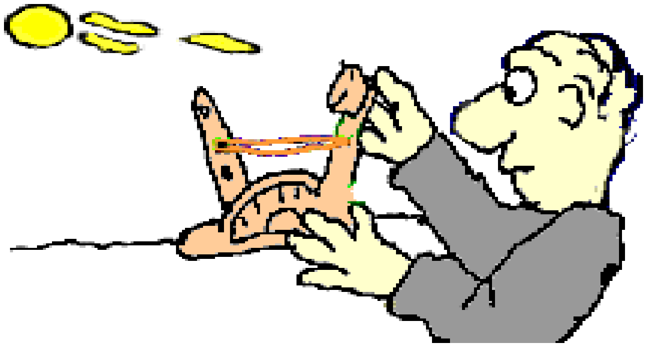
Agenda

- Necessary pre-requisites for testing
  - First Line of Defense Against Variation (PF/CE/CNX/SOP)
  - Measurement System Analysis (MSA)
- Hypothesis Testing
- Design of Experiments (DOE)
- Regression Analysis

Catapulting Power into Knowledge Gain

Statapult ® Catapult

Gathering Baseline Data



Using the statapult as a metaphor

- Statapult = Delivery Process
- Ball = Service Provided
- Setup = Job prep
- Measurement = Outcome of the service

Rapid-Fire Statapult® Exercise #1

Each team member will shoot the Statapult® X times using the following steps:

(1) Insure all pins are at position #3

(2) Pull the arm to 177° and launch the rubber ball

(3) Have someone measure your distance

(4) Disconnect rubber band between shots

(5) Your standard is no more than 15 seconds between shots

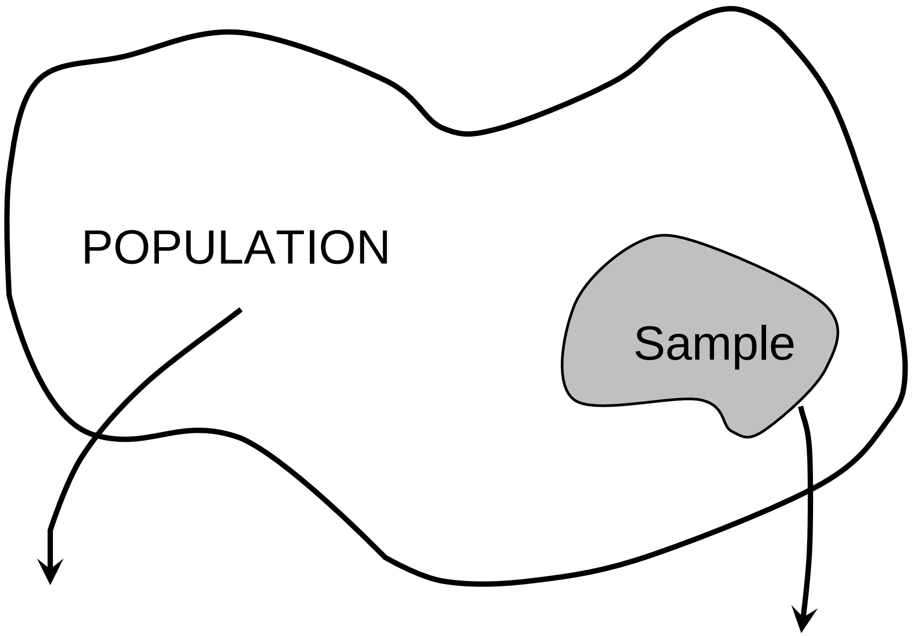
(6) Record distances; Calculate Range

Team Member:

	#1	#2	#3	#4	#5	#6	#7	#8
Shot #1	______	______	______	______	______	______	______	______
Shot #2	______	______	______	______	______	______	______	______
Shot #3	______	______	______	______	______	______	______	______
Shot #4	______	______	______	______	______	______	______	______
Shot #5	______	______	______	______	______	______	______	______

Range = Longest - Shortest = __________

Some Basic Definitions

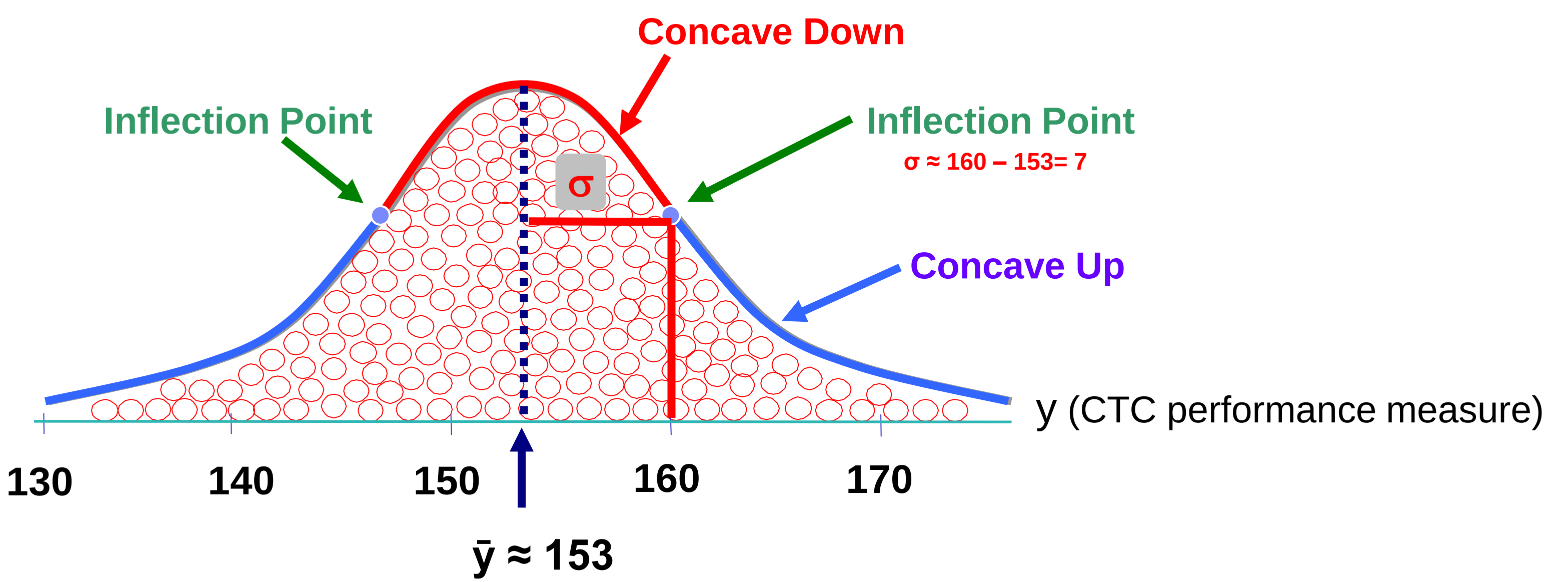


Population Parameters		Sample Statistics
μ	**Mean**	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
σ^2	**Variance**	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$
σ	**Standard Deviation**	$s = \sqrt{s^2}$

Graphical Meaning of $\bar{y}$ and σ
(for continuous data)

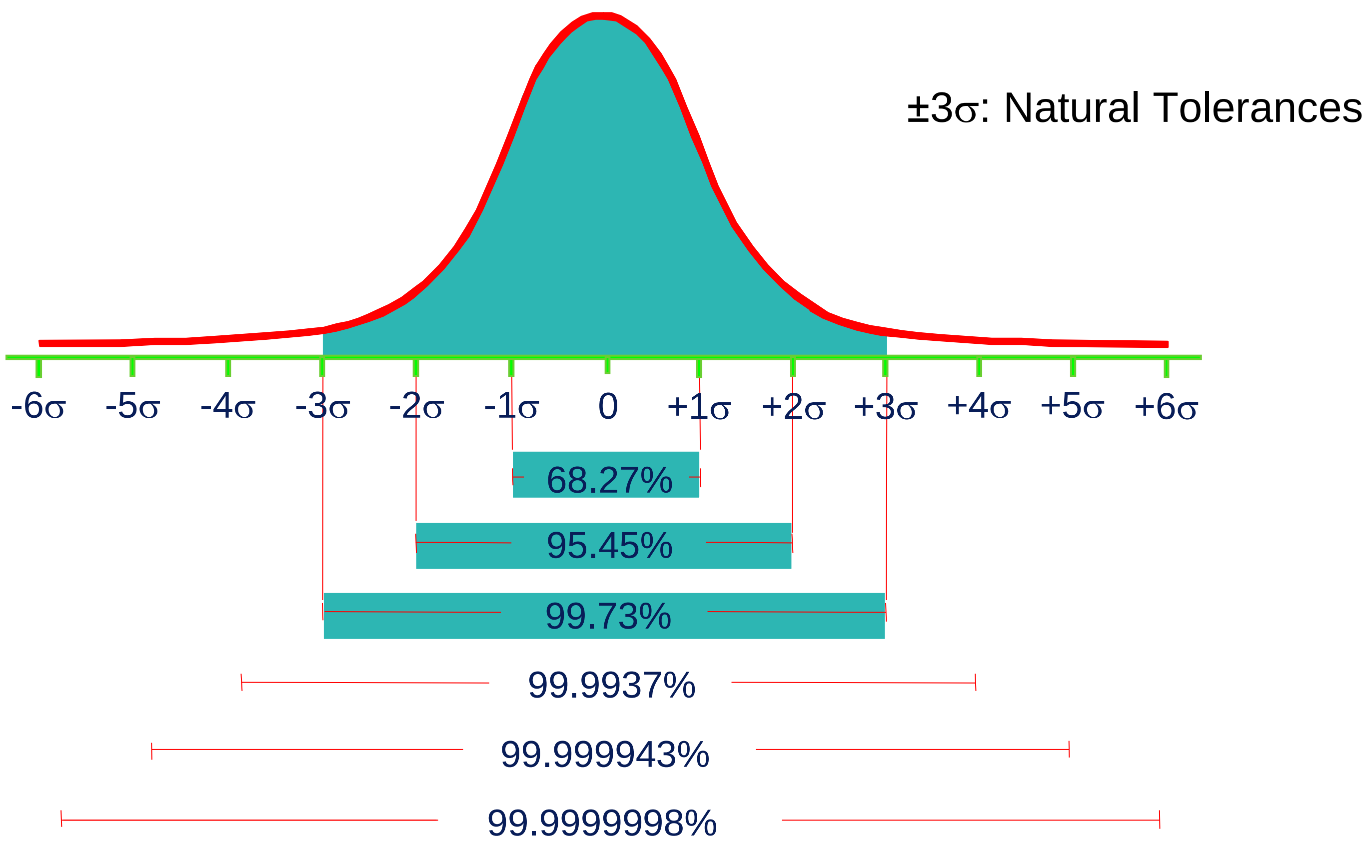
$\bar{y}$ = Average = Mean = Balance Point

σ = Standard Deviation



σ ≈ average distance of points from the centerline

Graphical View of Variation



Typical Areas under the Normal Curve

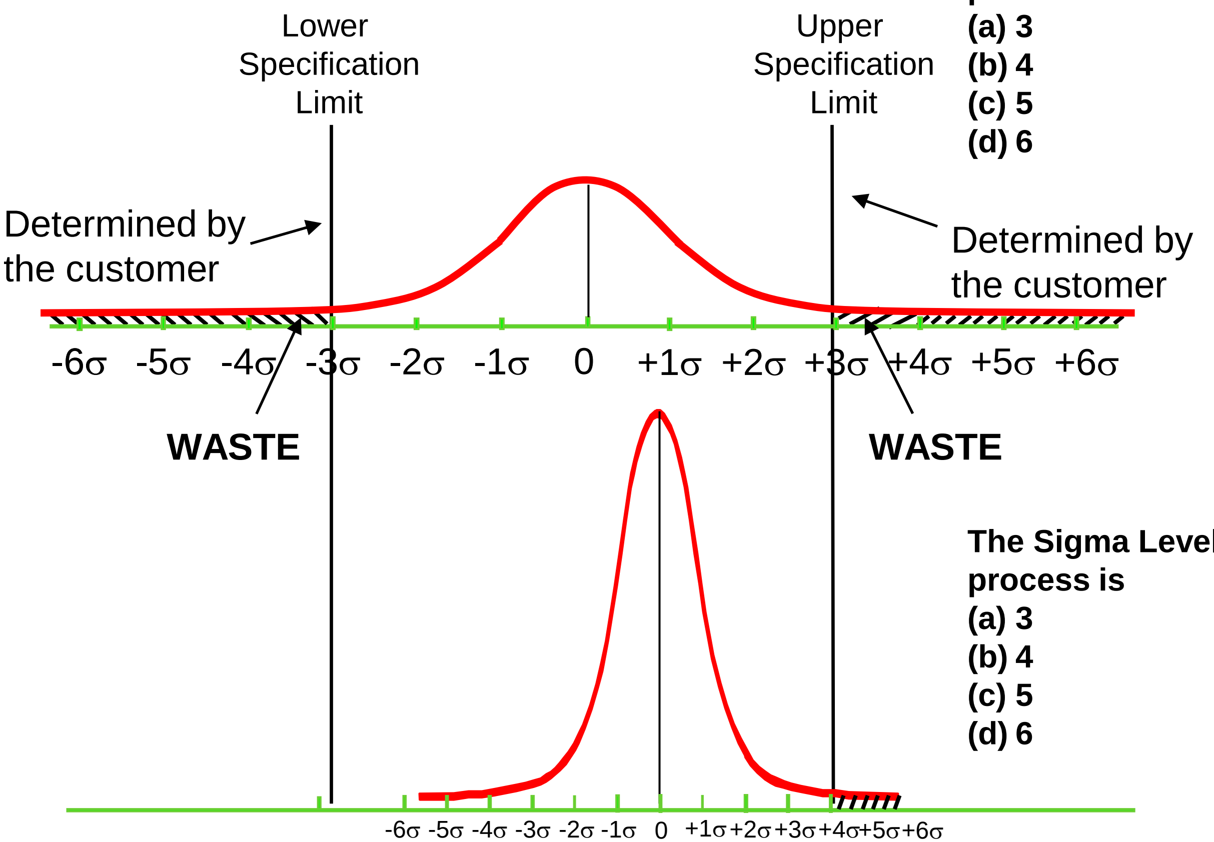
The Normal Distribution is even on the money

Graphical View of Variation and Process Capability
(for continuous data)

The ***Sigma Level*** of a process performance measure is the result of comparing the **Voice of the Process** with the **Voice of the Customer**, and it is defined as follows:

The **number of Sigmas** between the center of a process performance measure's distribution and the nearest specification limit

The Sigma Level of this process is
(a) 3
(b) 4
(c) 5
(d) 6



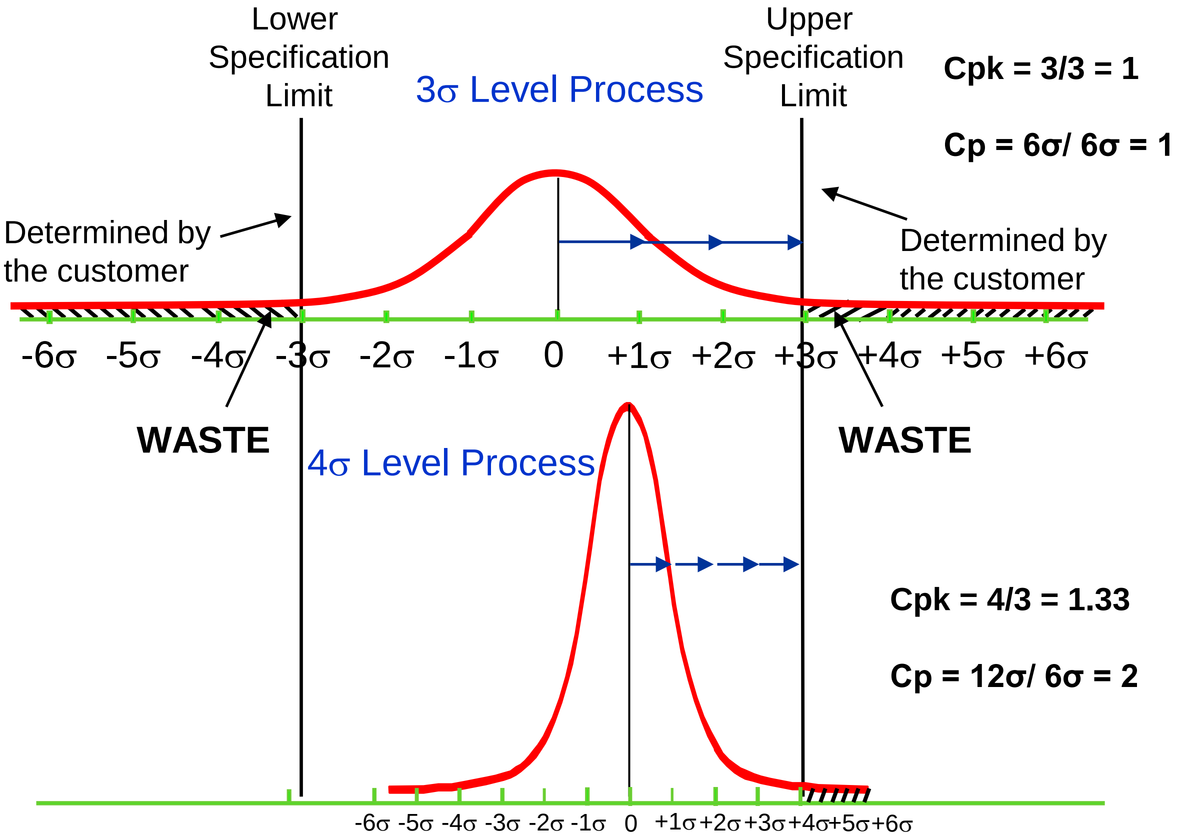
The Sigma Level of this process is
(a) 3
(b) 4
(c) 5
(d) 6

Other Measures of Process Capability
(for continuous variables)

Cpk = Sigma Level/3

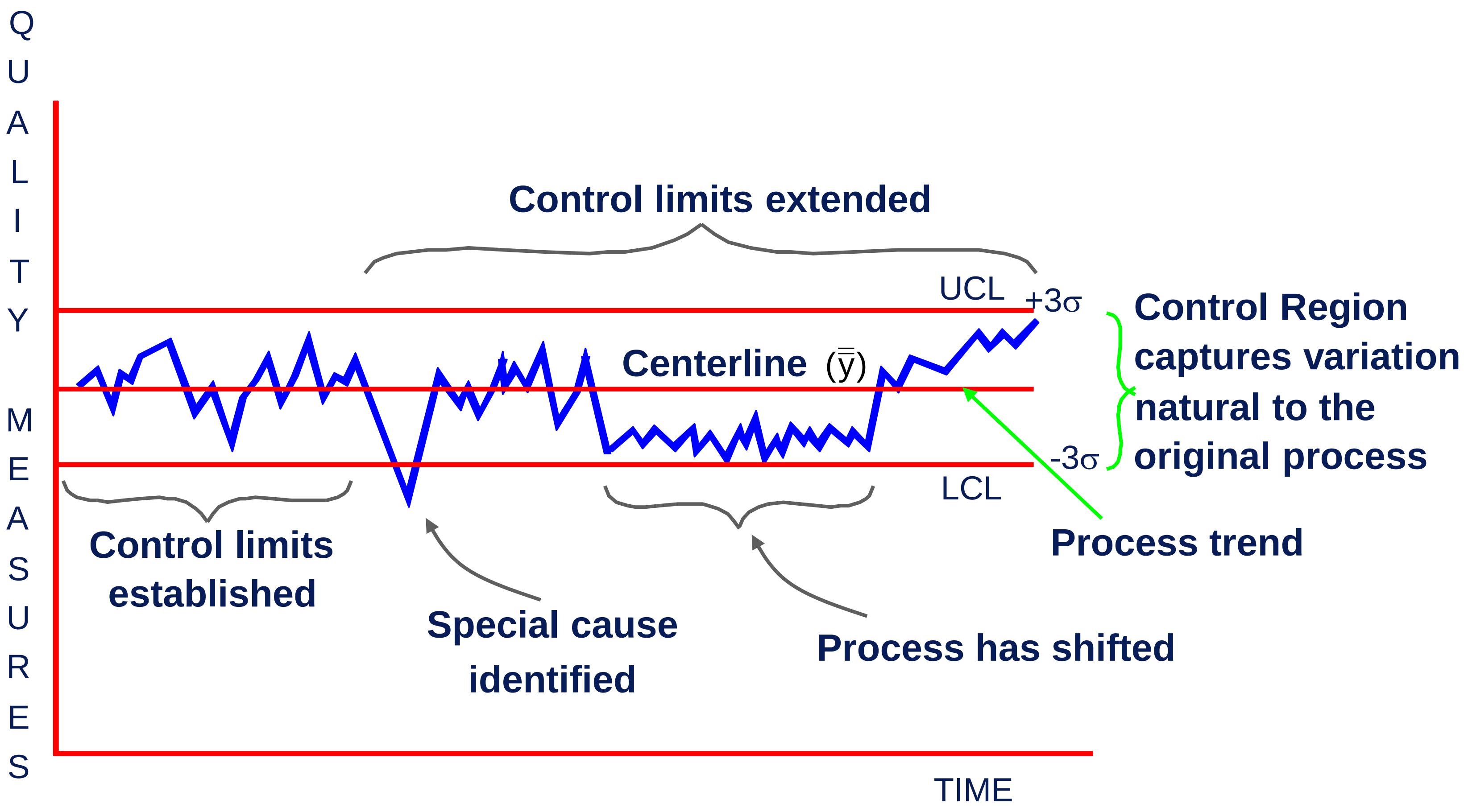
Cp = (USL – LSL)/6σ = Spec Width/Process Width

Note: Cpk = Cp whenever the process is centered between the two specs



Control Chart

… represents the "Voice of the Process"



Variation …the “Insidious” Enemy

- Variation leads to waste and poor quality
- Examples of variation:
  - Inputs to processes are inconsistent
  - Process steps are not performed the same way each time
  - Products and services vary in quality
- Impact of variation:
  - Customer sees and feels the variation (doesn’t know what to expect)
  - Customer never sees the average
  - Mistakes or defects (services not performed in accordance with customer specifications)
  - Commit dates are missed
  - Cost overruns
  - Need for excess capacity
  - Impact on lead times

How do we measure variation?

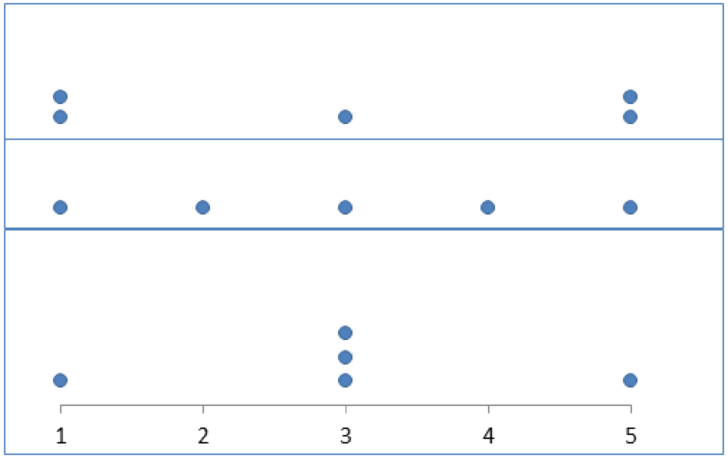
- Standard Deviation (Sigma = σ) is the best measure.
- Range is another measure of variation but not as good as σ.
- Example (all data sets have a sample size of n = 5, mean of 3 and range of 4):
  - Data Set 1: 1, 1, 3, 5, 5 Mean = 3, Range = 4, σ = ?
  - Data Set 2: 1, 2, 3, 4, 5 Mean = 3, Range = 4, σ = ?
  - Data Set 3: 1, 3, 3, 3, 5 Mean = 3, Range = 4, σ = ?

Dot Plots

Data Set 1: $\sigma = 2$

Data Set 2: $\sigma \approx 1.6$

Data Set 3: $\sigma \approx 1.4$



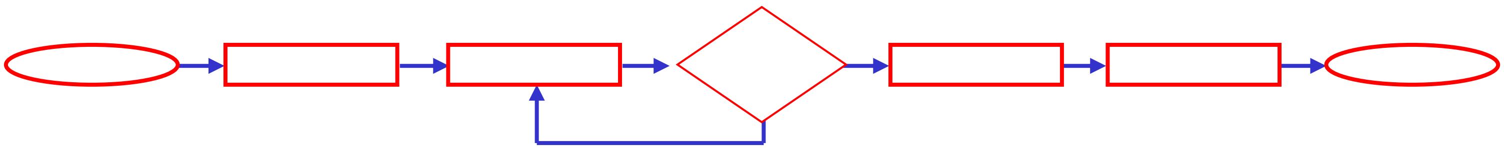
- In words, the standard deviation (or σ) of a data set is a measure of the variability of the values in the data set. Specifically, it measures how far the values "deviate" (on average) from the mean which graphically represents the center or balance point of the data.

PF/CE/CNX/SOP

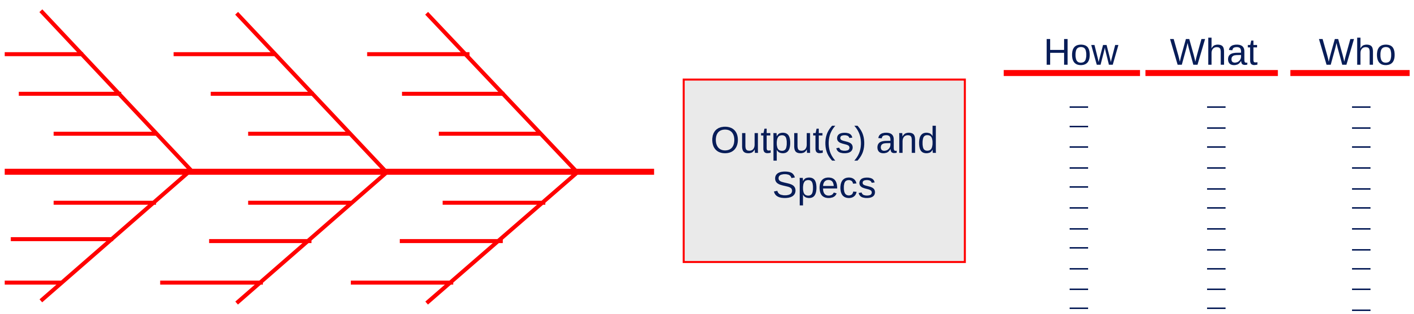
A Rapid Improvement Event and a Management Tool that

... removes waste, reduces variation, and decreases cycle time

(1) PROCESS FLOW (PF) OR PROCESS MAP



(2) CUSTOMER DRIVEN CAUSE AND EFFECT (CE)



C = Constants ⟵ Standard Operating Procedures (SOPs)
N = Noise
X = Experimental

(3) Partitioning the Variables into CNX

C = Controlled

- To hold a variable as constant as possible requires controlling the variable via Mistake Proofing and SOPs to eliminate errors and reduce variation.
- Controlling a variable or holding it as constant as possible doesn't just happen. It must be “engineered” into the process.
- Mistake Proofing: The process of eliminating conditions that lead to variation in the CTCs and ultimately cause errors.

N = Noise

- Noise variables are those that are not being controlled or held as constant as possible
- Mistake Proofing is needed to change an "N" variable to a "C" variable.

X = Experimental

- These are key variables that can be controlled and held constant at different levels or settings for the purpose of determining the effect of this variable on the CTC.

(4) Standard Operating Procedures (SOPs)

- Define the interaction of people and their environment when processing a product or service
- Detail the action and work sequence of the operator
- Provide a routine to achieve consistency of an operation
- Specify the best process we currently know and understand for controlling variation and eliminating waste
- Provide a basis for future improvements
- Validate mistake proofing in the process
- Strongly impact compliance to QMS such as ISO 9000, CMM, Sarbanes-Oxley, etc.

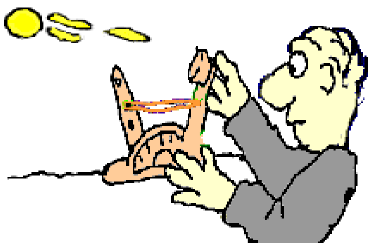
Statapult® Exercise #2: Reducing Variation

Process flow "Shooting the Statapult®"

Complete Cause-and-Effect Diagram

Label inputs as C or N

Use SOPs with mistake proofing to change N's into C's



Re-shoot Statapult® using the first example instructions (all pins at #3; pull angle = 177°; 15 sec. between shots; etc.)

Record data taken after PF/CE/CNX/SOPs and evaluate

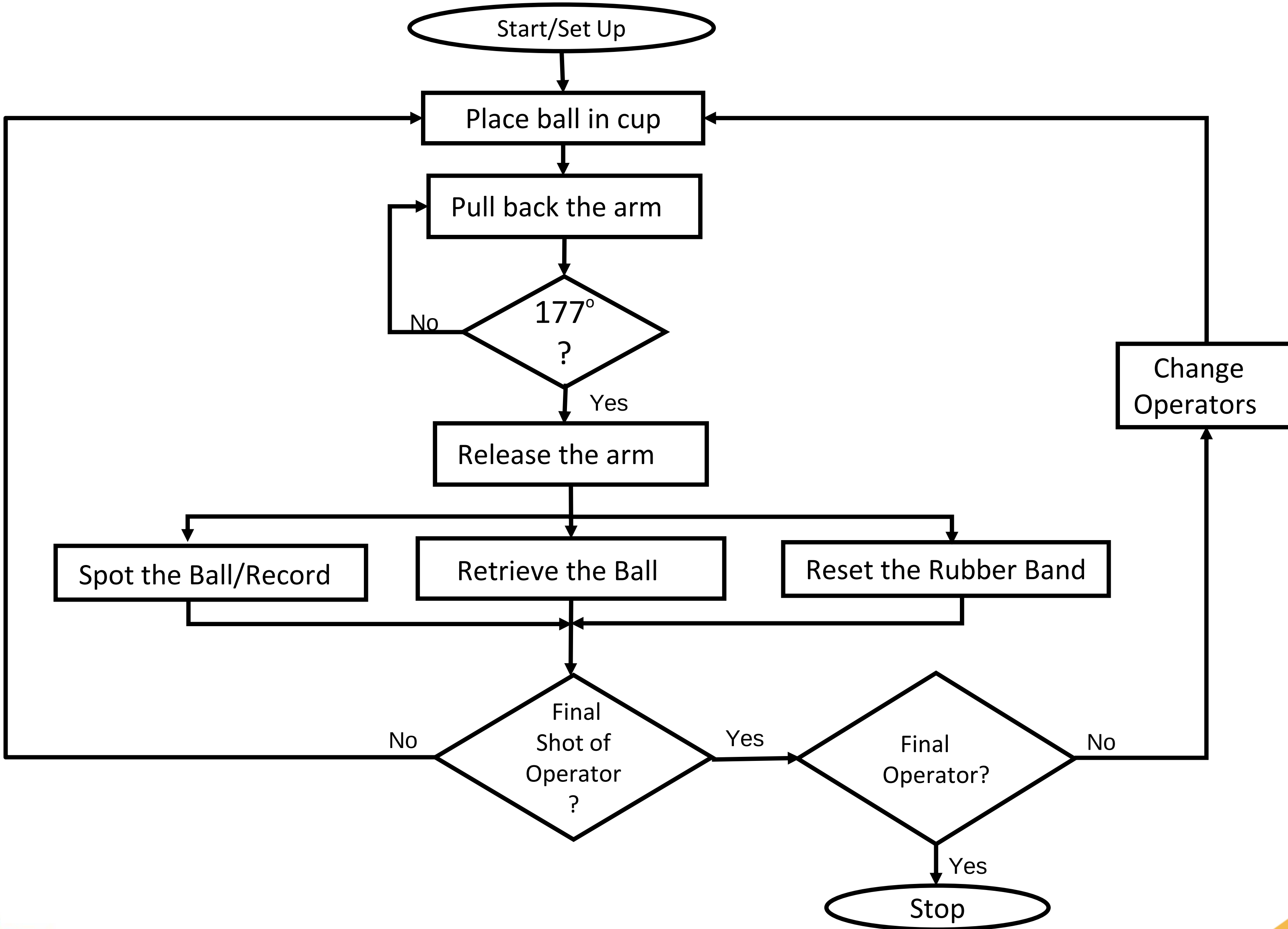
If improved, develop control plans

Team Member:

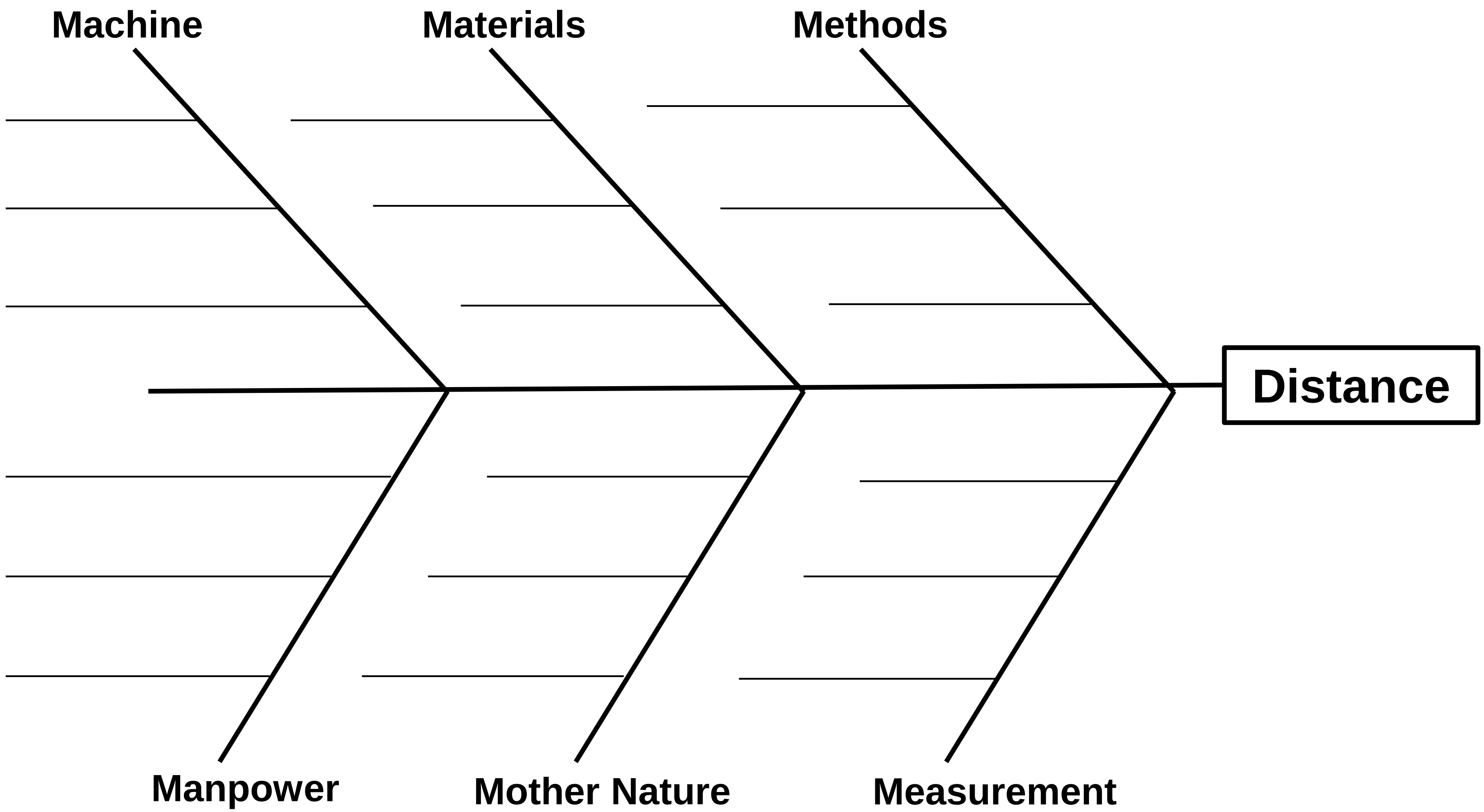
	#1	#2	#3	#4	#5	#6	#7	#8
Shot #1								
Shot #2								
Shot #3								
Shot #4								
Shot #5								

Range = Longest - Shortest = ____________

Process Flow



Cause-and-Effect Diagram Worksheet



Why do we need to measure?

To make better decisions, gauge true performance and determine if, in fact, our business objectives are being achieved.

"If you are not keeping score, you are just practicing."

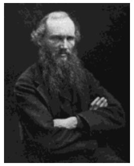
Vince Lombardi

Importance of Measurement
(Beginning of Process/Product Improvement)

- To assist with good decision making
- To identify/verify problem areas
- Because perception and intuition are not always reality
- To baseline performance
- To see if value streams and processes are improving

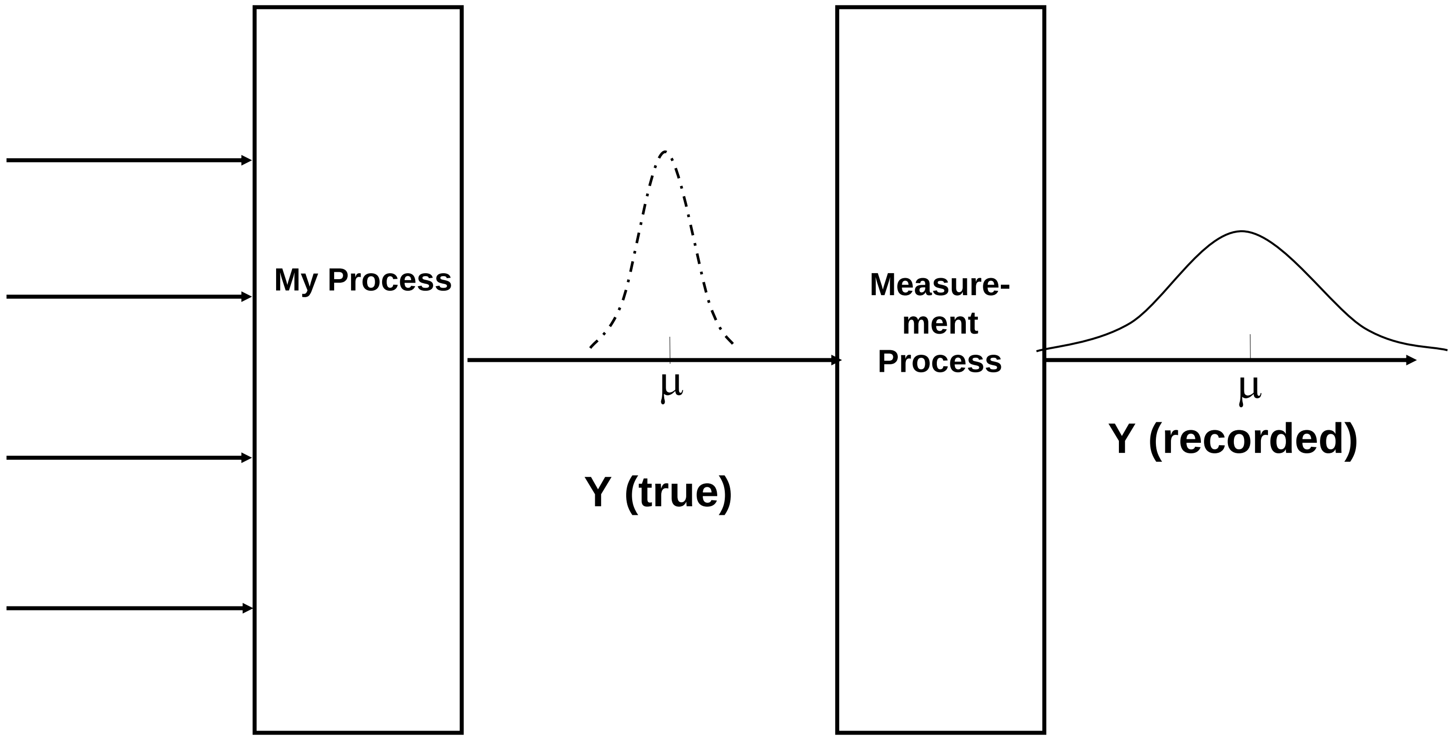
"To measure it, is to know."

"If you cannot measure it, you cannot improve it."



Lord Kelvin
(1821 – 1907)

Measurement is a Process



Where **DOES** the variation that we see in Y come from? Is it from the process itself or the measurement system? Which one of those two variations is stronger?

Is the measurement system –

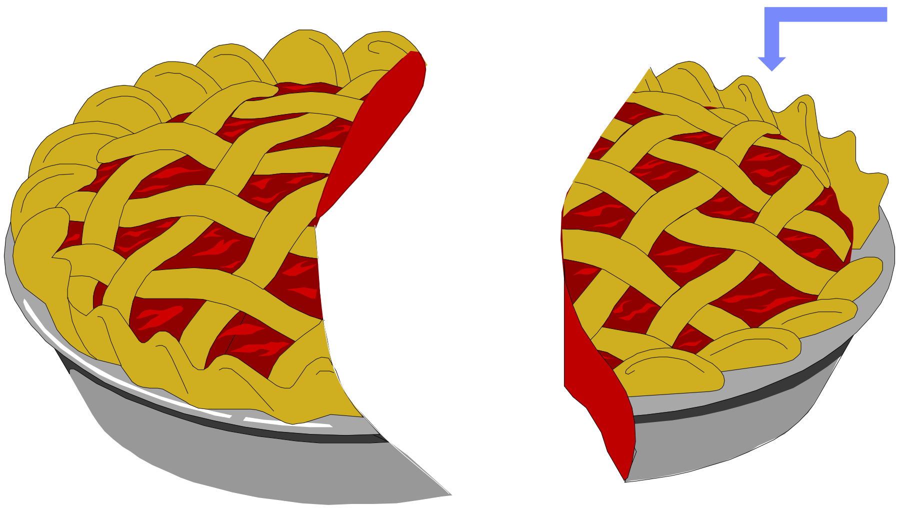
Accurate?

Precise?

Stable?

Measurement System Analysis (MSA)

Total (recorded) Variability is broken into two major pieces:



This piece of the pie will be further divided into two smaller pieces called Repeatability and Reproducibility.

Product / Service Variability

Measurement Variability

- **MSA identifies and quantifies the different sources of variation that affect a measurement system.**
- **Variation in measurements can be attributed to variation in the product/service itself or to variation in the measurement system.**
- **The variation in the measurement system itself is measurement error.**

Measurement Variability Broken Down Further

Purpose:

To assess how much variation is associated with the measurement system and to compare it to the total process variation or tolerances.

$$\sigma^2_{total} = \sigma^2_{product} + \sigma^2_{measurement}$$

$$\sigma^2_{measurement} = \sigma^2_{repeatability} + \sigma^2_{reproducibility}$$

Repeatability:

Variation obtained by the same person using the same procedure on the same product, transaction or service for repeated measurements (variability ***within*** operator).

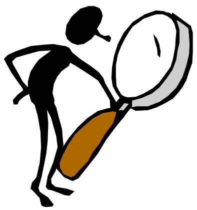
Reproducibility:

Variation obtained due to differences in people who are taking the measurements (variability ***between*** operators).

More on Repeatability and Reproducibility

For the scenario below, look at the data and indicate which of the following you think is true:

a. Repeatability appears to be more of a problem than reproducibility
b. Reproducibility appears to be more of a problem than repeatability
c. Repeatability and Reproducibility appear to be about the same



	Operator 1		Operator 2	
	Rep 1	Rep 2	Rep 1	Rep 2
Part 1	21	23	26	28
Part 2	19	18	24	24
Part 3	20	23	27	24
Part 4	19	22	21	20

- MSA will help <u>quantify</u>, more exactly, the capability of the measurement system and answer questions about repeatability, reproducibility, and capability with respect to the customer specs.

- **Rule of Thumb for MSA Sample Size:**
  - Variables (Continuous) Data: (# operators)*(number of parts) ≥ 20
  - Attribute (Binary) Data: (# operators)*(number of parts) ≥ 60

Measurement System Diagnostics

1. Precision-to-Tolerance Ratio (P/TOL)

$$P/TOL = \frac{6\sigma_{meas}}{USL - LSL}$$ (Specification Limits are needed)

ROT: If $P/TOL \leq .10$: Very Good Measurement System

$P/TOL \geq .30$: Unacceptable Measurement System

2. Precision-to-Total Ratio (P/TOT)

$$P/TOT = \frac{\sigma_{meas}}{\sigma_{total}}$$

ROT: If $P/TOT \leq .10$: Very Good Measurement System

$P/TOT \geq .30$: Unacceptable Measurement System

3. Discrimination or Resolution $= \left(\frac{\sigma_{product}}{\sigma_{meas}}\right) \times 1.41$

(# of truly distinct measurements that can be obtained by the measurement system) ROT: Resolution ≥ 5

Graphical View of Variance Components

MSA ANOVA Method Results

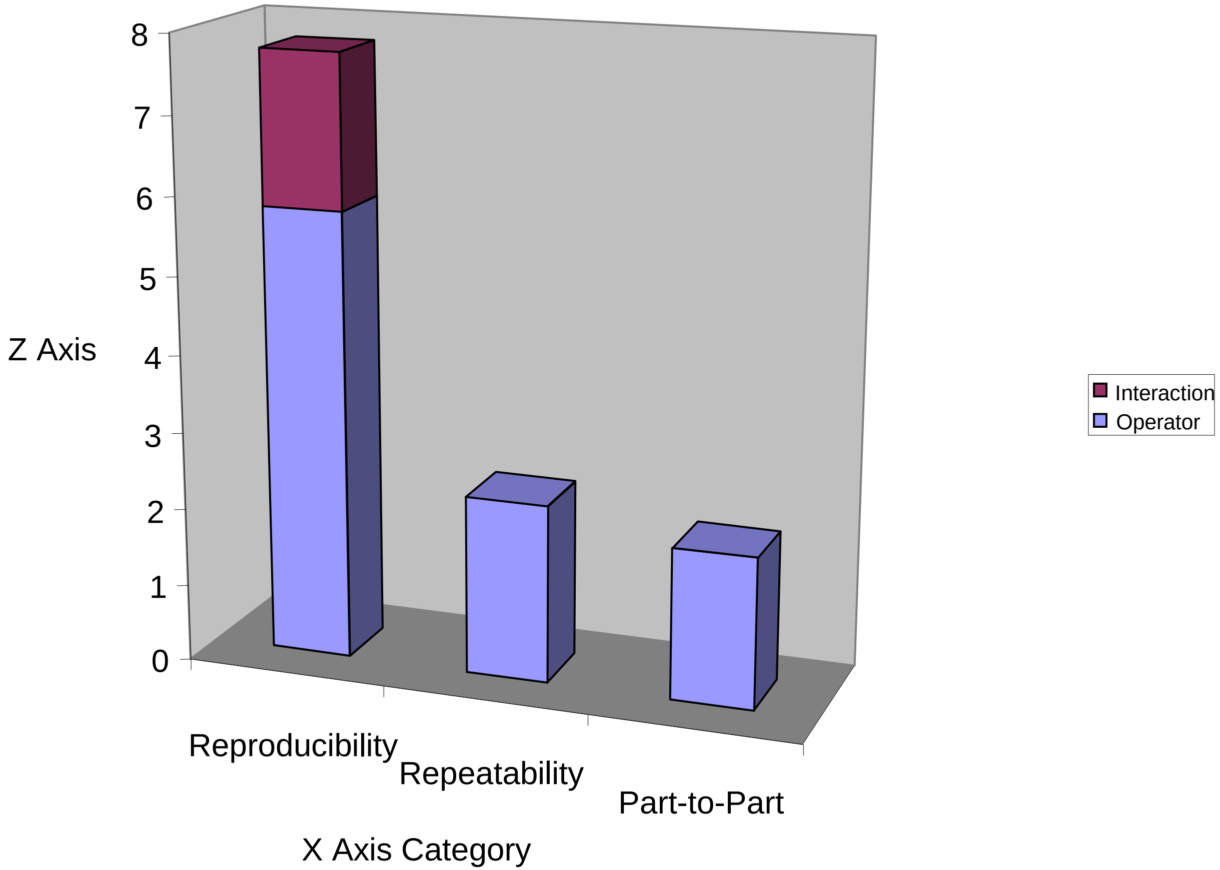
Source	Variance	Standard Deviation	% Contribution	p Value
Total Measurement (Gage)	10.0625	3.172144385	83.71%	
Repeatability	2.3125	1.520690633	19.24%	
Reproducibility	7.75	2.783882181	64.47%	
Operator	5.79166667	2.40658818	48.18%	
Oper * Part Interaction	1.95833333	1.399404635	16.29%	0.1167
Product (Part-to-Part)	1.95833333	1.399404635	16.29%	
Total	12.0208333	3.467107344	100.00%	

USL	35
LSL	10
Precision to Tolerance Ratio	0.76131465
Precision to Total Ratio	0.91492535
Resolution	0.6

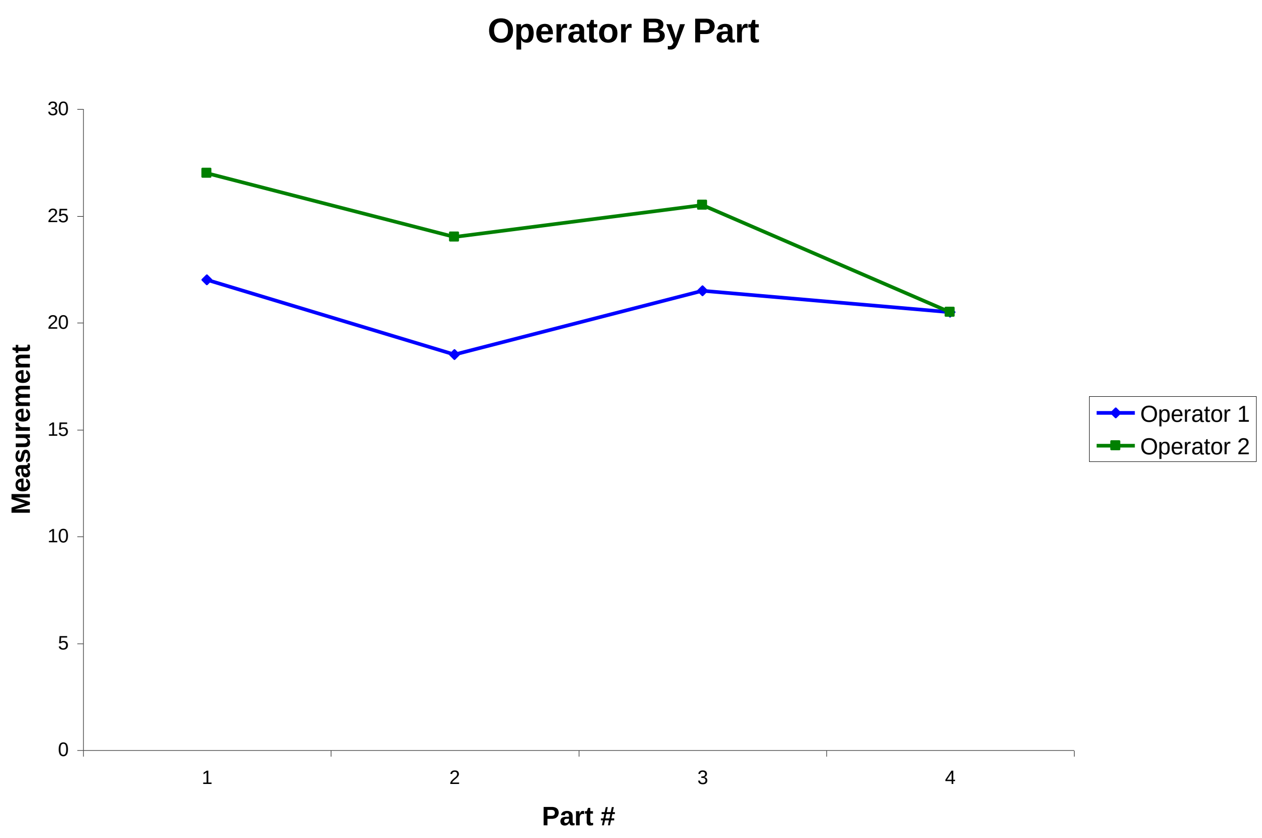
BIAS ANALYSIS	
Reference	Bias
Not Available	

- P/TOL and P/TOT are too high.
- Resolution is unacceptable.
- Reproducibility is significantly larger than Repeatability and appears to be the biggest problem with this measurement process.

Graphical View of Variance Components



Graphical View of Operator by Part Interaction



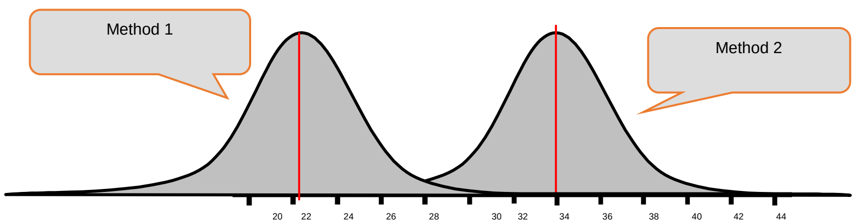
Hypothesis Testing

- A method for looking at data and comparing results
  - Method 1 vs. Method 2
  - Option A vs. Option B
  - Before vs. After Project results

- Helps us make good decisions and not get fooled by random variation:
  - "Is a difference we see REAL, or is it just random variation and no real difference exists at all?"

- We set up 2 hypotheses
  - Example:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2$$



- Based on the sample data we collect to estimate the population, we must decide in favor of either H_0 or H_1. We assume the two population means are equal. Which hypothesis does the evidence support?

Nature of Hypothesis Testing

H_0: Defendant is Innocent (assumed to be true)

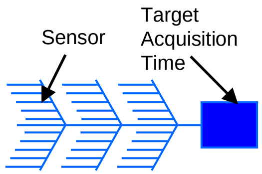
H_1: Defendant is Guilty (trying to show)

- Since verdicts are arrived at with less than 100% certainty, either conclusion has some probability of error. Consider the following table.

		True State of Nature	
		H_0	**H_1**
Conclusion Drawn	**H_0**	**Conclusion is Correct**	**Conclusion results in a Type II error**
	H_1	**Conclusion results in a Type I error**	**Conclusion is Correct**

- Type I or II Error Occurs if Conclusion Not Correct
  - The probability of committing a Type I error is defined as α ($0 \le \alpha \le 1$) and is often called the false detection error. In this example, sending an innocent person to jail.
  - The probability of committing a Type II error is β ($0 \le \beta \le 1$) and is often called missed detection error. Power is the complement of β, $(1 - \beta)$. In this example, letting a guilty person go.
  - The most critical decision error is usually a Type I error, but we should be concerned about the Type II error as well. Sample size controls both errors!

2-Sample Hypothesis Test Example



- In a target kill chain process, a team suspects that one of the important contributors to destroying the target is the amount of time it takes a sensor to acquire the target. The team further suspects that there might be a significant difference in the average amount of time it takes two different sensors to acquire the target. The team decides to conduct a test, to see if “sensor type” really is an important factor. In other words, is there a significant difference in the average target acquisition times between the first and second sensors?

- A random sample of 9 data points for each of the two sensors, shown below, was collected to help answer this question.

Sensor Target Time

Hypothesis Testing Data Files

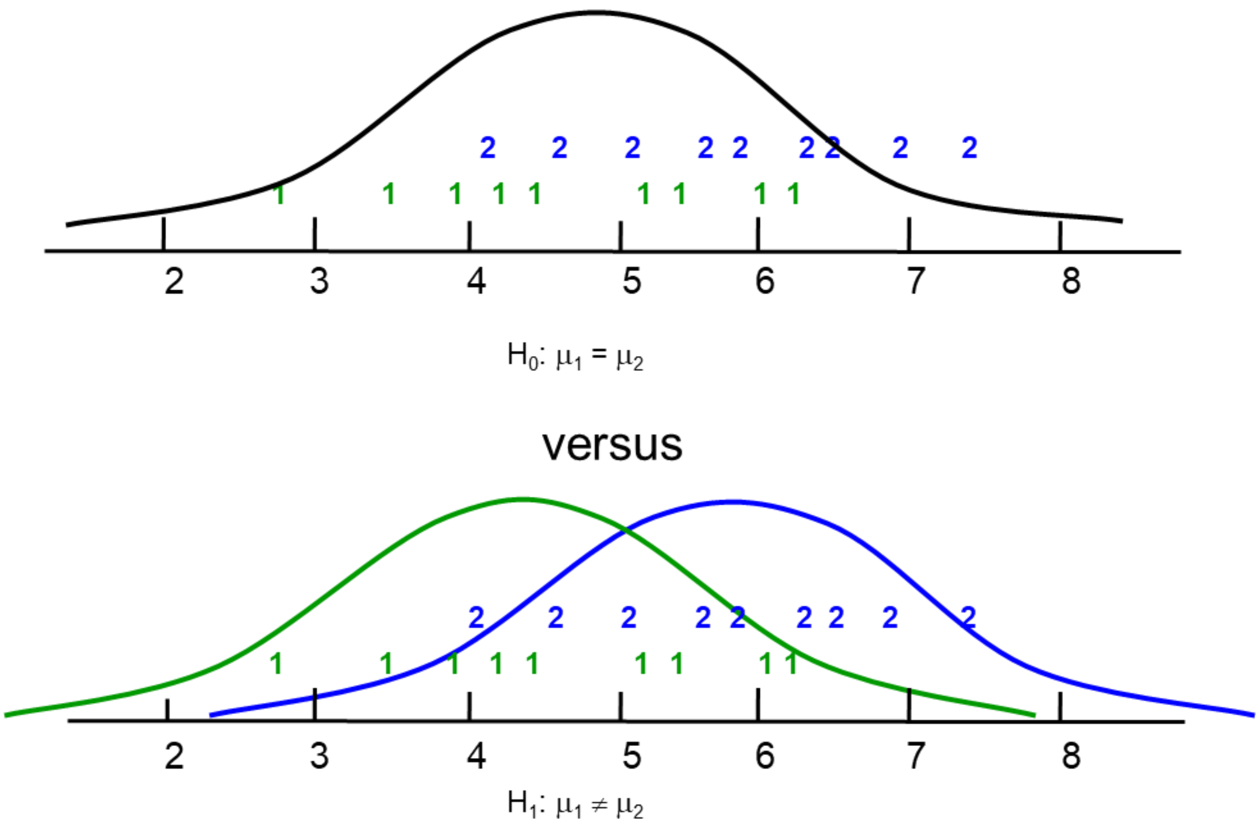
$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

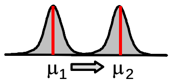
Target Acquisition Time (in seconds)											
Sensor	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	$\bar{y}$	**s**
1	2.8	3.6	6.1	4.2	5.2	4.0	6.3	5.5	4.5	4.6889	1.17
2	7.0	4.1	5.7	6.4	7.3	4.7	6.6	5.9	5.1	5.8667	1.08

2-Sample Hypothesis Test Example (cont.)

- The graphical interpretation of the hypotheses to be tested are:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$





Testing for Differences in Averages (t-test)

- The statistical test for detecting a shift in average is called the t-test. The result of the test is a p-value, which indicates the probability of making a type I error. P-values are derived from the data.

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

- Rule of Thumb:
  - If p-value < 0.05 (red), highly significant difference in the averages (H_1).
  - If 0.05 < p-value < 0.10 (blue), moderately significant difference in the averages. Perhaps get more data!
  - If p-value > 0.10 (black), no significant difference in the averages (H_0).
  - (1 – p-value) • 100% is our percent confidence that there is a significant difference in the averages (H_1).

- For video instruction on hypothesis testing, p-value, and conducting a t test in SPC XL, go to: https://airacad.com/our-insights/training-videos/spc-xl/

Testing for Differences in Averages (SPC XL)

SPC XL > Analysis Tools > t-Test matrix (StdDev)

t-Test Result		
Hypothesis Tested:	H0: Sensor 1 Mean = Sensor 2 Mean H1: Sensor 1 Mean not equal to Sensor 2 Mean	
	p-value (probability of Type I Error)	0.041
	Confidence that Sensor 1 Mean not equal to Sensor 2 Mean	95.9%

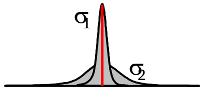
Summary Statistics	Sensor 1	Sensor 2
Mean	4.6889	5.8667
StDev	1.1731	1.0759
Count	9	9

SPC XL results

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

- Since the p-value = 0.041 (and red!), we can be at least (1 – p-value) • 100% confident—in this case 95.9% confident—that the two population averages are different. This is very strong evidence in support of H_1 and is called a statistically significant result.

Testing for Differences in Standard Deviations (F-test)



- The statistical test to detect differences in standard deviations is the F-test. The result of the test is a p-value, which indicates the probability of making a type I error. P-values are derived from the data.

- Rule of Thumb:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

  - If p-value < 0.05 (red), highly significant difference in the standard deviations (H_1).
  - If 0.05 < p-value < 0.10 (blue), moderately significant difference in the standard deviations. Perhaps get more data!
  - If p-value > 0.10 (black), no significant difference in the standard deviations (H_0).
  - (1 – p-value) • 100% is our percent confidence that there is a significant difference in the standard deviations (H_1).

- For video instruction on conducting an F test in SPC XL, go to:

https://airacad.com/our-insights/training-videos/spc-xl/

Hypothesis Testing Data Files

Sensor Target Time

Testing for Differences in Standard Deviations (SPC XL)

SPC XL > Analysis Tools > F-Test matrix (StdDev)

F-Test Result		
Hypothesis Tested:	H0: Sensor 1 Variance = Sensor 2 Variance H1: Sensor 1 Variance not equal to Sensor 2 Variance	
	p-value (probability of Type I Error)	**0.813**
Confidence that Sensor 1 Variance not equal to Sensor 2 Variance		**18.7%**

Summary Statistics	Sensor 1	Sensor 2
Mean	4.6889	5.8667
StDev	1.1731	1.0759
Count	9	9

SPC XL results

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

- Since the p-value = 0.813 (and black!), we can be at least (1 – p-value) • 100% confident—in this case 18.7% confident—that the two population standard deviations are different. This is not very strong evidence to support H_1. This is a statistically insignificant result. Our conclusion would be to stay with H_0. The data has failed to reject the null hypothesis. We assumed H_0 was true to start with and the weak evidence requires us to stay with that assumption!

Hypothesis Test Exercise

Hypothesis Testing Data Files

Paint – Dry Time

- The data below represents the drying time (in seconds) from samples of two different paints (L, H) that were tested.

Paint	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	Y_{11}	Y_{12}	Y_{13}	Y_{14}	$\bar{Y}$	S
L	201	209	215	221	211	213	217	205	218	208	203	214	212	215	211.6	5.8
H	218	225	217	222	223	220	222	216	221	224	224	221	220	219	220.9	2.7

- Use the appropriate hypothesis tests to determine:

1. Is there a significant difference in the average drying time between paint L and paint H? Why or why not?

2. Is there a significant difference in the drying time standard deviation between paint L and paint H? Why or why not?

3. The box plot is a great graphical tool to show, visually, the location and spread differences for two groups. It is a great visual for the group's location and spread differences but it has no statistical significance associated with it! For video instruction on generating a box plot in SPC XL, go to:

https://airacad.com/our-insights/training-videos/spc-xl/

DOE is a Process

PLAN
the experiment

Factor	A	B	C
Row #	pH	Rea Conc	Time
1	4.5	2	1
2	4.5	2	5
3	4.5	5	1
4	4.5	5	5
5	7.5	2	1
6	7.5	2	5
7	7.5	5	1
8	7.5	5	5

DESIGN
the experiment

CONDUCT
the experiment
(collect the data)

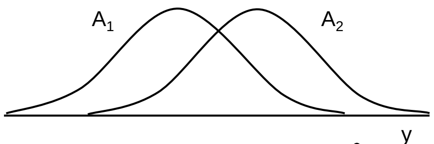
ANALYZE
the data and
draw conclusions

CONFIRM
the results

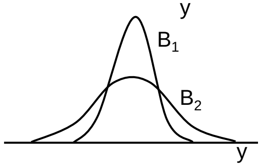
Objectives of a DOE

- Obtain the maximum amount of information using a minimum amount of resources.
- Determine which factors (inputs) shift the average response, which shift the variability and which have no effect.
- Build empirical models relating the response of interest to the input factors.
- Find factor settings that optimize the response and minimize the cost.
- Validate (confirm) results.
- DOE will help identify the following types of factors:

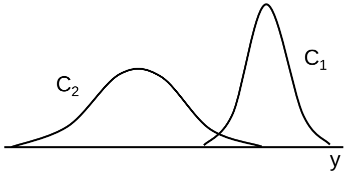
i) Factor A affects the average



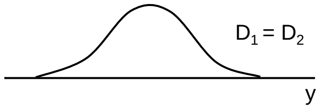
ii) Factor B affects the standard deviation



iii) Factor C affects the average and the standard deviation

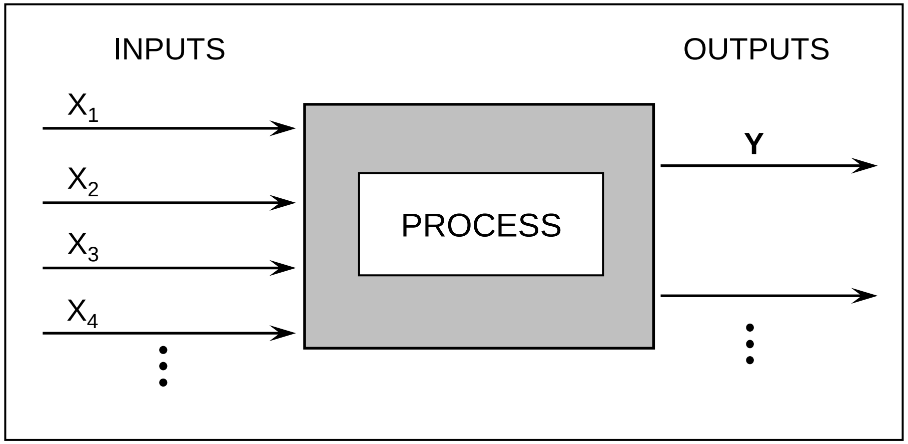


iv) Factor D has no effect



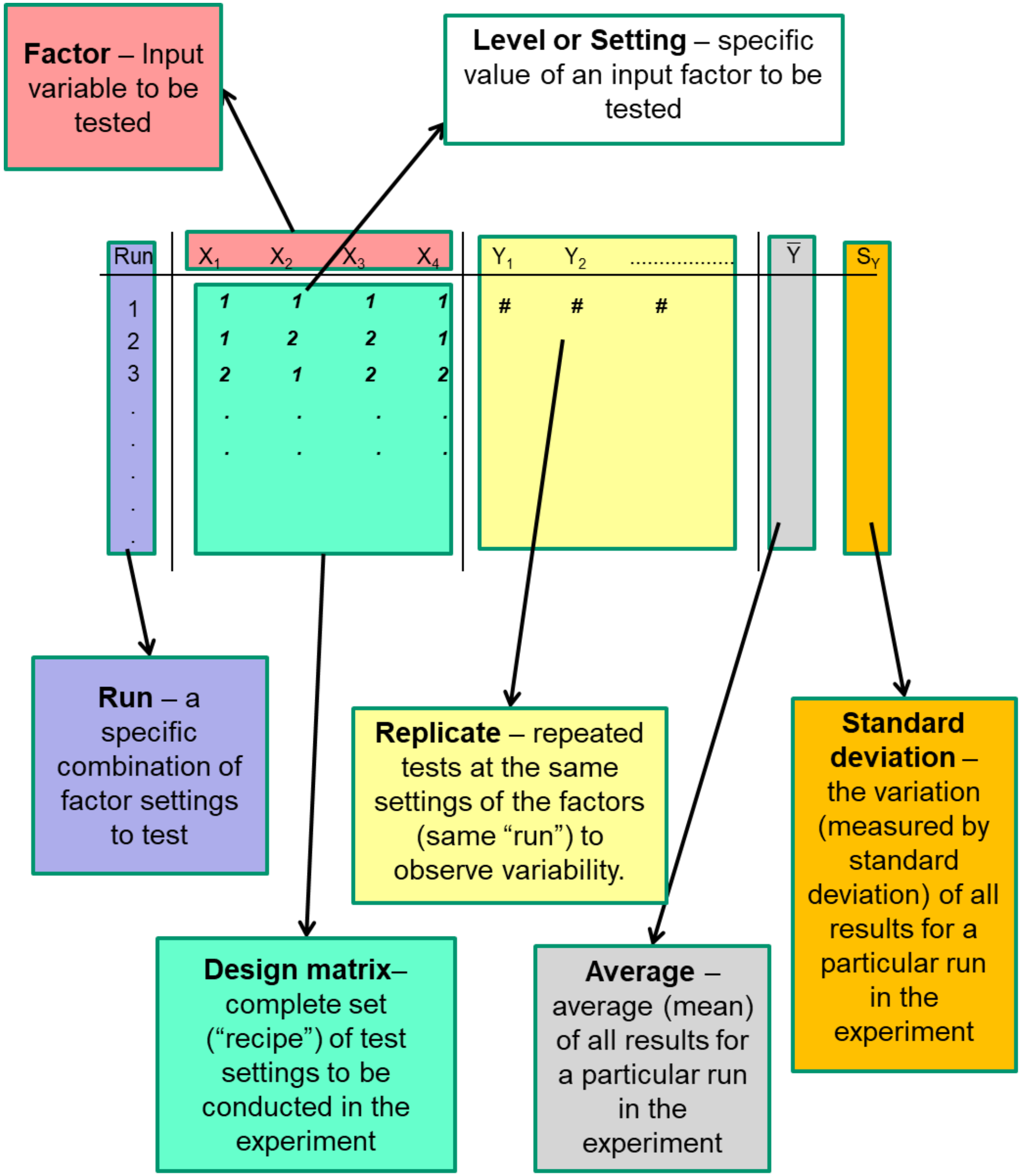
What is a Designed Experiment?

Purposeful changes of the inputs (factors) in order to observe corresponding changes in the output (response).



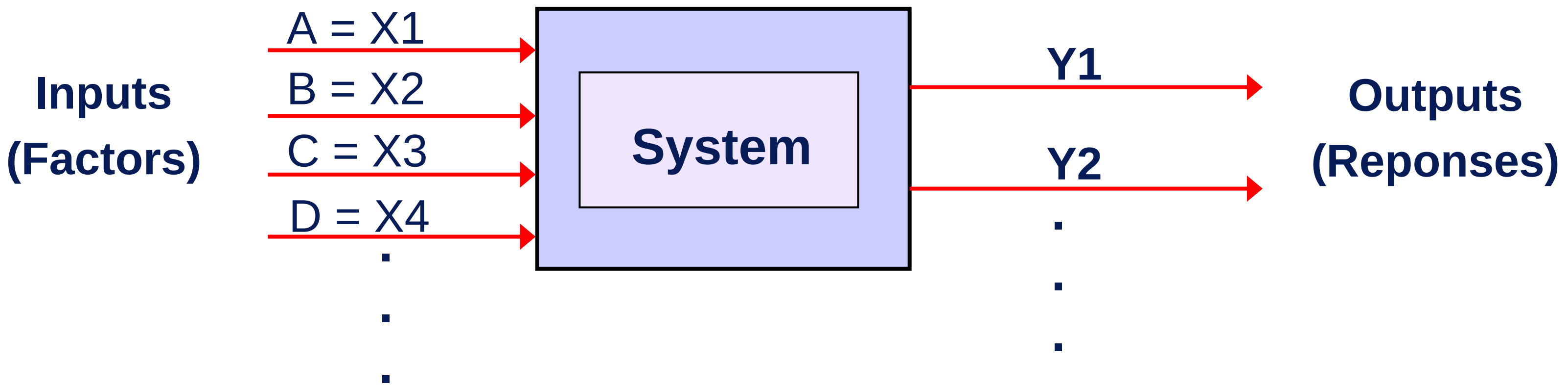
Run	X_1	X_2	X_3	X_4	Y_1	Y_2	....	$\bar{Y}$	S_Y
1									
2									
3									
.									
.									
.									
.									
.									

The Structure of a Designed Experiment



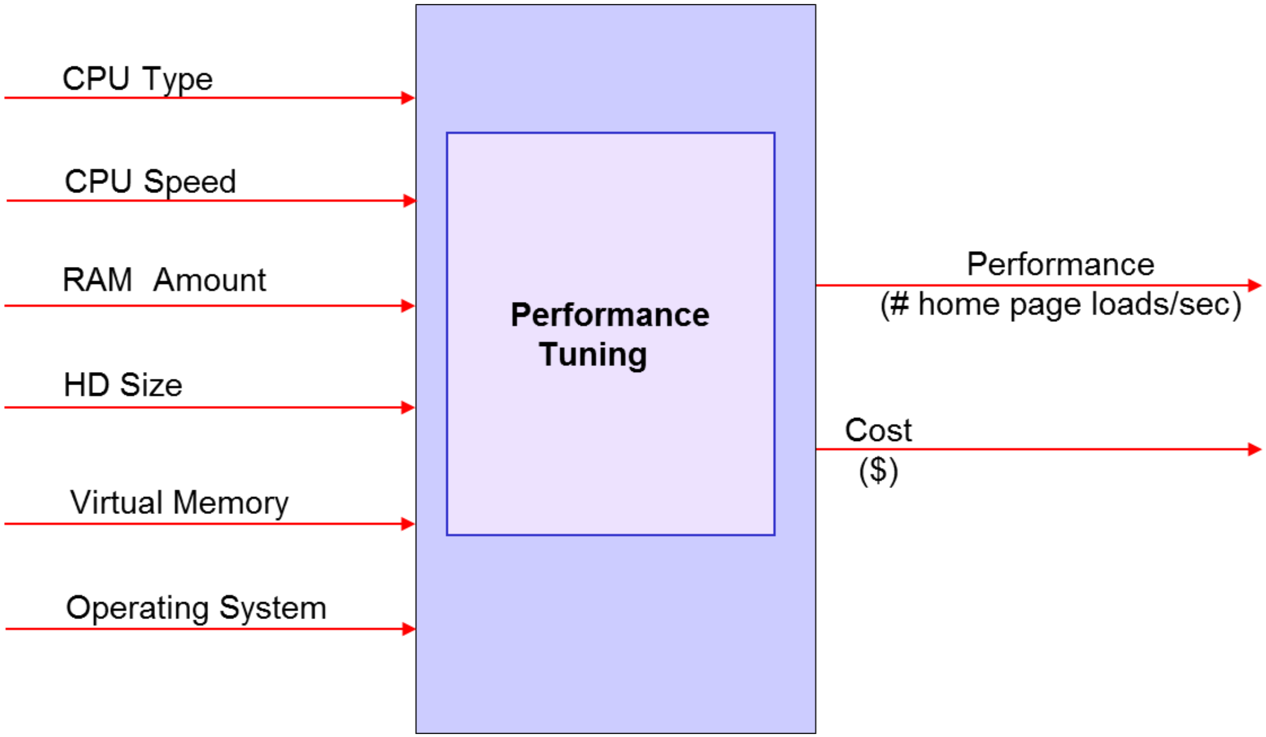
Who should use DOE?

- Anyone who wants to understand the causal relationships between the inputs to a system and the resulting outputs



- DOE is applicable to both physical processes and computer simulation models.

Example of a Web-Based Test Scenario



Subject Matter Experts Must Be Involved

Factors/Inputs (X's)	Levels (Choices)	Response/Outputs (Y's)
CPU Type	Itanium, Xeon	# home page loads/sec
CPU Speed	1 GHz, 2.5 GHz	Cost
RAM Amount	256 MB, 1.5 GB	
HD Size	50 GB, 500 GB	
VM	J2EE, .NET	
OS	Windows, Linux	

- Which factors are important? Which are not?
- Which combination of factor choices will create operational problems?
- How do you know for sure? Show me the data.

Terminology of DOE

Y: Output, response variable, dependent variable

X: Input, factor, test parameter, independent variable (a measurable entity that is purposely changed during a test)

Level: A unique value or choice of a factor (X)

Run: An experimental combination of the levels of the X's; a test case

Test Design Matrix: the collection of all test cases, also known as a covering array in software testing

Replication: Doing or repeating a test case

Effect: The difference or impact on Y when changing X

Interaction: When the effect of one factor depends on the level of another factor; also known as a combination effect

Test Design Matrix for the Web-Based Test Scenario

Test Design Matrix

Run	CPU Type	CPU Speed	RAM Amount	HD Size	VM	OS
1	Itanium	1 GHz	256 MB	50 GB	J2EE	Windows
2	Itanium	1 GHz	256 MB	50 GB	J2EE	Linux
3	Itanium	1 GHz	1.5 GB	500 GB	.NET	Windows
4	Itanium	2.5 GHz	256 MB	500 GB	.NET	Windows
5	Itanium	2.5 GHz	1.5 GB	50 GB	.NET	Linux
6	Itanium	2.5 GHz	1.5 GB	500 GB	J2EE	Linux
7	Xeon	1 GHz	1.5 GB	500 GB	J2EE	Windows
8	Xeon	1 GHz	1.5 GB	50 GB	.NET	Linux
9	Xeon	1 GHz	256 MB	500 GB	.NET	Linux
10	Xeon	2.5 GHz	1.5 GB	50 GB	J2EE	Windows
11	Xeon	2.5 GHz	256 MB	500 GB	J2EE	Linux
12	Xeon	2.5 GHz	256 MB	50 GB	.NET	Windows

Response Variables

Page Loads/second	Cost
9	37

Test Design Optimization

- The number of test cases or runs will depend on
  - Number of input factors and their types (qualitative or quantitative)
  - Number of levels we want to test each factor at (2-level designs are the simplest)
  - Purpose of the test
  - Other constraints that may be imposed on the test scenario

- Note: the number of test cases does NOT depend on the number of outputs or response variables

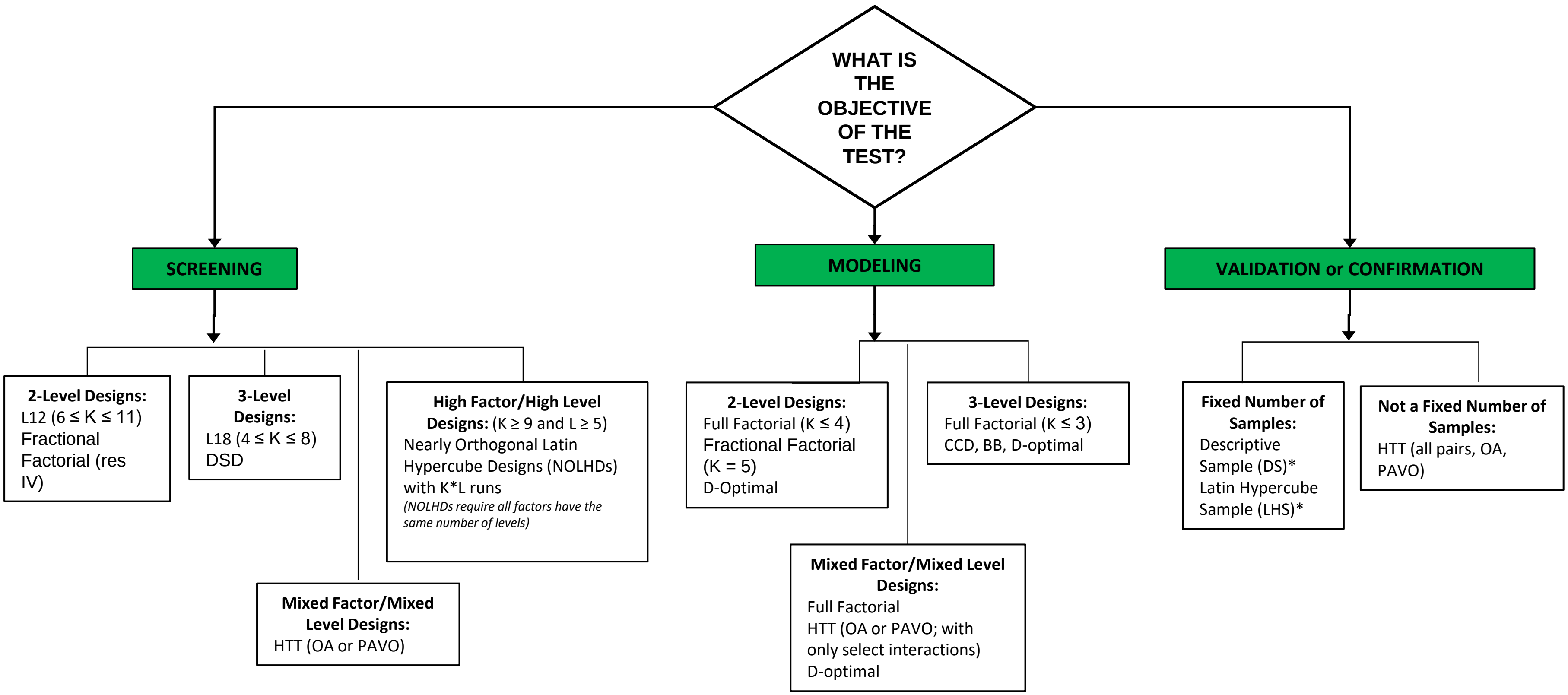
Major Reasons for Using a DOE

- **Screening**
  - For testing many factors in order to **separate** the vital few critical factors from the trivial many

- **Modeling**
  - For building **functions** that can be used to predict outcomes, assess risk, and optimize performance
  - These **include** the ability to evaluate **interaction and higher order effects**.

- **Performance Validation and Verification**
  - For **confirming** that a system performs in accordance with its specifications/requirements.

Various Options for Design Selection



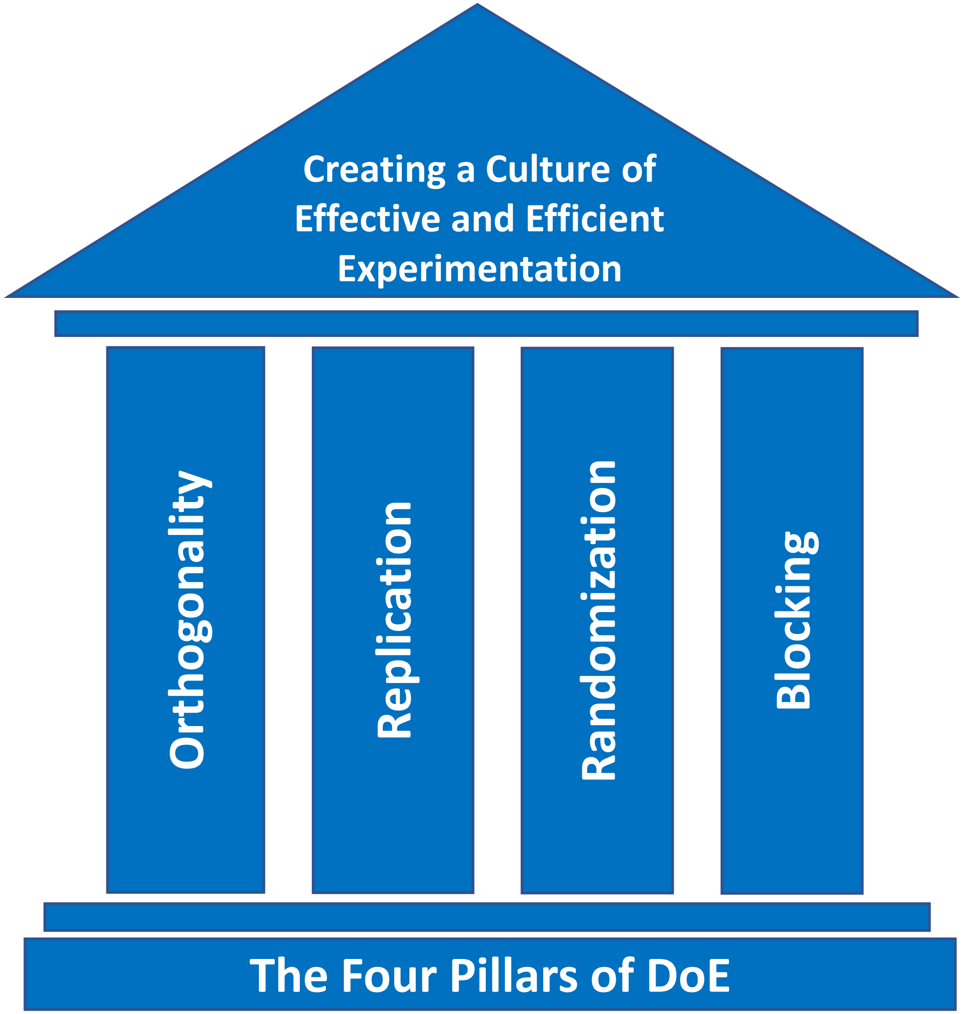
Notes:

1. “Mixed” factors means a combination of quantitative and qualitative (categorical)
2. “Mixed” levels means that not all factors have the same number of levels (settings)
3. “K” = Number of Factors and “L” = Number of Levels
4. HTT = High Throughput Testing
5. DSD = Definitive Screening Design
6. “OA” stands for Orthogonal Array; “PAVO” = Pairwise Value Ordering
7. Software such as HD Tools™, rdExpert™ Lite, Pro-Test™ and Quantum XL™ generate some or all of these designs

* DS and LHS are sampling techniques to generate representative samples according to a specified distribution and a specified sample size

* Representative samples do <u>not</u> give orthogonal designs. They are often used for getting test coverage, validating performance/ determining capability, or creating noise combinations for test

The Foundations of DOE



Orthogonality

- This is the feature of a test design that allows for the ***independent*** evaluation of the effects of factors and their interactions – and nonlinear effects as well, depending on the type of design chosen.
- Why is independent evaluation so important? It gets us much closer to cause and effect relationships, and it makes the subsequent analysis of the data much easier.
- The difference between DOE and an observational study (historical data analysis) is the ability to do independent evaluation and arrive at causal relationships.
- Most leaders of organizations do not know this. They may know that DOE is important but they really don't know why. Orthogonality is a major reason why.
- Sometimes the terms orthogonal, independent, or completely uncorrelated are used interchangeably.
- In a coded test design matrix, orthogonality means perfect vertical and horizontal balance.

Replication

- Since DOE is about the study of variation, replicating or getting repeated measures for the same test condition allows us to study variation.
- There are different types of variation of interest (e.g., within and between setup, etc.), so it behooves us to know how to take the replications or repeated measures in order to do a proper evaluation of the variation in the response variable (y).
- One of the most common questions in all of statistics and process improvement is, "what should my sample size be?" That is, how much data do I need?
- The power of a statistical test is based on the sample size or the number of replications.
- In some testing scenarios, we don't have the luxury to "replicate" the test cases and in other situations (deterministic simulators) replication is a waste.
- Replication is a major attribute of any test design and thus a stalwart in DOE, because it increases the precision of the test.

Randomization

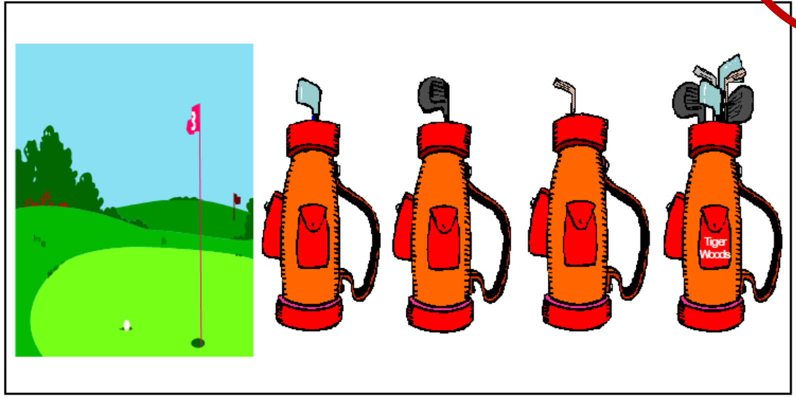
- The basic reason we randomize the test cases is to spread the noise (from factors we cannot control) as evenly as possible across the entire design space.
- In that way, noise factors do not become confounded or correlated with factors that are involved and controlled in the experiment.
- There are other major benefits of randomization
  - It minimizes selection bias. A good randomization procedure will be unpredictable in the sense that one cannot predict what the next set of test conditions will be based on knowledge of the previous test cases.
  - It facilitates blinding or masking of the test case identity from investigators, participants, and assessors; that is, it prevents bias.
  - It permits the use of probability to express the likelihood of any differences in outcomes between test cases to be due merely to chance.
- Depending on the test scenario, complete randomization may not be achievable because it may cost too much. Factors that are difficult to change or very expensive to change make randomization a real challenge.
- Complete randomization is hardly ever achieved in a DOE, so the practitioner needs to be able to make educated trade-offs.

Blocking

- Blocking is the arranging of test cases in groups (blocks) that are similar to one another. There are a variety of reasons why we might want to block.
- Oftentimes the runs in an experiment are completed under different conditions. This may lead to the consideration of variables that are not part of the designed experiment but still could be important and influence the results. These are called nuisance variables. Nuisance variables include things like operator, time of day, room temperature, and lot number when they themselves are not factors in the experiment.
- Blocking can be used to remove or estimate the effect of a nuisance variable.
- A factor in a DOE may be very difficult or expensive to change. Thus, we block on that variable while doing all of the runs or tests cases (randomly) at one level of the factor before changing the level and conducting all of the remaining runs (randomly) when that factor is at a second level.
- Any blocking used in an experiment should be well thought out before the experiment is run, because the analysis techniques used will depend on the factors and the blocking variables.
- In general, “block (the effect of) important nuisance variables when possible, and randomize when you can’t.”

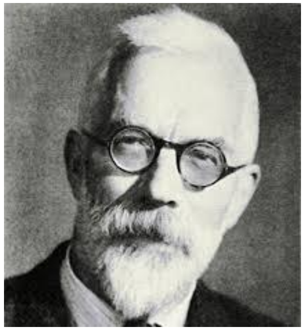
Important Contributors to DOE

	TAGUCHI	SHAININ	CLASSICAL	BLENDED APPROACH
Loss Function	*			*
Emphasis on Variance Reduction	*			*
Robust Designs	*			*
KISS	*	*		*
Simple Significance Tests		*		*
Component Swapping		*		
Multivariate Charts		*		*
Modeling			*	*
Sample Size			*	*
Efficient Designs			*	*
Optimization			*	*
Confirmation	*			*
Response Surface Methods			*	*



Which bag would a world class golfer prefer?

A Little History

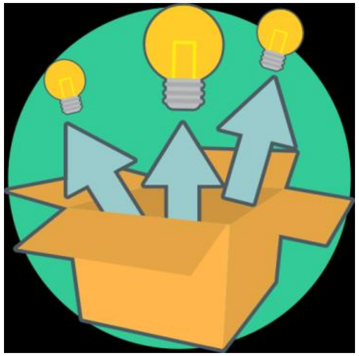


- **Genesis**: Sir R.A. Fisher; Rothamsted Laboratory, 1920's
- **Initial Applications**:
  - agriculture (Fisher and Yates) (20's)
  - cotton and woolen industries (Tippett and Daniels) (40's)
  - chemical industries (Davies and Box) (50's)
  - Japan (Deming, Juran, Ishikawa and Taguchi) (50's)
- **Applications Today**:
  - design and development of processes and products (medical devices, pharma, consumer products, etc.)
  - software testing
  - marketing and understanding voice of the customer
  - process/product simulation
  - and so many more …
- **Knowledge Gained from DOE:**
  - sensitivity
  - characterization
  - optimization
  - robustness
  - tolerance design for X's

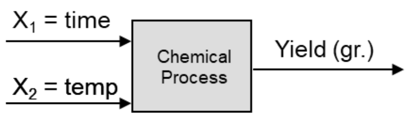
Approaches to Testing Multiple Factors

- **Traditional Approaches**
  - One Factor at a Time (OFAT)
  - Oracle (Best Guess)
  - All possible combinations (full factorial)

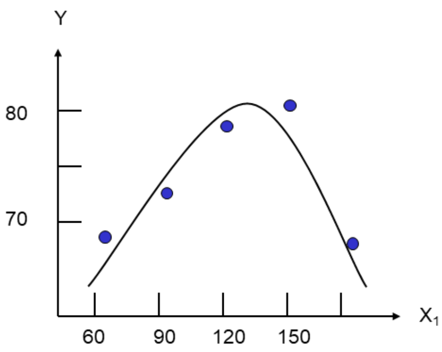
- **Modern Approach**
  - Statistically designed experiments (DOE) … factorial designs plus other selected DOE designs, depending on the situation



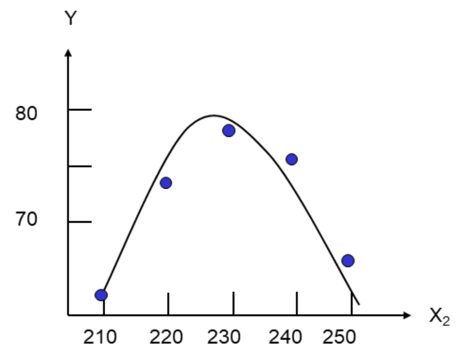
One Factor at a Time (OFAT) Testing



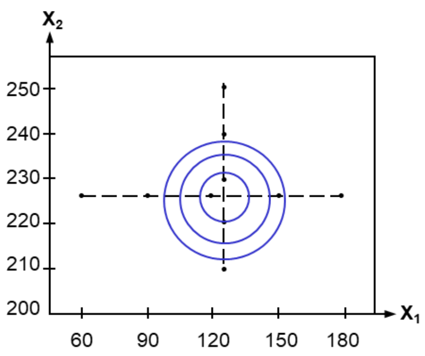
1. Hold X_2 constant and vary X_1
   Find the "best setting" for X_1



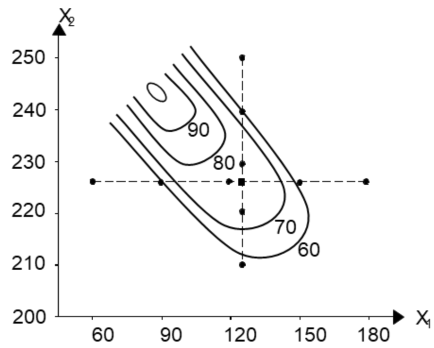
2. Hold X_1 constant at "best setting" and vary X_2.
   Find the "best setting" for X_2.



3. One factor at a time results



4. One factor at a time results versus optimal results



OFAT Summary

Good News

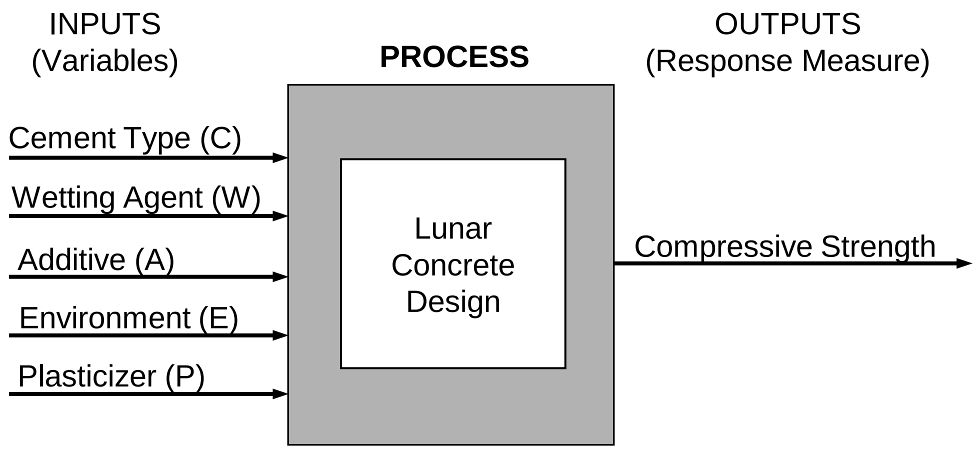
- Quick and simple
- Intuitive
- The way we are often taught

Bad News

- Will not be able to estimate factor interaction effects
- Will not be able to generate good prediction models with interactions

Oracle (Best Guess) Approach

- Subject matter experts or technical experts often use a “Best Guess” (Oracle) approach
- Example – Suppose we have a process with 5 inputs and 1 output. The objective is to learn how to maximize the output (adapted from case study in DOE Text: Chapter 8)



- Experts choose settings and a test plan for the inputs that make sense, based on their knowledge of the process to learn about the factor effects and optimize the output

C_1 - Portland Type III Cement
C_2 - Calcium Aluminate Cement

W_1 - Wetting Agent - 0.07 ml
W_2 - No Wetting Agent

A_1 - No Reinforcement
A_2 - Steel

E_1 - Semi-Evacuated
E_2 - Ambient Mixing

P_1 - Plasticizer - 1 ml
P_2 - No Plasticizer

Best Guess Test Plan

	Run	C	W	A	E	P	Y
Expert #1	1	1	1	1	1	1	5
Expert #2	2	1	1	1	1	2	6
	3	1	2	1	1	1	5
	4	1	2	2	1	2	6
	5	2	2	2	2	2	7
	6	2	2	2	2	1	8
	7	2	1	2	2	2	10
	8	2	1	1	2	1	11

- The 8 "best guesses" above are tried, based on the experts' current knowledge. At this point, what knowledge has been gained?
  - Does Factor C (cement type) affect Y (strength)?
  - Does Factor A (additive) affect Y (strength)?

More Terms and Definitions

Run	C	W	A	E	P
1	1	1	1	1	1
2	1	1	1	1	2
3	1	2	1	1	1
4	1	2	2	1	2
5	2	2	2	2	2
6	2	2	2	2	1
7	2	1	2	2	2
8	2	1	1	2	1

Aliased (perfectly confounded)

- Two columns are identical (same test pattern)
- Cannot learn about the effects of the factors independently

Run	C	W	A	E	P
1	1	1	1	1	1
2	1	1	1	1	2
3	1	2	1	1	1
4	1	2	2	1	2
5	2	2	2	2	2
6	2	2	2	2	1
7	2	1	2	2	2
8	2	1	1	2	1

Partially confounded

- Two columns are not identical, but also not balanced
- There is some correlation between the two columns
- Cannot learn about the effects completely independent

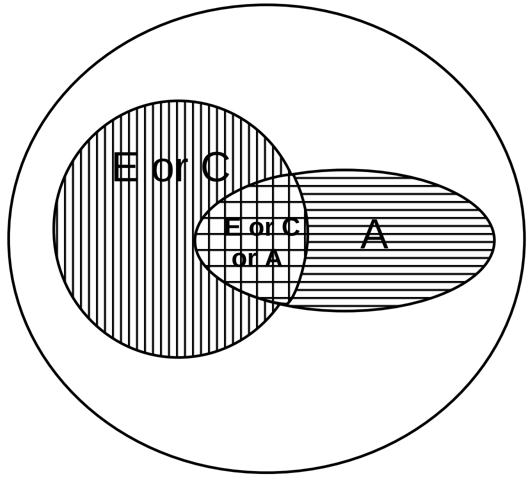
Run	C	W	A	E	P
1	1	1	1	1	1
2	1	1	1	1	2
3	1	2	1	1	1
4	1	2	2	1	2
5	2	2	2	2	2
6	2	2	2	2	1
7	2	1	2	2	2
8	2	1	1	2	1

Balanced

- Two columns are uncorrelated
- Within each level of one factor, the other factors test settings are evenly divided between its low and high test settings. In this example, when C is tested at its “1” setting, W is tested twice at its “1” setting and twice at its “2” setting. The same holds true when C is tested at its “2” setting.
- We can learn about the effects independently

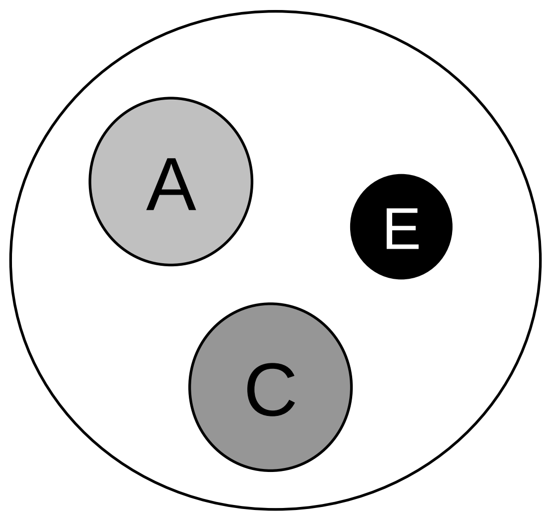
Where is the variation of Y coming from?

What we have is:



What we need is:

A design to provide independent estimates of effects.



How do we obtain this independence of variables?

Best Guess (Oracle) Summary

Good News

- Quick and simple for multiple inputs
- Intuitive
- The way we are often taught
- If the response is optimized, everyone is happy!

Bad News

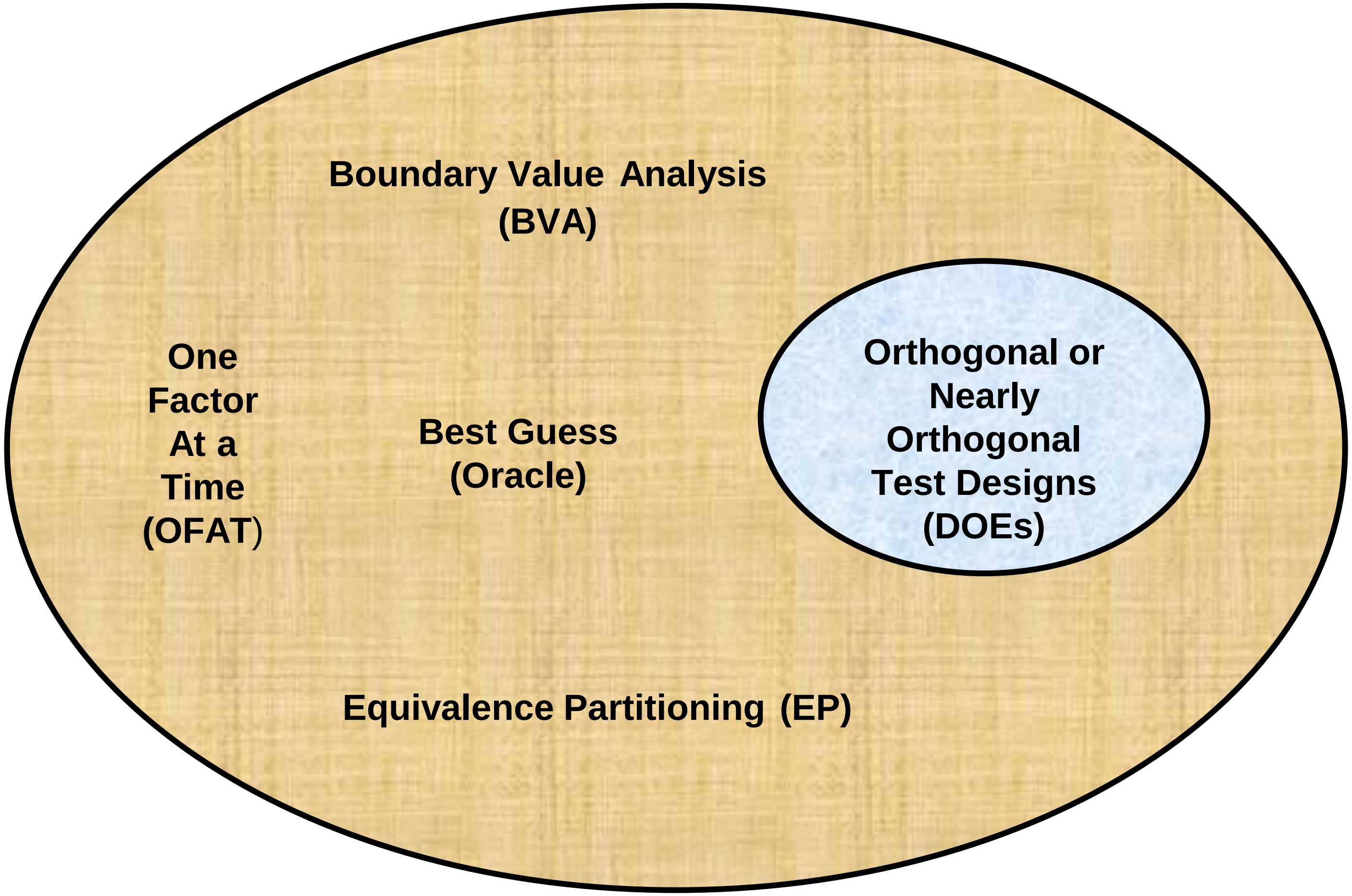
- Poor (limited) knowledge
- Will not be able to determine which inputs had any effect on the output response
- Oftentimes, factors will be confounded or aliased with other factors
- Will not be able to estimate interaction effects with multiple factors
- Will not be able to generate good prediction models

Famous Quote

"All experiments (tests) are designed;
some are poorly designed,
some are well designed."

George Box (1919-2013), Professor of Statistics, DOE Guru

DOE: a subset of all possible test designs



The Set of All Possible Test Designs

DOE Means Statistically Designed Experiments

Orthogonal or Nearly Orthogonal Designs

- **Full Factorials** (for modeling a small numbers of factors)
- **Fractional Factorials** (for screening or modeling at 2 levels)
- **Placket - Burman**
- **Latin Squares**
- **Hadamard Matrices**

  } Taguchi Designs (for screening)
- **Box - Behnken Designs**
- **Central Composite Designs (CCD)**

  } Response Surface Designs (for modeling nonlinear effects)
- **High Throughput Testing (All Pairs)** (for validation testing)
- **Nearly Orthogonal Hypercube Designs** (for screening or modeling computer simulators)

Prerequisites for Successful Testing

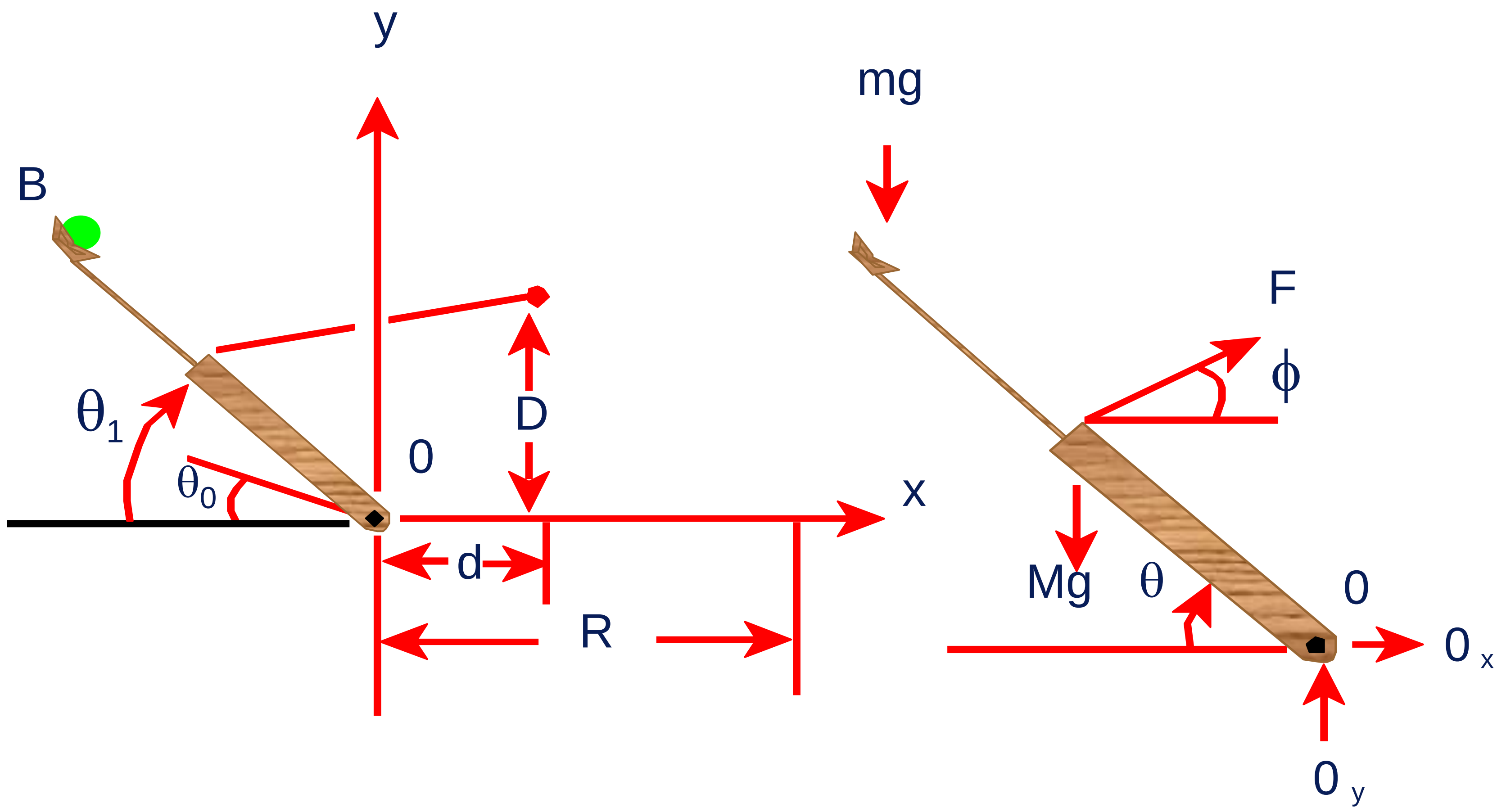
- **Remove excessive variation from the system**
  - Look at each step in the process and see where we can reduce variation (Process Flow)
  - Document all factors that could possibly impact the results (Cause and Effect)
  - Use Standard Operating Procedures (SOPs) to remove as much noise as possible
  - Apply the famous PF/CE/CNX/SOP variance reduction methodology

- **Perform a Measurement System Analysis (MSA)**
  - To determine how much variation is coming from the measurement system itself
  - To ensure the measurement system is capable
  - To improve the measurement system if it is not capable

Catapulting Power into Test and Evaluation

Statapult ® Catapult

The Theoretical Approach



The Theoretical Approach (cont.)

$$I_0 \ddot{\theta} = r_F F(\theta) \sin\theta \cos\varphi - (Mgr_G + mgr_B) \sin\theta \qquad \tan\phi = \frac{D - r_F \sin\theta}{d + r_F \cos\theta},$$

$$\frac{1}{2} I_0 \dot{\theta}^2 = r_F \int_{\theta_0}^{\theta} F(\theta) \sin\theta \cos\varphi \, d\theta - (Mgr_G + mgr_B)(\sin\theta - \sin\theta_0)$$

$$\frac{1}{2} I_0 \dot{\theta}_1^2 = r_F \int_{\theta_0}^{\theta_1} F(\theta) \sin\theta \cos\varphi \, d\theta - (Mgr_G + mgr_B)(\sin\theta_1 - \sin\theta_0).$$

$$x = v_B \cos\left(\frac{\pi}{2} - \theta_1\right) t - \frac{1}{2} r_B \cos\theta_1 \qquad y = r_B \sin\theta_1 + v_B \sin\left(\frac{\pi}{2} - \theta_1\right) t - \frac{1}{2} g t^2.$$

$$r_B \sin\theta_1 + (R + r_B \cos\theta_1) \tan\left(\frac{\pi}{2} - \theta_1\right) - \frac{g}{2V_B^2} \frac{(R + r_B \cos\theta_1)^2}{\cos^2\left(\frac{\pi}{2} - \theta_1\right)} = 0.$$

$$\frac{gI_0}{4r_B} \frac{(R + r_B \cos\theta_1)^2}{\cos^2\left(\frac{\pi}{2} - \theta_1\right)\left[r_B \sin\theta_1 + (R + r_B \cos\theta_1) \tan\left(\frac{\pi}{2} - \theta_1\right)\right]}$$

$$= r_F \int_{\theta_0}^{\theta_1} F(\theta) \sin\theta \cos\phi \, d\theta - (Mgr_G + mgr_B)(\sin\theta_1 - \sin\theta_0).$$

Statpult DOE Demo

	Actual Factors		Coded Factors			Response Values			
Run	A	B	A	B	AB	Y_1	Y_2	$\bar{Y}$	S
1	160	2	-1	-1	+1				
2	160	3	-1	+1	-1				
3	180	2	+1	-1	-1				
4	180	3	+1	+1	+1				

Simplified Table for Determining Sample Size Based on Confidence and Power

Percent Confidence that a term identified as significant, truly does belong in $\hat{s}$ [$\hat{y}$]	Percent chance of finding a significant variance [average] shifting term if one actually exists	Number of Runs in 2 Level Portion of the Design				
		2	4	8	12	16
		Sample Size per Experimental Condition				
95% (α = .05)	40% (β = .60)	5 [3]	3 [2]	2 [1]	N/A	N/A
95% (α = .05)	75% (β = .25)	9 [5]	5 [3]	3 [2]	2 [1]	2 [1]
95% (α = .05)	90% (β = .10)	13 [7]	7 [4]	4 [2]	3 [2]	N/A
95% (α = .05)	95% (β = .05)	17 [9]	9 [5]	5 [3]	4* [2]	3 [2]
95% (α = .05)	99% (β = .01)	21 [11]	11 [6]	6 [3]	5* [3]	4* [2]

Best Practices for “Operationalizing” DOE
(i.e., changing the culture to one of habitually using DOE)

1. Coaching on projects is an absolute must.

2. A Keep-It-Simple-Statistically (KISS) approach with easy-to-comprehend materials and easy-to-use software.

3. Gaining and propagating quick-hitting successes.

4. Getting leadership on board and continuously re-invigorating them is necessary.

5. Developing a culture of continuously generating transfer functions for the purpose of optimization, prediction, and risk assessment.

Key Take-Aways

- STAT includes various techniques such as hypothesis testing, MSA, DOE, and regression analysis.

- Hypothesis testing allows us to control, via sample size, the P(false detection) and P(missed detection), the Type I (alpha risk) and Type II (beta risk) errors, respectively.

- Measurement System Analysis (MSA) is a test on the measurement system itself.

- MSA will help <u>quantify</u>, more exactly, the capability of the measurement system and answer questions about repeatability, reproducibility, and capability with respect to the customer specs.

- There are various approaches to testing many factors simultaneously, to include OFAT and Oracle (Best Guess).

- DOE brings orthogonal or nearly orthogonal designs into play and can be used to screen, model, or perform validation testing.

- DOE is the key link between Test and Evaluation, because it allows us to evaluate the effects of factors and their interactions independently from one another.

- Learning about STAT/DOE and making it practical in an organization does NOT have to be difficult. Following the 5 Best Practices for "operationalizing" DOE in an organization can make it happen.

CAUTION!!

This tutorial is designed to be an overview of the topic of STAT. It is not meant to make anyone an expert in the subject matter. Becoming a practitioner will require more education than what this tutorial provides.

Air Academy Associates offers a live week-long class on Scientific Test and Analysis Techniques (STAT). This course can also be taken online at one's own pace or also virtually, or even some combination of the live, online and virtual venues depending on organizational needs.

For more information, please contact

Air Academy Associates, LLC

Toll Free: (800) 748-1277 or (719) 531-0777

Email: aaa@airacad.com
or mkiemele@airacad.com

Website: www.airacad.com