

# Data Collection

# No-Data better than Bad-Data

- We use data to make decisions
- Bad-data leads to bad decisions
- No-Data demands that new data be taken
  - Opportunity for data relevance, accuracy and integrity

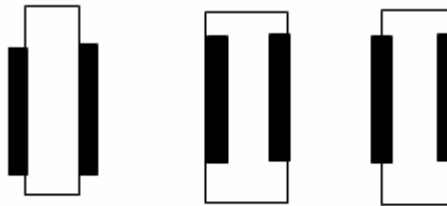
# What is Bad-Data

- Incomplete: e.g. missing time periods
  - Not representative
- Not relevant to problem at hand
- Collected without operational definition
- Old and not representative of current conditions
- Errors using measuring equipment (related to operational definition)

# Operational Definition Important

Example: Line width measurement

Wide Instrument cursor



Which is correct?

- Depends on Application
- How does instrument makes measurement
- How does operator use instrument

# Good Data

- Representative of problem at hand
  - completeness
- Data is relevant to problem
  - Right metrics
- Data is accurate
- Data is consistently taken and recorded
- Data can be analyzed and presented to tell the true story

## References used

- CSSBB 2007 Primer – *QCI*
- Lean Six sigma Pocket ToolBook - *George*

# Regarding Accuracy & Integrity



© QUALITY COUNCIL OF INDIANA  
CSSBB 2007

VI-42 (440)

## VI. MEASURE - DATA DATA COLLECTION / SAMPLING METHODS

### **Data Accuracy and Integrity**

**Bad data corrupts the decision-making process. Some considerations include:**

- **Avoid emotional bias relative to tolerances when counting, measuring, or recording data.**
- **Avoid unnecessary rounding. Rounding often reduces measurement sensitivity.**
- **If data occurs in time sequence, record the order of its capture.**
- **If an item characteristic changes over time, record the measurement after its manufacture, as well as after a stabilization period.**

# More from QCI

- Screen or filter data to detect and remove data entry errors such transposition or misplaced decimal points.
- Avoid removal by hunch. Use objective statistical tests to identify outliers.
- Each important classification identification should be recorded along with the data.

It is important to select a sampling plan appropriate for the purpose of the use of the data.

# Creating a Data Collection Plan

1. Decide what data to collect:
2. Develop Operation Definitions
3. Decide on the correct Sample Size for the measurements..
4. Now decide if the existing data will be adequate, or whether new data will be needed.

# Creating a Data Collection Plan (2)

5. Develop data collection forms
6. Identify who will take the data.
7. Train the data collectors.
8. Decide what analysis you intend to perform and who will do it.

These steps should all be done before the data collection and/or experimental work begins. Data collection can be costly and time-consuming. You want to be sure the effort will produce a conclusion.

# 1. Decide What Data To Collect

- Types of Data
- Measurement Scales

Identify the metrics that best describe the problem (measure the right thing). This would involve Input parameters and Output parameters. Identify any parameters for which descriptive information would help determine important patterns in the data, e.g. root causes, patterns of wear, etc.

# Types of Data

## – Attribute Data

- Countable
- Discrete

## – Variables Data

- Measurable
- Continuous
- More information than attribute data
- Often cost more to collect than attribute data

## – Locational Data

When you can collect either Attribute or Variables, collect Variables data because it intrinsically contains more information. Less data collection may be required. You can also change some variables data into attribute data

# Types of Data

	Variables	Attribute
Characteristics	measurable continuous may derive from counting	countable discrete units or occurrences good/bad
Types of Data	length volume time	no. of defects no. of defectives no. of scrap items
Examples	width of door lug nut torque fan belt tension	audit points lost paint chips per unit defective lamps
Data Examples	1.7 inches 32.06 psi 10.542 seconds	10 scratches 6 rejected parts 25 paint runs

Source: QCI CSSBB Primer, pVI-37

# Locational Data

The third type of data does not fit into either category above. This data is known as locational data which simply answers the question “where.” Charts that utilize locational data are often called “measles charts” or concentration charts.

# Measurement Scales

- **Nominal**
  - Numerical assignment to something non-numerical
  - More a form of classification than measurement
  - Can only use = or  $\neq$
- **Ordinal**
  - Can order items in terms of whether they have more or less of an attribute
  - Can use =,  $\neq$ ,  $<$ ,  $>$
- **Interval**
  - Difference between any two successive points is equal
  - Zero point on scale is arbitrary
  - Can add/subtract
- **Ratio**
  - The zero point indicates absence of an attribute
  - Can add/subtract/multiply/divide

# Measurement Scales – Examples

Scale	Description	Example
Nominal	Data consists of names or categories only. No ordering scheme is possible.	A bag of candy contained the following colors: Yellow 15 Red 10 Orange 9 Green 7
Ordinal (Ranking)	Data is arranged in some order but differences between values cannot be determined or are meaningless.	Product defects, where A type defects are more critical than D type defects are tabulated as follows: A 16 B 32 C 42 D 30
Interval	Data is arranged in order and differences can be found. However, there is no inherent starting point and ratios are meaningless.	The temperatures of three ingots were 200°F, 400°F and 600°F. Note, that three times 200°F is not the same as 600°F as a temperature measurement.
Ratio	An extension of the interval level that includes an inherent zero starting point. Both differences and ratios are meaningful.	Product A costs \$300 and product B costs \$600. Note, that \$600 is twice as much as \$300.

Source: QCI CSSBB Primer, pVI-39

# Key Characteristics

<b>Scale</b>	<b>Central Location</b>	<b>Dispersion</b>	<b>Significance Tests</b>
Nominal	Mode	Percentages	Chi Square
Ordinal	Median	Percentages	Sign or Run Test
Interval	Arithmetic Mean	Standard or Average Deviation	t-test, F-test Correlations
Ratio	Arithmetic, Geometric or Harmonic Mean	Standard Deviation	t-test, F-test Correlations

Source: QCI CSSBB Primer, pVI-40

# Measurement System Matrix

	Output Measures						
● Strong							
○ Medium							
△ Weak							
<b>Customer requirements</b>							

Perform VOC to define Critical Customer requirements (CTQ)

Identify Output Measures you are or should collect and relate these to the CTQ

Collect Data on Outputs most strongly related to CTQ

# Creating a Data Collection Plan

1. Decide what data to collect:
2. **Develop Operational Definitions**
3. Decide on the correct Sample Size for the measurements..
4. Now decide if the existing data will be adequate, or whether new data will be needed.

## 2. Develop Operational Definitions

- Describes how measurement is made.
- Describes what “good” looks like
- Should describe the metric being studied
- Procedure should be detailed enough that anyone doing the measurement can do it as intended. Detail may depend on who does the measurements. Visual Cues & standards?
- Who defines the “standard”
- Test the OD with people involved in the procedure devel. and then with people not involved in the procedure devel.
- Repeat until consistent results are obtained
- Procedure can be used as an ongoing training document.

Measurement Procedure can affect outcome.

# Creating a Data Collection Plan

1. Decide what data to collect:
2. Develop Operation Definitions
3. Decide on the correct Sample Type & Size for the measurements.
4. Now decide if the existing data will be adequate, or whether new data will be needed.

# What is Sampling

Sampling refers to the practice of evaluating (inspecting) a portion -the sample - of a lot – the population – for the purpose of inferring information about the lot.

Statistically speaking, the properties of the sample distribution are used to infer the properties of the population (lot) distribution.

# Why Sample?

- Economy
  - Less inspection labor
  - Less time
- Less handling damage
- Destructive testing

# 3. Select Sample Type & Size

## Types of Sampling

- Random Sampling
- Stratified Sampling
  - Random samples selected from each group or process

# Required Sample Size variables data

Select Risk (Confidence) Level

Smallest difference in Means to be detected

Variation in characteristic measured ( s or  $\sigma$  )

$$n = Z_{\alpha/2}^2 \sigma^2 / \delta^2$$

in which

n = sample size

$Z_{\alpha/2}$  = sigma for desired confidence level

s,  $\sigma$  = std. Deviation for sample/population

$\delta$  = smallest difference in mean to be detected

# Example

We want to determine if an equipment adjustment will alter hourly output by as much as 4 lbs/hour. What is minimum sample size, at a confidence level of 95%, that would confirm a shift in the mean output by more than 4 lbs/hr?

The historic standard deviation for the process is 15 lb/hr.

# Required Sample Size attribute data

This formula also works for Poisson data using  $\bar{c}$  for  $\sigma$ .

In which

$\bar{c}$  = average number of defects

For binomial data, use

$$n = Z_{\alpha/2}^2 (p)(1-p)/(\delta p)^2$$

In which

$\delta p$  = desired proportion difference to be detected

$p$  = proportion defective

# Sampling Caveats

- Size of sample is more important than percentage of lot
- Only random samples are statistically valid
- Access to samples does not guarantee randomness
- Misuse of sampling plans can be costly and misleading.
- Stratification
- No such thing as a single representative sample

# Representative Sample?

There is no such thing as a single representative sample

Why?

- Draw repeated samples of 5 from a normally distributed population.
- Record the  $\bar{X}$  (mean) and  $s$  (std.dev) for each sample
- What is the result?

# The Random Sample

At any one time, each of the remaining items in the population has an equal chance of being the next item selected

One method is to use a table of Random Numbers.

Enter the table randomly ( like pin-the-tail-on-the-donkey)

- Proceed in a predetermined direction – up, down, across
- Discard numbers which cannot be applied to the sample

# Random Number Table

10	09	73	25	33	76	52	01	35	86	34	67	35	48	76	80	95	90	91	17	39	29	27	49	45
37	54	20	48	05	64	89	47	42	96	24	80	52	40	37	20	63	61	04	02	00	82	29	16	65
08	42	28	89	53	19	64	50	93	03	23	20	90	25	60	15	95	33	47	64	35	08	03	36	06
99	01	90	25	29	09	37	67	07	15	38	31	13	11	65	88	67	67	43	97	04	43	62	76	59
12	80	79	99	70	80	15	73	61	47	64	03	23	66	53	98	95	11	68	77	12	17	17	68	33
66	06	57	47	17	34	07	27	68	50	36	69	73	61	70	65	81	33	98	85	11	19	92	91	70
31	06	01	08	05	45	57	18	24	06	35	30	34	28	14	86	79	90	74	39	23	40	30	97	32
85	25	97	76	02	02	05	16	56	92	68	66	57	48	18	73	05	38	52	47	18	62	38	85	79
63	57	33	21	35	05	32	54	70	48	90	55	35	75	48	28	46	82	87	09	83	49	12	56	24
73	79	64	57	53	03	52	96	47	78	35	80	83	42	82	60	93	52	03	44	35	27	38	84	35
08	52	01	77	67	14	90	56	86	07	22	10	94	05	58	60	97	09	34	33	50	50	07	39	98
11	80	50	54	31	39	80	82	77	32	50	72	56	82	48	29	40	52	42	01	52	77	56	78	51
83	45	29	96	34	06	28	89	80	83	13	74	67	00	78	18	47	54	06	10	68	71	17	78	17
88	68	54	02	00	86	50	75	84	01	36	76	66	79	51	90	36	47	64	93	29	60	91	40	62
99	59	46	73	48	87	51	76	49	69	91	82	60	89	28	93	78	56	13	68	23	47	83	41	13

Source: *Statistical Quality Control* by Grant & Leavenworth

# Stratified Sampling

- Random samples are selected from a homogeneous “lot”. Often, the parts may not be homogeneous because they were produced on different machines, by different operators, in different plants, etc.
- With stratified sampling, random samples are drawn from each “group” of processes that are different from other groups.

# Selecting the Sample

- Wrong way to select sample
  - Judgement: often leads to Bias
  - Convenience
- Right ways to select sample
  - Randomly
  - Systematically: e.g. every  $n$ th unit; risk of bias occurs when selection routine matches a process pattern

# Creating a Data Collection Plan

1. Decide what data to collect:
2. Develop Operation Definitions
3. Decide on the correct Sample Size for the measurements..
4. Now decide if the existing data will be adequate, or whether new data will be needed.

## 4. Will Existing Data be Useable?

- Will it be representative of problem at hand
  - Completeness
- Is it relevant to problem
  - Right metrics
- Is it accurate
- Has it been consistently taken and recorded
- Can it be analyzed and presented to tell the true story, identify the root cause, etc. Is there enough data.

A measurement R&R study may be required

# Creating a Data Collection Plan (2)

5. **Develop data collection forms**
6. Identify who will take the data.
7. Train the data collectors.
8. Decide what analysis you intend to perform and who will do it.

These steps should all be done before the data collection and/or experimental work begins. Data collection can be costly and time-consuming. You want to be sure the effort will produce a conclusion.

# 5. Develop Data Collection Forms

- Check Sheets
  - Defect
  - Location
  - Histogram
- Recording Sheets
  - Manual
- Automatic

# Check Sheets

- Facilitate data collection process

Time in System	Frequency
0-15 Min	
15-30 Min	//
30-45 Min	<del>////</del>
45-60 Min	//
60-75 Min	
75-90 Min	
> 90 Min	

Process Check Sheet

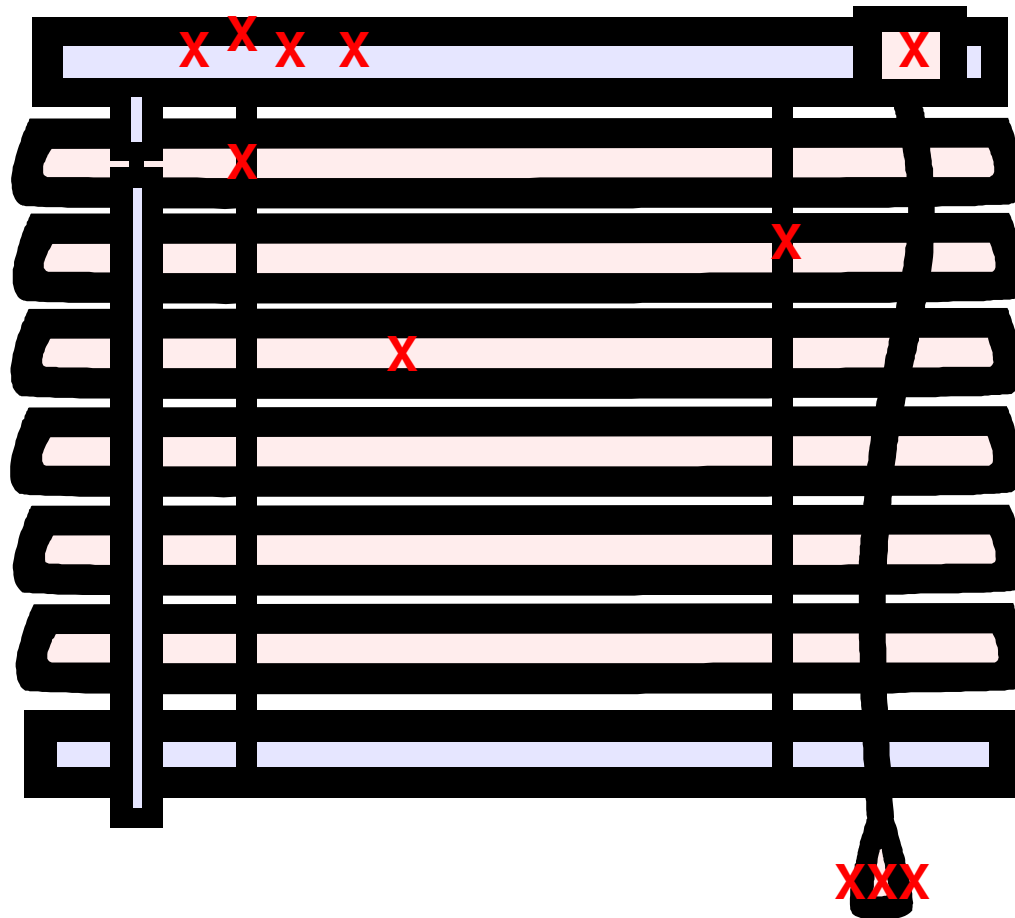
# Defect Check Sheets

Defect	Frequency
Wrong Diagnosis	
Lost Record	//
Waited > 30 Min	<del>////</del>
Wrong Medication	//

# Stratified Defects Check Sheet

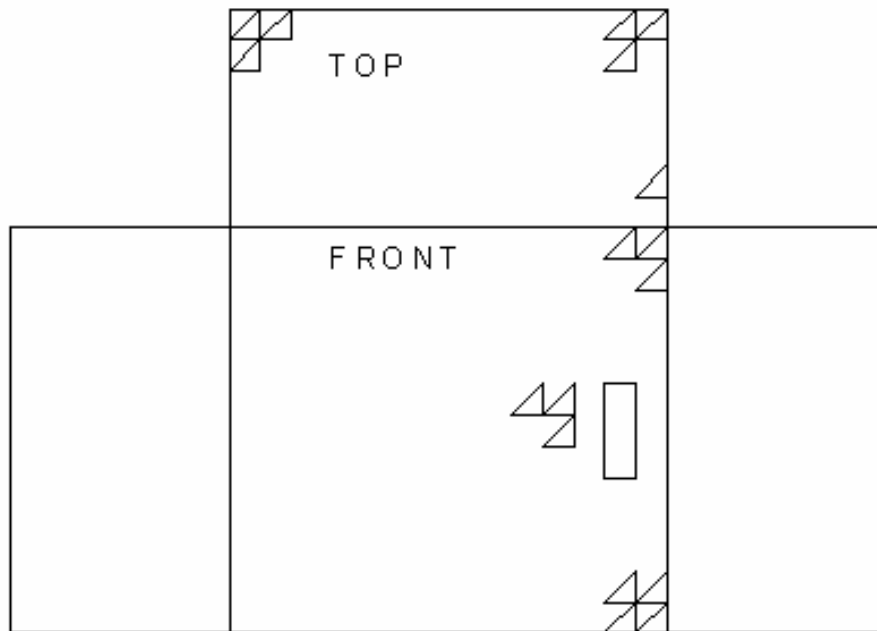
Defect	NP	MD
Wrong Diagnosis		
Lost Record	//	/
Waited > 30 Min	<del>////</del>	/
Wrong Medication		

# Defect Location Check Sheets (Measles Chart)

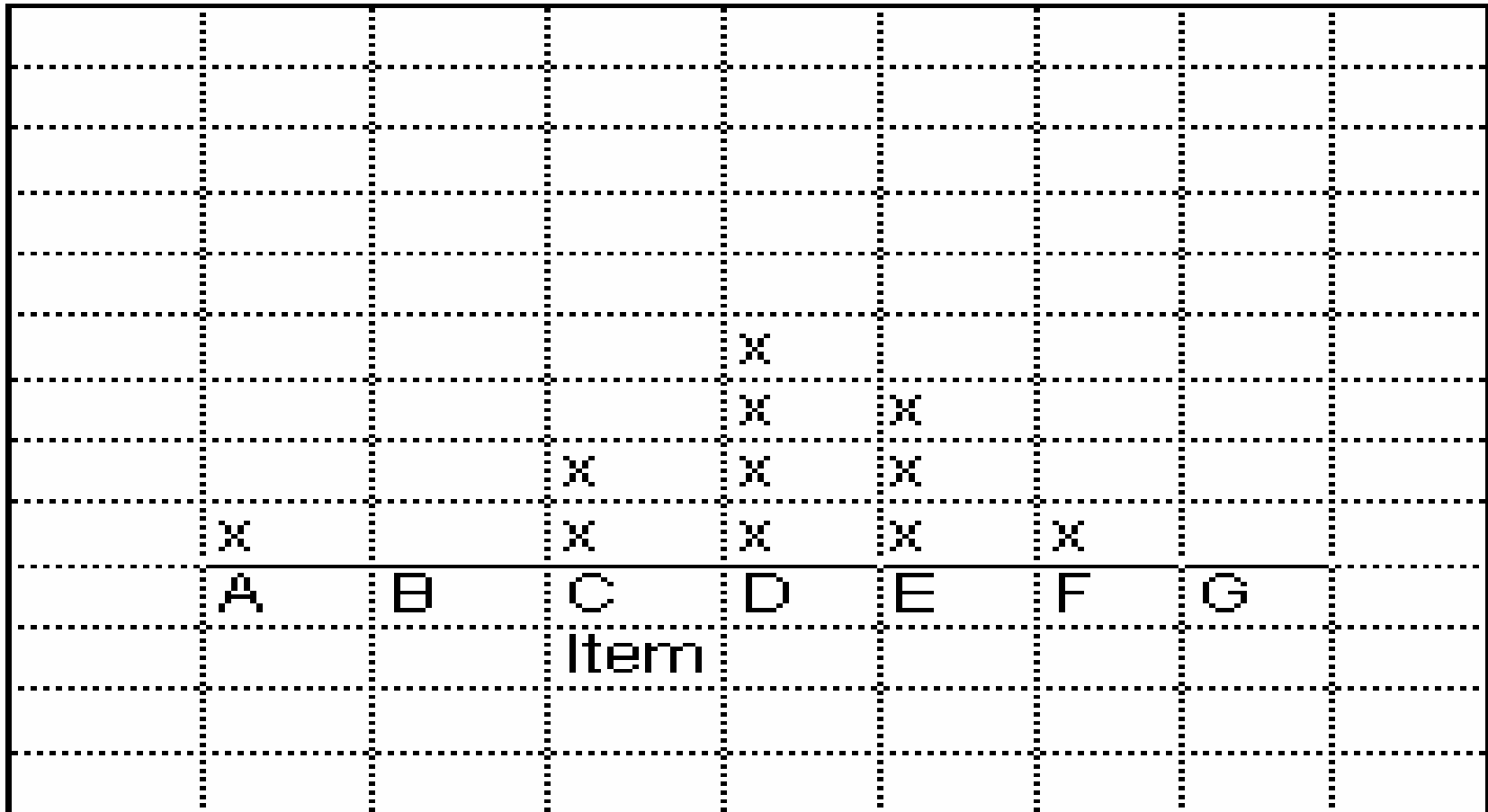


# Another Measles Chart

CHIPPED PAINT ON FRIDGE



# Histogram (Frequency Chart)





# Data Collection Forms

Tensile Stress Log

Date	Shift	Oper Init	Machine	Lot	Time In	Time Out	Break Tension

# Data Coding

- Used to increase efficiency of data entry and analysis
- Example Problems Data Coding Helps Avoid
  - Trying to squeeze too many digits into fields on data collection form
  - Data entry errors resulting from needing to enter large sequences of digits for single observation
  - Excessive time entering data because large sequences of digits must be entered for each observation

# Coding with Simple Math Operators

Coding by adding or subtracting a constant or by multiplying or dividing by a factor:

Let the subscript, lowercase c, represent a coded statistic; the absence of a subscript represents raw data; uppercase C indicates a constant; and lowercase f represents a factor. Then:

Code: $X_c = X + C$	Decode: $\bar{X} = \bar{X}_c - C,$	$\sigma = \sigma_c$
Code: $X_c = X - C$	Decode: $\bar{X} = \bar{X}_c + C,$	$\sigma = \sigma_c$
Code: $X_c = fX$	Decode: $\bar{X} = \bar{X}_c/f,$	$\sigma = \sigma_c/f$
Code: $X_c = X/f$	Decode: $\bar{X} = f\bar{X}_c,$	$\sigma = f\sigma_c$

Source: QCI CSSBB Primer, pVI-45

# Coding by Substitution

- Parts inspected with ruler with 1/8 inch increments.
- An alternative to recording actual measurements is to record the number of  $\pm 1/8$  inch deviations from the target value
- For example, assume a blind has a target width of 36 inches.
  - A blind with a width of 36.5 inches would be recorded as ?
  - A blind with a width of 37 inches would be recorded as ?
  - A blind that was 35.25 inches would be recorded as ?

# Coding by Substitution

- Parts inspected with ruler with 1/8 inch increments.
- An alternative to recording actual measurements is to record the number of  $\pm$  1/8 inch deviations from the target value
- For example, assume a blind has a target width of 36 inches.
  - A blind with a width of 36.5 inches would be recorded as +4
  - A blind with a width of 37 inches would be recorded as +8
  - A blind that was 35.25 inches would be recorded as -6

# Coding by Truncation of Repetitive Place Values

- Consider measurements such as 0.67512, 0.67543, 0.67589
- The digits 0.675 repeat
- In these cases the last two digits can be recorded as two digit integers.

# Creating a Data Collection Plan (2)

5. Develop data collection forms
6. Identify who will take the data.
7. Train the data collectors.
8. Decide what analysis you intend to perform and who will do it.

These steps should all be done before the data collection and/or experimental work begins. Data collection can be costly and time-consuming. You want to be sure the effort will produce a conclusion.

# 6. Identify Who Will Take the Data

- Technicians
- Engineers
- Incoming
- Supplier Outgoing
- Other
- Data Correlation & Validation
- Measurement R&R study
- Familiar with process
- Conflict of interest

# Creating a Data Collection Plan (2)

5. Develop data collection forms
6. Identify who will take the data.
7. **Train the data collectors.**
8. Decide what analysis you intend to perform and who will do it.

These steps should all be done before the data collection and/or experimental work begins. Data collection can be costly and time-consuming. You want to be sure the effort will produce a conclusion.

# 7. Train the Data Collectors

- Data Collectors may help with Collection Form design
- Practice Run
- Show how data will be analyzed; helps collectors see importance of collection procedure

# Creating a Data Collection Plan (2)

5. Develop data collection forms
6. Identify who will take the data.
7. Train the data collectors.
8. Decide what analysis you intend to perform and who will do it.

These steps should all be done before the data collection and/or experimental work begins. Data collection can be costly and time-consuming. You want to be sure the effort will produce a conclusion.

## 8. Decide the Analysis Methodology

- This provides feedback to Collection Forms design
- Will the data and analysis plan answer the desired questions?





