AMU | AMERICAN MERIDIAN UNIVERSITY

# Measurement System Analysis (MSA)

Dr. Bob Gee
Dean Scott Bonney
Professor William G. Journigan
American Meridian University

# Learning Objectives

Upon successful completion of this module, the student should be able to:

- Understand Measurement Systems Analysis validates tool accuracy, precision and stability
- Understand the importance of good measurements
- Understand the language of measurement
- Understand the types of variation in measurement systems
- Learn how to conduct and interpret a measurement system analysis with normally distributed continuous data
- Learn how to conduct an MSA with Attribute data

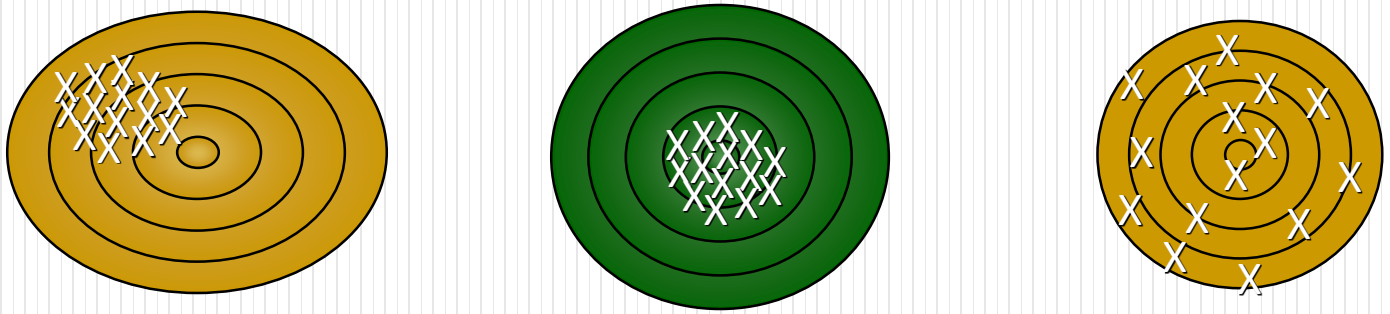
# Measurement System Analysis

- Measurement System Analysis (MSA) – Ability to measure and validate the accuracy of a measuring device against a recognized quantifiable standard
- Ability to assess process performance is only as good as the ability to measure it
- MSA is our *eyes* and *ears*
  - Must clearly see and hear process performance in order to improve it
  - Sometimes, improving the ability to measure our process results in immediate process improvements

**If you cannot measure, you cannot improve! ~ Taguchi**

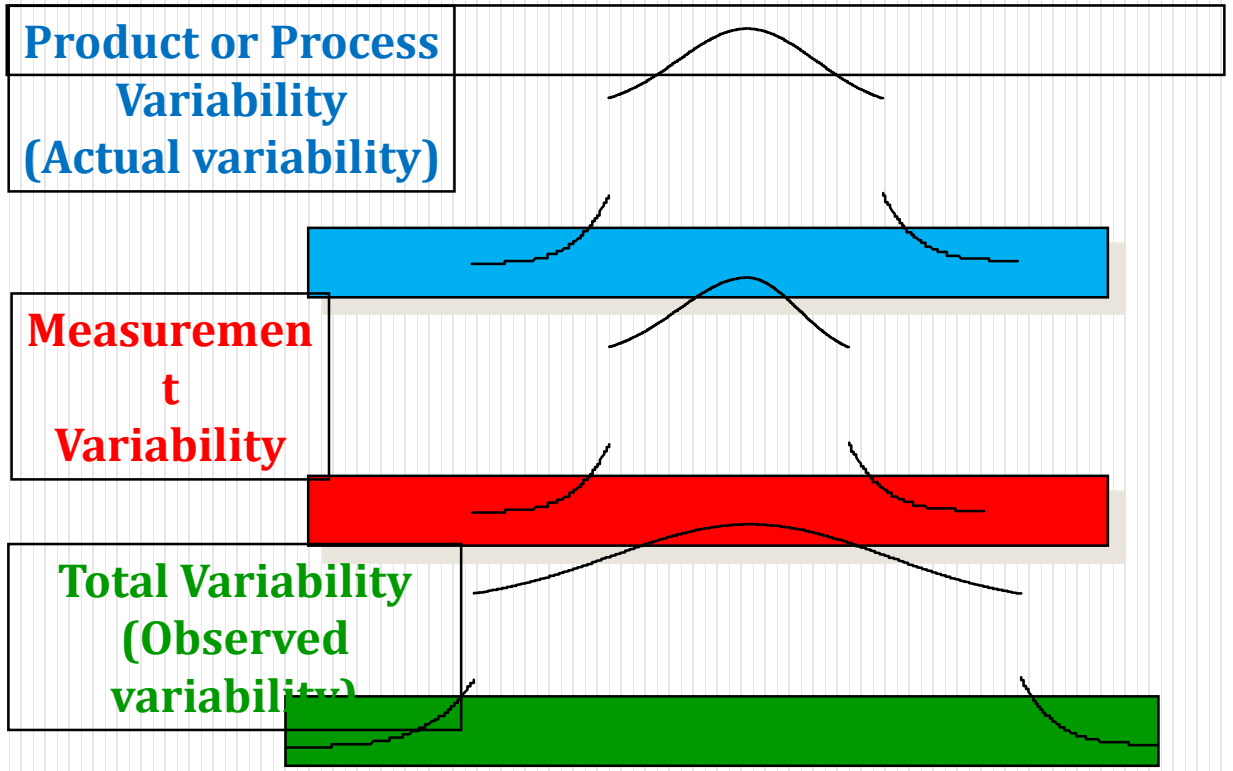
# Measurement Systems Analysis and Variation

- Measurement System Analysis (MSA) identifies and quantifies different sources of variation that affect a measurement system
- Variation in measurement attributes
  - Variation in the product itself
  - Variation in the measurement system
- Variation in measurement system itself is measurement error



# Measurement Variation

- This is the primary Measurement System issue in observed variation:




$$\hat{\sigma}^2\ Product$$

$$+$$

$$\hat{\sigma}^2\ ms$$

$$=$$

$$\hat{\sigma}^2\ Total$$

# Measurement Variation Concerns

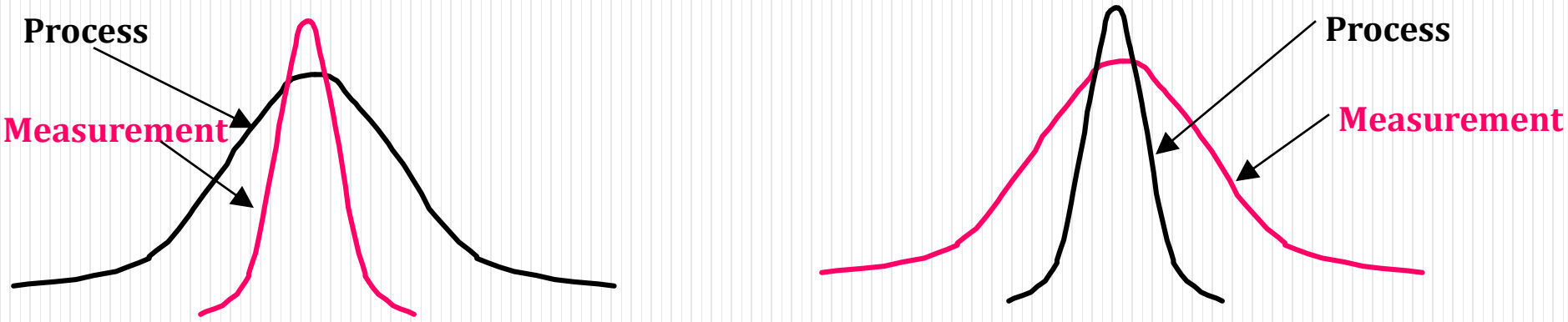
- Consider the reasons why we measure:

**Verify Conformity to Specifications (Product / Process)**

**Assist Continuous Improvement Activities**

**How might measurement variation affect these decisions?**

**What if the amount of measurement variation is unknown?**




**Measurement variation can make process capabilities appear worse than they are**

# Acceptable Measurement System Properties

- Measurement system must be in control
- Variability must be small:
  - Relative to process variation
  - Compared with specification limits
- Measurement increments must be small relative to the smaller of:
  - Process variability or
  - Specification limits
    - Rule of Thumb: Increments are no greater than $^{1}/_{10}$ th of the smaller of:
      - a) Process variability or
      - b) Specification limits

# Reducing Measurement Errors

- Piloting
- Train all people involved
- Double-check all data thoroughly
- Use statistical procedures to adjust for measurement error
- Use multiple measures of the same construct

# MSA Definitions

- Accuracy (Bias) — the difference between observed average measurement and a standard.
- Stability — variation obtained with a measurement system on the same parts over an extended period of time.
- Linearity — the difference of bias throughout the expected operating range of the equipment.
- Discrimination- the amount of change from a reference value that an instrument can detect.
- Repeatability (Precision) — variation when one person repeatedly measures the same unit with the same measuring system.
- Reproducibility — variation when two or more people measure the same unit with the same measuring system.

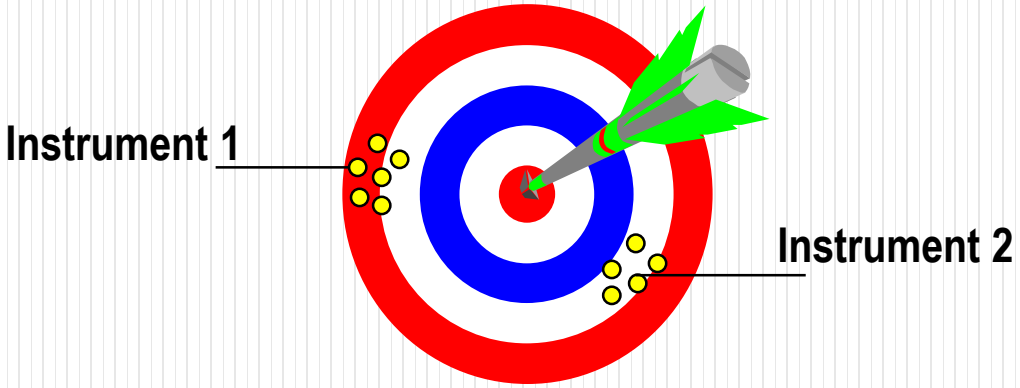
# Accuracy

- Accuracy is the difference (or offset) applied between the observed average of measurements and the true value. Establishing the true average is best determined by measuring the parts with the most accurate measuring equipment available or using parts that are of known value (i.e., standard calibration equipment).
- Instrument Accuracy differences between observed average measurement values and master value
- Master Value – determined by precise measurement based upon an accepted, traceable reference standard

# Potential Bias Problems

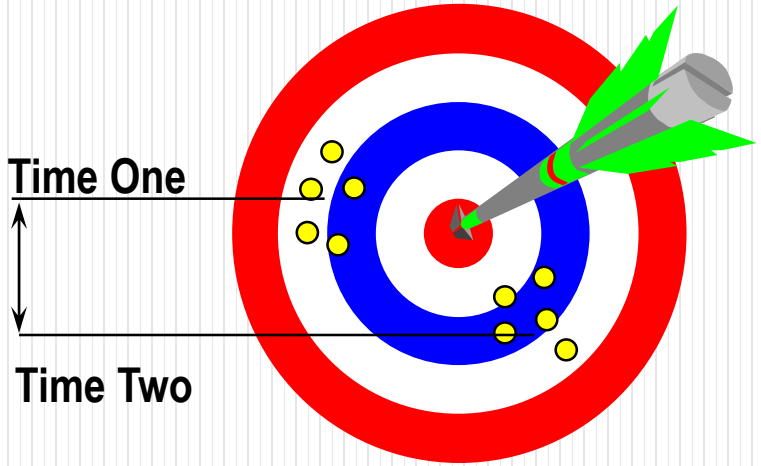
Measurement averages are different by fixed amount

- Bias culprits include:
  - Operator – Different operators get detectable different averages for the same value
  - Instrument – Different instruments get detectable different averages for the same measurement
  - Other – Day-to-day (environment), fixtures, customer, and supplier (sites)



# Stability

- Stability refers to the difference in the average of at least two sets of measurements obtained with the same Gage on the same parts taken at different times.
- If measurements do not change or drift over time, the instrument is considered to be stable



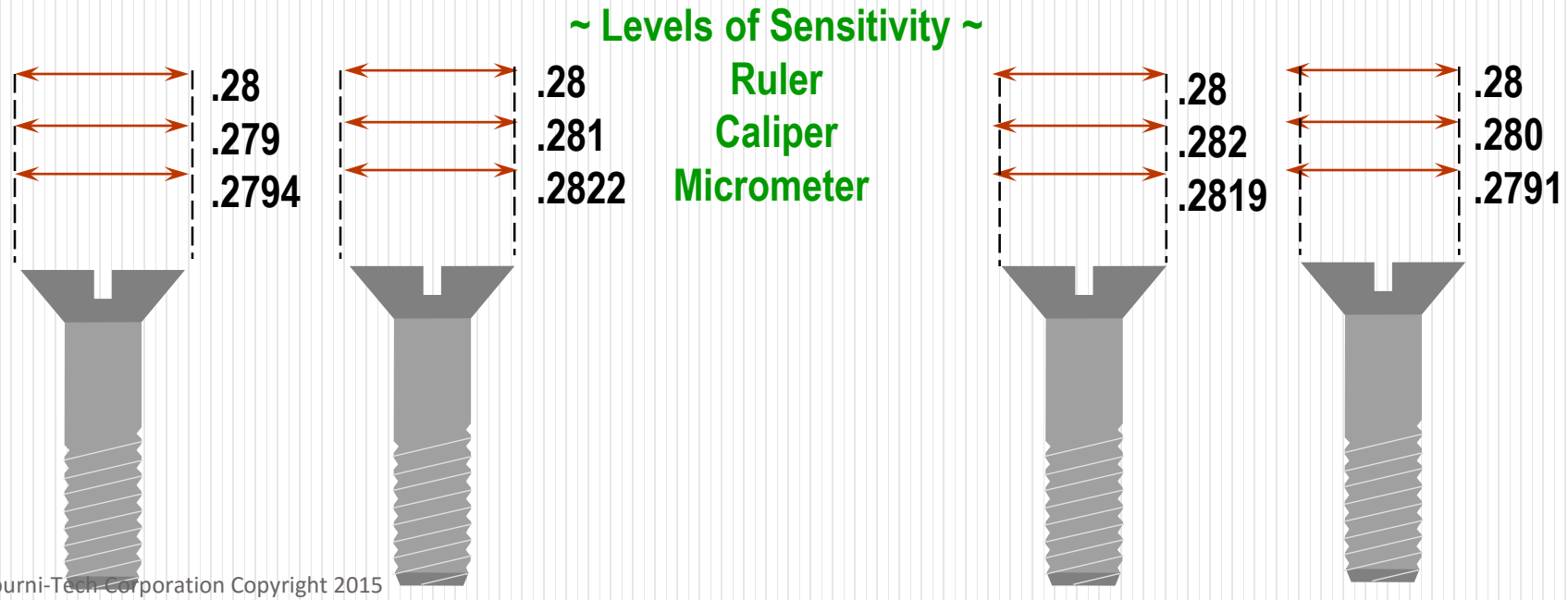
# Linearity

- Linearity is the difference in the accuracy of values throughout the expected operating range.

- Adequate Gage selection criteria (or Gage qualification) will eliminate linearity issues. The Gage qualification should incorporate selecting a Gage that is linear throughout the range of the specification

# Discrimination

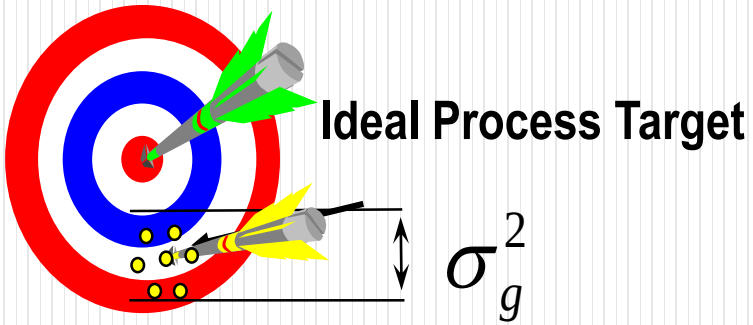
- Discrimination is the capability of detecting small measurement characteristic changes (gage sensitivity)
- Instrument may not be appropriate to identify process variation or quantify individual part characteristic values if discrimination is unacceptable
- If instrument does not allow process differentiation between common and special cause variations, it is unsatisfactory



# Repeatability

- Repeatability of the instrument is a measure of the variation obtained when one operator uses the same device to "repeatedly" measure the identical characteristic on the same part. Repeatability must also account for repeat measurements taken on an automated piece of test equipment (i.e., no operator).
- Goes to gage precision
- Variation between successive measurements of:
  - Same part / service
  - Same characteristic
  - By the same person using the same equipment (gage)

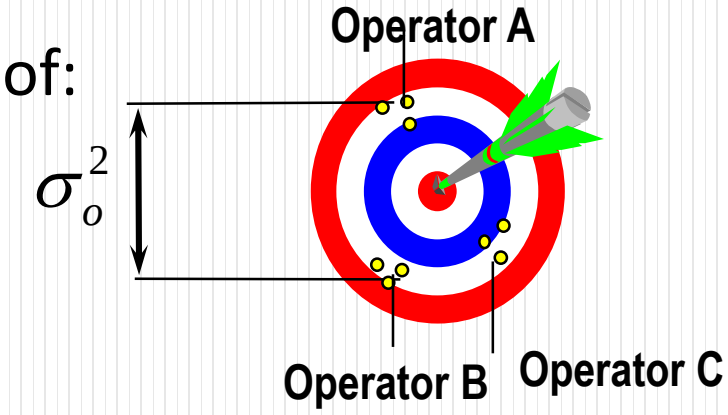
**Quantifies the repeatability of the instrument**



# Reproducibility

- Reproducibility is the variation in the averages of measurements made by different operators using the same device when measuring identical characteristics of the same parts. Reproducibility must also account for variation between different measuring devices (not only different appraisers).
- Operator Precision is the variation in the average of:
  - Measurements made by different operators
  - Using the same measuring instrument
  - When measuring the identical characteristic on the same part

**Quantifies the differences between the operators**

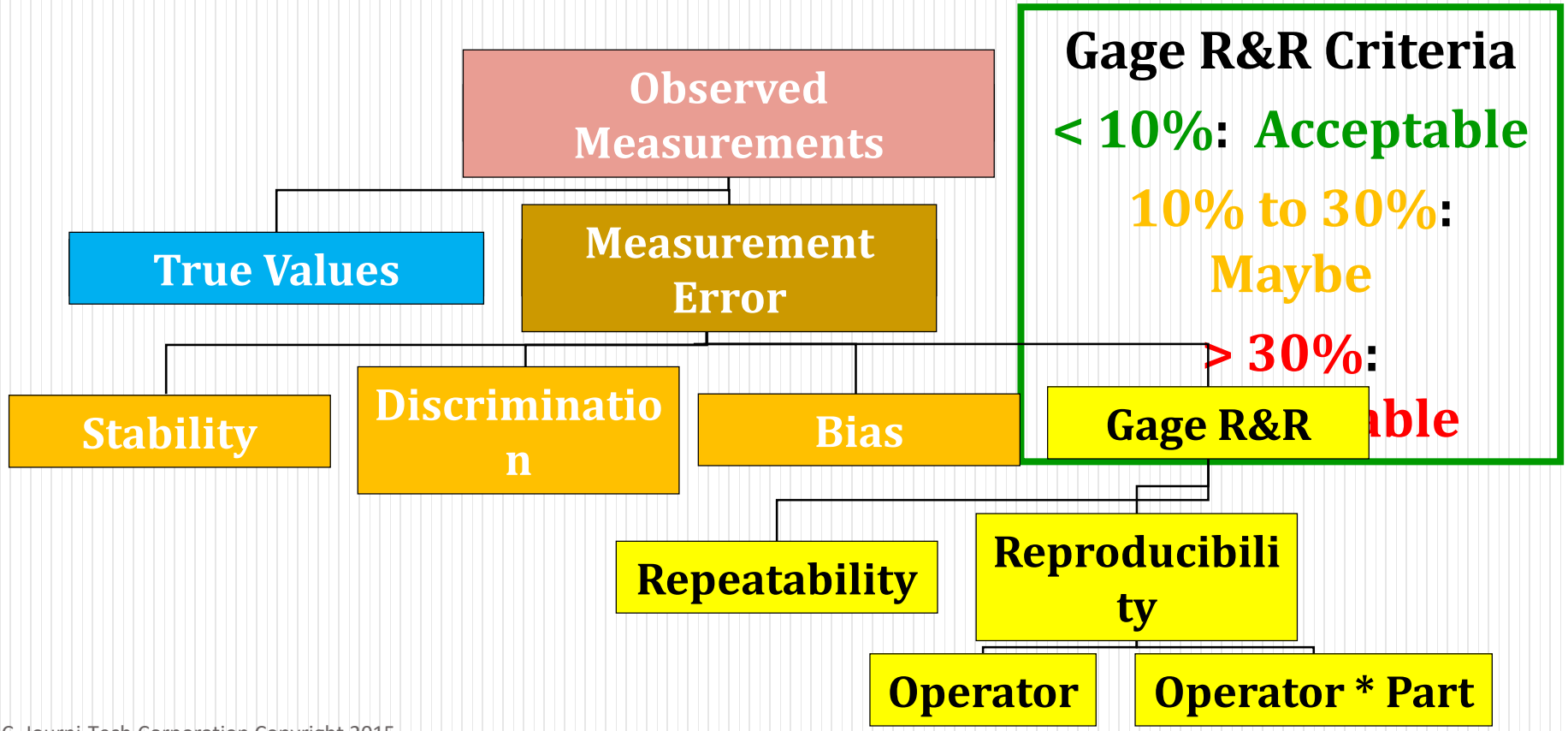


# Measurement Variation

- Measurement Variation relates to the instrument or gage
  - Consists of two components:  (2 R's of Gage R&R)
- Repeatability (Equipment / Gage Variability)
  - Given individual gets different measurements for the same thing when measured multiple times
- Reproducibility (Operator Variability)
  - Different individuals get different measurements for the same thing
- Tool used to determine the magnitude of these two sources of measurement system variation is called Gage R&R

# Measurement Error

**Gage R&R variation is the percentage that measurement variation (Repeatability & Reproducibility) represents of observed process variation**



# Acceptance Guidelines (By Method)

- There are three common methods used to qualify a measurement system:
  - % contribution
  - % study variation
  - Distinct categories
- We will use % contribution.
- The guidelines for each method are shown below.

| | % Contribution | % Study Variation | Distinct Categories |
|---|---|---|---|
| No issues with the measurement system | <5% | <10% | >10 |
| Depends on criticality and cost | 5% to 15% | 10% to 30% | 5 to 9 |
| Reject the measurement system | >15% | >30% | <5 |

# AIAG Gage R&R Standards

- The Automotive Industry Action Group (AIAG) has two recognized standards for Gage R&R:
  - Short Form – Five samples measured two times by two different individuals.
  - Long Form – Ten samples measured three time each by three different individuals.

# Measurement System Study Plan

- Select number of appraisers, number of samples, and number of repeat measures.
  - Use at least 2 appraisers and 5 samples, where each appraiser measures each sample at least twice (all using same device).
  - Select appraisers who normally do the measurement.
  - Select samples from the process that represent its entire operating range. Label each sample discretely so the label is not visible to the operator.
- Check that the instrument has a discrimination that is equal to or less than 1/10 of the expected process variability or specification limits.

# Running the Measurement Study

- Each sample should be measured 2-3 times by each operator.
- Make sure the parts are marked for ease of data collection but remain “blind” (unidentifiable) to the operators.
- Be there for the study.  Watch for unplanned influences.
- Randomize the parts continuously during the study to preclude operators influencing the test.

# Running the Study – Guidelines

- We are unsure of how noise can affect our measurement system, so use the following procedure:
  - Have the first operator measure all the samples once in random order.
  - Have the second operator measure all the samples once in random order.
  - Continue until all operators have measured the samples once (this is Trial 1).
  - Repeat steps 2 - 4 for the required number of trials.
  - Use a form to collect information.
  - Analyze results.
  - Determine follow-up action, if any.

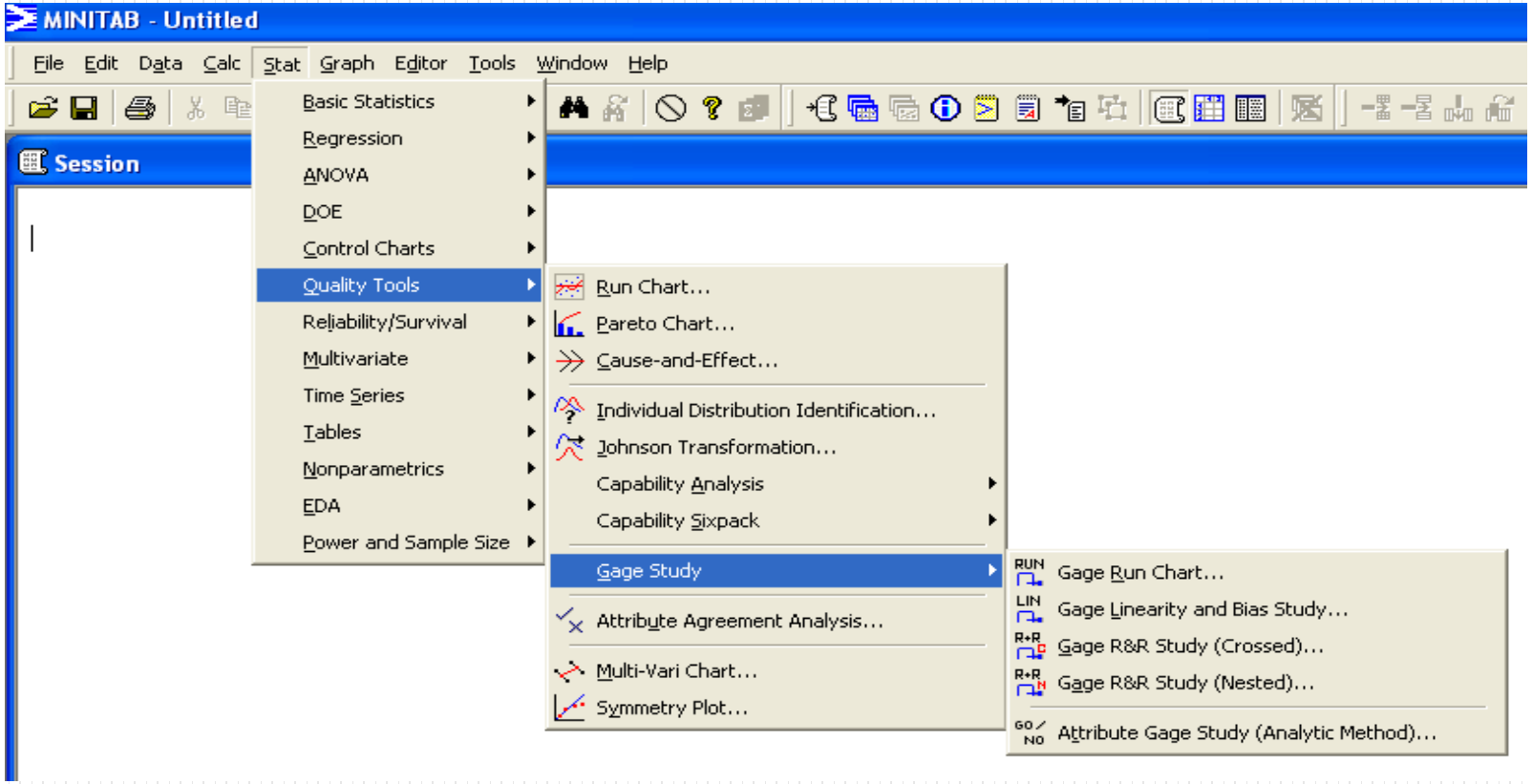
# MSA Example in Minitab

A project is looking at controlling the thickness of steel from a rolling process. A Gage R&R study has been completed on 10 pieces of steel using 3 different appraisers. The data can be found in "C:/Program Files (X86)/minitab/minitab17/English/Sample Data/Thickness.mtw."

| Column | Name | Description |
|---|---|---|
| C1 | Part | Steel Part Number |
| C2 | Appraiser | Appraiser Number |
| C3 | Measurement | Steel Thickness |

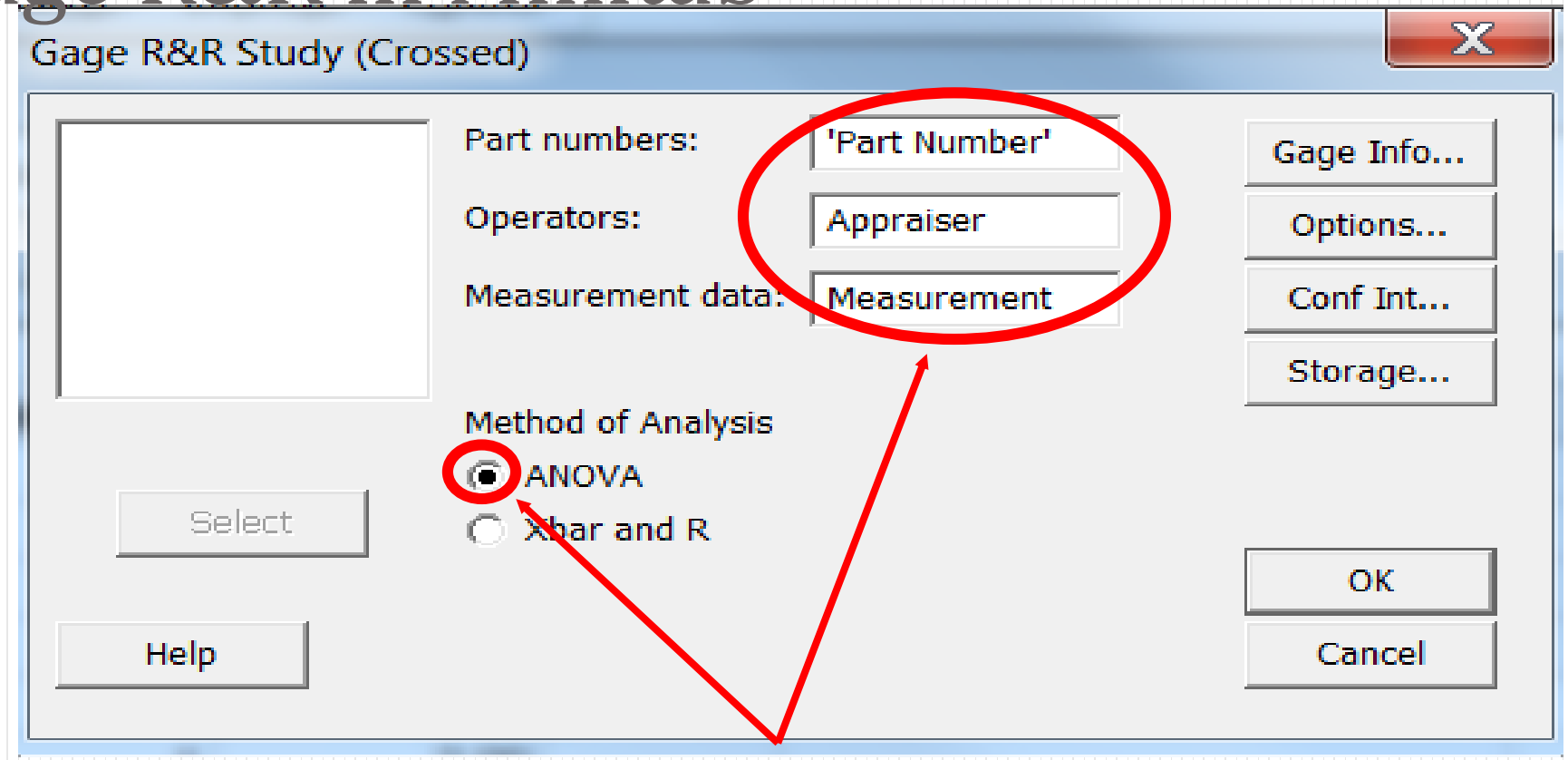
# MSA – Gage R&R in Minitab

## Stat > Quality Tools >Gage Study > Gage R&R Study (Crossed)




**Note**: **Gage R&R Study (Crossed) is the most commonly used method for Variables (Continuous Data). It is used when the *same parts* can be tested multiple times.**

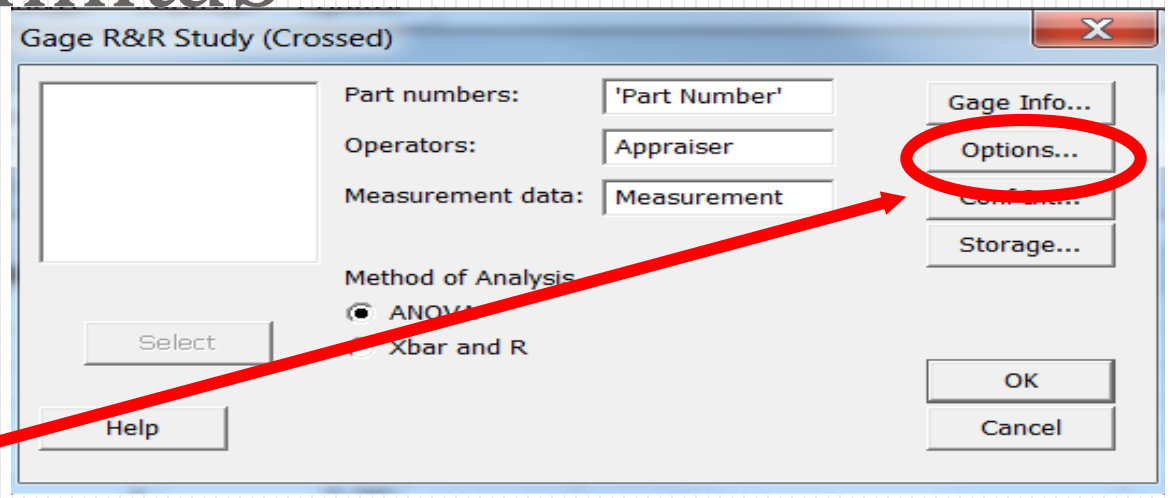
# Gage R&R in Minitab

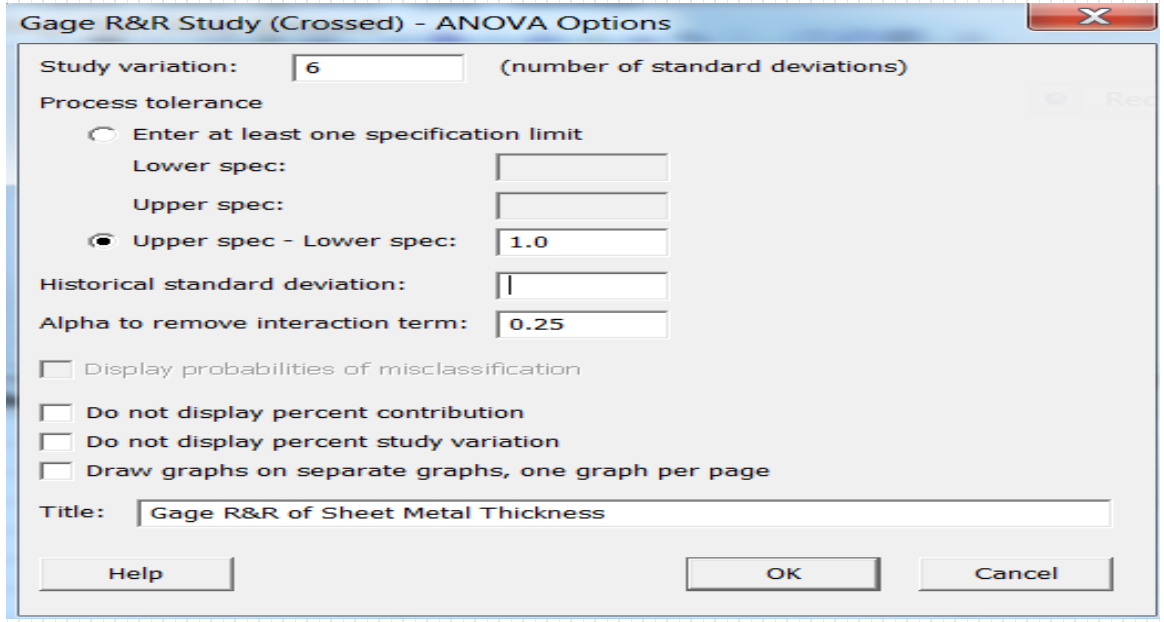



**Enter the variables (circled fields) in the above dialogue box and keep the ANOVA method of analysis checked**

# Gage R&R in Minitab

**After entering the variables in this dialog box, click on Options**



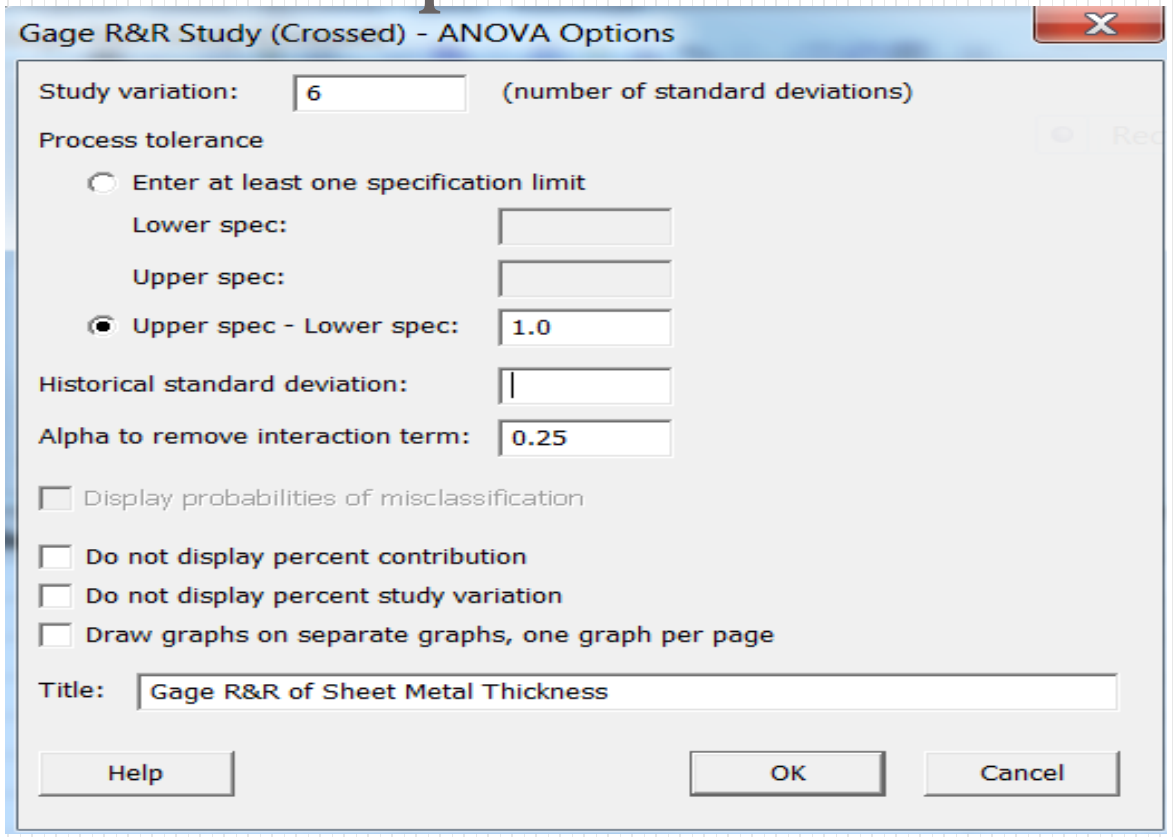




**Options dialog box**

# Gage R&R in Minitab – Options

**6.0 is the default for the <u>Study variation</u>.**

**This is the Z value range that calculates a 99.73% potential Study Variation based on the calculated Standard Deviation of the variation seen in the parts chosen for the study.**




**The Spec Limits for the process are 2.3 as the USL and 1.3 as the LSL. The Upper Spec- Lower Spec (process tolerance) is 2.3 – 1.3 = 1.0.**

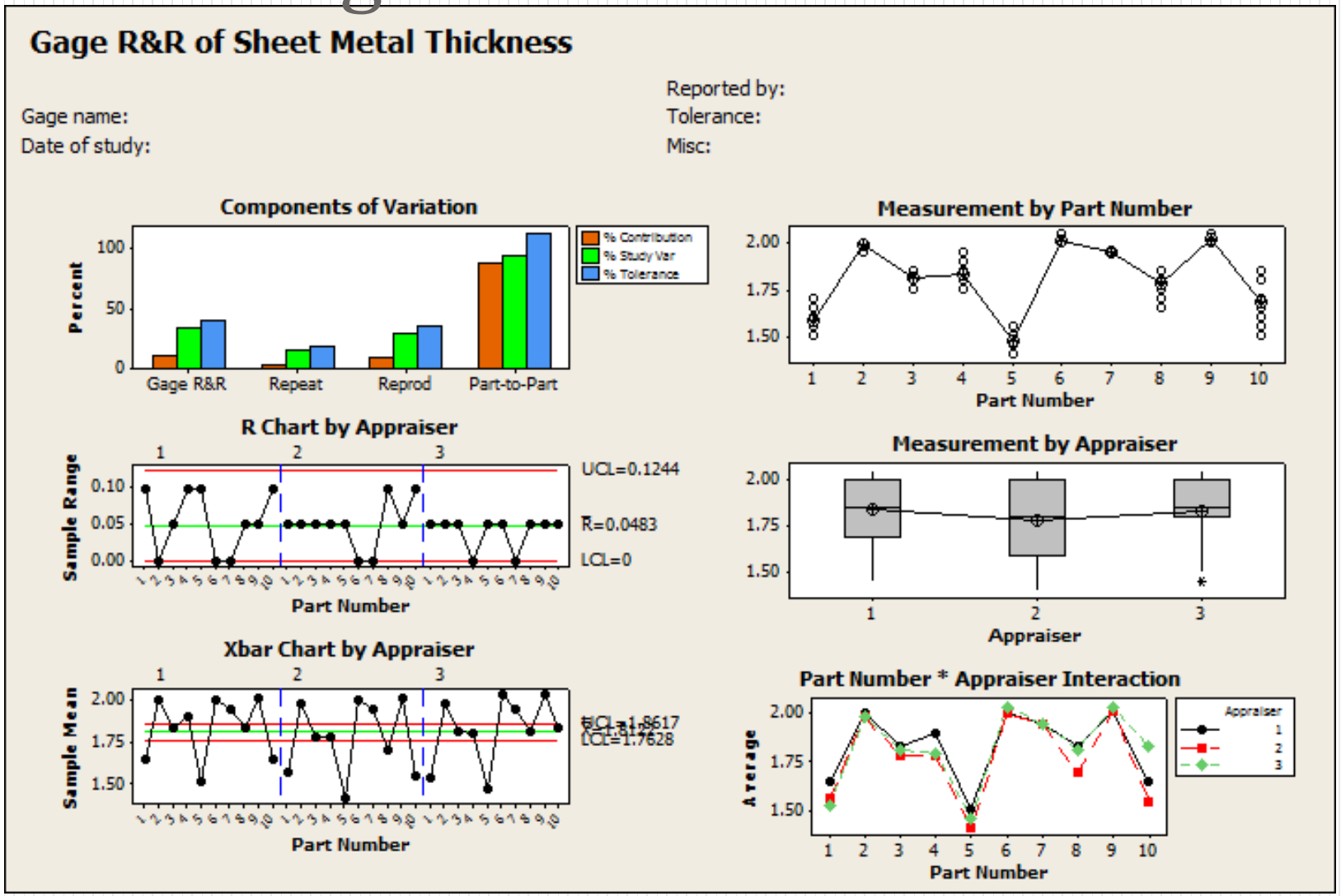
**Enter the Title of the Graph**

# Acceptability

Remember that the guidelines are:

- < 10 % – Acceptable
- 10 - 30 % – Marginal
  - May be acceptable based upon the risk of the application, cost of measurement device, cost of repair, etc.
- > 30 % – Not Acceptable
  - Every effort should be made to improve the measurement system.

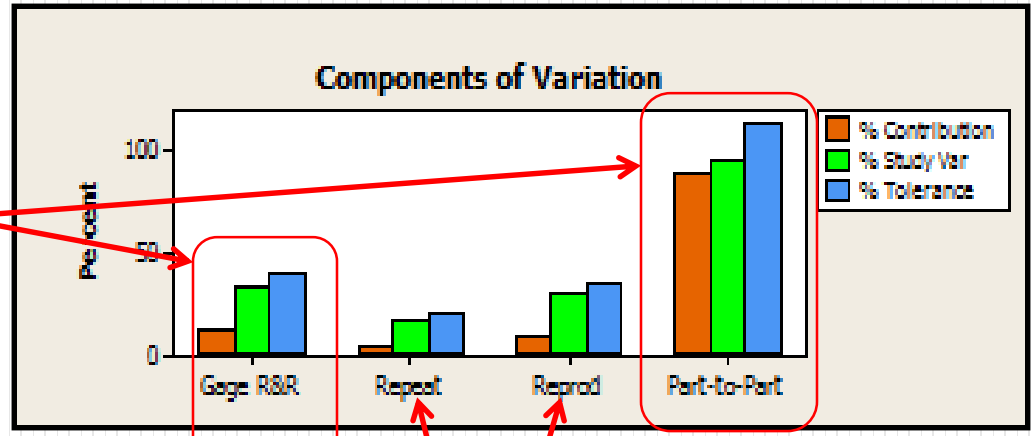
# Minitab – Gage R&R – Six-Pack



# Components of Variation

**The Gage R&R Bars should be small in comparison to the Part-to-Part Bars:**

- **First Bar- % Contribution**
- **Second Bar- % Study Variation (Total Variation)**
- **Third Bar- % of Tolerance**

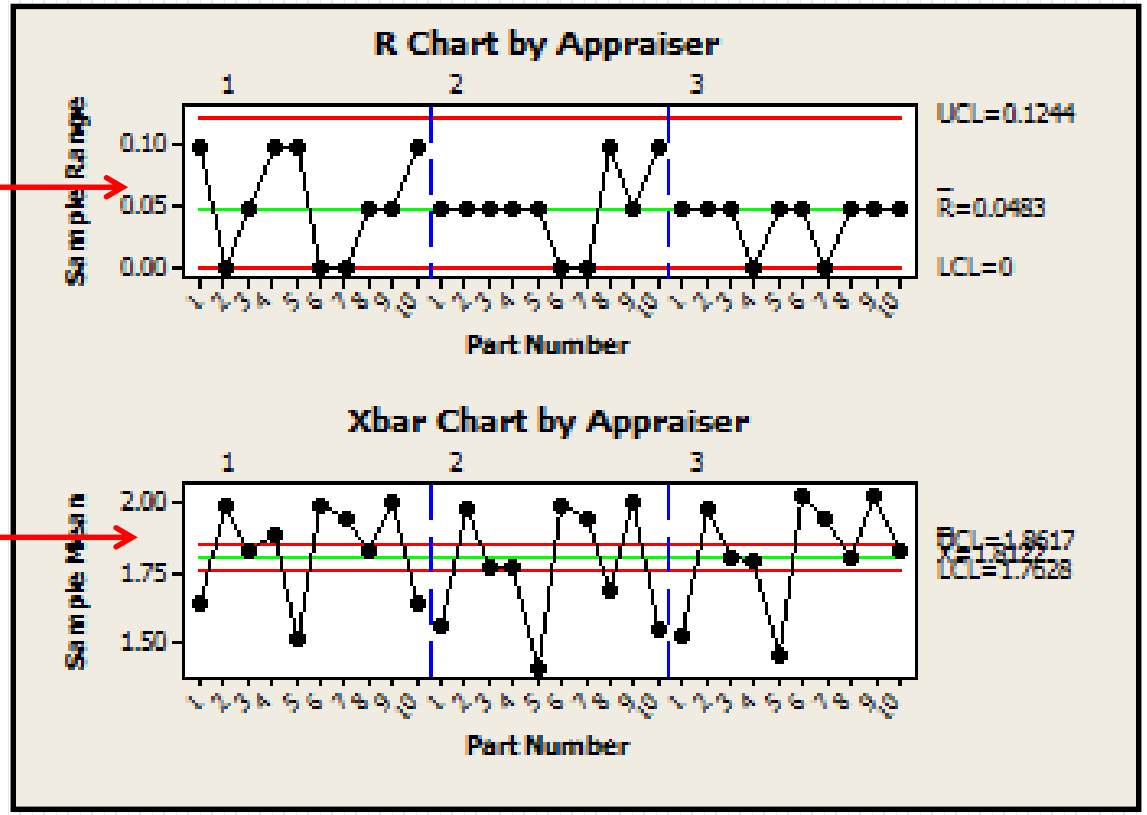



**Reproducibility is larger than Repeatability, indicating that improvements should focus on reducing the differences between appraisers first.**

# R Chart and Xbar Chart

If any points are outside the red lines, check for problems with the part.

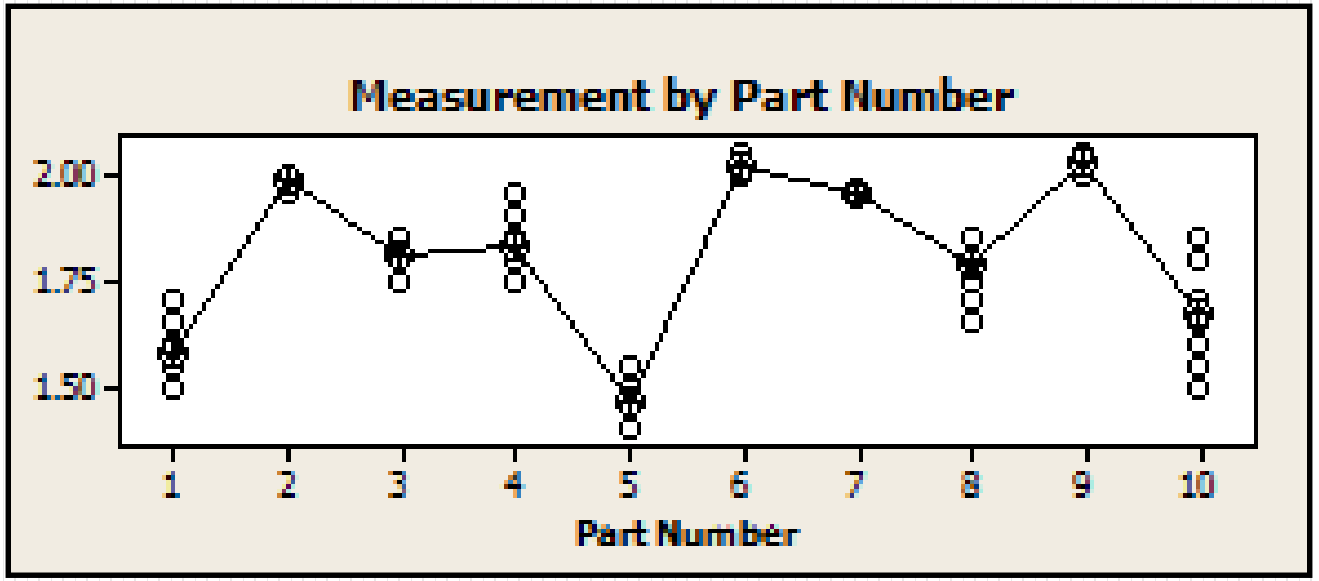
In contrast, this chart should have points outside the lines, which indicates the Gage R&R is low.



# Measurement by Part Number

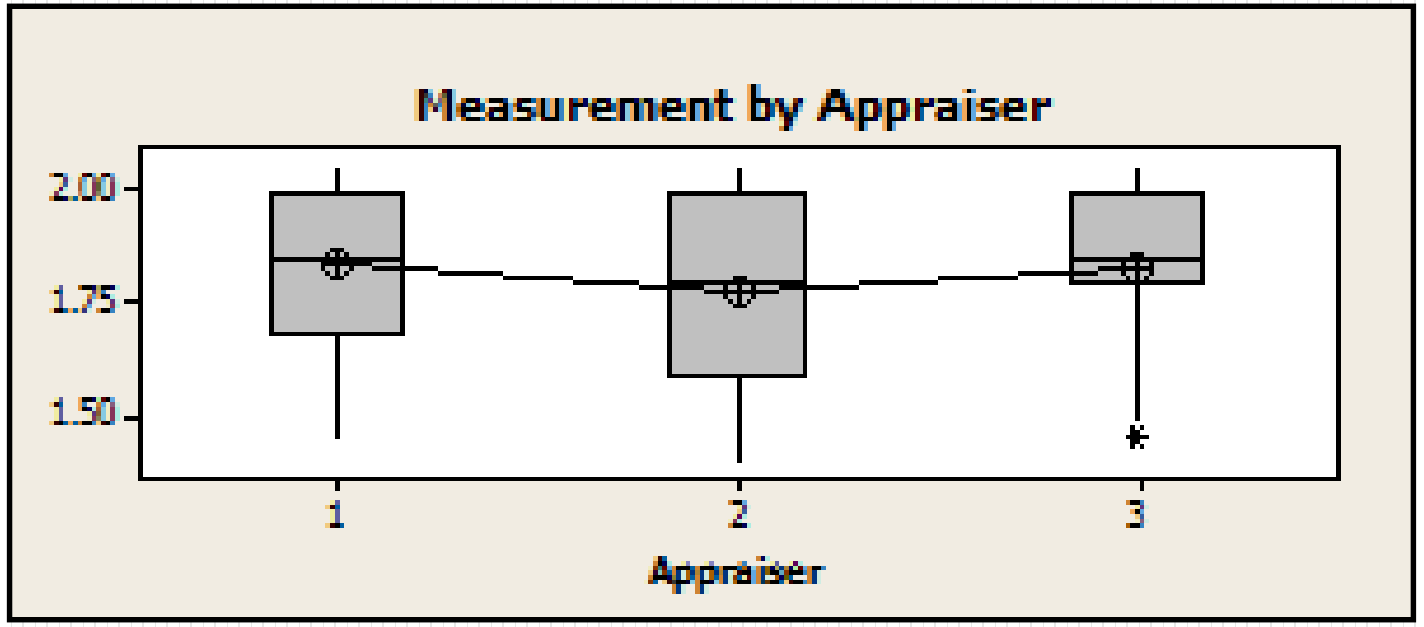
- **This chart shows the results of each part in order (1-10) to see if particular parts were hard to measure.**

- **Part 10 has the most variability.**



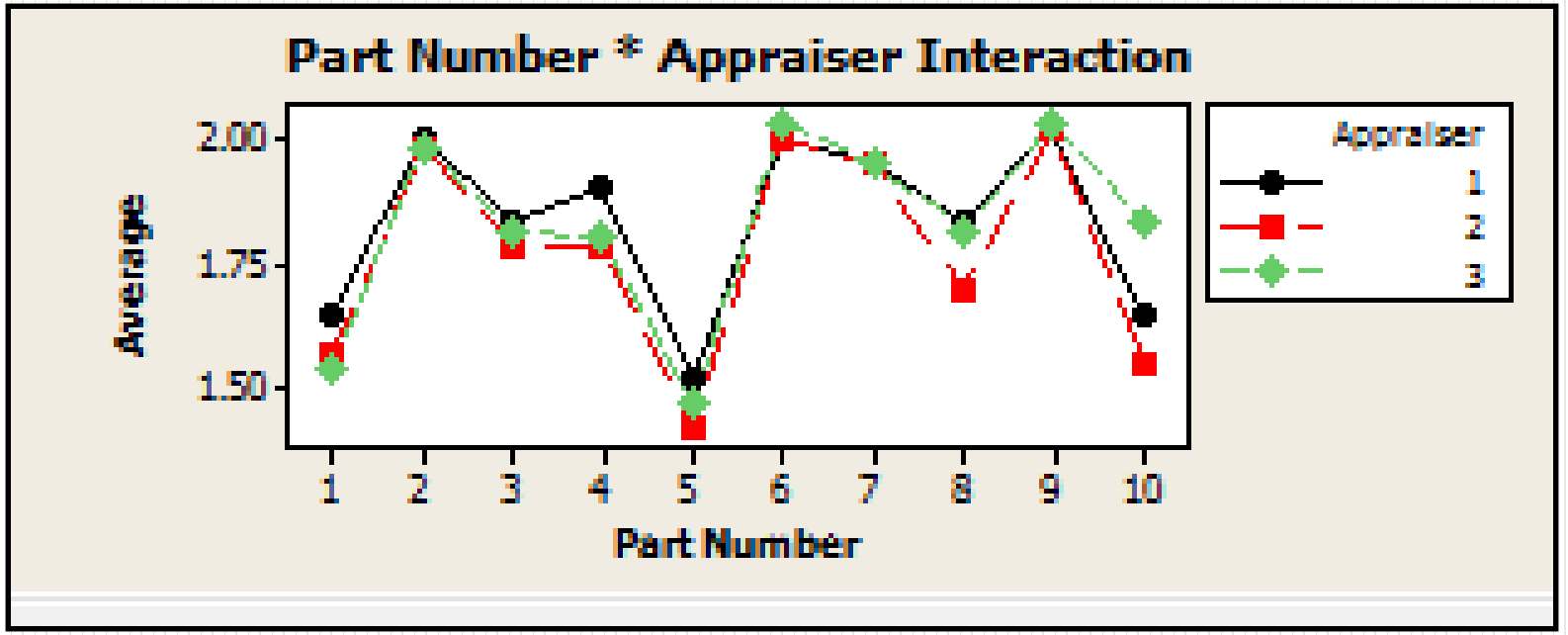
# Measurement by Appraiser

- This chart shows reproducibility for each appraiser.
- Appraiser 2 has lower measurements on average which may require some investigation.



# Part Number * Appraiser Interaction

- This chart is the same as the Measurement by Part Number chart, however, the results by appraiser are separated out.



# Gage R&R Study - ANOVA Method

## Gage R&R Study - ANOVA Method

### Two-Way ANOVA Table With Interaction

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Part Number | 9 | 2.92322 | 0.324802 | 36.5530 | 0.000 |
| Appraiser | 2 | 0.06339 | 0.031694 | 3.5669 | 0.050 |
| Part Number * Appraiser | 18 | 0.15994 | 0.008886 | 8.8858 | 0.000 |
| Repeatability | 60 | 0.06000 | 0.001000 | | |
| Total | 89 | 3.20656 | | | |

Alpha to remove interaction term = 0.25

**The ANOVA table assess which sources of variation are statistically significant.**

**The appraiser does have an affect on the result and there is an interaction between part number and appraiser (both p-values are .05 or less).**

# Gage R&R Output

## Gage R&R

| Source | VarComp | %Contribution (of VarComp) |
|---|---|---|
| Total Gage R&R | 0.0043889 | 11.11 |
| Repeatability | 0.0010000 | 2.53 |
| Reproducibility | 0.0033889 | 8.58 |
| Appraiser | 0.0007603 | 1.93 |
| Appraiser*Part Number | 0.0026286 | 6.66 |
| Part-To-Part | 0.0351019 | 88.89 |
| Total Variation | 0.0394907 | 100.00 |

**The Total Gage R&R variation is 11.11%, which is composed of the Repeatability of 2.53% plus the Reproducibility of 8.58%.**

**The part-to-part variability across all measurements is 88.89%.**

**Ideally, very little variability should come from Repeatability and Reproducibility.**

# Gage R&R Output

Process tolerance = 1

| Source | StdDev (SD) | Study Var (6 * SD) | %Study Var (%SV) | %Tolerance (SV/Toler) |
|---|---|---|---|---|
| Total Gage R&R | 0.066249 | 0.39749 | 33.34 | 39.75 |
| Repeatability | 0.031623 | 0.18974 | 15.91 | 18.97 |
| Reproducibility | 0.058214 | 0.34928 | 29.29 | 34.93 |
| Appraiser | 0.027573 | 0.16544 | 13.88 | 16.54 |
| Appraiser*Part Number | 0.051270 | 0.30762 | 25.80 | 30.76 |
| Part-To-Part | 0.187355 | 1.12413 | 94.28 | 112.41 |
| Total Variation | 0.198723 | 1.19234 | 100.00 | 119.23 |

Number of Distinct Categories = 3

**The number 3 is the Number of Distinct Categories that the measurement system is capable of discriminating within the process variation. An acceptable target is 5, so this reinforces the conclusion that the measurement system needs improvement.**

**The Gage R&R is 33.34% of the Total Variation and 39.75% of the Tolerance, which is > 30%, indicating improvement is required with the measurement system.**

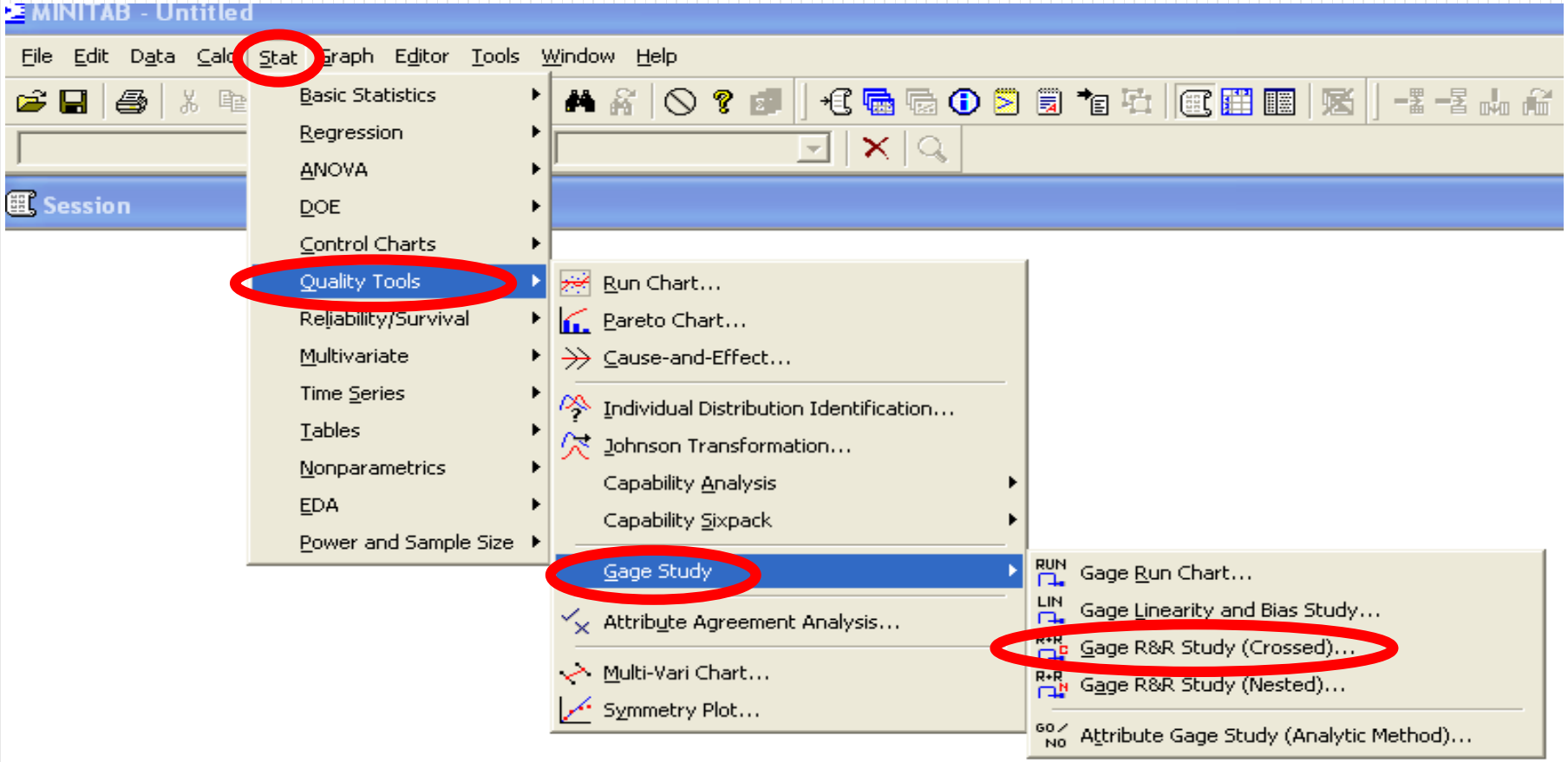
# Let's Do It Again

- Three parts were selected that represent the expected range of the process variation.
- Three operators measured the three parts, three times per part, in a random order.
- No History of the process is available and Tolerances are not established.
  - Open Minitab file "C:/Program Files (X86)/minitab/minitab17/English/Sample Data/Gage2.mtw"
- This data set is used to illustrate Gage R&R Study and Gage Run Chart.

| Column | Name | Count | Description |
|---|---|---|---|
| C1 | Part | 27 | Part number |
| C2 | Operator | 27 | Operator number |
| C3 | Response | 27 | Measurement value |
| C4 | Trial | 27 | Trial number |

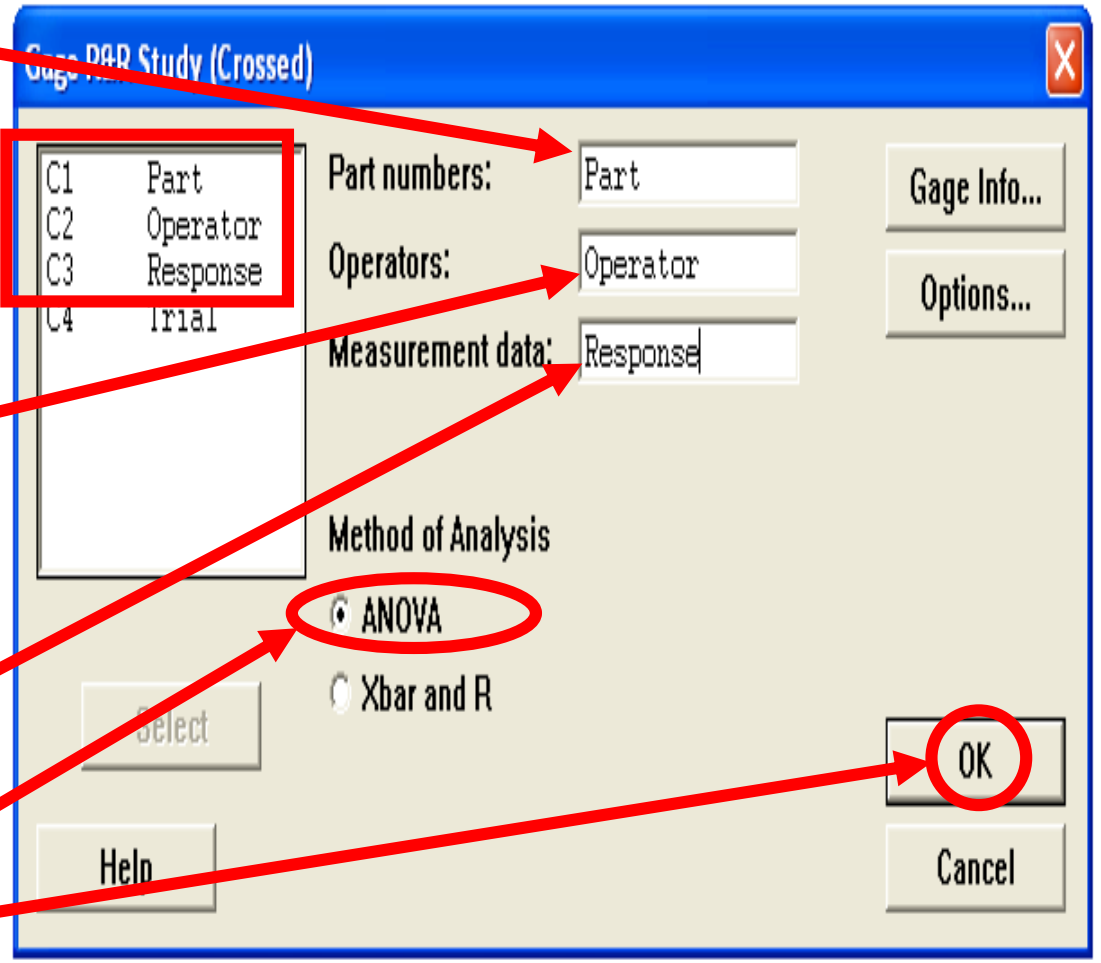
# Minitab – Gage R&R

**Stat > Quality Tools > Gage Study > Gage R&R Study (Crossed)**

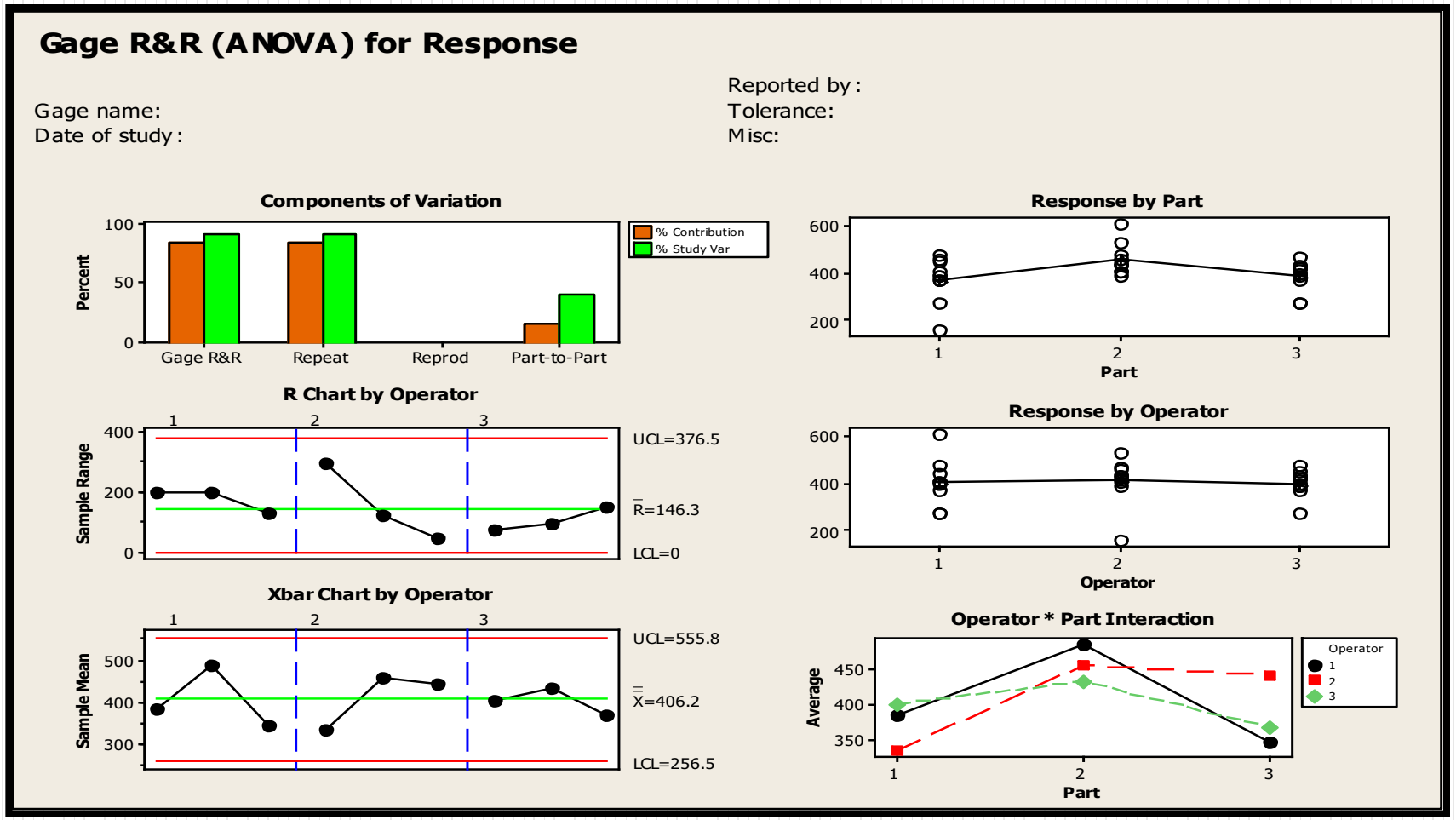


# Filling in the Dialogue Boxes

1. Set cursor in *Part numbers* box and double click on C-1 Part.
2. Set cursor in *Operators* box and double click on C-2 Operator.
3. Set cursor in *Measurement data* box and double click on C-3 Response.
4. Make sure *ANOVA* is selected and click on OK.



# Is This Study Unacceptable?



# Gage2.mtw – Results

**Gage R&R**

| Source | VarComp | %Contribution (of VarComp) |
|---|---|---|
| Total Gage R&R | 7304.67 | 84.36 |
| Repeatability | 7304.67 | 84.36 |
| Reproducibility | 0.00 | 0.00 |
| Operator | 0.00 | 0.00 |
| Part-To-Part | 1354.50 | 15.64 |
| Total Variation | 8659.17 | 100.00 |

| Source | StdDev (SD) | Study Var (6 * SD) | %Study Var (%SV) |
|---|---|---|---|
| Total Gage R&R | 85.4673 | 512.804 | 91.85 |
| Repeatability | 85.4673 | 512.804 | 91.85 |
| Reproducibility | 0.0000 | 0.000 | 0.00 |
| Operator | 0.0000 | 0.000 | 0.00 |
| Part-To-Part | 36.8036 | 220.821 | 39.55 |
| Total Variation | 93.0547 | 558.328 | 100.00 |

Number of Distinct Categories = 1

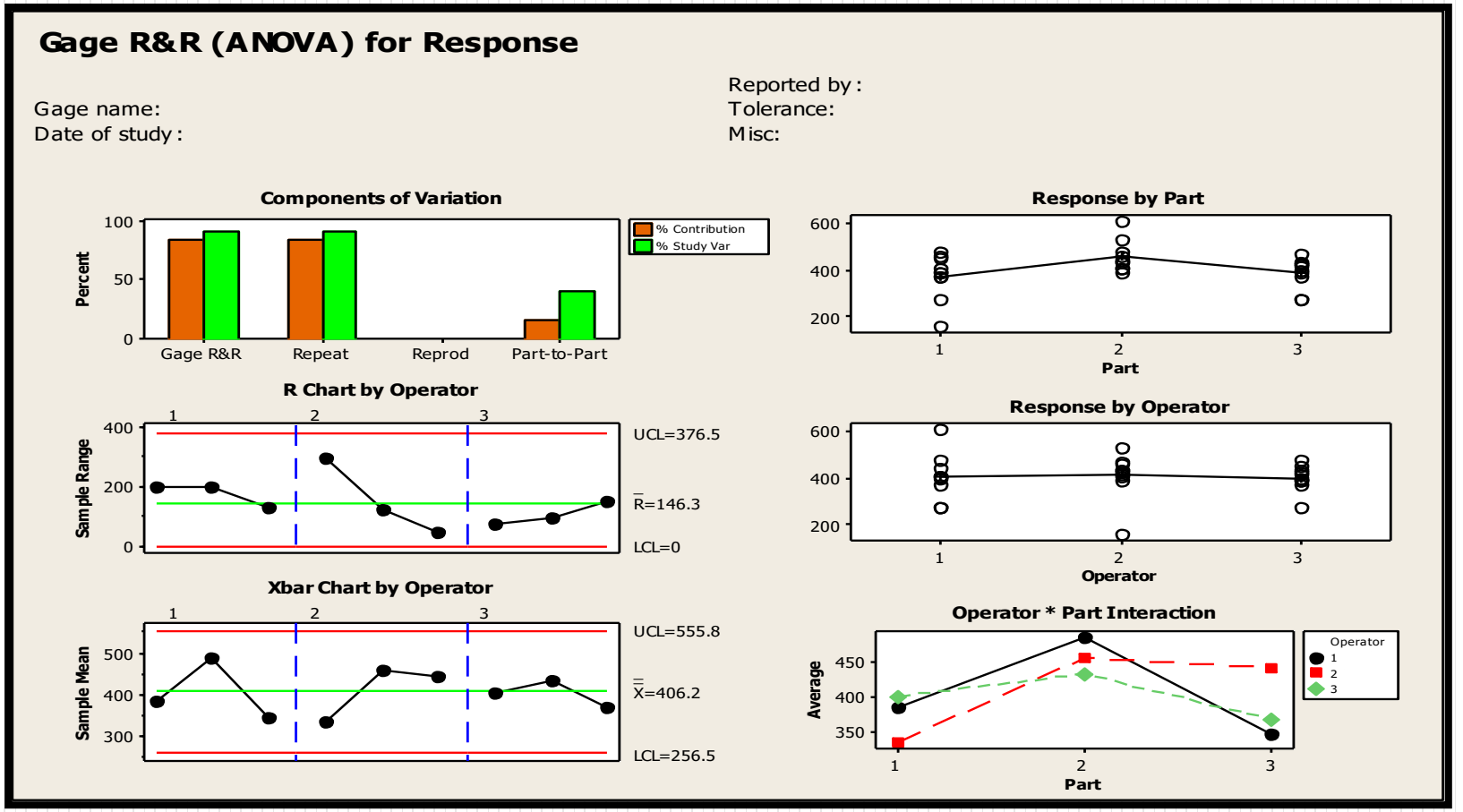
This should be less than 30% for process improvement efforts

What does this tell you?

Remember this? What does this mean?

# Gage2.mtw – Conclusions

What needs to be addressed first? Where do we begin improving this measurement system?



# Example: Price Quoting Process

- Work orders are called in by customers to a repair facility. An analyst looks at the work orders and tries to estimate a price to complete the work order. The price is then quoted to the customer.
- Bill Black Belt believed that the variability in the price quoting process was a key factor in customer satisfaction.
- Bill had received customer feedback that the pricing varied from very competitive to outrageous. It was not uncommon for a customer to get a job quoted one week, submit a near-identical job the next week and see a 35% difference in price.
- Help Bill determine how he might estimate the amount of error in the quoting process, especially with respect to repeatability and reproducibility.

# Example: Price Quoting Process

- Bill decided to set-up 10 fake customer pricing requests and have three different inside salespeople quote each one three times over the next two weeks.
- Due to the large variety of products the organization offered, Bill chose pricing requests that the sales manager calculated to be at $24,000.
- The department had enough volume coming through that Bill felt comfortable they would not recognize the quote, but he altered some unimportant customer information just to be sure.
- What would the AIAG call Bill's MSA?
- How else might Bill have conducted his study?

# Price Quoting Process

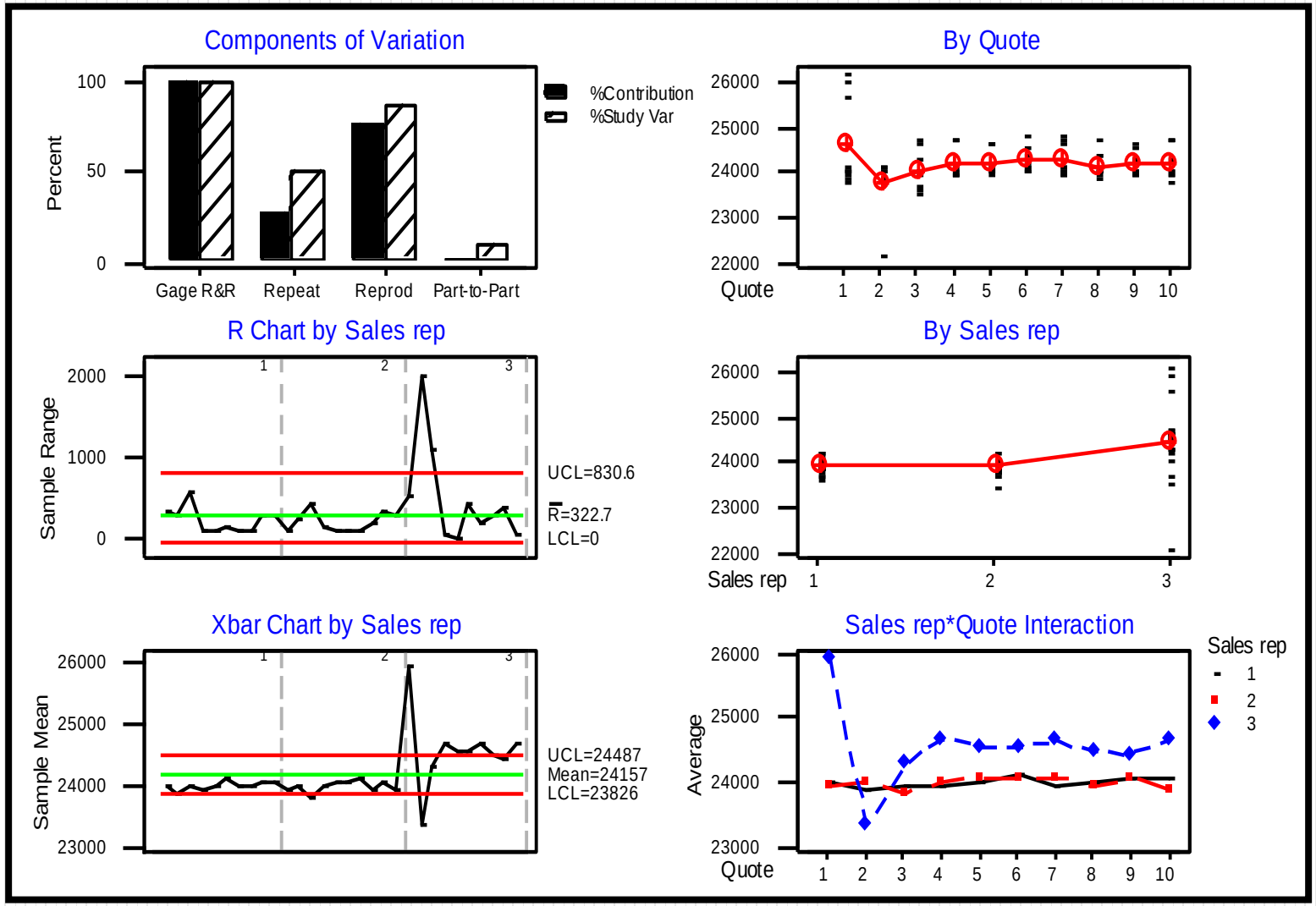
MINITAB - Untitled - [msa tranactional.MTW ***]

File Edit Manip Calc Stat Graph Editor Window H

| ↓ | C1 | C2 | C3 | C4 | C5 |
|---|---|---|---|---|---|
| | Quote | Sales rep | attempt | price | |
| 1 | 1 | 1 | 1 | 24000 | |
| 2 | 1 | 1 | 2 | 23750 | |
| 3 | 1 | 1 | 3 | 24100 | |
| 4 | 1 | 2 | 1 | 23900 | |
| 5 | 1 | 2 | 2 | 23850 | |
| 6 | 1 | 2 | 3 | 23950 | |
| 7 | 1 | 3 | 1 | 25600 | |
| 8 | 1 | 3 | 2 | 26125 | |
| 9 | 1 | 3 | 3 | 26000 | |
| 10 | 2 | 1 | 1 | 24000 | |
| 11 | 2 | 1 | 2 | 23850 | |
| 12 | 2 | 1 | 3 | 23700 | |
| 13 | 2 | 2 | 1 | 24100 | |
| 14 | 2 | 2 | 2 | 24000 | |
| 15 | 2 | 2 | 3 | 23825 | |
| 16 | 2 | 3 | 1 | 22100 | |

Here is the data Bill collected
(Partial data set shown)

# MSA Transactional Graphs… Your Thoughts?



*MSA Transaction:*

# What Do We Work on First?

| Source | % Contribution VarComp | % Contribution (of VarComp) |
|---|---|---|
| Total Gage R&R | 278,556 | 99.08 |
| *Repeatability* | 70,466 | 25.06 |
| *Reproducibility* | 208,091 | 74.01 |
| *Sales Rep* | 99,794 | 35.49 |
| *Sales Rep * Quote* | 108,296 | 38.52 |
| Part-To-Part | 2,597 | 0.92 |
| Total Variation | 281,154 | 100.00 |

| Source | StdDev (SD) | Study Var (5.15 * SD) | % Study Var (% SV) |
|---|---|---|---|
| Total Gage R&R | 527.785 | 2,718.09 | 99.54 |
| *Repeatability* | 265.454 | 1,367.09 | 50.06 |
| *Reproducibility* | 456.170 | 2,349.27 | 86.03 |
| *Sales Rep* | 315.902 | 1,626.90 | 59.58 |
| *Sales Rep * Quote* | 329.084 | 1,694.78 | 62.06 |
| Part-To-Part | 50.963 | 262.46 | 9.61 |
| Total Variation | 530.239 | 2,730.73 | 100.00 |

This value should be less than 30% for process improvement efforts

Number of Distinct Categories = 0

What does this mean?

# Attribute Gage R&R

- All the same principles of Variable Gage R&R can be applied to the Attribute data world as well.
- The target for an Attribute MSA is for it to reach the correct decision, every time.
- Key differences of Attribute Gage R&R studies are:
  - More data is required, because the Attribute data world has less resolution. At least 20 parts should be assessed at least 3 times by each appraiser.
  - You should ensure your selection of parts includes some borderline products or services that will really challenge the capability of the measurement system.

# Why Use Attribute Gage R&R?

- To determine if inspectors across all shifts, all machines and so on, use the same criteria to determine “good” from “bad”
- To assess your inspection or workmanship standards against your customer’s requirements
- To identify how well these inspectors are conforming to themselves
- To identify how well these inspectors are conforming to a “known master,” which includes:
  - How often operators decide to ship truly defective product
  - How often operators do not ship truly acceptable product
- To discover areas where:
  - Training is needed
  - Procedures are lacking
  - Standards are not defined

# MSA Attribute Classroom Exercise

**Purpose:** Practice attribute measurement analysis

**Agenda:** 1. Remain in your seats

2. Individually and in silence follow the instructions on each of the Inspection Exercise slides

Materials: Inspection Exercise slides

**Limit:** Exercise: 30 minutes

Discussion: 10 minutes

# Inspection Exercise

**Count the number of times the 6th letter of the alphabet appears in the following text:**

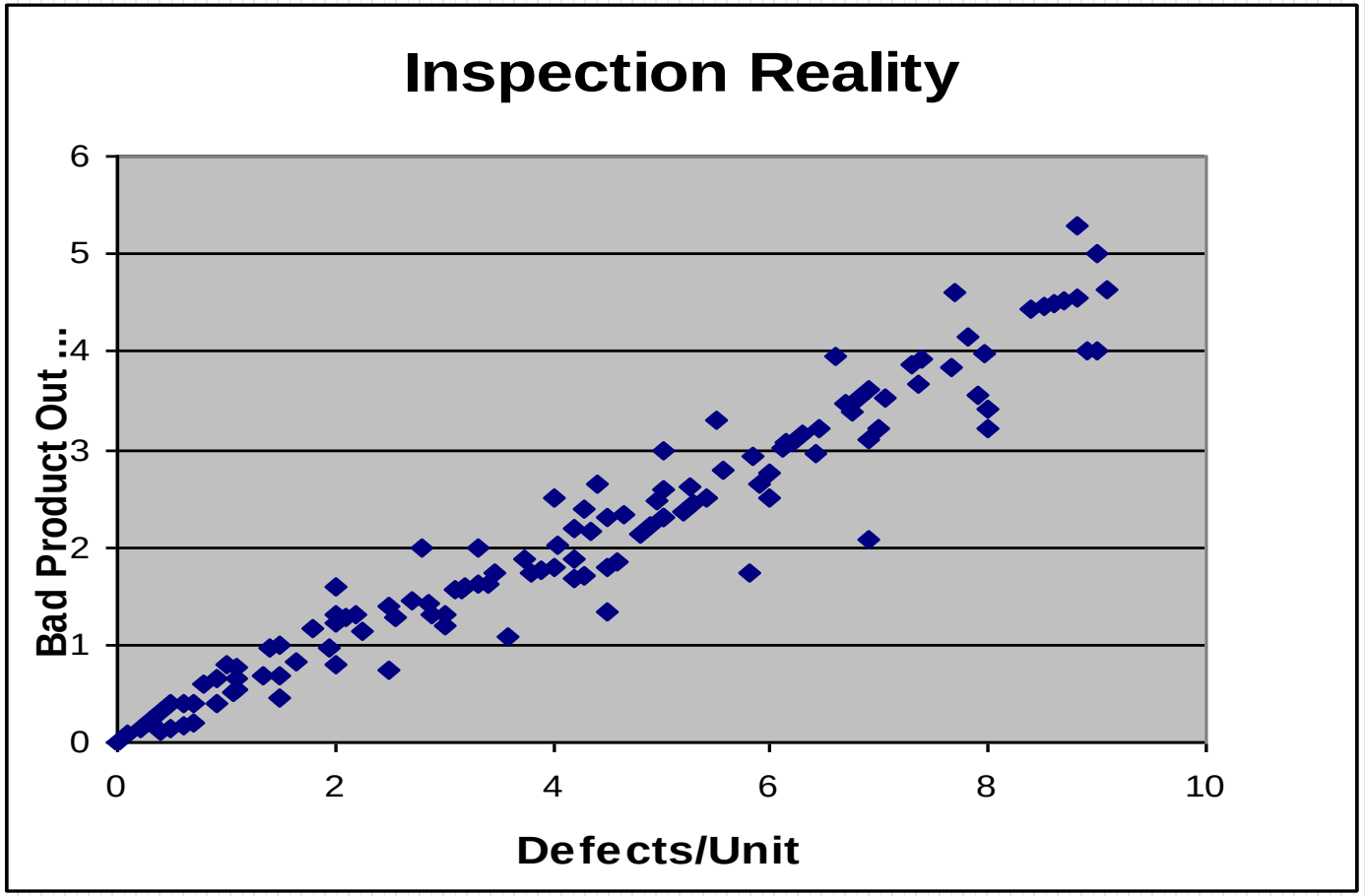
fatherly feeding of fresh farm raised fish because they believe it is the basis of good fundamental farm management.

# Inspection Exercise

**Count the number of times the letter "m" appears in the following text:**

**The need of training fish feeders for the first class fishing farms in the finest feeding methods of fresh fish is foremost in the eyes of the most famous fish farm owners. Since the forefathers of the current farm owners trained the first fresh fish feeders of all first class farms in the fatherly feeding of fresh fish, the farm owners felt they should carry on with the family tradition of training farm hands of the first class in the fatherly feeding of fresh farm raised fish because they believe it is the basis of good fundamental farm management.**

# Concept of Escaping Defects




**No matter how good you think your quality testing or audit plan is, the more defects you create, and the more defects you ultimately ship to your customer**

# How to Run an Attribute Gage R&R

- Select a minimum of 30 parts from the process.
  - 50% of the parts in your study should have defects.
  - 50% of the parts should be defect free
- If possible, select borderline (or marginal) good and bad samples
- Identify the inspectors who should be qualified
- Have each inspector independently and in random order assess these parts and determine whether or not they pass or fail (judgment of good or bad)

# How to Run an Attribute Gage R&R

- Use an Excel spreadsheet to report the effectiveness and efficiency of the attribute measurement system (inspectors and the inspection process)
- Document and implement appropriate actions to fix the inspection process (if necessary)
- Re-run the study to verify the fix

# Attribute Gage Terms

- **Attribute Measurement System:** compares parts to a specific set of limits and accepts the parts if the limits are satisfied.
- **Screen:** 100% evaluation of output using an attribute measurement system.
- **Screen Effectiveness (%):** ability of the attribute measurement system to properly discern good parts from bad.

# Attribute Gage Study

- Attribute data (Good/Bad)
- Compares parts to specific standards for Accept/Reject decisions
- Must screen for effectiveness to discern good from bad
- At least two associates and two trials each

# X-Ray Chart Illustrative Example

- X-rays are read by two technicians.
- Twenty X-rays are selected for review by each technician.
- Some X-rays have no problems and others have bone fractures.
- Objective: Evaluate the effectiveness of the measurement system to determine if there are differences in the readings.

# X-Ray Illustrative Example

- Twenty X-rays were selected that included good (no fracture) and bad (with fractures).
- Two technicians independently and randomly reviewed the 20 X-rays as good (no fracture) or bad (with fractures).
- Data are entered in spreadsheet and the Screen Effectiveness score is computed.

# X-Ray Illustrative Example

| | Associate A | | Associate B | | |
|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | Standard |
| 1 | G | G | G | G | G |
| 2 | G | G | G | G | G |
| 3 | NG | G | G | G | G |
| 4 | NG | NG | NG | NG | NG |
| 5 | G | G | G | G | G |
| 6 | G | G | NG | G | G |
| 7 | NG | NG | G | NG | NG |
| 8 | NG | NG | G | G | NG |
| 9 | G | G | G | G | G |
| 10 | G | G | G | NG | G |
| 11 | G | G | G | G | G |
| 12 | G | G | G | G | G |
| 13 | G | NG | G | G | G |
| 14 | G | G | G | G | G |
| 15 | G | G | G | G | NG |
| 16 | G | G | G | G | G |
| 17 | G | G | G | G | G |
| 18 | G | G | NG | G | G |
| 19 | G | G | G | G | G |
| 20 | G | G | G | G | G |

# X-Ray Measurement System Evaluation

- Do associates agree with themselves?
  - (Individual Effectiveness)
- Do associates agree with each other?
  - (Group Effectiveness)
- Do associates agree with the Standard?
  - (Department Effectiveness)

# X-Ray Example

**Individual Effectiveness:**

**Associate A:**
18/20 = .90
90%

**Associate B:**

?

| | Associate A | | Associate B | | |
|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | Standard |
| 1 | G | G | G | G | G |
| 2 | G | G | G | G | G |
| 3 | ~~NG~~ | ~~G~~ | G | G | G |
| 4 | NG | NG | NG | NG | NG |
| 5 | G | G | G | G | G |
| 6 | G | G | NG | G | G |
| 7 | NG | NG | G | NG | NG |
| 8 | NG | NG | G | G | NG |
| 9 | G | G | G | G | G |
| 10 | G | G | G | NG | G |
| 11 | G | G | G | G | G |
| 12 | G | G | G | G | G |
| 13 | ~~G~~ | ~~NG~~ | G | G | G |
| 14 | G | G | G | G | G |
| 15 | G | G | G | G | NG |
| 16 | G | G | G | G | G |
| 17 | G | G | G | G | G |
| 18 | G | G | NG | G | G |
| 19 | G | G | G | G | G |
| 20 | G | G | G | G | G |

# X-Ray Example

**Individual Effectiveness:**

**Associate A:**
18/20 = .90
90%

**Associate B:**
16/20 = .80
80%

| | Associate A | | Associate B | |
|---|---|---|---|---|
| | 1 | 2 | 1 | 2 |
| 1 | G | G | G | G |
| 2 | G | G | G | G |
| 3 | ~~NG~~ | ~~G~~ | G | G |
| 4 | NG | NG | NG | NG |
| 5 | G | G | G | G |
| 6 | G | G | ~~NG~~ | ~~G~~ |
| 7 | NG | NG | ~~G~~ | ~~NG~~ |
| 8 | NG | NG | G | G |
| 9 | G | G | G | G |
| 10 | G | G | ~~G~~ | ~~NG~~ |
| 11 | G | G | G | G |
| 12 | G | G | G | G |
| 13 | ~~G~~ | ~~NG~~ | G | G |
| 14 | G | G | G | G |
| 15 | G | G | G | G |
| 16 | G | G | G | G |
| 17 | G | G | G | G |
| 18 | G | G | ~~NG~~ | ~~G~~ |
| 19 | G | G | G | G |
| 20 | G | G | G | G |

# X-Ray Example

**Group Effectiveness:**

| | Associate A | | Associate B | |
|---|---|---|---|---|
| | 1 | 2 | 1 | 2 |
| 1 | G | G | G | G |
| 2 | G | G | G | G |
| 3 | NG | G | G | G |
| 4 | NG | NG | NG | NG |
| 5 | G | G | G | G |
| 6 | G | G | NG | G |
| 7 | NG | NG | G | NG |
| 8 | NG | NG | G | G |
| 9 | G | G | G | G |
| 10 | G | G | G | NG |
| 11 | G | G | G | G |
| 12 | G | G | G | G |
| 13 | G | NG | G | G |
| 14 | G | G | G | G |
| 15 | G | G | G | G |
| 16 | G | G | G | G |
| 17 | G | G | G | G |
| 18 | G | G | NG | G |
| 19 | G | G | G | G |
| 20 | G | G | G | G |

# X-Ray Example

**Group Effectiveness:**

13/20 = .65

65%

| | Associate A | | Associate B | |
|---|---|---|---|---|
| | 1 | 2 | 1 | 2 |
| 1 | G | G | G | G |
| 2 | G | G | G | G |
| 3 | NG | G | G | G |
| 4 | NG | NG | NG | NG |
| 5 | G | G | G | G |
| 6 | G | G | NG | G |
| 7 | NG | NG | G | NG |
| 8 | NG | NG | G | G |
| 9 | G | G | G | G |
| 10 | G | G | G | NG |
| 11 | G | G | G | G |
| 12 | G | G | G | G |
| 13 | G | NG | G | G |
| 14 | G | G | G | G |
| 15 | G | G | G | G |
| 16 | G | G | G | G |
| 17 | G | G | G | G |
| 18 | G | G | NG | G |
| 19 | G | G | G | G |
| 20 | G | G | G | G |

# X-Ray Example

**Departmental Effectiveness:**

***Compare every observation with the standard,***

$$\frac{\text{\# correct}}{\text{Total Obs.}}$$

| | Associate A | | Associate B | | |
|---|---|---|---|---|---|
| | 1 | 2 | 1 | 2 | Standard |
| 1 | G | G | G | G | G |
| 2 | G | G | G | G | G |
| 3 | NG | G | G | G | G |
| 4 | NG | NG | NG | NG | NG |
| 5 | G | G | G | G | G |
| 6 | G | G | NG | G | G |
| 7 | NG | NG | G | NG | NG |
| 8 | NG | NG | G | G | NG |
| 9 | G | G | G | G | G |
| 10 | G | G | G | NG | G |
| 11 | G | G | G | G | G |
| 12 | G | G | G | G | G |
| 13 | G | NG | G | G | G |
| 14 | G | G | G | G | G |
| 15 | G | G | G | G | NG |
| 16 | G | G | G | G | G |
| 17 | G | G | G | G | G |
| 18 | G | G | NG | G | G |
| 19 | G | G | G | G | G |
| 20 | G | G | G | G | G |

# X-Ray Example

**Departmental Effectiveness:**

$$\frac{20-8}{20} = \frac{12}{20}$$

= .60

60%

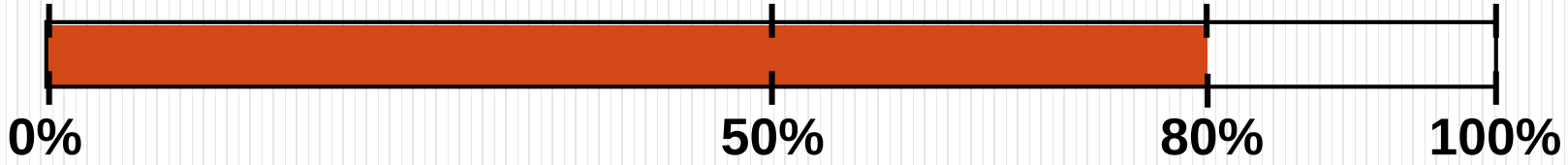
| | 1 | 2 | 1 | 2 | Standard |
|---|---|---|---|---|---|
| 1 | G | G | G | G | G |
| 2 | G | G | G | G | G |
| 3 | ~~NG~~ | ~~G~~ | ~~G~~ | ~~G~~ | ~~G~~ |
| 4 | NG | NG | NG | NG | NG |
| 5 | G | G | G | G | G |
| 6 | ~~G~~ | ~~G~~ | ~~NG~~ | ~~G~~ | ~~G~~ |
| 7 | ~~NG~~ | ~~NG~~ | ~~G~~ | ~~NG~~ | ~~NG~~ |
| 8 | ~~NG~~ | ~~NG~~ | ~~G~~ | ~~G~~ | ~~NG~~ |
| 9 | G | G | G | G | G |
| 10 | ~~G~~ | ~~G~~ | ~~G~~ | ~~NG~~ | ~~G~~ |
| 11 | G | G | G | G | G |
| 12 | G | G | G | G | G |
| 13 | ~~G~~ | ~~NG~~ | ~~G~~ | ~~G~~ | ~~G~~ |
| 14 | G | G | G | G | G |
| 15 | ~~G~~ | ~~G~~ | ~~G~~ | ~~G~~ | ~~NG~~ |
| 16 | G | G | G | G | G |
| 17 | G | G | G | G | G |
| 18 | ~~G~~ | ~~G~~ | ~~NG~~ | ~~G~~ | ~~G~~ |
| 19 | G | G | G | G | G |
| 20 | G | G | G | G | G |

# Another Statistical Approach to Measuring Agreement

- Kappa is a measure of agreement that has several desirable characteristics, as well as a few undesirable ones.
- It is a correlation coefficient that is adjusted for expected values and has the following general properties.:
  - If there is perfect agreement, then Kappa = 1
  - If the observed agreement is greater than the expected value (chance agreement), then Kappa is greater than 0—ranging between 0 and 1 depending on the degree of agreement.
  - If the observed agreement is less than the expected value, then Kappa is less than 0, ranging between 0 and -1 depending on the degree of disagreement.

# What is Kappa?

- Kappa normalizes the scale of agreement such that it starts at the expected value for the study that is being done.
- The illustration below shows the relationship between Kappa and % Agreement for a simple two trial or two alternative decision.




Scale of % Agreement

**Decision Table for Kappa**

| | | | |
|---|---|---|---|
| | ≥ | 0.90 | Best-Case Human Capability |
| 0.75 | - | 0.90 | Excellent Performance |
| 0.40 | - | 0.75 | Marginal Performance |
| | < | 0.40 | Poor Performance |

# KAPPA

- Certain data collection conditions need to be met for this technique to be effective:
  - Inspectors make decisions independent of each other
  - All classifications are independent of each other
  - One classification may be used more frequently than another
  - The categories are mutually exclusive and exhaustive
- Kappa (K) is defined as the proportion of agreement between evaluators after agreement by chance has been removed and while also combining the Alpha and Beta risk error into the collected data.

# Attribute Measurement Systems

- Most physical measurement systems use measurement devices that provide continuous data.
  - For continuous data Measurement System Analysis we can use control charts or Gage R&R methods.
- Attribute/ordinal measurement systems utilize accept/reject criteria or ratings (such as 1 - 5) to determine if an acceptable level of quality has been attained.
  - Kappa techniques can be used  to evaluate these Attribute and Ordinal Measurement Systems.

# Are You Really Stuck With Attribute Data?

- Many inspection or checking processes have the ability to collect continuous data, but decide to use attribute data to simplify the task for the person taking and recording the data.
- Examples:
  - On-time Delivery can be recorded in 2 ways:
    - in hours late, or
    - whether the delivery was on-time or late
- Many functional tests will evaluate a product on a continuous scale (temperature, pressure drop, voltage drop, dimensional, hardness, etc.) and record the results as pass/fail.

# Attribute and Ordinal Measurements

- Attribute and Ordinal measurements often rely on subjective classifications or ratings.
  - Examples include:
    - Rating different features of a service as either good or bad, or on a scale from 1 to 5
    - Rating different aspects of employee performance as excellent, satisfactory, needs improvement
- Should we evaluate these measurement systems before using them to make decisions on our Lean Six Sigma project?
- What are the consequences of not evaluating them?

# Scales

- **Nominal:** Contains numbers that have no basis on which to arrange in any order or to make any assumptions about the quantitative difference between them.
  - In an organization: Dept. 1 (Accounting), Dept. 2 (Customer Service), Dept. 3 ( Human Resources)
  - Modes of transport: Mode 1 (air), Mode 2 (truck), Mode 3 (sea)
- **Ordinal:** Contains numbers that can be ranked in some natural sequence but cannot make an inference about the degree of difference between the numbers.
  - On service performance: excellent, very good, good, fair, poor
  - Customer survey: strongly agree, agree, disagree, strongly disagree

# Kappa Techniques

- Kappa for Attribute Data:
  - Treats all misclassifications equally
  - Does not assume that the ratings are equally distributed across the possible range
  - Requires that the units be independent and that the persons doing the judging or rating make their classifications independently
  - Requires that the assessment categories be mutually exclusive

# Operational Definitions

- There are some quality characteristics that are either difficult or very time consuming to define.
- To assess classification consistency, several units must be classified by more than one rater or judge.
- If there is substantial agreement among the raters, there is the possibility, although no guarantee, that the ratings are accurate.
- If there is poor agreement among the raters, the usefulness of the rating is very limited.

# Consequences?

- What are the important concerns?
  - What are the risks if agreement within and between raters is not good?
  - Are bad items escaping to the next operation in the process or to the external customer?
  - Are good items being reprocessed unnecessarily?
  - What is the standard for assessment?
  - How is agreement measured?
  - What is the Operational Definition for assessment?

# What Is Kappa?

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

- P observed
  - Proportion of units on which both Judges agree = proportion both Judges agree are good + proportion both Judges agree are bad.
- P chance
  - Proportion of agreements expected by chance = (proportion Judge A says good * proportion Judge B says good) + (proportion Judge A says bad * proportion B says bad)

Note: equation applies to a two category analysis, e.g., good or bad.

# Kappa

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

- For perfect agreement, P observed = 1 and K = 1
  - As a rule of thumb, if Kappa is lower than .7, the measurement system is not adequate.
  - If Kappa is .9 or above, the measurement system is considered excellent.
- The lower limit for Kappa can range from 0 to -1
  - For P observed = P chance, then K = 0.
  - Therefore, a Kappa of 0 indicates that the agreement is the same as would be expected by random chance.

# Attribute Measurement System Guidelines

- When selecting items for the study consider the following:
  - If you only have two categories, good and bad, you should have a minimum of 20 good and 20 bad
  - As a maximum, have 50 good and 50 bad.
  - Try to keep approximately 50% good and 50% bad.
  - Have a variety of degrees of good and bad.

# Attribute Measurement System Guidelines

- If you have more than two categories, with one of the categories being good and the other categories being different error modes, you should have approximately 50% of the items being good and a minimum of 10% of the items in each of the error modes.
- You might combine some of the error modes as “other”.
- The categories should be mutually exclusive or, if not, they should also be combined.

# Within Rater/Repeatability Considerations

- Have each rater evaluate the same item at least twice.
- Calculate a Kappa for each rater by creating separate Kappa tables, one for each rater.
  - If a Kappa measurement for a particular rater is small, that rater does not repeat well within self.
  - If the rater does not repeat well within self, then he won't repeat well with the other raters and this will hide how good or bad the others repeat between themselves.
- Calculate a between-rater Kappa by creating a Kappa table from the first judgment of each rater.
- Between-rater Kappa will be made as pairwise comparisons (A to B, B to C, A to C).

# Kappa Example #1

- Bill Blackbelt is trying to improve an Auto Body Paint and Repair branch that has a high rejection rate for its paint repairs.
- Early on in the project, the measurement system becomes a concern due to obvious inspector to inspector differences as well as within inspector differences.
- The data on the following slide were gathered during a measurement system study.
- Kappa for each inspector as well as Kappa between inspectors need to be calculated.

# Consider the Following Data

| | First Mea. | Second Mea. | First Mea. | Second Mea. | First Mea. | Second Mea. |
|---|---|---|---|---|---|---|
| Item | Rater A | Rater A | Rater B | Rater B | Rater C | Rater C |
| 1 | Good | Good | Good | Good | Good | Good |
| 2 | Bad | Bad | Good | Bad | Bad | Bad |
| 3 | Good | Good | Good | Good | Good | Good |
| 4 | Good | Bad | Good | Good | Good | Good |
| 5 | Bad | Bad | Bad | Bad | Bad | Bad |
| 6 | Good | Good | Good | Good | Good | Good |
| 7 | Bad | Bad | Bad | Bad | Bad | Bad |
| 8 | Good | Good | Bad | Good | Good | Bad |
| 9 | Good | Good | Good | Good | Good | Good |
| 10 | Bad | Bad | Bad | Bad | Bad | Bad |
| 11 | Good | Good | Good | Good | Good | Good |
| 12 | Good | Good | Good | Bad | Good | Good |
| 13 | Bad | Bad | Bad | Bad | Bad | Bad |
| 14 | Good | Good | Bad | Good | Good | Good |
| 15 | Good | Good | Good | Good | Good | Good |
| 16 | Bad | Good | Good | Good | Good | Good |
| 17 | Bad | Bad | Bad | Good | Bad | Good |
| 18 | Good | Good | Good | Good | Good | Good |
| 19 | Bad | Bad | Bad | Bad | Bad | Bad |

# Contingency Table for Rater A

**Populate Each Cell with the Information Collected**

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 10 | 2 | 12 |
| | Bad | 1 | 7 | 8 |
| | | 11 | 9 | |

# Contingency Table

**The first cell represents the number of times Rater A judged an item 'Good' in both the first and second evaluation**

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 10 | 2 | 12 |
| | Bad | 1 | 7 | 8 |
| | | 11 | 9 | |

# Contingency Table

**The second cell represents the number of times Rater A judged an item 'Bad' the first time and 'Good' the second time**

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 10 | 2 | 12 |
| | Bad | 1 | 7 | 8 |
| | | 11 | 9 | |

# Contingency Table

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 10 | 2 | 12 |
| | Bad | 1 | 7 | 8 |
| | | 11 | 9 | |

**The third cell represents the number of times Rater A judged an item 'Good' the first time, and 'Bad' the second time**

# Contingency Table

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 10 | 2 | 12 |
| | Bad | 1 | 7 | 8 |
| | | 11 | 9 | |

**The fourth cell represents the number of times Rater A judged an item 'Bad' the first time, and 'Bad' the second time**

# Contingency Table

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 10 | 2 | 12 |
| | Bad | 1 | 7 | 8 |
| | | 11 | 9 | |

**The numbers on the margins represent the totals of the rows and columns**

# Contingency Table – Proportions

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 10 | 2 | 12 |
| | Bad | 1 | 7 | 8 |
| | | 11 | 9 | |

**The lower table represents the data in the top with each cell being represented as a percent of total**

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 0.5 | 0.1 | 0.6 |
| | Bad | 0.05 | 0.35 | 0.4 |
| | | 0.55 | 0.45 | |

**Represents 10/20**

Rater A Proportion

# Contingency Table – Proportions

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 0.5 | 0.1 | 0.6 |
| | Bad | 0.05 | 0.35 | 0.4 |
| | | 0.55 | 0.45 | |

**Calculated from the sum of the rows and columns**

# Remember How to Calculate Kappa?

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

- $P_{observed}$
  - Proportion of items on which both Judges agree = proportion both Judges agree are 'Good' + proportion both Judges agree are 'Bad'
- $P_{chance}$
  - Proportion of agreements expected by chance = (proportion Judge A says 'Good' * proportion Judge B says 'Good') + (proportion Judge A says 'Bad' * proportion B says 'Bad')

# Calculate Kappa for Rater A

| | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater A Second Measure | Good | 0.5 | 0.1 | 0.6 |
| | Bad | 0.05 | 0.35 | 0.4 |
| | | 0.55 | 0.45 | |

**$P_{observed}$ is the sum of the probabilities on the diagonal:**

$$P_{observed} = (0.500 + 0.350) = 0.850$$

**$P_{chance}$ is the probabilities for each classification multiplied and then summed:**

$$P_{chance} = (0.600*0.55) + (0.400*0..45) = 0.51$$

**Then $K_{Rater\ A} = (0.85 - 0.51)/(1 - 0.51) = 0.693$**

# Calculate Kappa for Rater B

| | | Rater B First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater B Second Measure | Good | | | |
| | Bad | | | |
| | | | | |

**Number**

$$K_{\text{Rater B}} =$$

| | | Rater B First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater B Second Measure | Good | | | |
| | Bad | | | |
| | | | | |

**Proportion**

# Kappa Between Raters

- To estimate a Kappa for between Raters, we will use the same procedure.
- We will limit ourselves to the first judging of the pair of Raters we are interested in calculating Kappa for.
- If there is a Rater who has poor Within-Rater repeatability (less than 85%), there is no use in calculating a Between-Rater rating for him/her.

# Kappa – Rater A to Rater B

| Rater A to Rater B | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater B First Measure | Good | 9 | 3 | 12 |
| | Bad | 2 | 6 | 8 |
| | | 11 | 9 | |

**Number of times both Raters agreed the item was 'Good' (using their first measurement)**

# Kappa Between Raters

| Rater A to Rater B | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater B First Measure | Good | 9 | 3 | 12 |
| | Bad | 2 | 6 | 8 |
| | | 11 | 9 | |

**Number of times Rater A judged an item 'Bad' and Rater B judged an item 'Good' (using their first measurement)**

# Rater A to Rater B Kappa

| Rater A to Rater B | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater B First Measure | Good | 9 | 3 | 12 |
| | Bad | 2 | 6 | 8 |
| | | 11 | 9 | |

**Number of times Rater A judged an item 'Good' and Rater B judged an item 'Bad' (using their first measurement)**

# Between Rater Kappa

| Rater A to Rater B | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater B First Measure | Good | 9 | 3 | 12 |
| | Bad | 2 | 6 | 8 |
| | | 11 | 9 | |

**Number of times both Raters agreed the item was 'Bad' (using their first measurement)**

# Kappa Between Raters – The Numbers

| Rater A to Rater B | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater B First Measure | Good | 9 | 3 | 12 |
| | Bad | 2 | 6 | 8 |
| | | 11 | 9 | |

| Rater A to Rater B | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater B First Measure | Good | 0.45 | 0.15 | 0.6 |
| | Bad | 0.1 | 0.3 | 0.4 |
| | | 0.55 | 0.45 | |

**The lower table represents the data in the top with each cell being represented as a percent of the total**

# Remember How to Calculate Kappa?

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

- $P_{observed}$
  - Proportion of items on which both Judges agree = proportion both Judges agree are 'Good' + proportion both Judges agree are 'Bad'
- $P_{chance}$
  - Proportion of agreements expected by chance = (proportion Judge A says 'Good' * proportion Judge B says 'Good') + (proportion Judge A says 'Bad' * proportion Judge B says 'Bad')

# Calculate Kappa for Rater A to Rater B

| Rater A to Rater B | | Rater A First Measure | | |
|---|---|---|---|---|
| | | Good | Bad | |
| Rater B First Measure | Good | 0.45 | 0.15 | 0.6 |
| | Bad | 0.1 | 0.3 | 0.4 |
| | | 0.55 | 0.45 | |

**$P_{observed}$ is the sum of the probabilities on the diagonal:**

**$P_{observed}$ = (0.450 + 0.300) = 0.750**

**$P_{chance}$ is the probability for each classification multiplied and then summed:**

**$P_{chance}$ = (0.600*0.55) + (0.400*0..45) = 0.51**

**Then $K_{Rater\ A/B}$ = (0.75 - 0.51)/(1 - 0.51) = 0.489**

# Improvement Ideas

- How might we improve this measurement system?
  - Additional Training
  - Physical Standards/Samples
  - Rater Certification (and periodic Re-certification) Process
  - Better Operational Definitions

# Kappa Conclusions

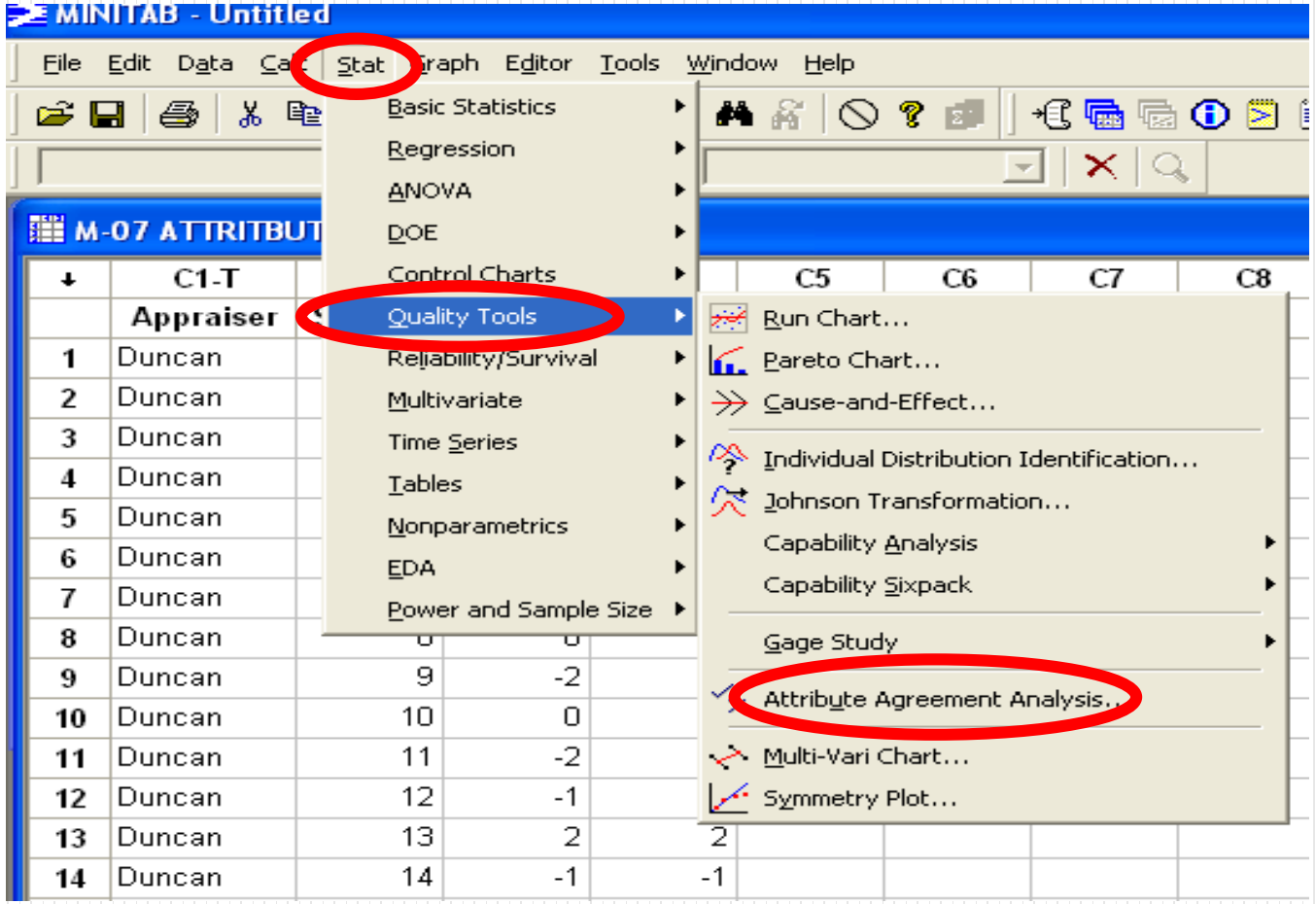
- Is the current measurement system adequate?
- Where would you focus your improvement efforts?
- What rater would you want to conduct any training that needs to be done?

# Minitab Example

- An educational testing organization is training five new appraisers for the written portion of the twelfth-grade standardized essay test.
- The appraisers' ability to rate essays consistent with the standards needs to be assessed.
- Each appraiser rated fifteen essays on a five-point scale (-2, -1, 0, 1, 2).
- The organization also rated the essays and supplied the "official score."
- Each essay was rated twice and the data captured in the file Minitab file "C:/Program Files (X86)/minitab/minitab17/English/Sample Data/Essay.mtw"
- Open the file and evaluate the appraisers performance.

# Minitab Example

**Stat > Quality Tools >Attribute Agreement Analysis**

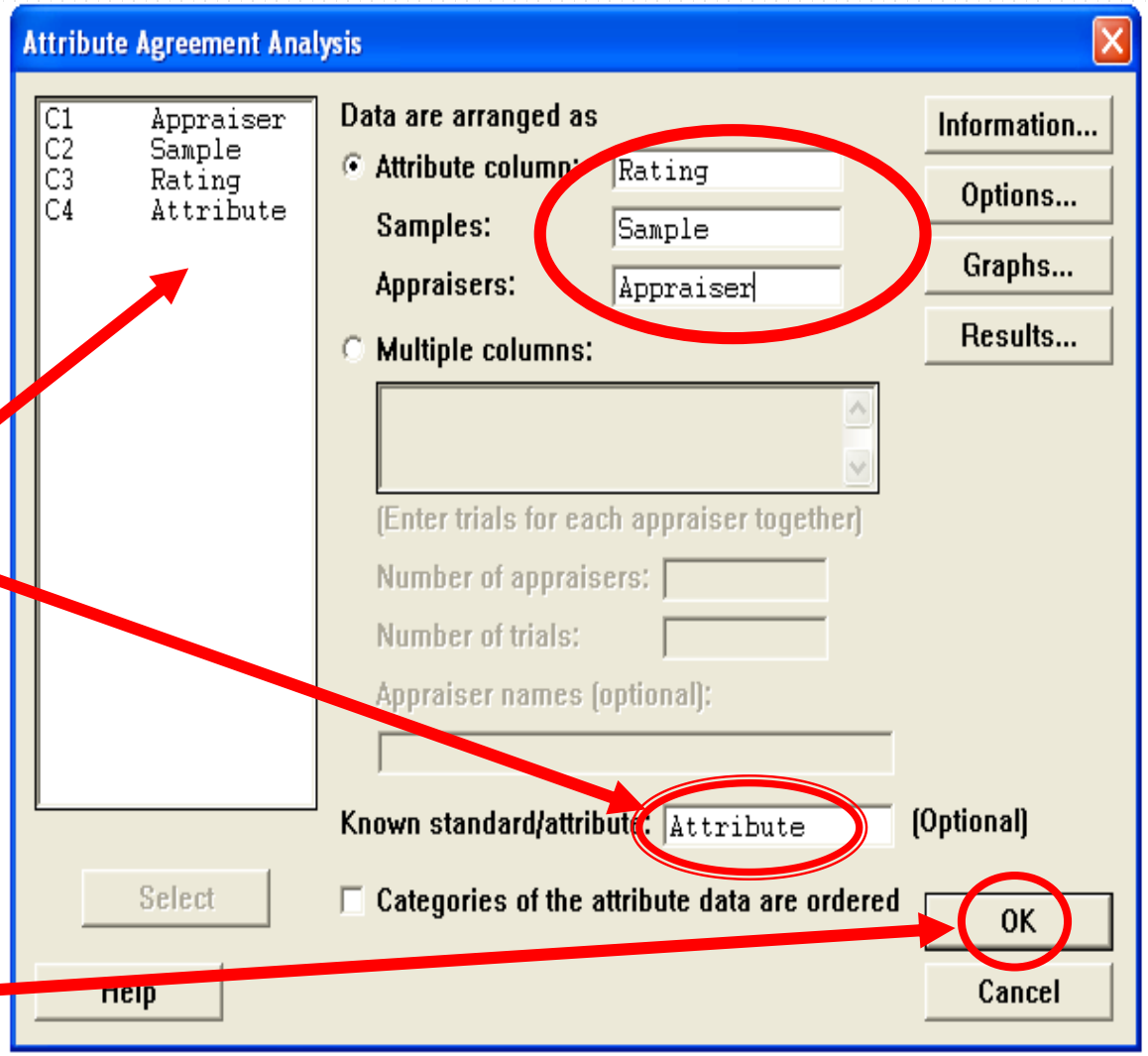


# Minitab Example

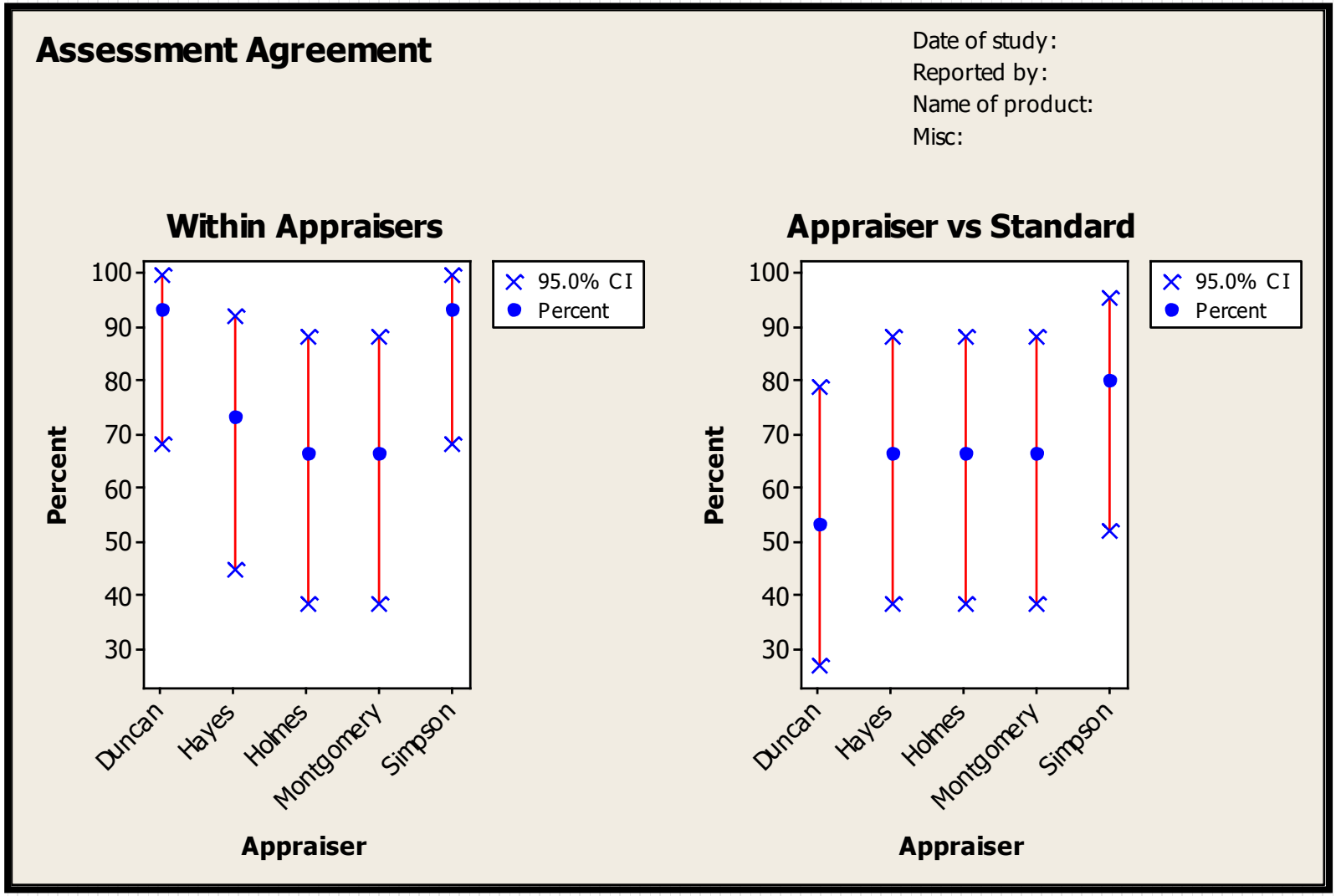
1. Double click on the appropriate variable to place it in the required dialog box. (same as before)

2. If you have a known standard (the real answer) for the items being inspected, let *Minitab* know what column that information is in.

3. Click on OK.



# Appraiser vs. Standard



# Within Appraiser

**Within Appraisers**

Assessment Agreement

| Appraiser | # Inspected | # Matched | Percent | 95 % CI |
|---|---|---|---|---|
| Duncan | 15 | 14 | 93.33 | (68.05, 99.83) |
| Hayes | 15 | 11 | 73.33 | (44.90, 92.21) |
| Holmes | 15 | 10 | 66.67 | (38.38, 88.18) |
| Montgomery | 15 | 10 | 66.67 | (38.38, 88.18) |
| Simpson | 15 | 14 | 93.33 | (68.05, 99.83) |

# Matched: Appraiser agrees with him/herself across trials.

**In addition to the Within-Appraiser graphic, Minitab will give percentages**

# Each Appraiser vs. Standard

**Each Appraiser vs Standard**

Assessment Agreement

| Appraiser | # Inspected | # Matched | Percent | 95 % CI |
|---|---|---|---|---|
| Duncan | 15 | 8 | 53.33 | (26.59, 78.73) |
| Hayes | 15 | 10 | 66.67 | (38.38, 88.18) |
| Holmes | 15 | 10 | 66.67 | (38.38, 88.18) |
| Montgomery | 15 | 10 | 66.67 | (38.38, 88.18) |
| Simpson | 15 | 12 | 80.00 | (51.91, 95.67) |

# Matched: Appraiser's assessment across trials agrees with the known standard.

**Some appraisers will repeat their own ratings well but may not match the standard well (look at Duncan)**

# More Session Window Output

**Between Appraisers**

Assessment Agreement

| # Inspected | # Matched | Percent | 95 % CI |
|---|---|---|---|
| 15 | 3 | 20.00 | (4.33, 48.09) |

# Matched: All appraisers' assessments agree with each other.

**All Appraisers vs Standard**

Assessment Agreement

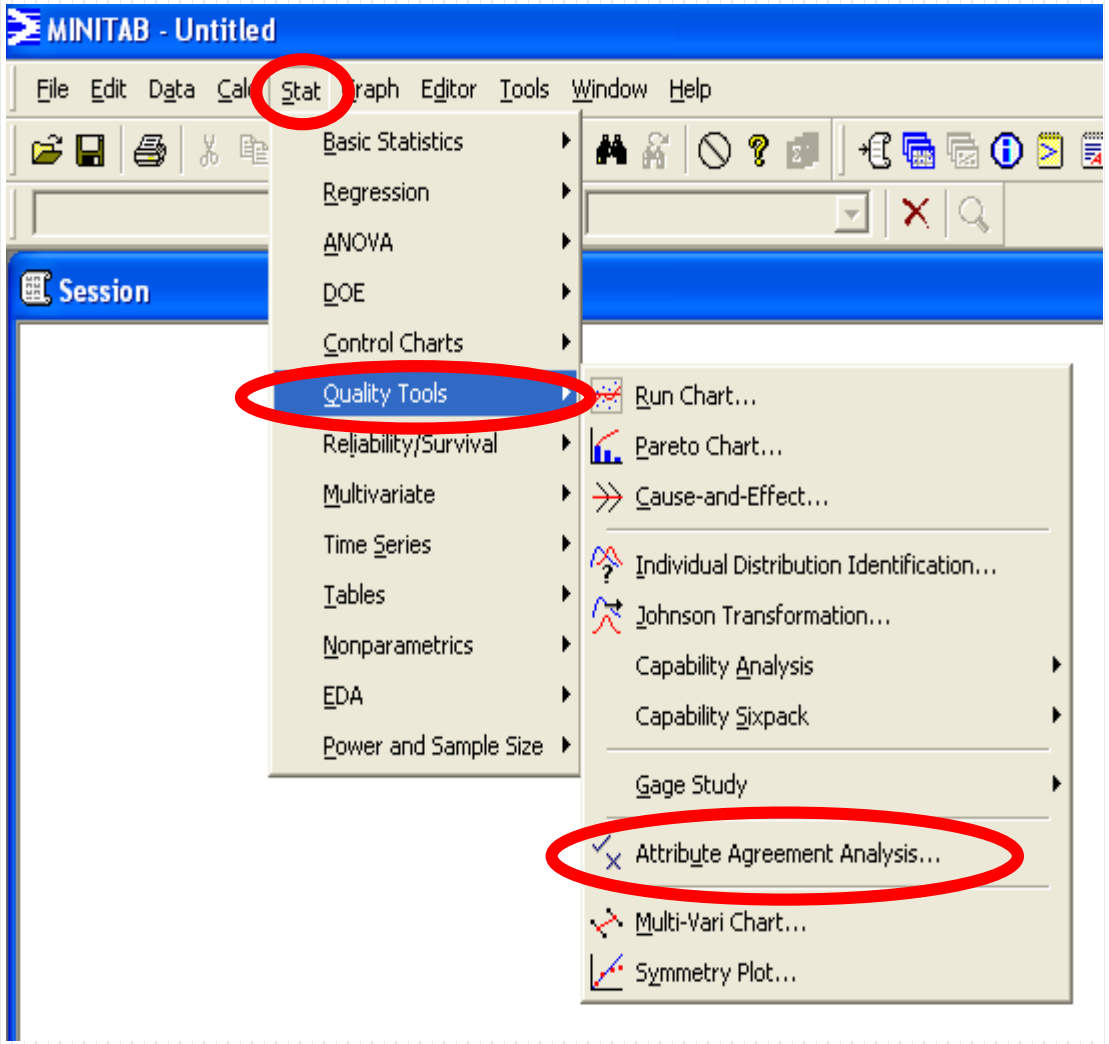
| # Inspected | # Matched | Percent | 95 % CI |
|---|---|---|---|
| 15 | 3 | 20.00 | (4.33, 48.09) |

# Matched: All appraisers' assessments agree with the known standard.

**The session window will give percentage data as to how all the appraisers did when judged against the standard**

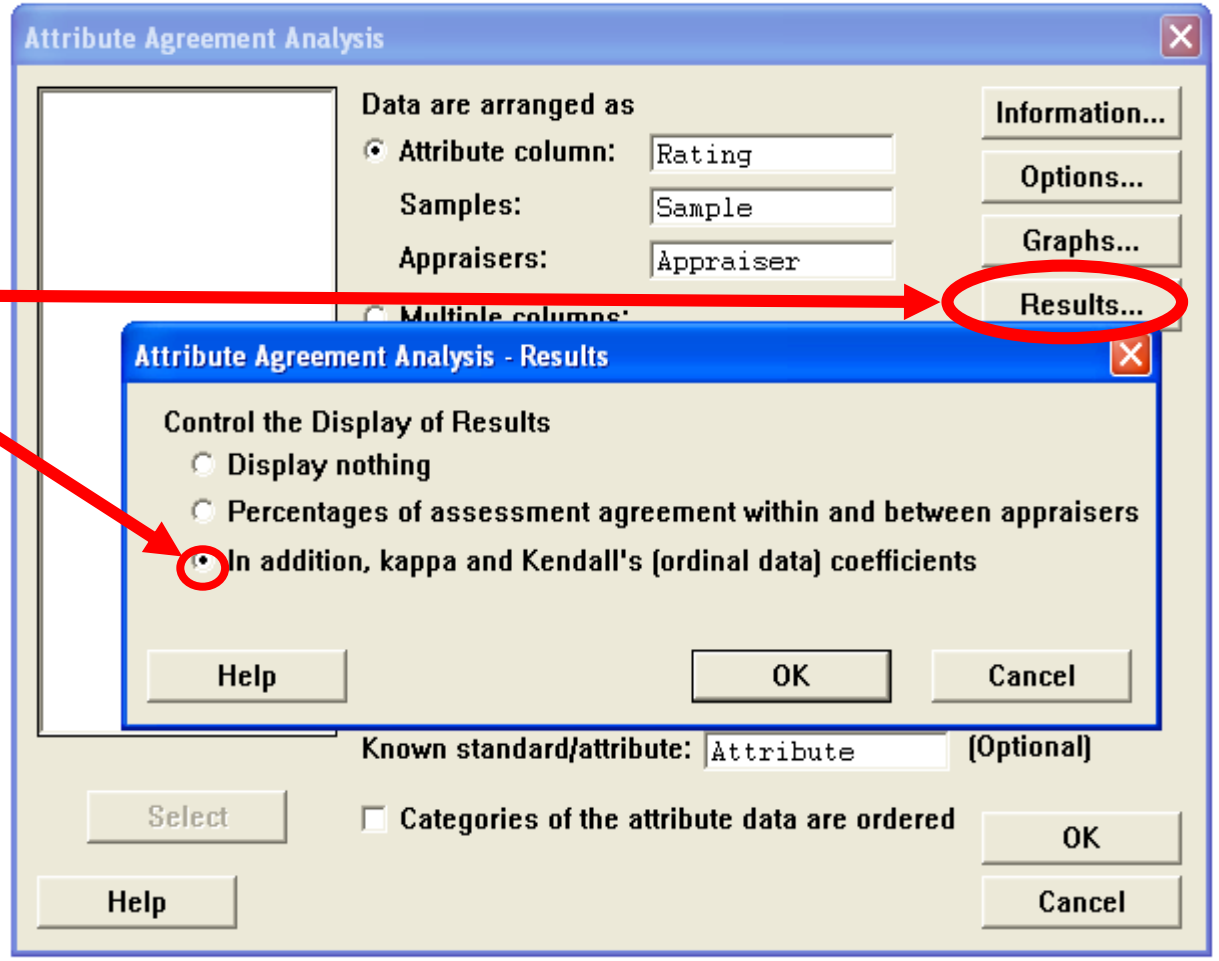
# How Do We Get Kappa from Minitab?

**Minitab can calculate Kappa for *Categorical Data* (pass/fail) as well as for *Ordinal Data.***



# How Do We Get Minitab to Report Kappa?

**Click on *Results* and ask for the additional output**




**This will add Kappa statistics to the session window output**

# Kappa and Minitab

**Minitab will calculate a Kappa for each (within) appraiser for each category**

**Within Appraisers**

Fleiss' Kappa Statistics

| Appraiser | Response | Kappa | SE Kappa | Z | P(vs > 0) |
|---|---|---|---|---|---|
| Duncan | -2 | 1.00000 | 0.258199 | 3.87298 | 0.0001 |
| | -1 | 1.00000 | 0.258199 | 3.87298 | 0.0001 |
| | 0 | 0.76000 | 0.258199 | 2.94347 | 0.0016 |
| | 1 | 0.84127 | 0.258199 | 3.25822 | 0.0006 |
| | 2 | 1.00000 | 0.258199 | 3.87298 | 0.0001 |
| | Overall | 0.91304 | 0.138858 | 6.57536 | 0.0000 |
| Hayes | -2 | -0.07143 | 0.258199 | -0.27664 | 0.6090 |
| | -1 | 0.65909 | 0.258199 | 2.55265 | 0.0053 |
| | 0 | 0.81366 | 0.258199 | 3.15131 | 0.0008 |
| | 1 | 0.42308 | 0.258199 | 1.63857 | 0.0507 |
| | 2 | 0.84127 | 0.258199 | 3.25822 | 0.0006 |
| | Overall | 0.65015 | 0.140344 | 4.63252 | 0.0000 |
| Holmes | -2 | 1.00000 | 0.258199 | 3.87298 | 0.0001 |
| | -1 | 0.76000 | 0.258199 | 2.94347 | 0.0016 |
| | 0 | 0.16667 | 0.258199 | 0.64550 | 0.2593 |
| | 1 | 0.40000 | 0.258199 | 1.54919 | 0.0607 |
| | 2 | 0.76000 | 0.258199 | 2.94347 | 0.0016 |
| | Overall | 0.57020 | 0.133786 | 4.26205 | 0.0000 |

**Note: This is only a part of the total data set for illustration.**

# Kappa vs. Standard

**Minitab will also calculate a Kappa statistic for each appraiser as compared to the standard.**

**Each Appraiser vs Standard**

Fleiss' Kappa Statistics

| Appraiser | Response | Kappa | SE Kappa | Z | P(vs > 0) |
|---|---|---|---|---|---|
| Duncan | -2 | 0.58333 | 0.182574 | 3.19505 | 0.0007 |
| | -1 | 0.16667 | 0.182574 | 0.91287 | 0.1807 |
| | 0 | 0.51216 | 0.182574 | 2.80524 | 0.0025 |
| | 1 | 0.55004 | 0.182574 | 3.01271 | 0.0013 |
| | 2 | 0.42308 | 0.182574 | 2.31729 | 0.0102 |
| | Overall | 0.45307 | 0.092712 | 4.88689 | 0.0000 |
| Hayes | -2 | 0.62963 | 0.182574 | 3.44862 | 0.0003 |
| | -1 | 0.81366 | 0.182574 | 4.45662 | 0.0000 |
| | 0 | 0.90683 | 0.182574 | 4.96693 | 0.0000 |
| | 1 | 0.52000 | 0.182574 | 2.84816 | 0.0022 |
| | 2 | 0.73638 | 0.182574 | 4.03331 | 0.0000 |
| | Overall | 0.74432 | 0.094868 | 7.84581 | 0.0000 |
| Holmes | -2 | 1.00000 | 0.182574 | 5.47723 | 0.0000 |
| | -1 | 0.88000 | 0.182574 | 4.81996 | 0.0000 |
| | 0 | 0.56063 | 0.182574 | 3.07072 | 0.0011 |

**Note: This is only a part of the total data set for illustration.**

# Kappa and Minitab

**Between Appraisers**

Fleiss' Kappa Statistics

| Response | Kappa | SE Kappa | Z | P(vs > 0) |
|---|---|---|---|---|
| -2 | 0.649317 | 0.0384900 | 16.8698 | 0.0000 |
| -1 | 0.552724 | 0.0384900 | 14.3602 | 0.0000 |
| 0 | 0.461511 | 0.0384900 | 11.9904 | 0.0000 |
| 1 | 0.449449 | 0.0384900 | 11.6770 | 0.0000 |
| 2 | 0.663756 | 0.0384900 | 17.2449 | 0.0000 |
| Overall | 0.543686 | 0.0195342 | 27.8326 | 0.0000 |

**All Appraisers vs Standard**

Fleiss' Kappa Statistics

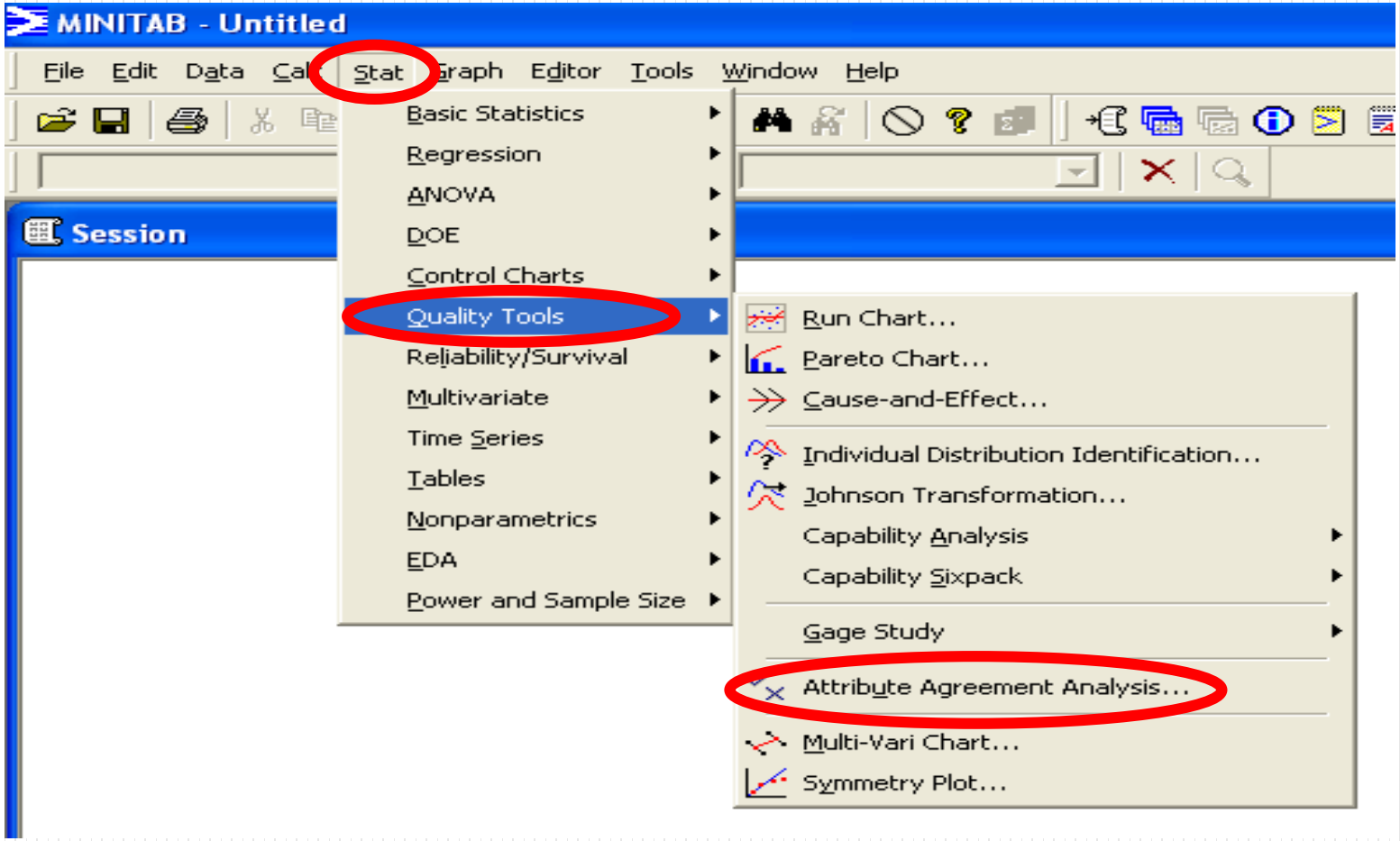
| Response | Kappa | SE Kappa | Z | P(vs > 0) |
|---|---|---|---|---|
| -2 | 0.805556 | 0.0816497 | 9.8660 | 0.0000 |
| -1 | 0.729433 | 0.0816497 | 8.9337 | 0.0000 |
| 0 | 0.653210 | 0.0816497 | 8.0002 | 0.0000 |
| 1 | 0.597464 | 0.0816497 | 7.3174 | 0.0000 |
| 2 | 0.783891 | 0.0816497 | 9.6007 | 0.0000 |
| Overall | 0.712428 | 0.0418820 | 17.0104 | 0.0000 |

**Minitab will not provide a Kappa between a specific pair of appraisers, but will provide an overall Kappa between all appraisers for each possible category of response**

**How might this output help us improve our measurement system?**

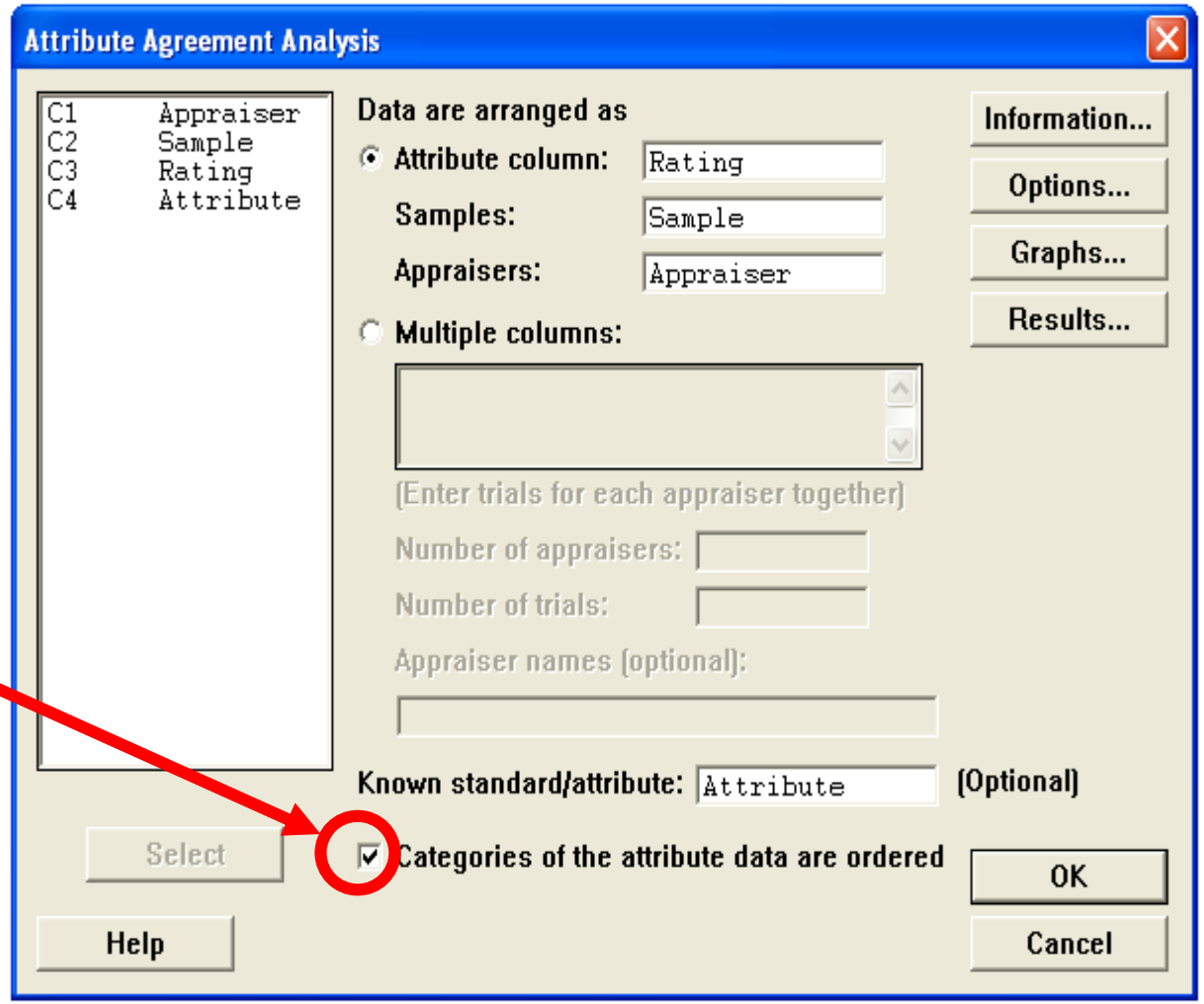
# What If My Data Is Ordinal?

**Stat > Quality Tools > Attribute Agreement Analysis**



# Ordinal Data

**If your data is *Ordinal,* you must also check this box.**



# What Is Kendall's

Within Appraiser

Kendall's Coefficient of Concordance

| Appraiser | Coef | Chi - Sq | DF | P |
|---|---|---|---|---|
| Duncan | 0.9901 | 27.7219 | 14 | 0.015 |
| Hayes | 0.9758 | 27.3226 | 14 | 0.017 |
| Holmes | 0.9540 | 26.7114 | 14 | 0.021 |
| Montgomery | 0.9471 | 26.5194 | 14 | 0.022 |
| Simpson | 0.9902 | 27.7263 | 14 | 0.015 |

**Kendall's coefficient can be thought of as an R-squared value, it is the correlation between the responses treating the data as attribute as compared to ordinal. The lower the number gets, the more severe the misclassifications were**

# Kendall's

Within appraiser versus standard

Kendall's Correlation Coefficient

| Appraiser | Coef | SE Coef | Z | P |
|---|---|---|---|---|
| Duncan | 0.9030 | 0.1361 | 6.6009 | 0.000 |
| Hayes | 0.9227 | 0.1361 | 6.7456 | 0.000 |
| Holmes | 0.9401 | 0.1361 | 6.8730 | 0.000 |
| Montgomery | 0.9288 | 0.1361 | 6.7900 | 0.000 |
| Simpson | 0.8876 | 0.1361 | 6.4878 | 0.000 |

**Kendall's coefficient can be thought of as an R-squared value, it is the correlation between the responses treating the data as attribute as compared to ordinal. The lower the number gets, the more severe the misclassifications were**

# Kendall's

**Between Appraiser**
**Kendall's Coefficient of Concordance**

| Coef | Chi - Sq | DF | P |
|---|---|---|---|
| 0.9203 | 128.8360 | 14 | 0.000 |

**Between appraiser as compared to standard**
**Kendall's Correlation Coefficient**

| Coef | SE Coef | Z | P |
|---|---|---|---|
| 0.9164 | 0.0609 | 15.0431 | 0.000 |

# Summary

In this module you have learned about:

- Measurement Systems Analysis as a tool to validate accuracy, precision and stability
- The importance of good measurements
- The language of measurement
- The types of variation in measurement systems
- Conducting and interpreting a measurement system analysis with normally distributed continuous data
- How to conduct an MSA with Attribute data