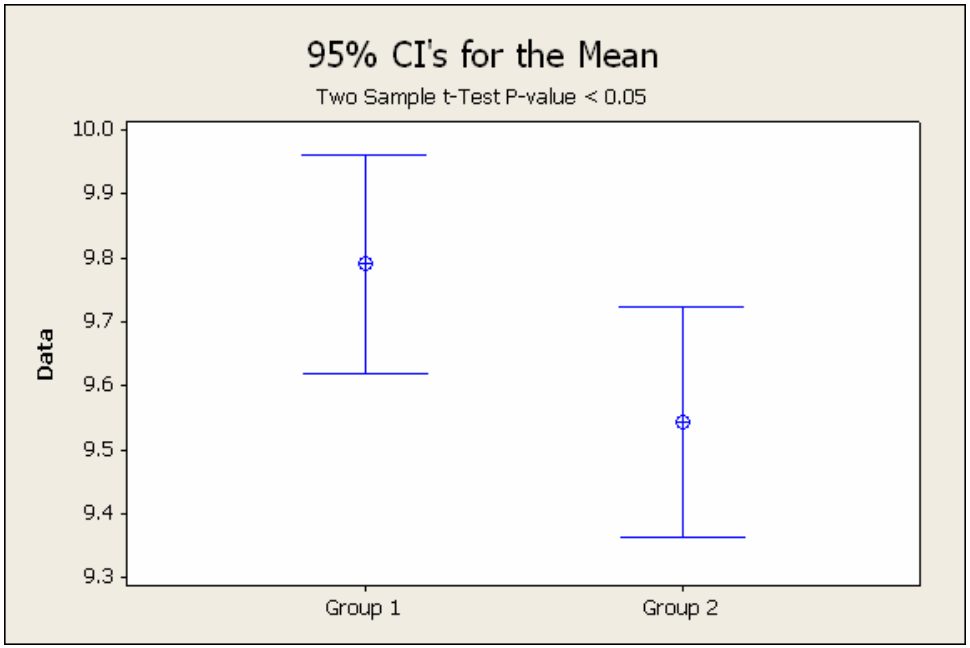
Ten Common and Dangerous Statistical Mistakes

Presenter: **Ross Farrelly**
Minitab Australia
61-2-9312 3763
rfarrelly@minitab.com.au

Minitab

Mistake 1: Misinterpreting Overlapping Confidence Intervals

- Two 95% confidence intervals that overlap may be significantly different at the 95% confidence level.

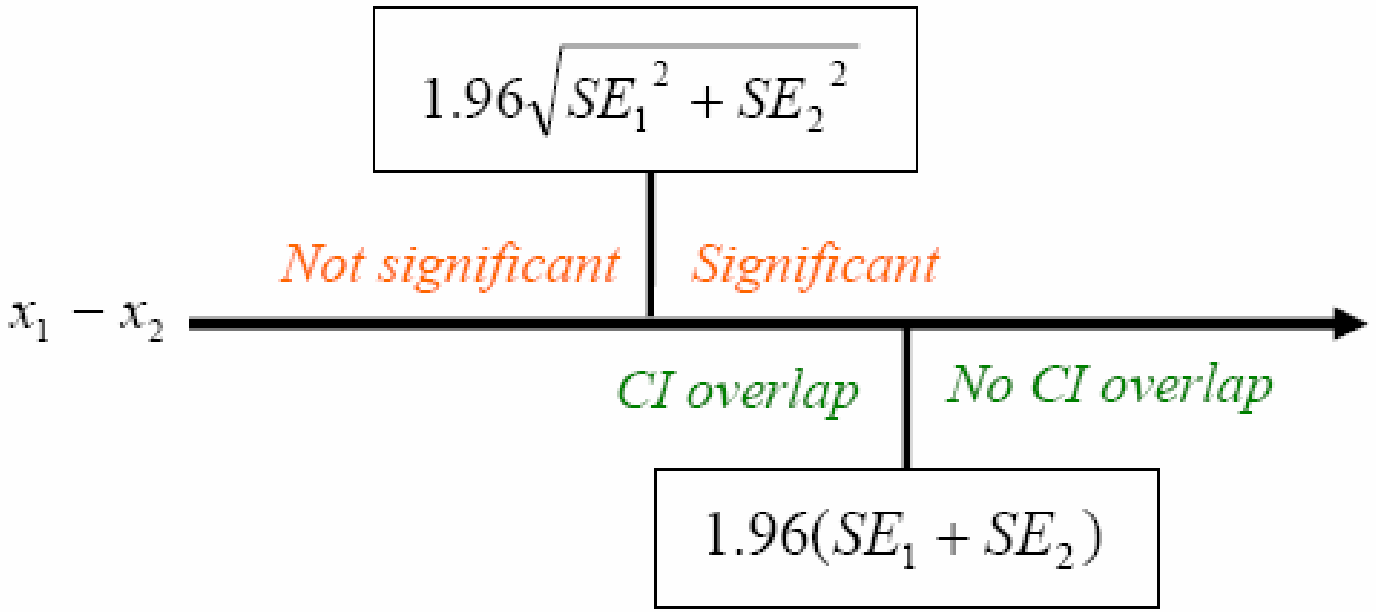


Paper: "A Misconception about Overlapping Confidence Intervals", *James A. Colton, Asia Pacific Engineering, October 2001*.

Mistake 1: Misinterpreting Overlapping Confidence Intervals

The means are significantly different when: $x_1 - x_2 > 1.96\sqrt{SE_1^2 + SE_2^2}$

There is no overlap between CI when: $x_1 - x_2 > 1.96(SE_1 + SE_2)$



Mistake 2: Not Distinguishing Between Statistical Significance and Practical Significance.

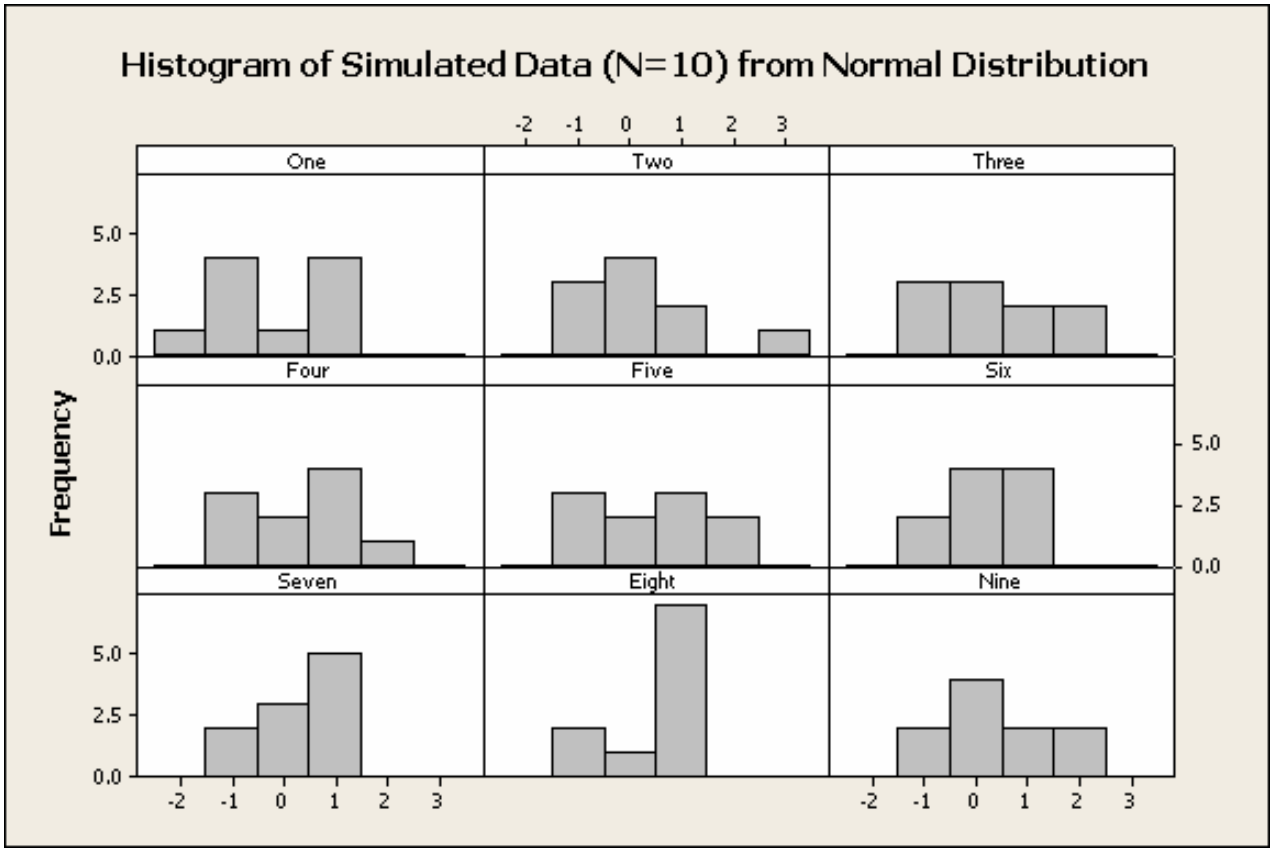
- Cereal Box Example
  - Every cereal box is weighted at the end of a cereal box filling line by an automated measuring system.
  - 18,000 boxes per shift are filled with a standard deviation of 2.5 grams.
  - The target fill weight is 360 grams.
  - A shift in the mean away from the target of 0.06 grams will be detected 90% of the time.

Mistake 2: Not Distinguishing Between Statistical Significance and Practical Significance.

- In most hypothesis tests, we know H_0 is not exactly true. We are just trying to see if there is a meaningful difference.
- Instead of a hypothesis test, use a confidence interval to see how large the difference might be and decide if action is needed.
- Use a hypothesis test when a meaningful difference has been specified in the sample size calculation.
- Mistake 2 is becoming more of an issue with automated data collection and huge databases.

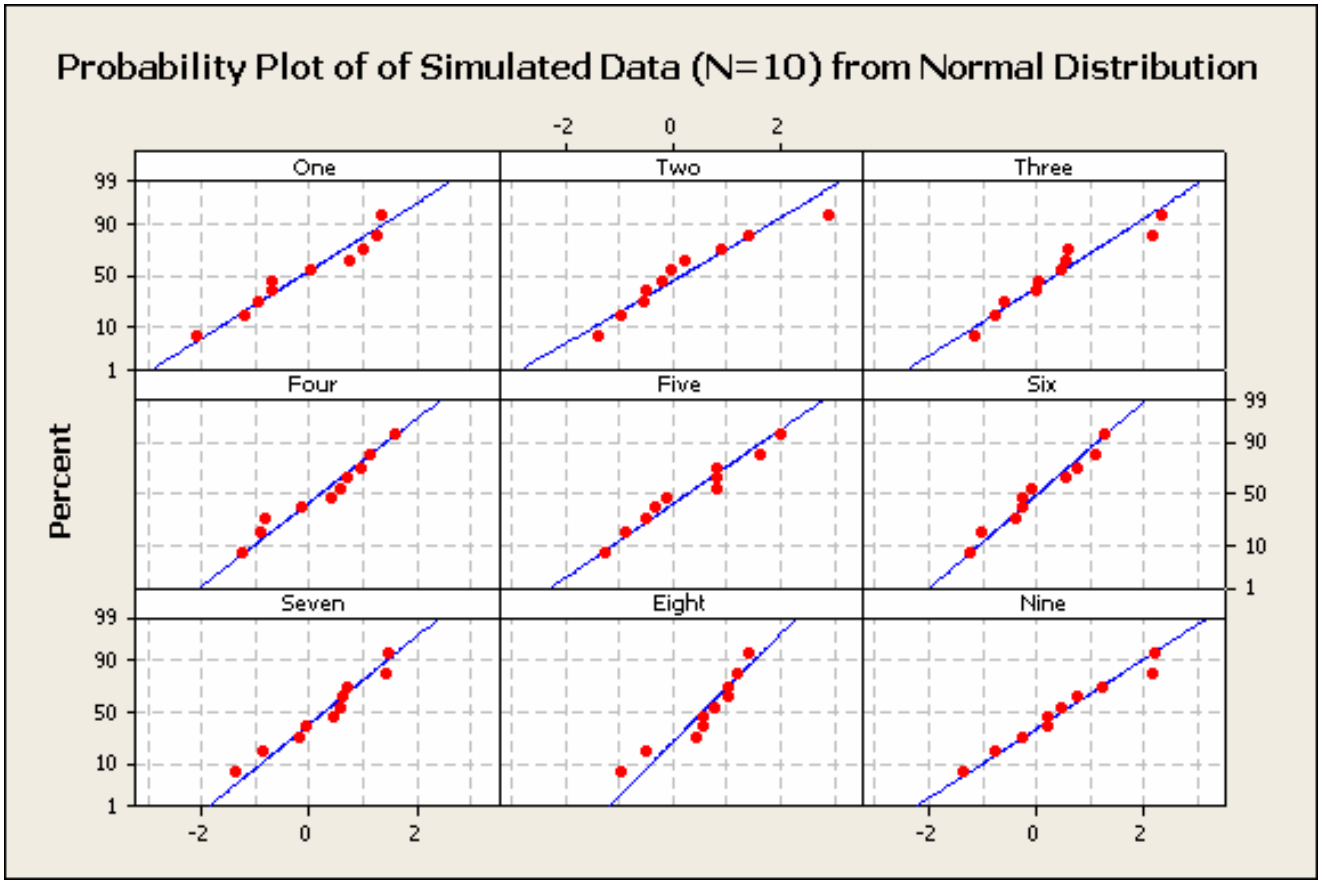
Mistake 3: Rejecting the Normality without Reason

- Avoid the histogram for small sample sizes.



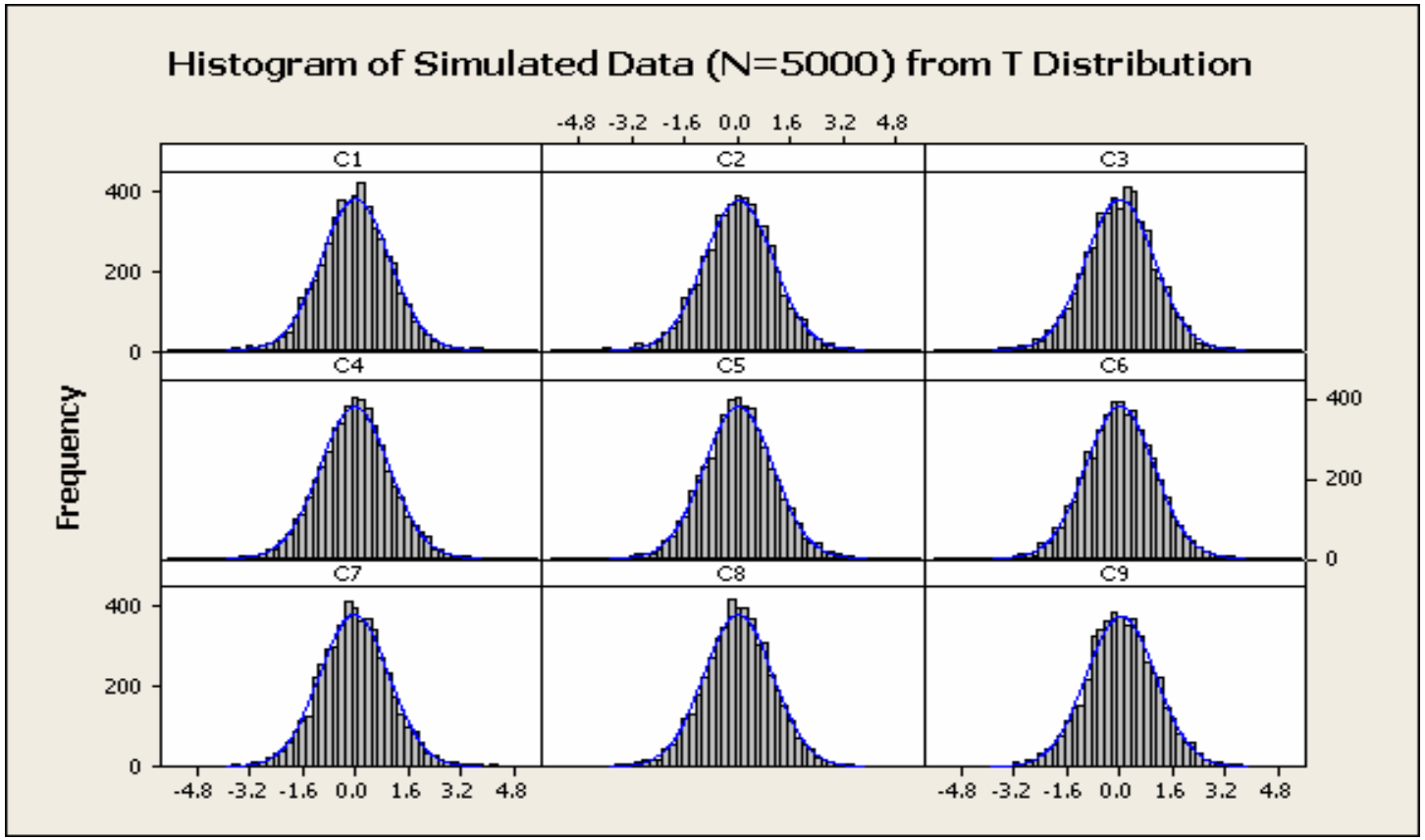
Mistake 3: Rejecting the Normality without Reason

- Use the probability plot instead.



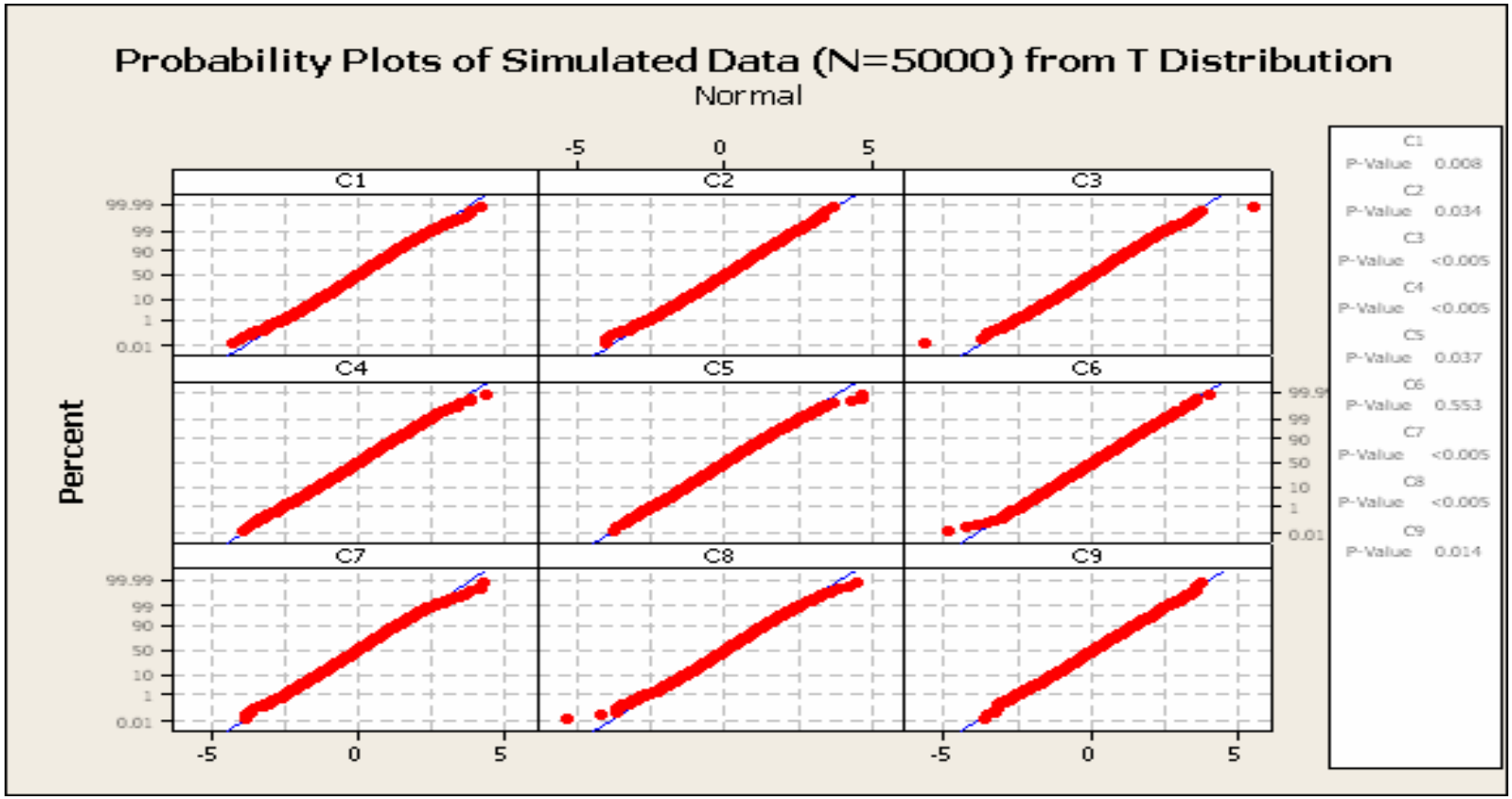
Mistake 3: Rejecting the Normality without Reason

- Be aware of Mistake 2 (detecting a meaningless difference) for large sample sizes.



Mistake 3: Rejecting the Normality without Reason

- Be aware of Mistake 2 (detecting a meaningless difference) for large sample sizes.



Mistake 3: Rejecting the Normality Assumption without Reason

- Avoid the histogram for small sample sizes.
- Be aware of Mistake 2 for large sample sizes.
- For DOE and Regression, don't check for Normality of the actual responses.
  - You expect the responses to be non-Normal because they are affected by the factors and covariates.
  - Check for Normality of the residuals.

Mistake 4: Stating you've proved the null hypothesis (H_0).

- A p-value above 0.05 indicates "there is not enough evidence to conclude H_1 at the 95% confidence level".
- Example: Flip a fair coin 3 times and test

  H_0: Proportion of Heads = 0.40

  H_1: Proportion of Heads ≠ 0.40 (truth)

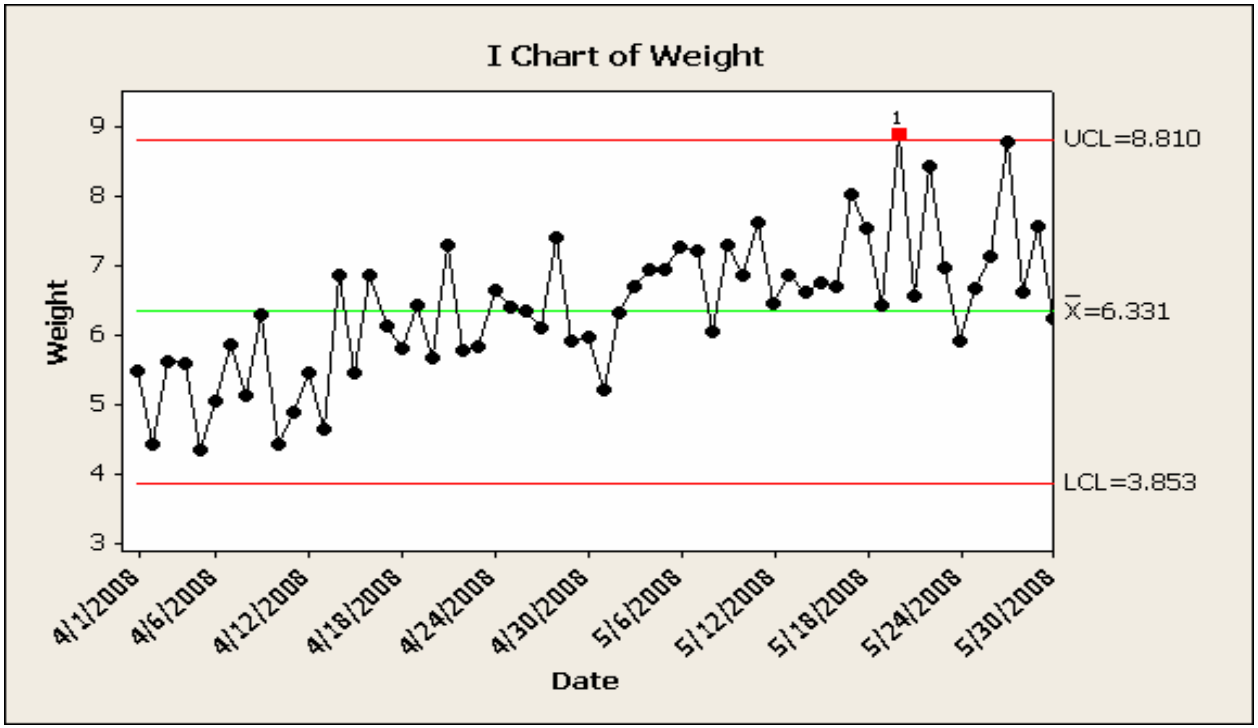
  - The p-value is guaranteed to be above 0.05.
  - You can not support H_1, but that does not mean you have proven H_0 either.

Mistake 5: Assuming Correlation = Causation

- Example 1: Ice cream sales and murder rate are correlated.
  - In the Summer months, both are high. In the Winter months, both are low.
  - One is not causing the other.
- Example 2: In a manufacturing process, the weight of a product changes over time along with other process/environmental variables.

Mistake 5: Assuming Correlation = Causation

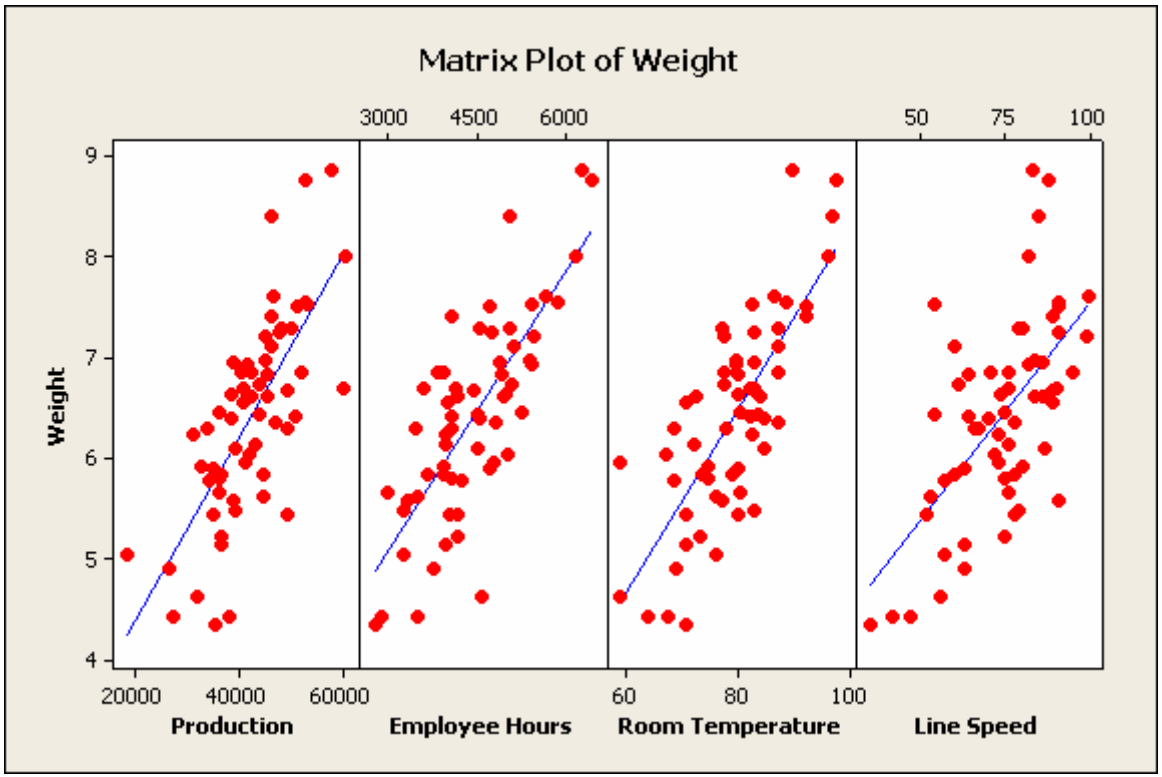
- The weight of a product measured over time:



- During April and May, the process was drifting.

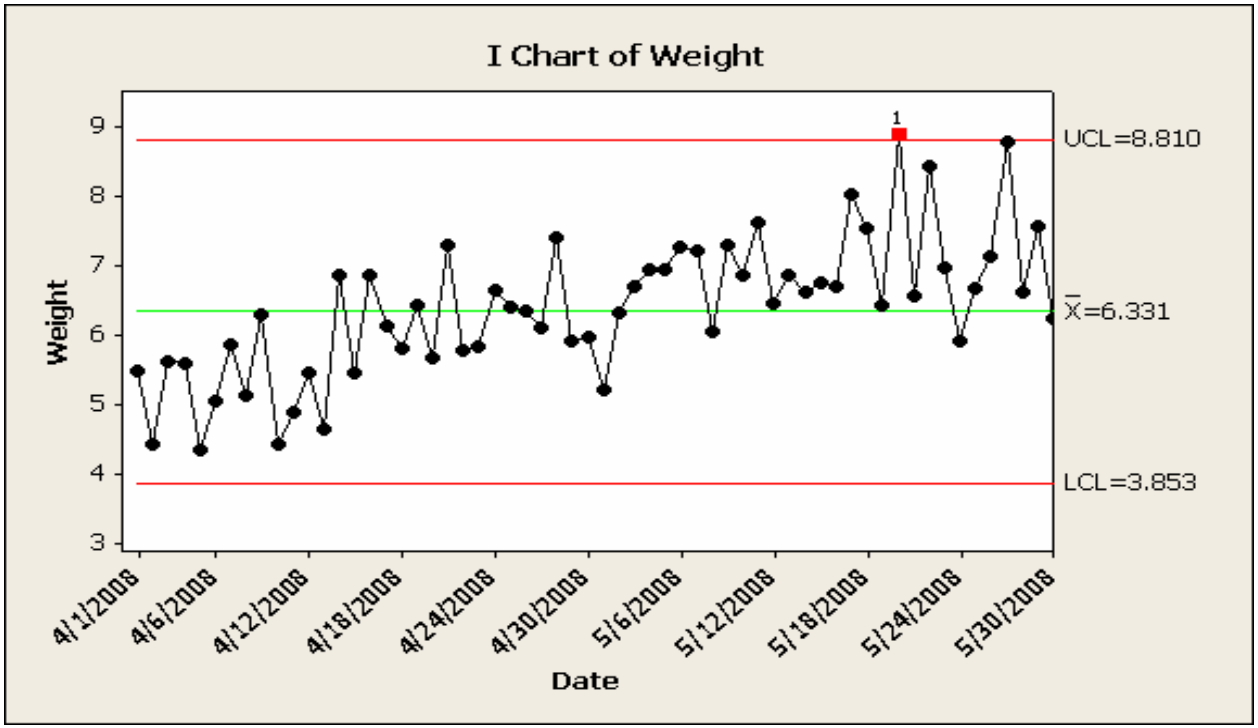
Mistake 5: Assuming Correlation = Causation

- There was also an increase in production, employee hours, facility air temperature, and the speed of the manufacturing line was increased several times during April and May.



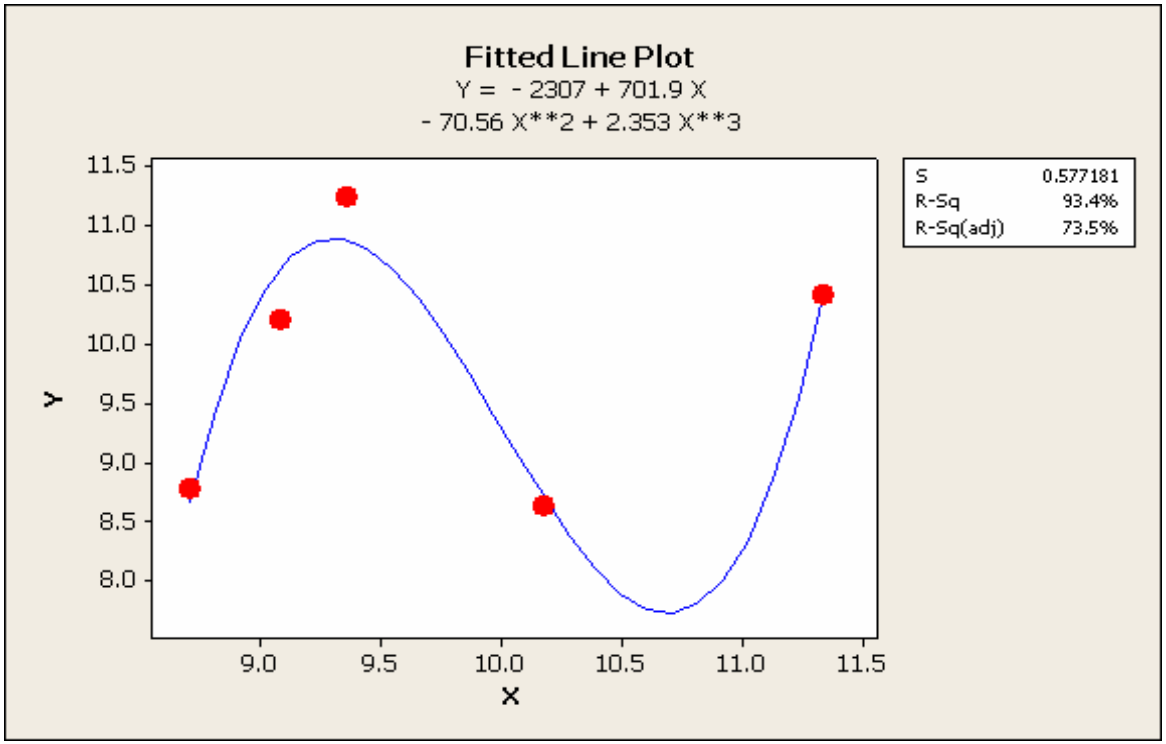
Mistake 5: Assuming Correlation = Causation

- It is difficult to tell which, if any, of those variables has a cause and effect relationship with Weight.



Mistake 6: Evaluating a Regression Model Exclusively on its R-squared value.

- R-squared represents the proportion of variation in the *sample* that is explained by the model.
- Example 1: Simulated X and Y data (no relationship)



Mistake 6: Evaluating a Regression Model Exclusively on its R-squared value.

- Example 2: 50 simulated X's and 1 simulated Y (no relationship). Final Model:

```
The regression equation is
C51 = 5.36 + 0.702 C34 - 0.545 C39 + 0.327 C46


Predictor      Coef  SE Coef      T      P
Constant      5.362    2.369   2.26  0.038
C34          0.7021   0.1568   4.48  0.000
C39         -0.5448   0.1594  -3.42  0.004
C46          0.3273   0.1366   2.40  0.029


R-Sq = 67.1%  R-Sq(pred) = 41.01%
```

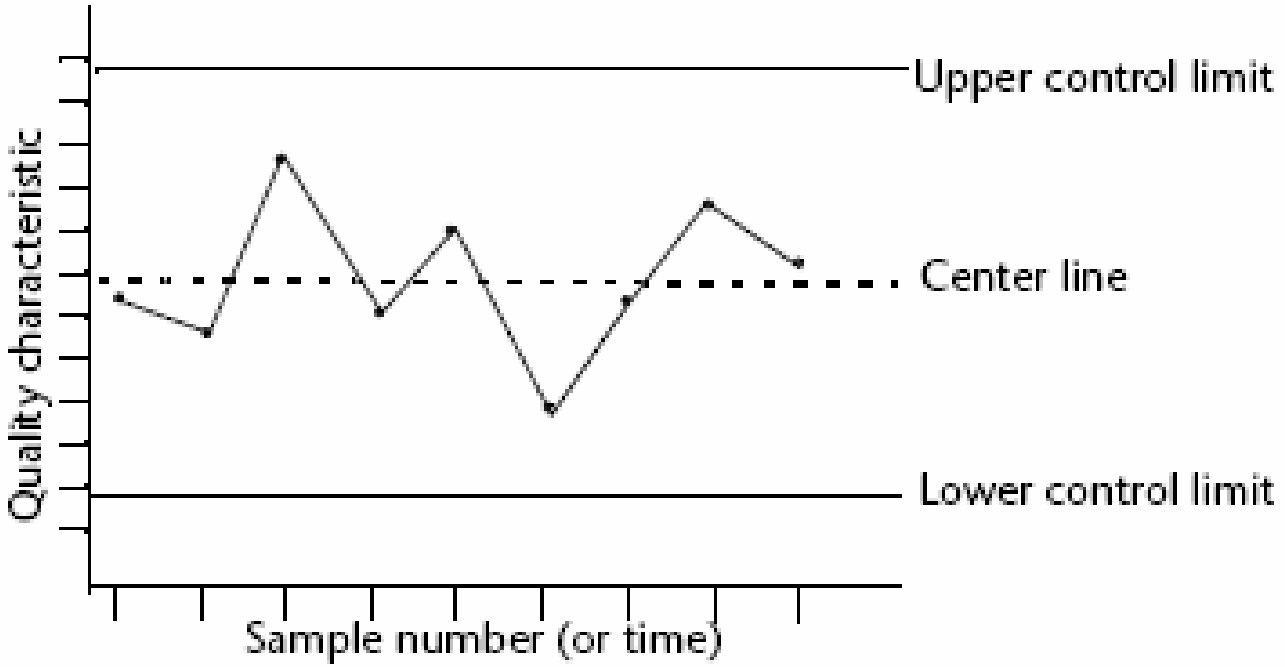
Mistake 6: Evaluating a Regression Model Exclusively on its R-squared value.

- These two examples show R-squared can be high even when the model is meaningless.
- R-squared predicted is a better measure.

7: Misuse of Control Charts

- Confusing specification limits and control limits

Structure of a control chart



7: Misuse of Control Charts

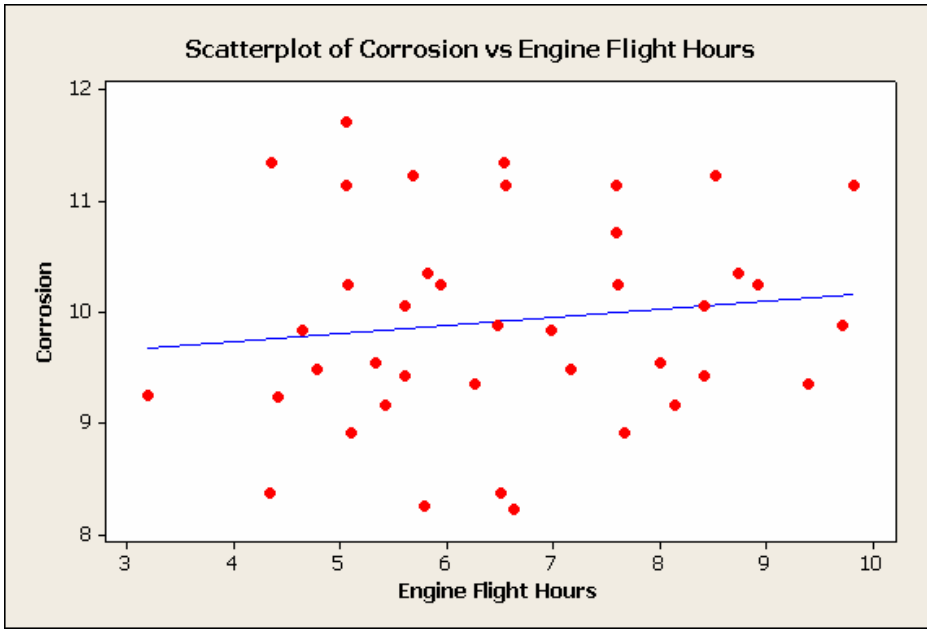
- Not including a large enough base period to get good estimate of the control limits
- Not fixing the control limits once there is a stable process

Mistake 8: Analyzing variables one-at-a-time.

- Variables not accounted for may bias the results.
- Example: Predicting amount of corrosion on an aircraft component from engine flight hours.
- There appears to be no relationship.

Analysis of Variance for Engine Flight Hours

Source	DF	Seq SS	Adj SS	Adj MS	F	P
FlightHours	1	1.825	1.825	1.825	0.66	0.423
Error	38	105.745	105.745	2.783		
Total	39	107.570				

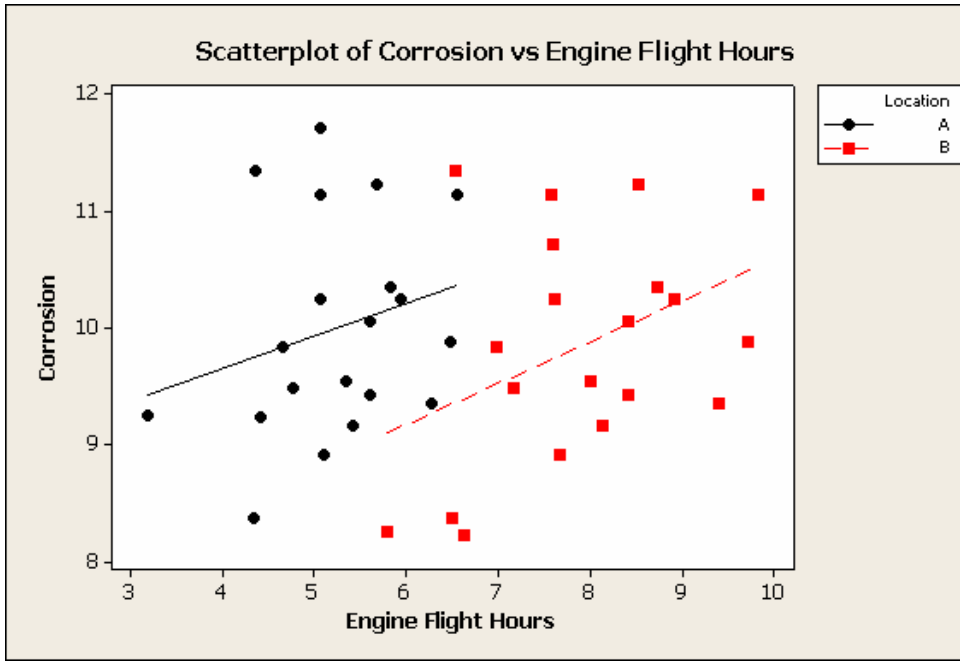


Mistake 8: Analyzing variables one-at-a-time.

- After including Location in the analysis, Flight Hours becomes significant.

Analysis of Variance for Engine Flight Hours

Source	DF	Seq SS	Adj SS	Adj MS	F	P
FlightHours	1	1.825	4.151	4.151	4.77	0.035
Location	1	73.559	73.559	73.559	84.56	0.000
Error	37	32.186	32.186	0.870		
Total	39	107.570				



Mistake 8: Analyzing variables one-at-a-time.

- Variables not accounted for will likely bias the results.
- The variation in the response due to variables not accounted for goes into the error.
- You can not look at interactions when analyzing one variable at-a-time.
- The type 1 error rate is not controlled.

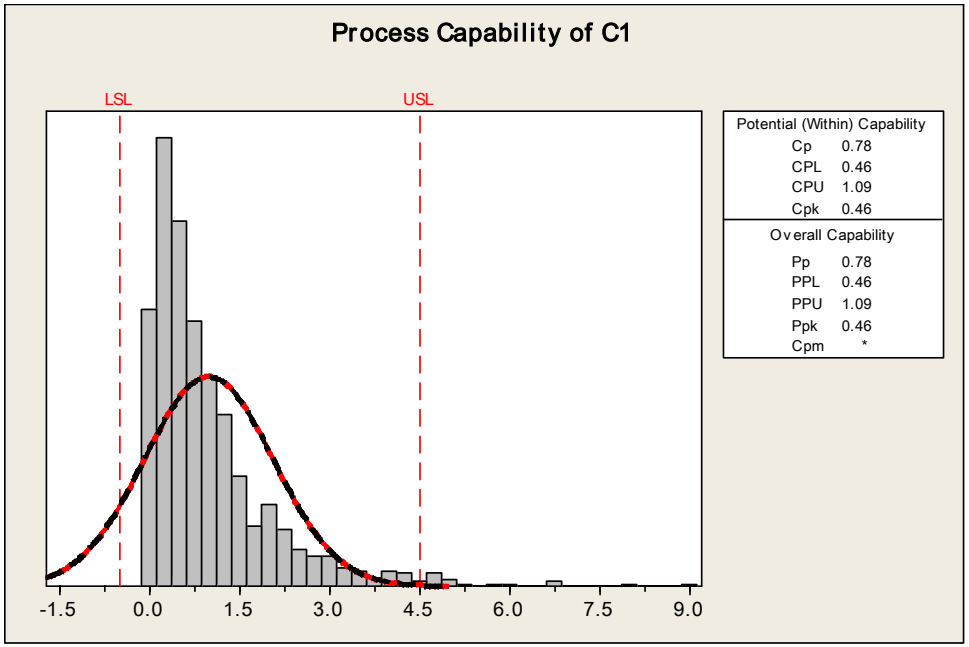
Mistake 9: Making inferences to a population that the sample does not represent.

- In capability analysis, data from one shift or one day is sometimes inappropriately used to estimate the capability of the entire manufacturing process.

- To avoid this, define the population before sampling and take a sample that represents the population.

10: Applying Normal Distribution to Nonnormal Data in a Capability Study

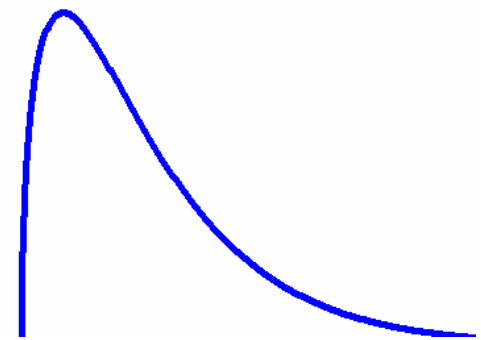
- The underlying distribution is not well modeled.
  - Capability metrics are highly sensitive to the assumed distribution model.
  - Finding an appropriate distribution to fit the data is ***extremely important.***

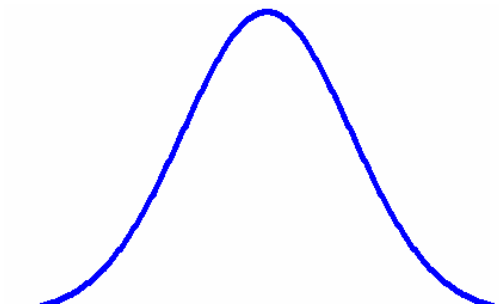


10: Applying Normal Distribution to Nonnormal Data in a Capability Study

- What to do with nonnormal data?

  - Transform the data
    - In statistics, data transformation is carried out in order to transform the data and ensure that it has a normal distribution.
    - Your data and spec limits will change.
    - Box Cox and Johnson





10: Applying Normal Distribution to Nonnormal Data in a Capability Study

- What to do with nonnormal data?
- Identify the distribution using the “Individual Distribution Identification”
- Your limits and specs will NOT change.
- Minitab provides 14 different distributions.

10.1: Not Seeking the Advice of an Expert

- Employees are sometimes placed in statistical training programs with the expectation they will come out experienced statisticians.
- While this training is excellent for the basic statistical projects, it's not enough to handle the more advanced issues that sometimes arise.
- Minitab's Mentoring service can be used to provide the necessary expert support.
- For more information, go to www.minitab.com/mentoring or contact Minitab Australia on 61-(0)2-93123700.

Summary

- 1: Misinterpreting overlapping confidence intervals.
- 2: Not distinguishing between statistical significance and practical significance.
- 3: Rejecting normality without reason.
- 4: Stating you've proved the null hypothesis (H_0).
- 5: Assuming Correlation = Causation.
- 6: Evaluating a regression model exclusively on its R-squared value.
- 7: Misuse of control charts
- 8: Analyzing variables one-at-a-time.
- 9: Making inferences to a population that the sample does not represent.
- 10: Applying normal distribution to nonnormal data in a capability study.
    - 10:1 – not seeking the advice of an expert

Ten Common and Dangerous Statistical Mistakes

Presenter: **Ross Farrelly**
Minitab Australia
61-2-9312 3763
rfarrelly@minitab.com.au

Minitab