NORTHROP GRUMMAN

Transforming Your Way to Control Charts That Work

November 19, 2009

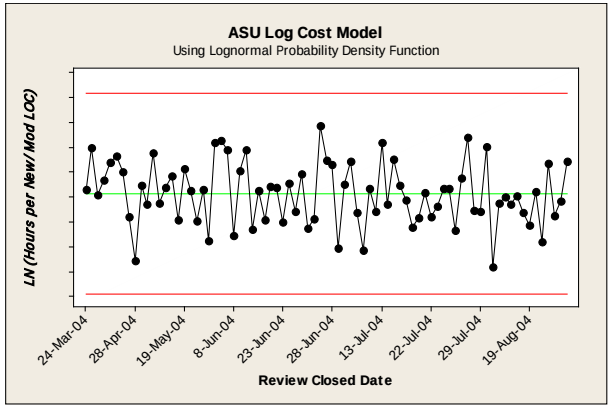
Richard L. W. Welch
Associate Technical Fellow
Robert M. Sabatino
Six Sigma Black Belt
Northrop Grumman Corporation

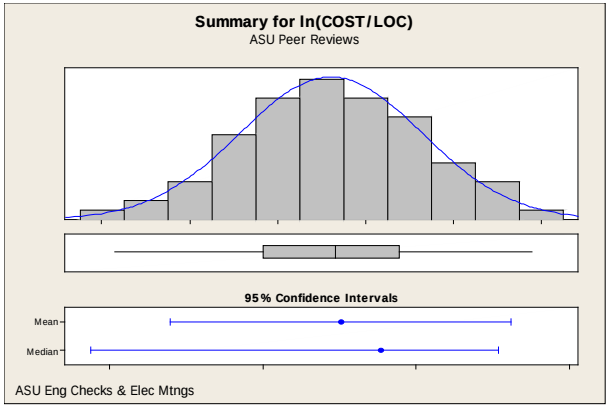
Outline

- The quest for high maturity
  - Why we use transformations
- Expert advice
- Observations & recommendations
- Case studies
  - Software code inspections
  - Drawing errors
- Counterexample
  - Software test failures
- Summary

The Quest for High Maturity

- You want to be Level 5
- Your CMMI appraiser tells you to manage your code inspections with statistical process control
- You find out you need control charts
- You check some textbooks. They say that “in-control” looks like this

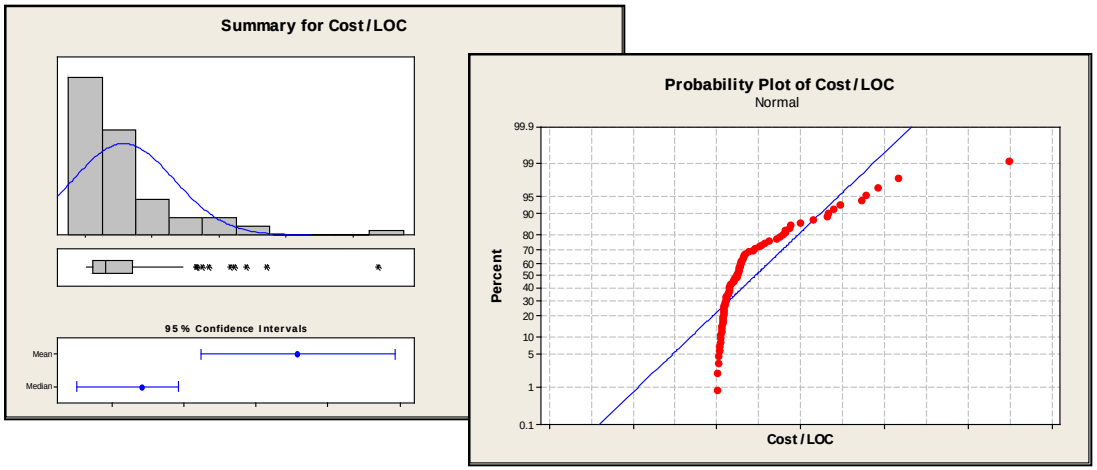




- You collect some peer review data
- You plot your data . . .

A typical situation (like ours, 5 years ago)

The Reality

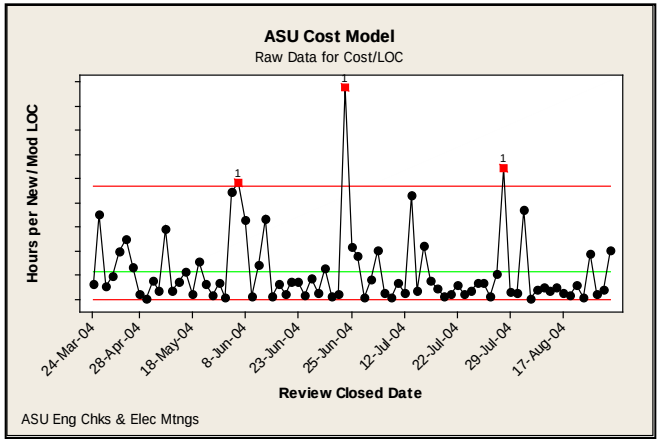


The data

- Highly asymmetrical
- Flunks Anderson-Darling test for normality ($p < 0.005$)

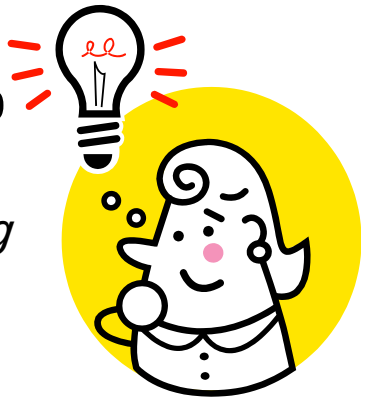
The control chart

- Penalizes due diligence in reviews
  - Up to 11% false alarm rate (Chebyshev's inequality)
- Doesn't flag superficial reviews
  - No lower control limit
- Skews the central tendency
  - Average cost looks like it busts the budget



What do you do?

An Idea



- Someone suggests that a data transformation might help
- You decide to hit the textbooks. Two of the best are
  - Donald Wheeler, *Understanding Variation: The Key to Managing Chaos*, 2nd edition, SPC Press, 2000
  - Douglas Montgomery, *Design and Analysis of Experiments*, 6th edition, Wiley, 2005
- We'll see what they have to say in a minute, but first let's look at some factoids about data transformations . . .

Data Transformation Factoids

- A data transformation is a mathematical function that converts your data into something else
  - For example, converting temperature from Fahrenheit to Celsius
- When data do not meet the necessary characteristics to apply a desired probability model, the right data transformation *can* help
- Data transformations are encountered in many statistical applications
  - Data analysis
  - Experimental statistics
  - Statistical process control (SPC)
  - Pattern recognition
- At Aerospace Systems, 40% of our control charts use a data transformation
  - We control about 50 different product development metrics

This presentation contains several slides intended to amplify the main discussion with technical detail meant to provide additional background information. We have labeled these “factoids.”

NORTHROP GRUMMAN

Expert Advice

Dueling Experts

NORTHROP GRUMMAN

- Unfortunately, they give conflicting advice

Wheeler favors the analysis of raw data

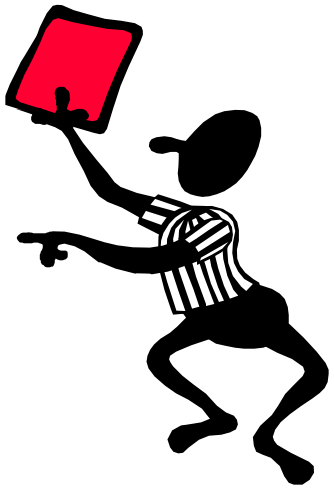
Montgomery advocates data transformations prior to analysis

Who can we believe?

Wheeler

- Doesn't favor transforming data
- Reasons that you cannot always meet the assumptions for specialized models like the binomial or Poisson
- Favors empirical use of individuals/moving range charts for SPC or equivalent techniques for other applications because of their simplicity

  - But this minimizes difficulties encountered with non-normal or non-symmetric data
  - Practitioners must take care not to overvalue the utility of 2-sigma and 3-sigma confidence bands/control limits for non-normal data



Calculus Factoids

- Tchebysheff's Inequality
  - For a random variable with finite variance (and no particular distribution), ± k-sigma limits will be exceeded $1/k^2$ * 100% of the time
    - For k=2, a 2-sigma confidence limit can be violated up to 25% of the time – not 5%!
    - For k=3, the 3-sigma control bands can be exceeded up to 11% of the time – not 0.3%!
  - This can result in an unworkable rate of false positives

For an extreme case of "dirty" data, you could need ± 18-sigma control limits to achieve to same degree of control as ± 3-sigma limits with well-behaved, normally distributed data.

But we never have dirty data, so why worry?

Montgomery

- Boldly recommends
  - Logarithmic transformation $y_{ij} = \ln x_{ij}$ for lognormal data
  - Square root transformation $y_{ij} = (x_{ij})^{0.5}$ or $y_{ij} = (1 + x_{ij})^{0.5}$ for Poisson data
  - Arcsine transformation $y_{ij} = \arcsin x_{ij}$ for point-binomial data (that is, binomial data expressed in fractional form)
  - Empirically-derived transformations for other cases (think Box-Cox)
- Focuses on stabilizing the variance

  - Transformations work when there is some underlying physical causality
    - How do you know?
    - What happens when you have small samples and take one more data point?
  - You may be sacrificing desirable decision-theoretic properties like unbiasedness, maximum likelihood, or uniformly most powerful

Decision Theory Factoids

NORTHROP GRUMMAN

- Unbiased
  - An unbiased estimator is, in the long run, accurate – that is, it has no non-random component of error
- Maximum Likelihood
  - A maximum likelihood estimator is the one most likely to be correct
- Uniformly Most Powerful
  - A uniformly most powerful decision rule has the best chance of flagging any discrepant or out-of-control condition

Statisticians do not abandon these properties without good reason.

NORTHROP GRUMMAN

Observations & Recommendations

Listen to both experts

Observations

- The Central Limit Theorem is a powerful ally
  - Any distribution with a finite variance can be approximated with a normal distribution by taking a large enough sample size *n*
  - Error of approximation varies as $1/n^{0.5}$
  - In a practical sense, mild excursions from normality can be mitigated by observing more data
- A transformation can be worthwhile
  - When a physical model explains why it should work
  - When an empirical model is well-founded in a large amount of data
    - Many studies of similar phenomena
    - Large data sets
- Transformations can be difficult
  - Mathematically messy
  - Hard to interpret
  - Hard to explain
  - Cultural resistance to their use

Recommendations

- Due diligence requires that we investigate and compare using raw data vs. transformed data
  - Modern tools limit us only by our imaginations
- Simpler is better unless there is harm
  - Statistical decision theory embraces the double negative: we do not abandon the easiest approach until we compile compelling evidence that it doesn't work
  - Consider both the incidence rates and costs of wrong decisions (false positives and negatives, Type I and II errors, etc.)
- The "correct" answer is not pre-determined
  - In 1936, R. A. Fisher derived the first multivariate normally based classifier
  - In 1976, the senior author showed Fisher's data were lognormally distributed and a transformation gave more accurate results
  - Fisher's error was 3.3%; the author's was 2%
  - Is a difference between 3.3% and 2% meaningful to your application?

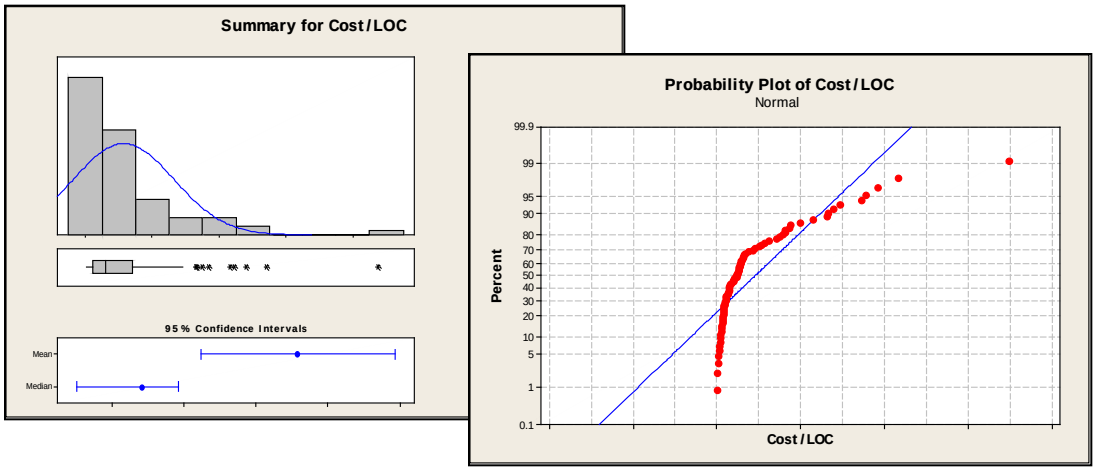
Fisher's paper is "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179-188. The lognormality of the data was shown in the senior author's PhD thesis *Studies in Bayesian Discriminant Analysis*, and published in the 1979 paper "Multivariate Non-Gaussian Bayes Discriminant Analysis", Statistica, 1979:1, pp 13-24, with C. P. Tsokos.

NORTHROP GRUMMAN

Software Code Inspections

Case Study #1

Reality, Part Two

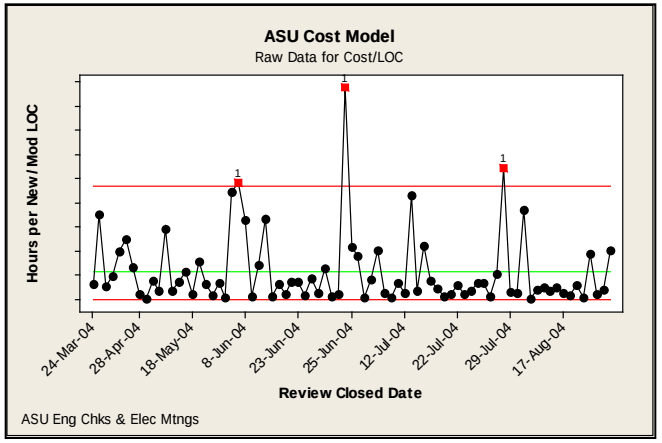


The data

- Highly asymmetrical
- Flunks Anderson-Darling test for normality ($p < 0.005$)

The control chart

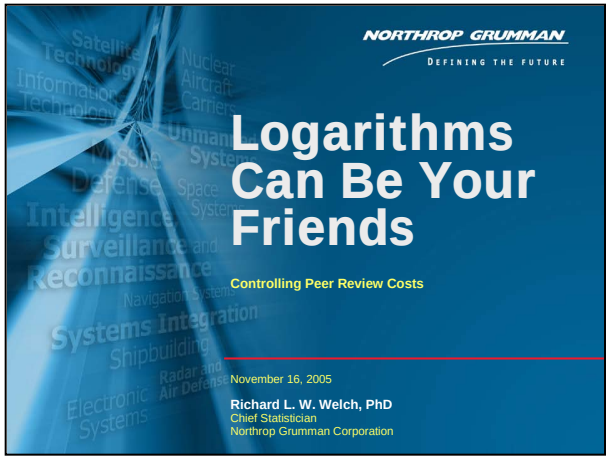
- Penalizes due diligence in reviews
  - 11% false alarm rate (Chebyshev's inequality)
- Doesn't flag superficial reviews
  - No lower control limit
- Skews the central tendency
  - Average cost looks like it is busting the budget



What do you do?

Stabilizing the Data

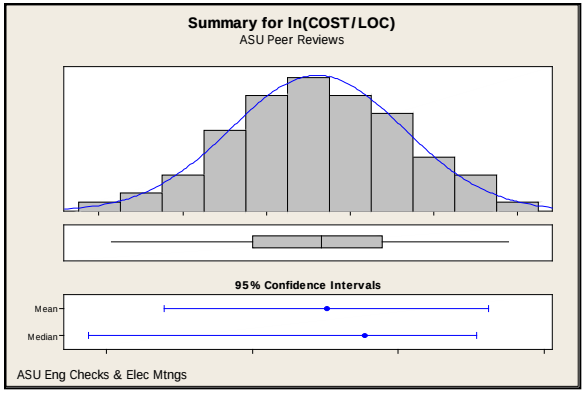
- Senior author's presentation at 2005 CMMI[SM] Technology Conference demonstrated how a log-cost model can successfully control software code inspections

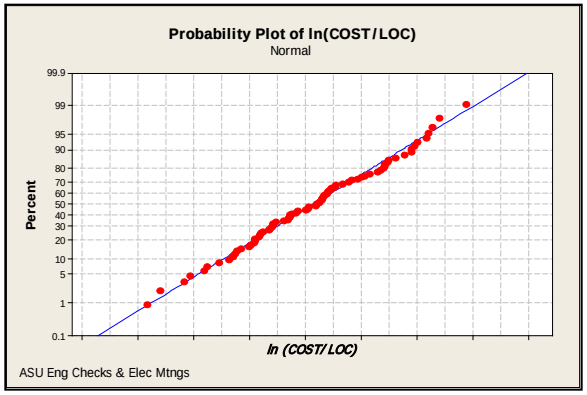


  - Peer review unit costs (hours per line of code) behave like commodity prices in the short term
  - Short term commodity price fluctuations follow a lognormal distribution
  - As a consequence, commodity prices follow a lognormal distribution
  - Therefore, taking the natural logarithm of a sequence of peer review costs transforms the sequence to a normally distributed series

Notes:
- Details on the log-cost model, "one of the most ubiquitous models in finance," can be found at riskglossary.com (http://www.riskglossary.com/articles/lognormal_distribution.htm)
- Prior CMMI Technology Conference & User Group papers are published on-line at: http://www.dtic.mil/ndia/

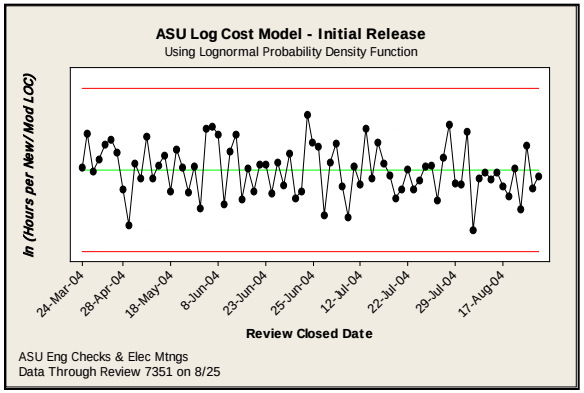
Our Data on Logs





Anderson-Darling test $p < 0.759$

- Impacts
  - Normality of the transformed data minimizes false alarms
  - We catch superficial reviews
  - Code reviews do not bust the budget
- Demonstrated utility & applicability
  - > 7,000 peer reviews over 6 years provide large sample validation



Demonstrating an in-control, stable process

Lognormal Factoid

Since the sample mean of the transformed data is

$$\bar{Y} = \frac{\sum_n \ln(x_i)}{n} = \frac{\ln \prod_n x_i}{n} = \ln \sqrt[n]{\prod_n x_i}$$

The inverse transformation results in the geometric mean of the untransformed data

$$e^{\bar{Y}} = \sqrt[n]{\prod_n X_i} = \bar{X}_{geo}$$

As a result of a similar derivation, the inverse transformation of the sample standard deviation is the geometric standard deviation of the untransformed data

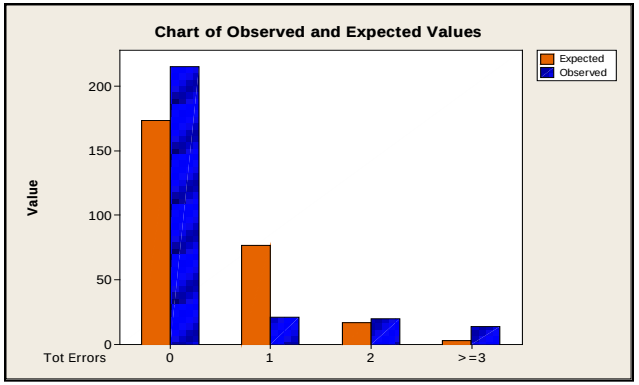
NORTHROP GRUMMAN

Drawing Errors

Case Study #2

Drawing Errors

- Drawing errors follow a Poisson distribution
  - Data are discrete counts of defects per drawing
  - C charts are commonly used to analyze Poisson-distributed defect data

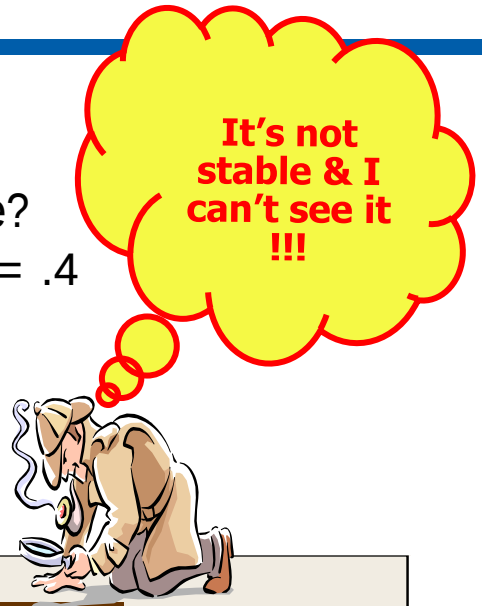


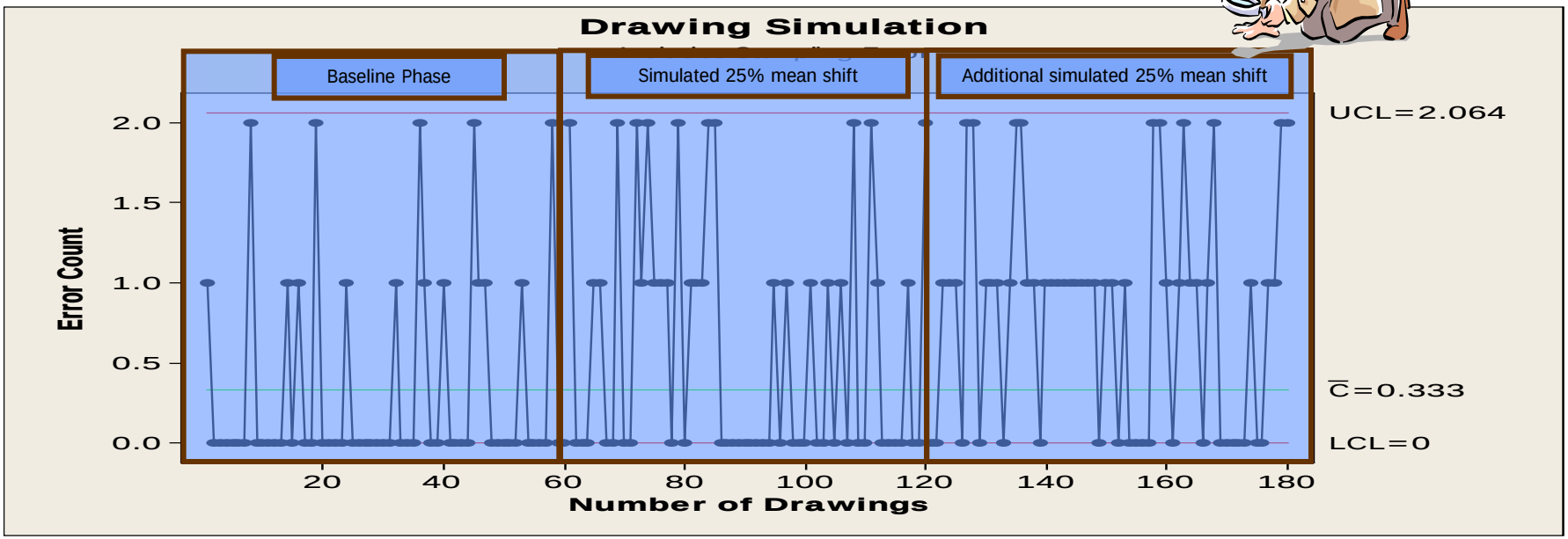
- C-charts lack insight
  - No ready indicator of deteriorating or improving process performance
    - Control limits establish the process capability for expected number of defects *per drawing*
  - Special causes are lagging indicators
  - Fractional control limits confuse data analysts
    - For example, UCL =2.4 defects. What are 2.4 defects? 3 defects are a special cause and 2 are not. But what about a run of consecutive drawings, each with two errors?

C charts detect performance changes poorly

C Chart Performance

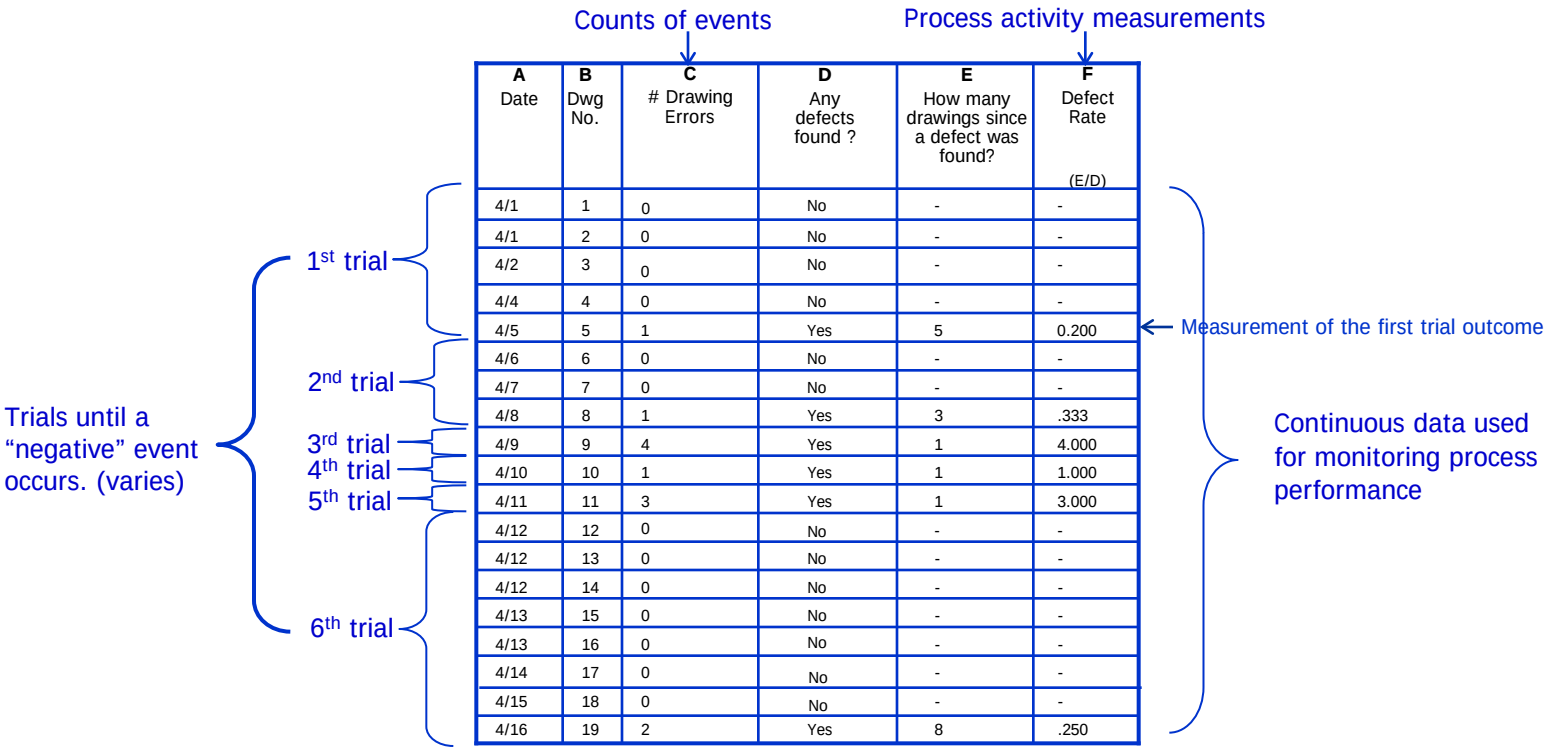
- Is this a stable process?
- How sensitive is the C Chart to changes in process performance?
  - Baseline phase simulated using Poisson distribution with mean = .4
  - Increased mean value by an additional 25% in two successive steps
  - Sample size = 60 drawings





Detecting Shifts in Process Performance

- Wheeler suggests transforming raw "counts of events" to "measurements of process activity" (i.e. rates)



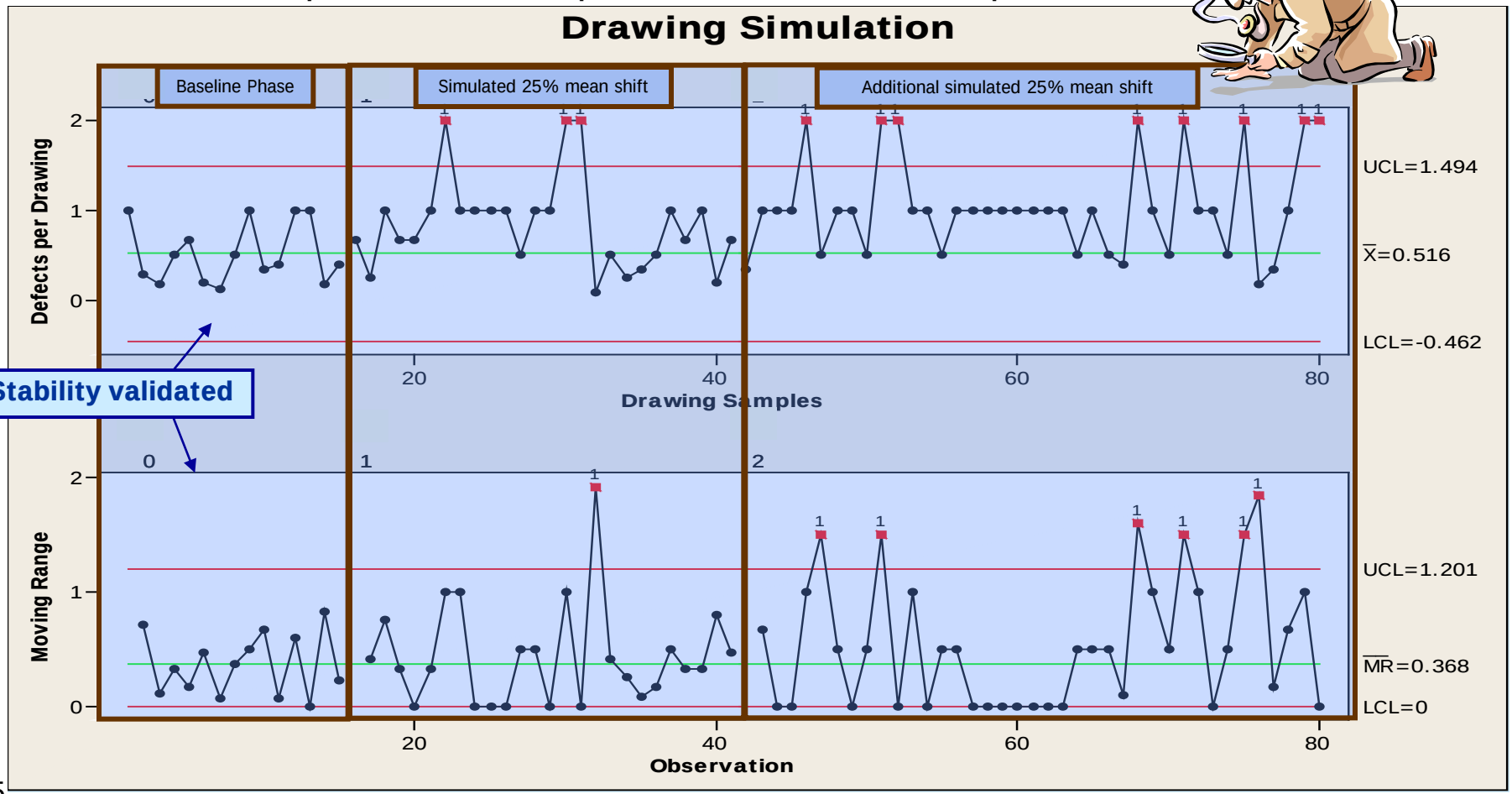
A Date	B Dwg No.	C # Drawing Errors	D Any defects found ?	E How many drawings since a defect was found?	F Defect Rate (E/D)
4/1	1	0	No	-	-
4/1	2	0	No	-	-
4/2	3	0	No	-	-
4/4	4	0	No	-	-
4/5	5	1	Yes	5	0.200
4/6	6	0	No	-	-
4/7	7	0	No	-	-
4/8	8	1	Yes	3	.333
4/9	9	4	Yes	1	4.000
4/10	10	1	Yes	1	1.000
4/11	11	3	Yes	1	3.000
4/12	12	0	No	-	-
4/12	13	0	No	-	-
4/12	14	0	No	-	-
4/13	15	0	No	-	-
4/13	16	0	No	-	-
4/14	17	0	No	-	-
4/15	18	0	No	-	-
4/16	19	2	Yes	8	.250

- The defect rate series is plotted on an Individuals & Moving Range (ImR) chart & analyzed

Yes, this was suggested by the same guy who doesn't like transformations. See Donald Wheeler, *Understanding Variation: The Key to Managing Chaos*, 2nd edition, SPC Press, 2000, pp. 100-104.

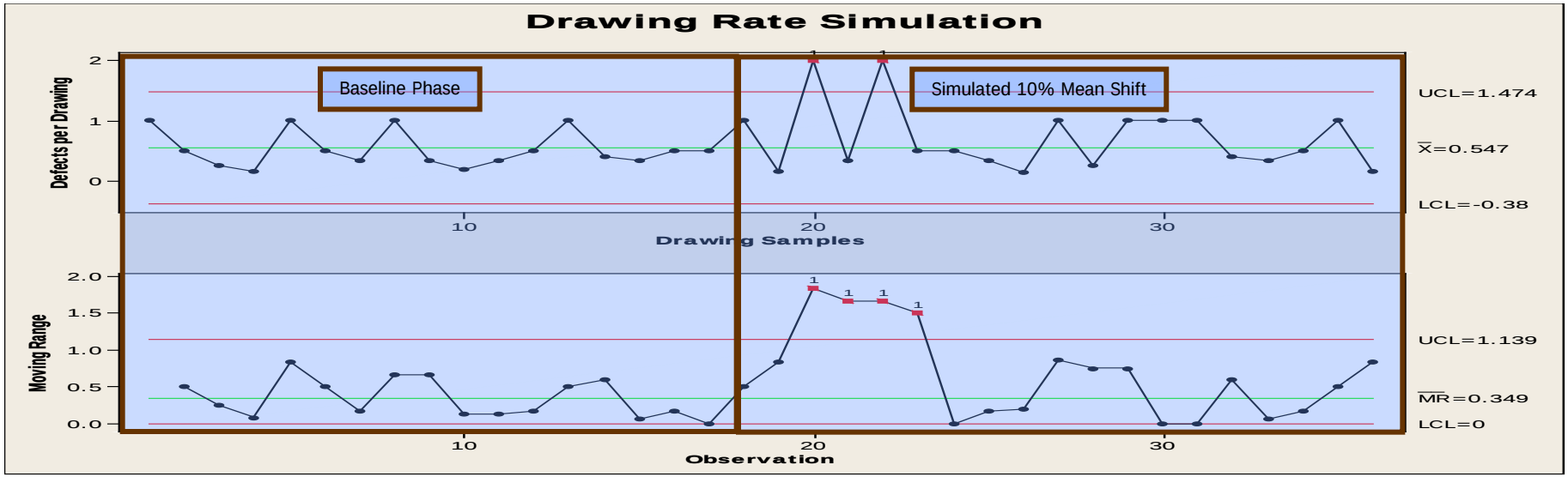
ImR Control Chart Sensitivity to Shifts in Process Performance

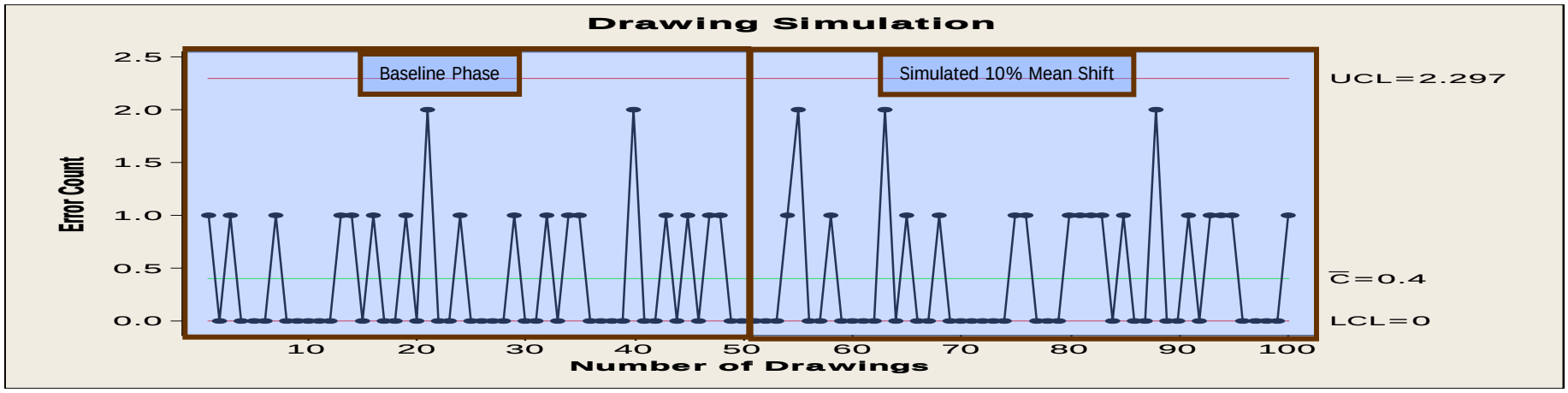
- Is this a stable process?
  - Same simulation as the C chart
  - Data was *transformed* to establish defect rates
  - Control chart parameters computed from the baseline phase



Sensitivity Comparison

- This simulation models a 10% process shift
- What chart would you use?





ImR Chart Factoids

- The mR chart portion monitors the process based on points representing the differences (i.e., range) between each 2 consecutive individual data points
  - Assesses whether the process variation is in control
  - The mR chart *should be in control* before you establish the baseline with the I chart
  - If the mR chart is not in control, then the control limits for the I chart will be inflated (too forgiving) and may fail to signal an out-of-control condition
- The I chart portion monitors process based on individual data points (defect rates)
  - Assesses whether the process center (x) is in control
  - The process centerline is calculated from the average of the individual data points
  - The control limits are set a distance of 3σ above and below the center line and provide a visual display for the amount of process variation expected
- The ImR chart will not be hyper sensitive to rare events containing numerous errors

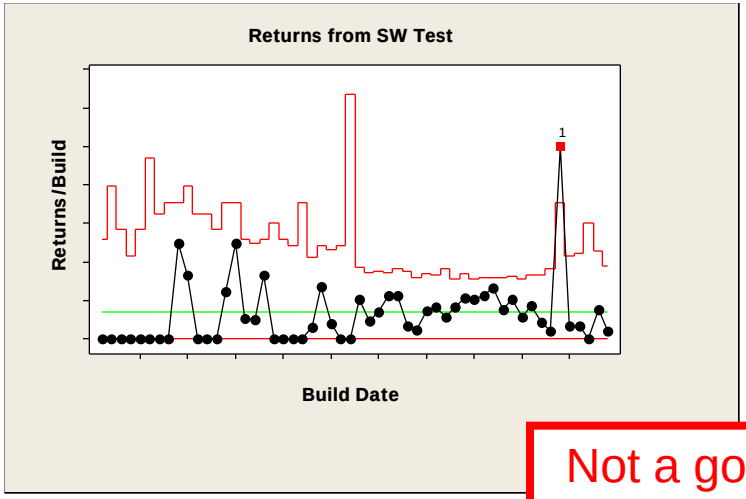
NORTHROP GRUMMAN

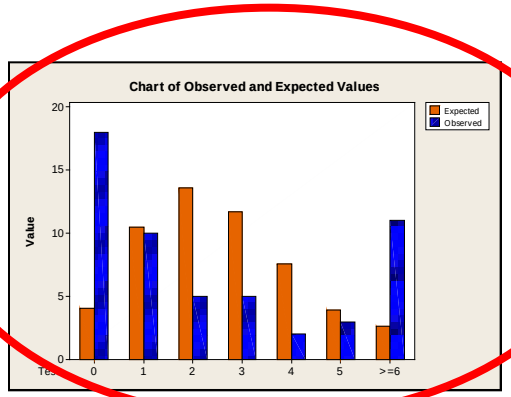
Software Test Failures

Counterexample

SW Test Returns – Poisson Model

- How often have you heard, as a universal truth
  - "Defects are Poisson distributed"?
- How about reality, as the Test group sees it?
- The u-chart plots SW test failures
  - This assumes SW test failures follow a Poisson distribution
  - Being diligent, we check the goodness of the model





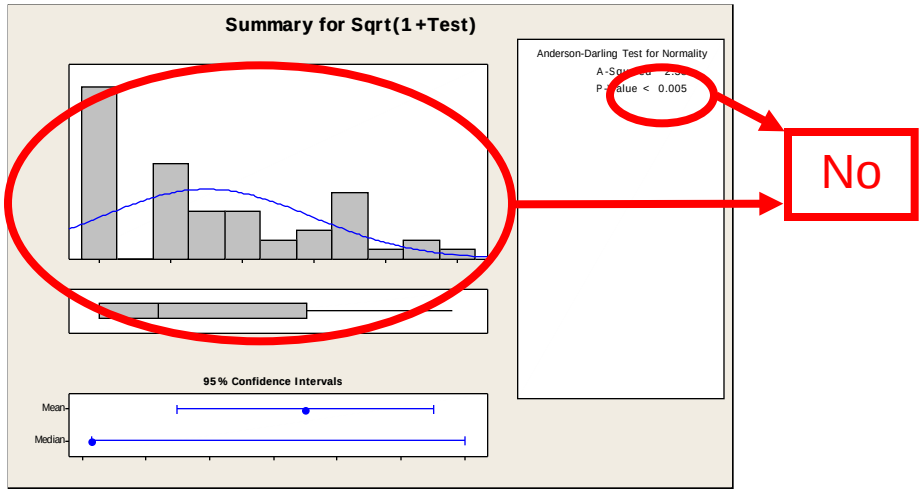
Not a good fit

χ^2 Test for Poisson

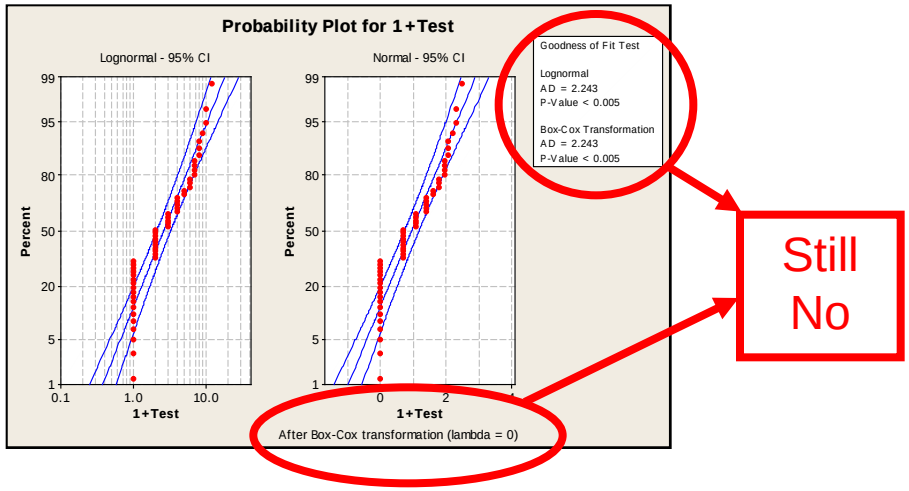
N	DF	Chi-Sq	p-Value
54	5	88.7177	0.000

Does Transforming the Data Help?

- Square root transformation
  - As recommended by Montgomery for data we thought should follow a Poisson model

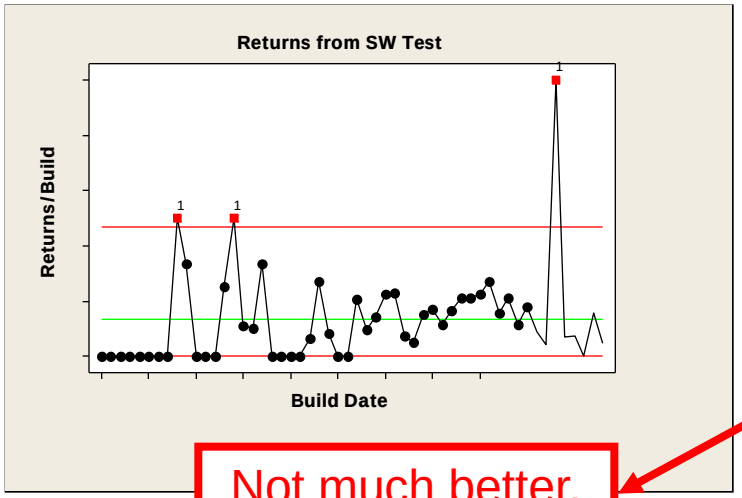


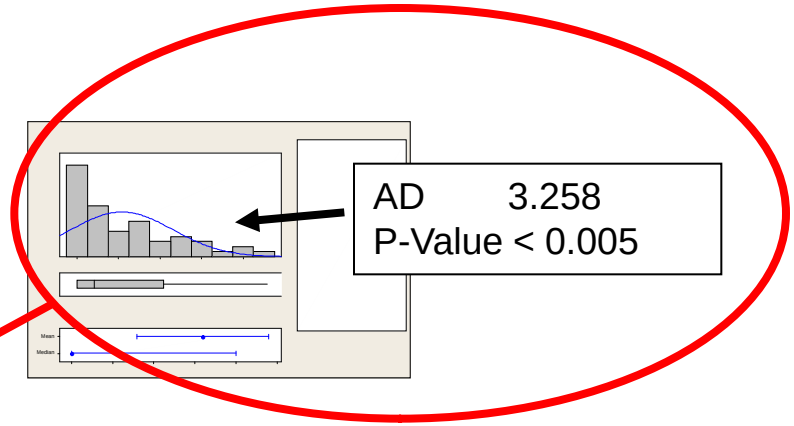
- Lognormal
  - When we try a Box-Cox transformation, we discover the optimal parameter lambda = 0
  - This is the same as a lognormal transformation



Discussion

- Sometimes no model seems to work
  - May have multiple failure mechanisms, with mixing of distributions
  - Need lots of data to sort out
- In this case, no obvious transformation suggests itself
- We revert to the basic principle that simpler is better
  - Individuals chart
  - Not great, but the simplest choice





Not much better, but simple

Summary

NORTHROP GRUMMAN

- The Central Limit Theorem is a powerful ally
- A transformation can be worthwhile
- Transformations can be difficult
- Due diligence requires that we investigate and compare using raw data vs. transformed data
- Simpler is better unless there is harm
- The “correct” answer is not pre-determined

When used carefully, transformations expand analysis capability

Richard L. W. Welch, PhD
Northrop Grumman Corporation
(321) 951-5072
Rick.Welch@ngc.com

Robert M. Sabatino
Northrop Grumman Corporation
(321) 726-7629
Robert.Sabatino@ngc.com