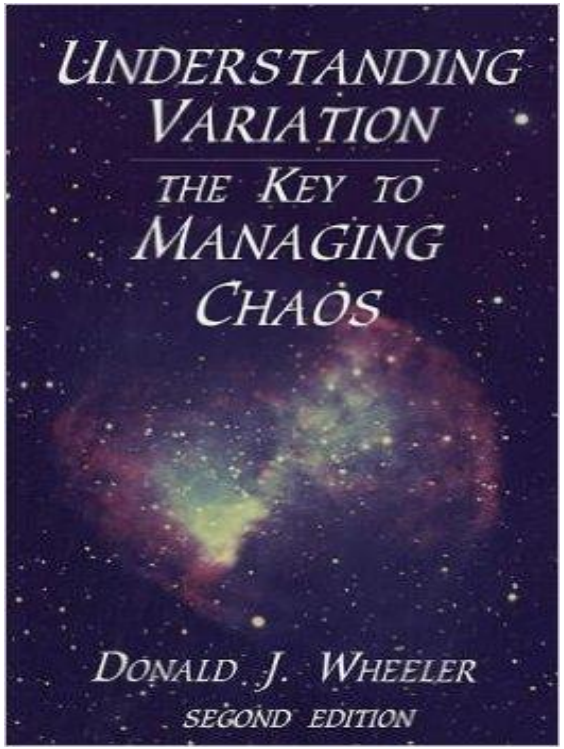
Business Data Analysis w/Excel

March 8th, 2017

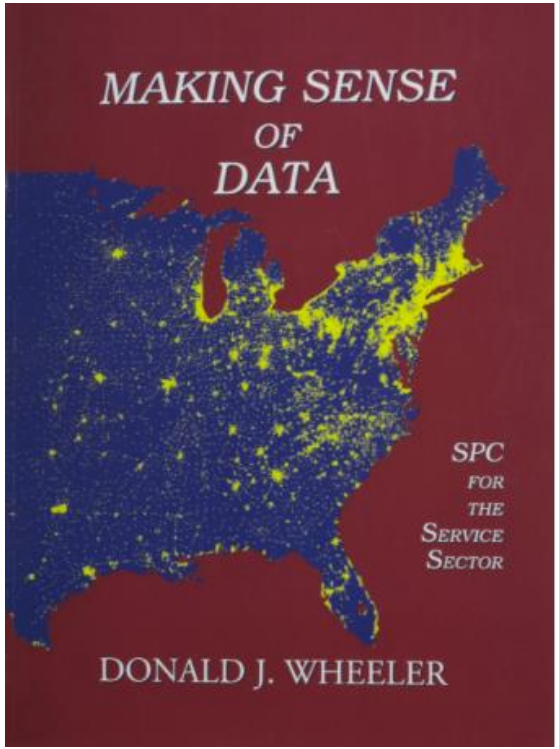
datasciencedojo
data science for everyone

Get these Books!



This one for your manager!

This one for you!



This deck uses many examples from this book!

datasciencedojo
data science for everyone

Who Am I?

- Dave Langer, VP of Data Science – Data Science Dojo

- 20+ years in technology:
  - Roles in development, architecture, & BI/DW/analytics.
  - Last job – Sr. Director, BI & Analytics @ Microsoft.

- Hooked on Data Science 5 years ago:
  - Extensive background in data and analytics.
  - Learned Machine Learning from 2nd place Netflix Prize winner.
  - #1 Data Scientist on YouTube.

- Joined Data Science Dojo to democratize Data Science.

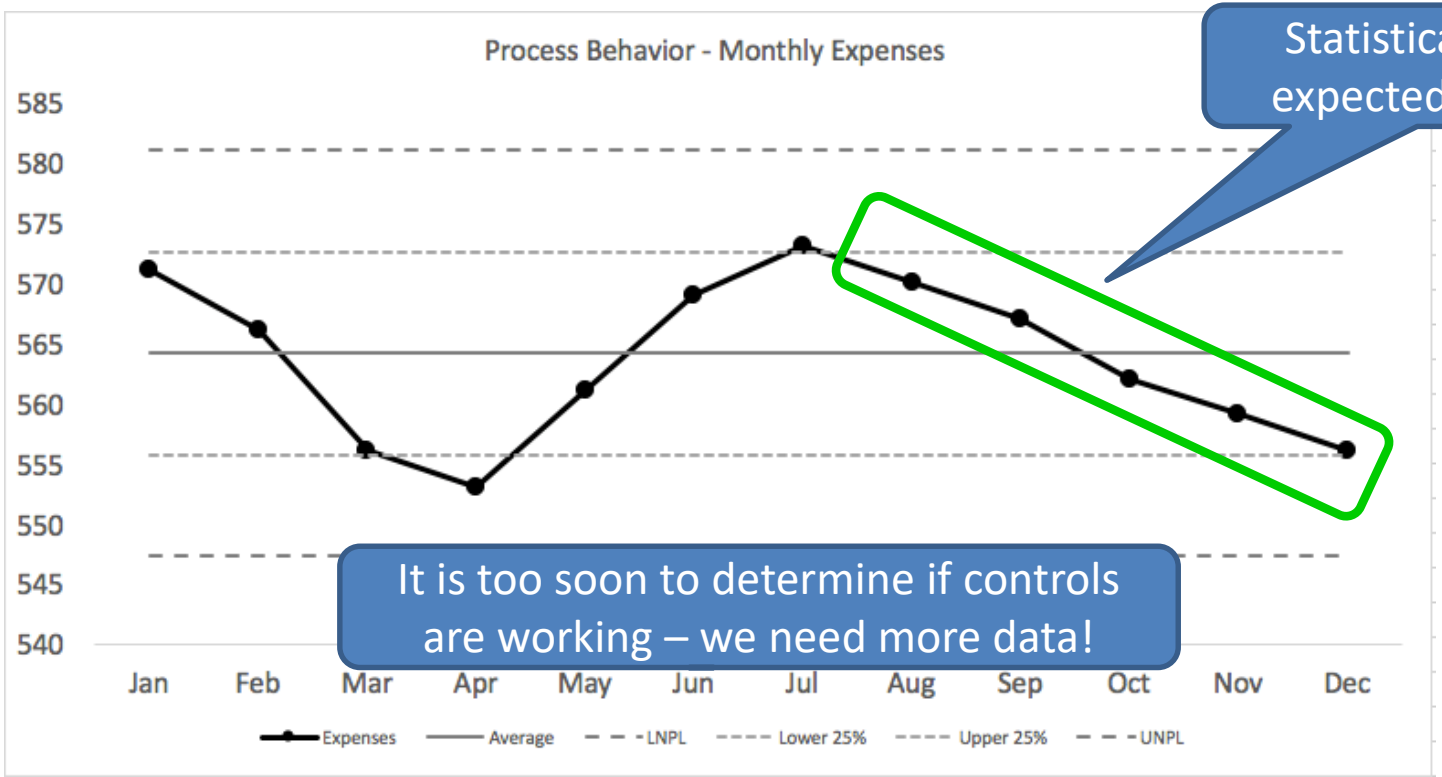
Motivation

- To deal with complexity and to compete effectively, business is increasingly data-driven:
  - Key Performance Indicators (KPIs)
  - Balanced Scorecards
  - Executive Dashboards

- However, skills to rigorously interpret, analyze, and understand this data is rare.

Can We Celebrate?

Not So Fast!



The Challenge

- Unfortunately, business data presents challenges:
  - Business data is often *aggregated*.
  - Business data is usually *autocorrelated*.
  - Small amounts of data is the norm.

Intuition

- Aggregated data:
  - Totaled by division, geography, time, etc.
- Autocorrelation – current values are related to previous values:
  - This quarter's revenue is related to (i.e., a function of) the previous quarter's revenue.
- If you have 2 years of quarterly data, you only have 8 values!

What We Need

- Tools that can deal with summarized business data.
- Techniques that accommodate small amounts of business data over time.
- Recipes, patterns and rules to use the tools and techniques to rigorously interpret, analyze, and understand business data.

Our Toolkit

- Histograms to understand the distributions of business data.
- Running records to identify trends over time.
- Process behavior charts to apply statistical rigor to analyze the changes and differences in our business.

Expectation Setting

- This presentation is very much the art of the possible:
  - Not enough time to teach Excel, resources in Appendix.

- You will not be an expert in these techniques:
  - We cannot cover all aspects of data analysis.
  - There are many gotchas and prerequisites for rigorous analysis of business data.
  - Good news – you don't need a PhD. in Statistics!

- Buy and study "Making Sense of Data"!

EXCEL WORKBOOK

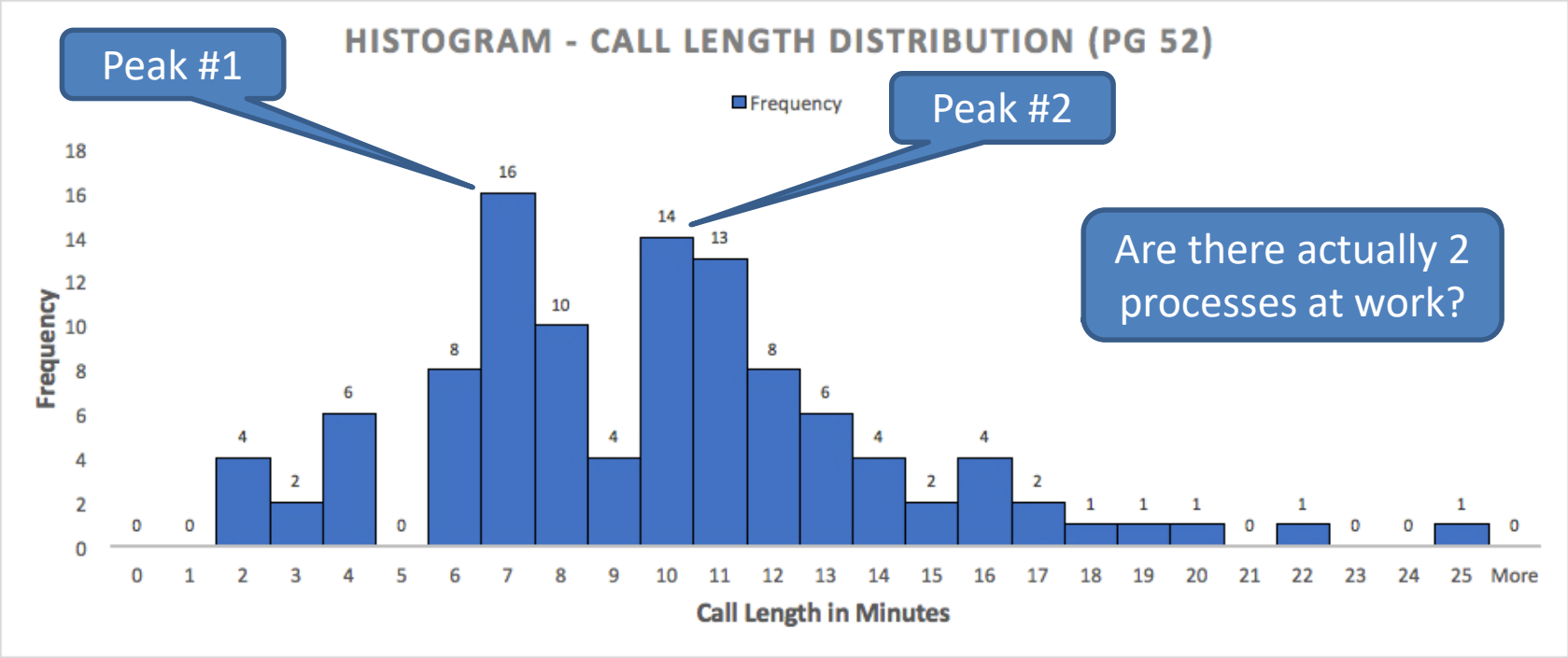
THE HISTOGRAM

Example Questions & Scenarios

- Questions:
  - “How is my business process actually executing?”
  - “Is my business process behaving as expected?”

- Scenarios:
  - You suspect that the nature of a business process is more complicated in practice.
  - Productivity isn’t what I expected, is something going on?

The Histogram



Source – “Making Sense of Data”, Page 52.

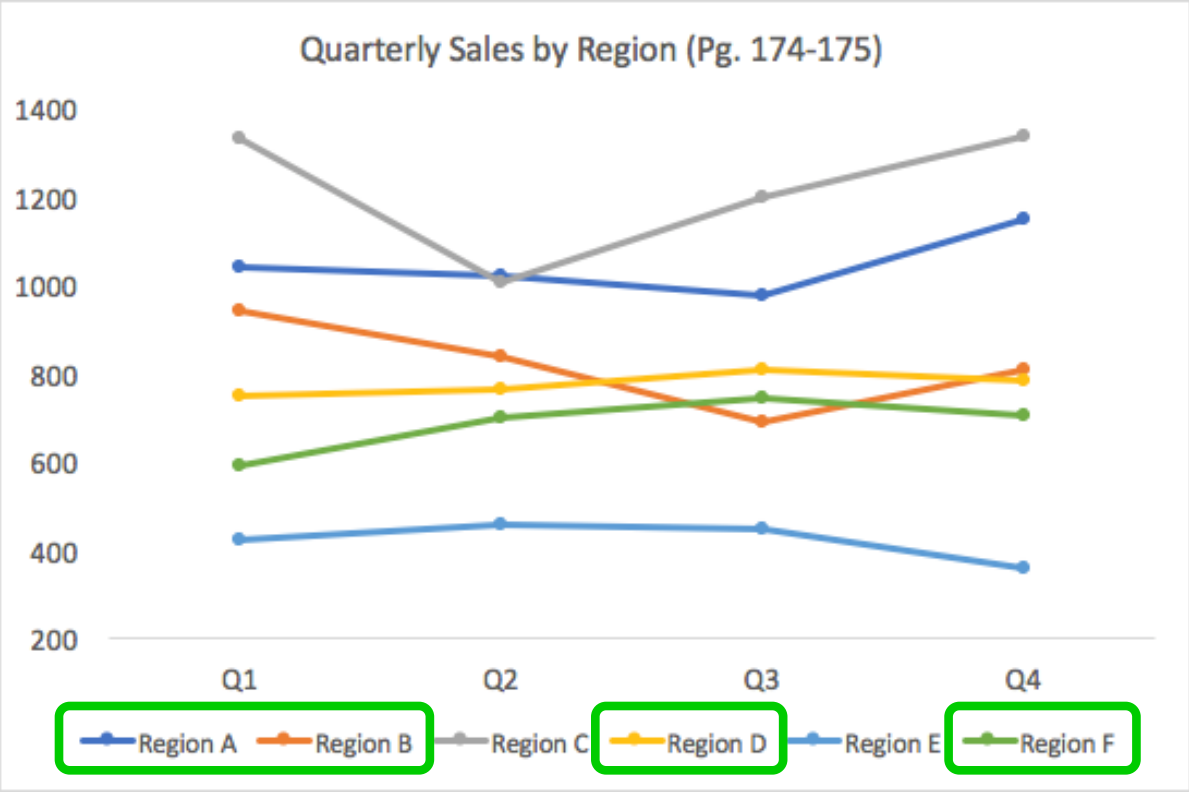
THE RUNNNING RECORD

Example Questions & Scenarios

- Questions:
  - "What has been happening over time?"

- Scenarios:
  - You've implemented a new marketing campaign, what has been happening?
  - Cost seem to be on the rise, how can we tell?

The Running Record



See trends over time

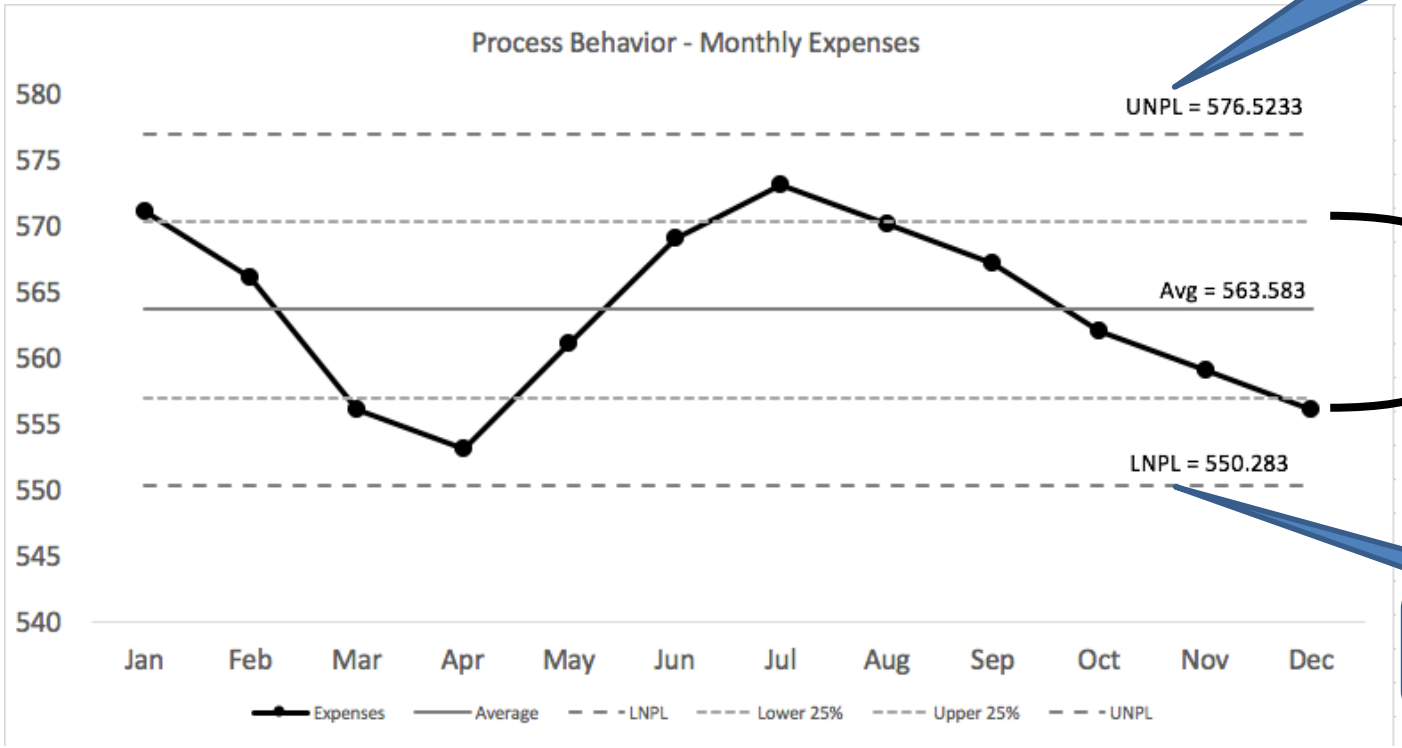
Any Associated Trends over time?

Can't be used to answer rigorous questions!

Source – "Making Sense of Data", Page 174-175.

PROCESS BEHAVIOR CHARTS

Understanding the Char



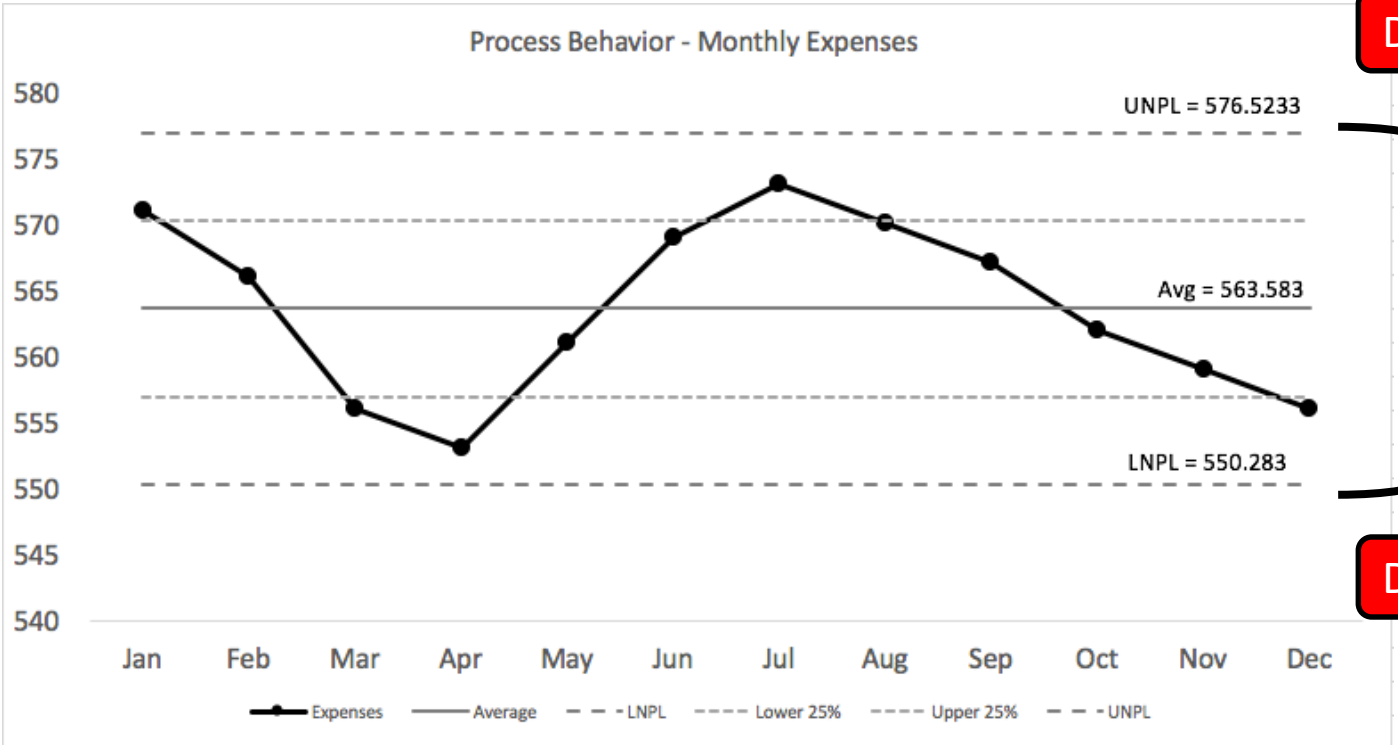
Upper Natural Process Limit

50+% of values

Lower Natural Process Limit

datasciencedojo
data science for everyone

Understanding the Chart



Danger Zone!

96+% of all values!

Danger Zone!

datasciencedojo
data science for everyone

Chart Requirements

- All the data being used was collected under similar conditions.
- The data can be logically compared.
- The charts used in this talk don't work with well with outliers.

TREND ANALYSIS

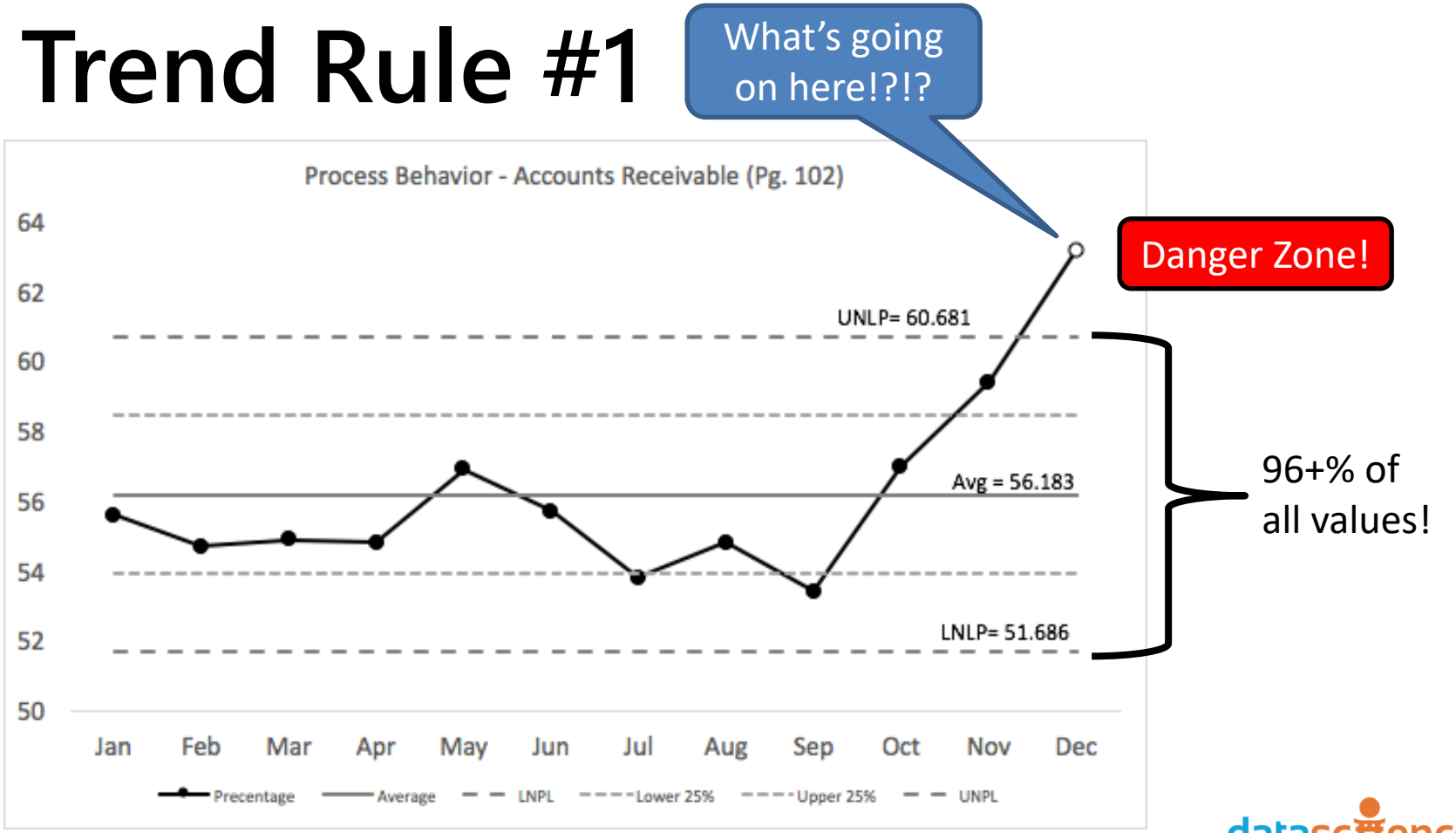
Example Questions & Scenarios

- Questions:
  - “Is it working?”
  - “Is there something going on here?”
- Scenarios:
  - You’ve implemented a new marketing campaign, is it actually moving the needle?
  - Cost seem to be on the rise, are they really?

The Rules of Trend Analysis

- Rule # 1 – Points Outside the Limits:
  - Single point outside the limits is an indication of a *dominant effect* that needs investigating.

Trend Rule #1

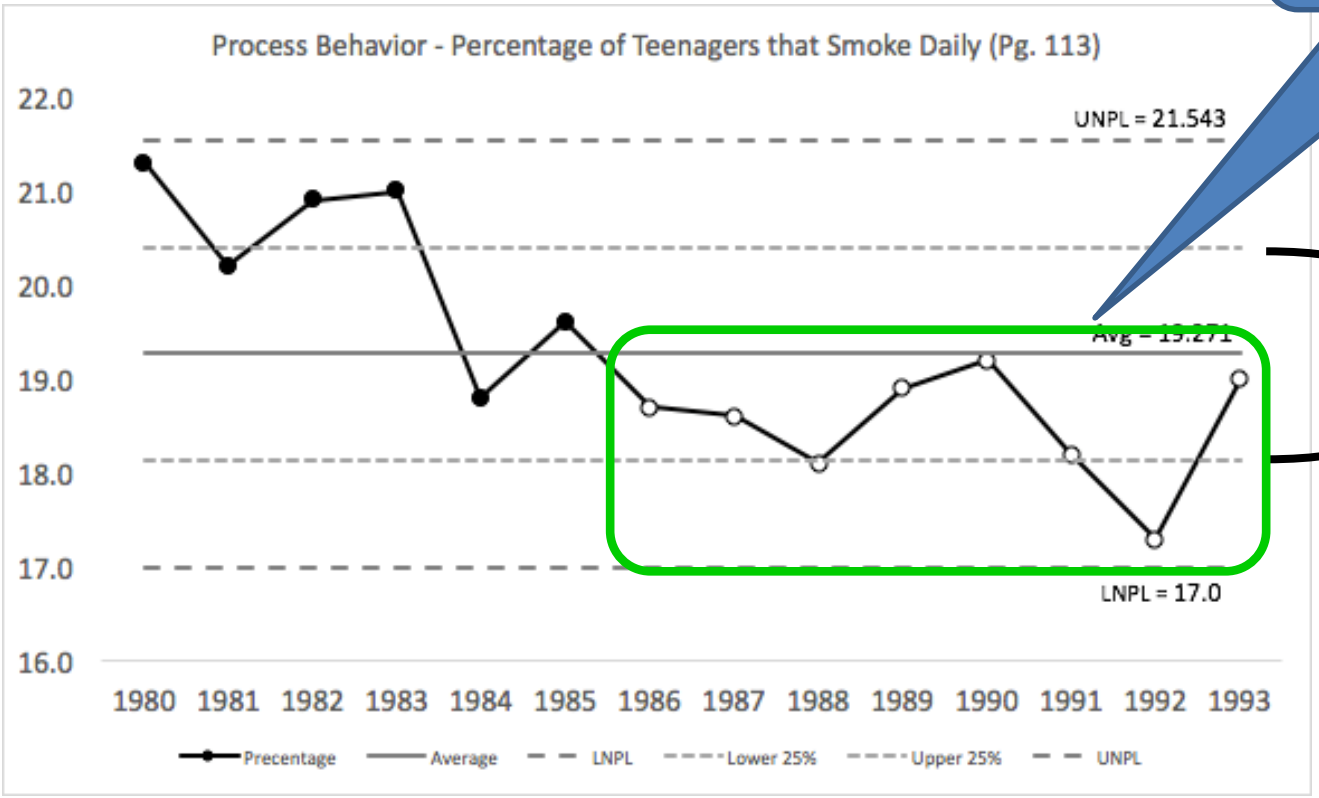


Source – "Making Sense of Data", Page 102.

The Rules of Trend Analysis

- Rule # 2 – Runs About the Central Line:
  - Eight successive values on the same side of the central line is an indication of a *weak sustained effect*.
  - Might want to investigate.

Trend Rule #2



What's changed?
PSAs?
After school specials?

50+% of values

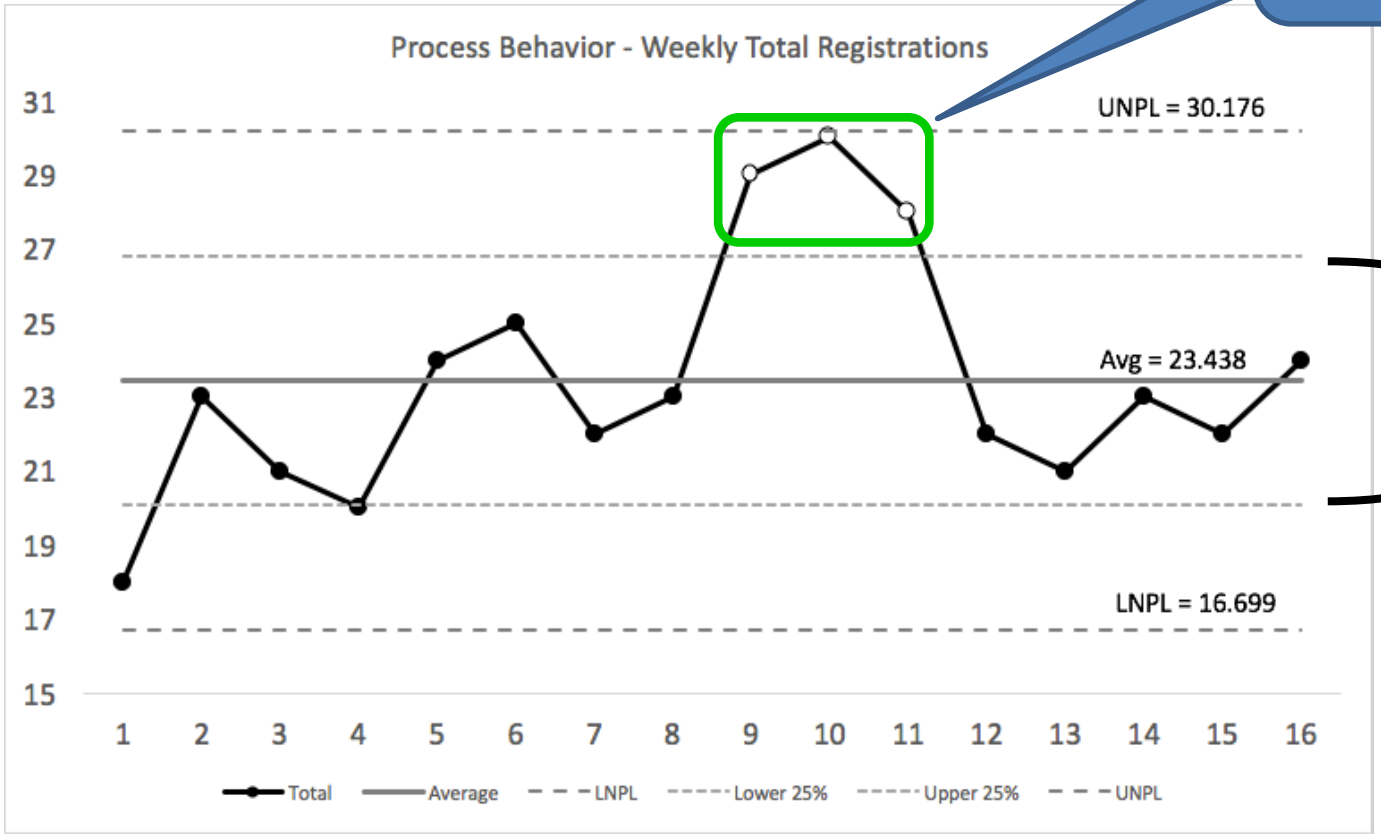
Source – "Making Sense of Data", Page 113.

datasciencedojo
data science for everyone

The Rules of Trend Analysis

- Rule # 3 – Runs Near the Limits:
  - Three out of four successive values within the upper 25% of the region between the limits or within the lower 25% region between the limits.
  - Could be indicative of a *moderate sustained effect*.

Trend Rule #3



Why the spike?
Marketing campaign?
Good economy?

50+% of values

datasciencedojo
data science for everyone

COMPARING GROUPS

Example Questions & Scenarios

- Questions:
  - "Are things different between these two groups?"
  - "Is the West division doing better than the East division?"
  - "Is org A doing worse than org B?"
- Scenarios:
  - Comparing on the job accident rates
  - Comparing employee attrition.
  - Comparing sales, expenses, profit, etc.

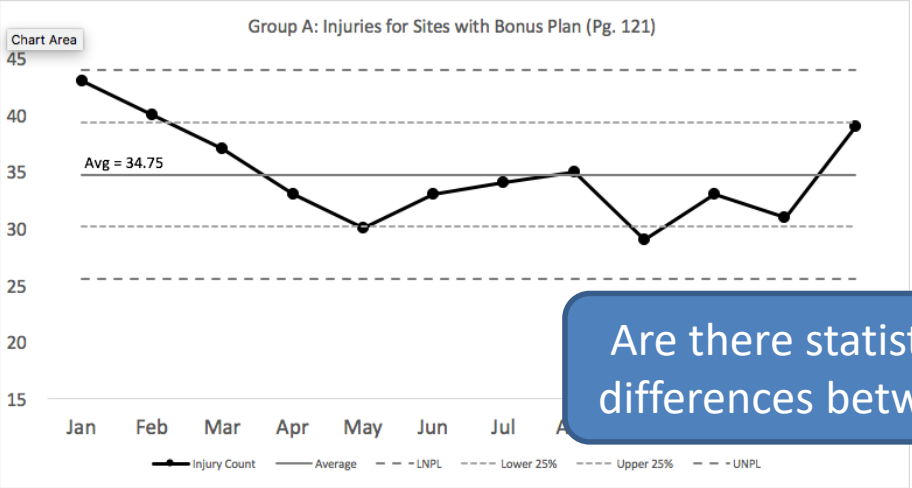
Let's use this as a hypothetical scenario.

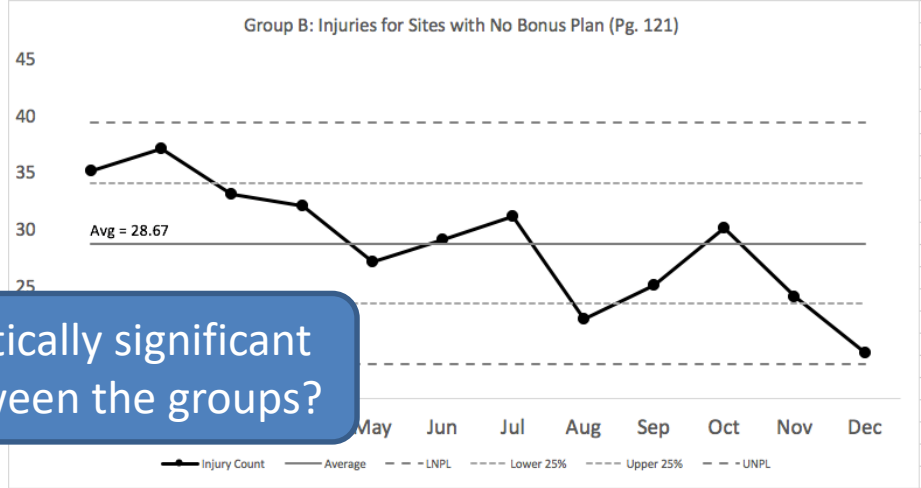
Hypothetical Scenario

- We run a trucking company and want boost productivity with a bonus plan.
- We conduct a year-long trial with select sites having the bonus.
- We are interested in the possible effects of the bonus on injury rates.

Comparing Groups

Average injury rate for bonus sites is 34.75. Average injury rate for no bonus sites is 28.67. Is the difference real?



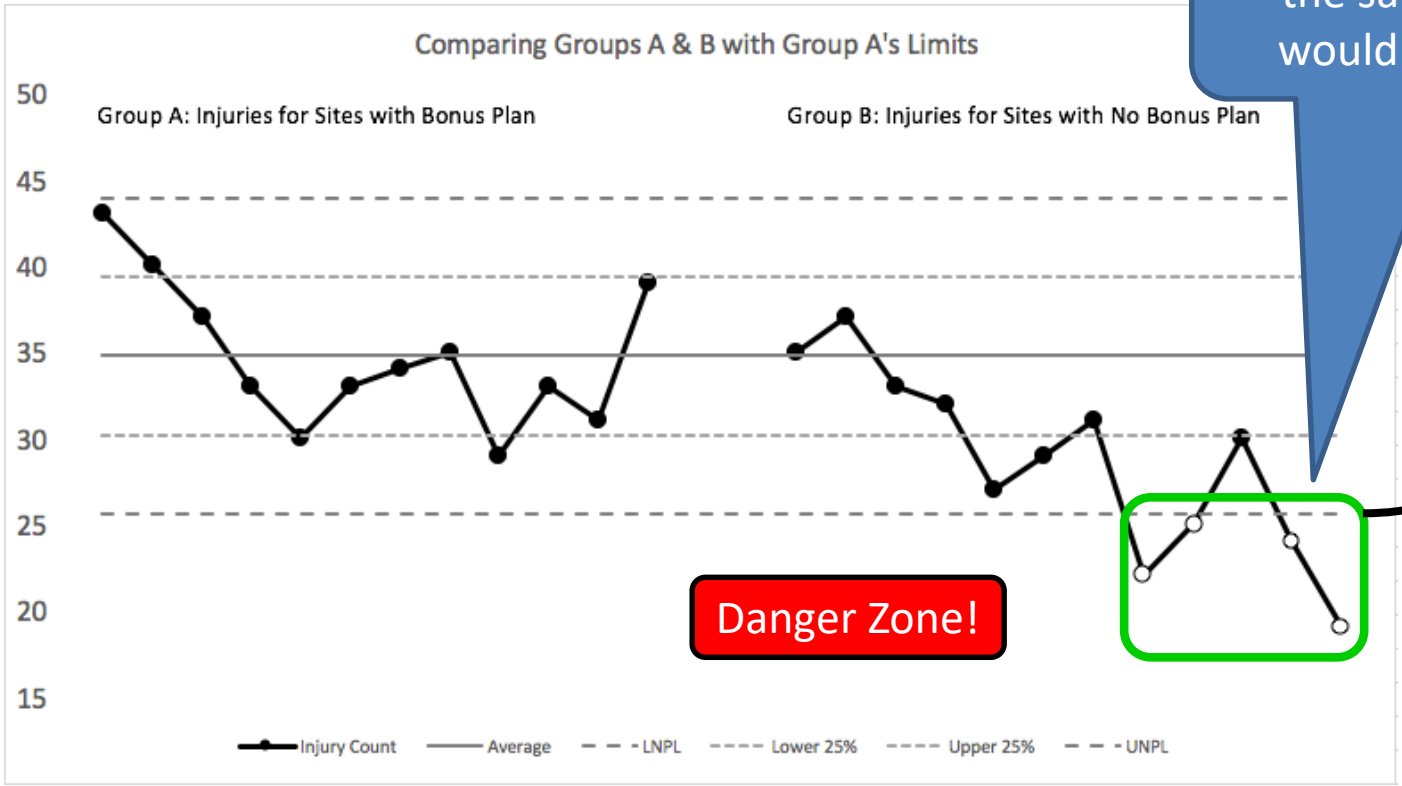


Are there statistically significant differences between the groups?

Source – "Making Sense of Data", Page 121.

datasciencedojo
data science for everyone

Comparing Groups



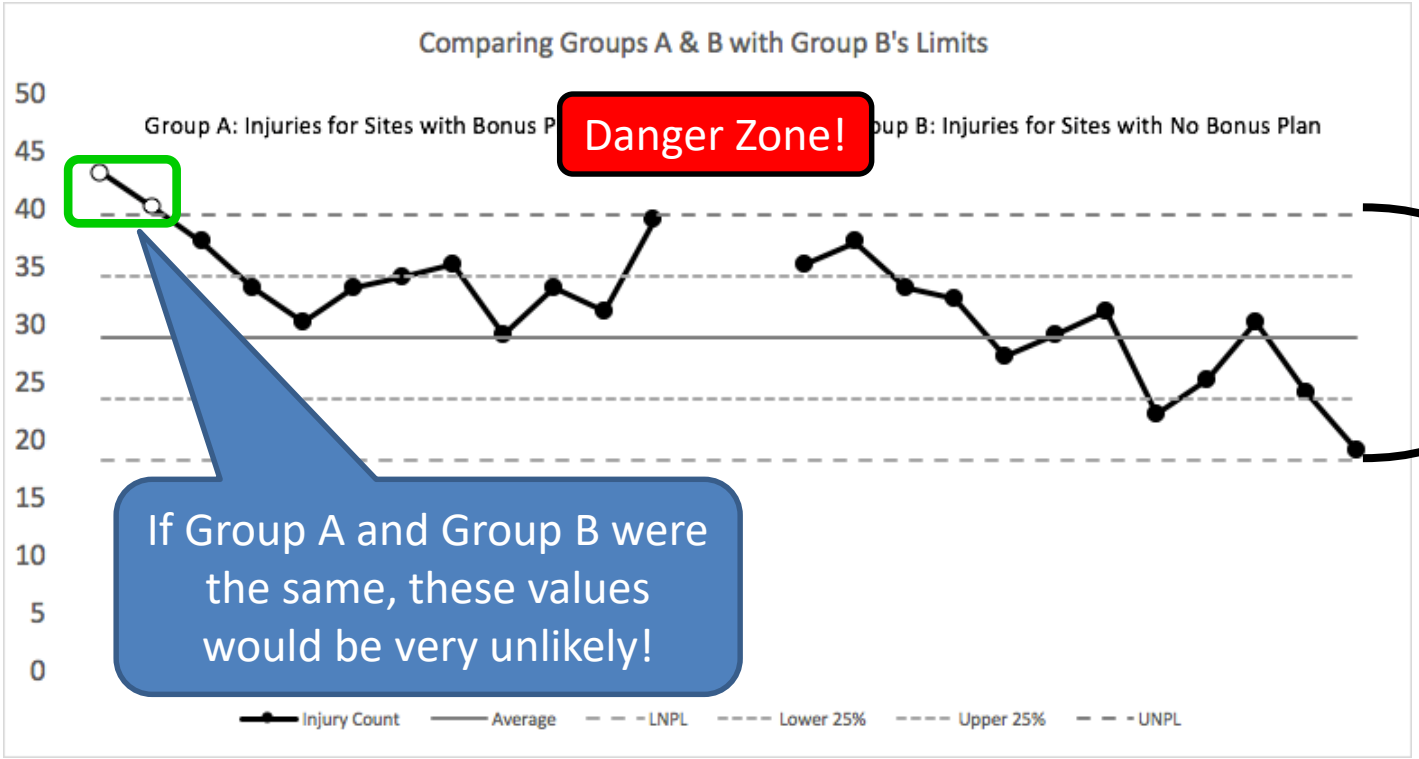
If Group A and Group B were the same, these values would be very unlikely!

96+% of all values!

Source – "Making Sense of Data", Page 121.

datasciencedojo
data science for everyone

Comparing Groups



96+% of all values!

Source – "Making Sense of Data", Page 121.

datasciencedojo
data science for everyone

Results

- By using process behavior chart limits, we can detect statistically significant differences between groups.
- While only one “Danger Zone” situation was sufficient to determine the groups are different, we saw two “Danger Zone” situations.
- We are safe in interpreting the differences in average injury rates are real – sites with the bonus average more injuries!

Summary

- Business data presents unique analytical challenges.
- However, histograms, running records, and process behavior charts work well with business data.
- Using process behavior charts we can apply rigorous analytics to our business data – including detecting statistically significant differences between groups!

QUESTIONS

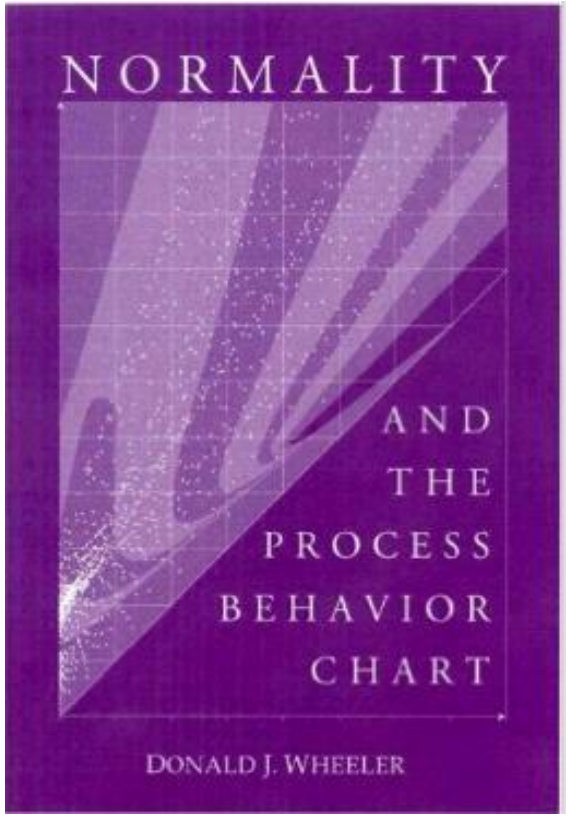
APPENDIX

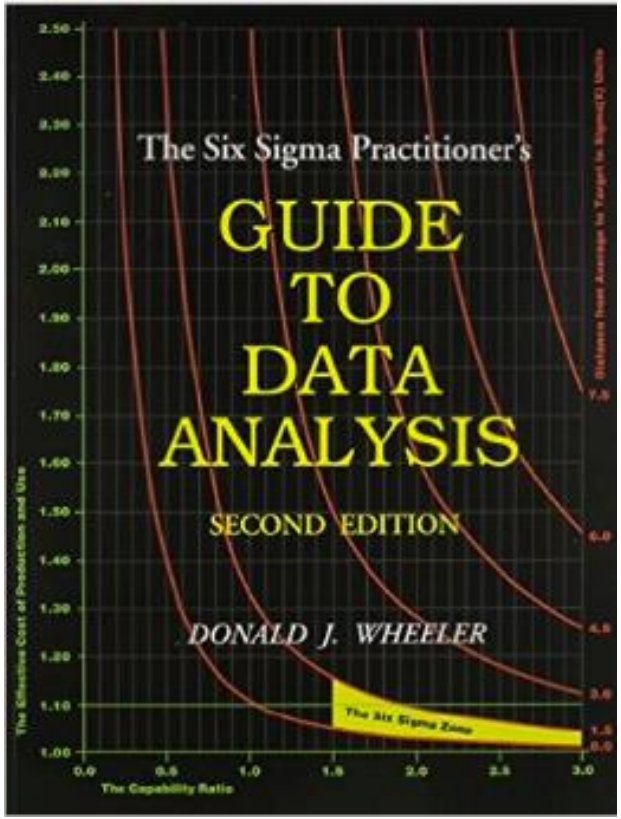
Get the Files!

- GitHub Repo:
  - https://github.com/datasciencedojo/meetup/tree/master/business_data_analysis_with_excel

- Direct links to files:
  - Excel
    - https://github.com/datasciencedojo/meetup/blob/master/business_data_analysis_with_excel/BusinessDataAnalysis.xlsx
  - PDF:
    - https://github.com/datasciencedojo/meetup/blob/master/business_data_analysis_with_excel/BusinessDataAnalysis.pdf

Want more goodness?





These books provide more in-depth mathematical details.

datasciencedojo
data science for everyone

Excel Resources

- Histograms:
  - Windows Excel 2010 and Mac:
    - https://www.youtube.com/watch?v=ujqgyrDUX1o
  - Windows Excel 2016:
    - https://www.youtube.com/watch?v=53DOu_vstvI
- Running Records:
  - Windows and Mac:
    - https://www.youtube.com/watch?v=mTnsxNfTFKo