

# Basic Data Analysis Principles

## *What to do once you get the data*

"When we reason about quantitative evidence, certain methods for displaying and analyzing data are better than others. Superior methods are more likely to produce truthful, credible, and precise findings. The difference between an excellent analysis and a faulty one can sometimes have momentous consequences."

-Edward R. Tufte, "Visual and Statistical Thinking: Displays of Evidence for Making Decisions"

*Visual Explanations*, Edward R. Tufte, Graphics Press, 1997.



# Acknowledgments



- ICEAA is indebted to TASC, Inc., for the development and maintenance of the Cost Estimating Body of Knowledge (CEBoK®)
  - ICEAA is also indebted to Technomics, Inc., for the independent review and maintenance of CEBoK®
- ICEAA is also indebted to the following individuals who have made significant contributions to the development, review, and maintenance of CostPROF and CEBoK®
- Module 6 Basic Data Analysis Principles
  - Lead authors: Megan E. Dameron, Bethia L. Cullis, Maureen L. Tedford
  - Senior reviewers: Richard L. Coleman, Jessica R. Summerville, John S. Smuck, Fred K. Blackburn
  - Reviewers: Samuel B. Toas, Kevin Cincotta, Matthew J. Pitlyk, Brian A. Welsh
  - Managing editor: Peter J. Braxton



# Unit Index

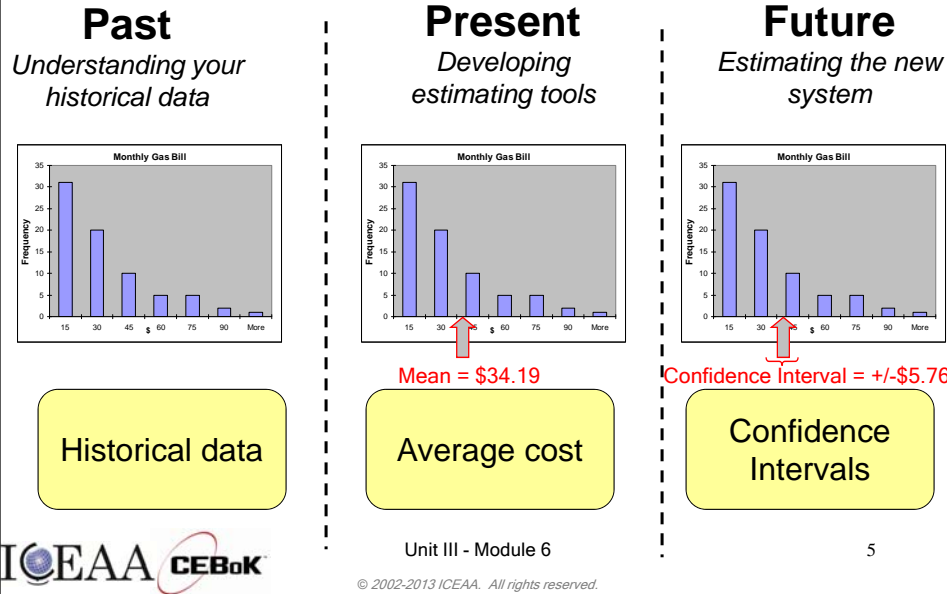
- Unit I - Cost Estimating
- Unit II - Cost Analysis Techniques
- Unit III - Analytical Methods
  - 6. Basic Data Analysis Principles**
  - 7. Learning Curve Analysis
  - 8. Regression Analysis
  - 9. Cost and Schedule Risk Analysis
  - 10. Probability and Statistics
- Unit IV - Specialized Costing
- Unit V - Management Applications

# Data Analysis Overview

- |  |   |
|--|---|
| <ul style="list-style-type: none"> <li>• Key Ideas               <ul style="list-style-type: none"> <li>- Visual Display of Information</li> <li>- Central Tendency of Data</li> <li>- Dispersion (Spread) of Data</li> <li>- Data accumulation</li> <li>- Outliers</li> </ul> </li> </ul>                                       | <ul style="list-style-type: none"> <li>• Practical Applications               <ul style="list-style-type: none"> <li>- Making sense of your data</li> </ul> </li> </ul>   |
| <ul style="list-style-type: none"> <li>• Analytical Constructs               <ul style="list-style-type: none"> <li>- Descriptive statistics                   <ul style="list-style-type: none"> <li>• Mean, median, mode</li> <li>• Variance, std deviation, CV</li> </ul> </li> <li>- Functional forms</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>• Related Topics               <ul style="list-style-type: none"> <li>- Parametrics </li> <li>- Distributions                   <ul style="list-style-type: none"> <li>• Normal, Chi, t, F </li> </ul> </li> <li>- Probability and Statistics</li> </ul> </li> </ul> |

# Data Analysis Within The Cost Estimating Framework

v1.2



# Data Analysis Outline

v1.2

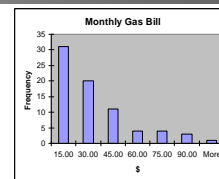
- Core Knowledge
  - Types of Data
  - Univariate Data Analysis
  - Scatter Plots
    - Variables
    - Axes and Function Types
  - Data Validation
    - Descriptive Statistics
    - Outliers
    - Rules of Thumbs
  - Two Cautionary Tales
- Summary
- Resources
- Related and Advanced Topics

## Types of Data

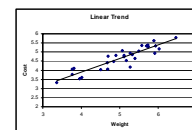
- Univariate
- Bivariate
- Multivariate
- Time Series

## Types of Data

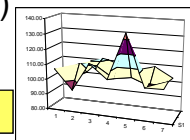
- Univariate
  - Single variable
  - Use descriptive and inferential statistics



- Bivariate
  - One independent variable and one dependent variable (i.e.,  $y$  is a function of  $x$ )
  - Use descriptive and inferential statistics



- Multivariate
  - Several independent variables and one dependent variable (i.e.,  $y$  is a function of  $x_1$ ,  $x_2$ , and  $x_3$ )
  - Use descriptive and inferential statistics



**Tip:** Univariate data plus a Nominal variable is really bivariate

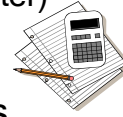
# Types of Data - Time Series

v1.2

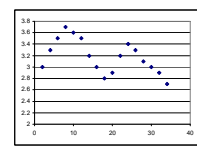


- Time as the independent variable
  - Interval matters! Make sure you use an XY (Scatter) and not a Line Chart in Excel unless intervals are equally spaced
- Smooth trends are rarely found in time series
  - Possible *rare* exceptions (e.g., corrosion over time)
  - “Standard” trends such as investment and inflation
- Look for paradigm shifts, cycles, autocorrelation
  - Use moving averages, divide data into groups and compare descriptive statistics
- Regression is often not useful as it only picks up smooth trends unless AR1/ARIMA
  - ANOVA and mean comparisons are more useful

11



2



# Univariate Data Analysis

v1.2

- Visual Display of Information
  - Histogram, stem-and-leaf, box plot
- Measures of Central Tendency
  - Mean (or median or mode)
- Measures of Variability
  - Standard deviation (or variance), coefficient of variation (CV)
- Measures of Uncertainty
  - Confidence Interval (CI)
- Statistical Tests
  - t test, chi square test, Kolmogorov-Smirnov (K-S) test

What does it look like?

What's your best guess?

How much remains unexplained?

How precise are you?

How can you be sure?

10

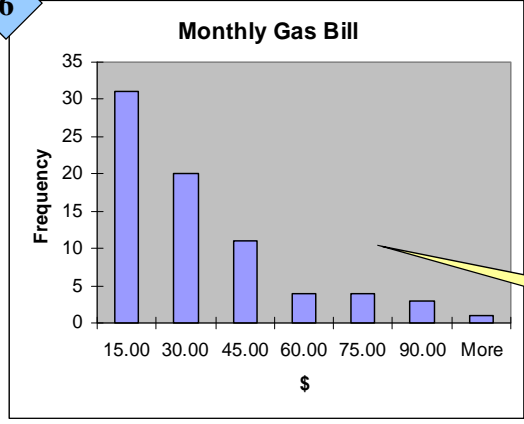
**Tip:** This analysis framework is mirrored in bivariate and multivariate analysis.

8

# 10 Visual Display - Histograms v1.2

- Histograms should be used to give an idea of the distribution of the data

6



**Warning:** Results of macros do not update if your data change!

Excel Data Analysis Add-In - Histogram.



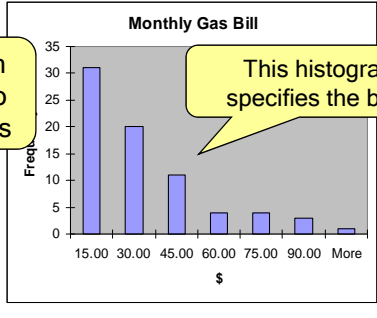
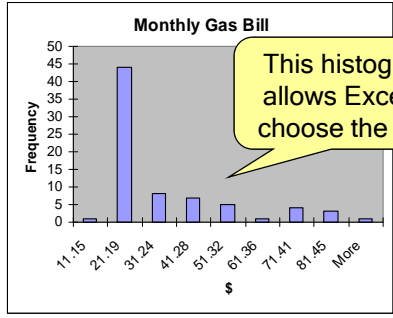
**Tip:** Create histogram manually using Chart type Column so that results *do* update when data change!

Skew-right distribution, possibly Exponential, Triangular, or Lognormal

10

# Histograms - Bins v1.2

- It is important to carefully consider the number of bins used in a histogram
  - Experiment with intervals to be sure you understand the data



6


5


**Warning:** Default bins in Excel histograms may not be optimal!

Which is clearer? Which sets a trap?



**Warning:** Histograms can be manipulated!


## Central Tendency - Mean

-  • The sample mean of the data set

  $\{x_1, x_2, \dots, x_n\}$  is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{(x_1 + x_2 + \dots + x_n)}{n}$$


-  - In Excel, use the “AVERAGE( )” function


- Means of example data sets:
  - Gas bill (74 months), \$26.52
  - Therms used (74 months), 14.8



The mean is the Expected Value of a random variable

## Central Tendency - Median

-  • The sample median is the “middle” data point, with 50% of the remaining observations falling under that point, and 50% above

 - If a data set has an odd number of points, the middle value is the median

- The median of the data set  $\{2, 5, 7, 9, 25\}$  is 7  AKA 50<sup>th</sup> Percentile

- If a data set has an even number of points, the two middle values are averaged

- The median of the set  $\{3, 6, 8, 11, 13, 30\}$  is 9.5 (average of 8 and 11)

- In general, the k<sup>th</sup> percentile is the point with k% of the data below and (100-k)% of the data above

- Quartiles (25, 50, 75), deciles (10, 20, ..., 80, 90), icosatiles (5, 10, 15, ..., 95)

- When there are extreme data points, the median may be more representative than the mean because robust outliers impact the mean more than the median

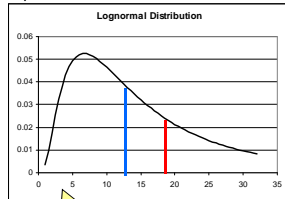
- “Representative” is a descriptive term, not a mathematical term
- There are many mathematical reasons to prefer mean over median

# Mean, Median, and Skew

v1.2

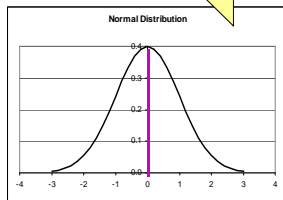
- The mean and the median are equal if the distribution is symmetric 15
- Unequal means and medians are an indication of skewness 19

**10**

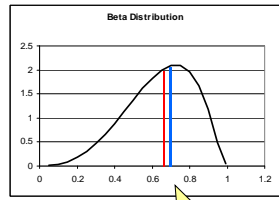


**Median < Mean**  
Skew(ed) Right

**Median = Mean**  
Symmetric



<http://en.wikipedia.org/wiki/Skewness>

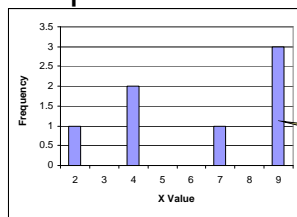


**Median > Mean**  
Skew(ed) Left

# Central Tendency - Mode

v1.2

- The sample mode is the most frequent point to occur in a data set 16
- The mode of a distribution is its peak
  - Value with the greatest probability mass (or density)
- The mode of the set {2,4,4,7,9,9,9} is 9
- The mode is a descriptive metric answering the question “what happens most frequently?”



- It can help give a visual idea of what the distribution looks like
- Most useful in discrete data

A histogram shows that the value 9 occurs most often ... this is the mode



v1.2

## Variability - Variance / Standard Deviation

- The sample variance measures the deviation of the data points from their mean

“easy to remember”

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

“easy to calculate”

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}{n-1}$$

**Tip:** Low variance indicates less dispersion, i.e., tighter data

- In Excel, use the “VAR( )” function

- The sample standard deviation is simply


10

$$s = \sqrt{s^2}$$

**Tip:** s is the estimator for the population parameter  $\sigma$

The standard deviation is expressed in the same units as the original data

- In Excel, use the “STDEV( )” function



Unit III - Module 6

17

© 2002-2013 ICEAA. All rights reserved.

v1.2

## Variability - Coefficient of Variation

- The Coefficient of Variation (CV) expresses the standard deviation as a percent of the mean

10


13

$$CV = \frac{s}{X}$$

**Tip:** Low CV indicates less dispersion, i.e., tighter data. 15% or less is desired

- Large CVs indicate that the mean is a poor estimator
  - Consider regression on cost drivers
  - Examine data for multiple populations (outliers)
- CVs of example data sets:
  - Gas bill, 74.4% (69.2%)
  - Therms used, 104.2% (102.5%)

Note that sums and averages tend to have smaller variances



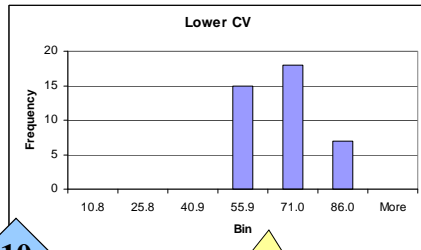
Unit III - Module 6

18

© 2002-2013 ICEAA. All rights reserved.

# Dispersion and CV

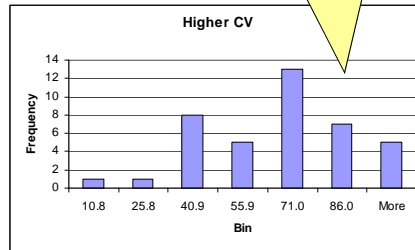
- These two data sets have the same mean, but different standard deviations



10

This data has a lower CV (17%) and is more tightly distributed

This data has a higher CV (38%) and has more dispersion



# Confidence Interval Illustration

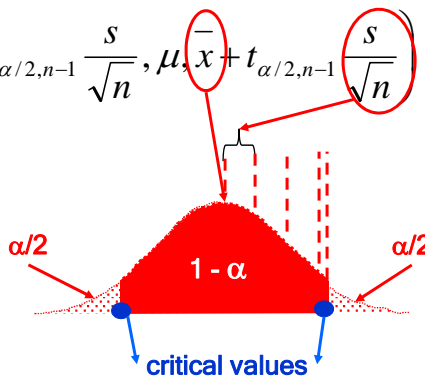
- A confidence interval (CI) suggests to us that we are  $(1-\alpha)*100\%$  confident that the true parameter value is contained within the calculated range\*

8

10

$$\left( \bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \mu, \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right)$$

\* Note this statement provides a general sense of what a confidence interval does for us in concise language, for ease of understanding. The specific statistical interpretation is that if many independent samples are taken where the levels of the predictor variable are the same as in the data set, and a  $(1-\alpha)$ -100% confidence interval is constructed for each sample, then  $(1-\alpha)$ -100% of the intervals will contain the true value of the parameter.



## Sample Sizes - Sufficiently Large n

- In general, we prefer n to be large ... how large is a function of our tolerance for error
  - The 68.3% CI for the mean is roughly  $CV/\sqrt{n}$
- So, for CVs ranging around 30%, we get the following 68.3% Confidence Interval with n:

4

n	+/-
4	15%
9	10%
16	8%
25	6%
36	5%

**Tip:** 30 is not a "magic number" of data points

10

- If we would like to be able to make judgments within about 5% points with a CV of 30%, we need  $n \approx 36$ 
  - We may have no choice but to deal with small n
  - In any case, we can calculate the range of estimated mean

8

## Prediction Intervals

- The previous confidence interval illustration gives the true *average* cost within a certain range
- If we want to know the *predicted cost of a new item* within a certain range, we need a prediction interval
- The PI suggests to us that we are  $(1-\alpha)*100\%$  confident that the next observation will be contained within the calculated range
- The larger standard error in the PI accounts for both the uncertainty in the mean (captured by the CI) and the uncertainty in individual observations

$$\left( \bar{x} - t_{\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}}, y_{n+1}, \bar{x} + t_{\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}} \right)$$

# Statistical Tests

- t test for mean
  - 10 - Is the Cost Growth Factor (CGF) for NAVAIR programs different than 1.0?
- Chi square test for variance
  - Is 30% a reasonable CV to use for this variable? Should t test for equal means assume equal variances?
- Chi square test for distribution
  - 17 - Are Line-Replaceable Unit (LRU) failures uniform across all deployed units?
- Kolmogorov-Smirnov test for distribution
  - Is the normal distribution appropriate for modeling uncertainty in design weight?

# Scatter Plots

- Variables
- Axes
- Function Types

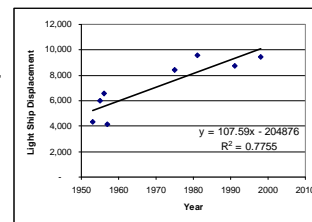
# Scatter Plots

- A picture *is* worth a thousand words!
- 📖 - A scatter plot can reveal a wealth of information about relationships present in the data
- Create scatter plots in Excel by using the Chart Wizard - XY (Scatter)
- Add a trend line in Excel by right clicking the plotted data and choosing Add Trend line
  - Helps link graph and equation
  - Look at inferential statistics later



8

**Tip:** Scatter plots are the single most useful tool in all of analysis ... they are “the gift of sight” to the analyst

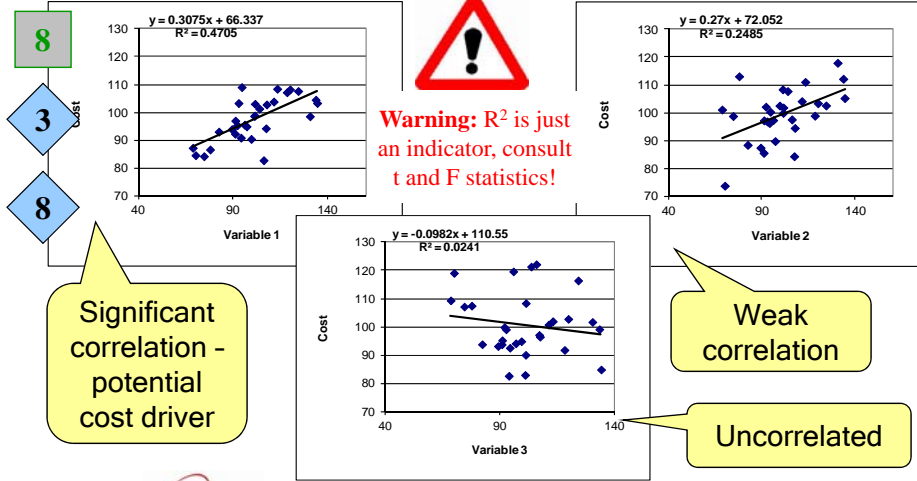


# Scatter Plots - Variables

- Plot cost (or other variable of interest, e.g., hours) as the dependent variable
- Look at a variety of different independent variables
  - 18 - Technical parameters such as weight, lines of code, etc.
  - Performance parameters such as speed, accuracy, etc.
  - Operational parameters such as crew size, flying hours, etc.
  - 8 - Cost of another element
- Think about which variables you *believe* should drive cost and collect that data!

# Scatter Plots - Cost Drivers

- Scatter plots can help identify cost drivers
- R<sup>2</sup> interpretation: % of variation in y explained (linearly) by variation in x




# Scatter Plots - Unit Space

- Data should first be plotted in unit space\*
- x is plotted on the horizontal axis (x-axis) and
- y is plotted on the vertical axis (y-axis)
- If the data have a non-linear relationship when plotted in unit space, investigate how the data can be “made” linear
  - Non-linear relationships can often be transformed to appear linear through the use of natural logs
  - Transformed data can then be regressed linearly
  - Before the widespread use of computers, non-linear data was graphed on semi-log or log-log paper

8

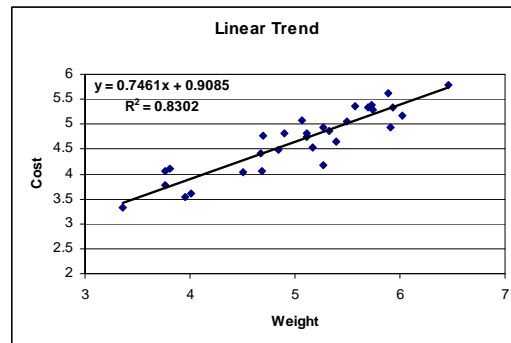
\* “Unit space” refers to the original, untransformed data.

## Scatter Plots - Linear Function

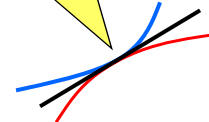
- The most common relationships are linear 
  - Of the form  $y = mx + b$  [ $m = \text{slope}$ ,  $b = \text{y-int.}$ ]
  - Plotted in unit space

11


8



**Tip:** Linear models are also the best approximations to non-linear models ... by which we mean, they take you "least far afield" if you guessed wrong.

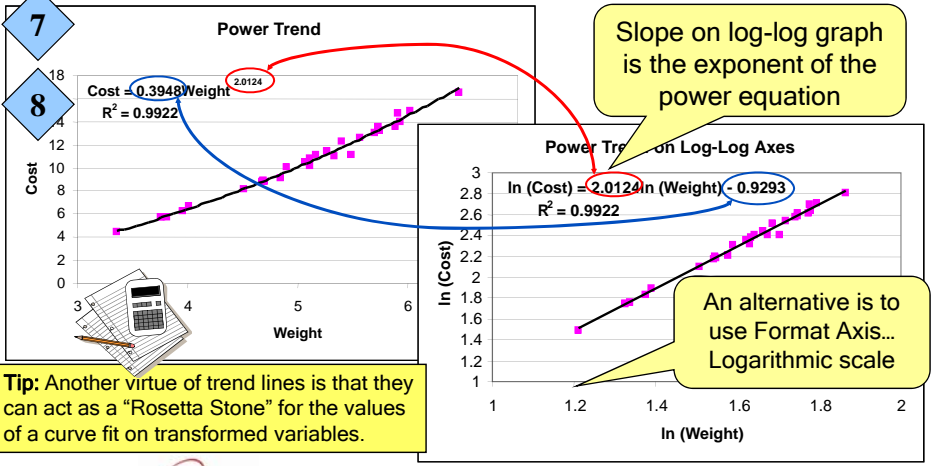


## Scatter Plots - Power Function

- Power functions are of the form  $y = ax^b$  
  - Can be transformed into linear functions
- Taking the natural log of both sides gives
 
$$\ln(y) = \ln(a) + b \ln(x)$$
- Plot  $\ln(x)$  on the horizontal axis and  $\ln(y)$  on the vertical axis and look for a linear trend
- This transformation is shown graphically on the next slide

### 3 Scatter Plots - Power Function

12 This function is most commonly used for learning curves, but can also be used for CERs



### Scatter Plots - Exponential Function

- Exponential functions are of the form  $y = ae^{bx} = a(e^b)^x = ak^x$
- Models of this form can be transformed and made to be linear
- Taking the natural log (ln) of both sides gives  $\ln(y) = \ln(a) + bx$

The natural log (ln) is the inverse function of the exponential:

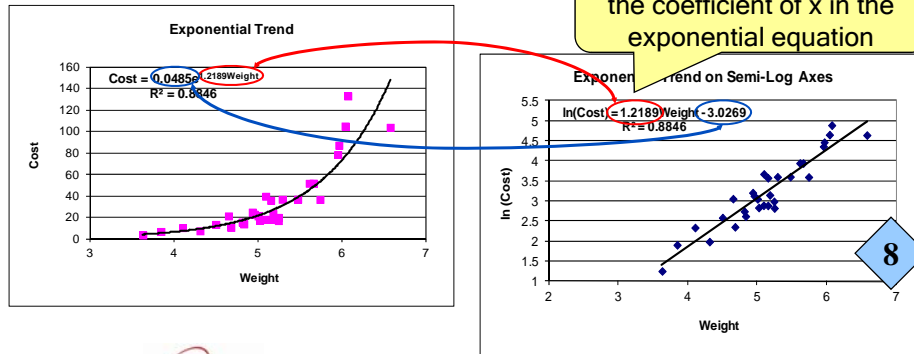
$y = e^x \iff x = \ln(y)$

Tip: Exponential functions are seldom encountered in cost estimation outside of inflation



# Scatter Plots - Exponential Function

- Then, x is plotted on the horizontal axis and ln(y) is plotted on the vertical axis
- This transformation is shown graphically below



# Scatter Plots - Constant Terms

- 2 • *Generalized* power and exponential equations are of the form:  $y = ax^b + c$ ,  $y = ae^{bx} + c$ 
  - Warning: Excel forces power and exponential trendlines to have  $c = 0$ !
- Power and exponential models usually assume a constant term of  $c = 0$
- However,  $c = 0$  is more common in theory than in practice
- 8 • If  $c = 0$  does not fit the data, consider using a model with  $c \neq 0$ 
  - Use the Excel Add-in Solver (or another, more robust optimization tool) to fit a curve to the data, where a, b, c are chosen simultaneously (GERM)
    - Minimize SSE or maximize unit-space  $R^2$

\*To b or Not to b\* The y-intercept in Cost Estimation, R. L. Coleman, J. R. Summerville, P. J. Braxton, B. L. Cullis, E. R. Druker, SCEA, 2007.

# Data Validation

- Scatter plotting gives you an idea of the relationships present in the data
- What's next?
  - Look at descriptive statistics
  - Look for outliers
  - Compare to historical studies, industry standards, or rules of thumb

# Descriptive Statistics

- Calculate descriptive statistics for each data group
  - Sample size
  - Raw mean
  - Standard deviation
  - Coefficient of variation (CV)
  - Weighted averages (e.g., dollar-weighted)
  - Moving averages (for time series data)
- In Excel, Tools - Data Analysis - Descriptive Statistics will easily calculate the most important descriptive statistics



**Warning:** Results of macros do not update if your data change!

**Tip:** Create formulae manually so that results *do* update when data change!

14

5

11

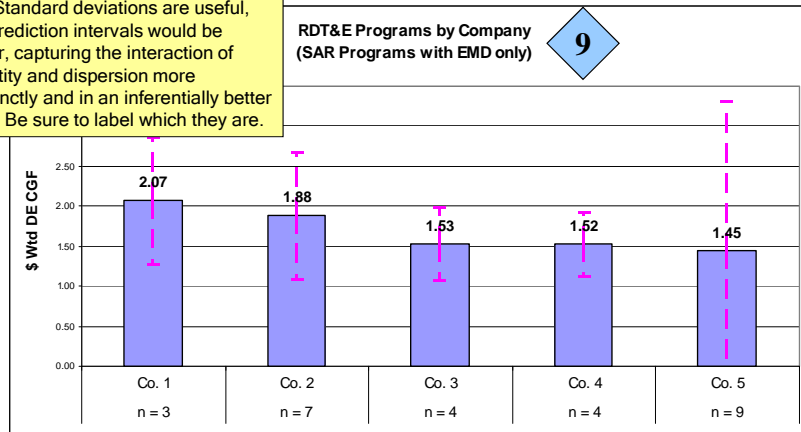


## Descriptive Statistics - Bar Charts

v1.2 

- Bar charts can be used to compare the descriptive statistics for different groups
- Y-error bars can be added to show the standard deviation

**Tip:** Standard deviations are useful, but prediction intervals would be better, capturing the interaction of quantity and dispersion more succinctly and in an inferentially better way. Be sure to label which they are.



## Bar Charts in Excel

v1.2

- Bar charts
  - Excel Chart Wizard - Column Chart
- Y-error bars
  - Format Data Series - Y-error bars (2003)
  - Chart Tools - Layout - Analysis - Error Bars (2007)
- Histogram
  - Excel Data Analysis Add-In - Histogram



**Tip:** It is recommended that you create your own dynamic histograms with flexible bin spacing using COUNTIF() and Column Charts.

# Outliers

- 📖 • Outliers are data points that fall far away from the center of the data *and* are not representative of the population you are trying to model
- 6 • For normally distributed data sets, about 95.45% of the data should fall within two standard deviations of the mean
- 10
  - So, we'd expect 4.55% to be outside two standard deviations
- 99.7% of the data should be within three standard deviations of the mean
  - If a data point is more than three standard deviations from the mean, it is a potential outlier

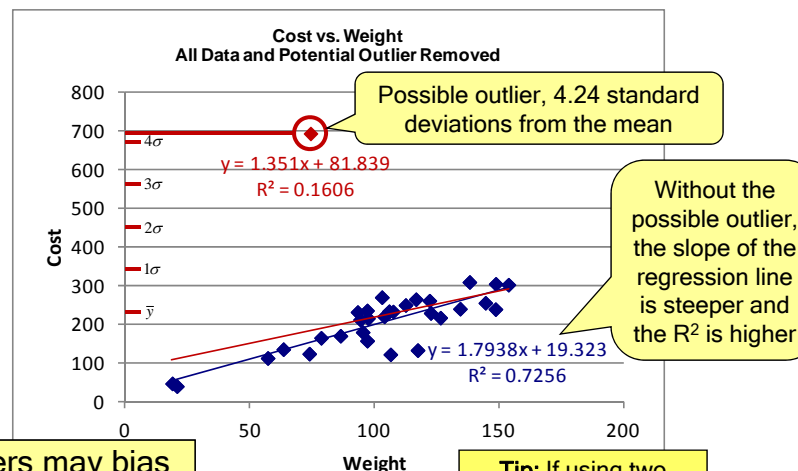
**Tip:** The normal distribution is a good first approximation, but if your data are significantly skewed, these rules of thumb should not be used to identify potential outliers.

# Outliers and Trend Lines

6

7

8



## Removing Outliers

- Do not remove an outlier from the data without a good reason!
  - Doing so removes some of the variation present in history
  - Doing so can be a form of “cooking the data”
- Good reasons for removing an outlier:
  - Program was restructured or divided
  - “One of these is not like the others”
    - e.g., a helo in a set of missile data
- Bad reasons for removing an outlier:
  - “Too high”
  - “2 standard deviations away from the mean” [!]

**Tip:** Outlier treatment separates the analysts from the spin meisters

4

## Rules of Thumb

- Compare your descriptive statistics to historical rules of thumb
  - NCCA Standard Factors handbook, for example
- Sanity check!

**Tip:** Comparison to history and cross checks separates the thorough from the sloppy

## Two Cautionary Tales

- “Expert’s Eyeball”
  - Descriptive Statistics and Visual Displays
- “Technical Hunch”
  - Outliers

## Engineering Judgments

- Suppose we are given an estimate that has “engineering judgment” as its basis

2

- Engineering judgments should never be accepted without validation!
- The analyst must find out if the “guess” is correct, or at least in the ballpark

14

- Experts often possess insight or intuition regarding systems that bears on cost, but it is the analyst’s job to make the estimate explicit and reproducible

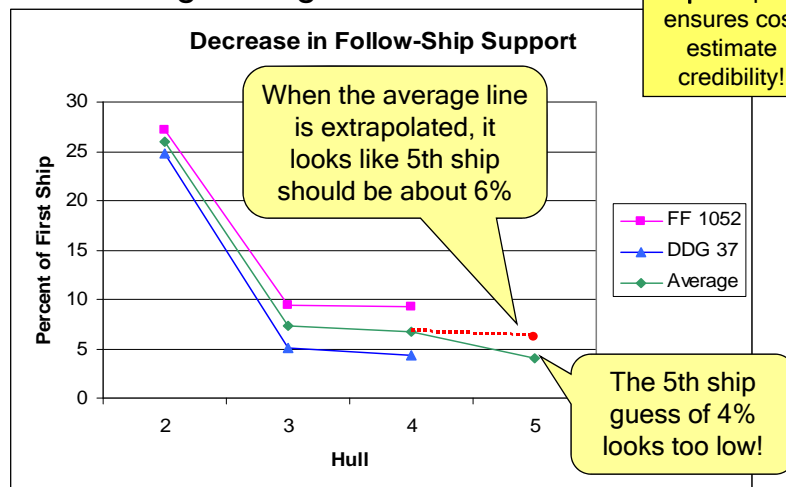
## Example: "Expert's Eyeball" v1.2

Follow Ship Support			
Hull	FF 1052	DDG 37	Average
2	27.1%	24.8%	25.95%
3	9.4%	5.1%	7.25%
4	9.2%	4.3%	6.75%
5	-	-	4.00%

Is the average a good idea?  
Is the 5<sup>th</sup> ship "guess" right?

## Example: "Expert's Eyeball" v1.2

- The average is a good number!



## Example: “*Technical Hunch*”

v1.2

- In this real-life example, we will look at the importance of correctly investigating outliers
- Scatter plots can be extremely useful in identifying potential outliers

## Example: “*Technical Hunch*”

v1.2

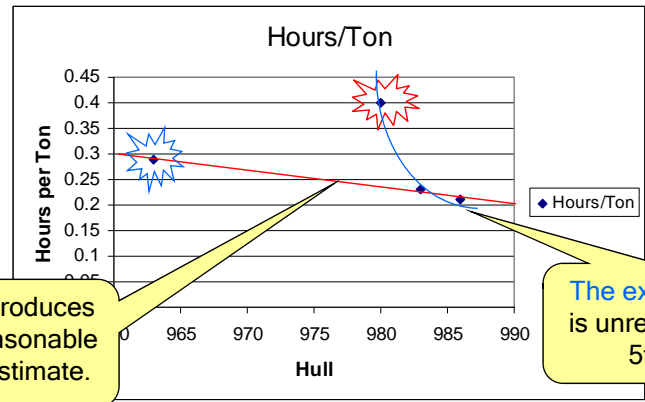
Shakedown	
Hull	Hours/Ton
DD 963	0.29
...	...
DD 980	0.40
...	...
DD 983	0.23
...	...
DD 986	0.21

“DD 963 is too low for a first ship”



# Wrong Outlier Rejected!

- Instead of DD 963, look into DD 980
  - That's the potential outlier!



This line produces a more reasonable 5th ship estimate.

The expert's curve is unrealistic at the 5th ship!

# Data Analysis Summary

- Steps of basic data analysis
  1. Scatter plot - visual depiction of the relationships in the data
  2. Descriptive statistics - calculate the means and CVs
    - If the CV is *under* 15%, the average may be a sufficient predictor, focus more attention on elements with higher CVs
    - If the CV is *over* 15%, focus on this element using regression analysis to look for a better predictor than the average (CER development)
  3. Look for outliers (data quality check)
  4. Compare to history

10

8