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.

© 2002-2013 ICEAA. All rights reserved.

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

© 2002-2013 ICEAA. All rights reserved.

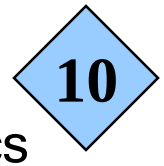
Unit Index

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

© 2002-2013 ICEAA. All rights reserved.

Data Analysis Overview

- Key Ideas
  - Visual Display of Information
  - Central Tendency of Data
  - Dispersion (Spread) of Data
  - Data accumulation
  - Outliers
- Practical Applications
  - Making sense of your data
- Analytical Constructs
  - Descriptive statistics
    - Mean, median, mode
    - Variance, std deviation, CV
  - Functional forms
- Related Topics
  - Parametrics
  - Distributions
    - Normal, Chi, t, F
  - Probability and Statistics

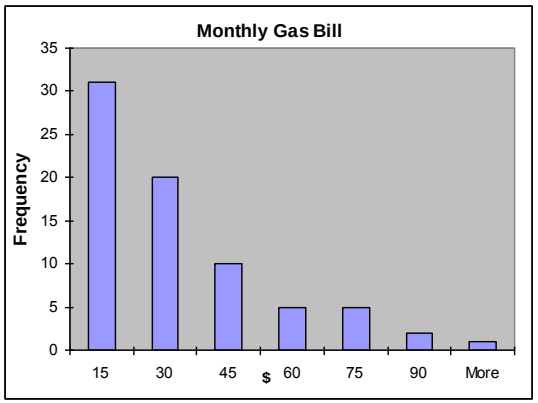


© 2002-2013 ICEAA. All rights reserved.

Data Analysis Within The Cost Estimating Framework

Past

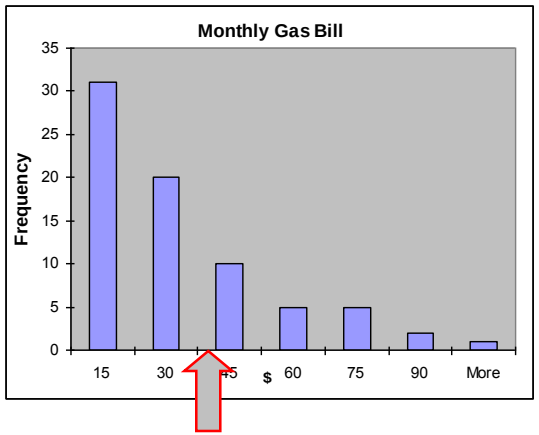
Understanding your historical data



Historical data

Present

Developing estimating tools

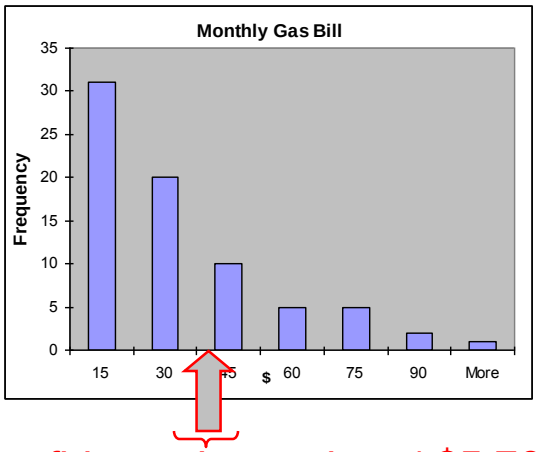


Mean = $34.19

Average cost

Future

Estimating the new system



Confidence Interval = +/-$5.76

Confidence Intervals

© 2002-2013 ICEAA. All rights reserved.

Data Analysis Outline

- 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

© 2002-2013 ICEAA. All rights reserved.

Types of Data

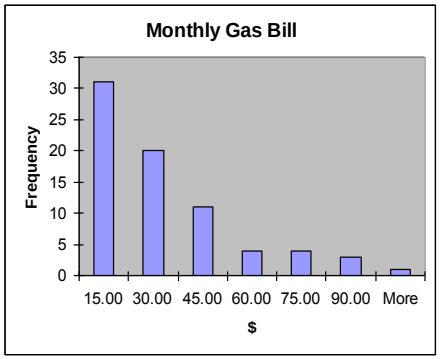
- Univariate
- Bivariate
- Multivariate
- Time Series

© 2002-2013 ICEAA. All rights reserved.

Types of Data

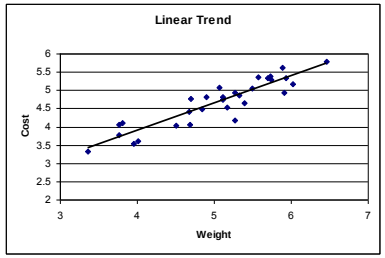
- **Univariate**
  - Single variable
  - Use descriptive and inferential statistics

10

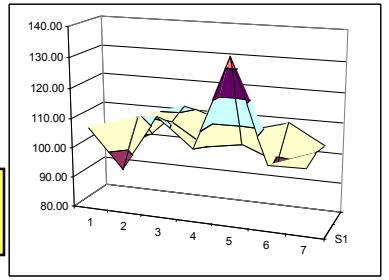


- **Bivariate**
  - One independent variable and one dependent variable (i.e., y is a function of x)
  - Use descriptive and inferential statistics

1



- **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

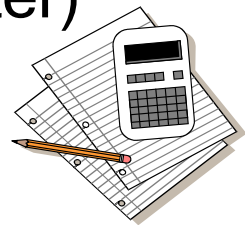


Tip: Univariate data plus a Nominal variable is really bivariate

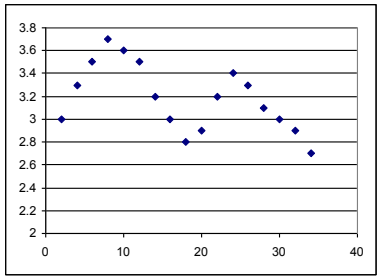
© 2002-2013 ICEAA. All rights reserved.

Types of Data - Time Series

- 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



© 2002-2013 ICEAA. All rights reserved.

Univariate Data Analysis

- 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 un-explained?

How precise are you?

How can you be sure?

10

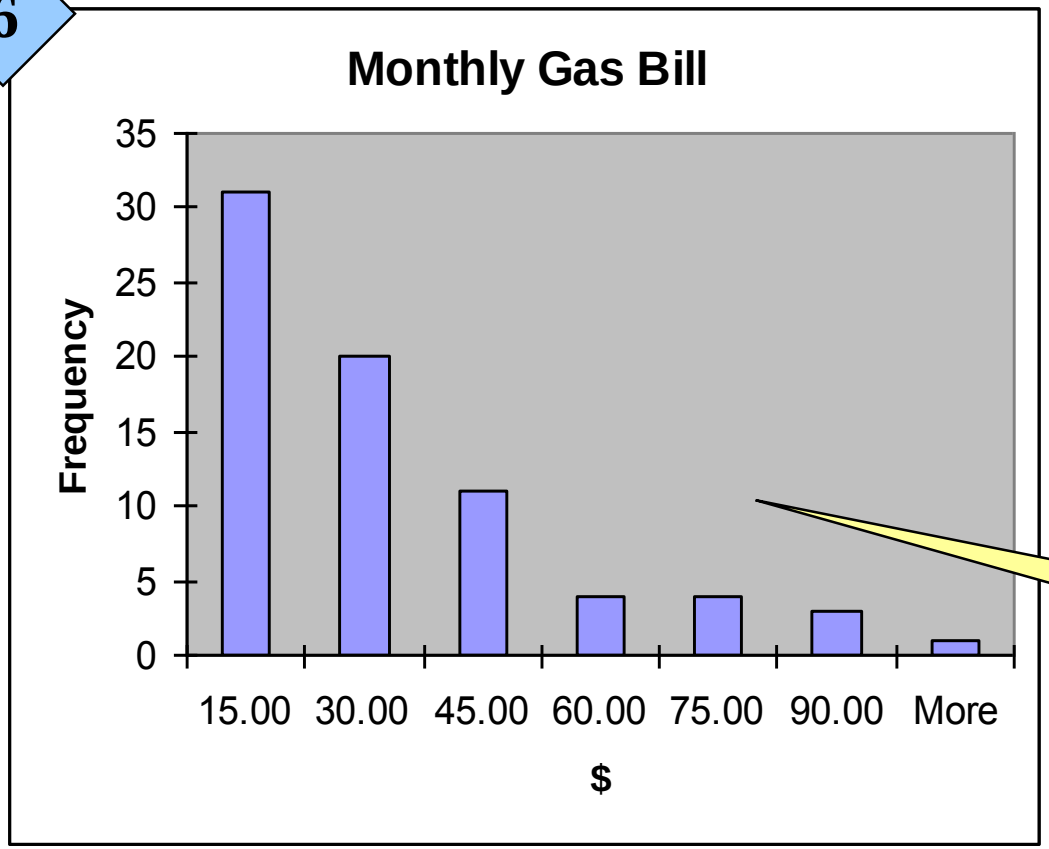
Tip: This analysis framework is mirrored in bivariate and multivariate analysis.

8

© 2002-2013 ICEAA. All rights reserved.

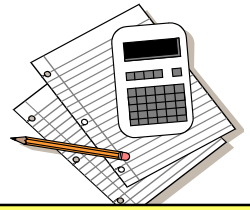
Visual Display - Histograms

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



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

Excel Data Analysis Add-In – Histogram.



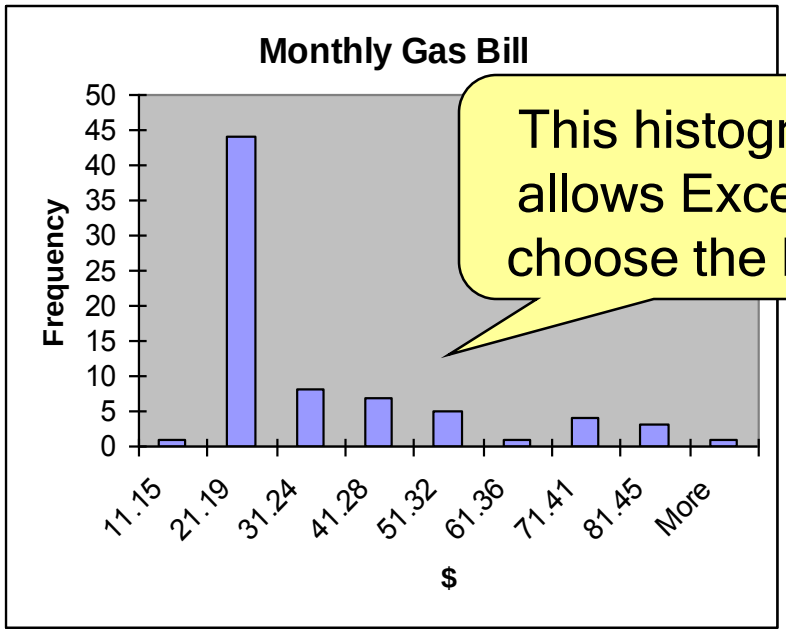
Tip: Create histogram manually using Chart type Column so that results *do* update when data change!

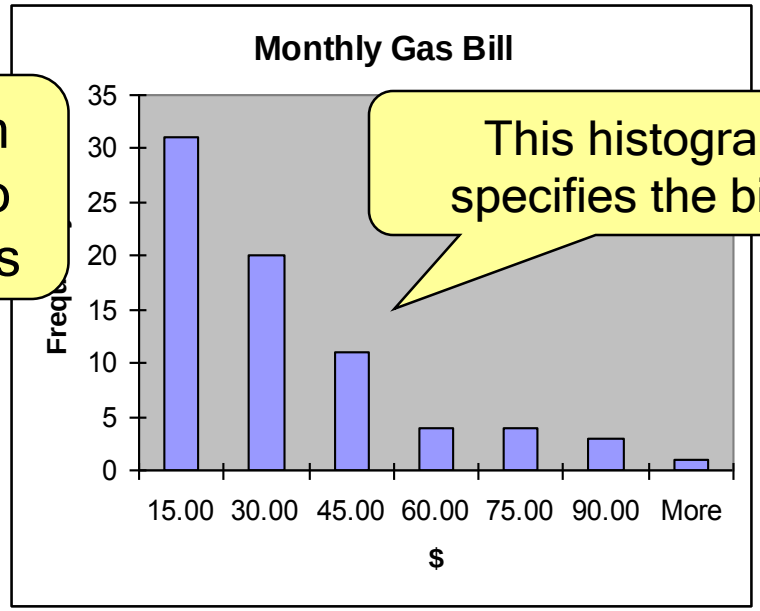
Skew-right distribution, possibly Exponential, Triangular, or Lognormal

© 2002-2013 ICEAA. All rights reserved.

Histograms - Bins

- 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!

© 2002-2013 ICEAA. All rights reserved.

Central Tendency - Mean

- The sample mean of the data set $\{x_1, x_2, \ldots, x_n\}$ is calculated as:

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

  - 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

© 2002-2013 ICEAA. All rights reserved.

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
  - 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^{th} 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

AKA 50th Percentile

© 2002-2013 ICEAA. All rights reserved.

Legend
Red = Extreme points
Blue = "Middle" points

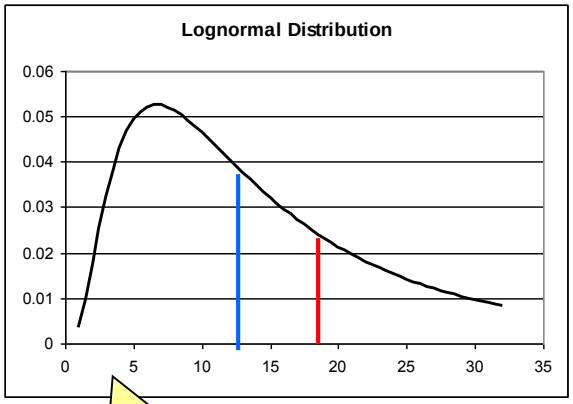
Mean, Median, and Skew

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

10

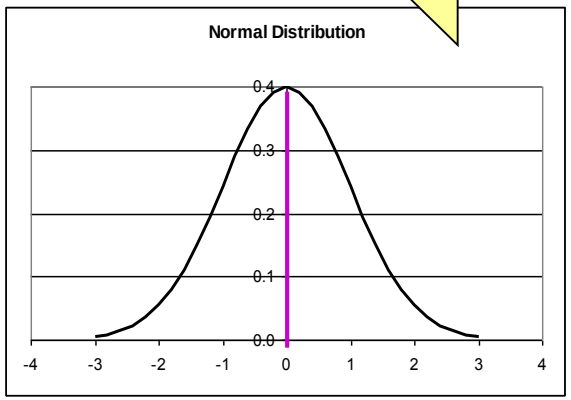
15

19

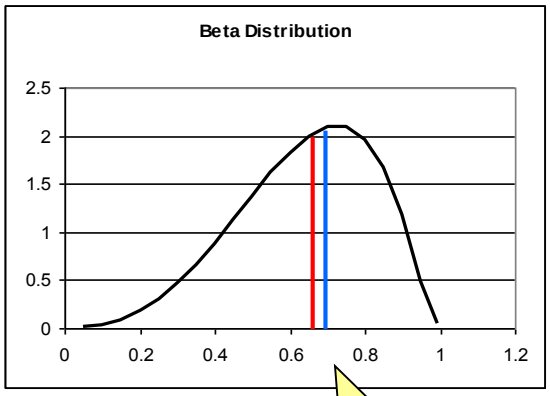


Median < Mean
Skew(ed) Right

Median = Mean
Symmetric



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



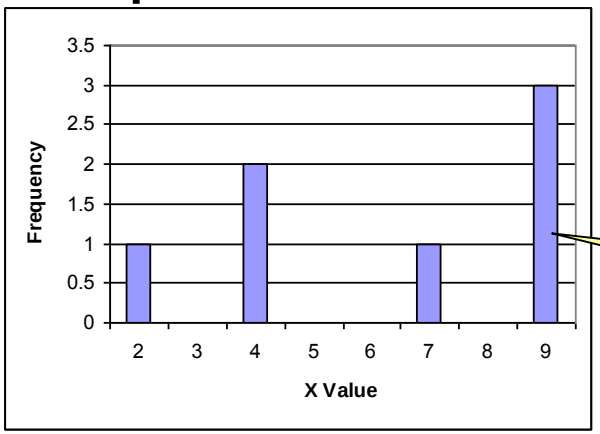
Median > Mean
Skew(ed) Left

© 2002-2013 ICEAA. All rights reserved.

Central Tendency - Mode

16

- The sample mode is the most frequent point to occur in a data set
  - 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

© 2002-2013 ICEAA. All rights reserved.

Variability - Variance / Standard Deviation

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

"easy to remember"

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

"easy to calculate"

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

  - In Excel, use the "VAR( )" function

- The sample standard deviation is simply

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

Tip: s is the estimator for the population parameter σ

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

  - In Excel, use the "STDEV( )" function

© 2002-2013 ICEAA. All rights reserved.

Variability - Coefficient of Variation

10

13

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

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

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)

8

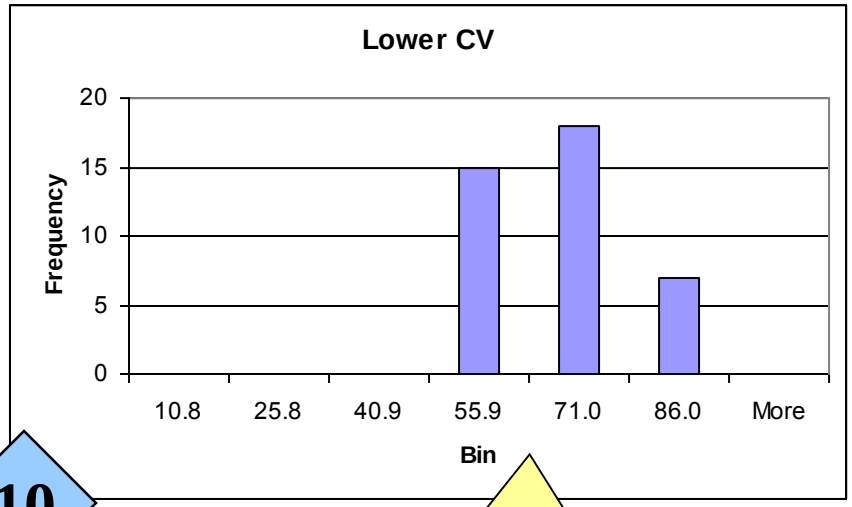
- CVs of example data sets:
  - Gas bill, 74.4% (69.2%)
  - Therms used, 104.2% (102.5%)

Note that sums and averages tend to have smaller variances

© 2002-2013 ICEAA. All rights reserved.

Dispersion and CV

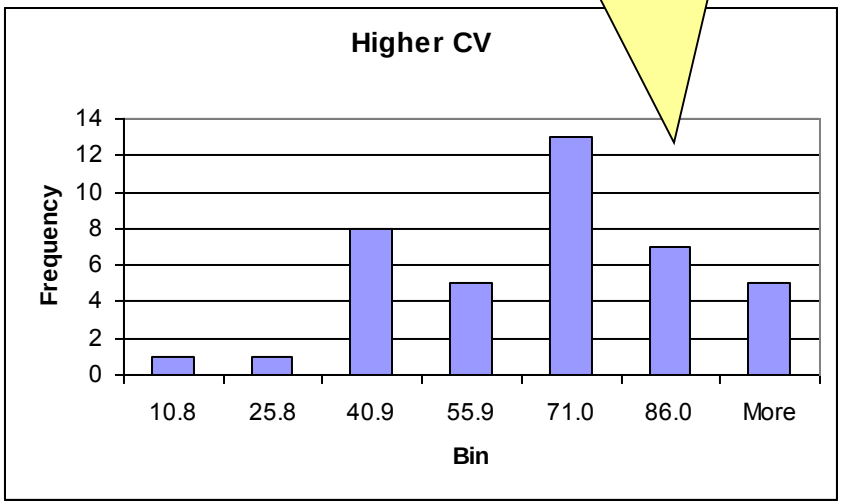
- These two data sets have the same mean, but different standard deviations



10

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

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



© 2002-2013 ICEAA. All rights reserved.

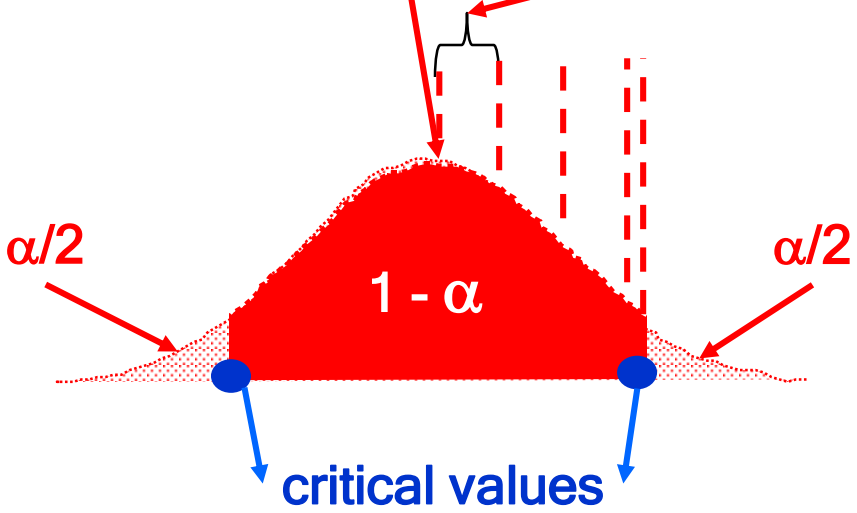
Confidence Interval Illustration

- A confidence interval (CI) suggests to us that we are (1-α)*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-α)·100% confidence interval is constructed for each sample, then (1-α)·100% of the intervals will contain the true value of the parameter.

© 2002-2013 ICEAA. All rights reserved.

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/√n
- So, for CVs ranging around 30%, we get the following 68.3% Confidence Interval with n:

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

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

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

© 2002-2013 ICEAA. All rights reserved.

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

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

© 2002-2013 ICEAA. All rights reserved.

Statistical Tests

- t test for mean
  - Is the Cost Growth Factor (CGF) for NAVAIR programs different than 1.0?

10

- 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
  - Are Line-Replaceable Unit (LRU) failures uniform across all deployed units?

17

- Kolmogorov-Smirnov test for distribution
  - Is the normal distribution appropriate for modeling uncertainty in design weight?

© 2002-2013 ICEAA. All rights reserved.

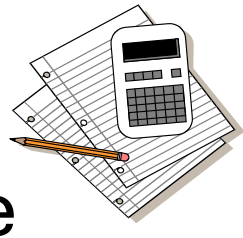
Scatter Plots

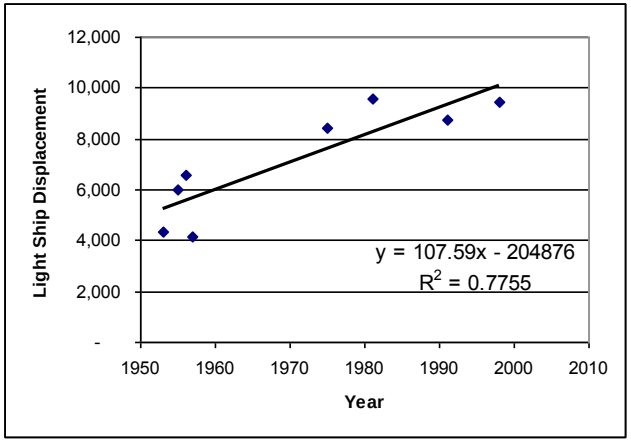
- Variables
- Axes
- Function Types

© 2002-2013 ICEAA. All rights reserved.

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





Tip: Scatter plots are the single most useful tool in all of analysis … they are "the gift of sight" to the analyst

© 2002-2013 ICEAA. All rights reserved.

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
  - 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.
  - Cost of another element
- Think about which variables you *believe* should drive cost and collect that data!

18

8

© 2002-2013 ICEAA. All rights reserved.

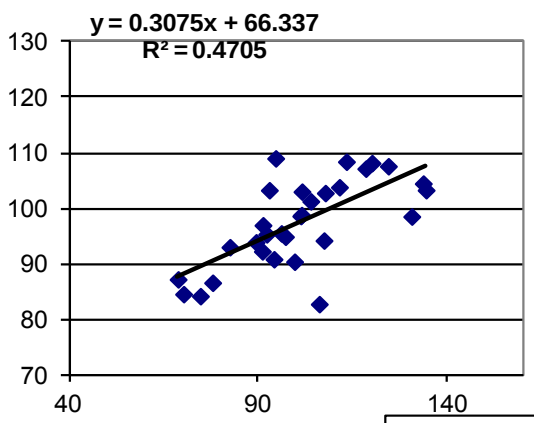
Scatter Plots – Cost Drivers

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

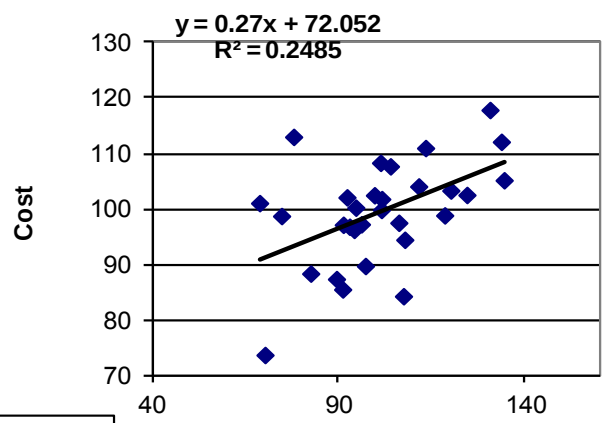
8

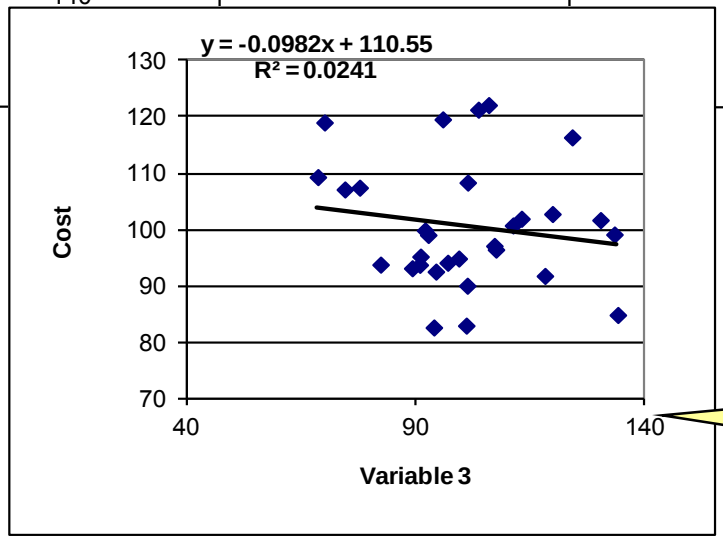
3

8



Warning: R^2 is just an indicator, consult t and F statistics!





Significant correlation – potential cost driver

Weak correlation

Uncorrelated

© 2002-2013 ICEAA. All rights reserved.

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.

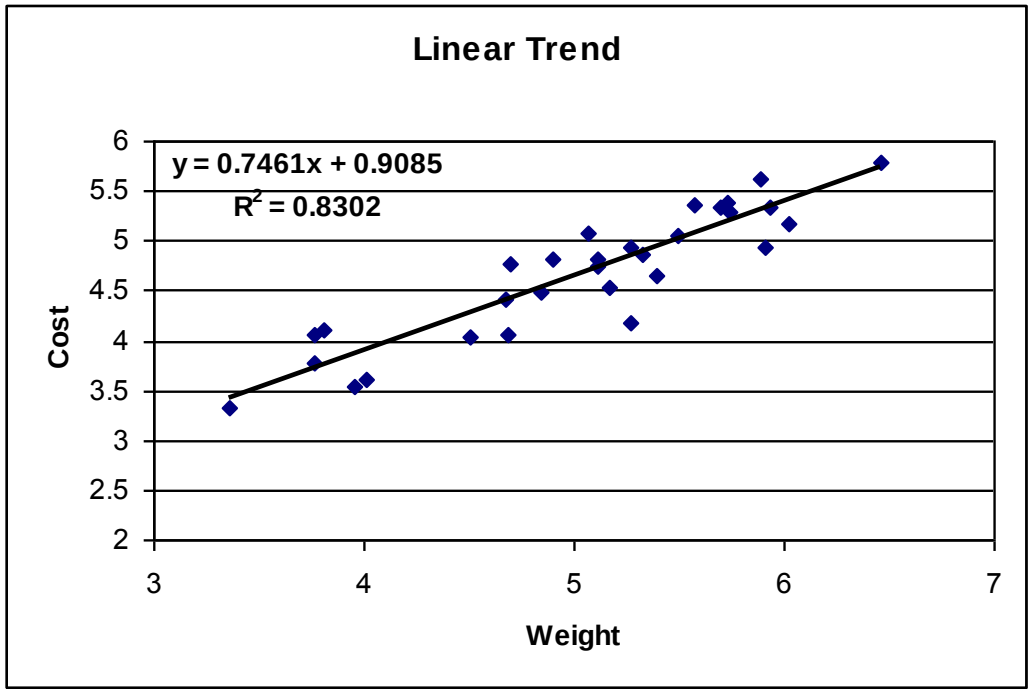
© 2002-2013 ICEAA. All rights reserved.

Scatter Plots – Linear Function

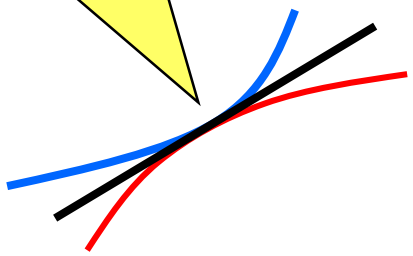
- The most common relationships are linear
  - Of the form y = mx + b [m = slope, b = y-int.]
  - Plotted in unit space

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.



© 2002-2013 ICEAA. All rights reserved.

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

$$\ln(y) = \ln(a) + b \ln(x)$$

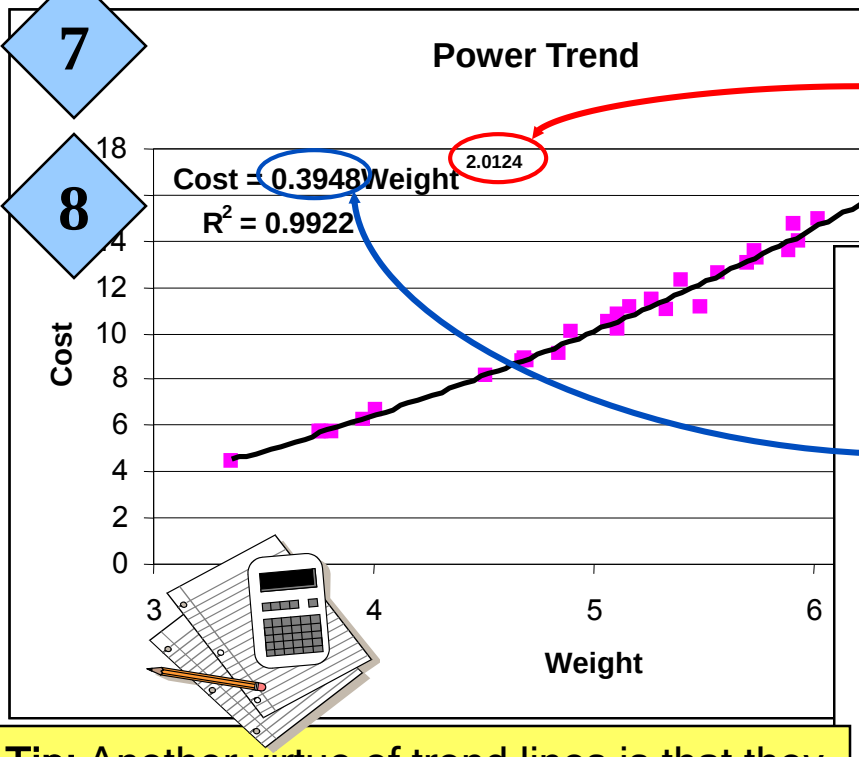
- Plot ln(x) on the horizontal axis and ln(y) on the vertical axis and look for a linear trend
- This transformation is shown graphically on the next slide

© 2002-2013 ICEAA. All rights reserved.

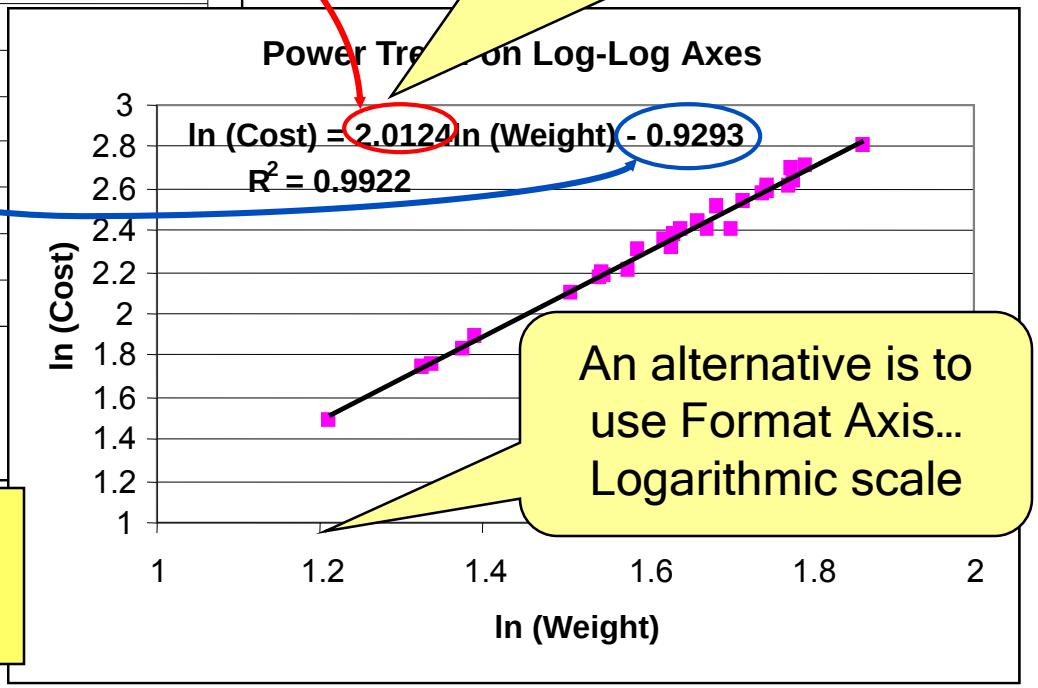
3 Scatter Plots – Power Function

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

7 8



Slope on log-log graph is the exponent of the power equation



An alternative is to use Format Axis… Logarithmic scale

Tip: Another virtue of trend lines is that they can act as a "Rosetta Stone" for the values of a curve fit on transformed variables.

© 2002-2013 ICEAA. All rights reserved.

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

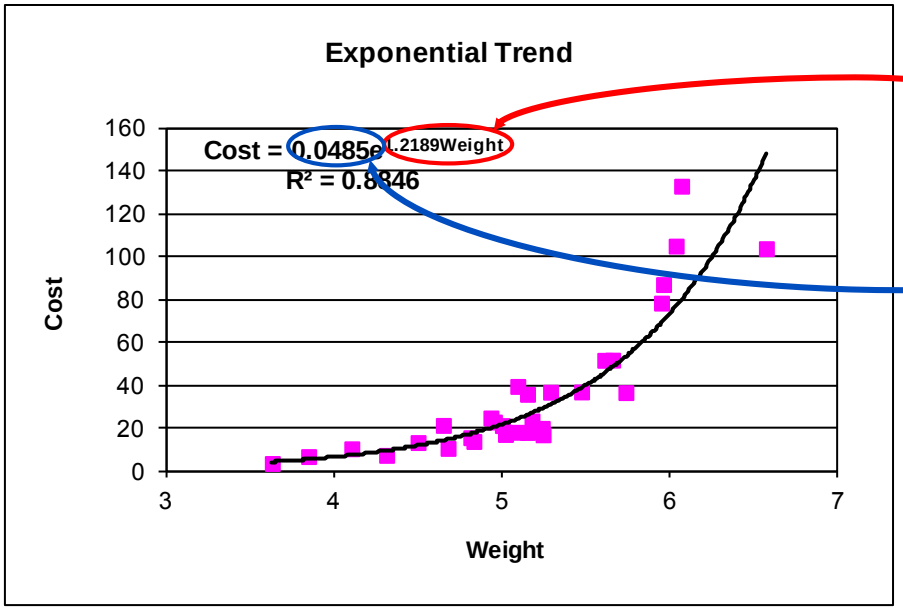
7

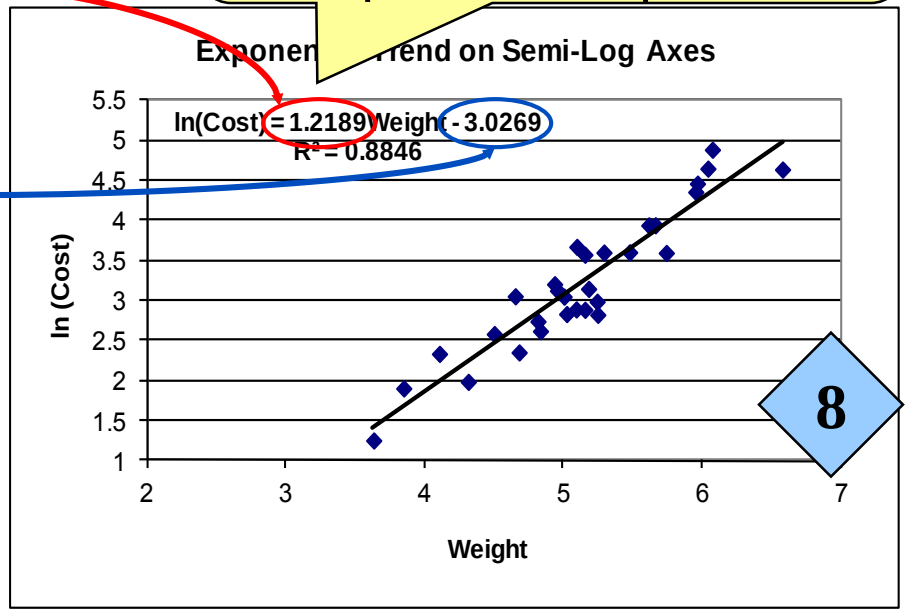
© 2002-2013 ICEAA. All rights reserved.

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

Slope on semi-log graph is the coefficient of x in the exponential equation





© 2002-2013 ICEAA. All rights reserved.

Scatter Plots – Constant Terms

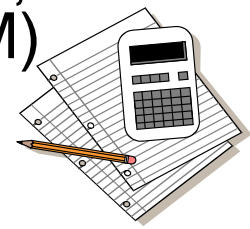
- *Generalized* power and exponential equations are of the form:

$$y = ax^b + c,\ y = ae^{bx} + c$$

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
- 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.



© 2002-2013 ICEAA. All rights reserved.

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

© 2002-2013 ICEAA. All rights reserved.

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

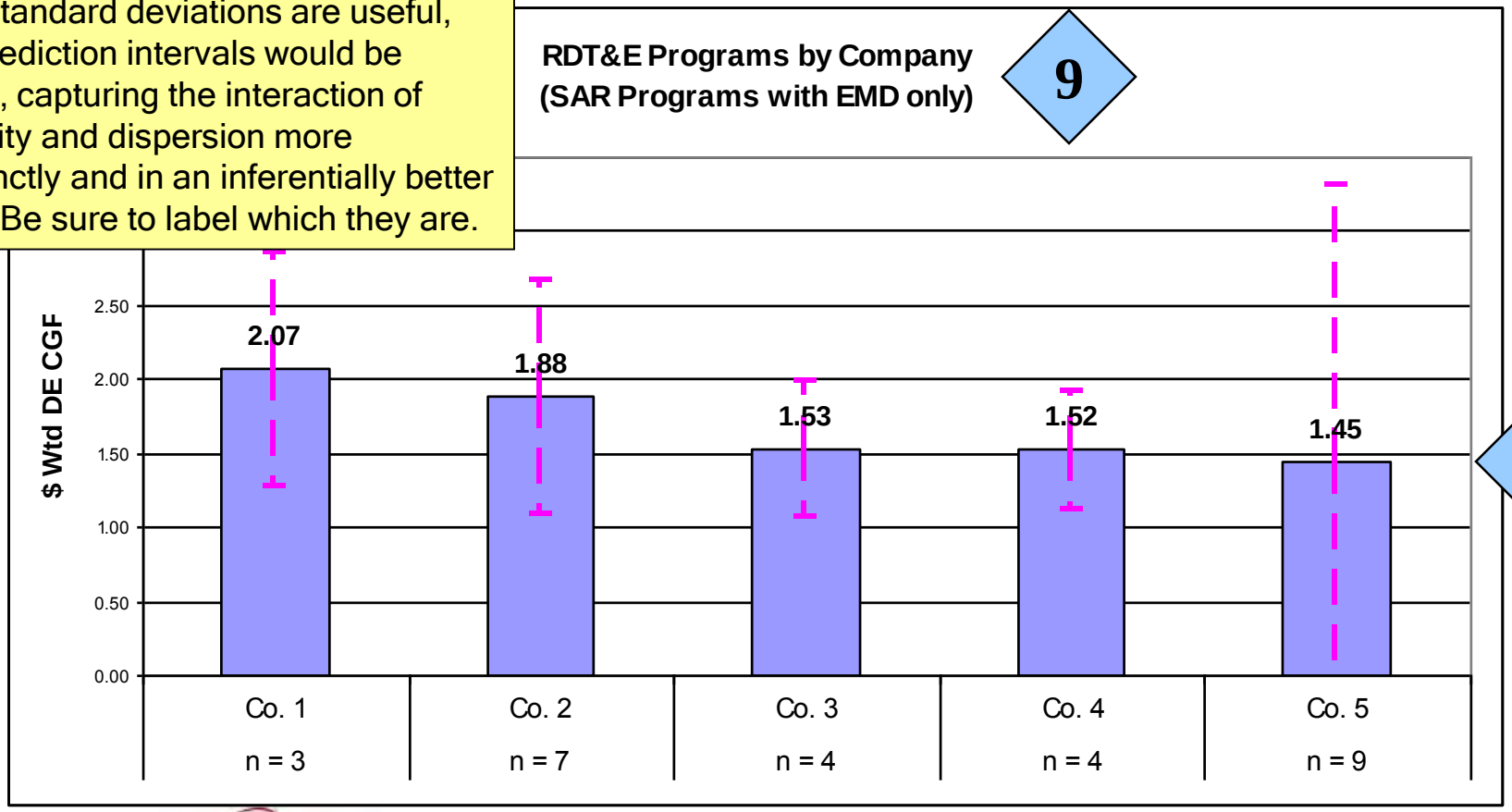
11

© 2002-2013 ICEAA. All rights reserved.

Descriptive Statistics - Bar Charts

- 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.



© 2002-2013 ICEAA. All rights reserved.

Bar Charts in Excel

- 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.

© 2002-2013 ICEAA. All rights reserved.

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

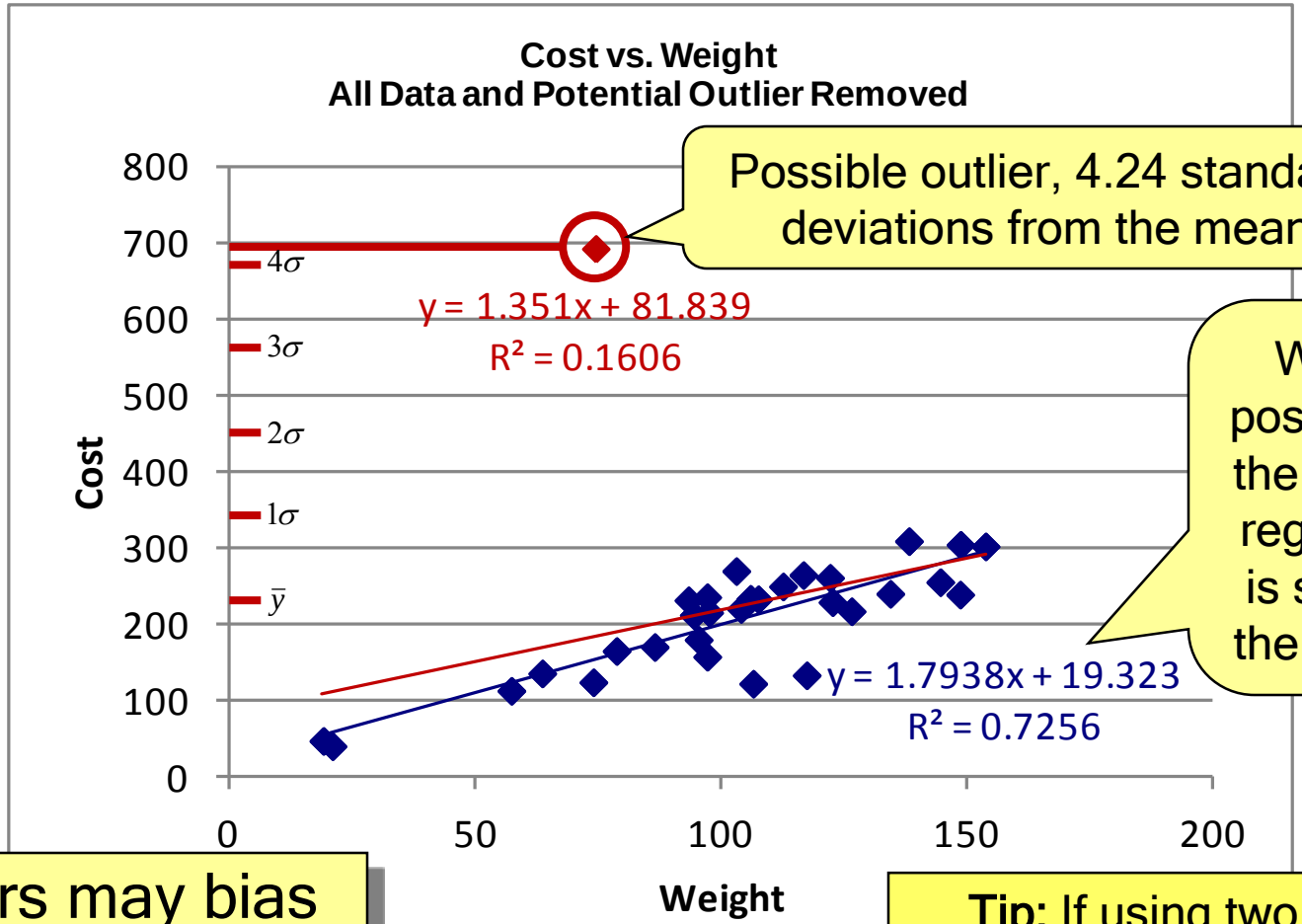
10

  - 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.

© 2002-2013 ICEAA. All rights reserved.

Outliers and Trend Lines



8

Outliers may bias the regression line

Tip: If using two graphs, do not change scale of axes when comparing!

© 2002-2013 ICEAA. All rights reserved.

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

© 2002-2013 ICEAA. All rights reserved.

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

© 2002-2013 ICEAA. All rights reserved.

Two Cautionary Tales

- "Expert's Eyeball"
  - Descriptive Statistics and Visual Displays
- "Technical Hunch"
  - Outliers

© 2002-2013 ICEAA. All rights reserved.

Engineering Judgments

- Suppose we are given an estimate that has “engineering judgment” as its basis
  - Engineering judgments should never be accepted without validation!
  - The analyst must find out if the “guess” is correct, or at least in the ballpark
  - 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

2

14

© 2002-2013 ICEAA. All rights reserved.

Example: *"Expert's Eyeball"*

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%

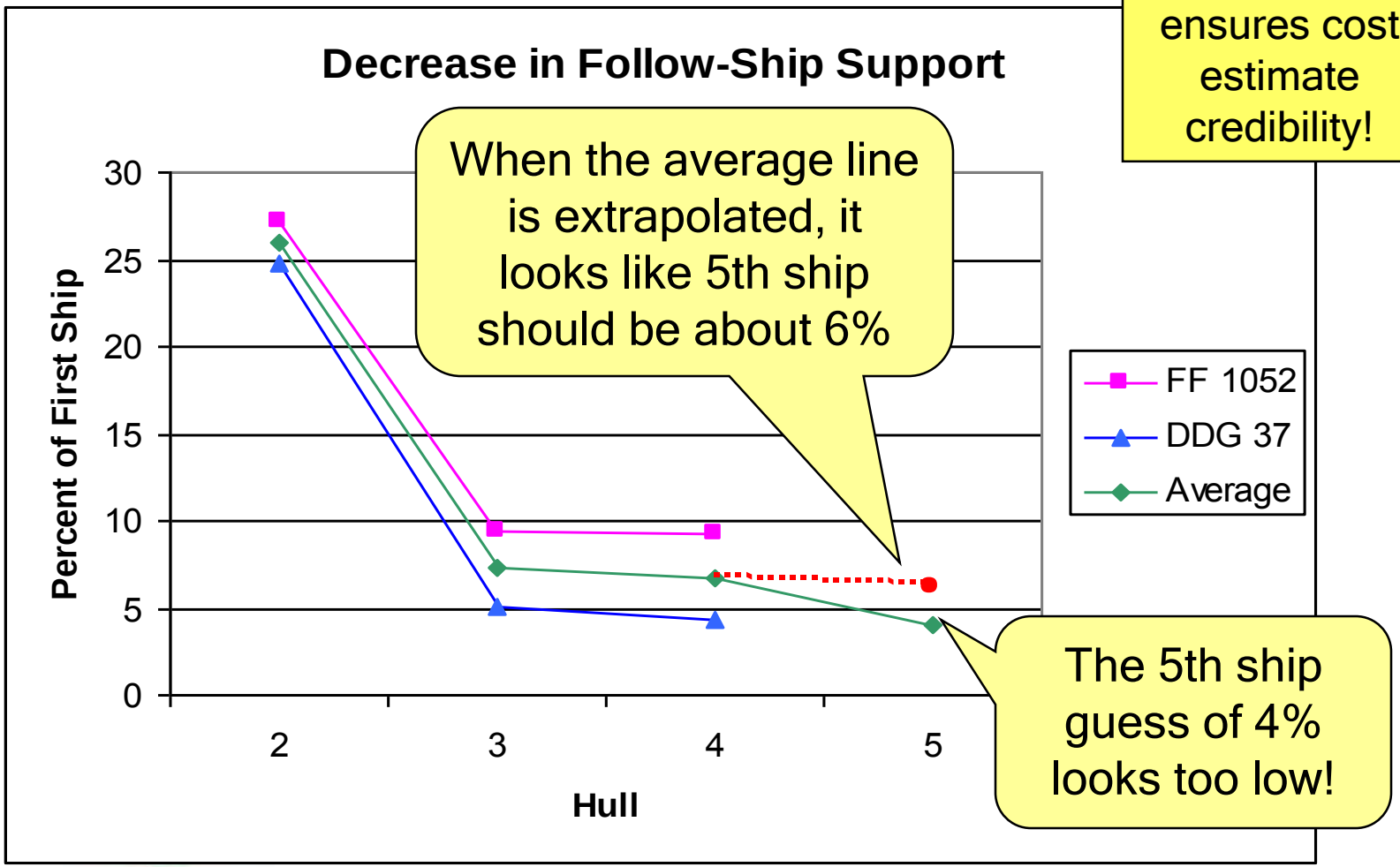
Is the average a good idea?
Is the 5th ship "guess" right?

© 2002-2013 ICEAA. All rights reserved.

Example: *"Expert's Eyeball"*

- The average is a good number!

Tip: Graphic ensures cost estimate credibility!



© 2002-2013 ICEAA. All rights reserved.

Example: *"Technical Hunch"*

- 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

© 2002-2013 ICEAA. All rights reserved.

Example: *"Technical Hunch"*

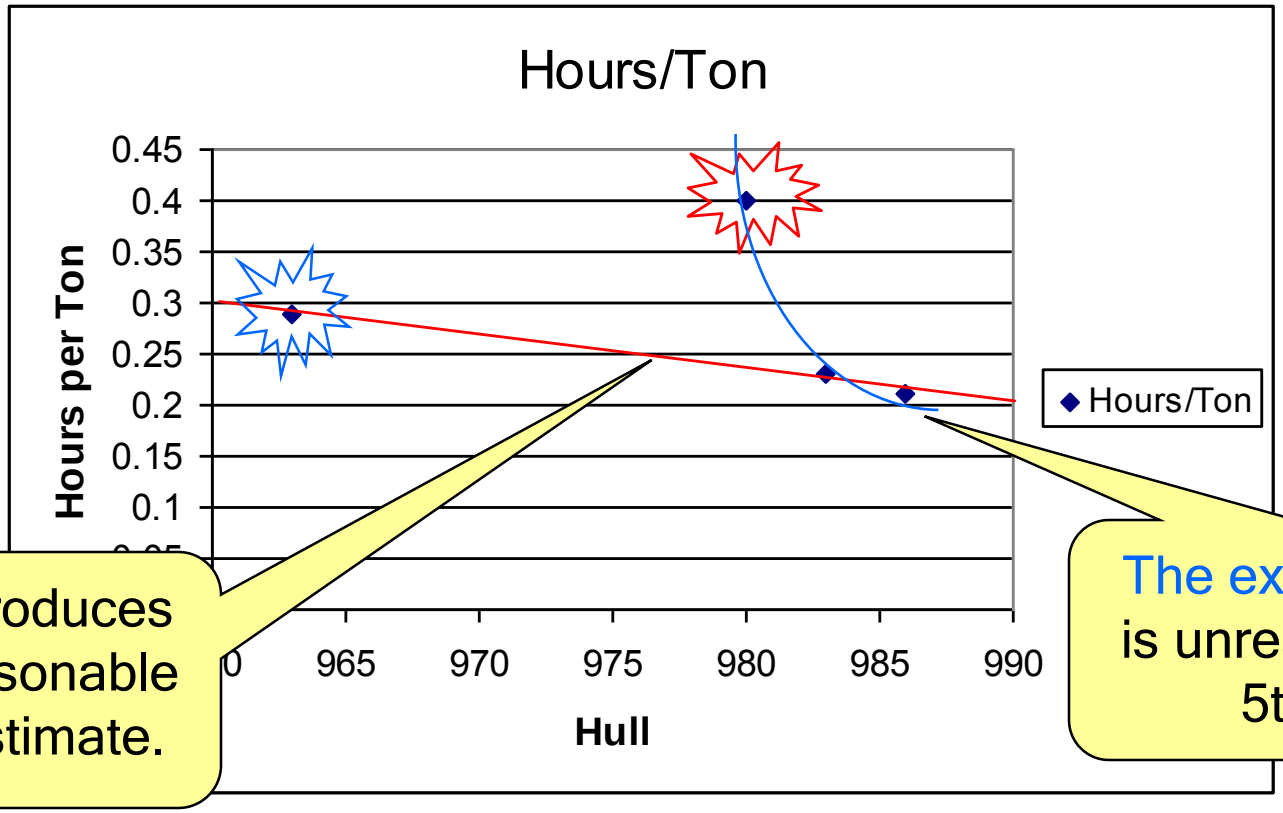
Shakedown	
Hull	**Hours/Ton**
DD 963	0.29
…	…
DD 980	0.40
…	…
DD 983	0.23
…	…
DD 986	0.21

"DD 963 is too low for a first ship"

© 2002-2013 ICEAA. All rights reserved.

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!

© 2002-2013 ICEAA. All rights reserved.

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

© 2002-2013 ICEAA. All rights reserved.

Resources

- *An Introduction to Mathematical Statistics and Its Applications*, 3rd ed., Richard J. Larsen and Morris L. Marx, Prentice Hall, 2000
- *Probability and Statistics for Engineering and the Sciences*, 5th ed., Jay L. Devore, Brooks/Cole Publishing, 1999
- *Calculus: Single Variable*, Deborah Hughes-Hallett and Andrew Gleason, John Wiley & Sons, 1998.
- *How to Lie with Statistics*, Darrel Huff, W.W. Norton & Company, 1954
- *The Visual Display of Quantitative Information*, Edward R. Tufte, Graphics Press, 1983
- *Envisioning Information*, Edward R. Tufte, Graphics Press, 1990
- *Visual Explanations,* Edward R. Tufte, Graphics Press, 1997
- *Beautiful Evidence*, Edward R. Tufte, Graphics Press, 2006

© 2002-2013 ICEAA. All rights reserved.

ODNI Data Analysis Efforts

- **Data**
  - Data Warehouse to store standardized technical, programmatic, cost in database
  - Separate data from estimates and models and organize at an enterprise level
  - Data Quality Scoring
  - Technical Data Collection ongoing in parallel (payloads, buses, software standards)

- **Analysis**
  - Quickly gather data from preapproved set in Warehouse
  - Designed template to analyze data, run regressions, document assumptions
  - Automatically implement many of the CEBoK methods, in addition to custom ones for our dataset

- **Historical Lessons Learned**
  - Tended to do a good job normalizing, collecting and analyzing
  - Not a good job of organizing; Data Warehouse greatly assists analysis
  - Earlier organization operated at highest-level factors
  - Do not have to operate at highest level
    - Can position ourselves to develop methods at lower levels instead of factors
    - Better tools, better use of time, established processes, streamlined documentation
    - Data already established; verified; demonstrated - - good pedigree
  - Where in cost shop lifecycle do we shift from analogies to parametrics
    - More serious about method development
    - Newer organizations use factors rather than parametrics
  - Data were decoupled from applications/models, but actually strengthen the bonds in the analysis

ODNI Data Warehouse

Structured Archival/Retrieval

IC Agencies

Cost Assessments
Joint Research
Program Reviews
Events (e.g., launches)

Non-IC

Data Collection Trips
Joint Research
Peer Reviews
Symposia

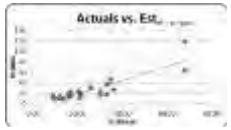
Industry

Data Collection Trips
One-on-One Sessions
Conferences

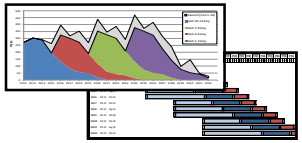
Normalization & Preparation

Data Warehouse

Cost Estimating Relationships



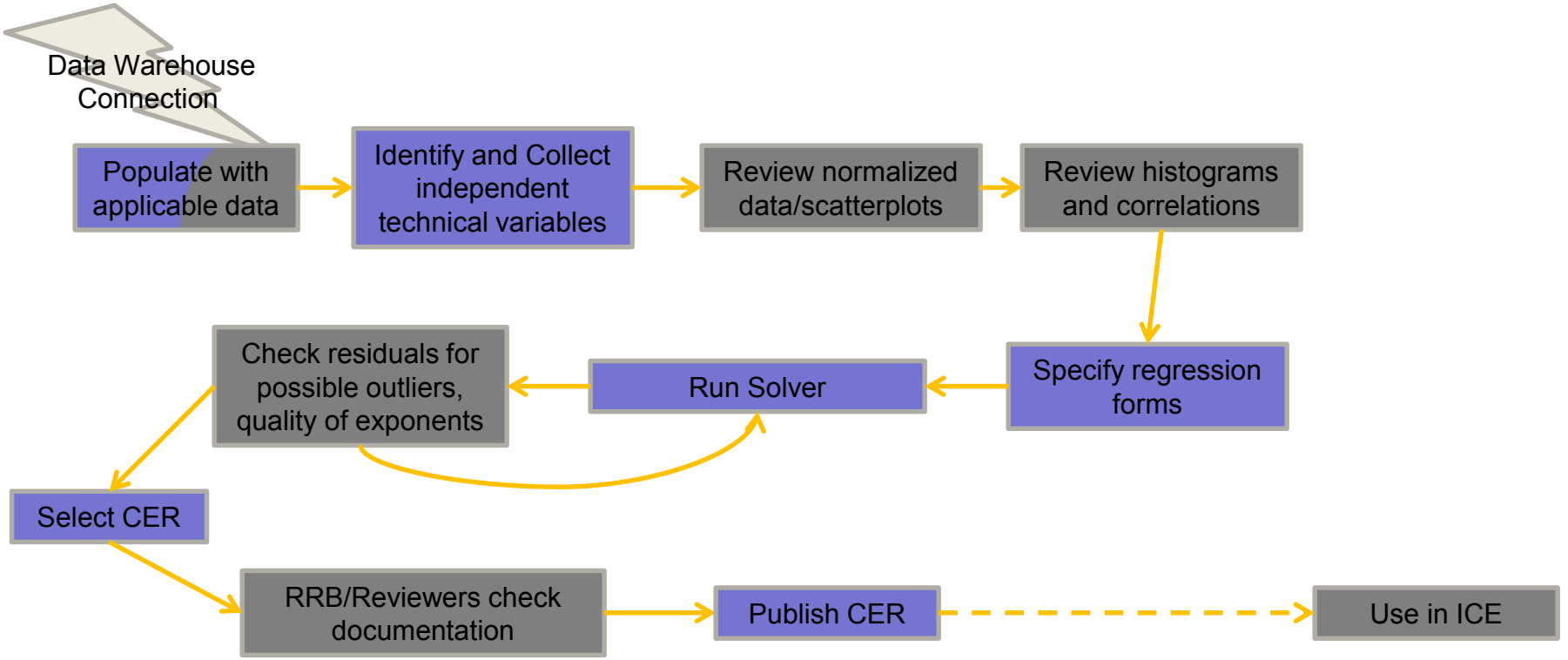
Analysis of Alternatives



Modeling & Analysis

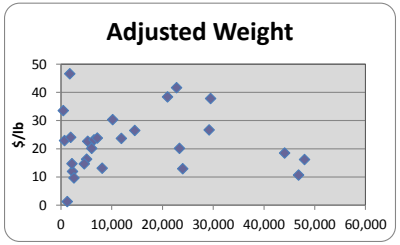
ICE Cost and Schedule Analysis

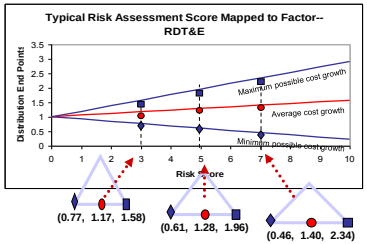
ODNI Standard Analysis Process

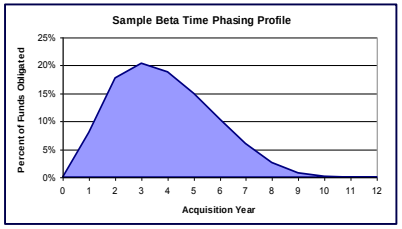


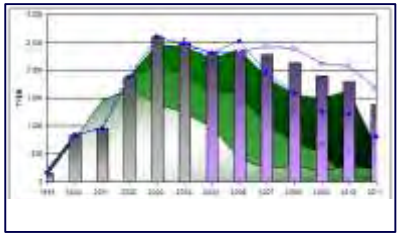
ODNI ICE Process

- **Customize model templates to fit current effort**
  - Maintains best practices while allowing tailored approach to task at hand
- **Develop methods based on combined available data and identify / develop cross-checks**
- **Incorporate risk assessments into model development**
  - Schedule/Technical risks identified through Evaluation phase
  - Cost estimating uncertainty as it pertains to methods (CER or factor standard deviation, etc.)
- **Integrate estimate inputs (e.g., space and ground components) into a single consolidated estimate**
  - Apply risk, integration factors to overall estimate
- **Time-phase estimate to support budget development**
  - Use phasing tools and historical profiles to determine “realistic” funding profile that reflects the independent schedule assessment
  - Apply cross-checks for achievability (e.g., staffing profiles)
- **Incorporate assessment of historical budget and future trends to adjust estimate profile**
  - Integrates program factors (existing budget constraints, execution) and external factors (sequestration, political decisions, etc.)









Related and Advanced Topics

- Visual Display of Information
- Additional Graph Types for Univariate
  - Stem-and-Leaf
  - Boxplots
- Bin Width and Number Rules
- Mean - Mental Math Trick
- Sample Sizes
  - Confidence Intervals
  - CI Simplified
  - Sufficiently Large n
  - Rules of Thumb
- Outlier Identification Rules

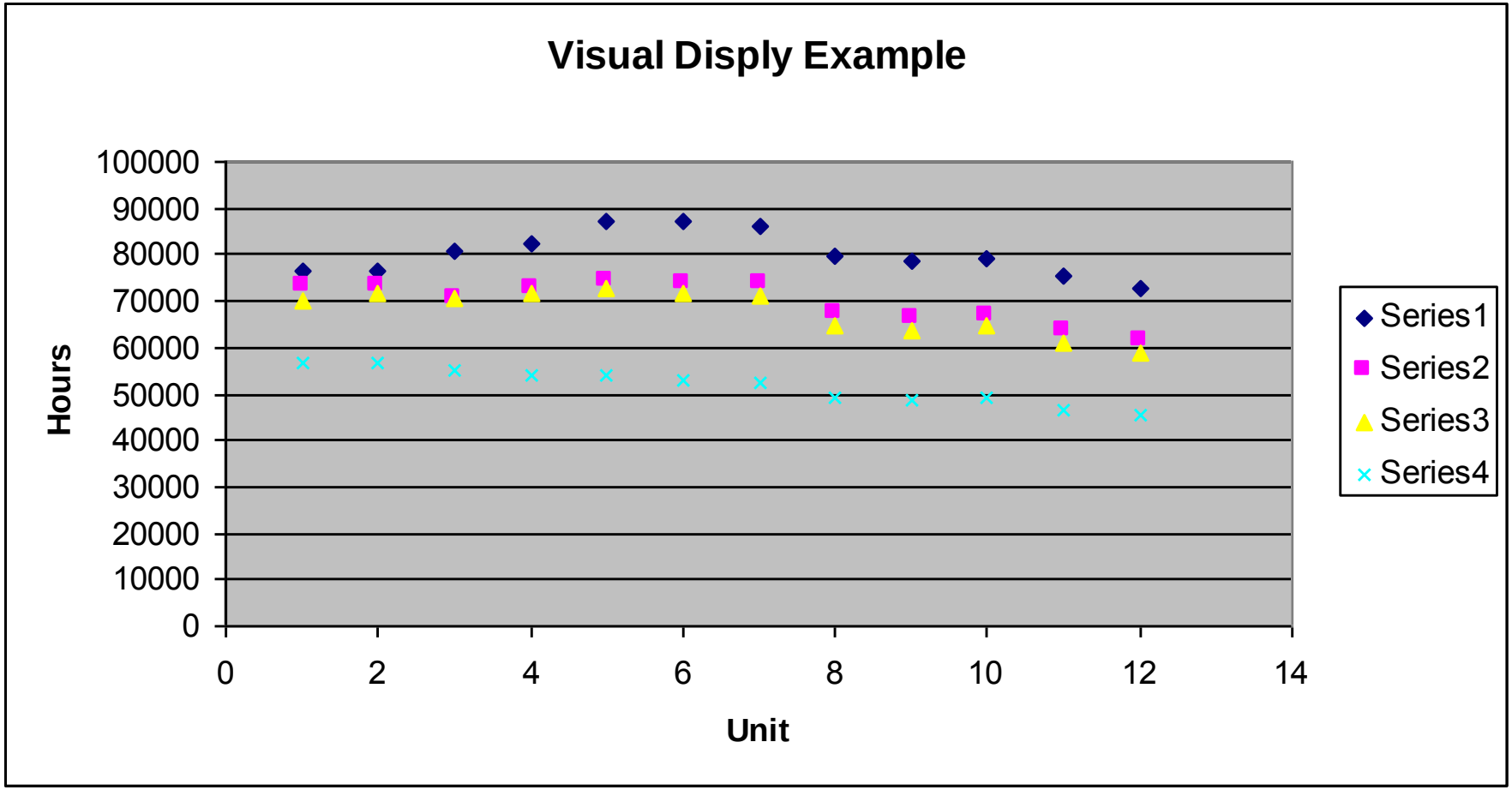
© 2002-2013 ICEAA. All rights reserved.

Visual Display of Information

- Poor visual displays of information hinder understanding
  - Excel's default scatter plot is not a "one-size-fits-all" information display
- Quick fixes ensure a graph can truly give "the gift of sight"
  - Use evocative colors to your advantage
  - Size matters
    - Make sure the graph fills the space – the data is the main event!
    - Check the scale
    - Choose a font size
  - Check the placement of the legend
- Two possible displays follow

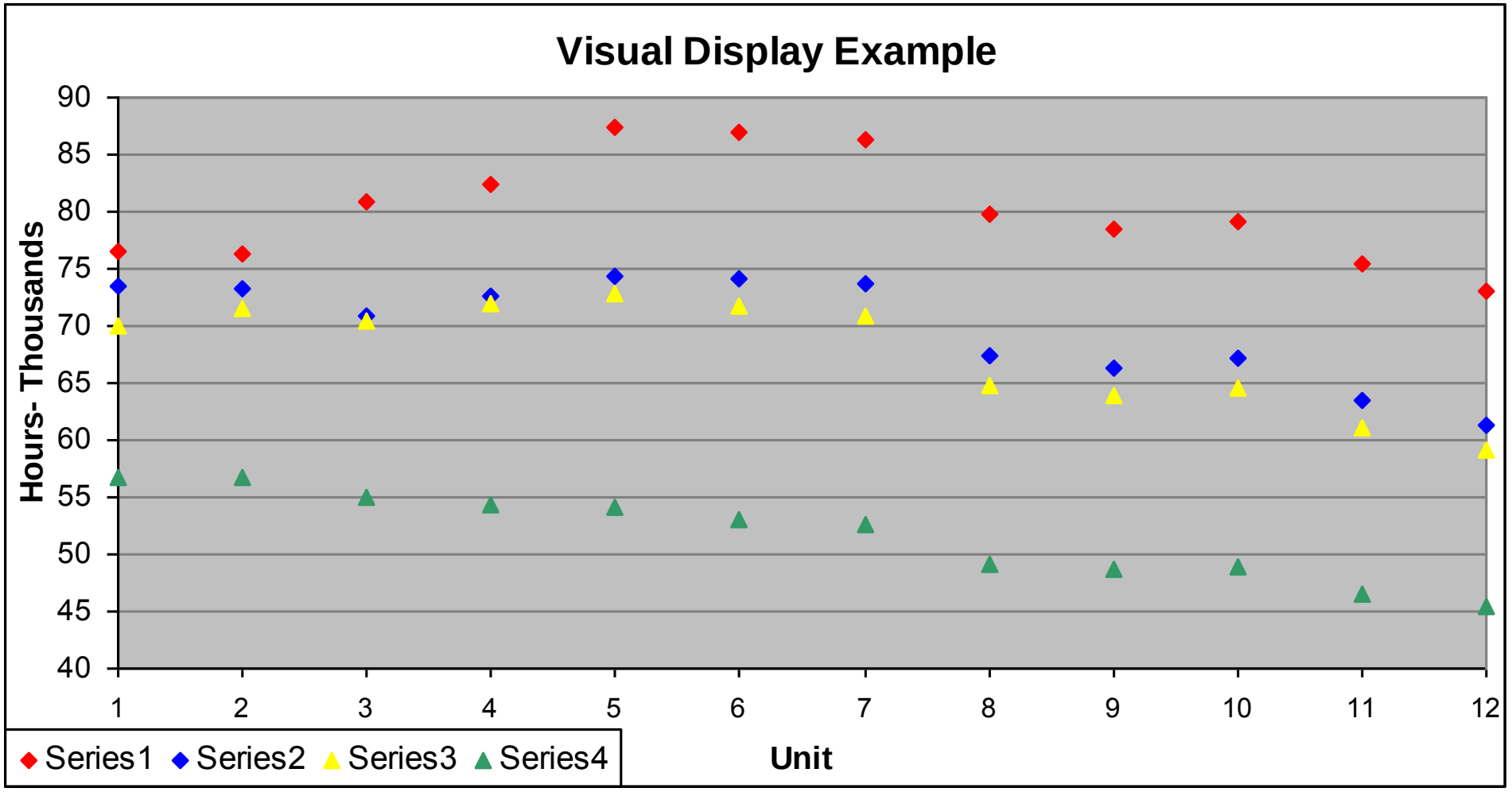
© 2002-2013 ICEAA. All rights reserved.

Visual Display of Information – Excel Default



© 2002-2013 ICEAA. All rights reserved.

Visual Display of Information – Another Display



© 2002-2013 ICEAA. All rights reserved.

Stem-and-Leaf Plots

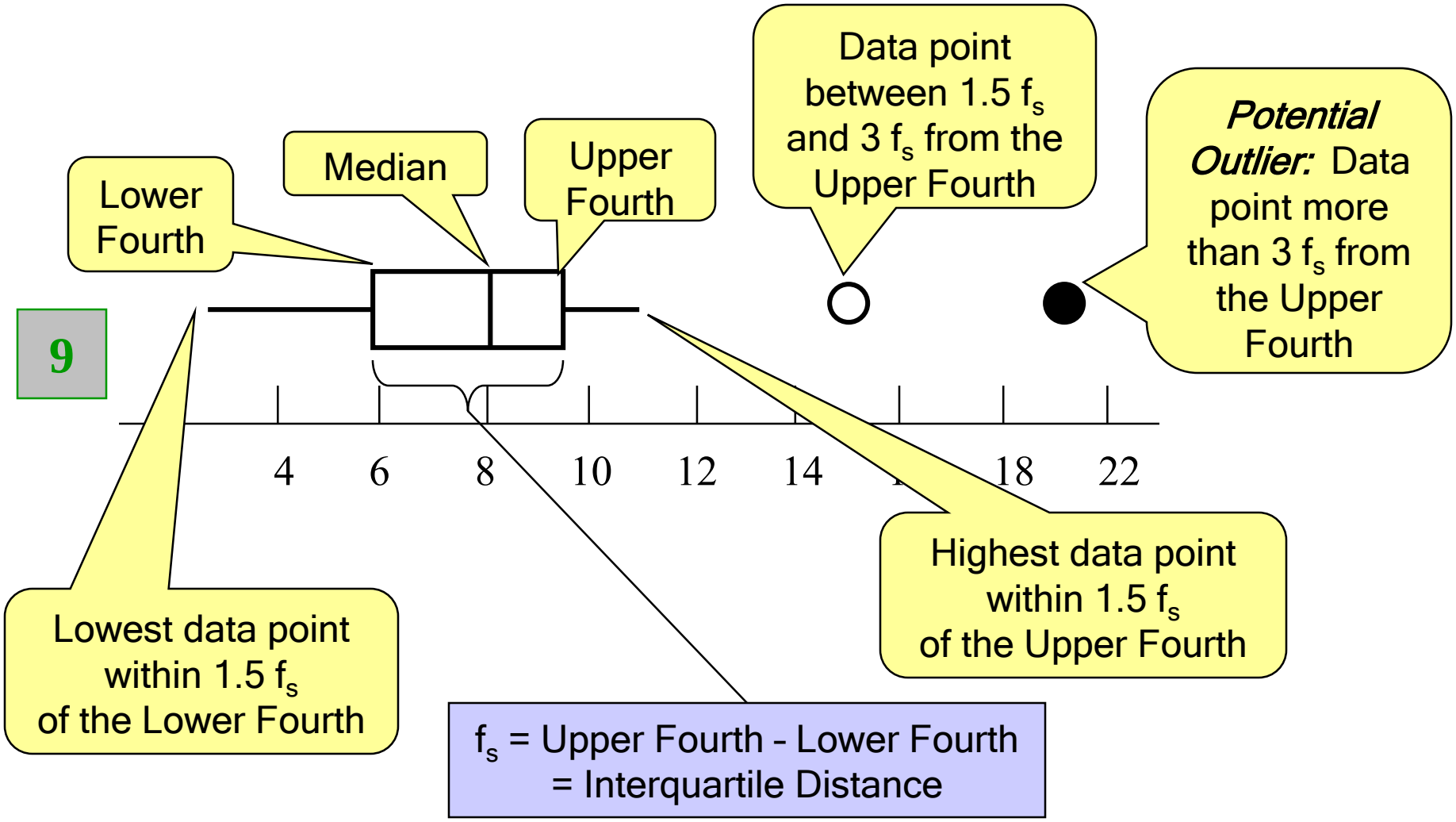
- Similar to a histogram
  - Horizontal numbers instead of vertical bars
- Example:
  - Therms of natural gas used

0	333333333333344444444444444555566677889999
1	005578
2	11135666679
3	2333889
4	8
5	46
6	02

Mode = 4 therms

© 2002-2013 ICEAA. All rights reserved.

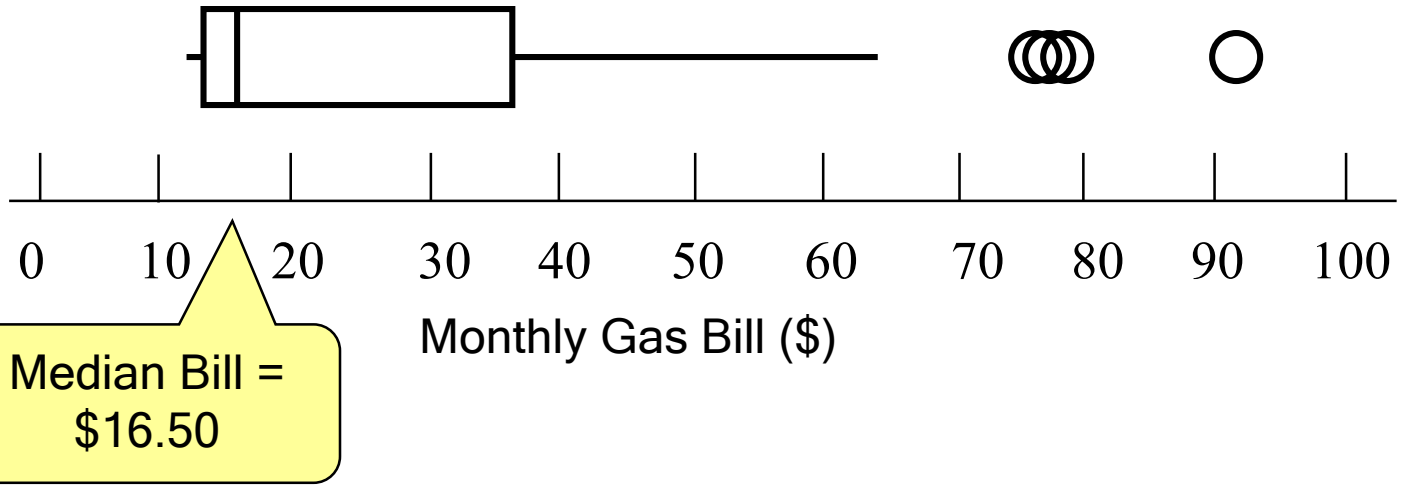
Box Plots



© 2002-2013 ICEAA. All rights reserved.

Box Plots Application

- Box plots can be used to:
  - Show the center, spread, and symmetry of the data
  - Identify outliers
- A sample box plot is shown on the previous slide, and a "real-world" one below:



© 2002-2013 ICEAA. All rights reserved.

Bin Widths and Number Rules

Various Rules for Bin Width (*h*) or Number of Bins (*k*) based on Number of Data Points (*n*), Sample Standard Deviation (*s*), and Interquartile Range (*IQR*)

Bin Width (*h*)	Number of Bins (*k*)	Assumptions	Comments
Square Root Rule			
$\frac{Max\{x_i\}-Min\{x_i\}}{k}$ ⇐	$\lvert\sqrt{n}\rvert$	-	Used in Excel Data Analysis Histogram tool
Sturges' Rule			
$\frac{Max\{x_i\}-Min\{x_i\}}{k}$ ⇐	$\lceil \log_2 n+1 \rceil$	$30 < n < 200$	Used by DAU
Scott's Rule			
$\frac{3.5s}{\sqrt[3]{n}}$ ⇒	$\left\lceil \frac{Max\{x_i\}-Min\{x_i\}}{h} \right\rceil$	Normal distribution	Reasonable default if data not too skewed
Freedman-Diaconis Rule			
$\frac{2 \cdot IQR}{\sqrt[3]{n}}$ ⇒	$\left\lceil \frac{Max\{x_i\}-Min\{x_i\}}{h} \right\rceil$	Normal distribution	Modifies Scott's Rule by focusing on *IQR* instead of *s*

© 2002-2013 ICEAA. All rights reserved.

Mean - Mental Math Trick

- The mean can also be an arbitrary number plus the average of the deviations from that number:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{n} \left(X_i - X^* + X^* \right)}{n} = \frac{nX^* + \sum_{i=1}^{n} \left(X_i - X^* \right)}{n} = X^* + \frac{\sum_{i=1}^{n} \left(X_i - X^* \right)}{n}$$

- Monthly average therms used data: {37, 26, 13, 3, 3, 3, 3, 3, 4, 7, 21, 40}
- Average = 10 + (27+16+3-7-7-7-7-7-6-3+11+30)/12 = 10 + 43/12 = 13.6

© 2002-2013 ICEAA. All rights reserved.

Sample Sizes – Confidence Interval

- How big a sample size do we need so that a 68.3% Confidence Interval (one standard deviation) about the estimate is +/-5% of the estimate?
  - i.e., there is 68.3% probability that the population mean is within 5% of our estimated mean.
- Consider the confidence interval for the mean of a normal distribution

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

- Note that the size of the range around the estimate of the mean is a function of:
  - the variability, captured by standard deviation, s, or coefficient of variation, CV
  - the sample size, n

Note: we are assuming a normal distribution for simplicity

© 2002-2013 ICEAA. All rights reserved.

Sample Sizes - CI Simplified

- Instead of working with standard deviations, we would like to shift to CVs
  - CVs are unit-less and more intuitive (expressed in percents)
- So, divide the range by $\bar{x}$

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

$$\frac{\bar{x}}{\bar{x}} \pm t_{\alpha/2,n-1} \frac{s}{\bar{x}\sqrt{n}}$$

$$1 \pm t_{\alpha/2,n-1} \frac{CV}{\sqrt{n}}$$

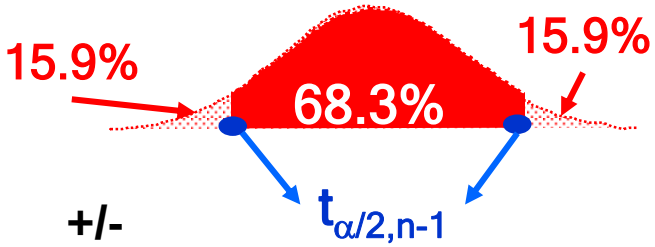
This shifts the range into percents. The range is relative to 100% of the estimate

$$\pm t_{\alpha/2,n-1} \frac{CV}{\sqrt{n}}$$

© 2002-2013 ICEAA. All rights reserved.

Sample Sizes – Sufficiently Large n

- What sample size is needed for judgments within 5%?
  - For a 68.3% two-tailed CI, we have α = 1 - 0.683 = 31.7% and thus $\alpha/2$ = 15.9%
  - Suppose we have a CV of 30%



n	CV	$t_{0.159,\,n}$	+/-
4	30%	1.141627	17%
9	30%	1.058728	11%
16	30%	1.032242	8%
25	30%	1.020404	6%
36	30%	1.014083	5%

$$t_{\alpha/2,n-1}\frac{CV}{\sqrt{n}}$$

Note: for a 95% CI we would use α = 0.05. The t multipliers would vary from 2.78 to 2.03

  - We would like to be able to make judgments within about 5% points, so with a CV of 30%, we need $n \approx 36$

© 2002-2013 ICEAA. All rights reserved.

Sample Sizes - Rule of Thumb

- For an easy rule of thumb, we can just round the *t* value to *t* = 1
  - Then, we use simply $\frac{CV}{\sqrt{n}}$

n	CV	$t_{0.159,\,n}$	Exact +/-	Thumb rule
4	30%	1.141627	17%	15%
9	30%	1.058728	11%	10%
16	30%	1.032242	8%	8%
25	30%	1.020404	6%	6%
36	30%	1.014083	5%	5%

$t_{\alpha/2,n-1}\frac{CV}{\sqrt{n}}$

Tip: For a 68.3% CI, use $\frac{CV}{\sqrt{n}}$. For a 95% CI, use $2\frac{CV}{\sqrt{n}}$.

© 2002-2013 ICEAA. All rights reserved.

Outlier Identification Rules

Rule	Outlier(s) Iff…	Rationale
Chauvenet's Criterion	$n \cdot \left[1 - \Phi\left(\frac{\lvert x - \bar{x} \rvert}{s} \right) \right] < 0.5$	Normal distribution properties
Grubbs' Test	$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}}$	Normal distribution properties, where $G = Max\left\{ \frac{\lvert x - \bar{x} \rvert}{s} \right\}$
Dixon's Q Test	Gap/Range > (critical value from table), where Gap = distance between outlier and its closest neighbor	Unclear. Will not detect two approximately equal outliers.
IQR-Based	x not in the interval $[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$	Can customize k based on choice of distribution, α, and n. For example, in a normal distribution, k = 3 implies that < 5% of points should fall outside the range.

© 2002-2013 ICEAA. All rights reserved.