



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”

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



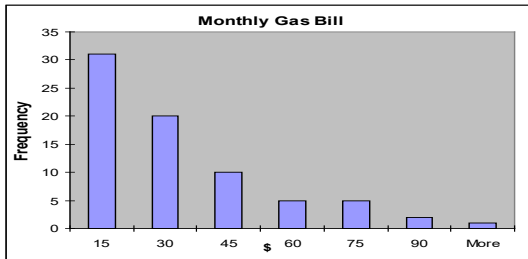
Key Ideas	Analytical Constructs	Practical Applications	Related topics
			
<ul style="list-style-type: none">• Visual Display of Information• Central Tendency of Data• Dispersion (Spread) of Data• Data Visualization• Machine Learning Algorithms	<ul style="list-style-type: none">• Descriptive statistics<ul style="list-style-type: none">• Mean, median, mode• Variance, std deviation, CV• Functional forms	<p>Making sense of your data</p>	<ul style="list-style-type: none">• Parametric• Distributions• Probability and Statistics

Data Analysis Within The Cost Estimating Framework

ICEAA International Cost Estimating and Analysis Association

Past

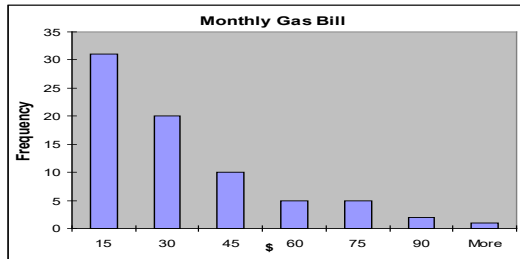
Understanding your historical data



Historical data

Present

Developing estimating tools

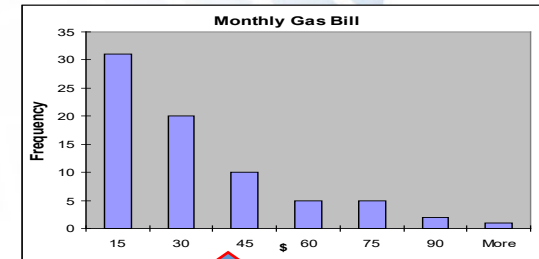


Mean = \$34.19

Average cost

Future

Estimating the new system



Confidence Interval = +/- \$5.76

Confidence Intervals

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
- Machine Learning Algorithms
 - Supervised Learning
 - Unsupervised Learning

Summary

Resources

Related and Advanced Topics

Types of Data



Univariate

Bivariate

Multivariate

Time
Series

Univariate Data Analysis



Visual Display of Information

- Histogram, stem-and-leaf, box plot

What does it look like?

Measures of Central Tendency

- Mean (or median or mode)

What's your best guess?

Measures of Variability

- Standard deviation (or variance), coefficient of variation (CV)

How much remains un-explained?

Measures of Uncertainty

- Confidence Interval (CI)

How precise are you?

Statistical Tests

- t test, chi square test, Kolmogorov-Smirnov (K-S) test

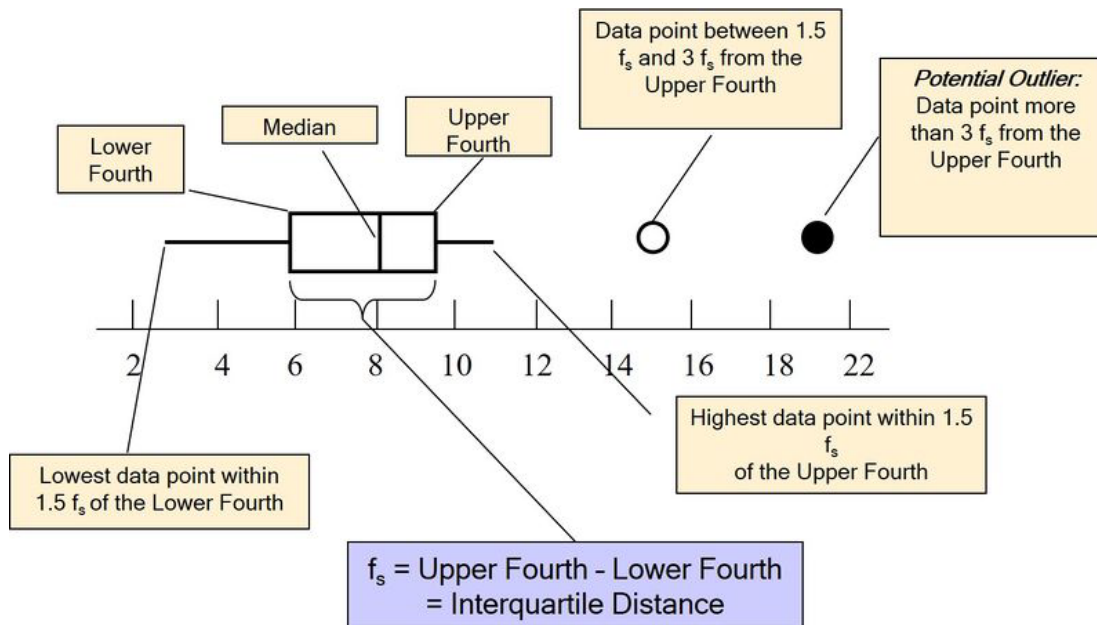
How can you be sure?

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

Visual Display – Box Plots

ICEAA International Cost Estimating and Analysis Association

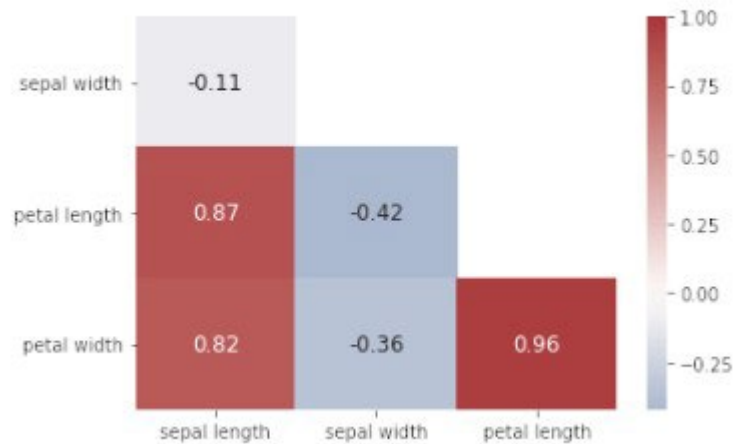
- Standardized way of summarizing distribution of data set



Visual Display – Heat Maps

ICEAA International Cost Estimating and Analysis Association

- Visualizes correlation between variables in a data set



Visual Display – Waterfall Chart



- Highlights changes between elements over time



(a)



(b)



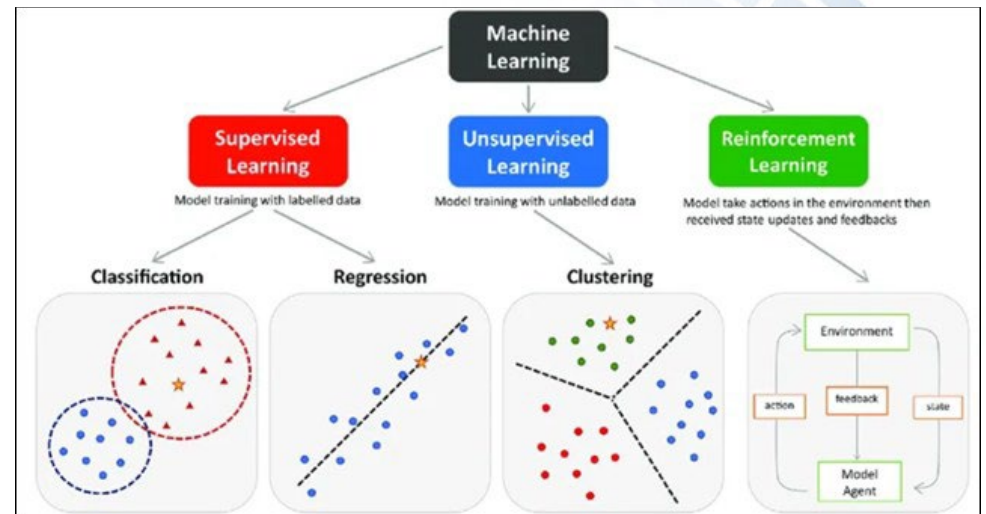
Machine Learning Algorithms

What's New?



Machine Learning (ML) is a field of artificial intelligence in which algorithms learn from data to make predictions and inform decisions.

- Supervised learning
- Unsupervised Learning
- Reinforcement Learning

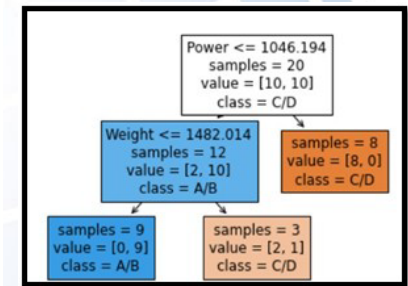
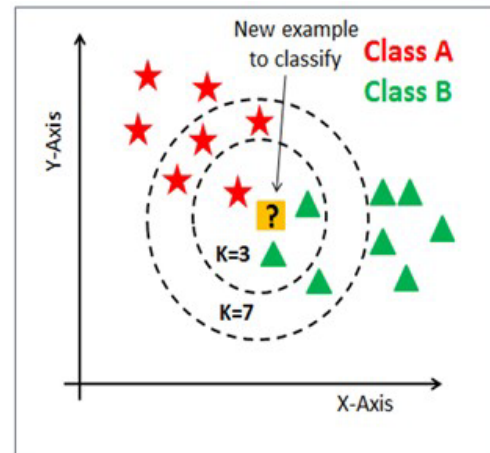
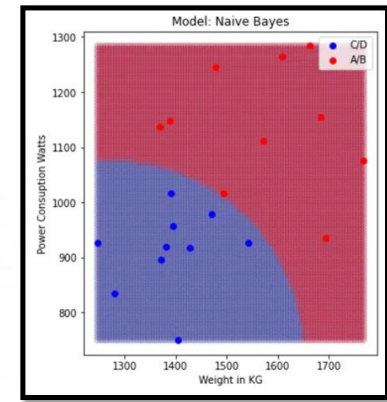
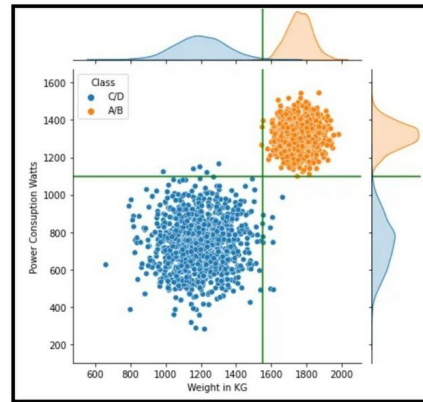


Supervised Learning



Classification:

- Naïve Bayes
- Decision Trees
- Random Forests
- k-Nearest Neighbors
- Logistic Regression

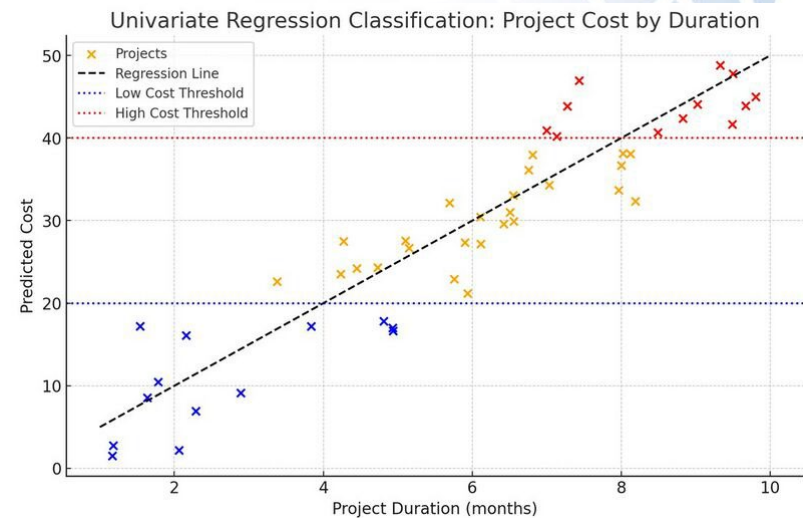
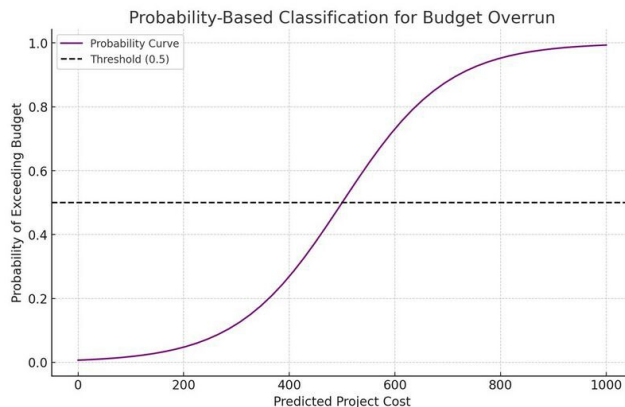
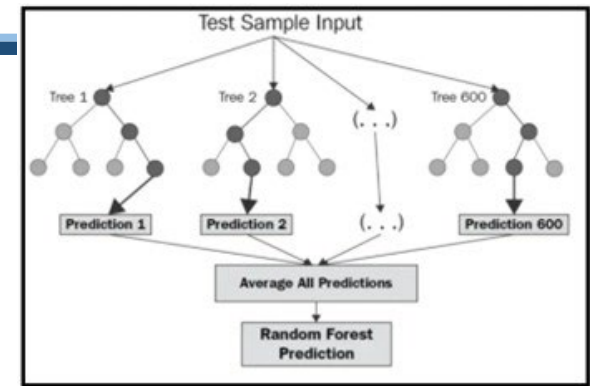


Supervised Learning



Regression:

- Linear Regression
- Decision Trees for Regression
- Random Forests for Regression
- k-NN for Regression

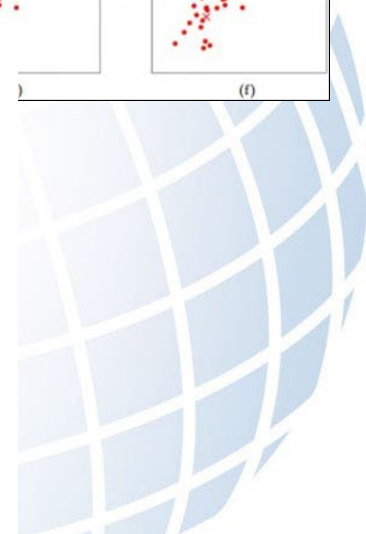
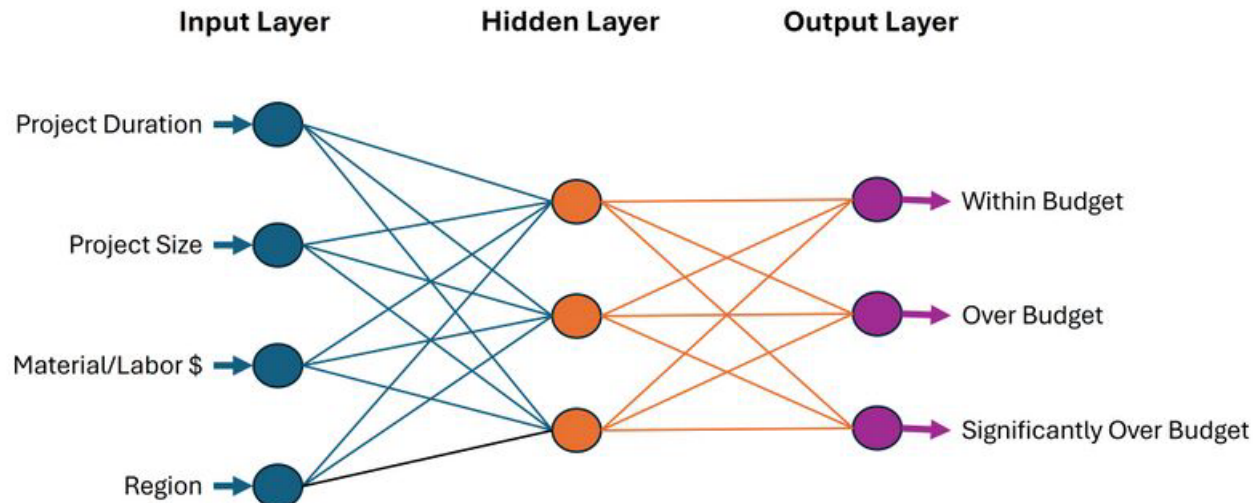
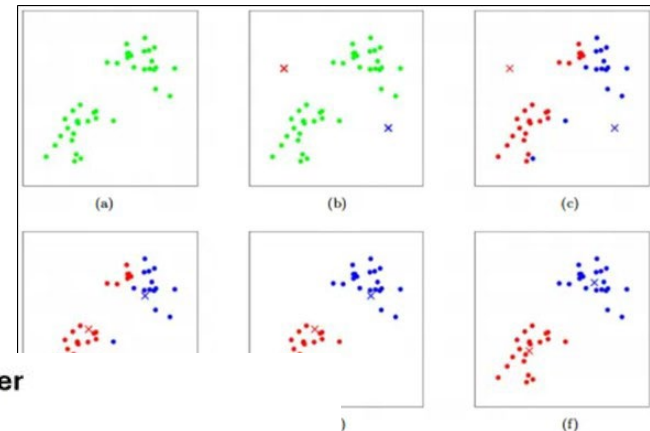


Unsupervised Learning



Regression:

- k-Means Clustering
- Neural Networks



Reinforcement Learning



Reinforcement learning is an advanced ML approach that responds to an environment in real time, with a specified goal. Examples include an autonomous vehicle with the goal to safely transport individuals from one location to another, or a computer capable of competing against a human in a game of chess with the goal to win. Some differences between reinforcement learning and supervised/unsupervised learning include:

- no supervisor to guide the training;
- no training with a large pre-collected dataset (data is provided dynamically via feedback from the real-world environment in which you are interacting);
- iterative decision making over a sequence of time-steps where inferences are run repeatedly, navigating through the real-world environment as you go.

New Example – Classifying Risk with Logistic Regression



Table 6.4: Logistic model coefficients

Feature	Coefficient (β)	Interpretation
Intercept	-8.0	Baseline log-odds when all features are zero
Project Size	0.9	Larger projects increase log-odds of high cost risk
Duration	0.5	Longer projects increase the log-odds of high cost risk
Team Experience	-0.8	More experienced teams decrease log-odds of high cost risk
Team Members	0.2	Larger teams slightly increase the log-odds of high cost risk
Complexity Score	1.5	Higher complexity significantly increases the log-odds of high cost risk
Market Volatility Index	0.05	Higher market volatility slightly increases the log odds of high cost risk

To predict the probability P of a new project being high risk, remember the logistic regression model equation:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

where:

- β_0 is the intercept,
- β_i are the coefficients, and
- X_i are the feature values.

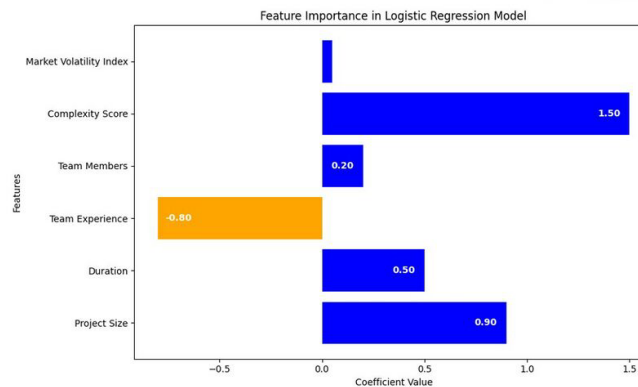
If the new project has a project size of \$4M, a duration of 10 months, team experience of 6 years, complexity of 6, and a market volatility index of 65, we can compute log odds:

$$\text{Log-odds} = -8.0 + (0.9 * 4) + (0.5 * 10) + (-0.8 * 6) + (0.2 * 9) + (1.5 * 6) + (0.05 * 65)$$

$$\text{Log-odds} = -8.0 + 3.6 + 5 - 4.8 + 1.8 + 9.0 + 3.25 = 0.85$$

and therefore:

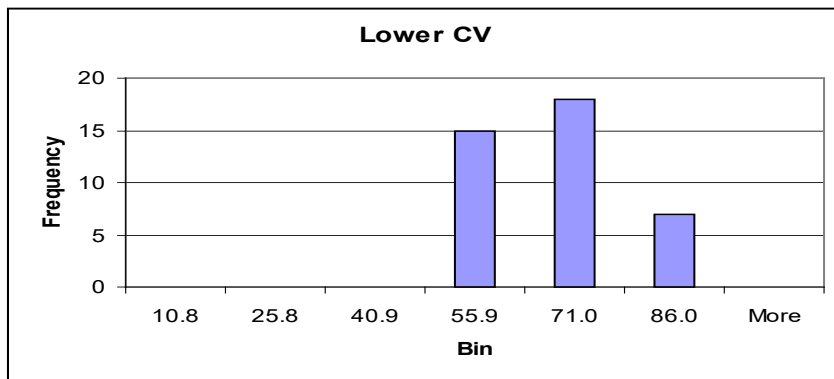
$$P = \frac{1}{1 + e^{-0.85}} = 0.99995$$



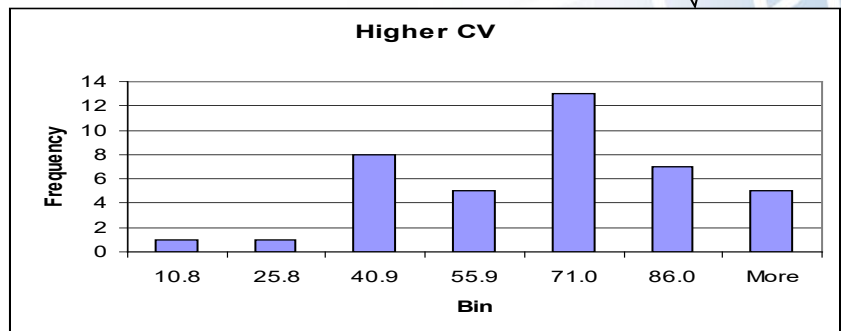
Dispersion and CV



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



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



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

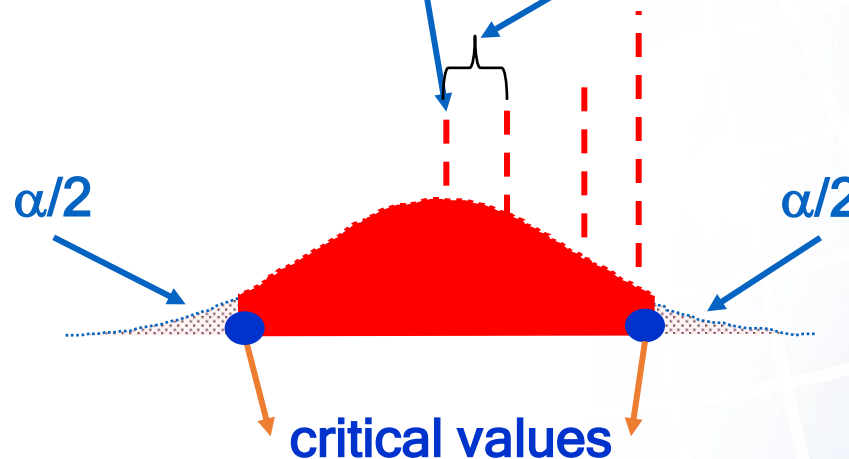
Confidence Interval Illustration



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*

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

* 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:

➤ 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 $n \gg 36$
 - We may have no choice but to deal with small n
 - In any case, we can calculate the range of estimated mean

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
 - 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
 - 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

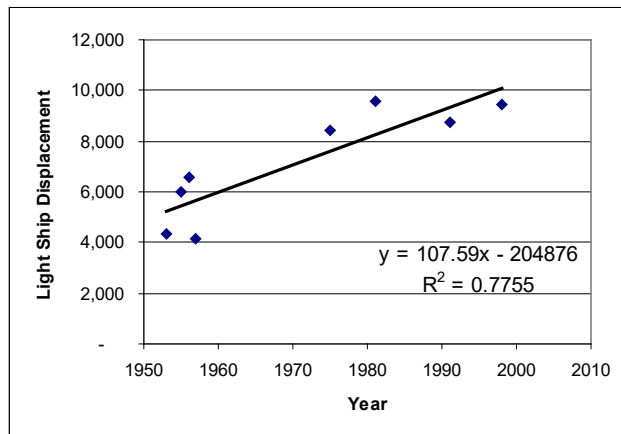


Variables

Actions

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

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



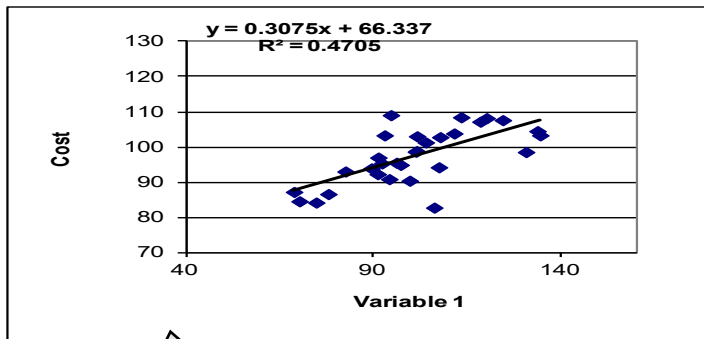
Scatter Plots - Variables

- Plot cost (or other variable of interest, e.g., hours) as the dependent variable
- Look at a variety of different independent variables
 - 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!

Scatter Plots – Cost Drivers



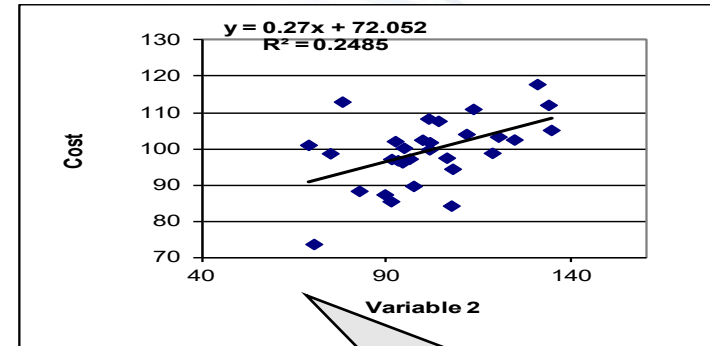
- Scatter plots can help identify cost drivers
- R^2 interpretation: % of variation in y explained (linearly) by variation in x



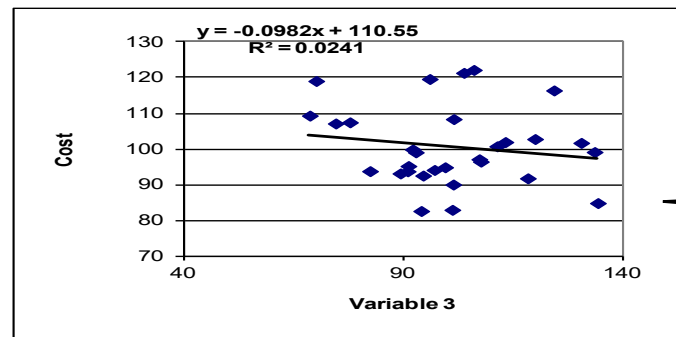
Significant correlation - potential cost driver



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



Weak correlation



Uncorrelated

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

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

Outlier Identification Rules



Rule	Outlier(s) Iff...	Rationale
Chauvenet's Criterion	$n \cdot \left[1 - \Phi \left(\frac{ x - \bar{x} }{s} \right) \right] < 0.5$	Normal distribution properties
Grubbs' Test	$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{\alpha/(2n), n-2}^2}{n-2 + t_{\alpha/(2n), n-2}^2}}$	Normal distribution properties, where $G = \text{Max} \left\{ \frac{ x - \bar{x} }{s} \right\}$
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.

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

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

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