Welcome to Lean and Six Sigma Training

Lean Six Sigma Black Belt
Summarized Material

Roles & Responsibilities

Leadership/ Sr. Executive
- Owns vision, direction, integration, results
- Leads change

Green Belts
- Supports Black Belts by participating on project teams

Black Belts
- Apply Breakthrough Strategy to specific projects, lead and direct teams to execute projects

Master Black Belt
- Trains and coaches Black Belts , Green Belts and leaders

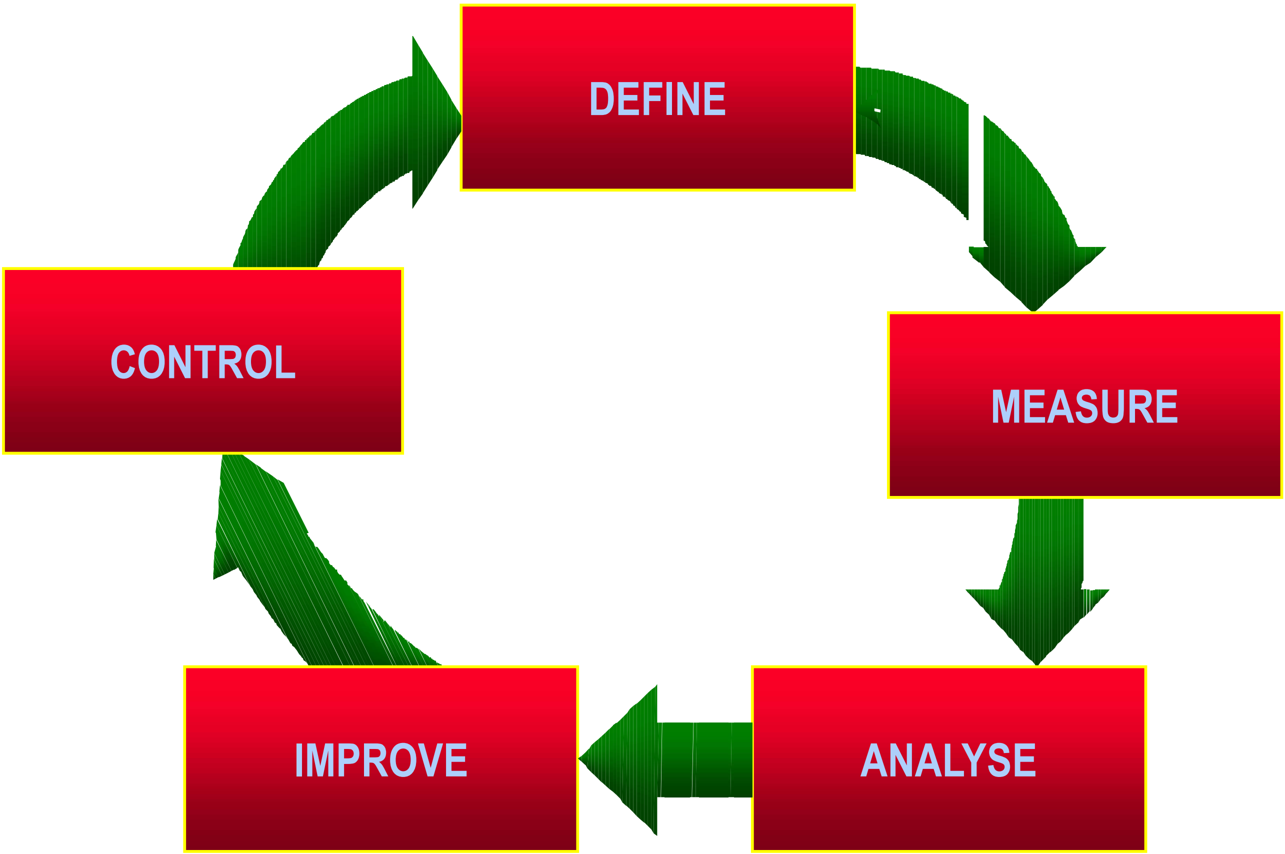
Process Owner
- Owns the process in which GB/BB works
- Challenges GB findings

Champion / Sponsor
- Develops deployment and Strategy
- Supports cultural change

Yellow Belts
- Support Six Sigma
- Ensure success of GB/BB projects

All Employees (White Belts)
- Understand vision
- Apply concepts to their job and work area

DMAIC : An Improvement Methodology

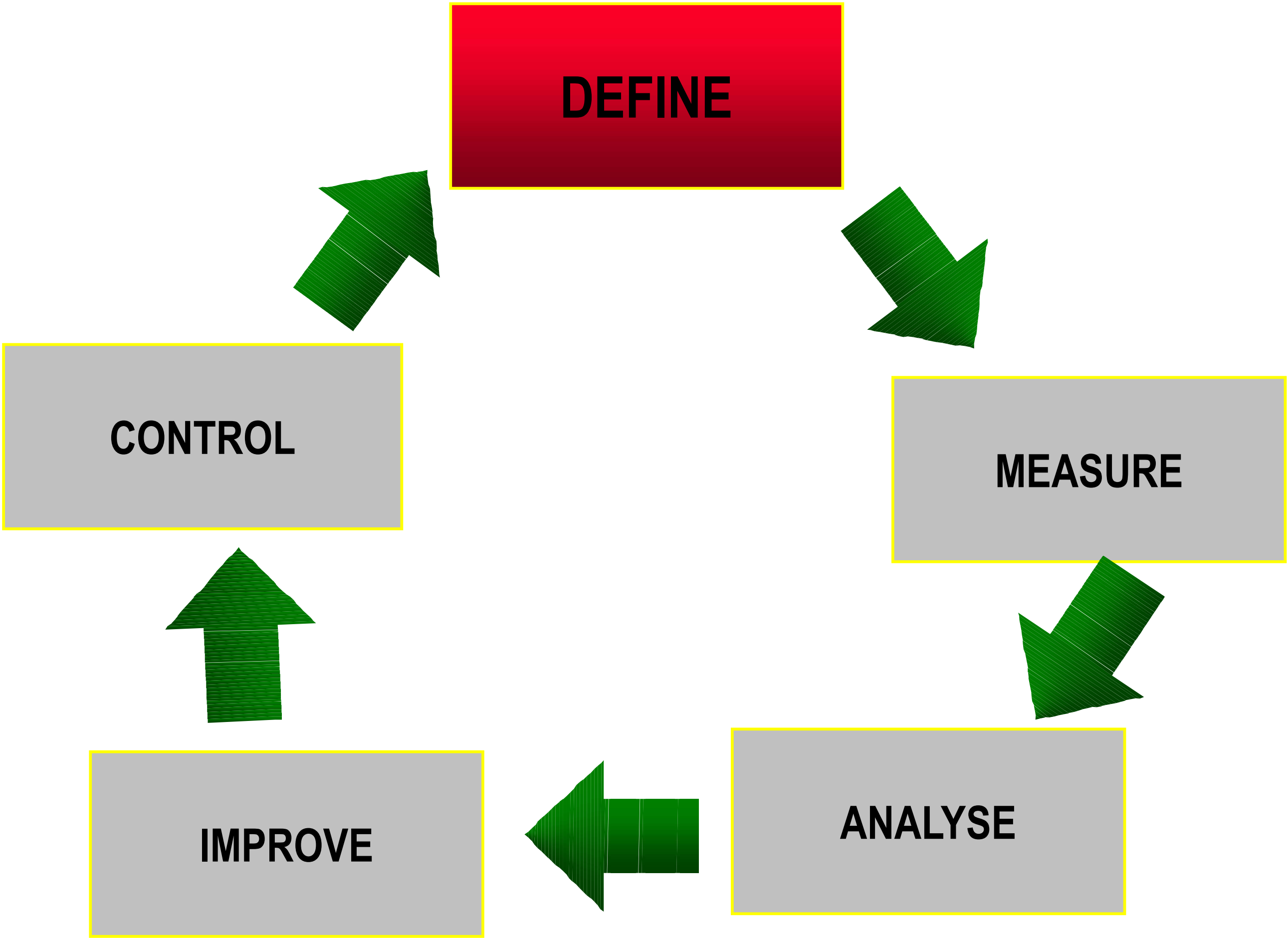


DMAIC : An Improvement Methodology

- **DEFINE**: Set direction for improvement
- **MEASURE**: Collect reliable data to understand current process performance
- **ANALYSE**: Identify problem's root causes through process and data analysis
- **IMPROVE**: Determine new improved process design
- **CONTROL**: Ensure improvement effectiveness over time

Module 2: Define Phase

DMAIC : An Improvement Methodology



DEFINE Roadmap

Define → Measure → Analyze → Improve → Control

High Level Process Map (SIPOC)

- Process Definitions
- Connecting the Customer to Your Process

Customer CTQs

- Types of customers
- Methods of collecting customer requirements
- Translate customer needs into specific requirement
- Customer requirements analysis and prioritization

Project Charter

- Business Opportunities
- Preliminary Problem Statement
- Goal statement
- Project Scope
- Milestones
- Roles

DMAIC Project Charter

Project No.:_____________

Project Name:

Process :

Resource Plan

Champion / Sponsor:
Green / Black Belt:
Functional Managers/Process Owner:
Coach / Master Black Belt:

Team Members

Text

Problem Statement

Text

Scope

Text

Goal Statement

Text

Customer CTQ's

Text

Estimate Financial Opportunities / Intangible Benefits

Text

High Level Project Milestone

Text

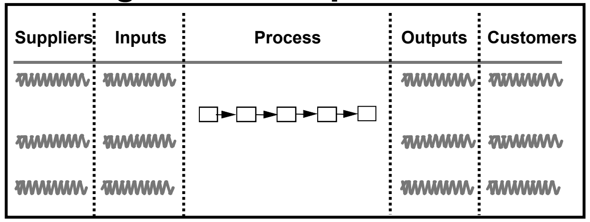
Validation

Green / Black Belt	Master Black Belt	Process Owner
CEO	Financial Analyst	Champion / Sponsor

DEFINE SUMMARY

Purpose: **To set set direction for improvement project** by developing a team charter. By defining the customers and their requirements (Critical To Quality = CTQs), mapping the high level business process to be improved.

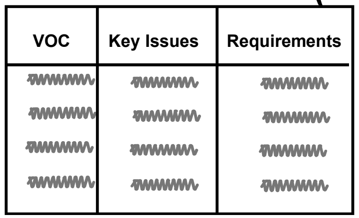
High Level Map - SIPOC



- Complete high level "as-is" process map, identifying suppliers, inputs, 5-7 high level activities, outputs & customers

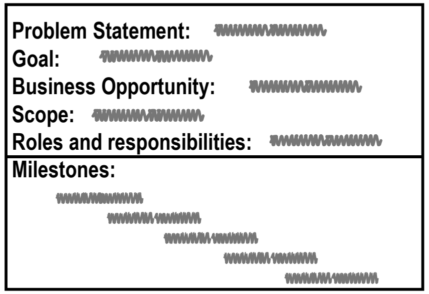
Use Survey or Focus Groups?

Voice of Customer (VOC)



- Gather and display data verifying customer requirements (CTQs)

Project Charter

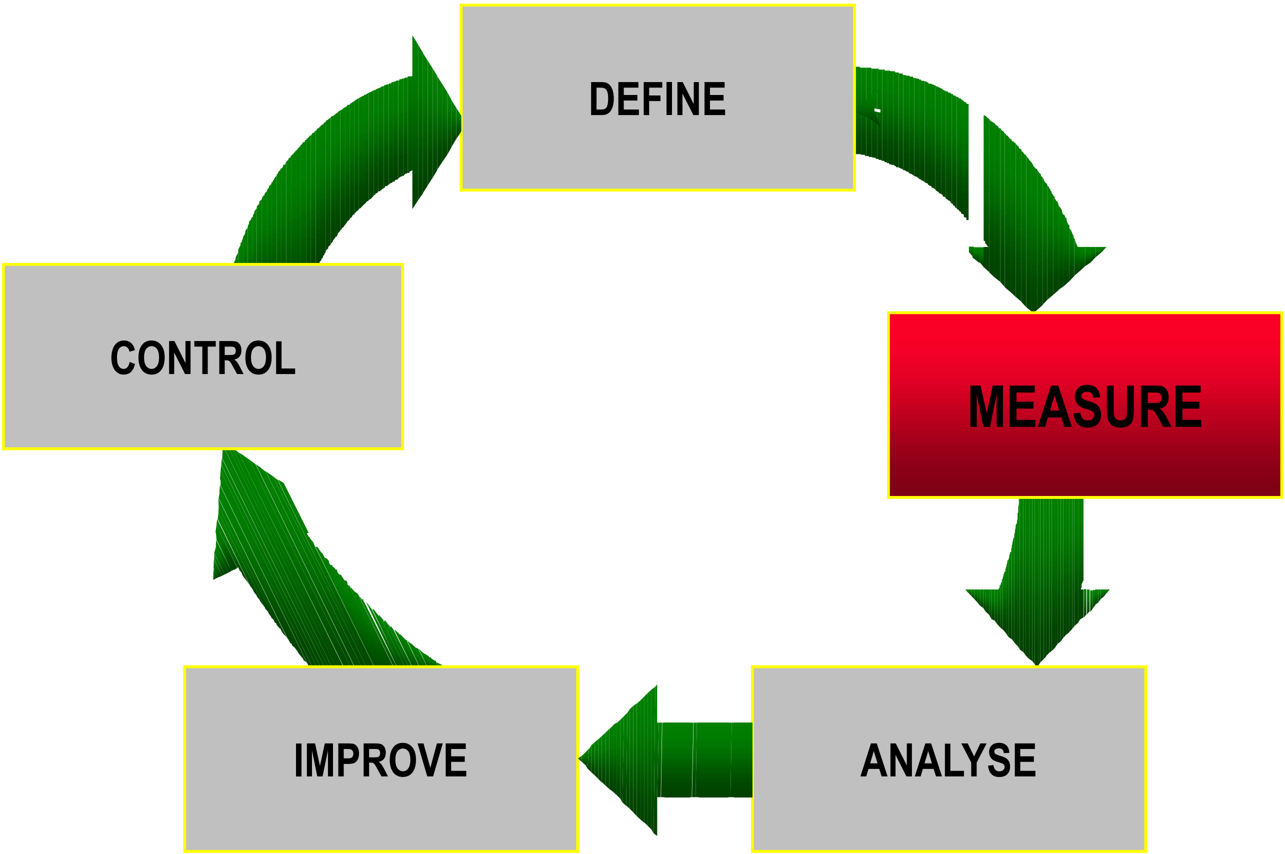


- Develop charter to include:
  - Problem statement
  - Goal for improvement
  - Business opportunity
  - Scope of project
  - Milestones for completion
  - Roles

anexas consultancy SE

Module 3: Measure Phase

DMAIC : An Improvement Methodology



Measure

Objective :

- Collect reliable data to understand current process performance

Steps :

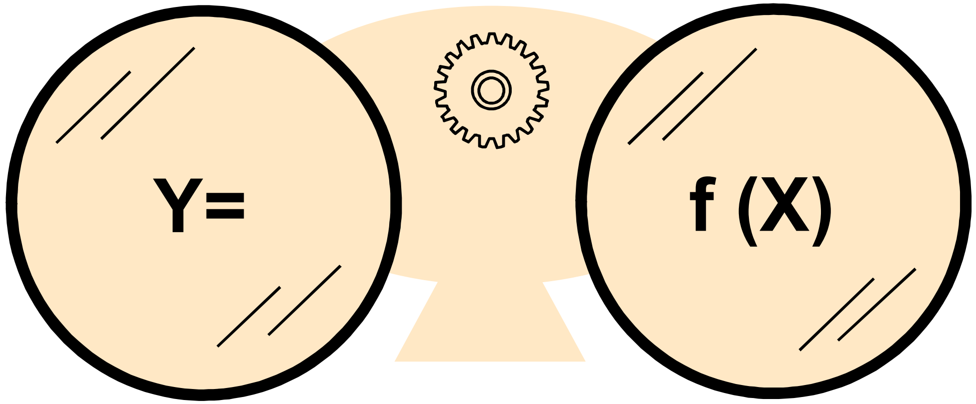
→ Choose the data to be collected (output measures, process and input measures)

→ Organize the data collection plan (What ? Why ? When? Who? How? How many ?)

→ Study process variation

→ Understand the capability of the process

Key principles for investigation



Response

- Y
- Dependent
- Output

- Effect
- Symptom
- Monitor

Predictor

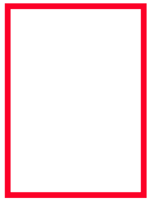
- $X_1 \ldots X_N$
- Independent
- Input-Process variables

- Cause
- Problem
- Control

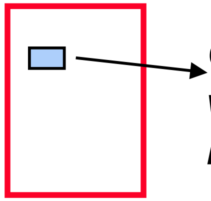
Compute Process Sigma

Key Definitions

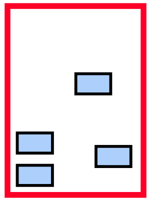
Unit: the item produced or processed



Defect: any event that does not meet the specification of a CTQ as defined by the customer



Defect opportunity: any event which can be measured that provides a chance of not meeting a customer requirement (specification)



Calculate process sigma : formula

Calculate the number of Defects Per Million Opportunities

$$\text{DPMO} = \frac{\text{(No. of Defects)}}{\text{No. Of Units x No. of opportunities}} \times 1\ 000\ 000$$

In the Sigma table, look at the Sigma value relating to the DPMO determined

Conversion Table

Long term Yield Rendement Long terme	Process Sigma Sigma du processus	Defects per 1,000,000 Défauts par 1.000.000	Long term Yield Rendement Long terme	Process Sigma Sigma du processus	Defects per 1,000,000 Défauts par 1.000.000
99.99966%	6.0	3.4	93.320%	3.0	66,800
99.9995%	5.9	5	91.920%	2.9	80,800
99.9992%	5.8	8	90.320%	2.8	96,800
99.9990%	5.7	10	88.50%	2.7	115,000
99.9980%	5.6	20	86.50%	2.6	135,000
99.9970%	5.5	30	84.20%	2.5	158,000
99.9960%	5.4	40	81.60%	2.4	184,000
99.9930%	5.3	70	78.80%	2.3	212,000
99.9900%	5.2	100	75.80%	2.2	242,000
99.9850%	5.1	150	72.60%	2.1	274,000
99.9770%	5.0	230	69.20%	2.0	308,000
99.9670%	4.9	330	65.60%	1.9	344,000
99.9520%	4.8	480	61.80%	1.8	382,000
99.9320%	4.7	680	58.00%	1.7	420,000
99.9040%	4.6	960	54.00%	1.6	460,000
99.8650%	4.5	1,350	50%	1.5	500,000
99.8140%	4.4	1,860	46%	1.4	540,000
99.7450%	4.3	2,550	43%	1.3	570,000
99.6540%	4.2	3,460	39%	1.2	610,000
99.5340%	4.1	4,660	35%	1.1	650,000
99.3790%	4.0	6,210	31%	1.0	690,000
99.1810%	3.9	8,190	28%	0.9	720,000
98.930%	3.8	10,700	25%	0.8	750,000
98.610%	3.7	13,900	22%	0.7	780,000
98.220%	3.6	17,800	19%	0.6	810,000
97.730%	3.5	22,700	16%	0.5	840,000
97.130%	3.4	28,700	14%	0.4	860,000
96.410%	3 .3	35,900	12%	0.3	880,000
95.540%	3.2	44,600	10%	0.2	900,000
94.520%	3.1	54,800	8%	0.1	920,000

anexas consultancy SE

Exercise

In plenary.

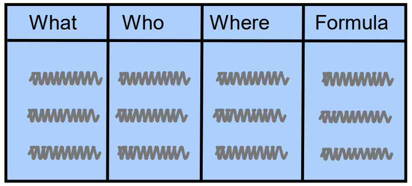
Calculate the Sigma of your process assuming the problem statement to be correct

- DPMO
- **Process Sigma** =

MEASURE

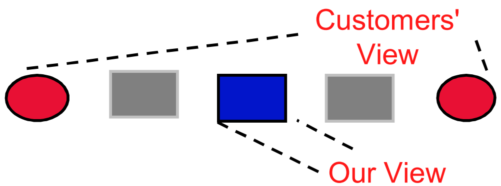
Purpose : **To measure and understand baseline performance** for the current process by collecting reliable data (quantitative & qualitative)

Data Collection



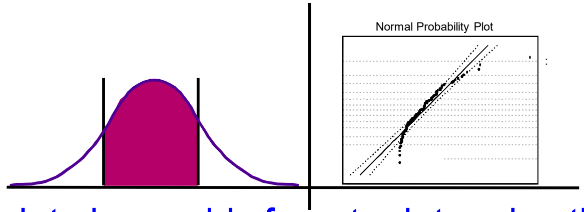
- Develop a data collection plan
  - Operational definition
  - Sampling

Customer oriented mindset



- Select the measure your customer uses to judge your performance (Key Output Measure Y)
- Plan to collect CONTINUOUS data

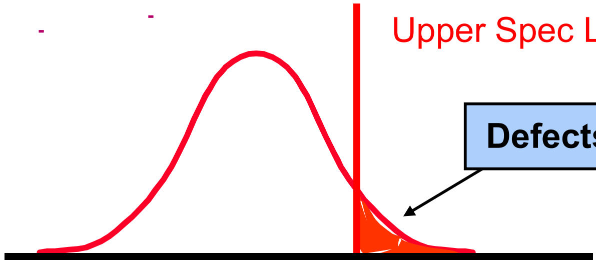
Graphical Display



- Display data in graphic form to determine the type of distribution, the metrics to understand variation and set goals for the improvement strategy.
  - Normal Distribution described by Mean and Standard deviation
  - Skewed Distribution described by Q1 (or Q3) and Inter Quartile Range
  - Long tailed distribution described by Median and Span 5-95

Calculate Process Sigma

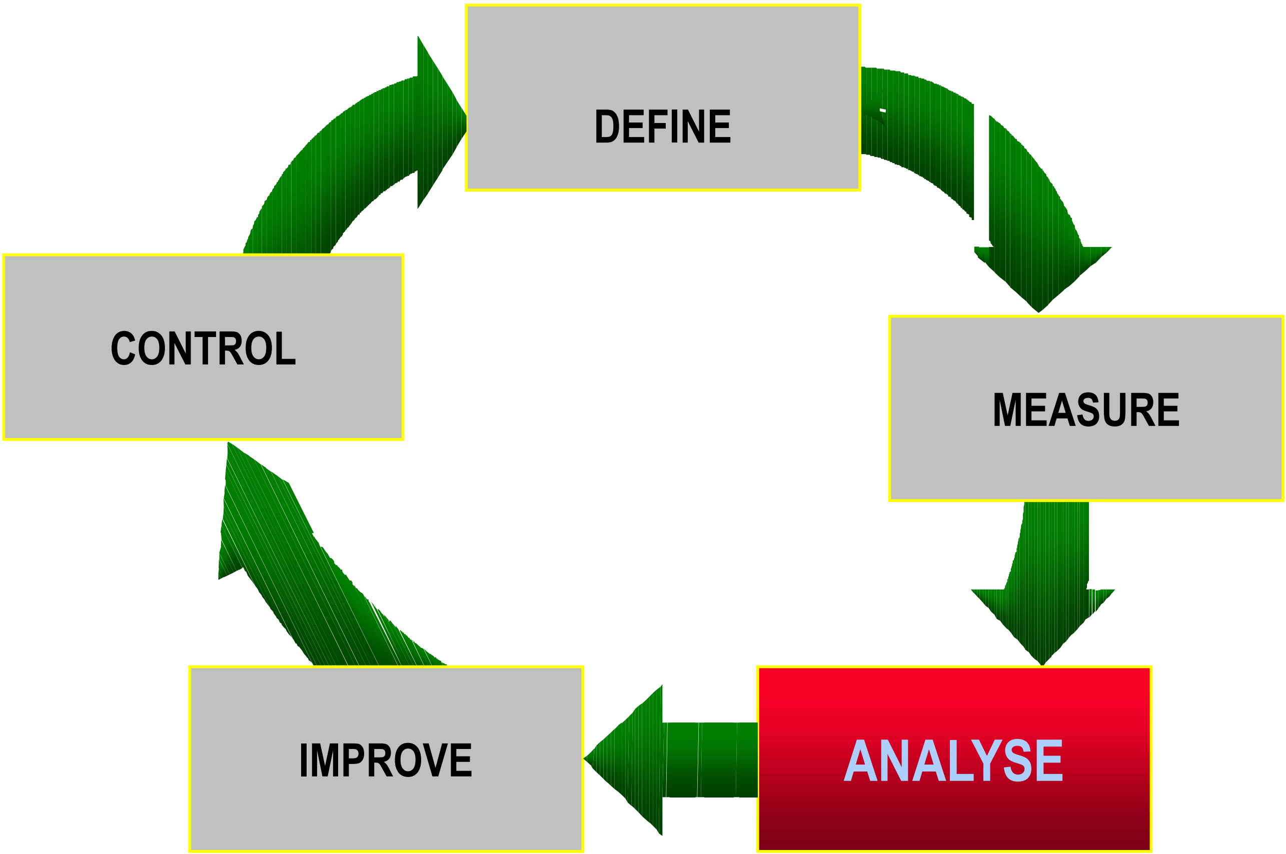
Defects "Outside" Spec Limit



- Compute baseline sigma

Module 4: Analyse Phase

DMAIC : An Improvement Methodology



Analyse Phase

Objective :

- Identify problem's root causes through process and data analysis

Steps :

- Cause and Effect Diagram
- Control Impact matrix
- Pareto chart
- Value analysis in using process map

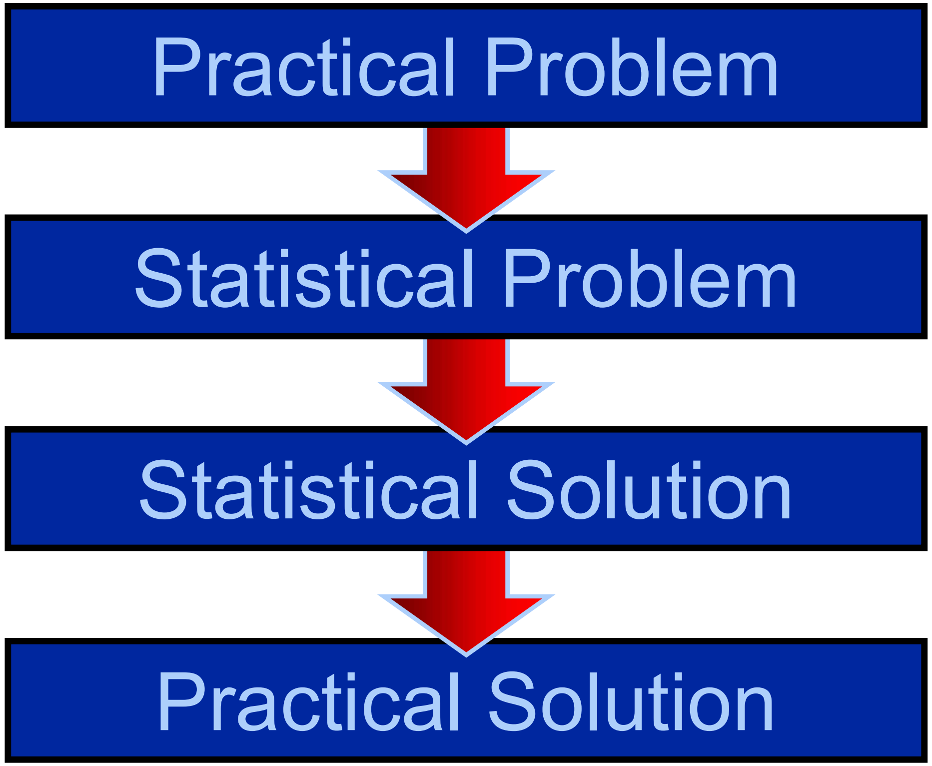
Introduction to Hypothesis Testing

Why Learn Hypothesis Testing?

- To identify sources of variability using historical or current data:
  - Passive: a process is sampled or historic sample data is obtained
  - Active: a modification is made to a process and then sample data is obtained
- Provides objective solutions to questions which are traditionally answered subjectively

What is Hypothesis Testing?

- A procedure for testing a claim about a population parameter
- Answers practical questions such as:
  - "Is there a real difference between ______ and ______ ?"



What is Hypothesis Testing?

Example:

Practical Problem

Is there a real difference between production costs of Server 1 and Server 2?

Statistical Problem

$H_o: \mu_T = \mu_A$
$H_a: \mu_T \neq \mu_A$

Statistical Solution

T-Test of difference = 0 (vs. not =): T-Value = -0.88 p-Value = 0.390 DF = 17
Fail to reject the null hypothesis

Practical Solution

There is no evidence of significant difference between production costs of Server 1 and Server 2

What is Hypothesis Testing?

- In hypothesis testing, relatively small samples are used to answer questions about population parameters (inferential statistics)
- There is always a chance that the selected sample is not representative of the population; therefore, there is always a chance that the conclusion obtained is wrong
- With some assumptions, inferential statistics allows the estimation of the probability of getting an “odd” sample and quantifies the probability (p-value) of a wrong conclusion

Parameters Versus Statistics

	Population Parameters	Sample Statistics
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	P	p

- Population parameters (values) are fixed, but unknown
- Sample statistics are used to estimate or infer population values

> Hypotheses are statements about population parameters, not sample statistics

Hypothesis Tests

Y	X	Hypothesis Test
Continuous / Variable Data	Attribute / Discrete Data	1-z, 1-t, 2-t, paired t, ANOVA
Attribute / Discrete Data	Attribute / Discrete Data	1-p, 2-p, Chi Square
Continuos / Variable Data	Continuos / Variable Data	Correlation, Regression, Multiple Regression
Attribute / Discrete Data	Continuous / Variable Data	Logistic Regression

anexas
consultancy SE

Significance Level

Goal: show observed values are so unlikely to come from the same population, that H_o must be wrong

However, even if the values are unlikely there is still a chance that they may occur. The chance they may occur is α.

This is called the significance level (α)

There is an α % chance that we are wrong when we say that Server 1 is more efficient than Server 2

α (Alpha) - Simplified Perspective

Null Hypothesis (H_o) assumed true

e.g., defendant assumed innocent

Prosecuting attorney must provide evidence beyond reasonable doubt that assumption is not true

Reasonable doubt = α

Alpha (α) & Beta (β) Risk

α-risk

- Risk of finding a difference when there really isn’t one
- Type I error or Producers’ risk

β-risk

- Risk of not finding a difference when there really is one
- Type II error or Consumers’ risk

Truth Table: α and β Risk

Decision

Truth

	Fail to reject H_o	Reject H_o
H_o true	Correct Decision CI = 1- α	Type I Error (α-Risk or *false positive)*
H_a true	Type II Error (β -Risk or *false negative)*	Correct Decision Power = 1- β

Producers' Risk

Consumers' Risk

What is p - value?

- The probability of getting sample statistics like the one we observed if our null hypothesis is true
- The chance you will be wrong if you rejected null hypothesis
- Based on an assumed or reference distribution (Z, t, F, etc.)

Decision Criteria

$p < \alpha$, reject the null hypothesis
$p > \alpha$, fail to reject the null hypothesis

Hypothesis Testing

Hypothesis Tests

Y	X	Hypothesis Test
Continuous / Variable Data	Attribute / Discrete Data	1 z, 1 t, 2 t, paired t, ANOVA
Attribute / Discrete Data	Attribute / Discrete Data	1 p, 2 p, Chi Square
Continuous / Variable Data	Continuos / Variable Data	Correlation, Regression, Multiple Regression
Attribute / Discrete Data	Continuos / Variable Data	Logistic Regression

Hypothesis Testing of Mean

Steps

Steps in Hypothesis tests:

1. State the null hypothesis (Ho)

   Null Hypothesis is:

   All means are equal (1-z, 1-t, 2-t, paired t, ANOVA) [Cont- Att]

   Y is independent of X (Regression) [Cont –Cont]

   Y is not related to X (1 p, 2p, Chi Square) [Att-Att]

   Y is not related to X (Binary Logistic Regression) [Att-Cont]

2. State the alternative hypothesis (Ha)

   At least one mean is different(1-z, 1-t, 2-t, paired t, ANOVA)

   Y is dependent on X (Regression)

   Y is related to X (1 p, 2p, Chi Square)

   Y is related to X (Binary Logistic Regression) [Att-Cont]

3. Choose alpha value (α = .05). Also known as level of significance. Confidence Level = 1-α

4. Collect data

Steps

Steps in Hypothesis tests:

5. Choose appropriate hypothesis test

6. Get p value

7. If p is < 0.05 , Reject Ho

   If p is > 0.05, Accept Ho

Remember :

If p is low Ho must go

If p is high, Ho must fly

Why Learn Hypothesis Tests of Mean?

- Make data driven decisions with defined confidence
- Determine if a statistically significant difference of means exists between:
    - A sample and a target
    - Two independent samples
    - Paired samples

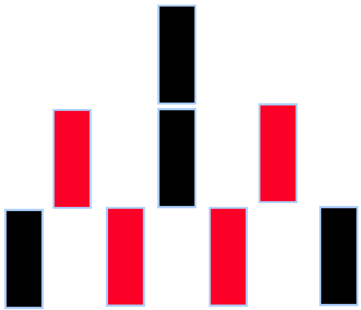
What are Hypothesis Tests of Mean?

Test	Method for analyzing the differences between:
1 Sample Z	a sample mean and a target value when population standard deviation is known
1 Sample t	a sample mean and a target value when population standard deviation is not known
2 Sample t	means obtained from two independent samples
Paired t	mean differences obtained from paired samples

Note: Above tests are used when the dependent variable (response) is continuous and the independent variable (factor) is discrete

1 Sample Z Test

Single Mean Comparison



μ vs target value

σ known

Practical Question
(example)

"Is the population statistically different from the target value?"

Statistical Question
H_o: μ = target value
H_a: $\mu \neq$ target value

anexas consultancy SE

Business Process Example: Rising Transaction Costs

A financial institution is concerned about rising costs per teller transaction. Leadership of the institution wants to take appropriate action if the population average cost per teller transaction is greater than $1.40.

A random sample of 45 costs per teller transaction produced an average value of $1.45. It is known from previous experience that the population standard deviation of the transaction cost is approximately $0.32.

Analyze the sample data from the file Tellercost.mtw and determine if we have evidence to show that the population mean cost per teller transaction cost is greater than $1.40.

Example: Rising Transaction Costs

- Practical Problem
  - Did the average cost per transaction increase?
  - Is the average cost per transaction greater than $1.40?
- Statistical Problem
  - Is there a shift in the mean cost per transaction from the historical average?
  - Null hypothesis: Average cost is $1.40
  - Alternate hypothesis: Average cost is greater than $1.40
  - Is there evidence (at a significance level of 5%) to show that the average cost per transaction has increased? Otherwise we maintain the current belief - i.e., the null hypothesis

Example: Rising Transaction Costs

- State the hypotheses and significance level

H_o: μ = \$1.40

H_a: μ > \$1.40

α = 0.05

- What hypothesis test is appropriate?

These hypotheses deal with mean values

Only one factor for examination – rising transaction cost

Comparing population mean against a target value using one sample data

Data follows a normal distribution

σ known, \$0.32

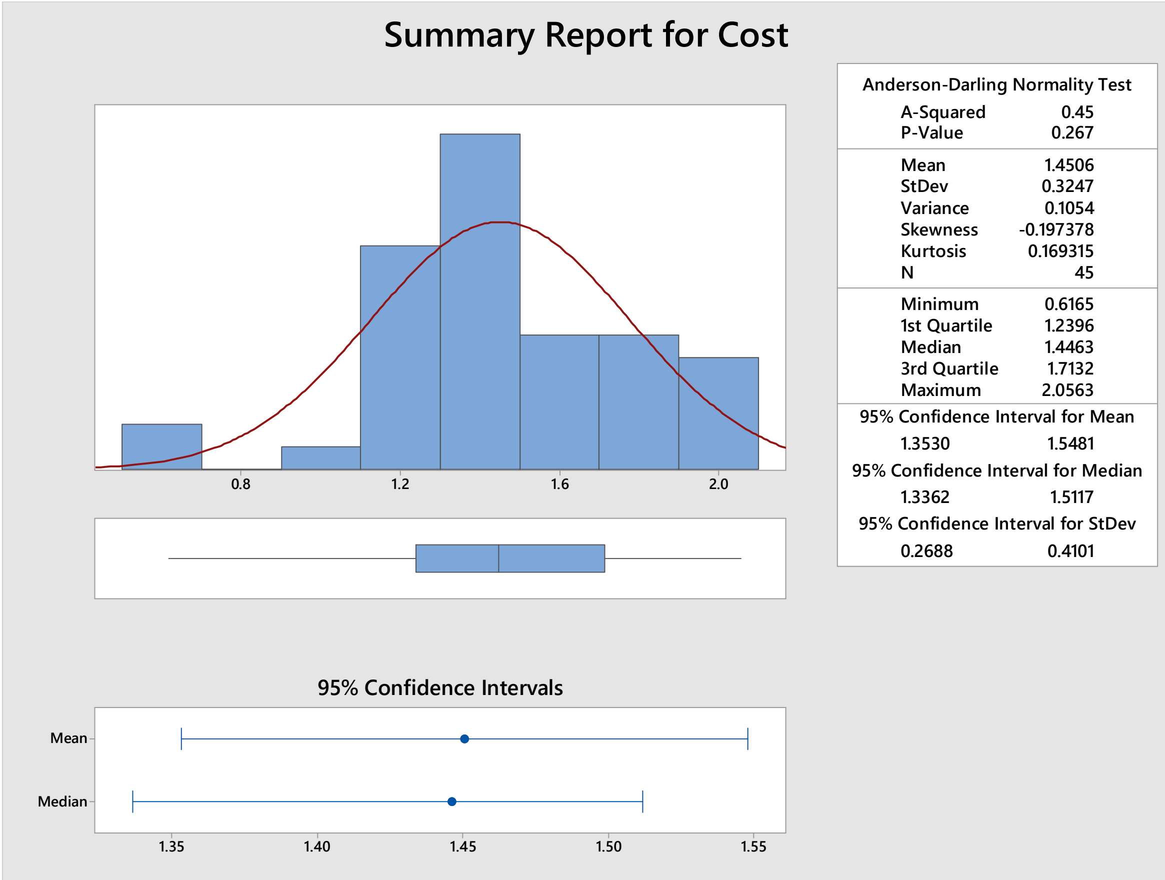
Use 1-Sample Z-Test

Example: Rising Packing Costs

Practical and Graphical

- Open the file ***Tellercost.mtw***
- What practical questions do you have about this data?
- Evaluate descriptive statistics

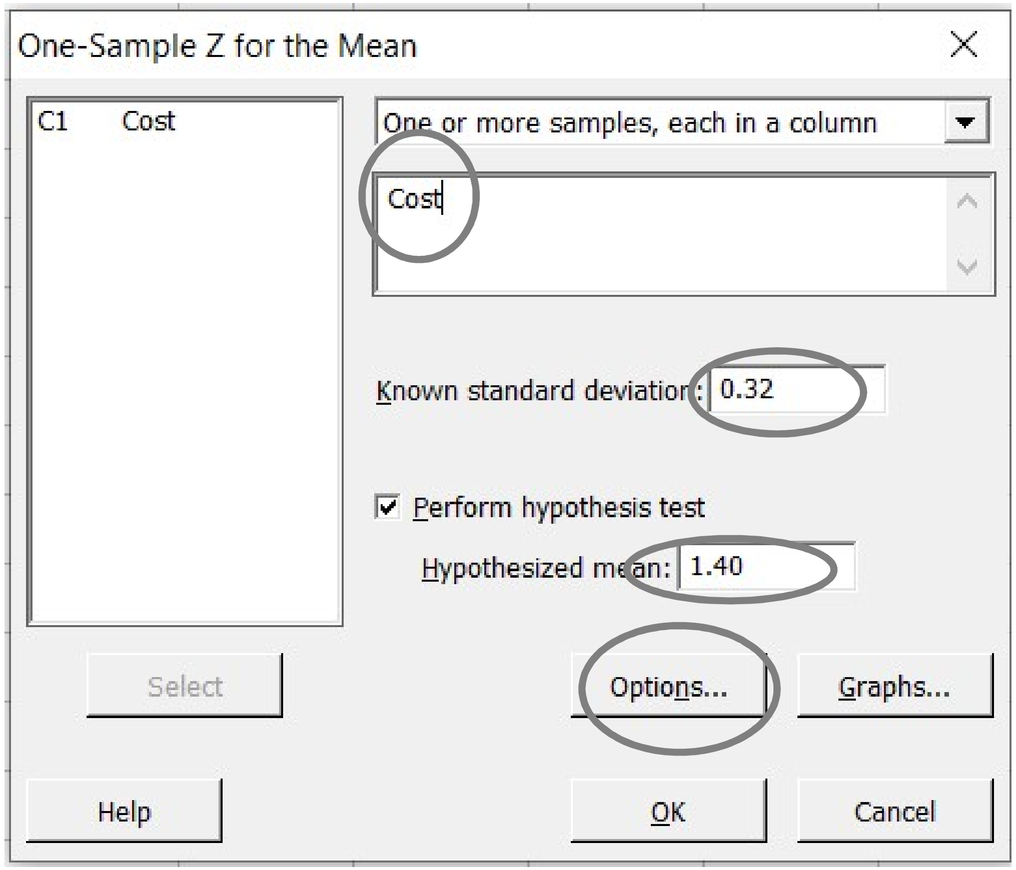
Descriptive Statistics



Example: Rising Packing Costs

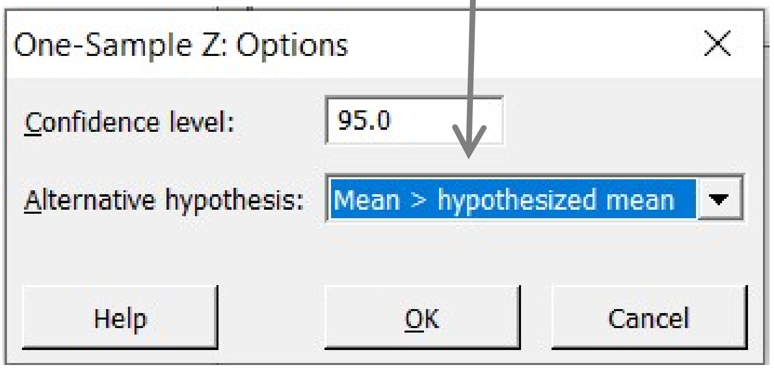
Tool Bar Menu > Stat > Basic Statistics > 1-Sample Z

Analysis through Minitab



Choose *greater than* for Alternative

(H_a: μ > \$1.40)



Example: Rising Transaction Costs

One-Sample Z: Cost

```
Test of mu = 1.4 vs mu > 1.4
The assumed sigma = 0.32

Variable            N       Mean      StDev    SE Mean
Cost               45     1.4506     0.3247     0.0477

Variable        95.0% Lower Bound         Z       P
Cost                       1.3721      1.06   0.145
```

Interpretation:

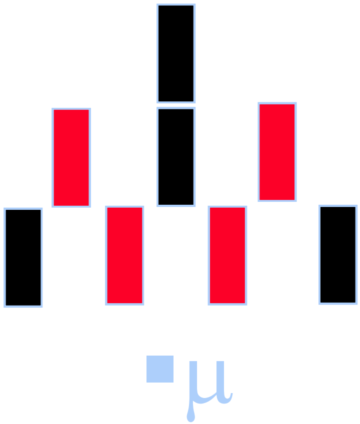
p-value = 0.145

Since p-value > α-value (0.05) fail to reject H_o

Infer H_o true: not enough evidence that average teller transaction cost is greater than $1.40

1 Sample t Test

Single Mean Comparison



vs.

target value

σ NOT known

Practical Question (example)

- *"Is the population*
- *statistically greater than*
- *the target value?"*

Statistical Question

H_o: μ = target value

H_a: μ > target value

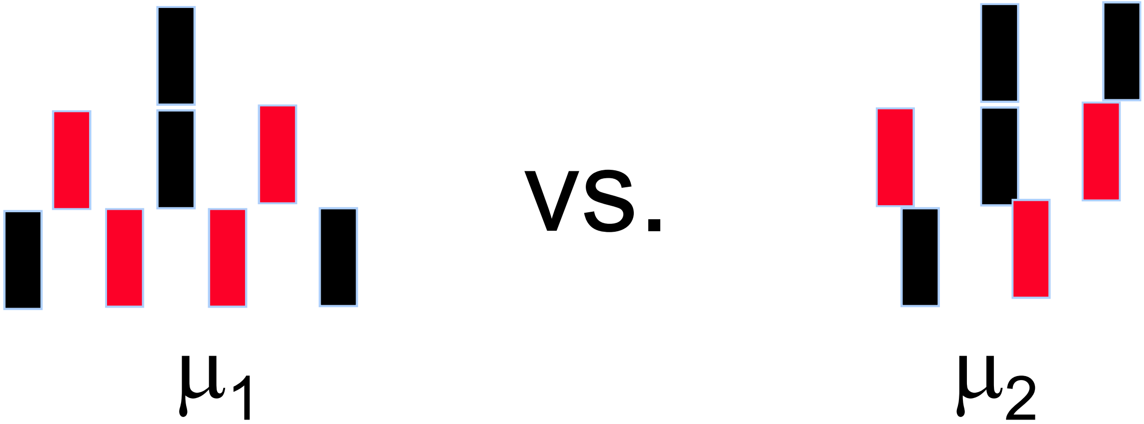
1-Sample t-Test

- Hypothesis test about the unknown population mean using information from one sample
- Population standard deviation not known and distribution is normal

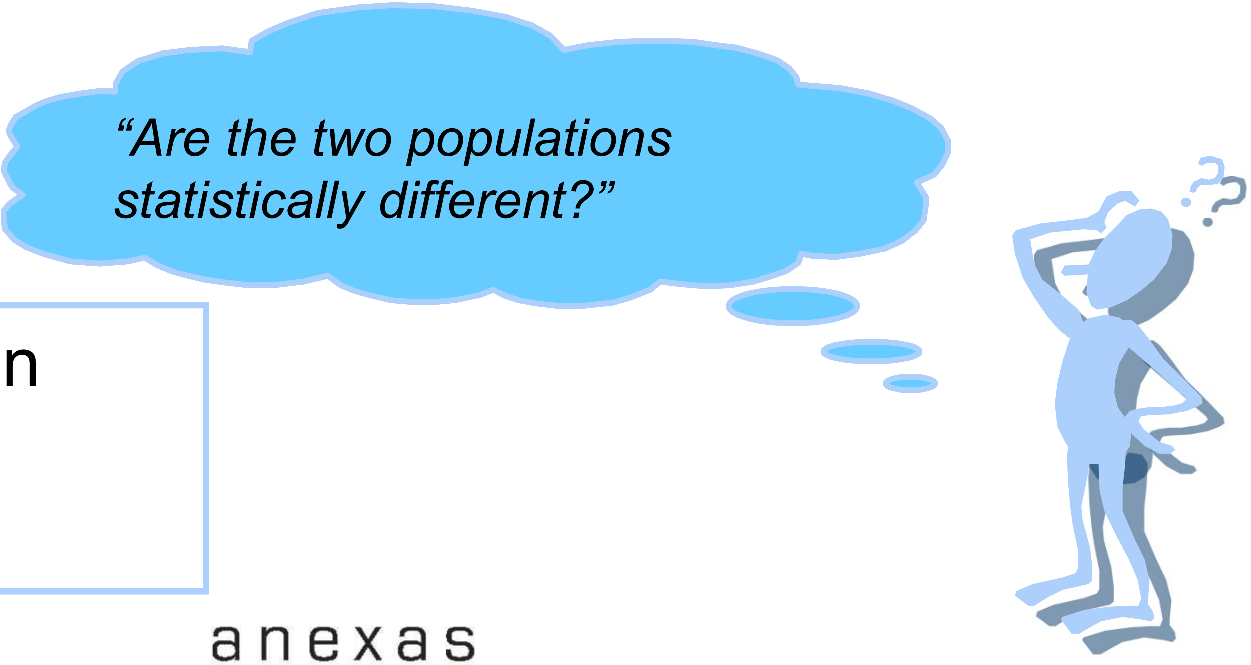
Note: Normality assumptions relaxed when the number of sample observations is large (generally true when sample size >30).

2 Sample t Test

Two Sample Comparison



Practical Question
(example)



Statistical Question

H_o: $\mu_1 = \mu_2$

H_α: $\mu_1 \neq \mu_2$

2-Sample t-Test

- Hypothesis test about the difference between two population means using two samples
- Distributions are normal
- Two independent samples
  - Can be of different size

Business Process Example:
Teller vs. ATM Costs

As part of an investigation to study the transaction costs of tellers versus ATMs, a bank has collected a random sample of 45 teller transaction costs and 53 ATM transaction costs.

The data is given in file ***ATMTeller.mtw.***

Perform a hypothesis test to determine if average value teller transaction cost is higher than ATM transaction costs by at least $0.35.

Example: Teller vs. ATM Costs

- Practical problem
  - –Is average cost of teller transactions higher than average cost of ATM transactions by at least $0.35?
- Statistical problem
  - –Is the population mean for teller transaction cost higher than the population mean of ATM transaction costs by at least $0.35?
  - –Null hypothesis: difference between mean value of teller transaction costs and mean value of ATM transaction costs is equal to $0.35
  - –Alternate hypothesis: difference between mean value of teller transaction costs and mean value of ATM transaction costs is greater than $0.35

Example: Teller vs. ATM Costs

- State the hypotheses and significance level

$$H_o: \mu_{Teller} - \mu_{ATM} = \$0.35$$

$$H_a: \mu_{Teller} - \mu_{ATM} > \$0.35$$

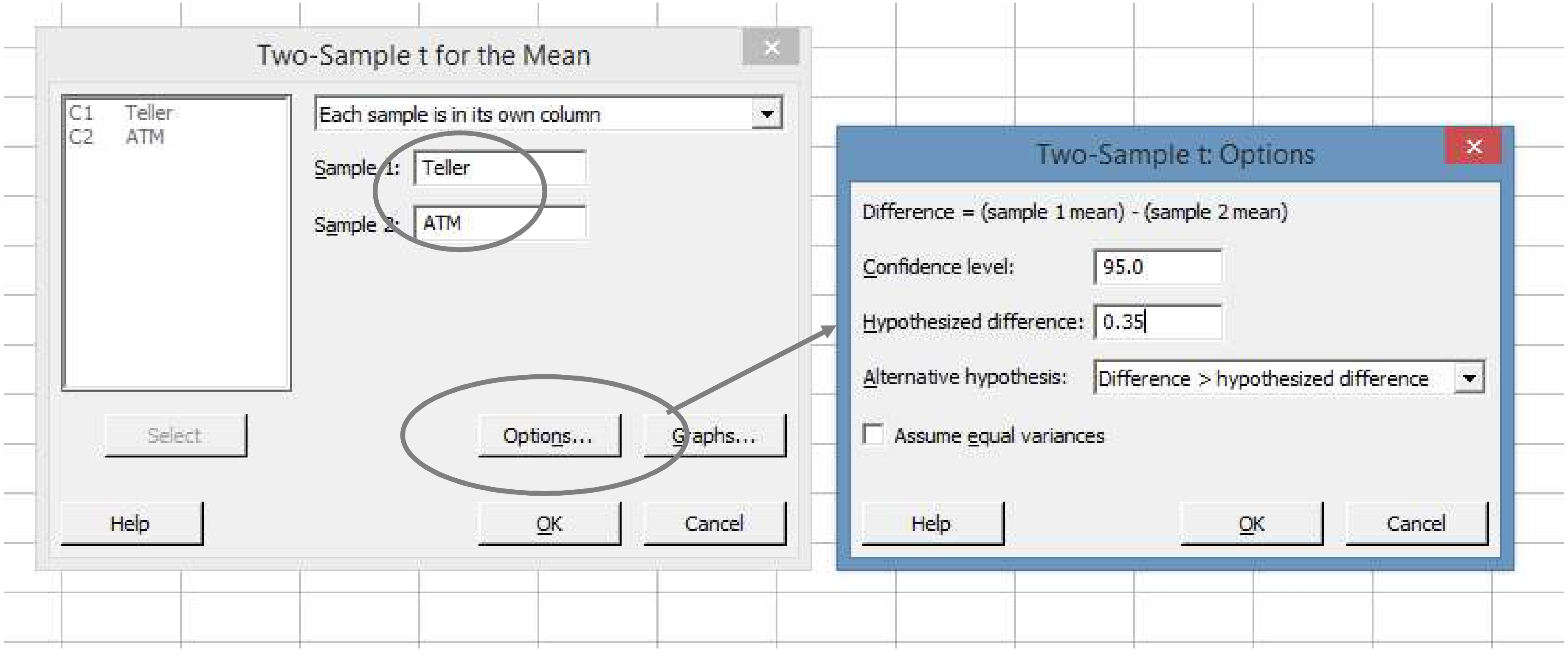
$$\alpha = 0.05$$

- What hypothesis test is appropriate?
  - These hypotheses deal with mean values
  - Only one factor for examination - transaction cost
  - Comparing population means based on two independent sets of sample data
  - Samples are normally distributed
  - Use 2-Sample t-Test

Example: Teller vs. ATM Costs

Tool Bar Menu > Stat > Basic Statistics > 2-Sample t

Analysis through Minitab



Example: Teller vs. ATM Costs

```
Two-sample T for Teller vs ATM

          N      Mean      StDev    SE Mean
Teller   45      1.451     0.325      0.048
ATM      53      0.985     0.210      0.029

Difference = mu Teller - mu ATM
Estimate for difference:  0.4654
95% lower bound for difference: 0.3716
T-Test of difference = 0.35 (vs >): T-Value = 2.05  P-Value = 0.022  DF = 72
```

Interpretation:

-p-value 0.022

-Since p-value < α-risk (0.05), reject the null hypothesis

-The difference between Teller cost and ATM costs is greater than $0.35

ANOVA
Analysis of Variance

Why Learn ANOVA?

ANOVA

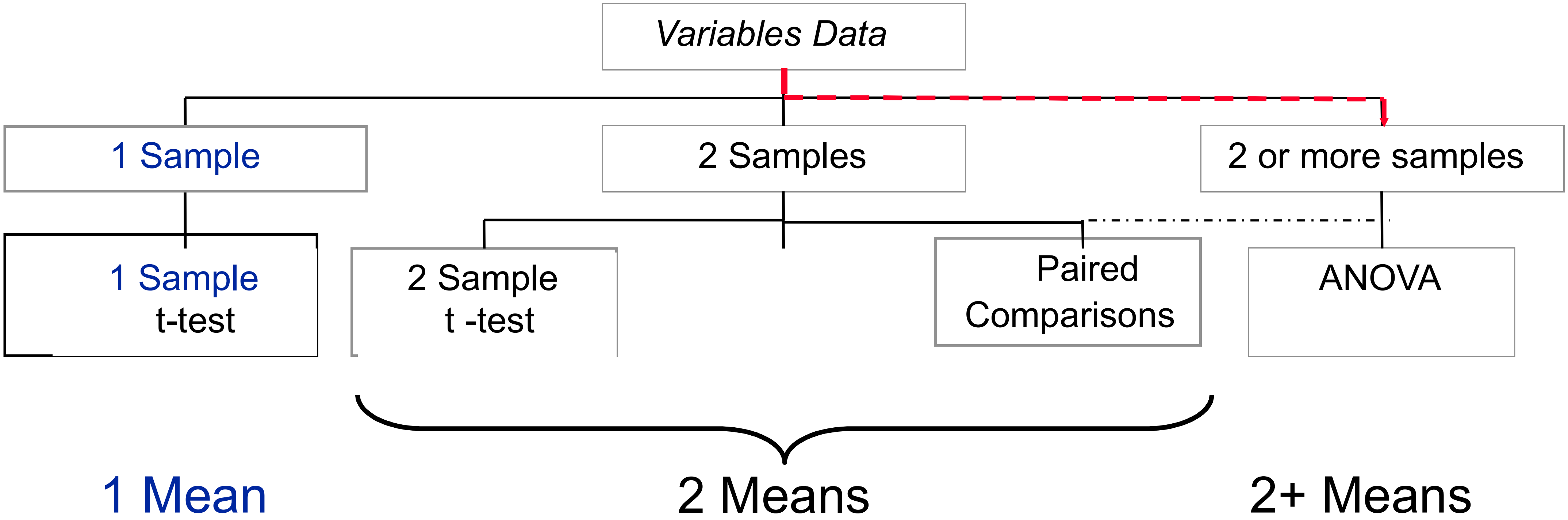
- Performs hypothesis testing for two or more means
- Evaluates several PIVs
- Handles multiple levels
- Shows sources of process variation
- Generates an underlying variability estimate

What is ANOVA?

- Hypothesis Test for MEANS
  - Uses two components of variance
    - within variance (no change)
    - between variance (after a change)
- Uses the F-distribution to test the variance components
- Comprehensive test for significance
- Backbone test statistic for subsequent complex analysis

When to Use ANOVA

Variables Road Map



ANOVA is used to test two or more means

anexas consultancy SE

Process Variation

- All processes are influenced by other factors
- Is variation due to a real factor effect or are the differences just random variation?
- t-tests are tools that offer some help, but are limited to testing two means
- Finding factors that are sources of variation are key to process improvement

▪ANOVA allows concurrent testing of several means

ANOVA in Minitab™

Setting Up the Data in Minitab™

Open worksheet ***VENDOR YIELD.MTW***

Vendor Yield.MTW ***

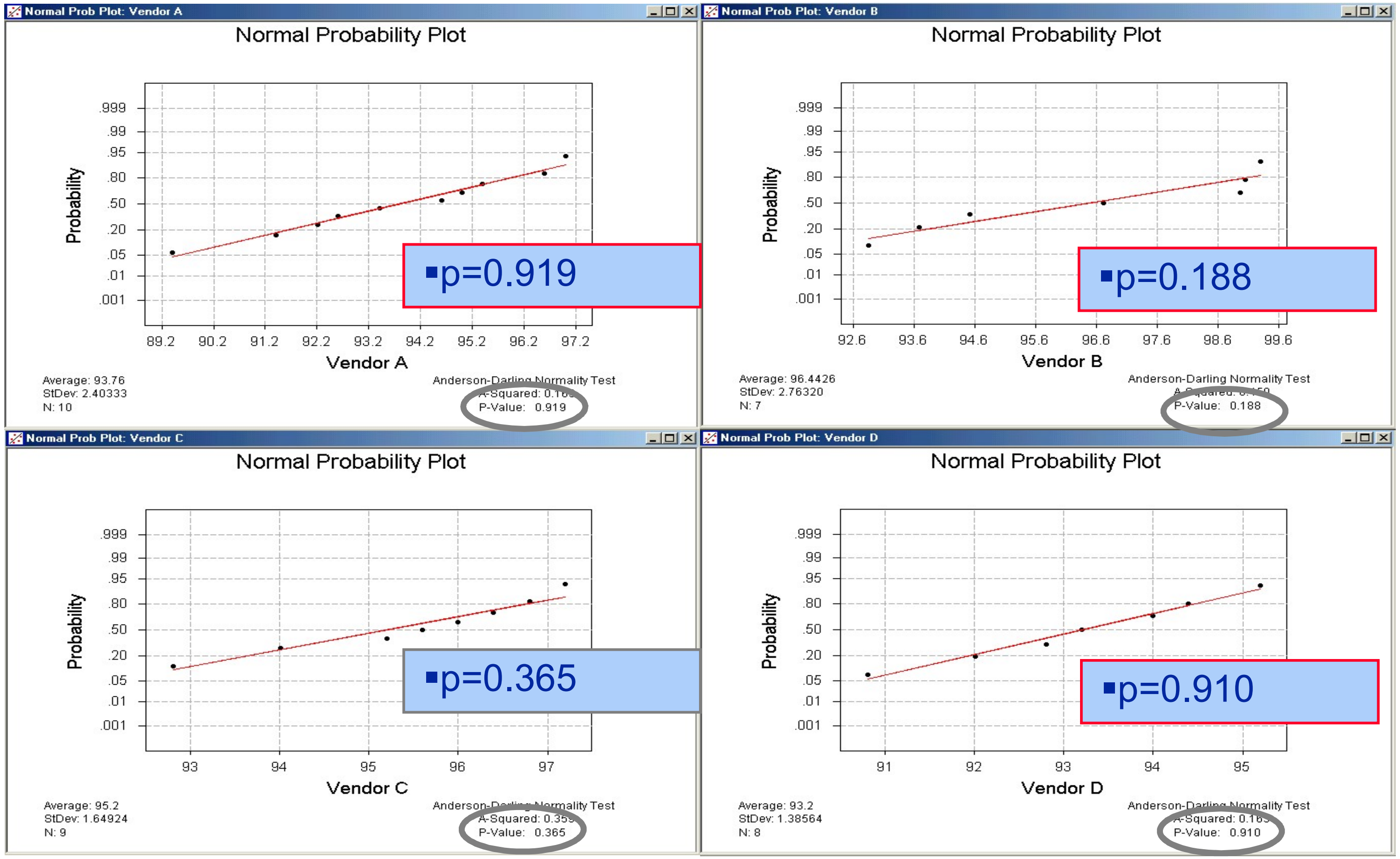
↓	C1	C2	C3	C4	C5-T	C6
	Vendor A	Vendor B	Vendor C	Vendor D	Vendor	Yield
1	91.4	99.3	92.8	94.4	Vendor A	91.4000
2	94.6	93.7	96.4	92.8	Vendor A	94.6000
3	92.6	99.1	96.0	90.8	Vendor A	92.6000
4	95.0	99.0	94.0	93.2	Vendor A	95.0000
5	92.2	92.8	92.8	95.2	Vendor A	92.2000
6	97.0	96.7	95.6	93.2	Vendor A	97.0000
7	89.4	94.5	96.8	92.0	Vendor A	89.4000
8	95.4		97.2	94.0	Vendor A	95.4000
9	93.4		95.2		Vendor A	93.4000
10	96.6				Vendor A	96.6000

Sorted data is in columns C1-C4. Stacked data is in column C5-T and C6

Prerequisites for ANOVA

- Every subgroup has a normal distribution
- Subgroups have statistically equal variances
- Residuals are independent and normally distributed about the mean

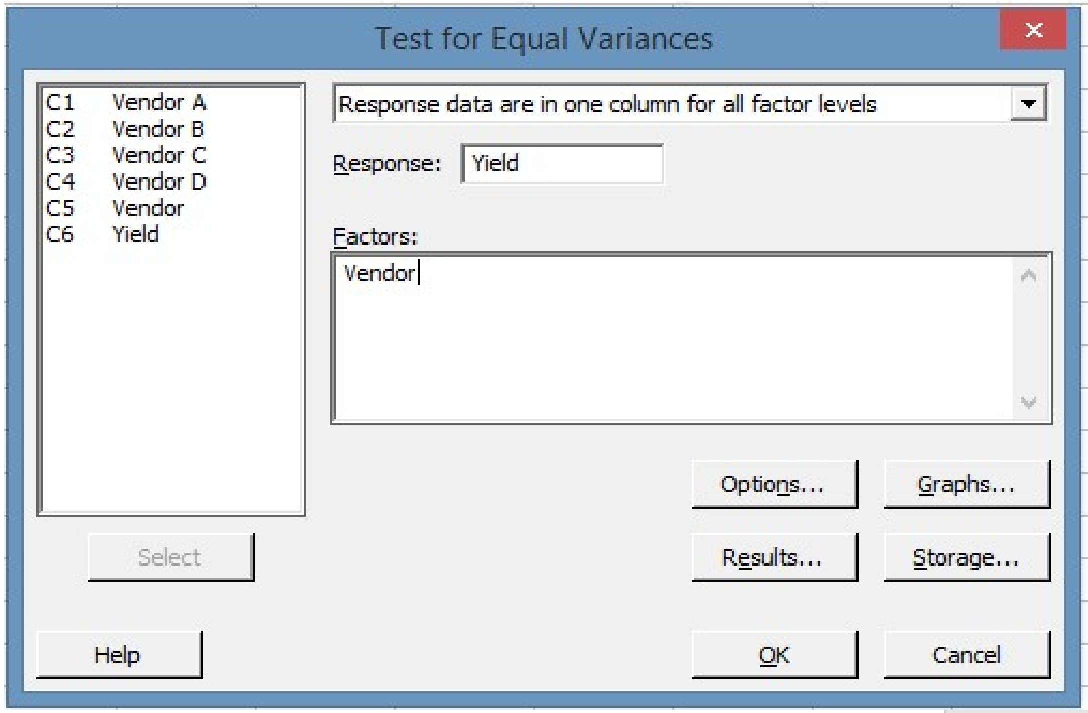
Testing Data for Normality

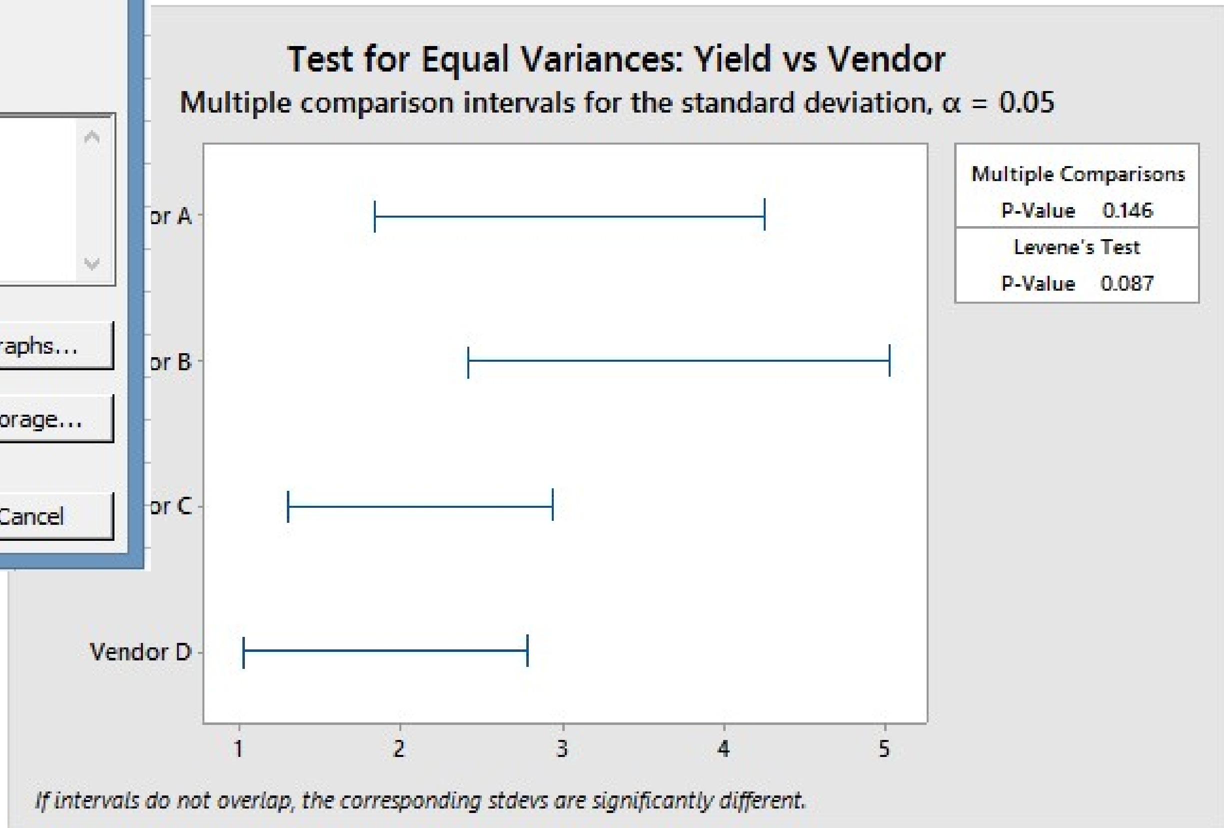


All subgroups have a normal distribution

Testing Data for Equal Variances

Tool Bar Menu > Stat > ANOVA > Test for Equal Variances

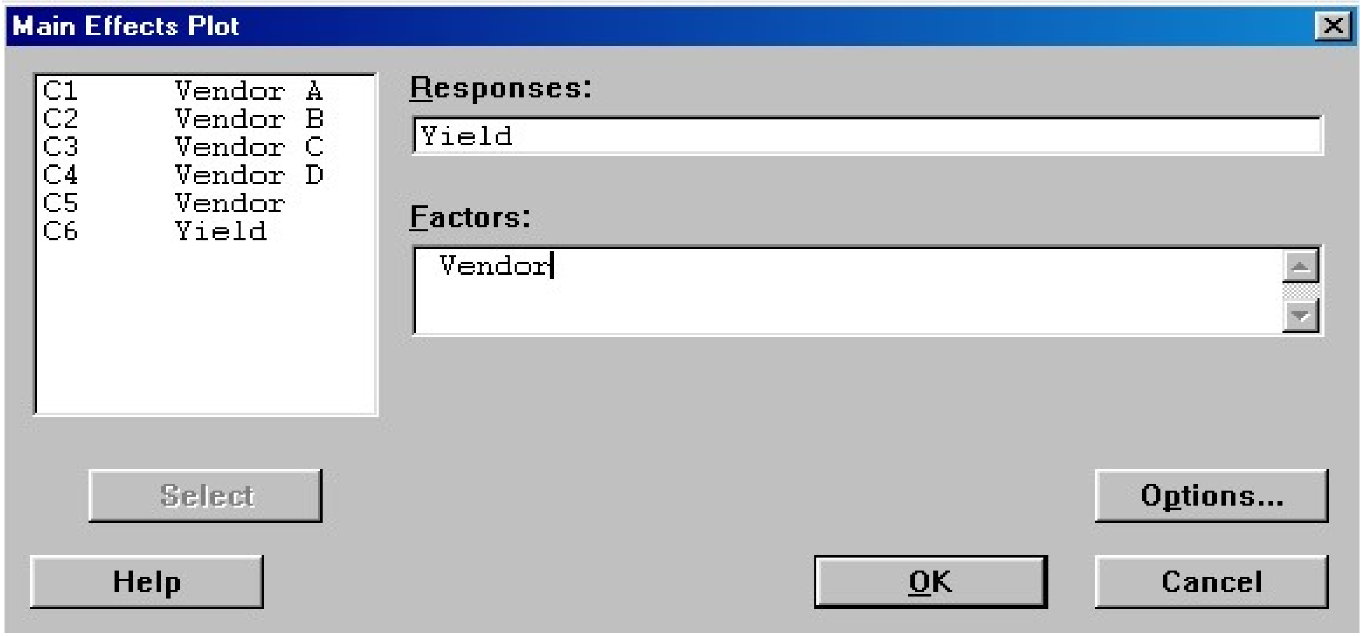




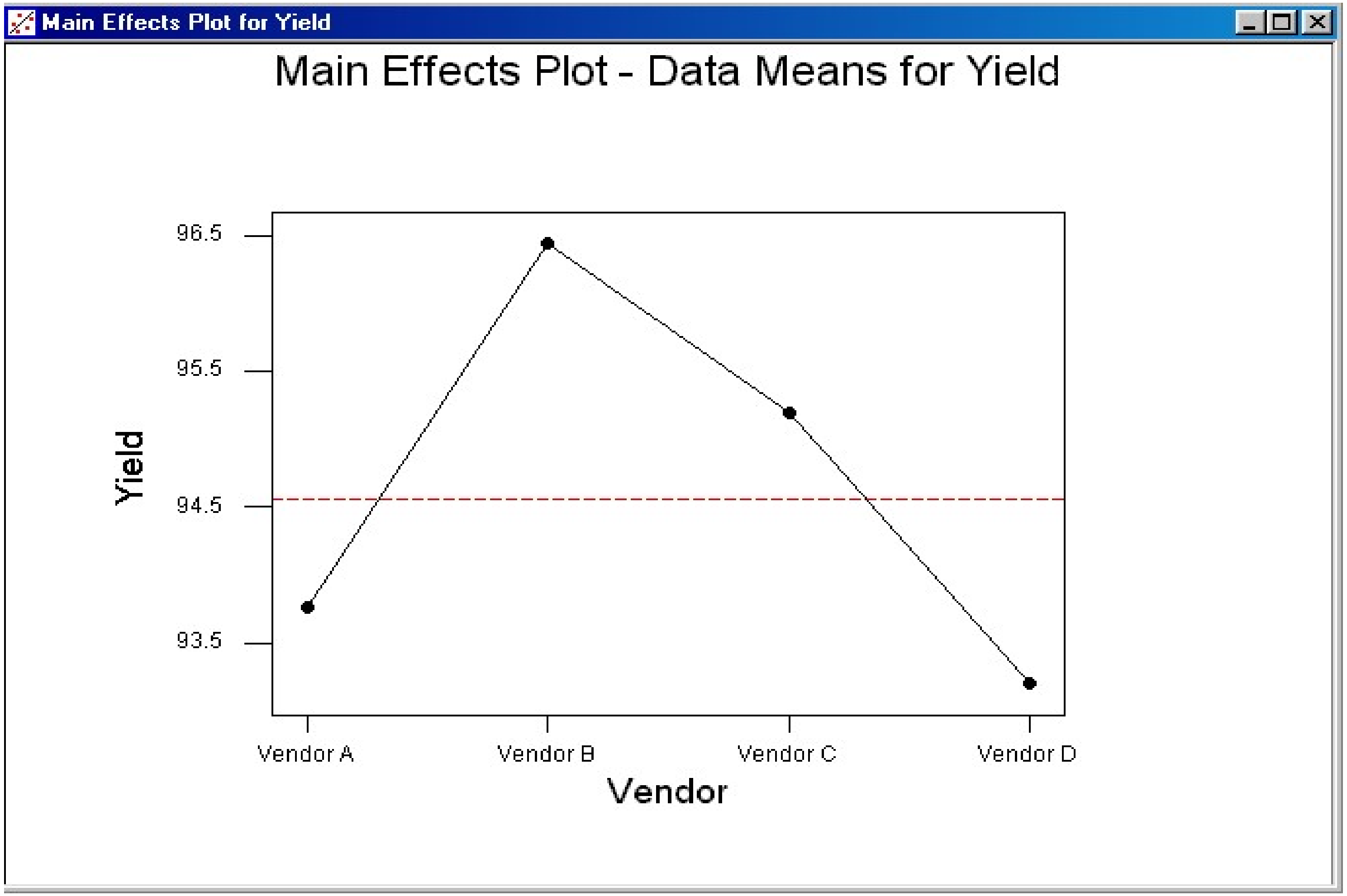
P=0.146 indicates variances are equal

Running a Main Effects Plot

Tool Bar Menu > Stat > ANOVA > Main Effects Plot



The Main Effects Plot

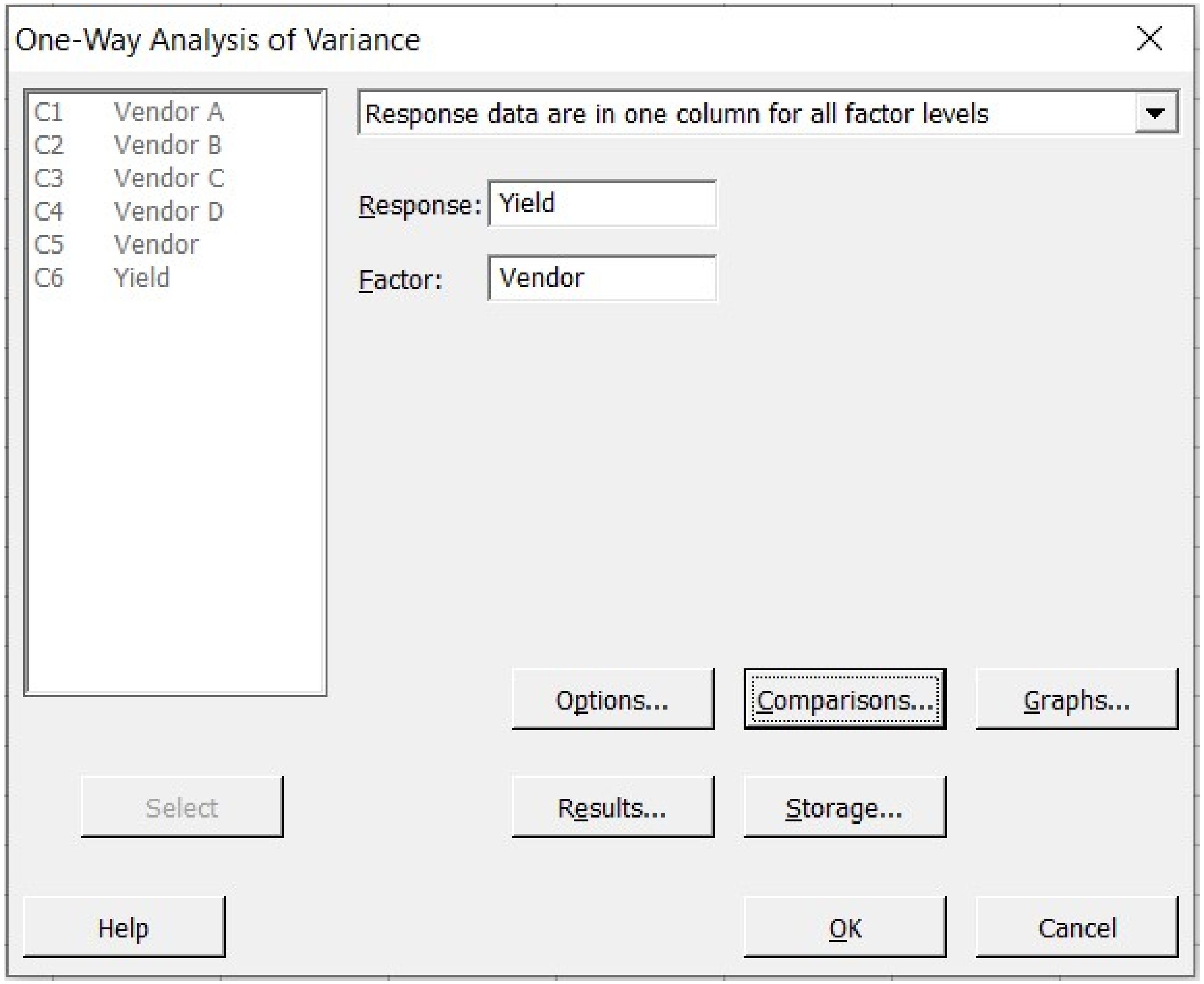


- The plot shows the output vs. the factor

anexas consultancy SE

Running a One Way ANOVA

Tool Bar Menu > Stat > ANOVA > One Way…



The One Way ANOVA

One-way ANOVA: Yield versus Vendor

Method

Null hypothesis	All means are equal
Alternative hypothesis	At least one mean is different
Significance level	α = 0.05

Equal variances were assumed for the analysis.

Factor Information

Factor	Levels	Values
Vendor	4	Vendor A, Vendor B, Vendor C, Vendor D

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Vendor	3	49.69	16.564	3.74	0.022
Error	30	133.00	4.433		
Total	33	182.69			

- $p < 0.05$: source is significant!

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.10551	27.20%	19.92%	6.05%

Means

Vendor	N	Mean	StDev	95% CI
Vendor A	10	93.760	2.403	(92.400, 95.120)
Vendor B	7	96.44	2.76	(94.82, 98.07)
Vendor C	9	95.200	1.649	(93.767, 96.633)
Vendor D	8	93.200	1.386	(91.680, 94.720)

Pooled StDev = 2.10551

Tests of Significance (Non- Parametric)

Tests of Significance (Non- Parametric)

	Test of Mean/ Median	Test of Variance	Test of Proportion
1 Sample	▪ 1 sample Z Test ▪ 1 Sample t Test ▪ 1 sample Sign ▪ 1 Sample Wilcoxon	Descriptive Statistics Bartlett's test Levene's test	1 Proportion Test
2 Sample	▪ 2 Sample t Test ▪ Mann-Whitney ▪ Paired T Test		2 Proportion Test
2 or more Samples	▪ ANOVA ▪ Mood's Median/ ▪ Kruskal Wallis Test		Chi Square

1 Sample Sign Test - Overview

- 1 sample sign tests allow you to compare the median of just one sample against a known median value, such as an industry benchmark or well established historical mean.

Example:-

- A recruitment consultancy has recently implemented a new salary negotiation process and a project team is trying to verify that it has improved (increased) the salaries that are being achieved. The salaries of the first 20 placements made using the new negotiation process have been recorded and the project team want to compare these results against the benchmark.

1 Sample Sign Test

Because the sample of salaries is not Normally distributed and there is only one sample to compare against the benchmark.

The median salary of the sample is 60K and the historical salary median was 48K, and so it appears that there has been an increase in the median of 12K. however its quite a small sample and there is lots of variation within the sample and it is difficult to be sure.

The data file for this example can be found in ***Salaries.mpj***

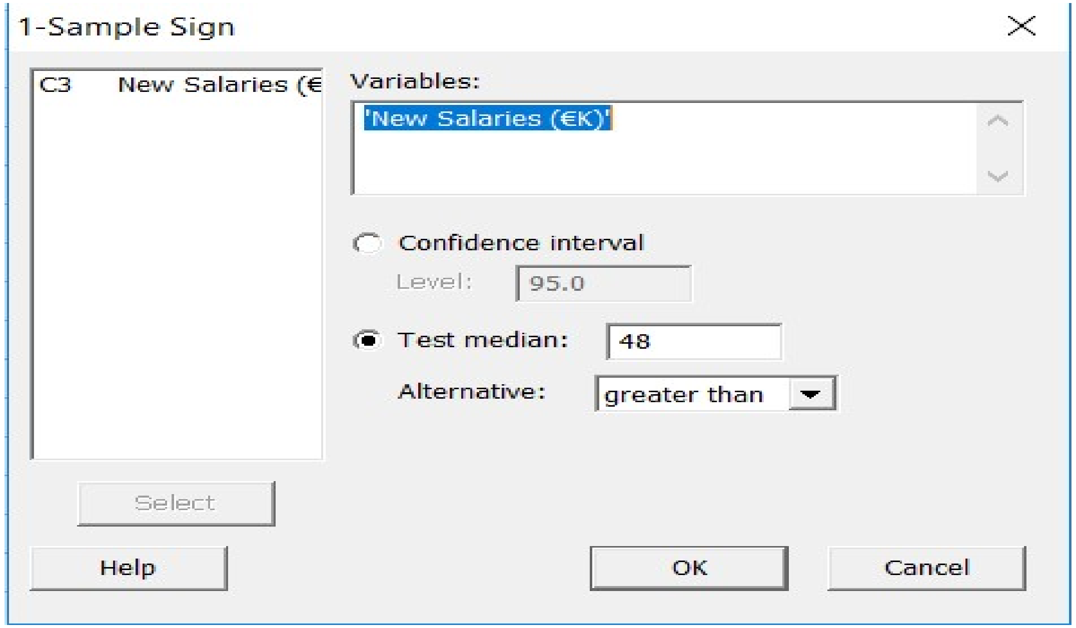
In this example, the sample median is being tested against a historical benchmark (the test median). Enter 48 (the historical benchmark level)

Alternate Hypothesis: Less than / Not equal / Greater than):

In this example because it appears that the sample median is greater than test median, the alternative greater than has been selected.

1 Sample Sign Test - Interpreting the Results

Stat > Nonparametrics > 1-Sample Sign



Sign Test for Median: New Salaries (€K)

```
Sign test of median =  48.00 versus > 48.00

                    N  Below  Equal  Above       P  Median
New Salaries (€K)  20      5      1     14  0.0318   60.00
```

Session Window Results:

The first line provides a summery of the Null and Alternate Hypothesis, in a technical shorthand. Translated into day-to-day language, this line says: These are the results of a Sign Test for the sample median being equal to 48 (median = 48.00) versus it being greater than 48 (versus > 48.00)

anexas consultancy SE

1 Sample Sign Test – Interpreting the Results

Sign Test for Median: New Salaries (€K)

```
Sign test of median =  48.00 versus > 48.00

                      N  Below  Equal  Above       P  Median
New Salaries (€K)    20      5      1     14  0.0318   60.00
```

How the Sign test works:

1. While the Sign test produces a p-value that can be interpreted in a similar way to other hypothesis tests, the mathematics behind the Sign Test are quite different.
2. The sign test works by classifying each result within the sample as either above, below or equal to the test median. If the null hypothesis were true, we would expect to see approximately half of the results above and half below, the test median
3. However in this case, the majority of the results (14 out od 20) were above the test median, and this was high enough for the test to indicate (with statistical confidence) that the sample median (60) is greater than test median (48)
4. It is interesting to note that one of the results was exactly equal to the test median as summarised in the session window output above. The row was 7 in the example data file.

anexas consultancy SE

1 Sample Sign Test – Interpreting the Results

Sign Test for Median: New Salaries (€K)

```
Sign test of median =  48.00 versus > 48.00

                     N  Below  Equal  Above       P  Median
New Salaries (€K)   20      5      1     14  0.0318   60.00
```

- The p-value is 0.0318 (which is lower than the Alpha Level of 0.05), and so you can reject the null, and conclude (with 95% confidence) than the median salary in the sample is greater than the historical median of 48, So the new negotiation process does increase placement salaries.
- The I Sample Sign test menu in Minitab does not offer any additional graphs. However using a Dot plot from the graph menu, the sample results can be visually compared to the historical benchmark of 48K. As can be seen, while some of the results are below 48, most of them are above and this is enough to conclude, statistically, that the median new salary is higher than 48K

One-Sample Wilcoxon Test - Overview

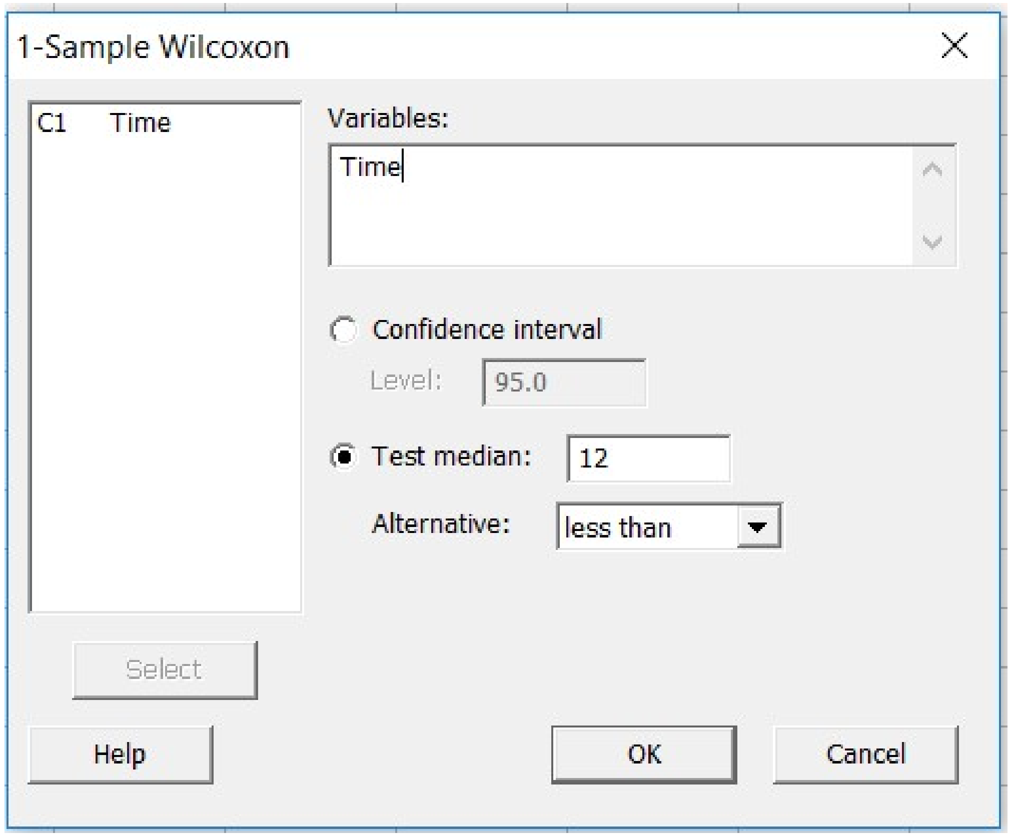
- Use the one-sample Wilcoxon (also called one-sample Wilcoxon signed rank) confidence interval and test procedures to make inferences about a population median based on data from a random sample.
- Use the 1-Sample Wilcoxon when you are unable to assume a distribution for the population from which the sample was drawn, but you can assume the distribution is symmetric . This is a nonparametric alternative to one-sample Z and one-sample t procedures.

One-Sample Wilcoxon Test - Overview

Example: A chemist wants to see if a newly developed antacid relieves pain in less than 12 minutes.

The data file for this example can be found in : ***Antacid.MTW***

Stat > Nonparametrics > One Sample Wilcoxon



One-Sample Wilcoxon Test – Interpreting the Results

Wilcoxon Signed Rank Test: Time

```
Test of median = 12.00 versus median < 12.00

          N for  Wilcoxon            Estimated
     N    Test   Statistic       P      Median
Time 16     16        26.5   0.017       10.05
```

Other results in the Session Window :

The Wilcoxon statistic is 26.5 and the Associated p-value is 0.017

Session Window Results:

Minitab produces only session window results for this test as follows:

Based on the sample data, you want to know if the newly developed antacid relieves pain in less than 12 minutes.

The hypotheses are

Null Hypothesis: H0: Median = 12.00 and
Alternate Hypothesis: H1: Median < 12.00

Estimated Median:

The median of the observations for each treatment. These medians provide an estimate of the population medians for each treatment.

p-value is 0.017, Since, the p-value is less than 0.05, therefore you should reject H0 and conclude that the antacid relieves pain significantly faster than 12 minutes.

Mann-Whitney Test - Overview

Use the two-sample Mann-Whitney (also called two-sample rank or two-sample Wilcoxon rank sum) confidence interval and test procedures to make inferences about the difference between two population medians based on data from two independent, random samples.

For example, you can determine whether

- The packing time of two packing machines is the same
- The time to relief is the same for two pain relievers

Assumptions:

- Samples are randomly drawn whose distributions have the same shape
- The two random samples are independent

The Mann-Whitney test is a nonparametric alternative to the two-sample t test with pooled sample variances.

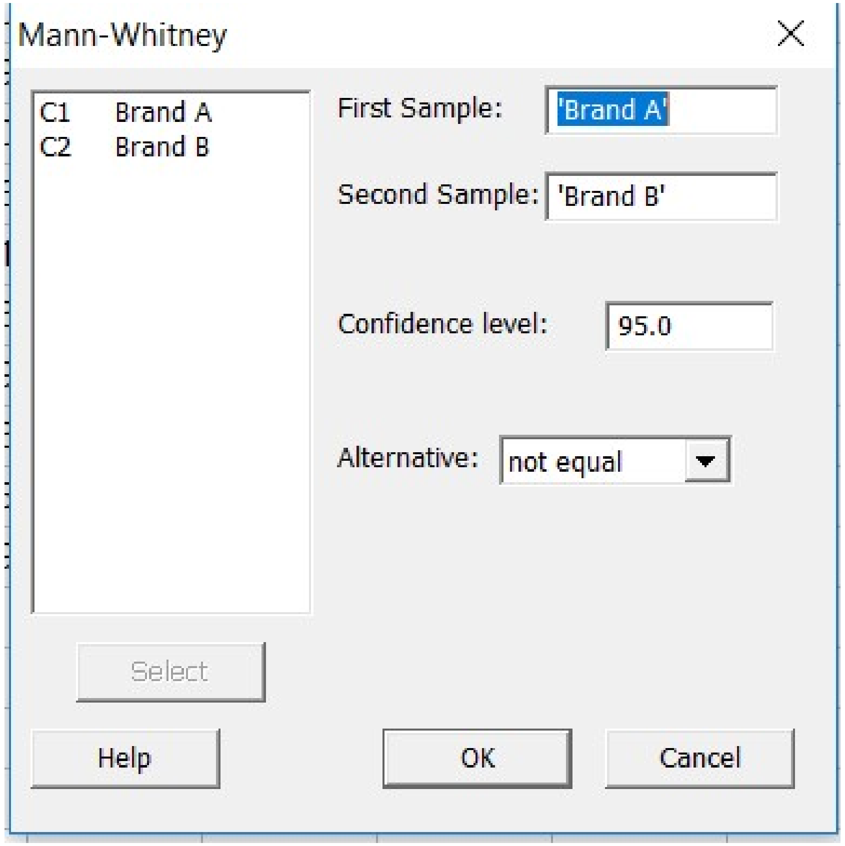
anexas consultancy SE

Mann-Whitney Test - Overview

Example: A state's highway department uses two brands of paint for striping roads. A highway official wants to know if a difference exists between the two brands of paint. To assess the problem, the official records the number of months that stripes applied with each brand of paint last on the highway.

The data file for this example can be found in : ***Srtipes.MTW***

Stat > Nonparametrics > Mann-Whitney



Mann-Whitney Test – Interpreting the Results

Mann-Whitney Test and CI: Brand A, Brand B

```
             N   Median
Brand A     11   36.000
Brand B     10   37.600

Point estimate for η1 - η2 is -1.850
95.5 Percent CI for η1 - η2 is (-3.000,-0.901)
W = 76.5
Test of η1 = η2 vs η1 ≠ η2 is significant at 0.0019
The test is significant at 0.0019 (adjusted for ties)
```

Other results in the Session Window :

The Mann-Whitney statistic is 76.5 and the associated p-value is 0.0019.

Confidence Interval:

Because you do not know the true value of the median, the confidence interval gives you a range of likely values based on the sample. In repeated sampling, the proportion of intervals that include the true value of the median is equal to 1 minus the chosen a-level.

Session Window Results:

Minitab produces only session window results for this test as follows:

Based on your sample, you want to know if the time that the paint stripes last on the highway is the same for the two brands

The hypotheses are

Null Hypothesis: H0: H0: $\eta 1 = \eta 2$ and
Alternate Hypothesis: H1: H1: $\eta 1$ not = $\eta 2$

Point Estimate:

The difference between the sample medians is a point estimate for the difference between population medians.

p-value is 0.0019. Because the p-value is less than 0.05, you should reject H0 and conclude that the median times are significantly different.

Kruskal-Wallis Test - Overview

The Kruskal-Wallis test compares the medians of different samples of data, and can be used where the data samples are not Normally distributed and do not have any obvious outliers.

Example:- A project is looking at the time to deliver different products (INGOT and BILLET). The box plot below shows that the INGOT product appears to be delivered quicker than BILLET, and the team are keen to validate this conclusion before other tools (such as detailed process mapping) are used to find out why.

Because the INGOT results do not appear to be normally distributed (a histogram and Box plot both indicate a skewed distribution), a Kruskal-Wallis test is being used to compare the median values of the two samples.

Kruskal-Wallis Test – Interpreting the Results

Kruskal-Wallis Test: Time to deliver versus Product

```
Kruskal-Wallis Test on Time to deliver

Product    N  Median  Ave Rank      Z
BILLET    25   7.500      29.7   0.73
INGOT     30   6.750      26.6  -0.73
Overall   55              28.0

H = 0.53  DF = 1  P = 0.467
H = 0.54  DF = 1  P = 0.463  (adjusted for ties)
```

Session Window Results:

Minitab produces only session window results for this test as follows:

Firstly the sample sizes and medians of the samples are summarised. The difference in the median values is 0.75 days (7.5-6.75), but this should be considered in combination with the following:

The size of the samples (25 for BILLET and 30 for INGOT) appears relatively low.

The resolution of the data was to the nearest 0.5 days (see previous page).

From these reasons, a hypothesis test is essential in order to decide if the difference in medians is statistically significant, as described on the left.

Kruskal-Wallis Test – Interpreting the Results

Kruskal-Wallis Test: Time to deliver versus Product

```
Kruskal-Wallis Test on Time to deliver

Product   N  Median  Ave Rank      Z
BILLET   25   7.500      29.7   0.73
INGOT    30   6.750      26.6  -0.73
Overall  55              28.0

H = 0.53  DF = 1  P = 0.467
H = 0.54  DF = 1  P = 0.463  (adjusted for ties)
```

Analysing the p-value:

The p-value from this test is 0.463. since this is higher than 0.05, we cannot say with confidence that there is a difference in the medians of the two samples.
In other words, the median delivery times of the two internet products (that the two samples of data represent) could be the same as each other.

Note: the two p-values usually very similar, but if not, use the value that is “adjusted for ties”

So, to summarise the results in day to day language:

Based on the data we have collected, we cannot say with confidence that there is a difference between the medians.
The difference of 0.75 hours between the sample medians could easily have occurred just by chance.

If there is a difference, more data will have to be collected to prove it.

Mood’s Median Test - Overview

The Mood’s Median test compares the medians(central position) of different samples of data, where the samples are not Normally distributed and where there are obvious outliers.in the data samples.

Example:- A project is looking at the time to deliver different products (INGOT and BILLET).
The data has been stratified into two groups INGOT and BILLETS, and the box plot below shows that the INGOT product appears to be delivered quicker than BILLET.

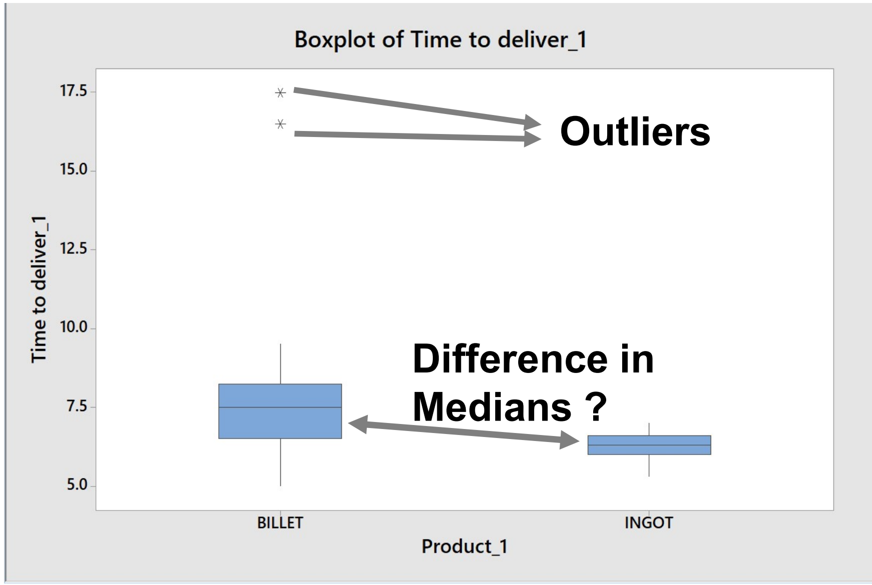
The team are keen to validate this before they set out to find and understand the root cause of this difference. Mood’s Median test is being used because the Billet data appears to be skewed and also has some outliers (the asterisks)

Mood’s Median Test - Overview

Data format:

Data for the Mood’s Median test must be stacked in one column, with the subgroup(factor) code alongside as shown on the right.

The data file for this example can be found in ***Time to deliver INGOT BILLETS.mpj and in Columns C4 &***



Stat > Nonparametrics > Mood’s Median

Mood's Median Test – Interpreting the Results

Mood Median Test: Time to deliver_1 versus Product_1

```
Mood median test for Time to deliver_1
Chi-Square = 7.73    DF = 1    P = 0.005

                                    Individual 95.0% CIs
Product_1  N≤  N>  Median  Q3-Q1  +---------+---------+---------+------
BILLET      9  16    7.50   1.75          (---------------*-------)
INGOT      22   8    6.30   0.60  (----*--)
                                  +---------+---------+---------+------
                                  6.00      6.60      7.20      7.80

Overall median = 6.50

A 95.0% CI for median(BILLET) - median(INGOT): (0.00,1.91)
```

Session Window Results:

Minitab produces only session window results for this test as follows:

The p-value for the test is 0.005, since this is less than Alpha Level of 0.05 we can say, with 95% confidence, that the medians of the subgroups are different.

A rough graph of the 95% Confidence Intervals (Cis) for the median of the subgroups is shown. Note that although the Cis are visually overlapping, the statistical conclusions is that they are different.

Other results in the Session Window :

Based on the data subgroups,
the difference between the subgroup
sub-group medians is 1.2 Days(7.50-6.30)

The confidence interval for the difference provides more detail and confirms (with 95% confidence), that the difference in medians is somewhere between 1.91 and 0.00 Days.

So, to summarise the results in day to day language:

We can be very confident that there is a difference in the median delivery time for INGOT & BILLETS.

The median delivery time of INGOTs is at least 0 days quicker than BILLETs but could be as much as 1.91 days quicker.

Hypothesis Testing of Proportions

Hypothesis Tests

Y	X	Hypothesis Test
Continuous / Variable Data	Attribute / Discrete Data	1 z, 1 t, 2 t, paired t, ANOVA
Attribute / Discrete Data	Attribute / Discrete Data	1 p, 2 p, Chi Square
Continuous / Variable Data	Continuos / Variable Data	Correlation, Regression, Multiple Regression
Attribute / Discrete Data	Continuos / Variable Data	Logistic Regression

anexas consultancy SE

Why Learn Hypothesis Tests of Proportion?

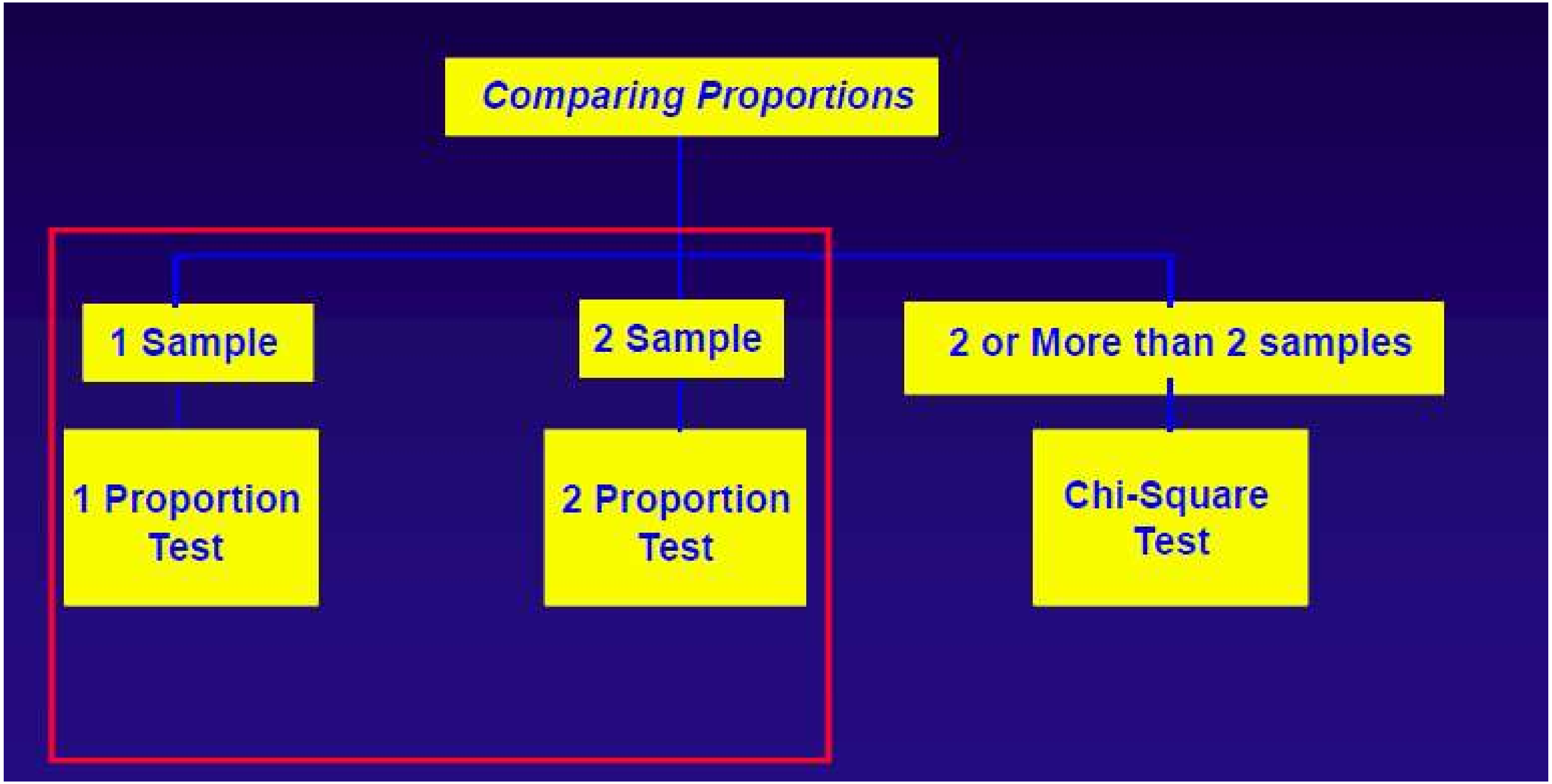
- Make data driven decisions with defined confidence
- Determine if a statistically significant difference of proportion exists between:
  - A sample and a target
  - Two independent samples
  - More than two independent samples

What are Hypothesis Tests of Proportion?

Test	Method for analyzing the differences between:
1 Proportion	a sample proportion and a target value
2 Proportion	proportion obtained from two independent samples

anexas consultancy SE

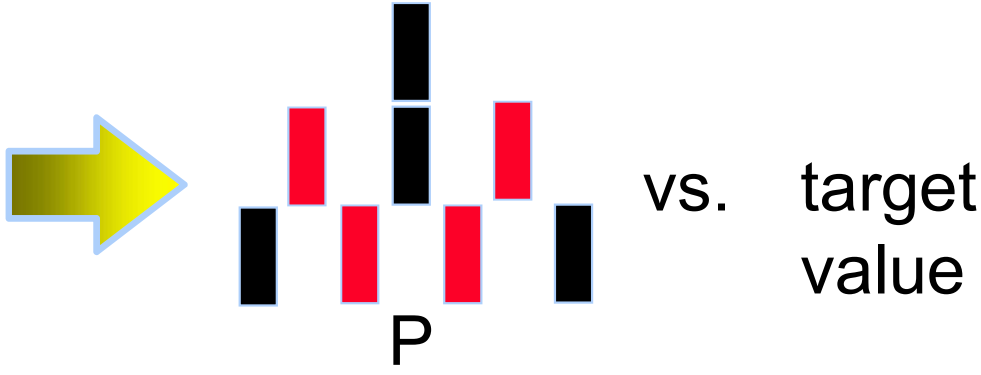
Hypothesis Testing of Proportion - Roadmap



Comparison of Proportion: 2 Scenarios

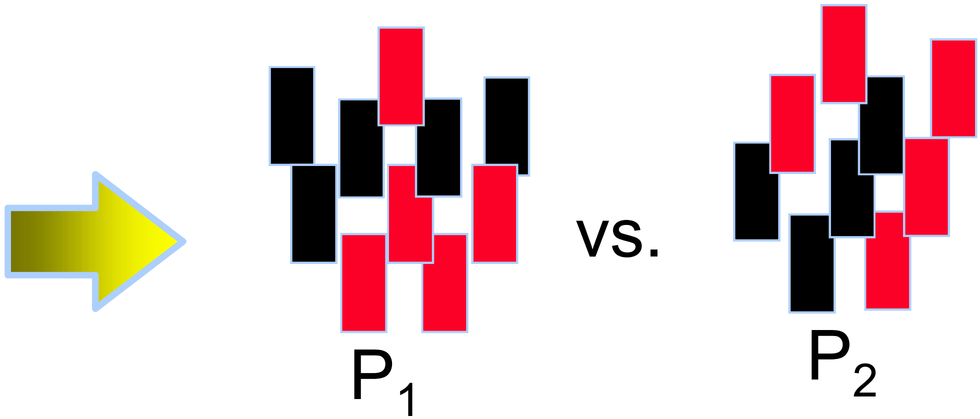
1) Single Proportion Comparison

One population proportion compared to a target value



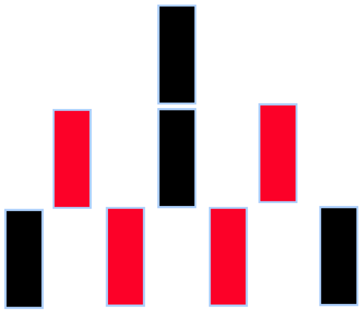
2) Two Sample Comparison

Proportions of two independent populations compared to each other



1 Proportion Test

Single Proportion Comparison



vs. target value

P

Practical Question
(example)

"Is the population proportion statistically different from the target value?"

Statistical Question

$H_0 : P = \text{target value}$

$H_a : P \neq \text{target value}$

1 Proportion Test

- Hypothesis test about the population proportion using information from one sample

Business Process Example
IPO Prospectus

A Black Belt[SM] is studying the effects of voluntary disclosure of earnings forecast in the Initial Public Offering (IPO) prospectus.

A random sample of 130 IPO prospectus revealed that 58 of them did not reveal their earnings forecast.

Test the hypothesis at 5% significance level that less than 50% of IPO prospectus do not disclose their earnings forecast.

Example: IPO Prospectus

- Practical Problem
  - Is the percentage of IPO prospectus disclosing their earnings forecast less than 50%?
- Statistical Problem
  - Is population proportion of IPO prospectus revealing their earnings forecast less than 50%?
  - Null hypothesis: population proportion is 50%
  - Alternate hypothesis: population proportion is less than 50%

Example: IPO Prospectus

State the hypotheses and significance level

H_o: P = 0.50

H_a: P < 0.50

$\alpha = 0.05$

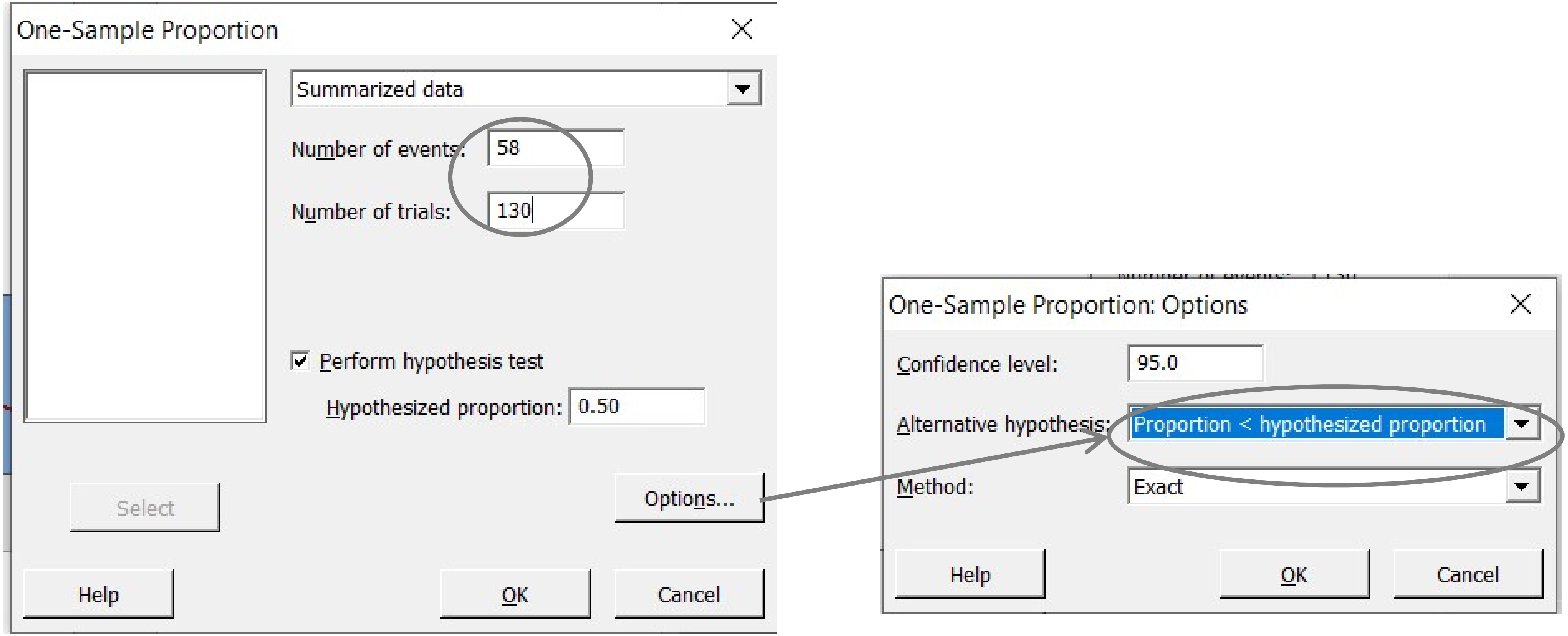
What hypothesis test is appropriate?

- These hypotheses deal with proportions
- Comparing population proportion against a target proportion using one sample data
- Use 1 Proportion Test

Example: IPO Prospectus

Tool Bar Menu > Stat > Basic Statistics > 1 Proportion

Analysis through Minitab



Example: IPO Prospectus

Test and CI for One Proportion

```
Test of p = 0.5 vs p < 0.5

                                                               Exact
Sample        X       N  Sample p  95.0% Upper Bound  P-Value
1            58     130  0.446154           0.522079    0.127
```

- Interpretation:
  - P-value = 0.127.
  - P-value > α-risk (0.05): Fail to reject H_o
  - Infer H_o: insufficient evidence that only less than 50% of IPO prospectus disclose their earnings forecast

anexas consultancy SE

Industrial Process Example
Migraine Medicine

A pharmaceutical company has invented a new medicine for relieving migraine headaches. The company wants to test the hypothesis that the drug is effective more than 73% of the time.

A clinical study of 111 out of 143 adults suffering from migraine headaches reported relief after using the drug. Is this sufficient evidence that the drug is effective more than 73% of the time?

Use a 5% significance level.

Example: Migraine Medicine

Practical Problem

- Is the migraine medicine effective more than 73% of the time?

Statistical Problem

- Is population proportion of medicine effectiveness greater than or equal to 73%?
- Null hypothesis: population proportion is 73%
- Alternate hypothesis: population proportion is greater than 73%

Example: Migraine Medicine

State the hypotheses and significance level

$$H_o: P = 0.73$$

$$H_a: P > 0.73$$

$$\alpha = 0.05$$

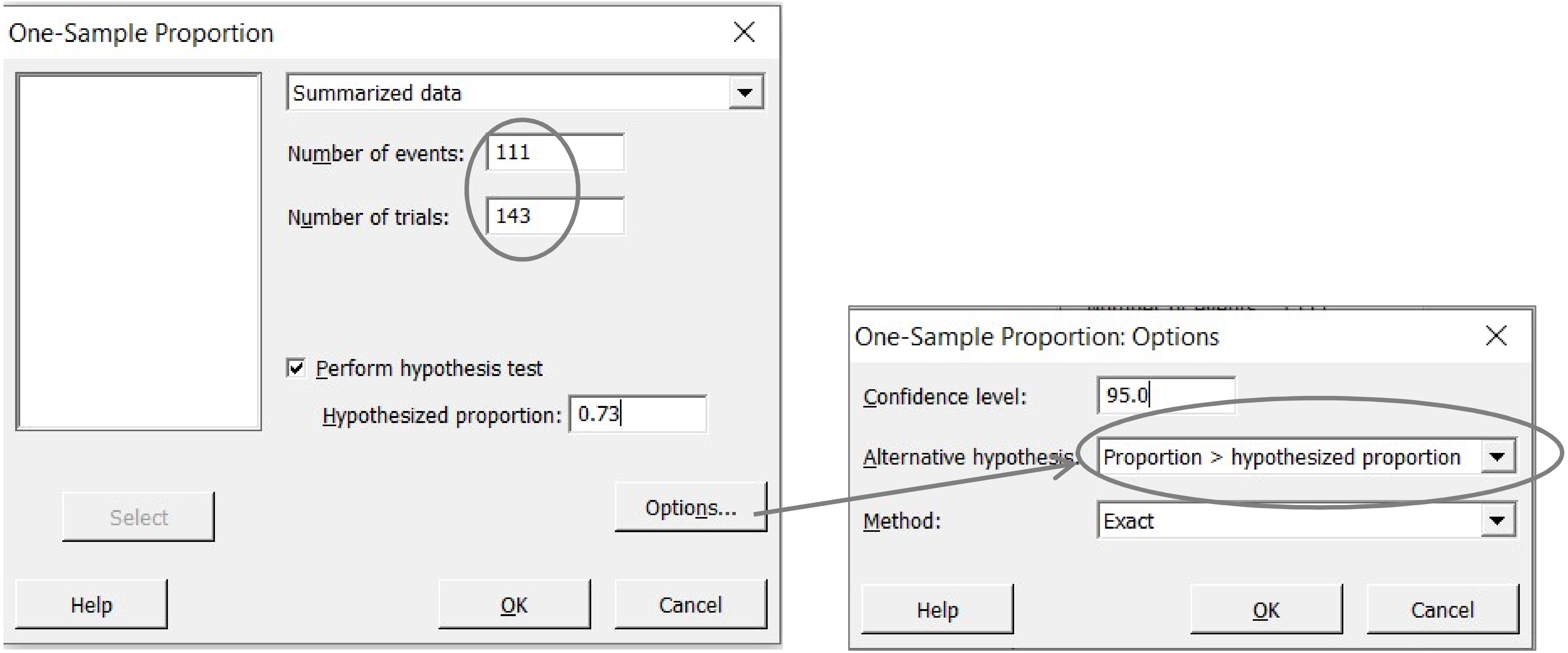
What hypothesis test is appropriate?

- These hypotheses deal with proportions
- Comparing population proportion against a target proportion using one sample data
- Use 1 Proportion Test

Example: Migraine Medicine

Tool Bar Menu > Stat > Basic Statistics > Proportion

Analysis through Minitab



anexas consultancy SE

Example: Migraine Medicine

Test and CI for One Proportion

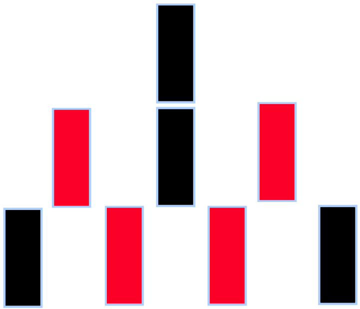
```
Test of p = 0.73 vs p > 0.73

                                                          Exact
Sample        X      N  Sample p  95.0% Lower Bound  P-Value
1           111    143  0.776224           0.711327    0.124
```

Interpretation:

- p-value = 0.124
- p-value > α-risk (0.05): fail to reject H_o
- Infer H_o: insufficient evidence that that the drug is more than 73% effective

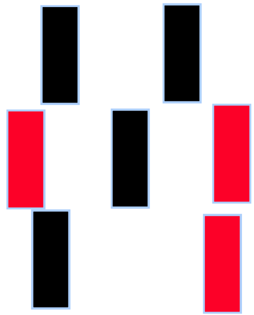
2 Proportion Test

Two Sample Proportion Comparison



P_1

VS.



P_2

Practical Question
(example)

Are the two populations' proportions
- statistically different?

Statistical Question

$$H_o: P_1 = P_2$$

$$H_a: P_1 \neq P_2$$

anexas consultancy SE

2 Proportion Test

- Hypothesis test about the difference between two population proportions using information from two samples
- Two sets of samples are statistically independent

Industrial Process Example: Comparing Medicines

MedChoice, Inc. distributes two identical brands of medicine for relieving migraine headaches.

It is found from controlled studies that 145 out of 200 people suffering from migraines reported relief through use of Brand A whereas 101 out of 150 people reported relief through the use of Brand B.

The company wants to know if we can conclude at the 5% level of significance that the percentage of people getting relief through use of Brand A is higher than through Brand B?

Example: Comparing Medicines

- Practical Problem
  - Is Brand A better than Brand B in providing relief from migraine headaches?

- Statistical Problem
  - Is population proportion of relief through Brand A greater than population proportion through Brand B?
  - Null hypothesis: population proportion for Brand A = population proportion for Brand B
  - Alternate hypothesis: population proportion for Brand A is greater than that of Brand B

Example: Comparing Medicines

State the Hypotheses and Significance Level

H_o: $P_A - P_B = 0$

H_a: $P_A - P_B > 0$

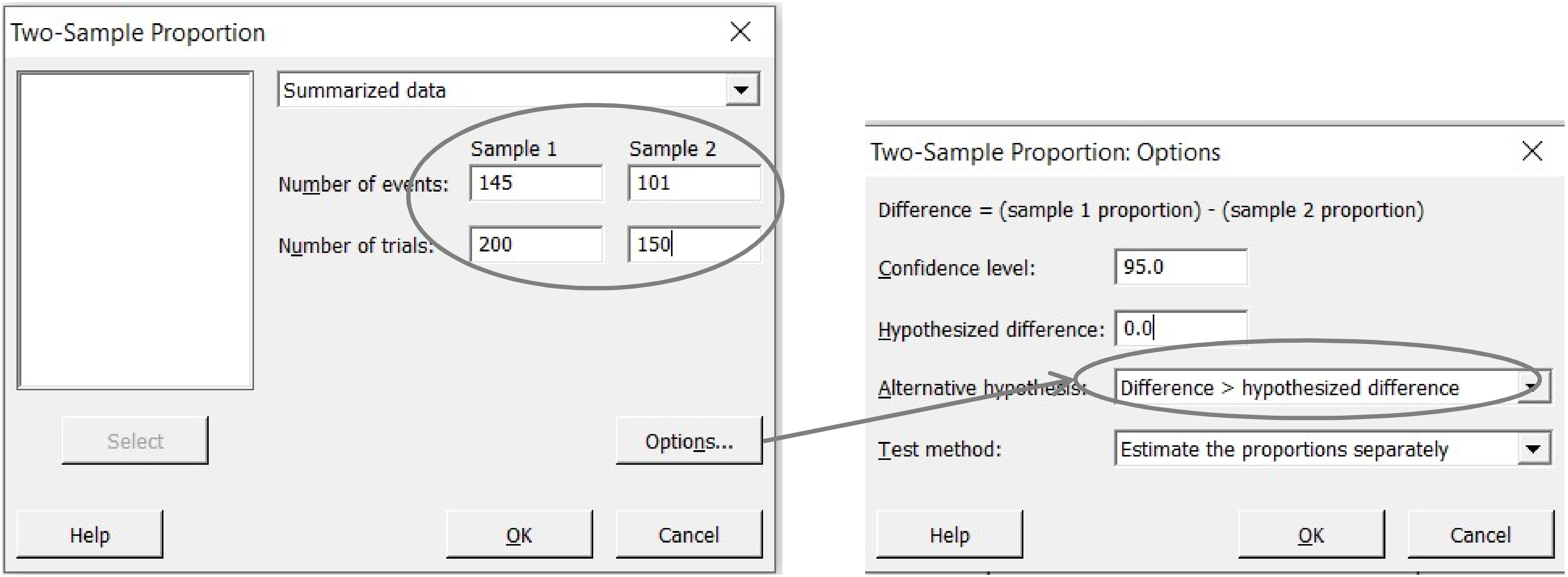
$\alpha = 0.05$

- What Hypothesis Test is Appropriate?
  - These hypotheses deal with proportion values
  - Comparing population proportions using two
  - sets of independent samples
  - Use 2 Proportion Test

Example: Comparing Medicines

Tool Bar Menu > Stat > Basic Statistics > 2 Proportion

Analysis through Minitab



anexas consultancy SE

▪Example: Comparing Medicines

```
Test and CI for Two Proportions

Sample     X    N   Sample p
1        145  200   0.725000
2        101  150   0.673333



Difference = p (1) - p (2)
Estimate for difference:  0.0516667
95% lower bound for difference:  -0.0299692
Test for difference = 0 (vs > 0):  Z = 1.04  P-Value = 0.149

Fisher's exact test: P-Value = 0.176
```

What is the Interpretation?

p-value = 0.149

p- value (0.149) > α-risk (0.05); fail to reject H_o

Infer H_o: insufficient evidence that brand A is more effective than brand B

Hypothesis Testing
Chi-Square Tests

Why Learn Chi-Square Tools?

Make data driven decisions with defined confidence

Determine if

- Two attribute variables are related
- A population fits a certain probability model (distribution)

What Are Chi-Square Tools?

Chi-Square Goodness-of-Fit Test

To test if a particular distribution (model) is a good fit for a population

Chi-Square Test for Association

To test if a relationship between two attribute variables exists

$$\chi^2 = \sum_{j=1}^{g} \frac{(f_o - f_e)^2}{f_e}$$

▪ *Chi-Square Statistic*

Both of these tools use the Chi-Square distribution, where f_o and f_e are the observed and expected frequencies, respectively.

Test for Association

Business Process Example: Black Belt℠ Projects

A sample of Black Belts℠ was asked to rate both their six sigma project performance and the average weekly hours spent with the Project Champion℠ discussing project details. The results are shown in the following table. Test at the 5% level the null hypothesis of no association between the two sets of ratings.

Data is given in ***Chi1.mtw***

Time with Champion

PROJECT PERFORMANCE

HOURS	Low	Medium	High
< 0.1	17	21	12
0.1 - 1	31	53	21
> 1	17	42	71

Example: Black Belt[SM] Projects

- Practical problem
  - Does the performance of Black Belt[SM] projects depend on time spent with Project Champions[SM]?
- Statistical problem
  - Is there an association between project performance and time spent with ChampionSM?
  - Null hypothesis: project performance is independent of the time spent with ChampionSM
  - Alternate hypothesis: project performance is dependent of the time spent with ChampionSM
- What hypothesis test is appropriate?
  - These hypotheses deal with relationship between two attribute variables
  - Use Chi-Square Test for Association

anexas consultancy SE

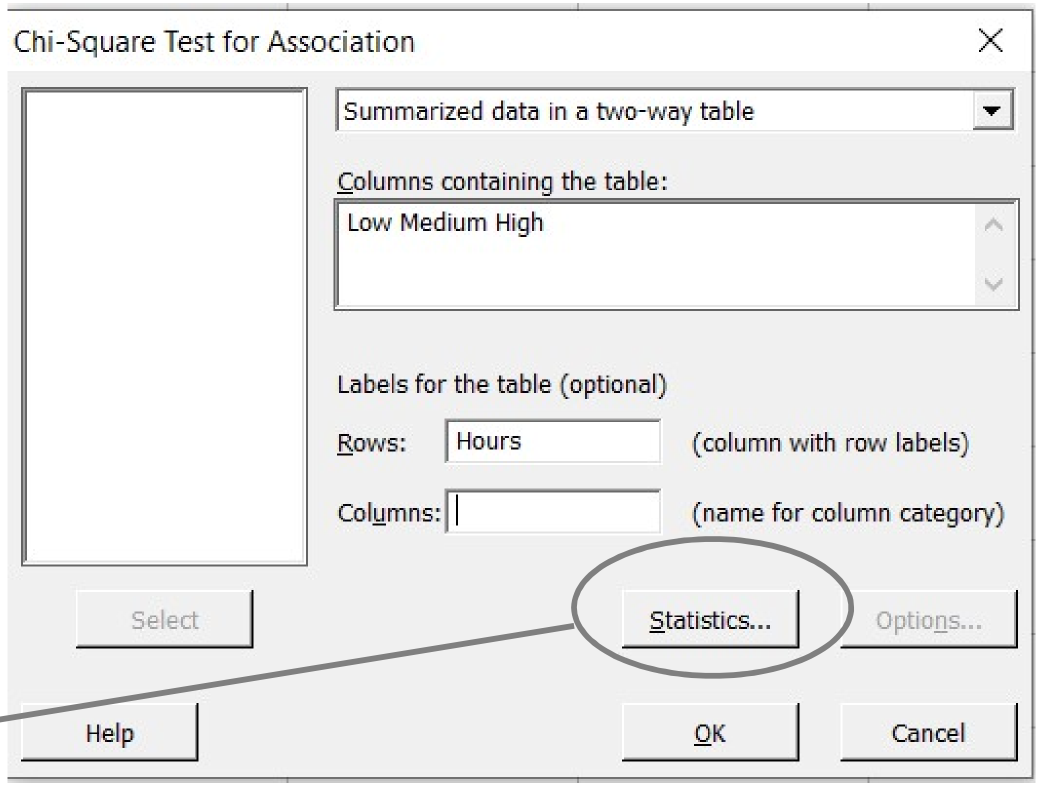
Example: Black Belt[SM] Projects

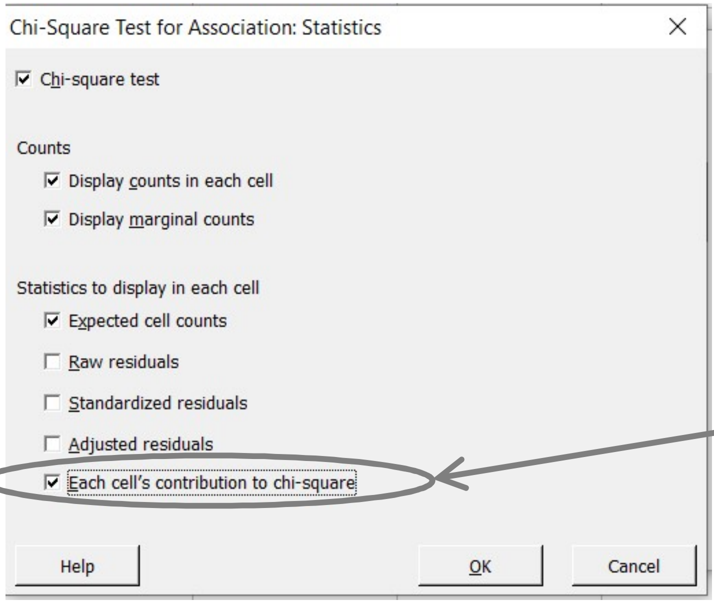
Tool Bar Menu > Stat > Tables > Chi-Square Test

1. Open data file Chi1.mtw

C1-T	C2	C3	C4
Hours	Low	Medium	High
<0.1	17	21	12
0.1-1.0	31	53	21
>1	17	42	71

2. *Stat > Tables > Chi-Square Test for Association*
3. Fill in the dialog as shown below:





4. Click Statistics and Enable Each cells' contribution to chi-square

anexas consultancy SE

Example: Black Belt[SM] Projects

Chi-Square Test: Low, Medium, High

Rows: Hours Columns: Worksheet columns

	Low	Medium	High	All
<0.1	17	21	12	50
	11.40	20.35	18.25	
0.1-1.0	31	53	21	105
	23.95	42.74	38.32	
>1	17	42	71	130
	29.65	52.91	47.44	
All	65	116	104	285

Cell Contents: Count
Expected count

Pearson Chi-Square = 36.622, DF = 4, P-Value = 0.000
Likelihood Ratio Chi-Square = 37.348, DF = 4, P-Value = 0.000

- Interpret output
- What is χ^2(calc)
- What is the p-value?
- What is its interpretation?

anexas consultancy SE

Example: Black Belt[SM] Projects

- Interpretation:
  - p-value = 0.000
  - p-value < a-risk (0.01): reject Ho
  - Infer Ha: sufficient evidence that Black BeltSM project performance and and time spent with ChampionSM are dependent

Hypothesis Testing-
Correlation and Regression

Hypothesis Tests

Y	X	Hypothesis Test
Continuous / Variable Data	Attribute / Discrete Data	1 z, 1 t, 2 t, paired t, ANOVA
Attribute / Discrete Data	Attribute / Discrete Data	1 p, 2 p, Chi Square
Continuous / Variable Data	Continuos / Variable Data	Correlation, Regression, Multiple Regression
Attribute / Discrete Data	Continuos / Variable Data	Logistic Regression

Why Learn Correlation and Regression?

- Explore the existence of relationship between variables with the aid of data
- Screen variables and determine which variable(s) has the biggest impact on the response(s) variable
- Describe the nature of relationship with the help of an equation and use it for prediction

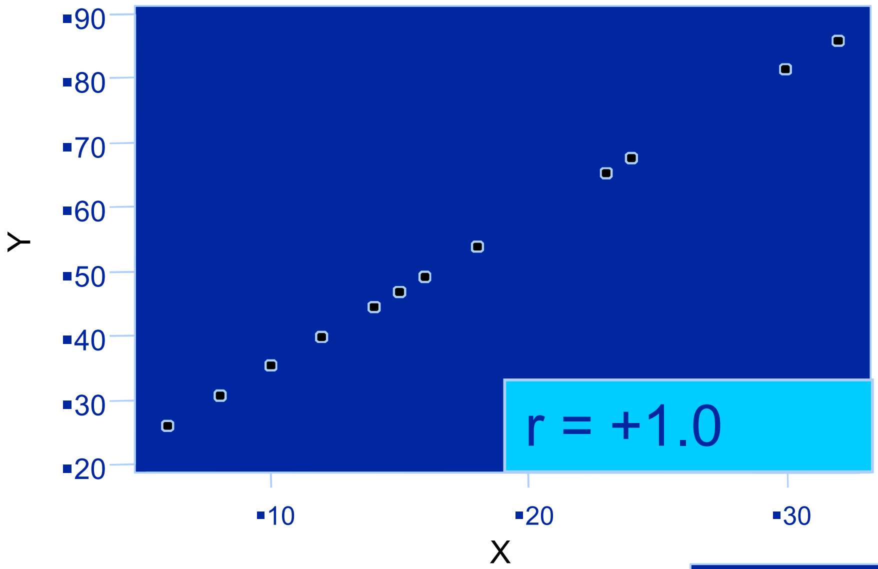
Correlation

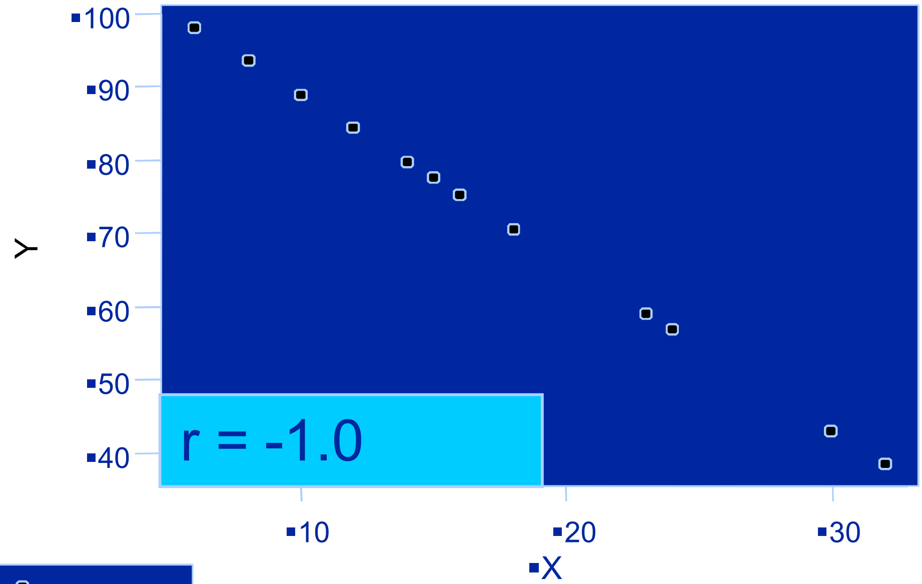
What is Correlation?

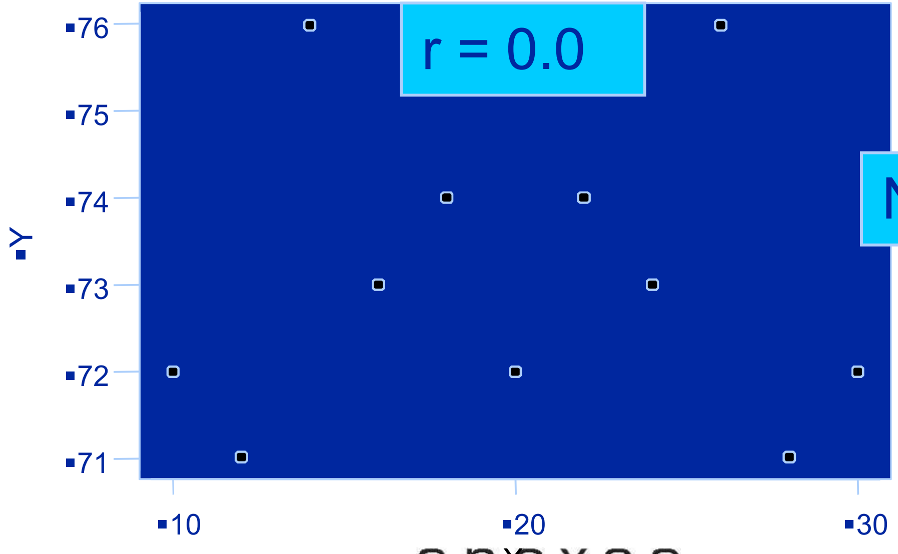
- Correlation is a measure of the strength of association between two quantitative variables
  (Ex: Pressure and Yield)
- Measures the degree of linearity between two variables assumed to be completely independent of each other

Correlation coefficient or P*earson* correlation coefficient is a way of measuring the strength of correlation

Correlation Coefficient



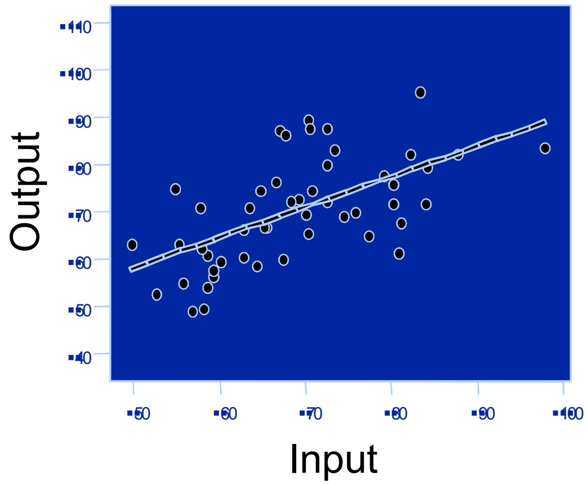




No correlation

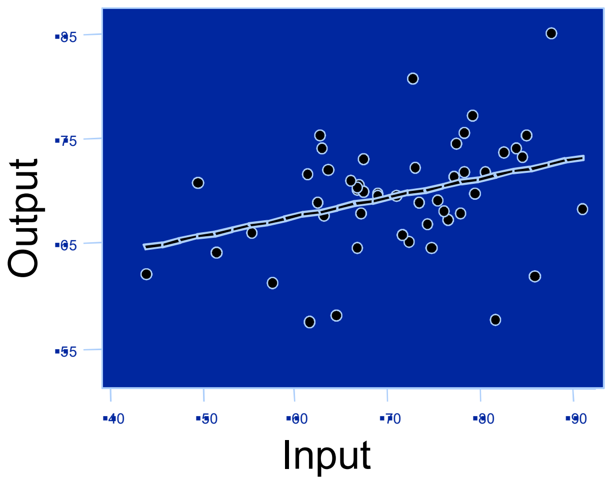
Strength and Direction of “+” Correlation

Moderate positive correlation



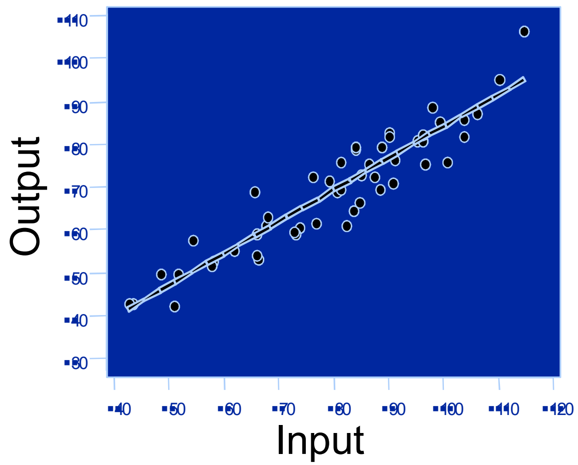
Y=25.7595+0.645418X
R Squared=0.369

Weak positive correlation



Y=56.6537+0.181987X
R Squared=0.115

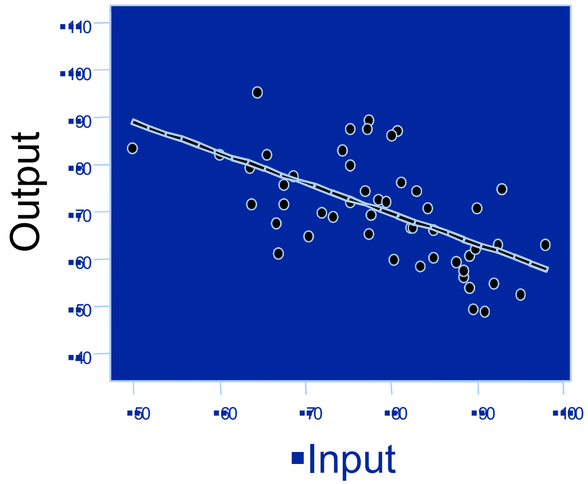
Strong positive correlation



Y=9.77271+0.745022X
R Squared=0.876

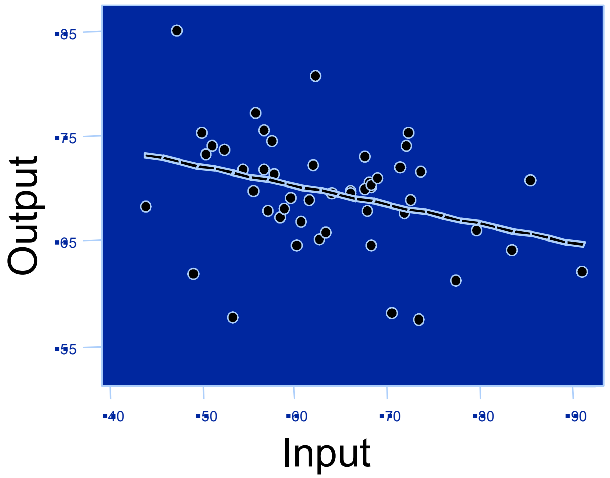
Strength and Direction of "-" Correlation

Moderate negative correlation



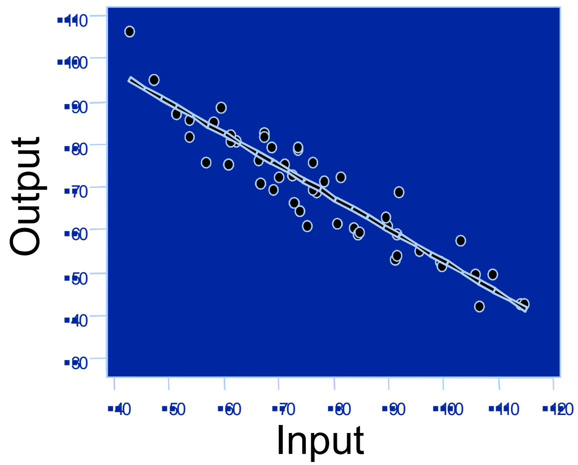
Y=90.3013-0.645418X
R Squared=0.369

Weak negative correlation



Y=74.8524-0.181987X
R Squared=0.115

Strong negative correlation

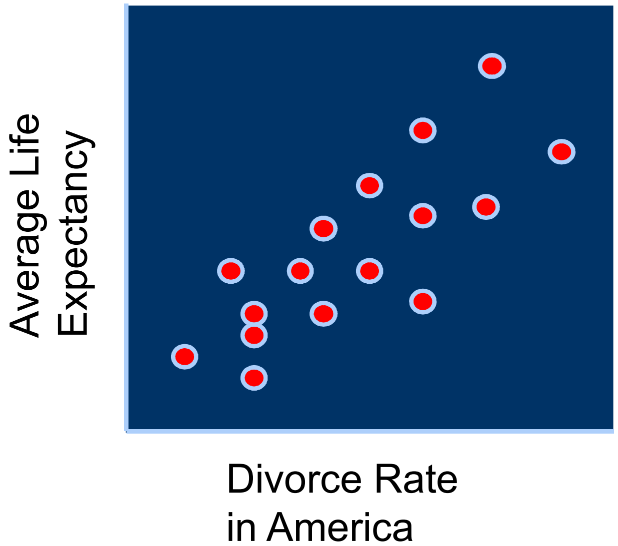


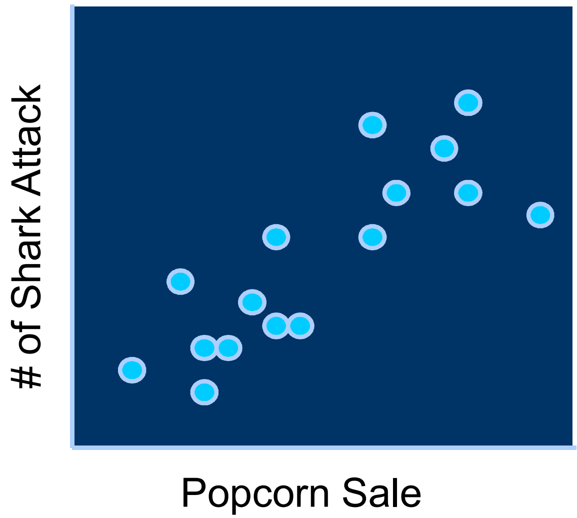
Y=99.1754-0.745022X
R Squared=0.876

Correlation vs. Causation

Data shows that average life expectancy of Americans increased when the divorce rate went up!

Is there a correlation between grass height and hair length?





Correlation does not imply causation! A third variable may be 'lurking' that causes both x and y to vary

Business Process Example: Cereal Sales

A market research analyst for a certain brand of cereal is interested in finding out if there is a relationship between the sales generated and shelf space used to display the cereal. As a result she conducted a study and collected data from 12 different stores selling this brand of cereal.

Shelf Space, Sq in	Sales, $
574	960
635	1779
533	651
560	831
628	1460
615	1370
540	851
587	1220
656	1889
594	1370
622	1609
567	1120

The data contains sales $ generated for a certain month and the shelf space dedicated to the product.

What would you do?

What questions might you ask?

Data in *Sales.mtw*

anexas consultancy SE

Example: Cereal Sales

- Practical Problem
  - Is there a relationship between sales $ from cereal and the shelf space used to display the cereal?
  - If there a relationship, how strong is that relationship?
- Statistical Problem
  - Are the variables ‘Sales’ and ‘Shelf Space’ correlated?
  - Null hypothesis: Sales and Shelf space are not correlated
  - Alternate hypothesis: Sales and Shelf space are correlated

Example: Cereal Sales

State the Hypotheses and Significance Level

H_o: $\rho = 0$

H_a: $\rho \neq 0$

$\alpha = 0.01$

Notice that the hypotheses are about a population parameter

What Hypothesis Test is Appropriate?

These hypotheses deal with correlation coefficient

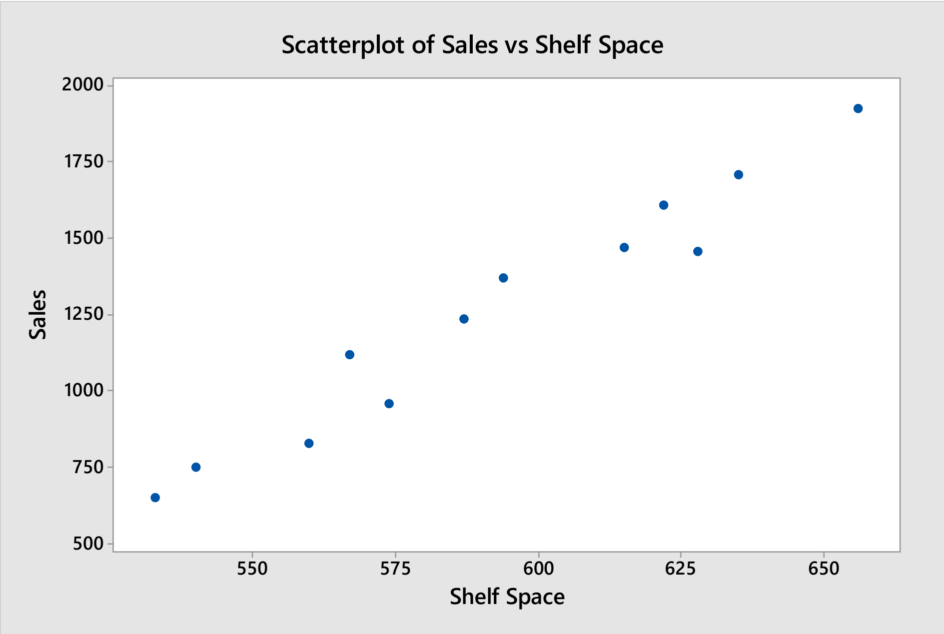
Make decisions based on Pearson correlation coefficient and ‘p-value’

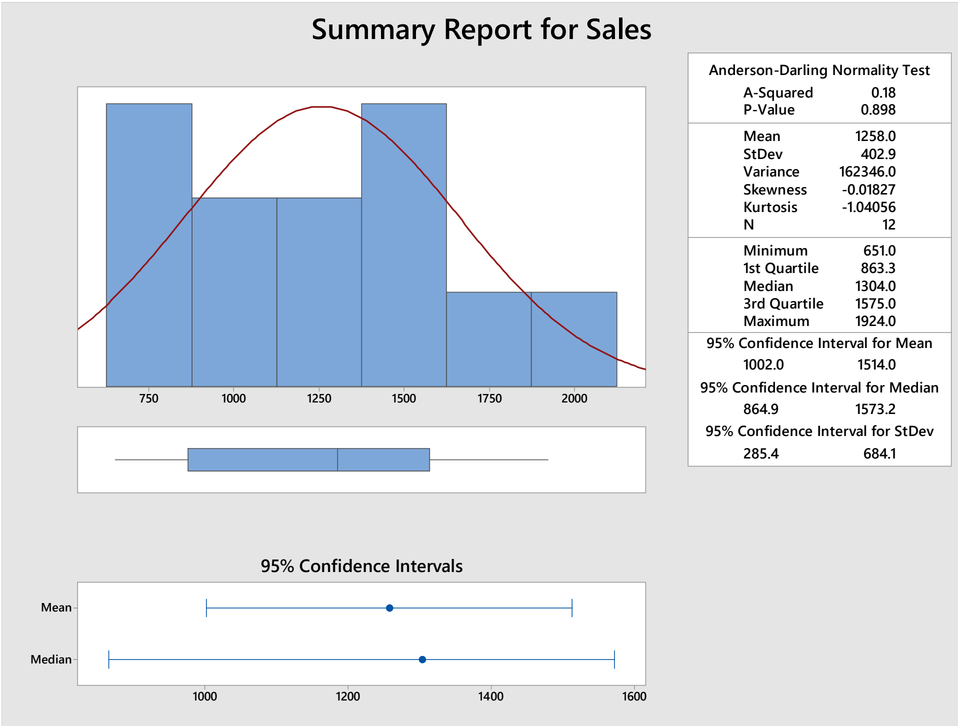
Example: Cereal Sales

Tool Bar Menu > Stat > Basic Statistics > Graphical Summery

- Practical and Graphical:
  - Practical questions about the data?
  - Plot the data using different techniques

Graph > Scatter Plot

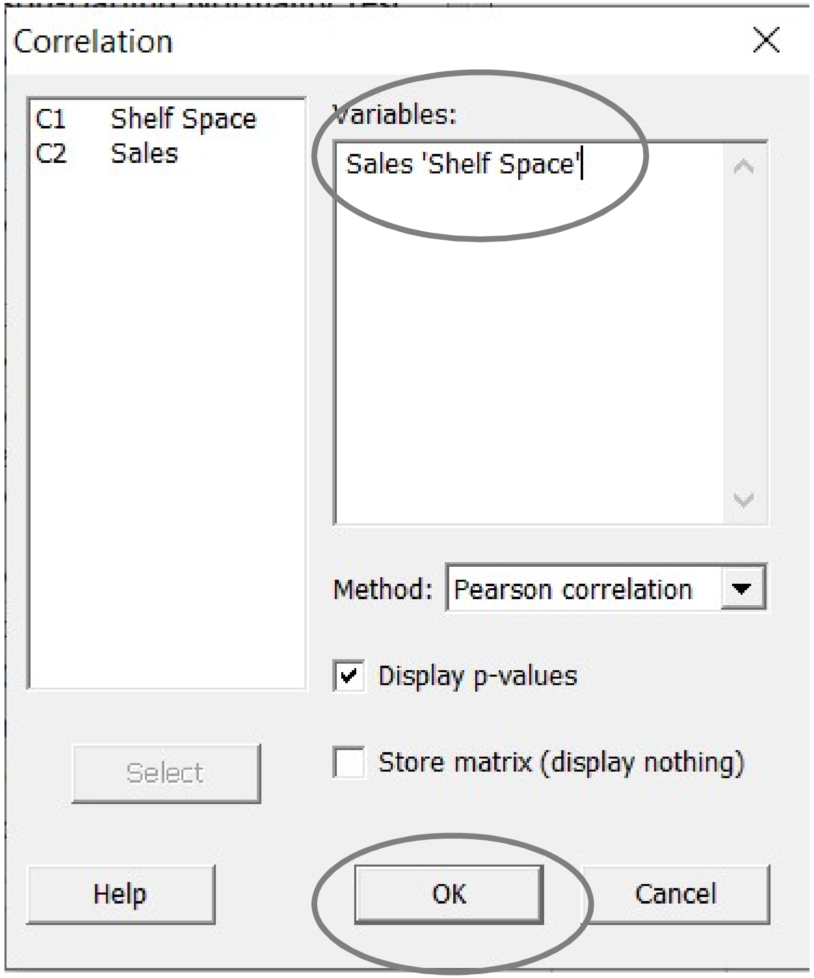




anexas consultancy SE

Example: Cereal Sales

Tool Bar Menu > Stat > Basic Statistics > Correlation



anexas consultancy SE

Example: Cereal Sales

Correlations: Shelf Space, Sales

Pearson correlation of Shelf Space and Sales = 0.978
p-value = 0.000

- What is the Decision?
  - Pearson correlation or correlation coefficient for the sample, r = 0.978
  - Does that mean ‘ρ’ is greater than zero? Or could it be that r = 0.978 due to chance variation while ‘ρ’ is still zero?
  - Answer this question using table next page

- What about
- the p-value?

Example: Cereal Sales

- What is the statistical interpretation?
  - p-value (0.000) < α-risk (0.01): reject the null hypothesis
  - Infer H_a: sufficient evidence that there is a correlation between sales $ and shelf space

Regression

Correlation and Regression

- Correlation tells how much linear association exists between two variables
- Regression provides an equation describing the nature of relationship

Correlations: Shelf Space, Sales

Pearson correlation of Shelf Space and Sales = 0.978

p-value = 0.000

Regression Analysis: Sales versus Shelf Space

The regression equation is Sales = - 4711 + 10.1 Shelf Space

Regression Terminology

- Response Variable
  - This is the uncontrolled variable - also known as dependent variable, output variable or Y variable
- Regressor Variable
  - Response depends on these variables - also known as independent variables, input variables, or X variables
- Noise Variable
  - Input variables (X) that are not controlled in the experiment
- Regression Equation
  - Equation that describes the relationship between independent variables and dependent variable
- Residuals
  - Difference between predicted response values and observed response values

Regression Objectives

- Determination of a Model
    - Explore the existence of relationship
- Prediction
    - Describe the nature of relationship using an equation and use the equation for prediction
- Estimation
    - To assess the accuracy of prediction achieved by the regression equation
- Determination of KPIV
    - Screen variables and determine which variable has the biggest impact on the response variable

Types of Regression

Simple Linear Regression

Single regressor (x) variable such as x_1 and model linear with respect to coefficients

Example 1: $y = a_0 + a_1x + \text{error}$

Example 2: $y = a_0 + a_1x + a_2 x^2 + a_3 x^3 + \text{error}$

Note: 'Linear' refers to the coefficients a_0, a_1, a_2 , etc. It implies that each term containing a coefficient is added to the model. In example 2, the relationship between x and y are cubic polynomial in nature, but the model is linear with respect to the coefficients.

Types of Regression

Multiple Linear Regression

Multiple regressor (x) variables such as x_1, x_2, x_3 and model linear with respect to coefficients

Example: $y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + error$

Simple Non-Linear Regression

Single regressor (x) variable such as x and model non-linear with respect to coefficients

Example: $y = a_0 + a_1 (1-e^{-a_2 x}) + error$

Multiple Non-Linear Regression

Multiple regressor (x) variables such as x_1, x_2, x_3 and model non-linear with respect to coefficients

Example: $y = (a_0 + a_1 x_1) / a_2 x_2 + a_3 x_3 + error$

Simple Linear Regression

Simple Linear Regression

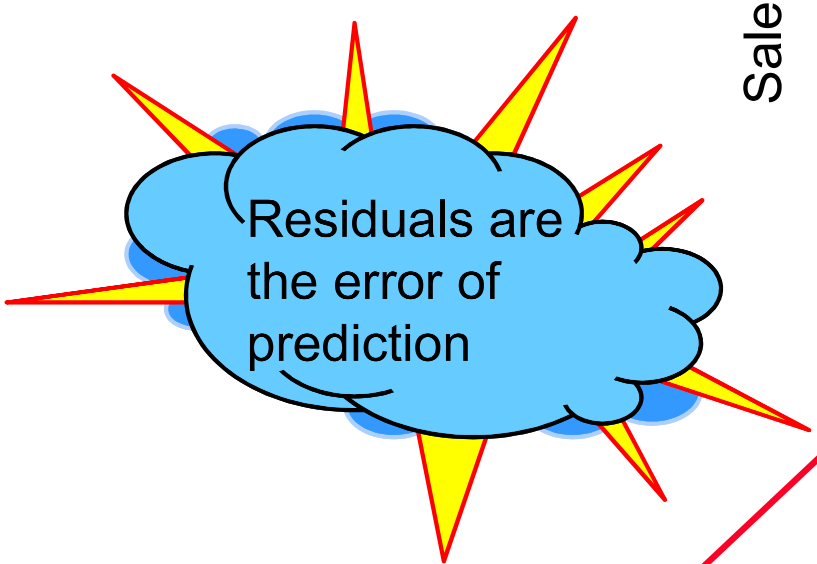
- Use one independent variable (x) to explain the variation in dependent variable (y)
  - Example 1: use shelf space to explain variation sales $
  - Example 2: amount of fertilizer applied to explain the yield of crop
- Method of Least Squares
  - Use the 'Method of Least Squares' to find the best fitting regression line

anexas consultancy SE

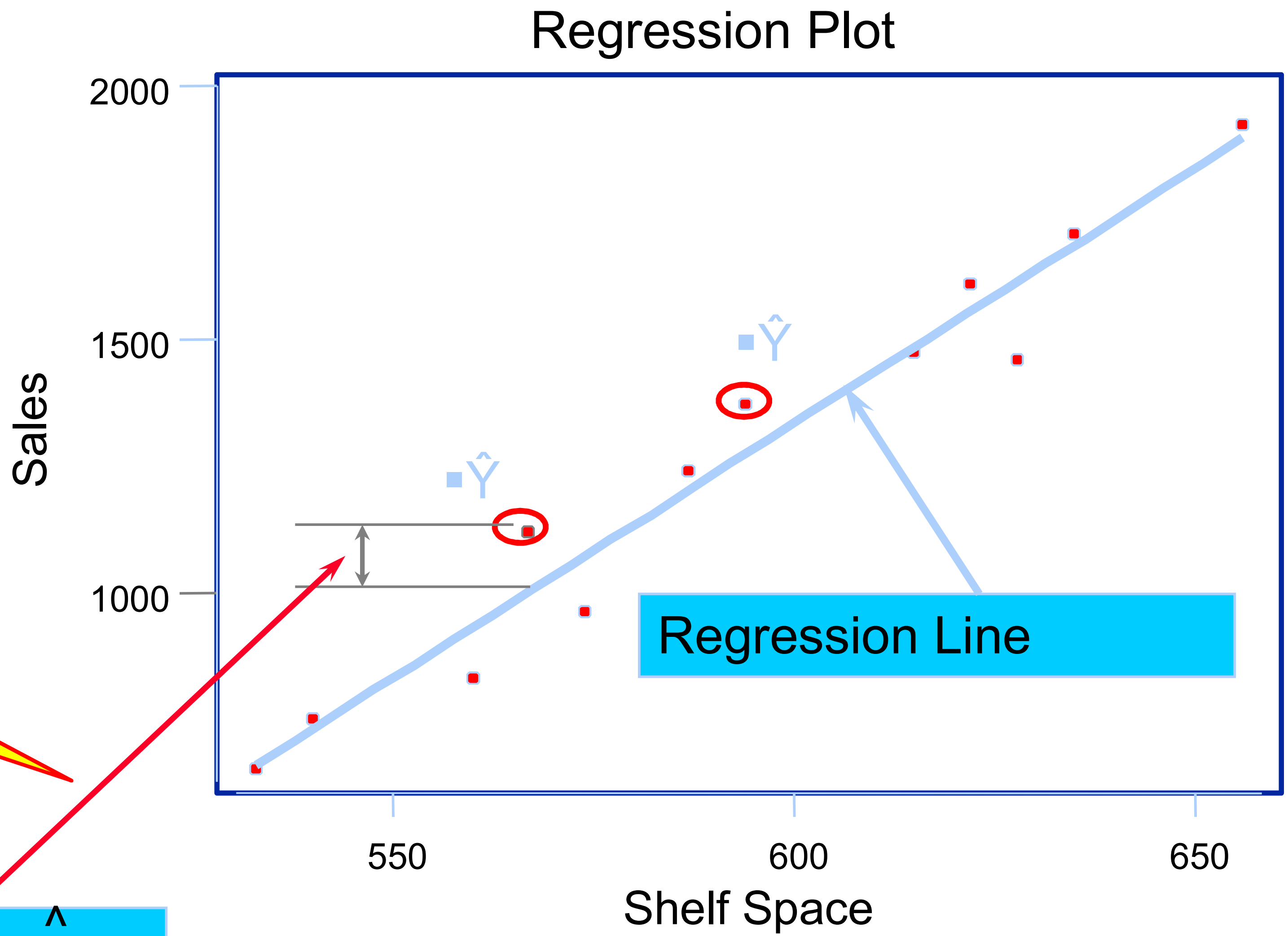
Method of Least Squares

Objective:

Find a line that will minimize sum of squares of residuals



Residual = Y - Ŷ



Business Process Example: Cereal Sales

A market research analyst for a certain brand of cereal is interested in predicting the sales generated from information on shelf space used to display the cereal. As a result she conducted a study and collected data from 12 different stores selling this brand of cereal

Shelf Space, Sq in	Sales, $
574	960
635	1779
533	651
560	831
628	1460
615	1370
540	851
587	1220
656	1889
594	1370
622	1609
567	1120

- The data contains sales $ generated for a certain month and the shelf space dedicated to the product
- How will we create a simple linear regression model for the two variables?
- Predict the sales $ using the regression equation when shelf space is 615 sq. in.

Data in ***Sales.mtw***

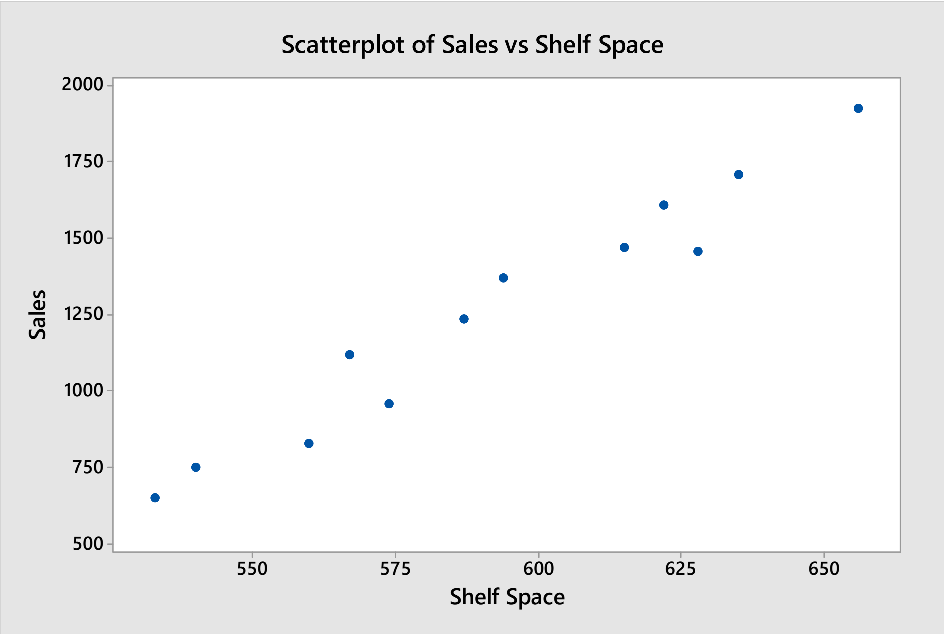
anexas consultancy SE

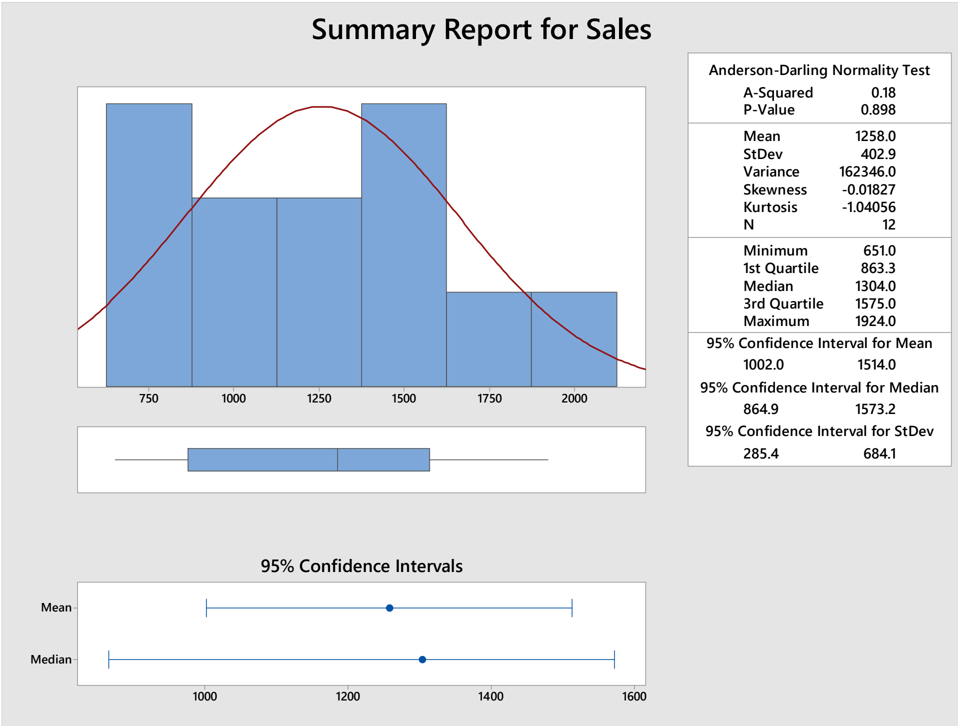
Example: Cereal Sales

Tool Bar Menu > Stat > Basic Statistics > Graphical Summery

- Practical and Graphical:
  - Practical questions about the data?
  - Plot the data using different techniques

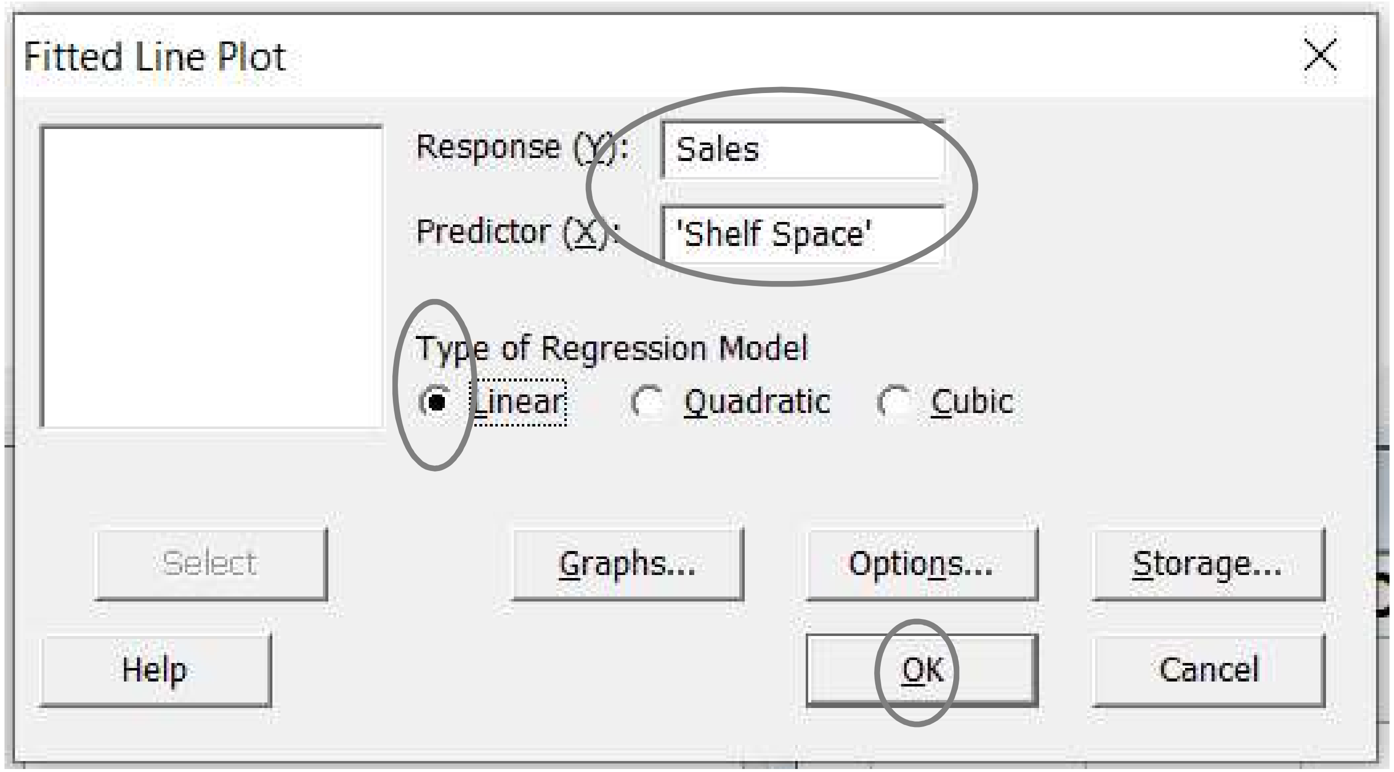
Graph > Scatter Plot





Example: Cereal Sales

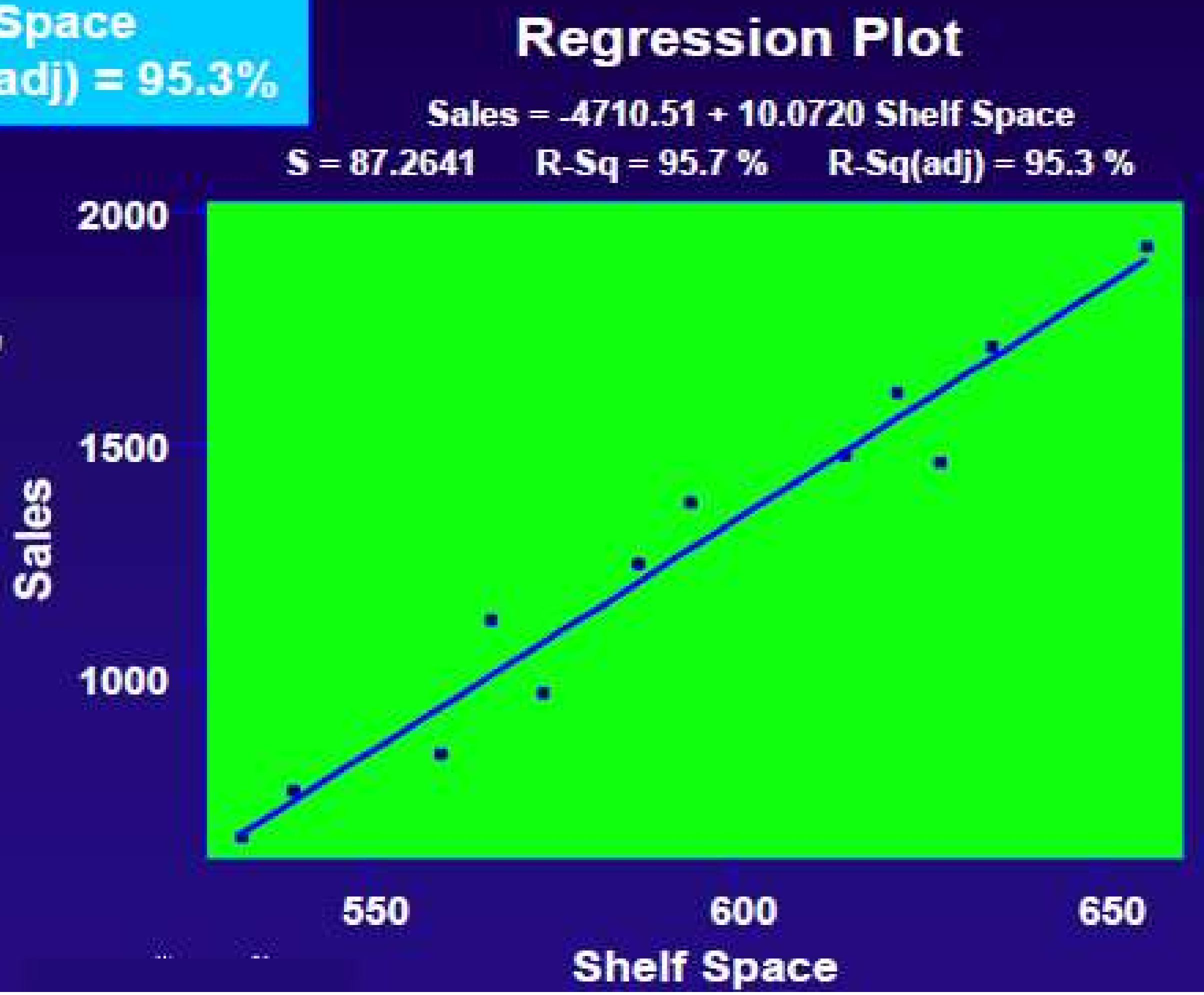
- Tool Bar Menu > Stat > Regression > Fitted Line Plot



Example: Cereal Sales

The regression equation is
Sales = -4710.51 + 10.0720 Shelf Space
S = 87.2641 R-Sq = 95.7 % R-Sq(adj) = 95.3%

- Also from previous, correlation coefficient, r = 0.978
- What do these numbers mean?



Example: Cereal Sales

Session Output from Minitab

Regression Analysis: Sales versus Shelf Space

The regression equation is

Sales = -4710.51 + 10.0720 Shelf Space

S = 87.2641 R-Sq = 95.7 % R-Sq(adj) = 95.3 %

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1709656	1709656	224.511	0.000
Error	10	76150	7615		
Total	11	1785806			

Regression is significant

What About R-squared?

- R-squared is a measure describing the quality of regression
- Measures the proportion of variation that is explained by the regression model
- $R^2 = SS_{regression} / SS_{total} = (SS_{total} - SS_{error}) / SS_{total} = 1 - [SS_{error}/SS_{total}]$

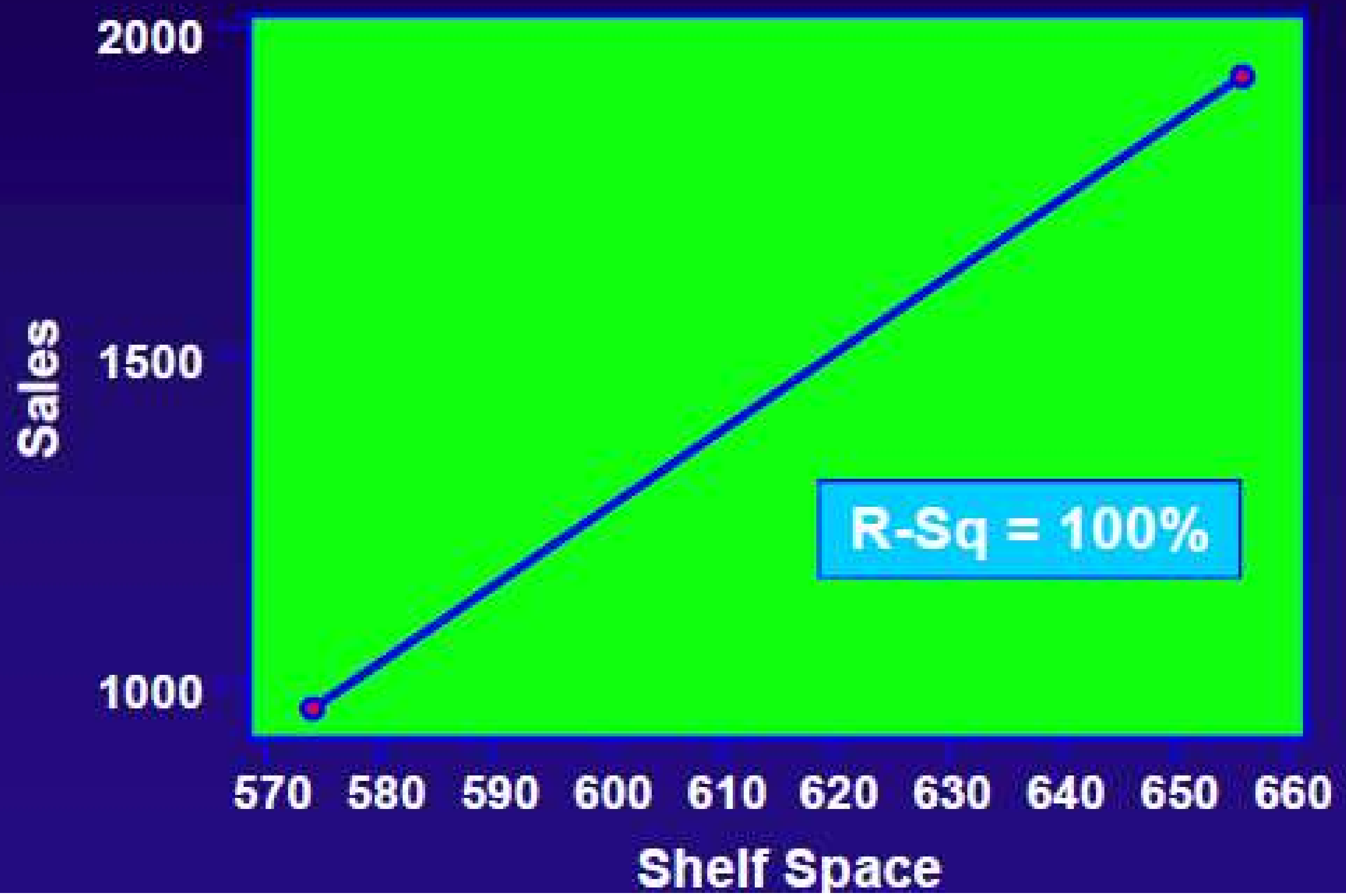
Source	DF	SS	MS	F	P
Regression	1	1709656	1709656	224.511	0.000
Error	10	76150	7615		
Total	11	1785806			

$$R^2 = 1709656 / 1785806 = 95.74\%$$

95.7% of variation in sales can be explained by variation in shelf space

What About R-Sq?

- What is the R-squared on a regression with two data points?
- Does that mean a model with two data points is better?

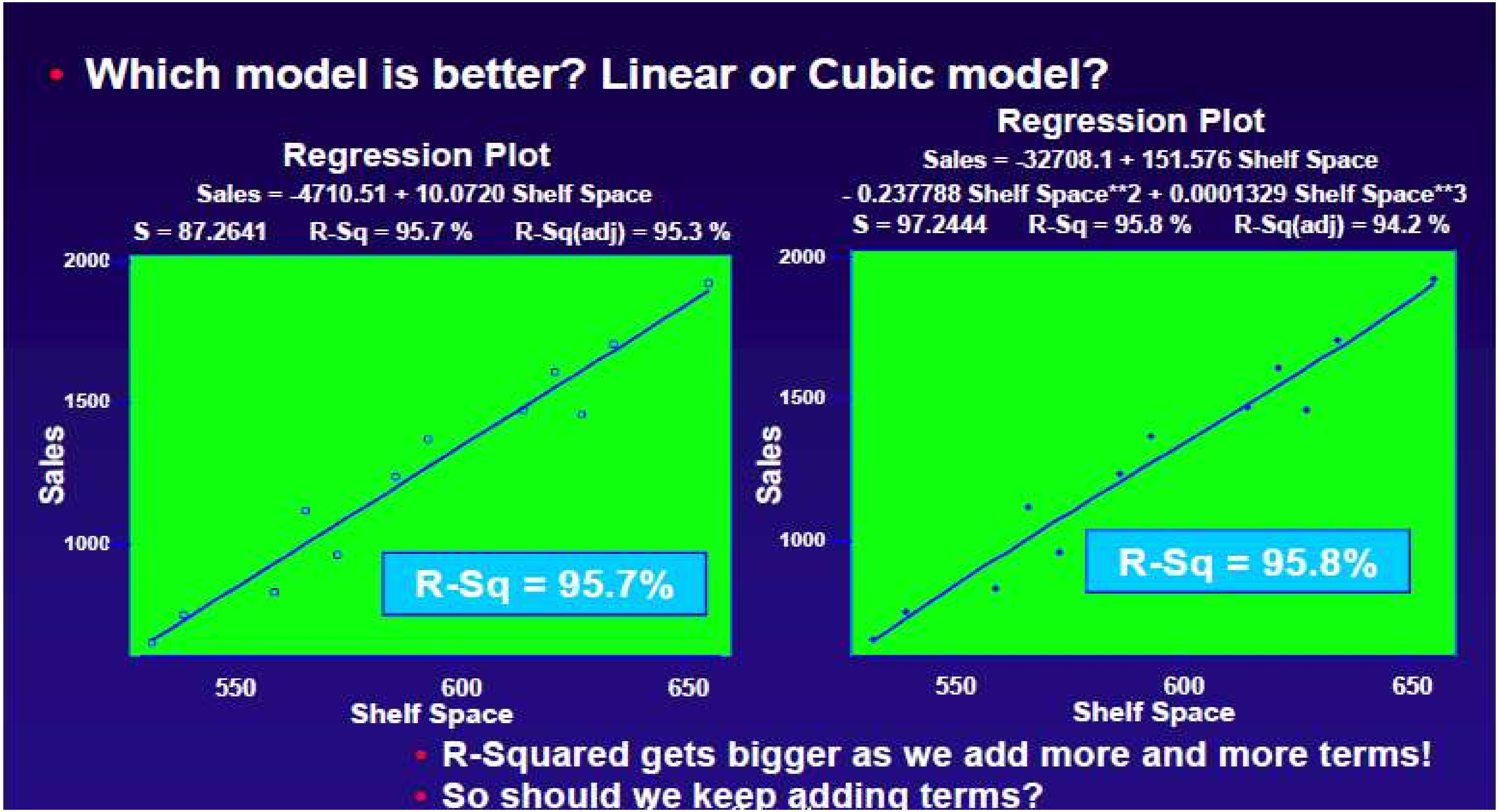


Example: Cereal Sales

Tool Bar Menu > Stat > Regression > Fitted Line Plot

- What is the R-squared if we choose a 'cubic' polynomial regression?

Example: Cereal Sales



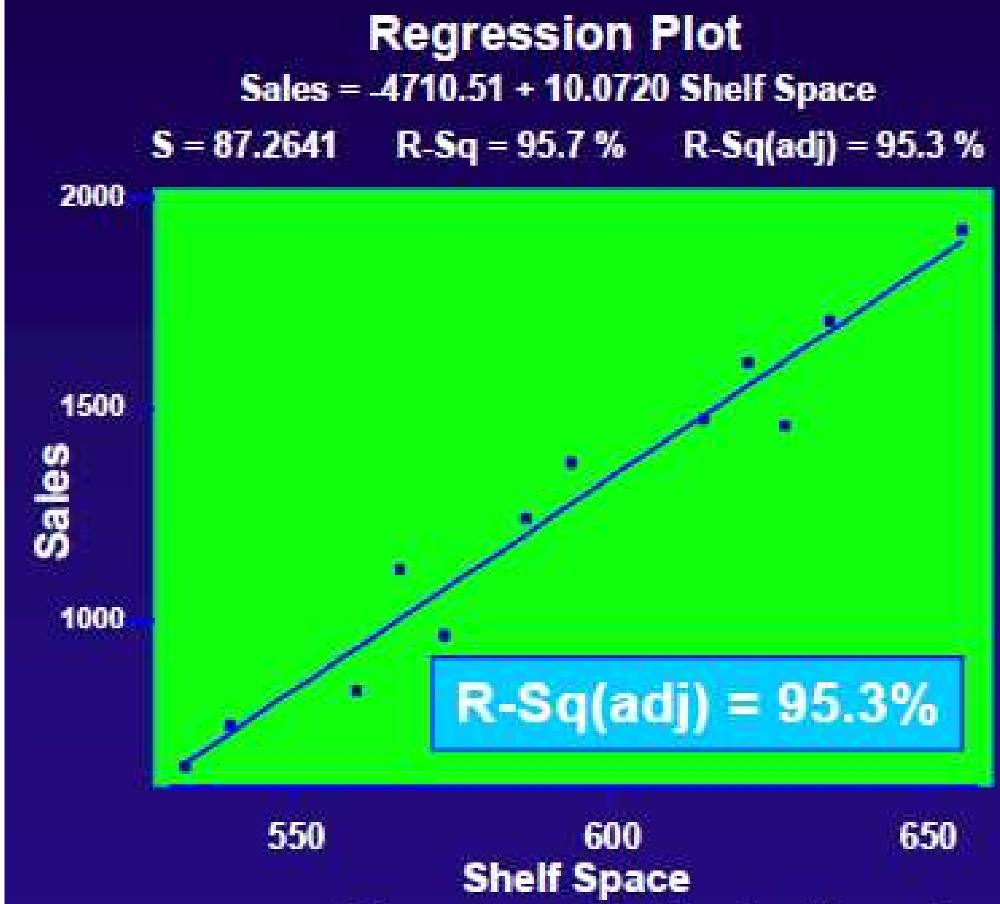
What is R-Sq (adj)?

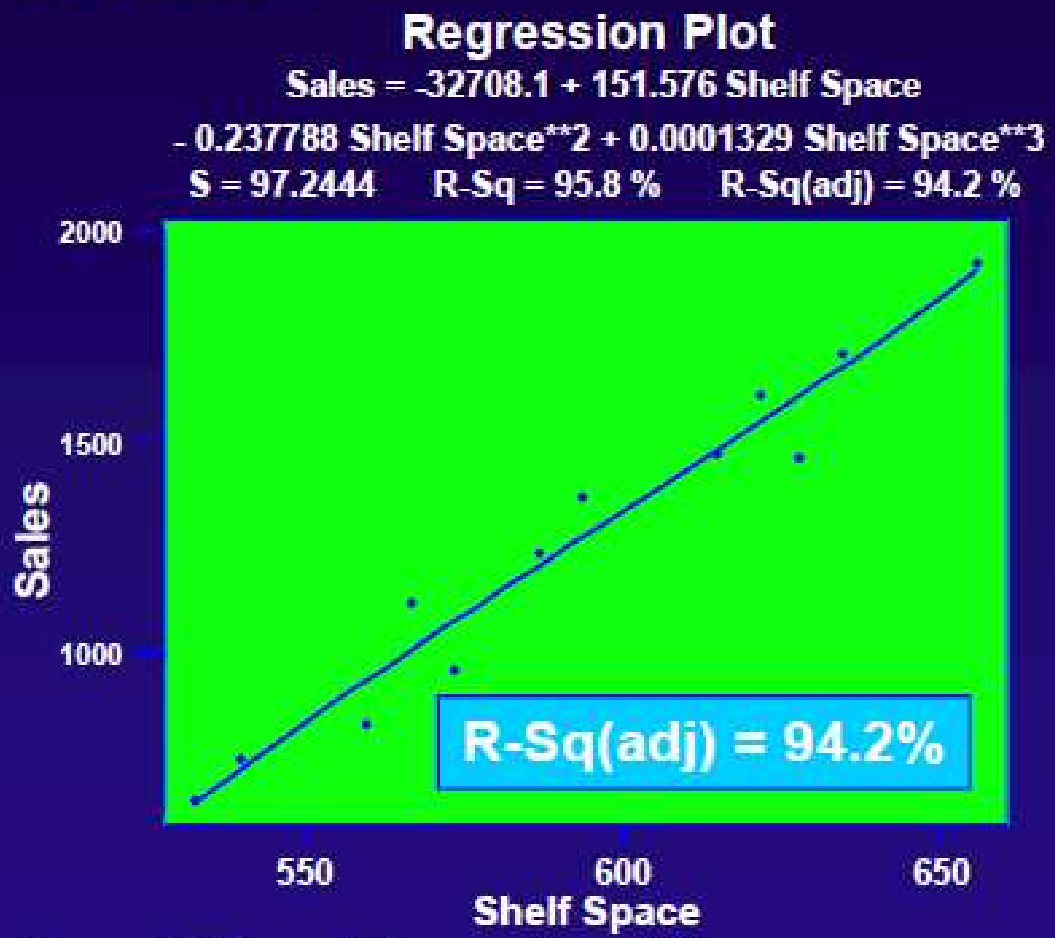
- More realistic measurement and is a modified measure of R-squared
- Takes into account of number of terms in the model and number of data points
- Adj $R^2 = 1 - [SS_{error} / (n-p)] / [SS_{total} / (n-1)] = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$

  where n = number of data points and p = number of terms in the model
- Becomes smaller when added terms provide little new information and as the number of model terms gets closer to the total sample size

Example: Cereal Sales

- Which model is better? Linear or Cubic model?





Linear model is better since the additional terms in cubic model did not add value. How about a quadratic model?

anexas consultancy SE

Example: Cereal Sales

The regression equation is

Sales = -4710.51 + 10.0720 Shelf Space

S = 87.2641 R-Sq = 95.7 % R-Sq(adj) = 95.3%

Predict 'Sales' for 615 'Shelf Space' in the above equation

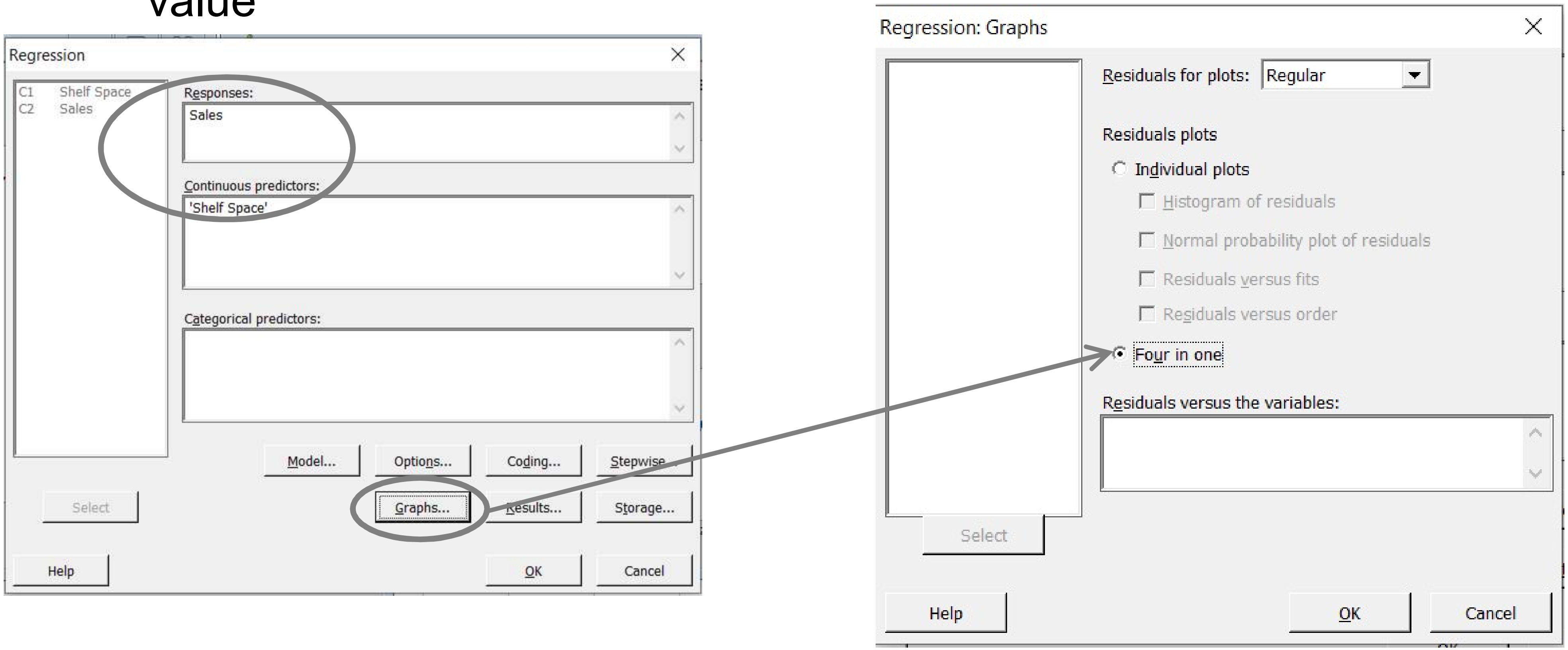
- Substitute the value for 'Shelf Space' in the above equation
- Sales = -4710.51 + 10.072 (615) = $1483.77
- What about the uncertainty around this prediction? Is sales expected to be exactly $1483.77?

Checking Assumptions

Tool Bar Menu > Stat > Regression > Regression

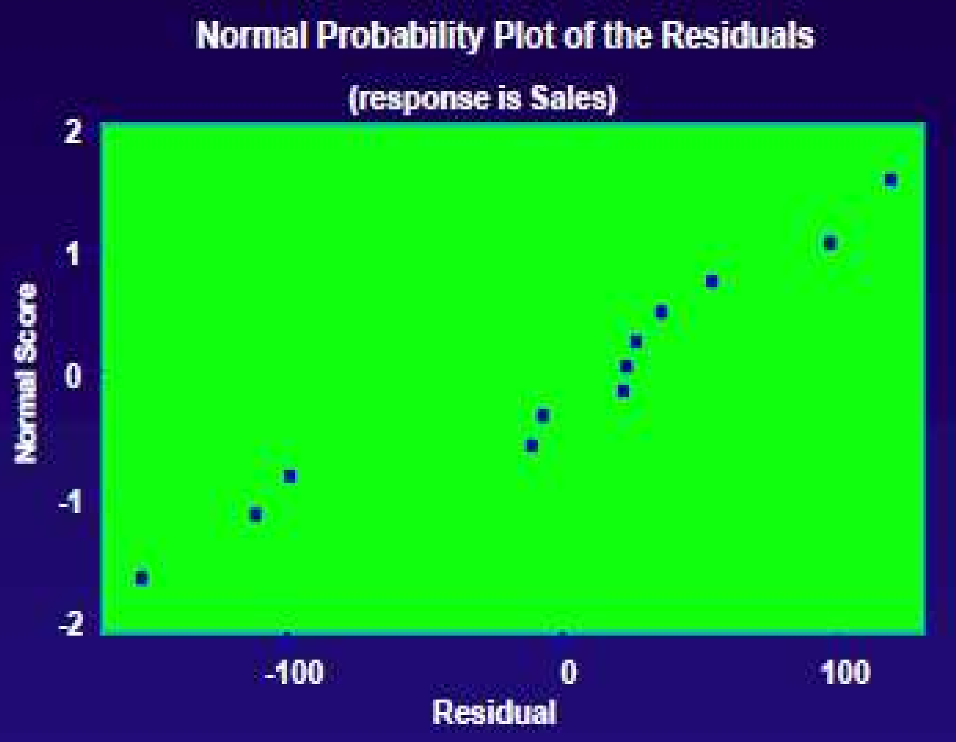
Residuals are error in the fit of regression line

- Difference between the observed value of response variable and fitted value

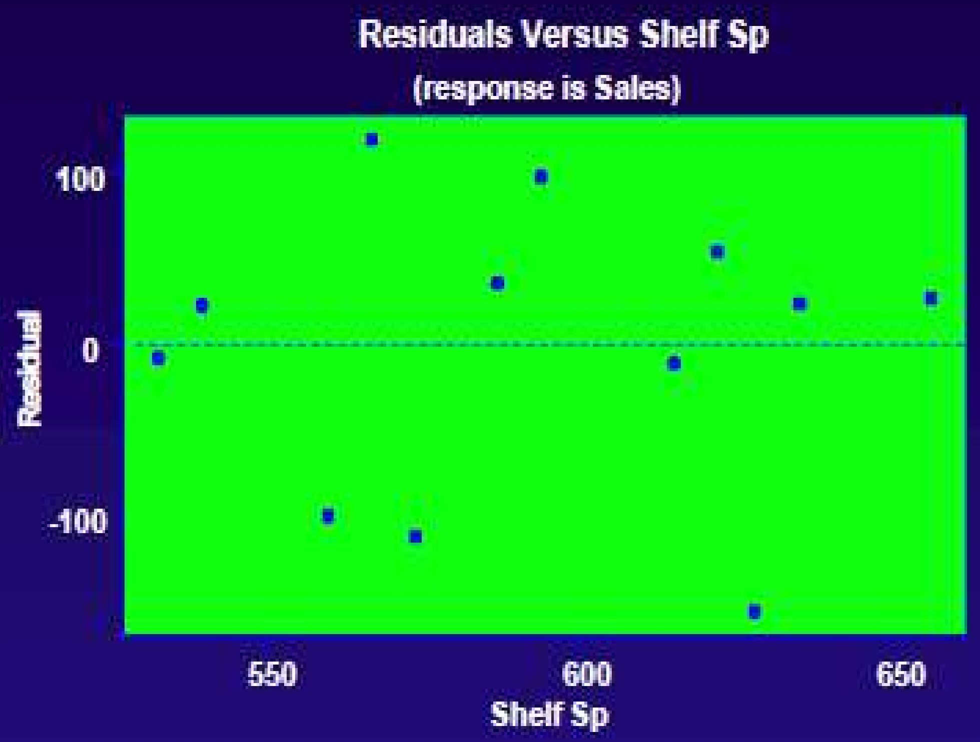


anexas consultancy SE

Assumptions for Regression



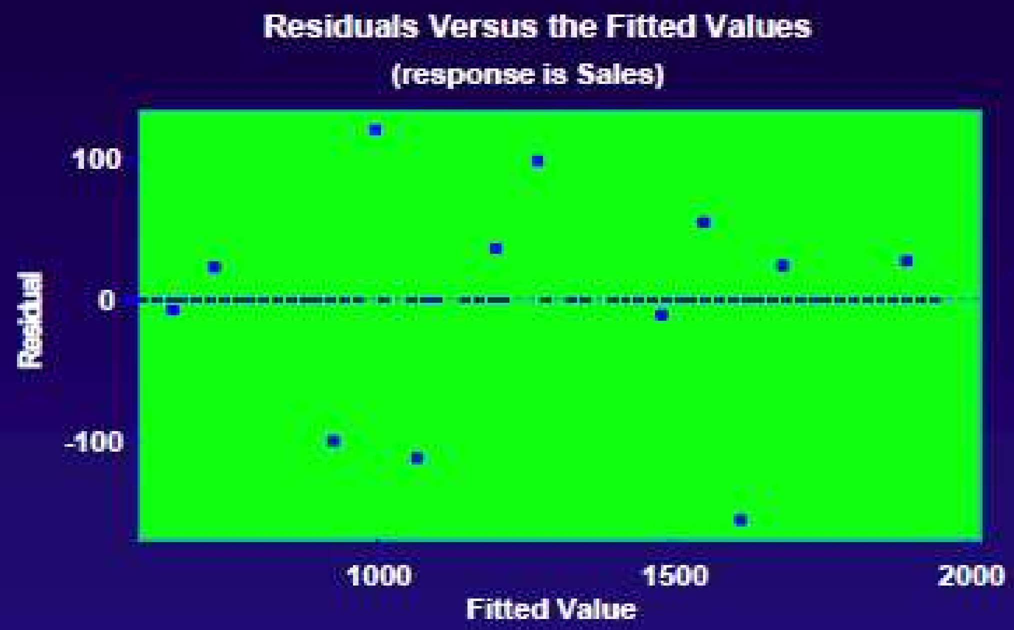
Residuals are normally distributed around mean of zero

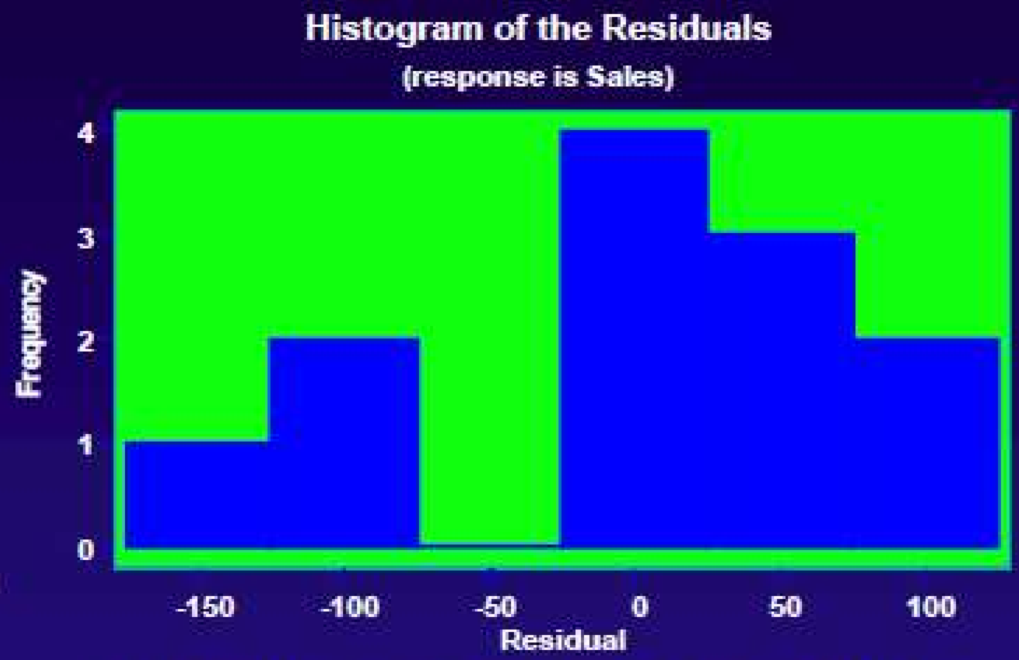


Residuals are independent of 'Shelf Space' variable

anexas consultancy SE

Assumptions for Regression





Residuals do not exhibit heteroscedasticity (they have homogenous variance)

Residuals are randomly distributed

Histogram of residuals resemble a normal distribution with mean of zero

No assumptions were violated; regression results are valid

Multiple Regression

Module Objectives

By the end of this module participant will be able to:

- Determine, for a given response variable, the key process input variables from a set of multiple input variables
- Perform multiple linear regression for a given set of response variable using several input variables
- Perform model diagnostics and validate assumptions
- Use regression model to predict the value of a response variable for given values of predictor variables

Why Learn Multiple Regression?

- Explore the existence of relationship between a dependant variable and several independent variables
- Screen multiple input variables and determine which variables have the biggest impact on the response variable
- Describe the nature of relationship with an equation and use it for prediction

What is Multiple Regression?

- Procedure of establishing relationship between a continuous type response variable and two or more independent variables
- Multiple regression equation can be used to predict a response based on values of predictor variables
- Multiple regression equation takes the form

$$Y = f\ (x_1, x_2, x_3, \dots)$$

Types of Multiple Regression

Multiple Linear Regression

Multiple regressor (x) variables such as x_1, x_2, x_3 and model linear with respect to coefficients

Example1: $y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \text{error}$

Example2: $y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_2^2 + \text{error}$

Multiple Non-Linear Regression

Multiple regressor (x) variables such as x_1, x_2, x_3 and model non-linear with respect to coefficients

Example: $y = (a_0 + a_1 x_1) / a_2 x_2 + a_3 x_3 + \text{error}$

This module focuses on multiple linear regression applying general least squares method

Multicollinearity

- A condition in which two or more independent variables (x variables) are correlated (pairwise and more complex linear relationships)
- When used in multiple regression model, they contribute to redundant information
- For example, fuel economy of a truck = f (truck load, engine horse power)
- But truck load may be correlated with engine horse power
- Truck load and horse power provide some overlapping information leading to potential problems

Problems Due to Multicollinearity

- Multicollinearity can cause severe problems
  - calculations of coefficients and standard errors are affected (unstable, inflated variances)
  - difficulty in assessing any particular variable's effect
  - opposite signs (from what is expected) in the estimated parameters
  - if two input variables x_1 and x_2 are highly correlated, then p-value for both might be high

Detecting Multicollinearity

- High values of pairwise correlation (generally > 0.8) provide warnings of potential multicollinearity problems
- If the above two variables are strongly correlated, one of them should be removed from regression model

Variance Inflation Factor

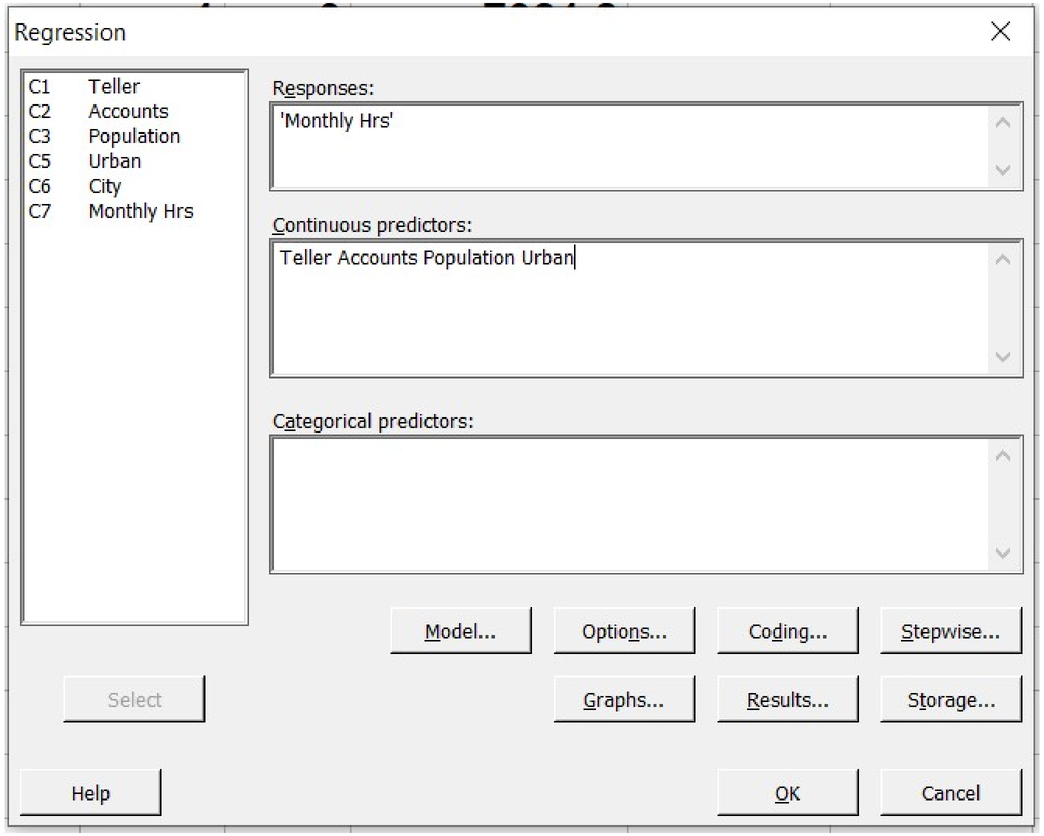
A metric, called variance inflation factor (VIF) calculates the degree of <u>multicollinearity</u>

$$VIF = \frac{1}{1 - R_i^2}$$

- R_i^2 is the R^2 value obtained when X_i is regressed against other X
- A large VIF implies that at least one variable is redundant
- VIF > 10: high degree of multicollinearity - cause for serious concern ($R_i^2 > .9$)
- VIF > 5: moderate degree of multicollinearity ($0.8 < R_i^2 < 0.9$)
- Guideline: Ensure that VIF < 5 when possible

Calculating VIF

Minitab displays VIF values in the session window through *Stat* > *Regression > Regression> Fit Regression Model* menu



Regression Analysis: Monthly Hrs versus Teller, Accounts, Population, Urban

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	34308448	8577112	5345.10	0.000
Teller	1	2719647	2719647	1694.83	0.000
Accounts	1	41	41	0.03	0.876
Population	1	8948965	8948965	5576.83	0.000
Urban	1	123813	123813	77.16	0.000
Error	12	19256	1605		
Total	16	34327704			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
40.0583	99.94%	99.93%	99.90%

Coefficients

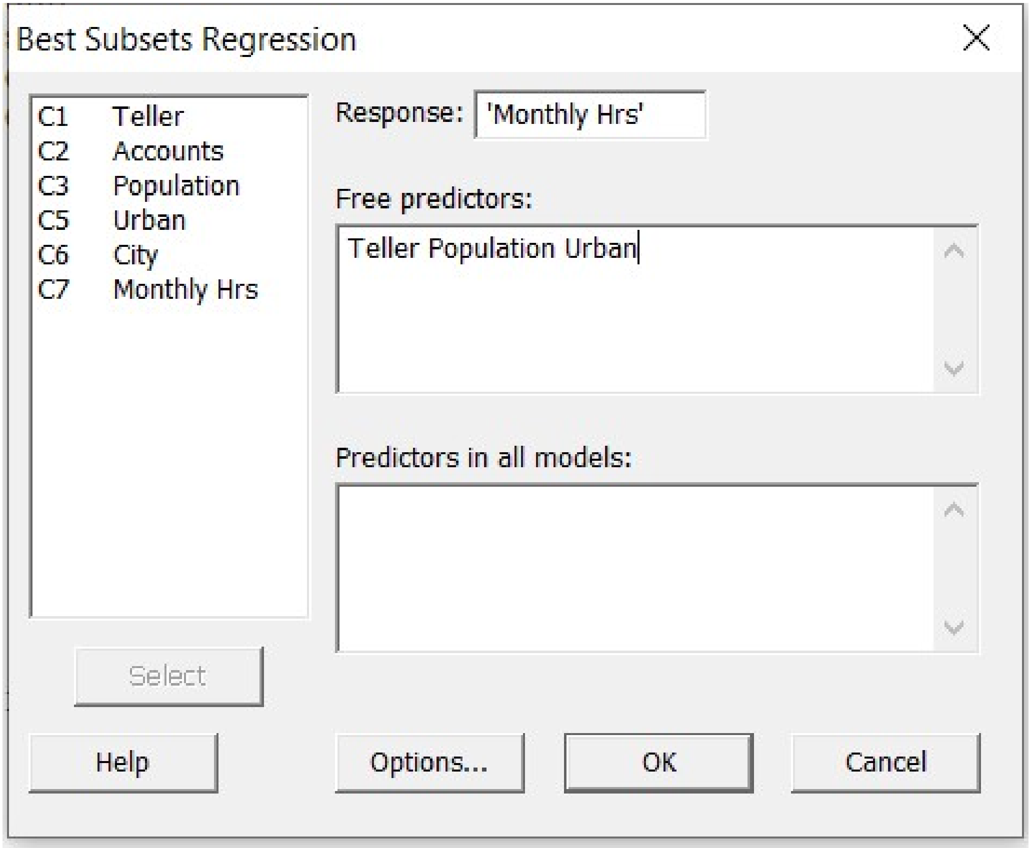
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1094	148	7.41	0.000	
Teller	0.9510	0.0231	41.17	0.000	10.59
Accounts	0.0091	0.0571	0.16	0.876	8.31
Population	0.056381	0.000755	74.68	0.000	1.04
Urban	238.5	27.2	8.78	0.000	1.95

Predictor Variable Selection

- What combination of predictor variables is best for the regression model?
- Three options in Minitab:
  - Stepwise: procedure to add and remove variables to the regression model to produce a useful subset of predictors
  - Best Subsets: procedure to give best fitting regression model that can be constructed with one variable, two variable, three variable, etc. models
  - Regression: once the best model is selected, use Regression to get more detailed diagnostics

Best Subsets

Tool Bar Menu > Stat > Regression > Regression> Best Subsets



Use 'Best Subsets' technique to select a group of likely models for further analysis.

Best Subsets Statistics

- Select the smallest subset that fulfills certain statistical criteria
- Minitab displays R^2, R^2 (adjusted), C-p, and s statistics
  - R^2 (large R^2 is desired; use to compare models with the same number of terms)
  - adjusted R^2 (large is desired; use to compare models with different number of terms)
  - s (standard deviation of error terms; small is desired)
  - Mallows' C-p statistic (small is desired; Guideline: want C-p $\leq$ number of terms in model)

Putting It All Together

Multiple regression objective: Establish a model with high prediction ability and minimum multicollinearity

Multiple regression steps:

1. Remove variables contributing to multicollinearity from the predictors
2. Use remaining variables and apply Best Subsets to evaluate best predictor candidates for the model
3. Choose the best candidate and complete regression analysis
4. Perform model diagnostics to identify outliers and unusual observations
5. Analyze residuals for violation of assumptions
6. Assess predictive capability using new observations

Banking Process Example: Bank Labour Hours

A banking institution wants to produce an empirical equation that will estimate personnel needs its branches. The following data was collected from its existing branches at various locations. The response variable (Y) for the study was monthly labor hours. The input variables were average number of daily teller transactions (x_1), average count of total number of accounts (x_2), location of branch (x_3), population within 20 Km radius (x_4). The data is recorded in the file Banking.mtw.

anexas consultancy SE

Business Process Example: Bank Labor Hours

- Establish a multiple regression model to predict monthly labor hours using the predictor variables
- Perform model diagnostics to detect any outliers or unusual observations
- Validate any assumptions used for creating the model

Example: Bank Labor Hours

- ***Banking.mtw***
  - Output variable: Monthly Hrs
  - Input variables: Teller Accounts, Population, Location
  - Practical questions about the data and process?

File Edit Manip Calc Stat Graph Editor Window Help Six Sigma

↓	C1	C2	C3	C4-T	C5
	Teller	Accounts	Population	Location	Monthly Hrs
1	2842.8	2967	58800	Urban	7381.8
2	2549.2	3243	65200	Urban	7465.0
3	2718.4	3519	70900	Urban	7925.3
4	3683.2	3818	77400	Urban	9265.2
5	5916.8	4439	79300	City	11282.6
6	5825.6	4347	81000	City	11156.2
7	4785.6	4025	71900	City	9714.0
8	5410.4	4278	63900	Urban	10123.0
9	5062.4	4370	54500	Urban	9269.0
10	5647.6	4301	39500	City	8746.2
11	6676.4	4485	44500	City	9963.6
12	5828.4	4738	43600	Urban	9346.3

Example: Bank Labor Hours

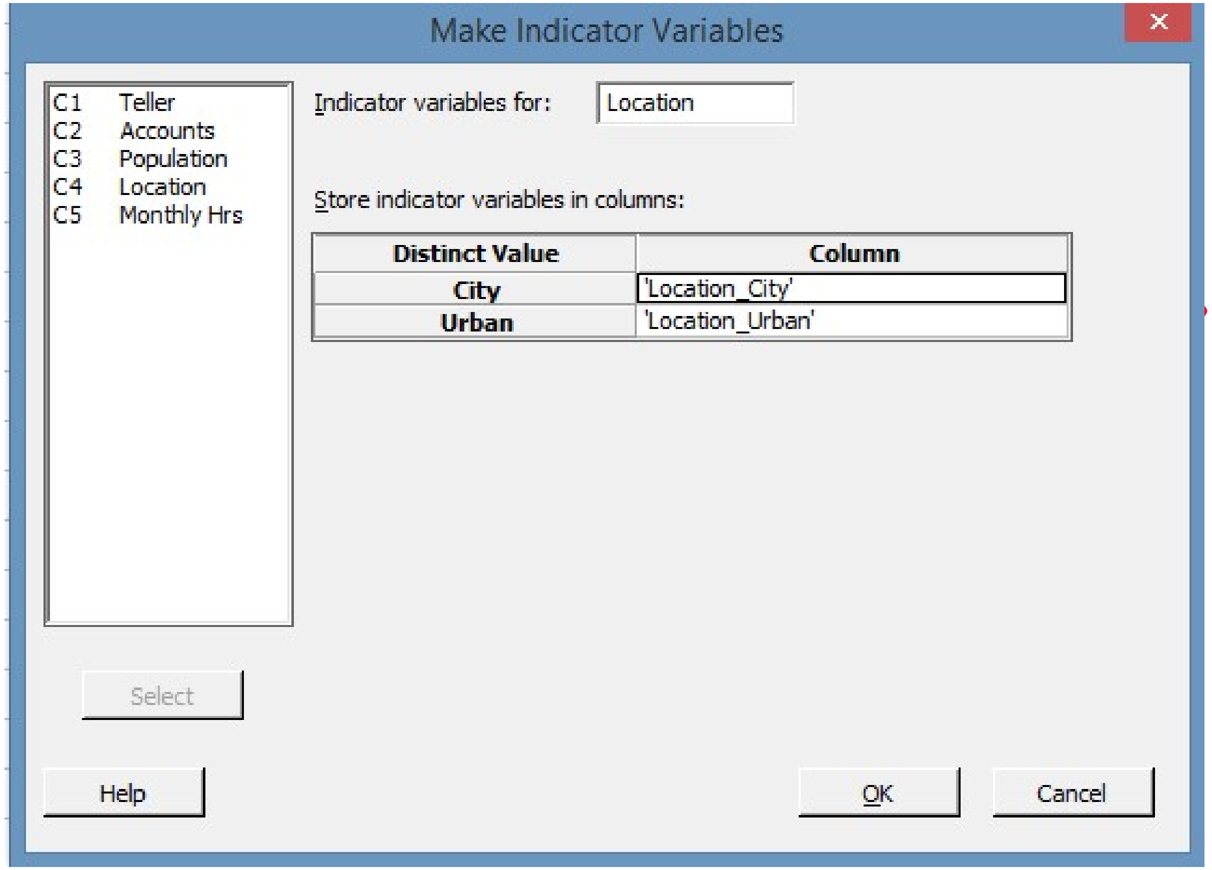
- Dealing with indicator variables (dummy variables)

  - To use independent variables which are categorical (e.g, location, gender) in regression, first create “indicator” variables (dummy variables)
  - Indicators are simply 1’s and 0’s which are used like binary code.
  - Each level of a categorical variable is assigned a column.
  - If a row of data is associated with that level of the variable, the value in that column for that row of data will be 1. If not associated with that level, the value will be 0.

Example: Bank Labor Hours

1. <u>Create an indicator variable for "Location"</u>

   Toolbar>Calc> Make indicator Variables



Example: Bank Labor Hours

- Result should follow the pattern below
- <u>Order</u> in which the columns are named is important. For alphanumeric data, the columns are created in the order specified as the value order (alphabetic, order of appearance, user defined).

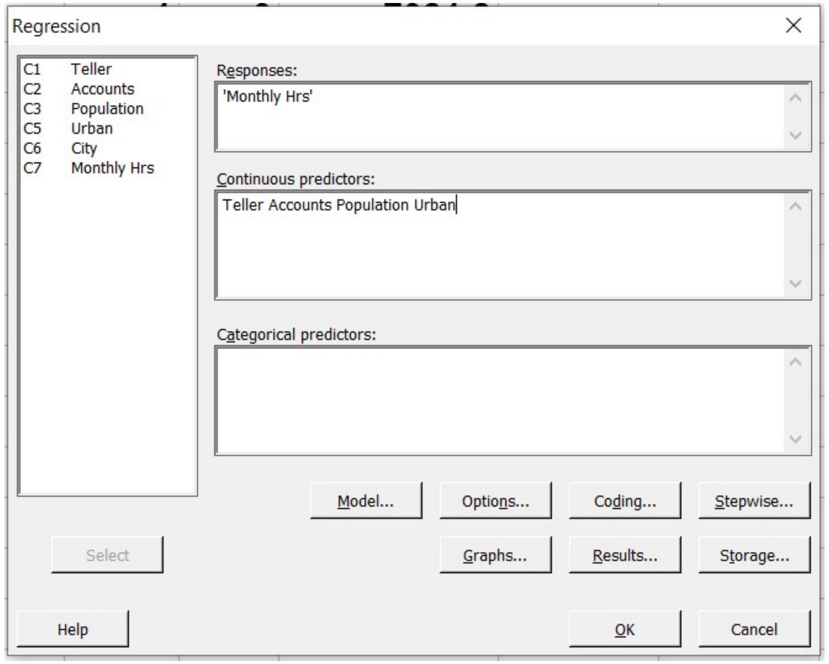
Location	Urban	City
Urban	1	0
Urban	1	0
Urban	1	0
Urban	1	0
City	0	1
City	0	1
City	0	1
Urban	1	0
Urban	1	0
City	0	1
City	0	1
Urban	1	0

Example: Bank Labor Hours

- <u>Create an indicator variable for “Location”</u>
- When categorical variables are used in the regression model, one column of indicator values is NEVER included.
- In the example, use either “Urban” or “City” since it would be redundant to use both. (If “Urban”=0, then the observation must be “City”)

Example: Bank Labor Hours

- <u>Multiple regression steps</u>:
- 1. Remove variables contributing to multicollinearity from the predictors
- Identify if multicollinearity is a problem by using variance inflation factors (VIF) measurements
- 1. *Stat> Regression> Regression> Fit Regression Model*
- 2. Fill in the dialog as shown below:



Example: Bank Labor Hours

Serious multicollinearity problem!
Want VIF <5

```
The regression equation is
Monthly Hrs = 1094 + 0.951 Teller + 0.0091 Accounts + 0.0564
Population + 238 Urban
```

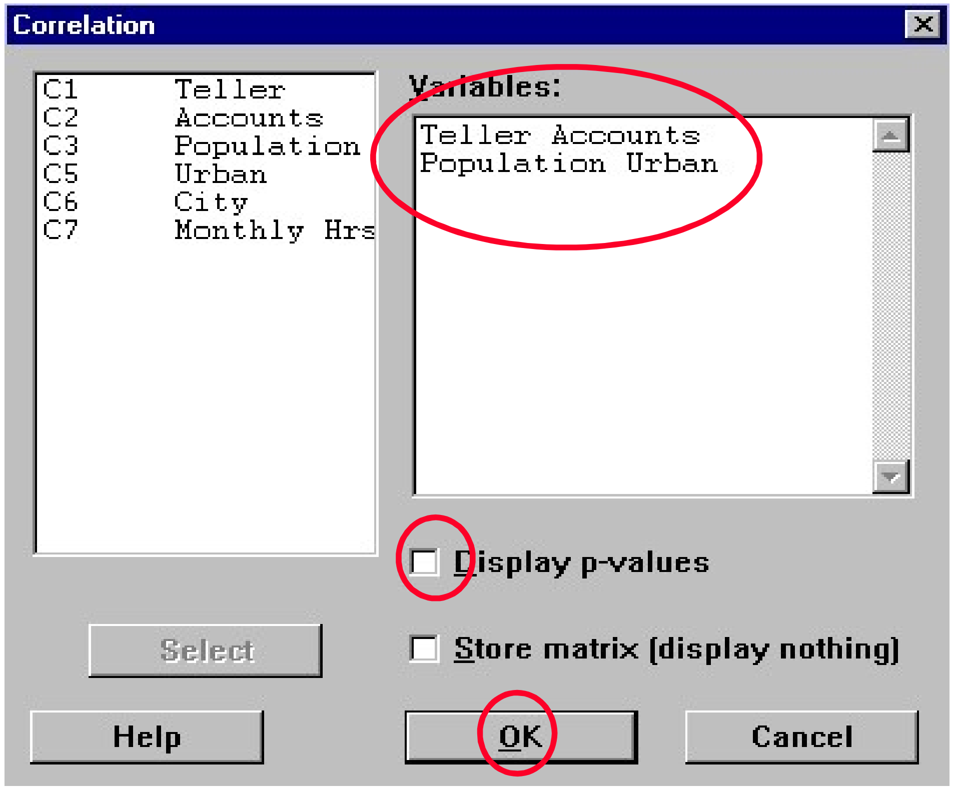
Predictor	Coef	SE Coef	T	P	VIF
Constant	1094.3	147.8	7.41	0.000	
Teller	0.95103	0.02310	41.17	0.000	10.6
Accounts	0.00913	0.05706	0.16	0.876	8.3
Populati	0.0563806	0.0007550	74.68	0.000	1.0
Urban	238.49	27.15	8.78	0.000	1.9

Example: Bank Labor Hours

b. Identify pairwise correlations between x variables

1. *Stat> Basic Statistics> Correlation..*
2. Fill in the dialog as shown below:

3: Press Ok



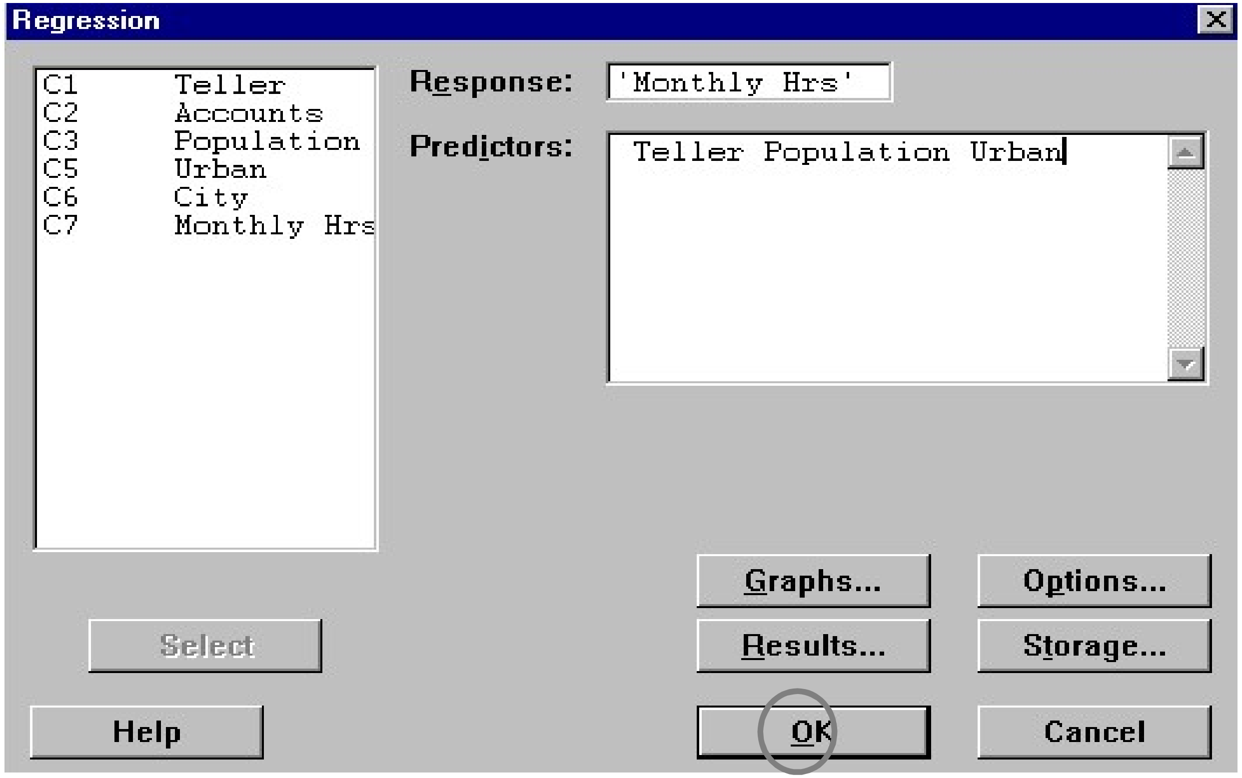
Example: Bank Labor Hours

Next step: remove Accounts and check multicollinearity

	Teller	Accounts	Populati
Accounts	0.918		
Populati	-0.024	-0.082	
Urban	-0.573	-0.372	-0.126

Example: Bank Labor Hours

Remove “Accounts” and repeat regression and check for VIF values



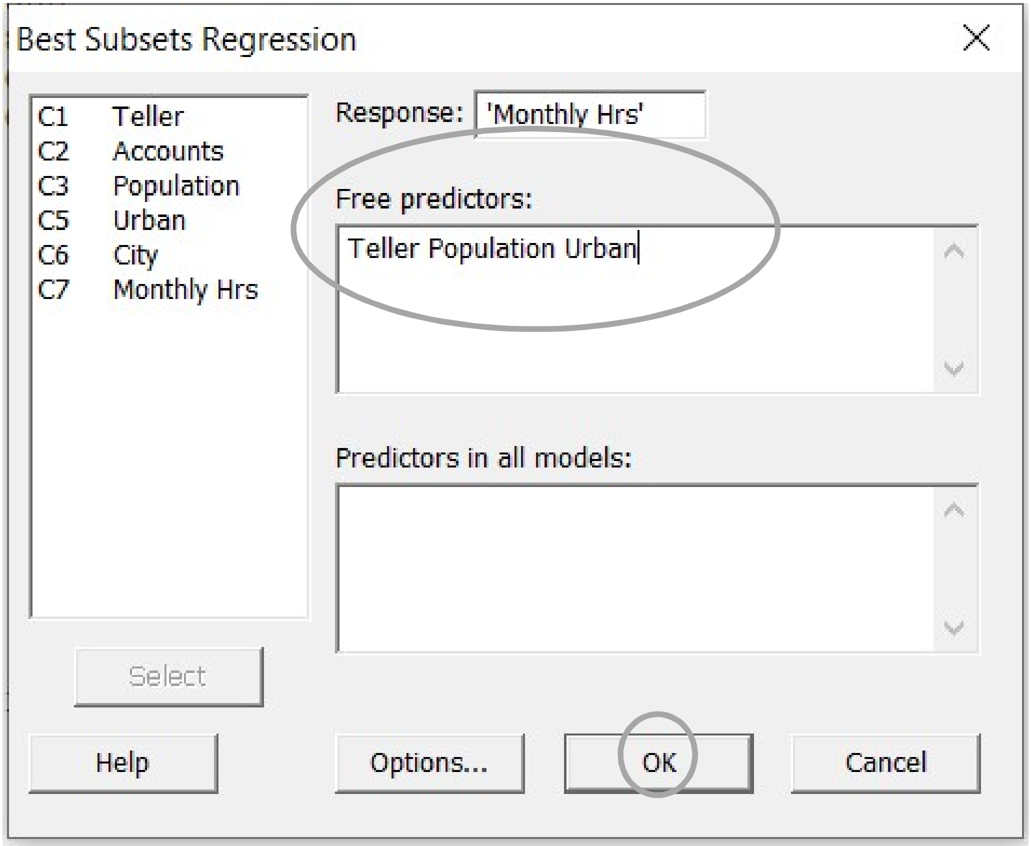
Example: Bank Labor Hours

All VIF <5

Predictor	Coef	SE Coef	T	P	VIF
Constant	1114.38	74.88	14.88	0.000	
Teller	0.954454	0.008388	113.78	0.000	1.5
Populati	0.0563707	0.0007237	77.89	0.000	1.0
Urban	240.49	23.18	10.38	0.000	1.5

Example: Bank Labor Hours

- Multiple regression steps:
- 2. Use remaining variables and apply Best Subsets to evaluate best predictor candidates for the model



- 1. *Stat> Regression> Best Subsets..*
- 2. Fill in the dialog as shown below:
- 3: Press Ok

anexas consultancy SE

Example: Bank Labor Hours

Response is Monthly

Vars	R-Sq	R-Sq(adj)	C-p	S	Teller	Population	Urban
1	73.7	71.9	6075.6	776.22	X		
1	25.7	20.8	2E+04	1303.6			X
2	99.5	99.4	109.7	113.11	X	X	
2	73.7	70.0	6069.0	802.90	X		X
3	99.9	99.9	4.0	38.528	X	X	X

anexas consultancy SE

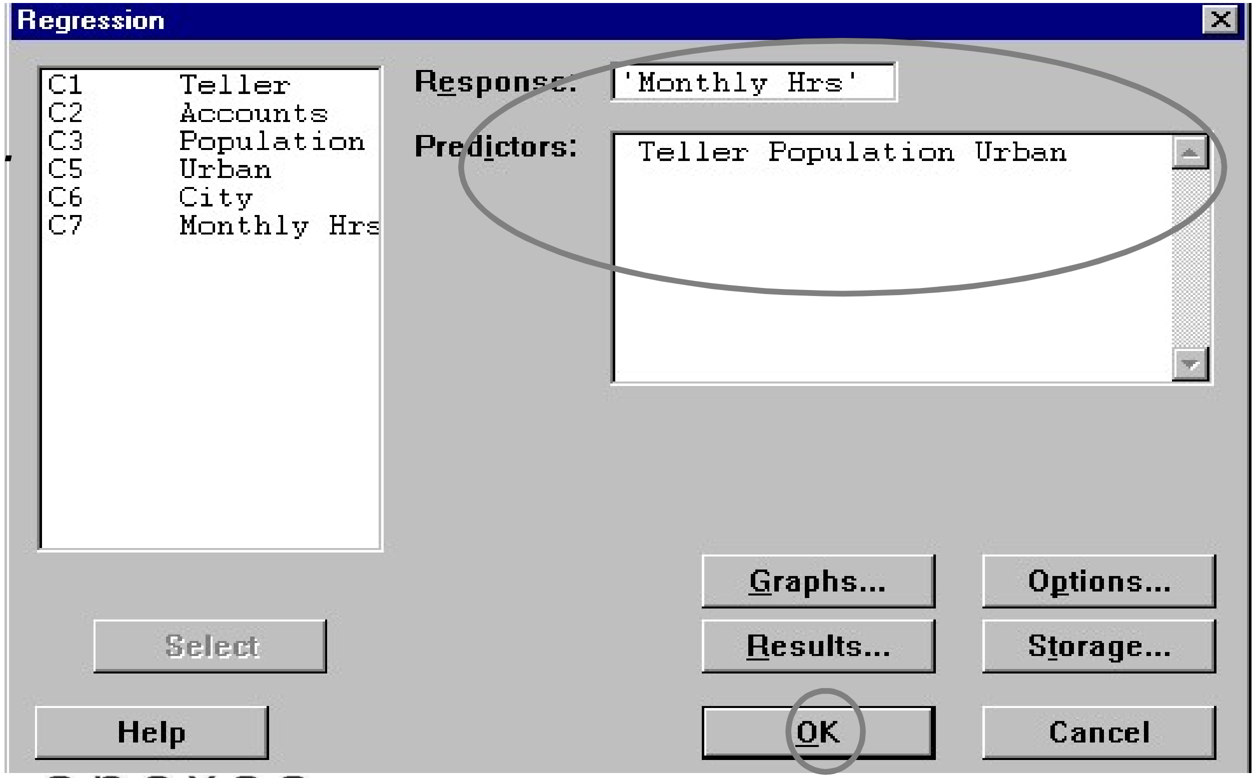
Example: Bank Labor Hours

- Multiple regression steps:

3. Choose the best candidate and complete regression analysis

1. *Stat> Regression> regression...*
2. Fill in the dialog as shown below:

3: Press Ok



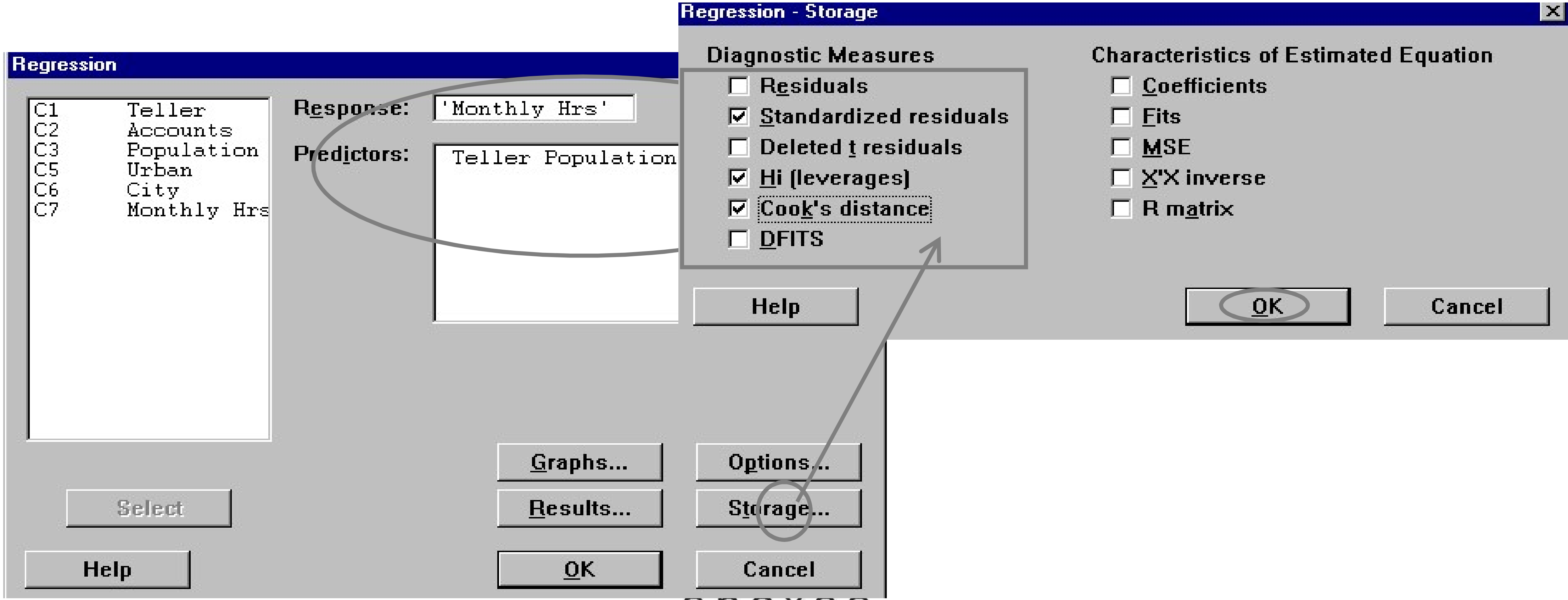
Example: Bank Labor Hours

```
The regression equation is
Monthly Hrs = 1114 + 0.954 Teller + 0.0564 Population +
240 Urban

Predictor        Coef     SE Coef          T        P
Constant      1114.38       74.88      14.88    0.000
Teller       0.954454    0.008388     113.78    0.000
Populati    0.0563707   0.0007237      77.89    0.000
Urban          240.49       23.18      10.38    0.000

S = 38.53       R-Sq = 99.9%     R-Sq(adj) = 99.9%
```

Example: Bank Labor Hours

Multiple regression steps:

4. Perform model diagnostics to identify outliers and unusual observations



Example: Bank Labor Hours

Outliers/ Unusual Observations

C7	C8	C9	C10
Monthly Hrs	SRES1	HI1	COOK1
7381.8	-0.03088	0.243202	0.000077
7465.0	0.05043	0.275047	0.000241
7925.3	-0.62889	0.256412	0.034096
9265.2	0.92022	0.197013	0.051941
11282.6	1.45481	0.181257	0.117139
11156.2	-2.44597	0.197162	0.367313
9714.0	-0.60984	0.197691	0.022910
10123.0	0.05897	0.158210	0.000163
9269.0	0.28358	0.152478	0.003617
8746.2	0.49623	0.398760	0.040829
9963.6	-0.98303	0.304931	0.105985
9346.3	-0.91801	0.315137	0.096947
9734.0	0.65238	0.172624	0.022199
10390.2	1.78362	0.127810	0.116546
10773.6	-0.36476	0.139764	0.005404
12673.2	0.62914	0.293429	0.041094
11776.1	-0.57219	0.389073	0.052128

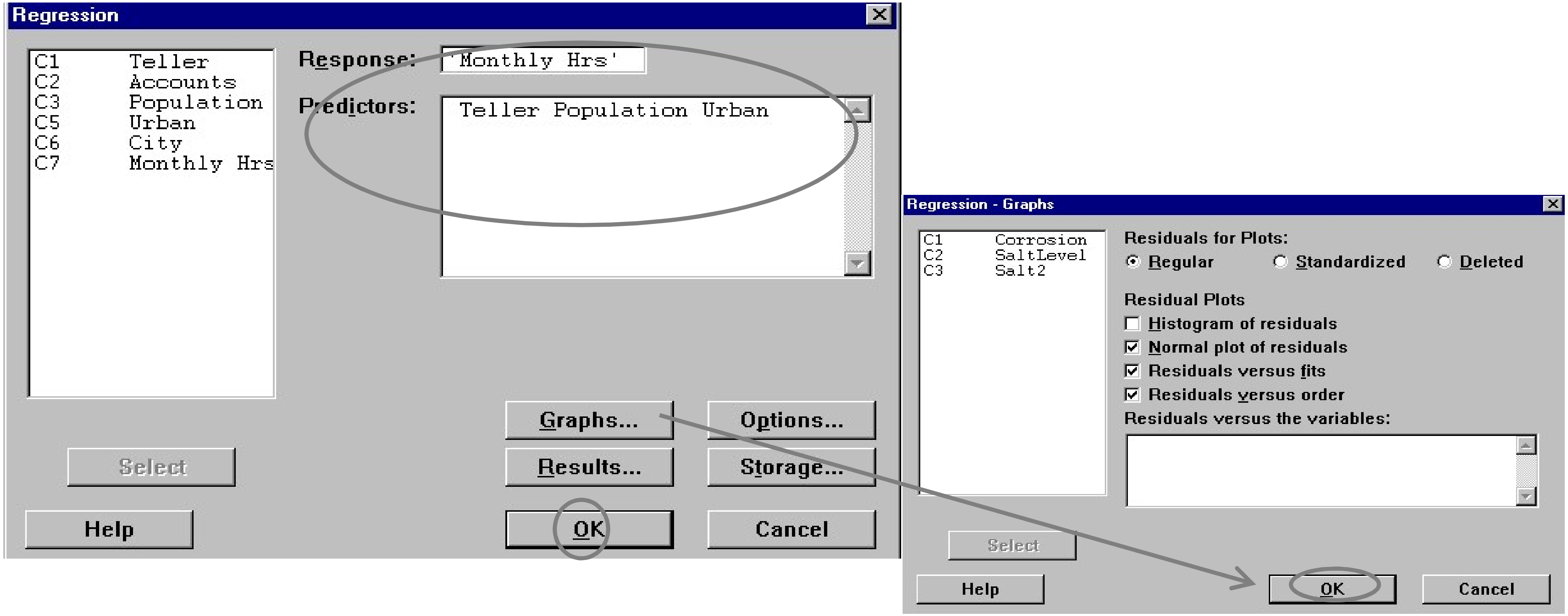
Potential trouble spots:
Standardized residual>2
Leverage > 2p/n = 0.4706
Cook's distance > 1

What actions should be taken with the outliers and influential observations, if any ?

Example: Bank Labor Hours

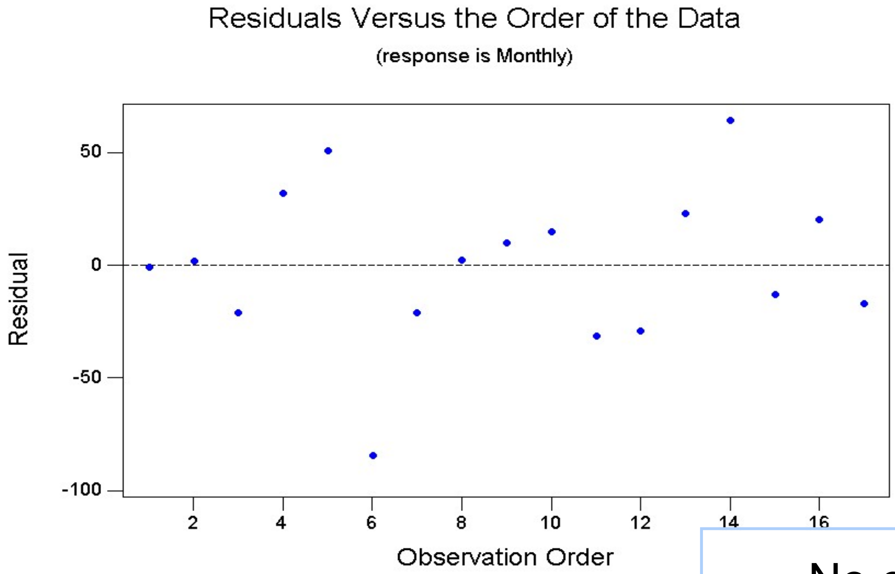
Multiple regression steps:

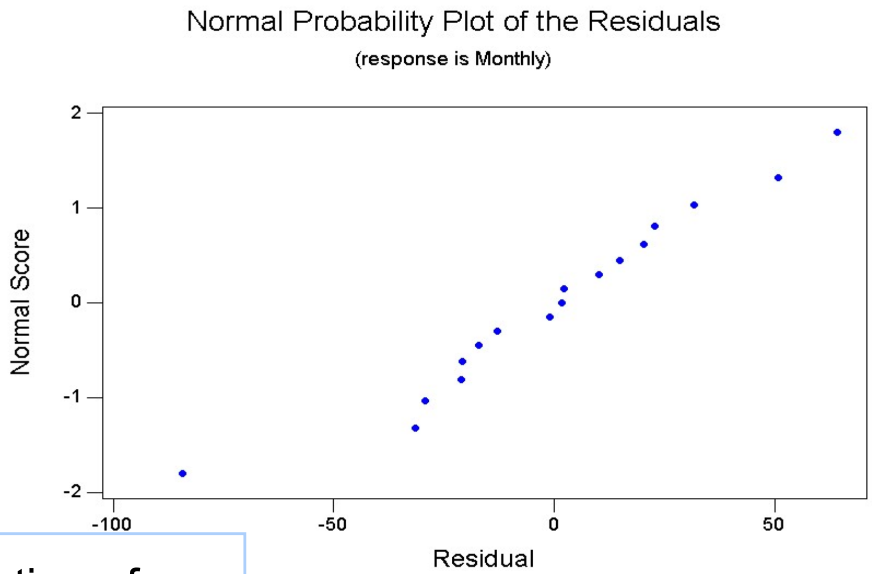
5. Analyze residuals for violation of assumptions



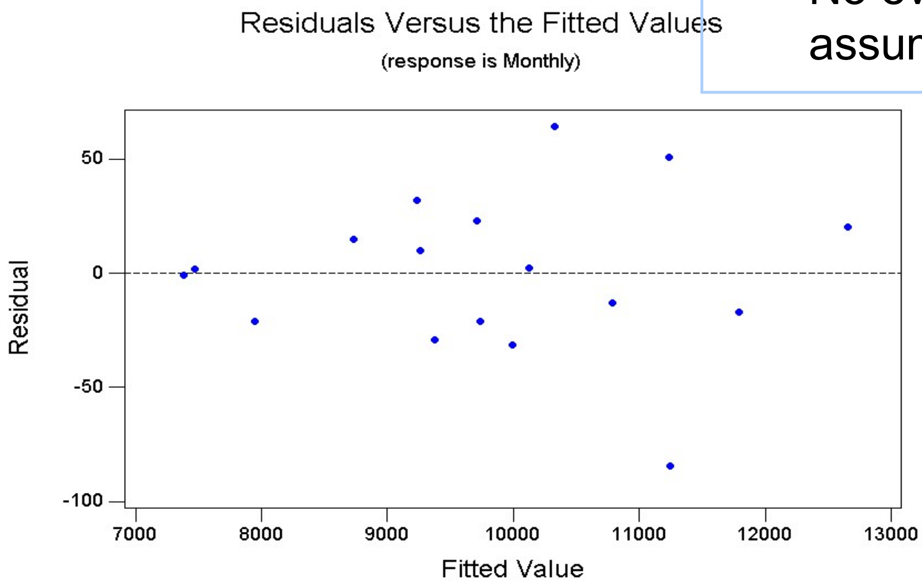
Example: Bank Labor Hours

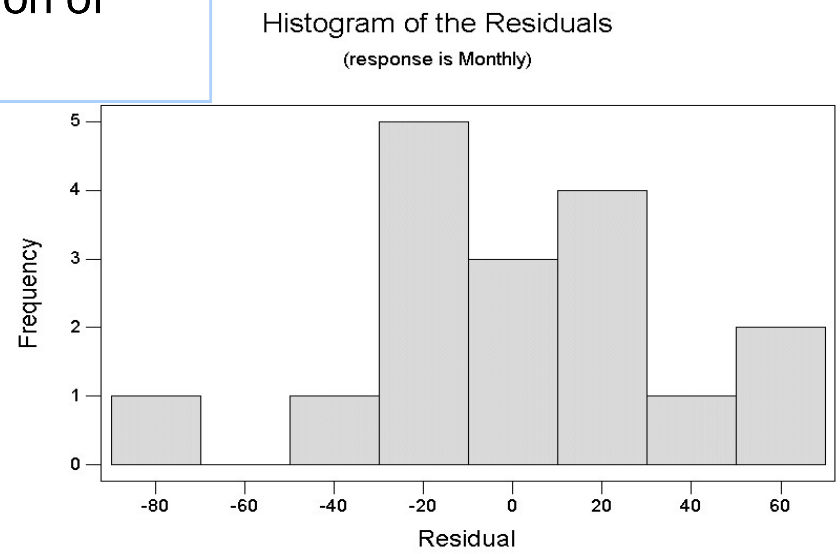
Residual Plot Analysis





No evidence of violation of assumptions





Logistic Regression Analysis

Discrete Ys

Days Between Maintenance and Chances of Good Start-Ups

- Suppose you work on a chemical blending process for a small-volume, highly regulated product
- It is a one-shift operation and you must shut down and clean the blender at the end of each day
- There is no preventive maintenance schedule in place: maintenance is called after certain mechanical problems are observed and procedures are done at the end of the shift
- You are interested in establishing a preventive maintenance schedule and would like to learn how many days you can go between maintenance procedures and still have a chance of a good start-up each day (machine starts on first attempt)

Days Between Maintenance and Chances of Good Start-Ups, cont.

- You check the log for the last 89 maintenance procedures
  - What data would you collect to answer this question?
  - How would you display them?
  - Can you predict an attribute outcome (yes/no) with a continuous variable?

Days Between Maintenance and Chances of Good Start-Ups, cont.

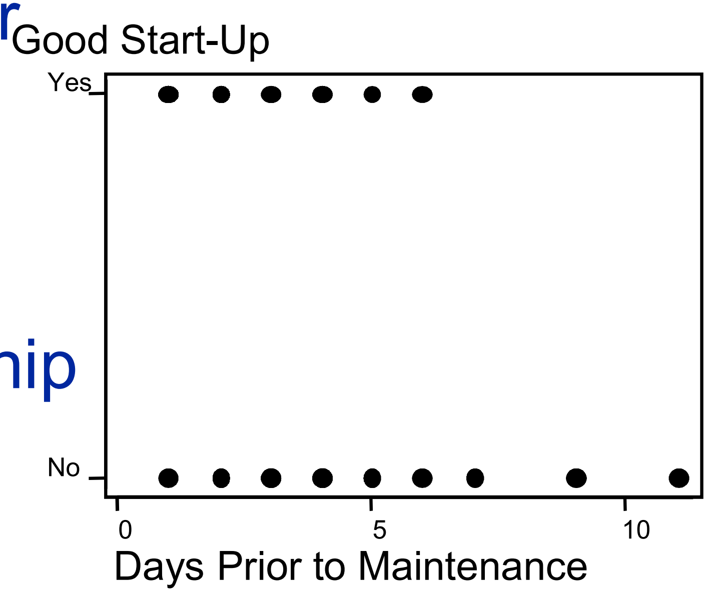
Maintenance Procedure	Days Prior to Maintenance	Good Start-Up First Run Next Day?
1	2	Yes
2	1	Yes
3	7	No
4	2	Yes
5	5	No
.	.	.
86	2	Yes
87	4	No
88	2	Yes
89	3	Yes

anexas consultancy SE

Possible Ways to Display Discrete Y-Data

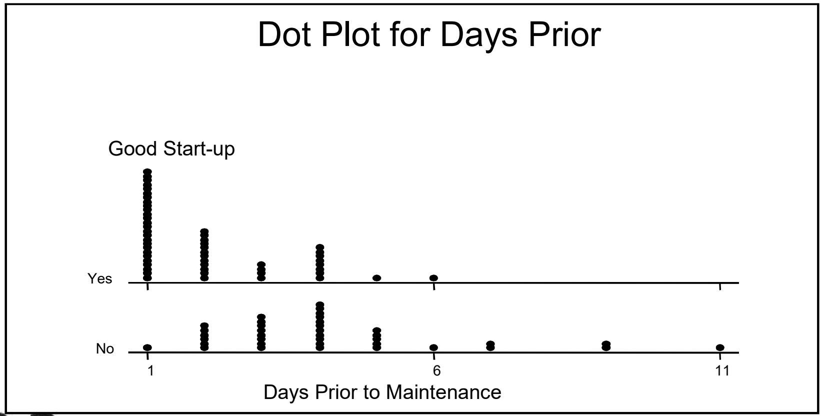
1. Scatter plot of raw data

- Inappropriate data for scatter plot—Y is discrete-attribute, not continuous
- Thus, the plot is not useful; cannot understand relationship or make predictions based on this pattern



2. Stratified dot plot

- Useful
- Can start to see patterns that reveal information about the relationship

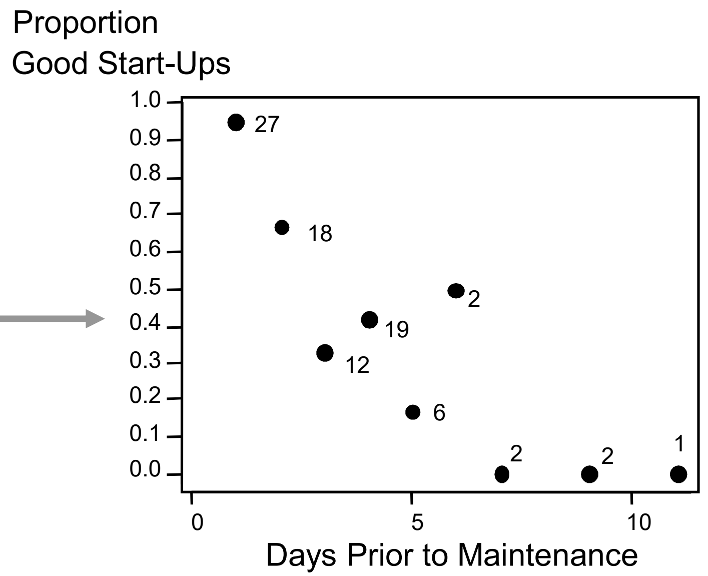


Possible Ways to Display Discrete Y-Data, cont.

3. Scatter plot of summarized data

- Very useful
- Can pick out patterns to help us answer our question

Each point is labeled with the number of maintenance procedures at those conditions



Use Proportions for Discrete Data

- Logistic regression predicts the **proportion** or **probability** of a particular Y attribute
- In other words, you would like to know the percent of times the chemical blender works right the first run of the day after a maintenance procedure

Raw Data

Maintenance Procedure	(X) Days Prior to Maint	(Y) Good Start-Up
1	2	Yes
2	1	Yes
3	7	No
.	.	.
87	4	No
88	2	Yes
89	3	Yes

A particular Y-attribute

X = continuous data

Y = discrete attribute data

Appropriate Data for Y in Logistic Regression

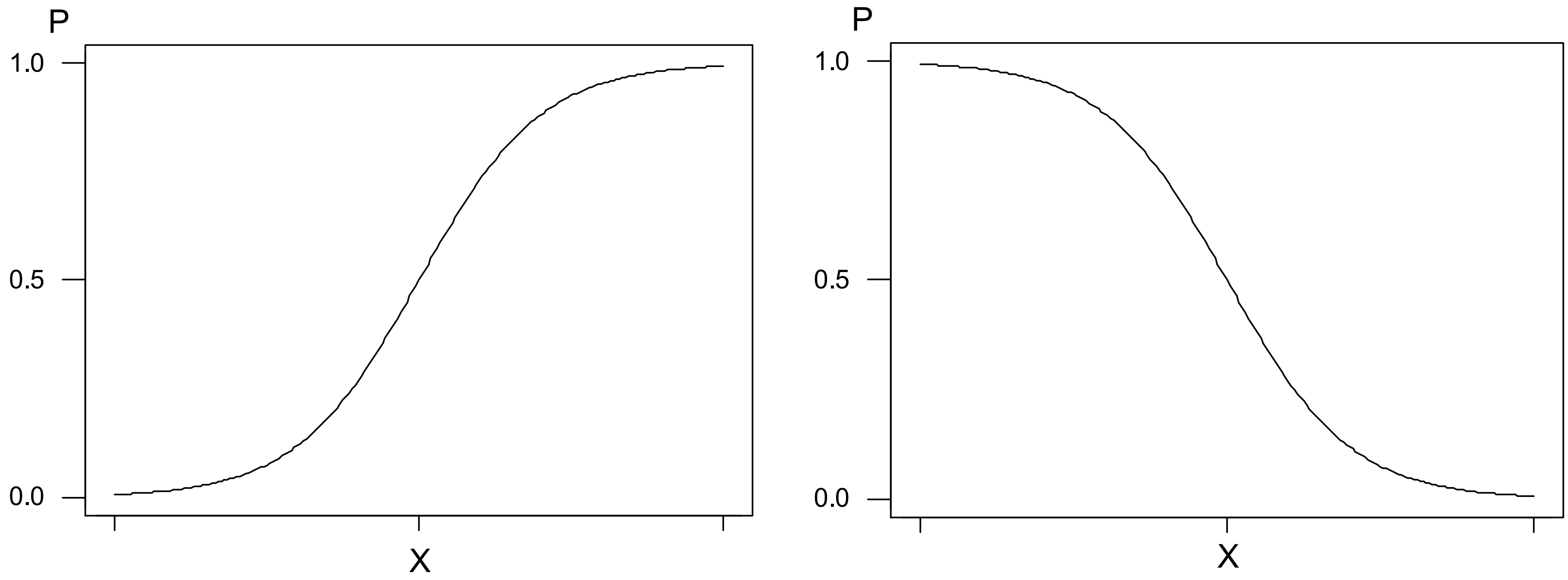
Type of data for Y

- Discrete, attribute
- Two levels only
    - “Success”
    - “Failure”
- Choose one and call it an **event**
    - People are usually more interested in successes, but you can just as easily predict failures
    - Minitab will want to know which level interests you and will call it the “event” (of interest)

Logistic Regression Fits a Curve to the Data

The relationship between an event probability (p) and a predictor (X) is usually curved, not straight

- It looks like a tilted "S," backward "S," or some portion of an "S"
- It has asymptotes at 0 and 1

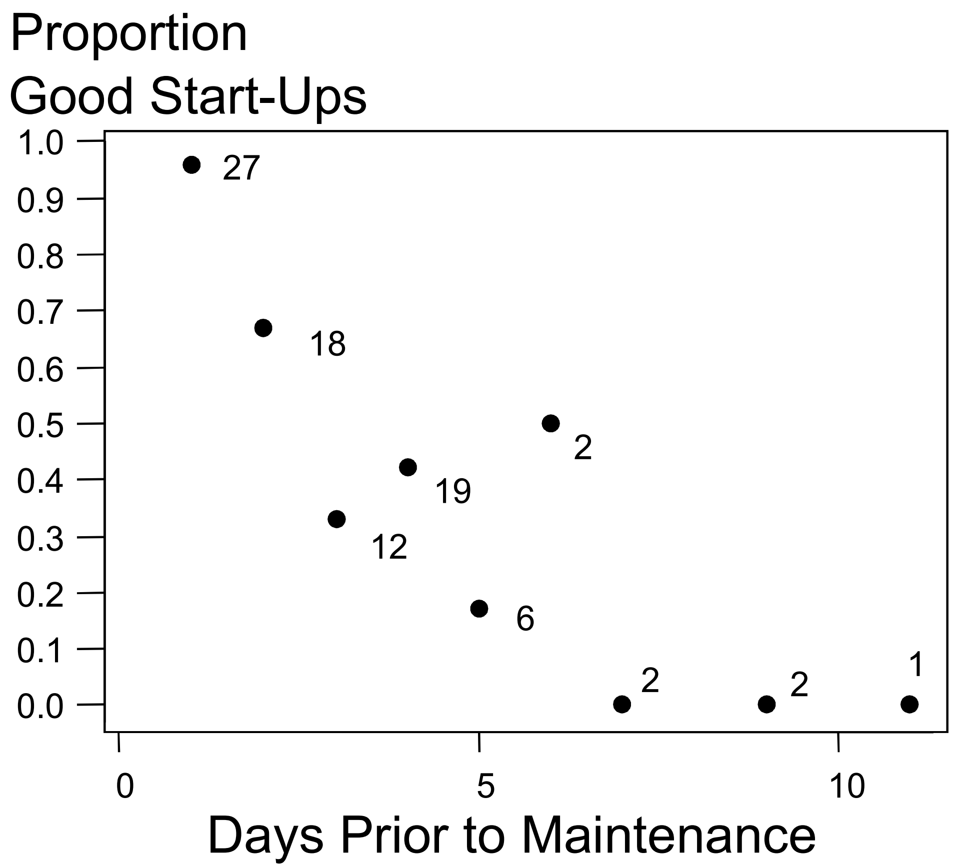


anexas consultancy SE

Logistic Regression Fits a Curve to the Data, cont.

Maintenance example

- The proportion of good start-ups decreases as the days prior to maintenance increase, but not at a constant rate
- Draw a curve that might fit the points on this plot

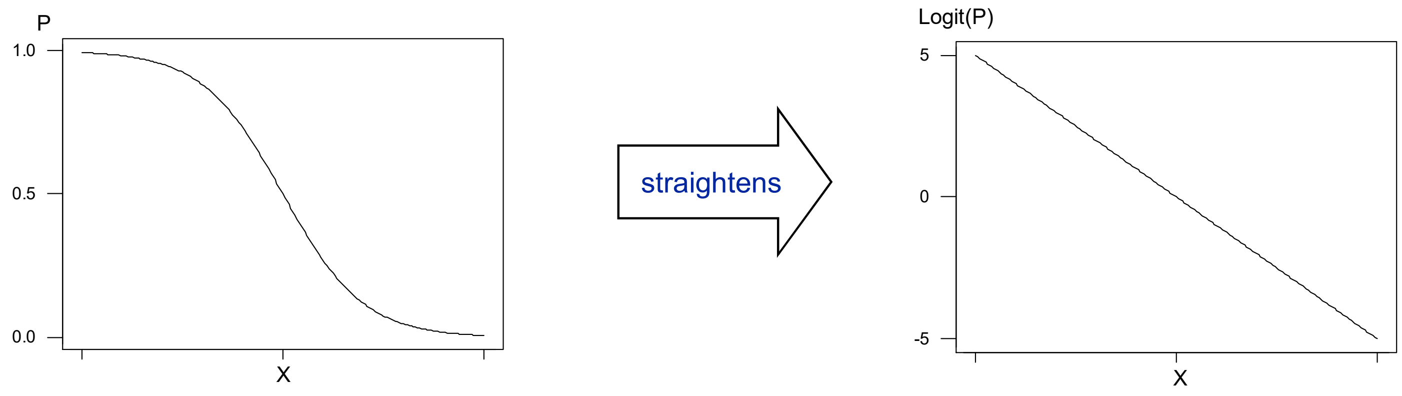


How Does Logistic Regression Work?

- The Y-variable is actually P, the proportion of Y-attributes at each X-value
- The logit transformation straightens the S-shaped relationship between P and X

P (proportion) vs. X

Logit P (transformation) vs. X



- Logistic regression gets its name from this transformation

Doing Logistic Regression

- We will use Minitab to:
  - Obtain b_0 and b_1
  - Back-transform the fits to obtain predicted P's (event probabilities or EPROs) for each value of X
- Logistic regression is an advanced topic
  - Much of Minitab's output is beyond the scope of this course
  - The interpretation of b_1 is not simple
  - We will use plots of the back-transformed fits vs. X to interpret
    the relationship

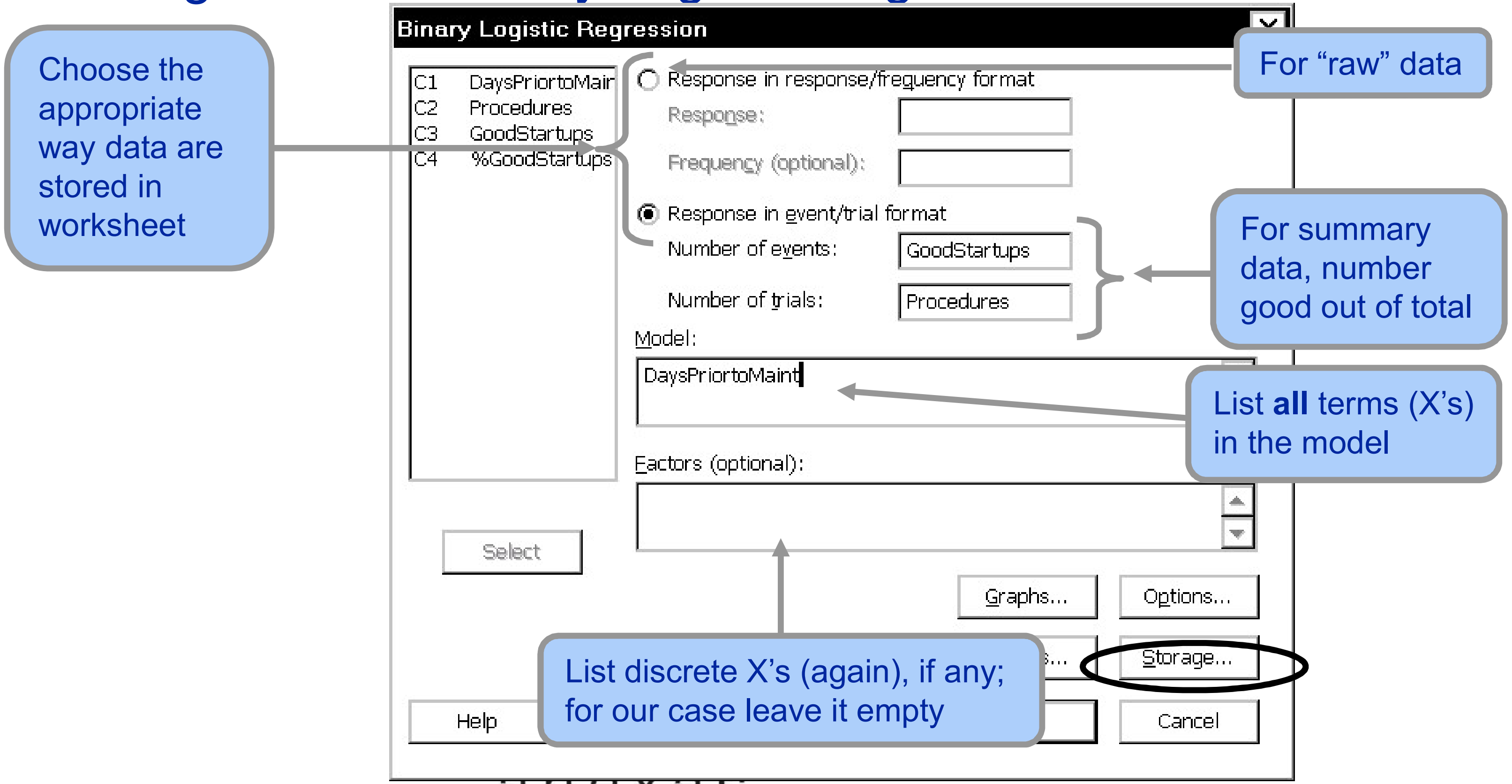
Regression Analysis

Logistic Regression with Discrete Y's

Minitab Follow-Along: Logistic Regression

- **Data: *PrevMain.mtw***
  1. Do a logistic regression analysis on the "Good Start-ups" predicted from "Days Prior to Maint.":

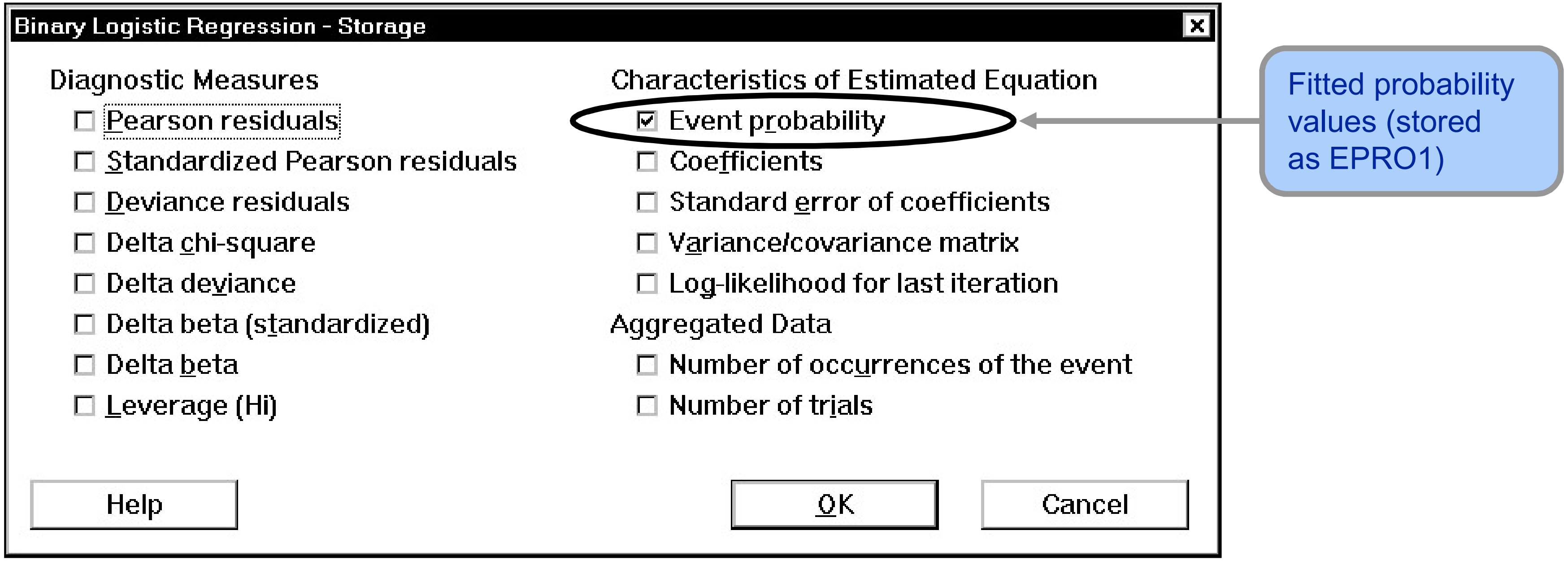
     Stat > Regression > Binary Logistic Regression



Minitab Follow-Along: Logistic Regression, cont.

2. Store the fitted Y's (known as event probabilities) to make a fitted line plot later:

   Stat > Regression > Binary Logistic Regression > Storage



anexas consultancy SE

Minitab Output: Logistic Regression

Binary Logistic Regression: GoodStartups, Procedures versus DaysPriortoM

```
Link Function:  Logit

Response  Information

GoodStartups  Event         52
              Non-event     37
Procedures    Total         89

Logistic Regression Table
                                                              Odds     95% CI
Predictor              Coef   SE Coef       Z      P      Ratio  Lower  Upper
Constant            2.82120  0.617574    4.57  0.000
DaysPriortoMaint  -0.858419  0.196212   -4.37  0.000      0.42   0.29   0.62

LogLikelihood = 44.643
Test that all slopes are zero: G = 31.555, DF = 1, PValue = 0.000

GoodnessofFit Tests
Method                 ChiSquare    DF      P
Pearson                    9.772     7  0.202
Deviance                   8.497     7  0.291
HosmerLemeshow             5.769     3  0.123

Table of Observed and Expected Frequencies:
(See HosmerLemeshow Test for the Pearson ChiSquare Statistic)
                      Group
Value      1     2     3     4     5   Total
Success
  Obs      2     8     4    12    26      52
  Exp    1.4   6.7   6.7  13.5  23.7
Failure
  Obs     11    11     8     6     1      37
  Exp   11.6  12.3   5.3   4.5   3.3

  Total   13    19    12    18    27      89

Measures of Association:
(Between the Response Variable and Predicted Probabilities)
Pairs          Number  Percent    Summary Measures
Concordant       1483    77.1%    Somers' D                 0.66
Discordant        217    11.3%    GoodmanKruskal Gamma     0.74
Ties              224    11.6%    Kendall's Taua           0.32
Total            1924   100.0%
```

Logistic regression equation:

$$\ln \frac{p}{(1-p)} = 2.82 - 0.858(\text{Days Prior})$$

Both slope and intercept are significant

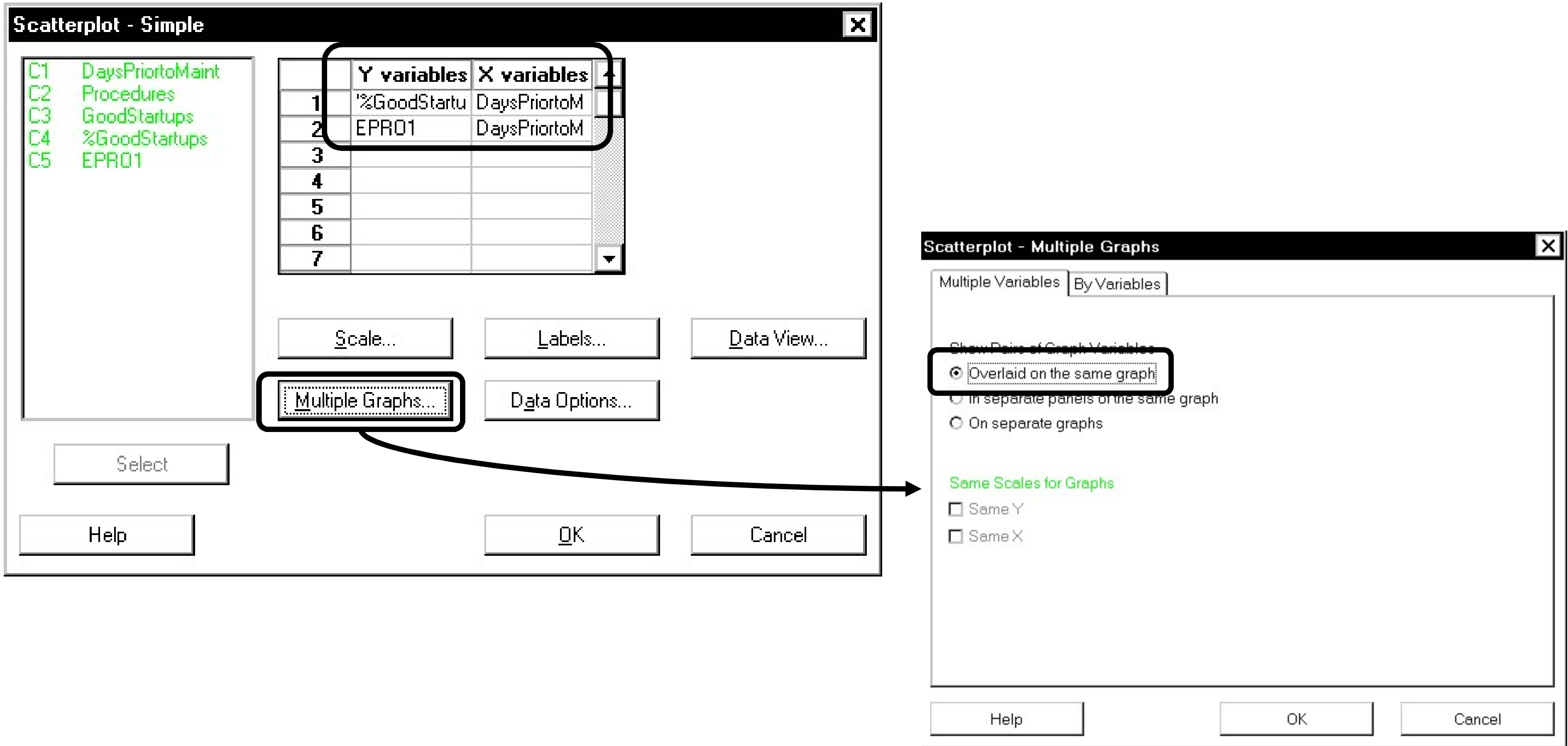
Testing for a lack of fit (curve through points). Since P > 0.05, conclude no lack of fit in any of three tests.

Beyond scope of course

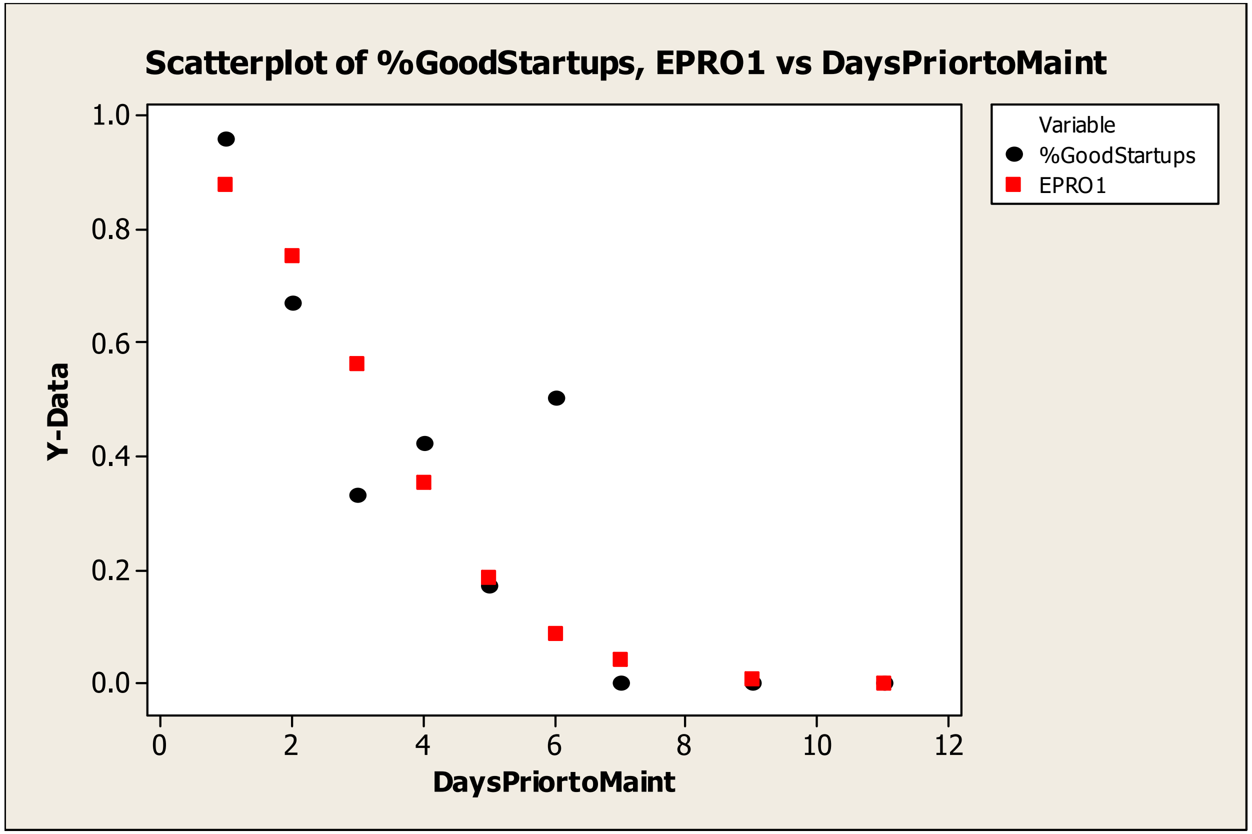
Minitab Follow-Along: Logistic Regression, cont.

3. Make a fitted curve plot for the data:

Graph > Scatter plot > Simple



Minitab Follow-Along: Logistic Regression, cont.



Conclusion: If you schedule maintenance every two days, you'll have a 75% chance of a good start-up the next day.

anexas consultancy SE

Regression Analysis

Logistic Regression with Discrete Y's

Multiple Logistic Regression

- You can try to improve the prediction of event probabilities by using several predictors (X's), not just one
  - The predictors (X's) can be either continuous or discrete
  - The multiple logistic regression equation is

$$\log_e\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots b_kX_k$$

  - If any of the X's are discrete (in groups), you must designate one group to be the reference group
  - The coefficient for a discrete X group compares its intercept with the reference group (see discussion in previous section)
- Multiple logistic regression is useful for identifying key drivers (X's), particularly those that affect Y (event probabilities) in combination with each other

Practice: Applications of Logistic Regression

- **Objective:** Practice identifying predictor variables for logistic regression situations
- **Instructions:**
  - Circle the outcome you would define as a "success" for each discrete Y
  - List some possible predictors
- **Time:** 5 minutes

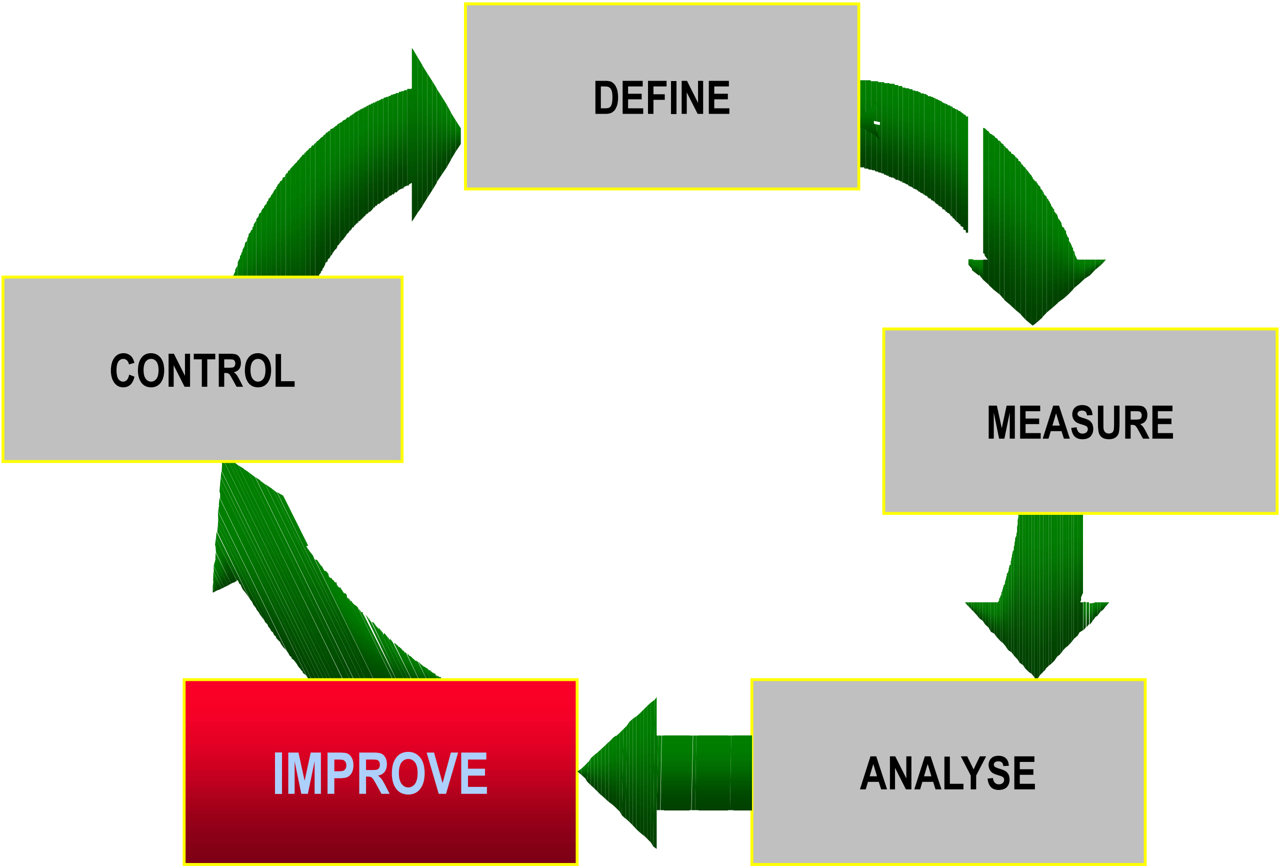
Discrete Y's	Predictors (X's)
Shipment is on time or late	
Customer buys product or not	
Deal was won or lost	
Customer retained this year or not	
Product works or not	
Machine breaks down or not	

Answers: Applications of Logistic Regression

Discrete Y's	Predictors (X's)
Shipment is **on time** or late	• Size of order • Number of other orders in queue
Customer **buys** product or not	• Product price • Presence of CTQ
Deal was **won** or lost	• Cost • Experience level of company representative
Customer **retained** this year or not	• Length of time with company • Amount price increased or decreased
Product **works** or not	• Number of defects • Quality of materials
Machine breaks down or **not**	• Type of maintenance procedure used • Experience level of maintenance operator

Module 5: Improve Phase

DMAIC : An Improvement Methodology



Improve

Objective :

Determine new improved process design

Steps :

Generate solutions

Select and test solutions

anexas consultancy SE

Idea Generation: Creativity approaches

- Process benchmarking
  - Compare the performance of an existing process against other companies' "best in class" practices (same market or not)
  - Determine how those companies are organised to deliver these performance level
- Best practices
  - Use company data
- Brainstorming
  - Brainstorming with post it notes, channelled brainstorming, anti-solution etc

Brainstorming

Types of Brainstorming

- Round Robin
- Anti Solution
- 6-3-5
- 6 Thinking Hats

Solution Selection Matrix
Select among Possible Solutions Using Objective Criteria

Criteria		Weight	Solution A		Solution B		Solution C	
			Score	Weighted Score	Score	Weighted Score	Score	Weighted Score
1				0		0		0
2				0		0		0
3				0		0		0
4				0		0		0
5				0		0		0
6				0		0		0
TOTAL				**0**		**0**		**0**

Where **weight** and **scores** on following scale : High = 9, Medium = 3 and Low = 1.

Conclusions:

Criteria are the requirements that you want your solution to meet. Some criteria are "must" criteria. Any solution that does not meet even one of the "must" criteria must be eliminated

anexas consultancy SE

Solution Selection Matrix

Criteria		Weight	Solution A		Solution B		Solution C	
			Score	Weighted Score	Score	Weighted Score	Score	Weighted Score
1	cheap solution	3	3	9	9	27	9	27
2	quick to implement	3	9	27	1	3	3	9
3	high impact on CTQs	9	9	81	9	81	9	81
4	compliant	9	1	9	9	81	9	81
5				0		0		0
6				0		0		0
TOTAL				**126**		**192**		**198**

Where **weight** and **scores** on following scale : High = 9, Medium = 3 and Low = 1.

Example(s):

Example :

Solution A = outsource all data processing

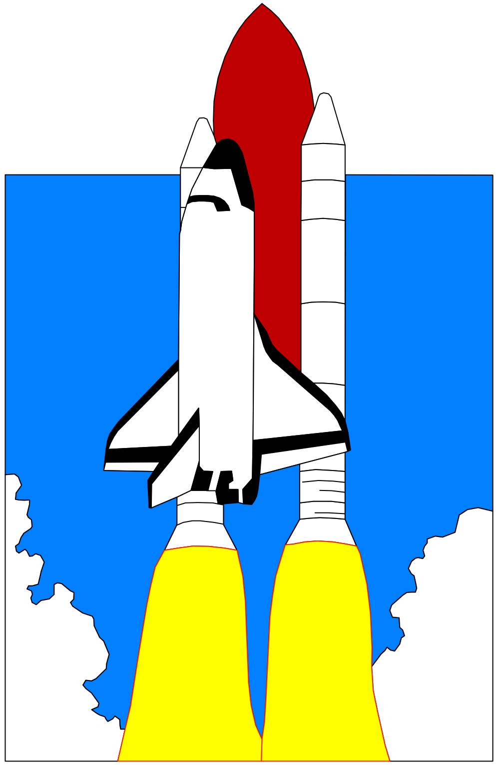
Solution B = development of our own software

Solution C = buy a software and adapt to our needs

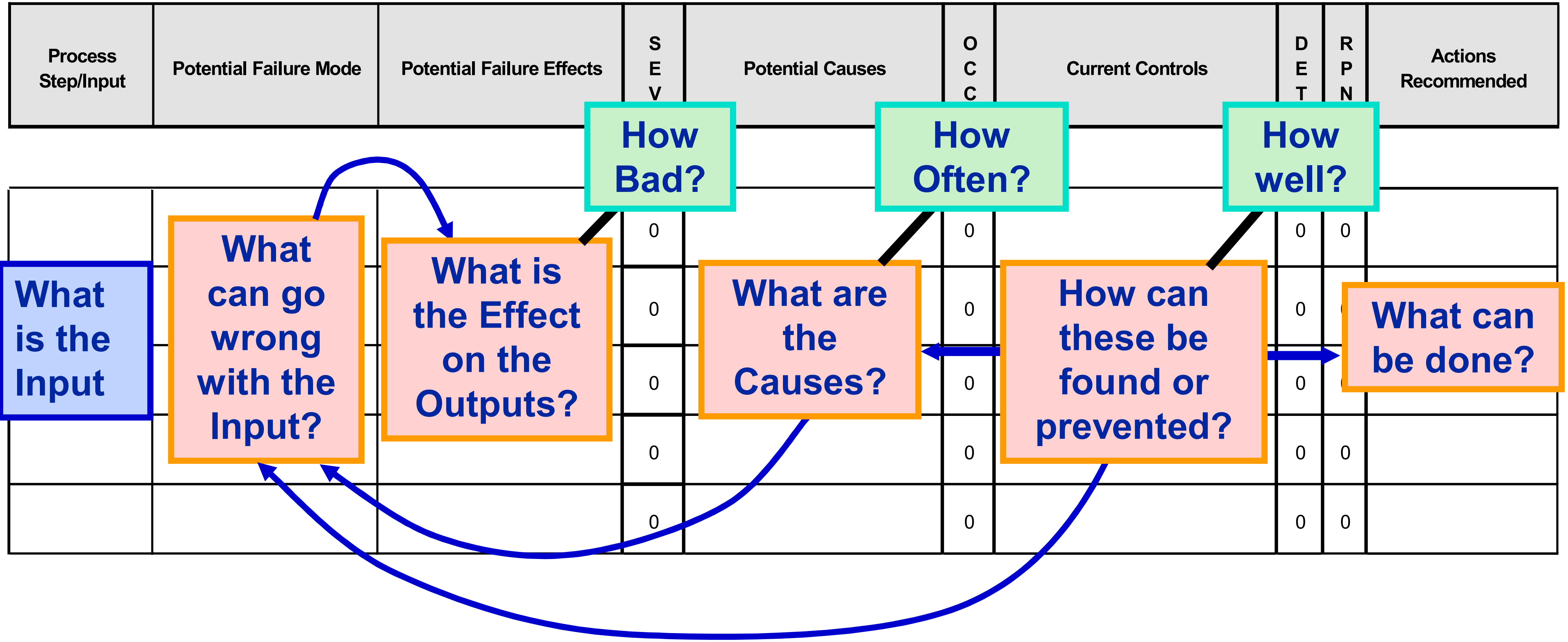
It seems here that solution C is the most satisfying. B also can be considered as an option.

Criteria are the requirements that you want your solution to meet. Some criteria are “must” criteria. Any solution that does not meet even one of the “must” criteria must be eliminated

Failure Modes and Effects Analysis



Overview



Process Step/Input
Potential Failure Mode
Potential Failure Effects
SEV
Potential Causes
OCC
Current Controls
DET
RPN
Actions Recommended
How Bad?
How Often?
How well?
What is the Input
What can go wrong with the Input?
What is the Effect on the Outputs?
What are the Causes?
How can these be found or prevented?
What can be done?
0

Introduction to the Design of Experiments (DOE)

Module Objectives

By the end of this module, the participant will:

- Understand the strategy behind Design of Experiments (DOE)
- Create a design matrix
- Code factors
- Understand the limitations of OFAT experiments
- Interpret interactions

Why Learn About Design of Experiments?

- Properly designed experiments will improve
- Efficiency in gathering information
  - Planning
  - Resources
- Predictive knowledge of the process
- Ability to optimize process
  - Response
  - Control input costs
- Capability of meeting customer CTQs

What is Design of Experiments ?

- A DOE is a planned set of tests on the response variable(s) (KPOVs) with one or more inputs (factors) (PIVs) each at two or more settings (levels) which will:

- Determine if any factor is significant
- Define prediction equations
- Allow efficient optimization
- Direct activity to rapid process improvements
- Create significant events for analysis

DOE Terminology

- Response (Y, KPOV): the process output linked to the customer CTQ
- Factor (X, PIV): uncontrolled or controlled variable whose influence is being studied
- Level: setting of a factor (+, -, 1, -1, hi, lo, alpha, numeric)
- Treatment Combination (run): setting of all factors to obtain a response
- Replicate: number of times a treatment combination is run (usually randomized)
- Repeat: non-randomized replicate
- Inference Space: operating range of factors under study

DOE Objectives

- Learning the most from as few runs as possible; efficiency is the objective of DOE
- Identifying which factors affect mean, variation, both or none
- Screening a large number of factors down to the vital few
- Modeling the process with a prediction equation:

  Y = f (A,B,C …)
- Optimizing the factor levels for desired response
- Validating the results through confirmation

Experimental Design Considerations

Experimental Objective	Experimental Design Types
Screen	OFAT (One Factor at a Time) Fractional Designs 2^k Full Factorial
Optimize *Model*	2^k Full Factorial (w center points/replication) Full Factorial (replication) RSM (Response Surface Method)

Many KPIV's ***Few***

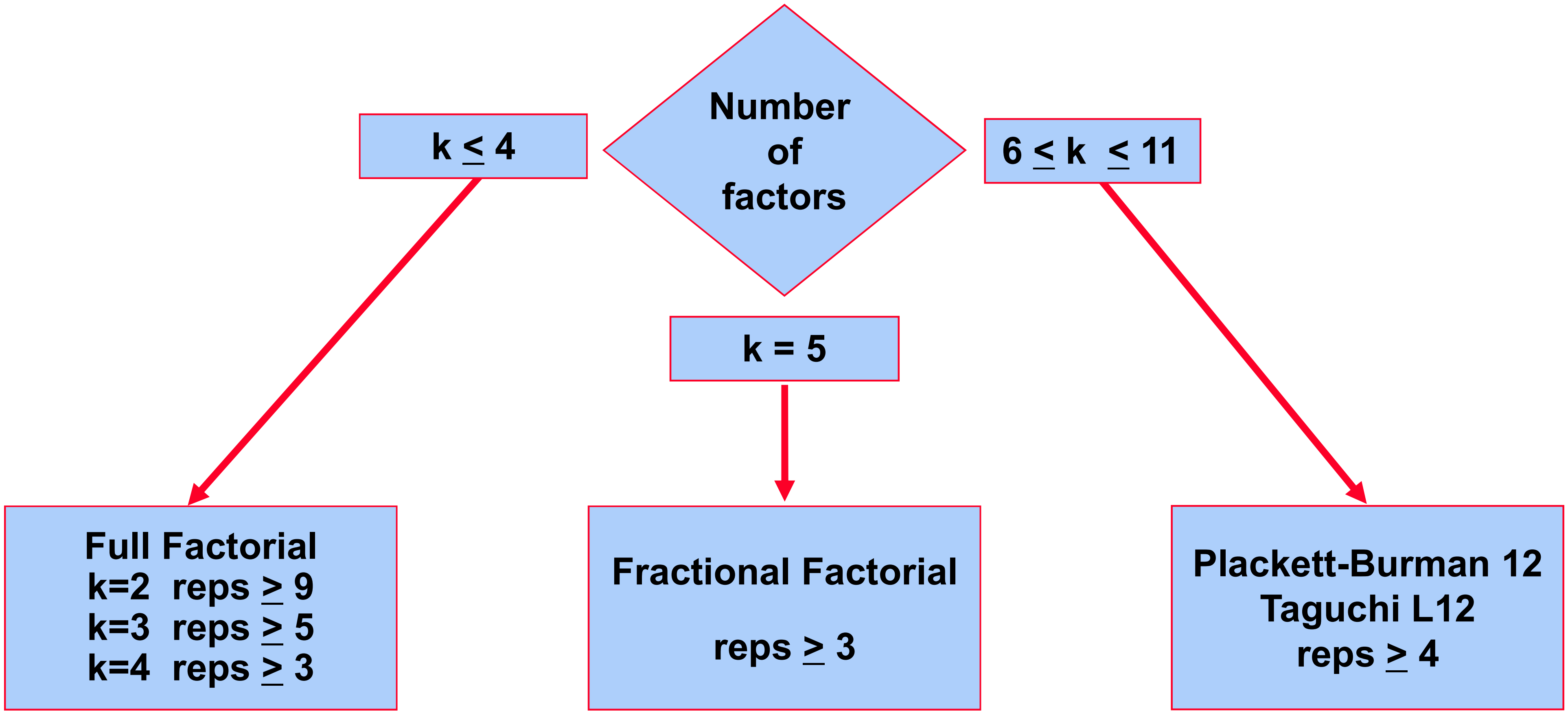
Less KNOWLEDGE ***More***

Less COST ***More***

Higher complexity designs offer greater knowledge at a higher price

anexas consultancy SE

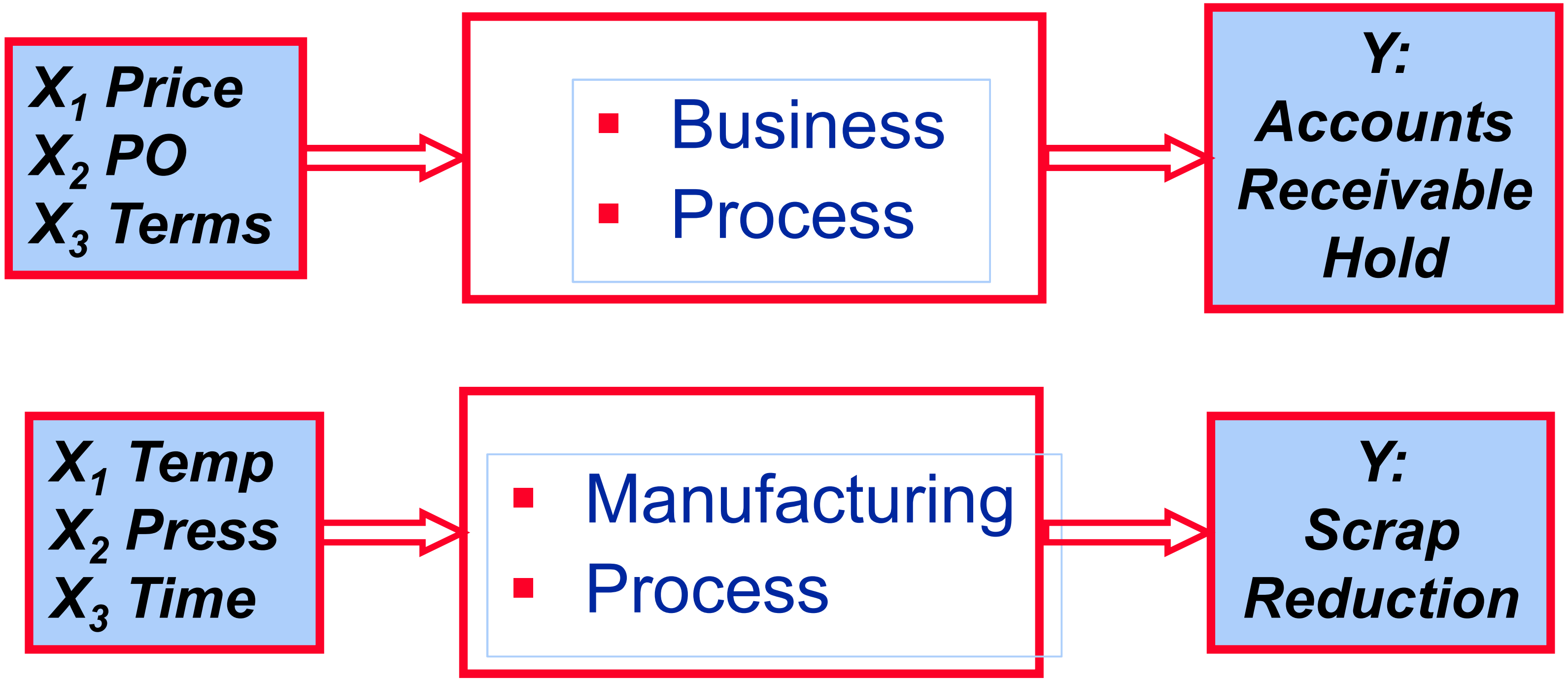
Two Level Sample Size



Provides greater than 99% confidence in mean and 95% confidence in variance prediction models

Factors

Factors and the Process



DOE can be applied to both business and industrial processes

anexas consultancy SE

One Factor at a Time (OFAT) Experiments

One Factor at a Time

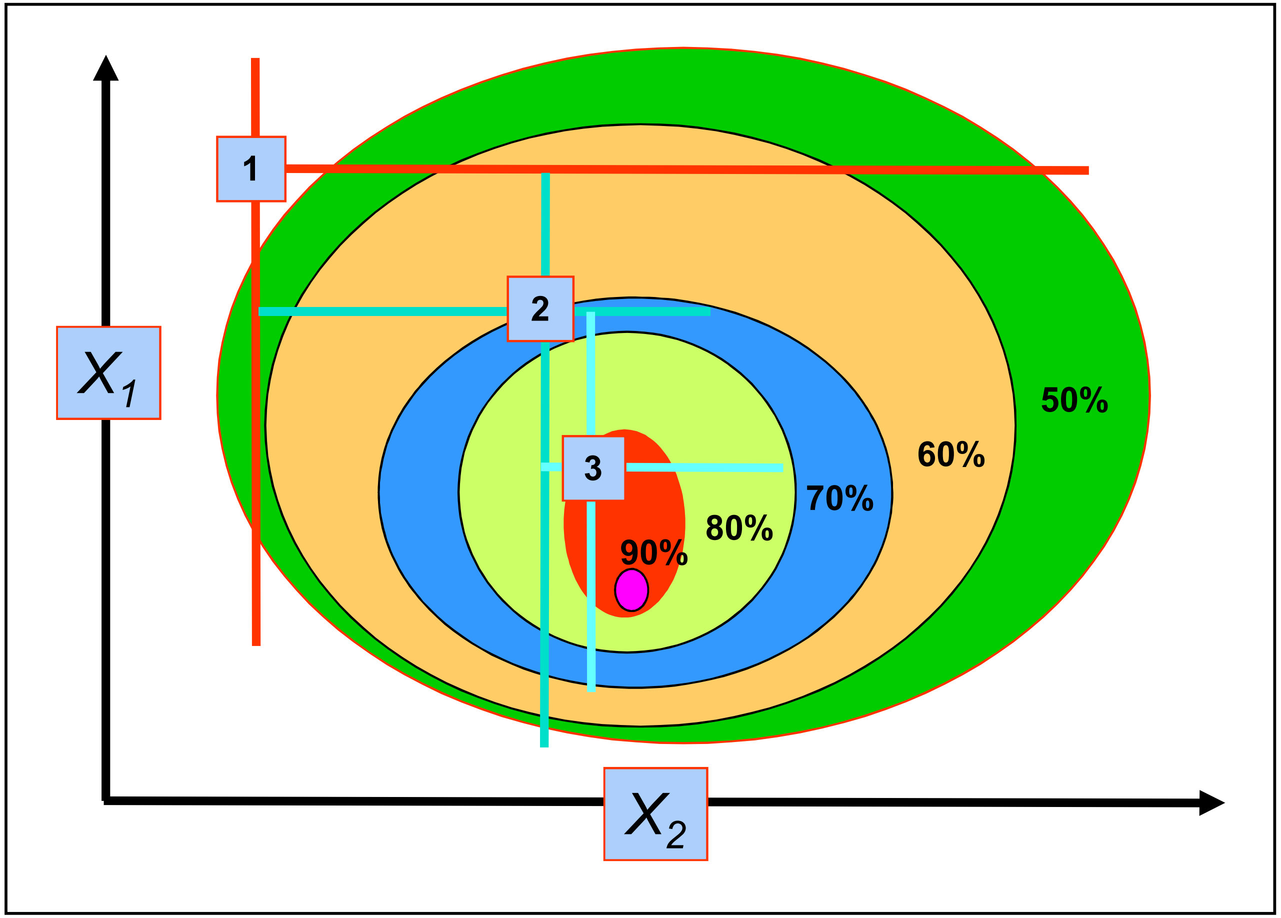
- A process has two variable inputs: flow and temperature. Historically the process has run with flow = 18 and temp = 160 with a yield of 81.7%.

Holding flow at 18, a range of temperatures is run:

Old setting →

Flow	Temp	Yield
18	140	71.4
18	145	74.8
18	150	78.3
18	155	79.4
18	160	81.7
18	165	82.8
18	170	84.0
18	175	84.5
18	180	84.2
18	185	83.3
18	190	80.5
18	195	78.3
18	200	76.0

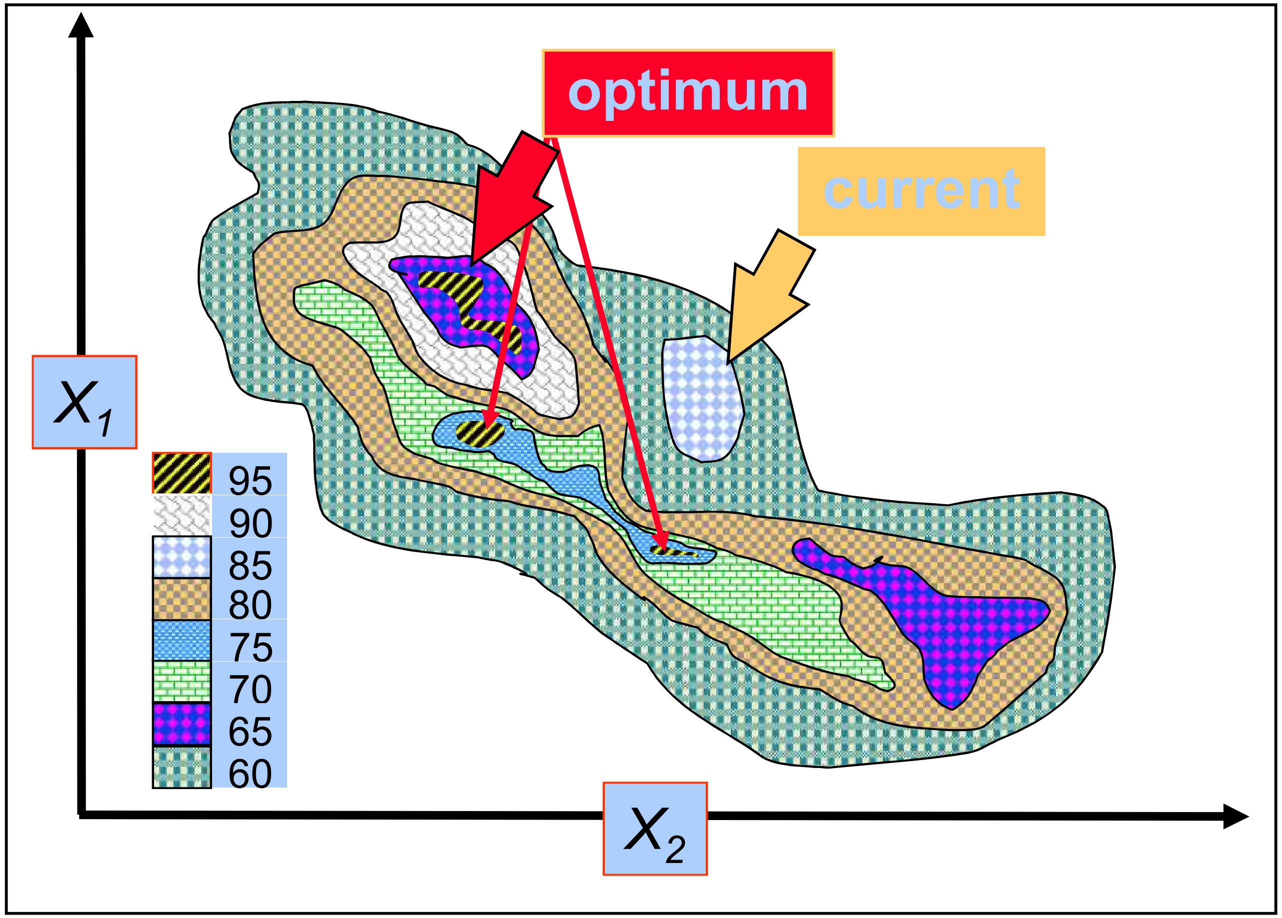
OFAT Prediction Contours



One factor at a time presumes a nested family of responses

Factor interactions are missed with OFAT

Process Response Contours



More often than not, the response contours of a process contain many peaks and valleys created by the interaction of factors

OFAT optimization can become inefficient and sometimes misleading

Prepared to Explore DOE

DoE 3 Factor 2 Level ***

↓	C1	C2	C3	C4	C5	C6	C7	C8
	StdOrder	RunOrder	CenterPt	Blocks	A	B	C	
1	6	1	1	1	1	-1	1	
2	3	2	1	1	-1	1	-1	
3	2	3	1	1	1	-1	-1	
4	1	4	1	1	-1	-1	-1	
5	8	5	1	1	1	1	1	
6	7	6	1	1	-1	1	1	
7	4	7	1	1	1	1	-1	
8	5	8	1	1	-1	-1	1	
9								

Minitab will be the tool for creating and analyzing DOE experiments in the Breakthrough Strategy ™

Full Factorial Design of Experiments

Linear Combinations of Factors for Two Levels

Combinations of Factors and Levels

- A process whose output Y is suspected of being influenced by three inputs A, B and C. The SOP ranges on the inputs are
  - A 15 through 25, by 1
  - B 200 through 300, by 2
  - C 1 or 2
- A DOE is planned to test all combinations

Is testing all combinations possible, reasonable and practical?

Combinations of Factors and Levels cont'd

- Setting up a matrix for the factors at all possible process setting levels will produce a really large number of tests.
- The possible levels for each factor are

  A = 11

  B = 51

  C = 2

- How many combinations are there?

A	B	C
15	200	1
16	200	1
17	200	1
18	200	1
19	200	1
20	200	1
21	200	1
22	200	1
23	200	1
24	200	1
25	200	1
15	202	1
16	202	1
17	202	1
.	.	.
.	.	.
.	.	.
.	.	.
22	300	2
23	300	2
24	300	2
25	300	2

We must make assumptions about the response in order to manage the experiment

anexas consultancy SE

Linear Response for Factors at Two Levels

- The team decides, from process knowledge, that the response is close to being linear throughout the range of factor level settings (inference space).
- A reasonable assumption for most processes
- The levels of the factors for the test would then be

  A 15 and 25

  B 200 and 300

  C 1 and 2

The design becomes much more manageable!

anexas consultancy SE

The Three Factor Design at Two Levels

- The revised experiment consists of all possible combinations of A, B and C each at the chosen low and high settings:

A	B	C
15	200	1
15	200	2
15	300	1
15	300	2
25	200	1
25	200	2
25	300	1
25	300	2

This is a 2^3 full factorial design (pronounced two to the three). It consists of all combinations of the three factors each at two levels

anexas consultancy SE

Naming Conventions

- The naming convention for full factorial designs has the level raised to the power of the factor:

$$\boldsymbol{\mathit{level}}^{\text{(factor)}}$$

- and is called “a (level) to the (factor) design”

- What would a two level, four factor design be called?
- How many combinations (runs) are in a 23 design?

Class Exercise

- Write the total number of combinations for the following designs

  2^3

  2^4

A	B	C	D

- Assume factors are named A, B, C, D, etc. and the levels are low "-" and high "+".

Did we all generate the same designs?

Replicates and Repeats

What are Replicates and Repeats?

- Replicate
- Total run of all treatment combinations
  - Usually in random order
- Requires factor level change between runs
- All experiments will have one replicate
  - Two replicates are two complete experiment runs
- Statistically best experimental scenario
- Repeat (also repetition)
- Additional run without factor level change

Minitab Design Replication

- Minitab easily handles replicating the design
- Replicate or repeat is treated same in design
- Actual factor level change between runs is at the discretion of the experimenter
  - Minitab provides treatment combination
  - Randomization or information needed is part of strategy of experiment

Coding the Design

Coding the design by transforming the low factor level to a “-1” and the high factor level to a “+1” offers analysis advantages

Coding Review Exercise

Fill in the coded design based upon the uncoded design

	Uncoded Factors			Coded Factors		
Runs	A	B	C	A	B	C
1	15	200	1			
2	25	200	1			
3	15	300	1			
4	25	300	1			
5	15	200	2			
6	25	200	2			
7	15	300	2			
8	25	300	2			

Any uncoded design can be transformed into a coded design

Requirement of Factor Independence

- Factors are mathematically independent when only the response is a function of the factors
- A factor is not a function of another factor
- The coded design is orthogonal
  - Factors will be independent

DOE analysis requires that the factors be independent

anexas consultancy SE

Main Effects and Interactions

Calculating Main Effects

A DOE is run:

	Coded Factors		Response
	A	B	Y
	-1	-1	48
	1	-1	96
	-1	1	72
	1	1	36
Y_{ave} at $FACTOR_{high}$	66	54	
Y_{ave} at $FACTOR_{low}$	60	72	
Effect	6	-18	

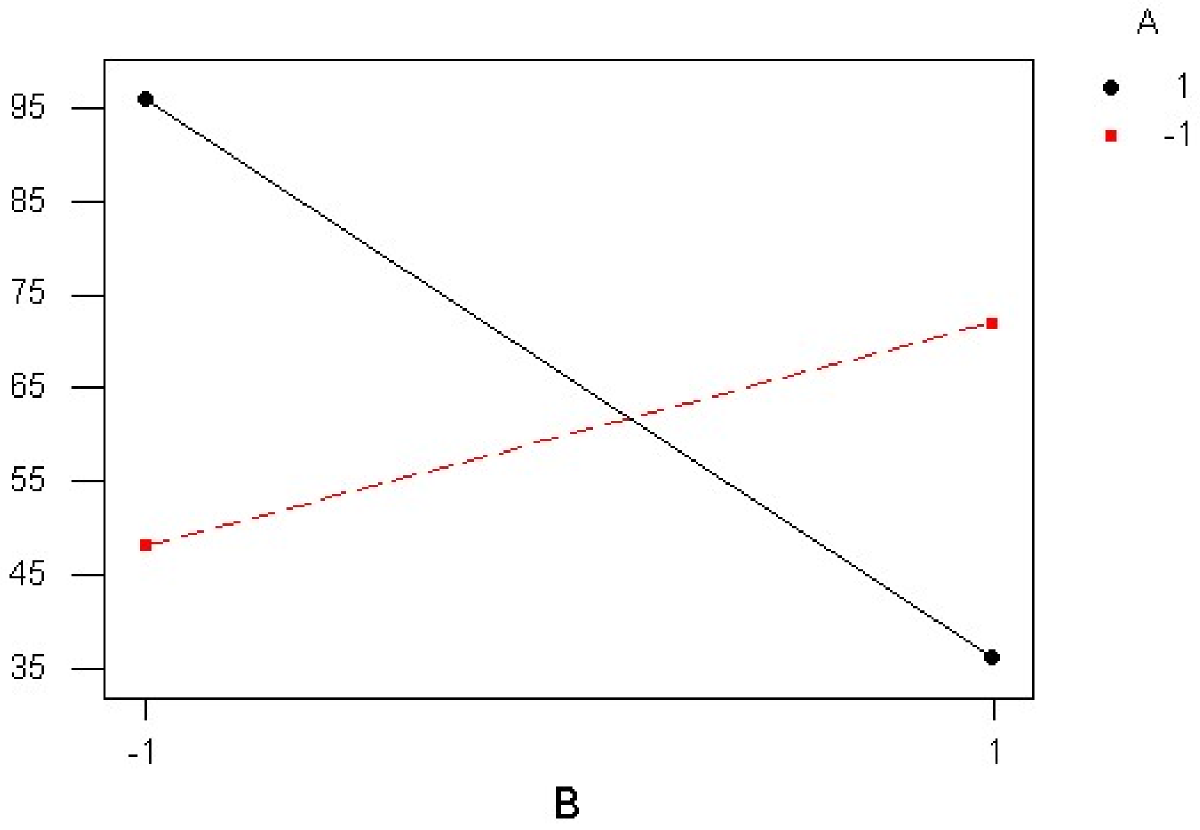
$$\frac{96+36}{2}=66$$

What does a non-zero effect mean?

$$effect=\overline{Y}(@\ factorhigh)-\overline{Y}(@\ factorlow)$$

The factor (or main) effects are easily calculated

Discovering Interactions



A response change due to both A and B changing is called an interaction

Main Effects, Interactions and Cube Plots in Minitab

Create the Experiment

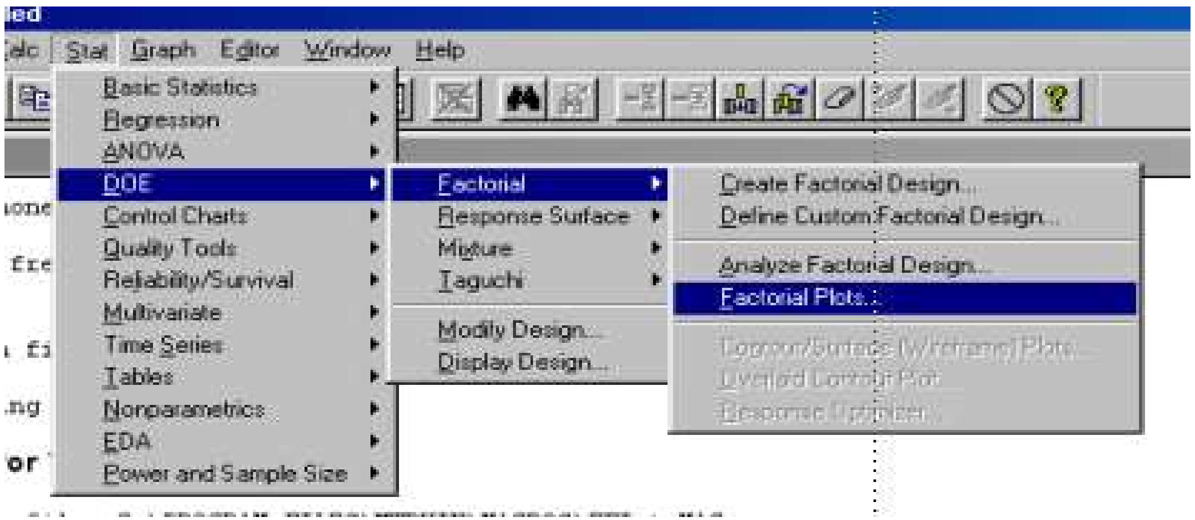
Tool Bar Menu > Stat > DOE > Factorial > Factorial Plots

Create a 2^2 coded design for factors A and B.

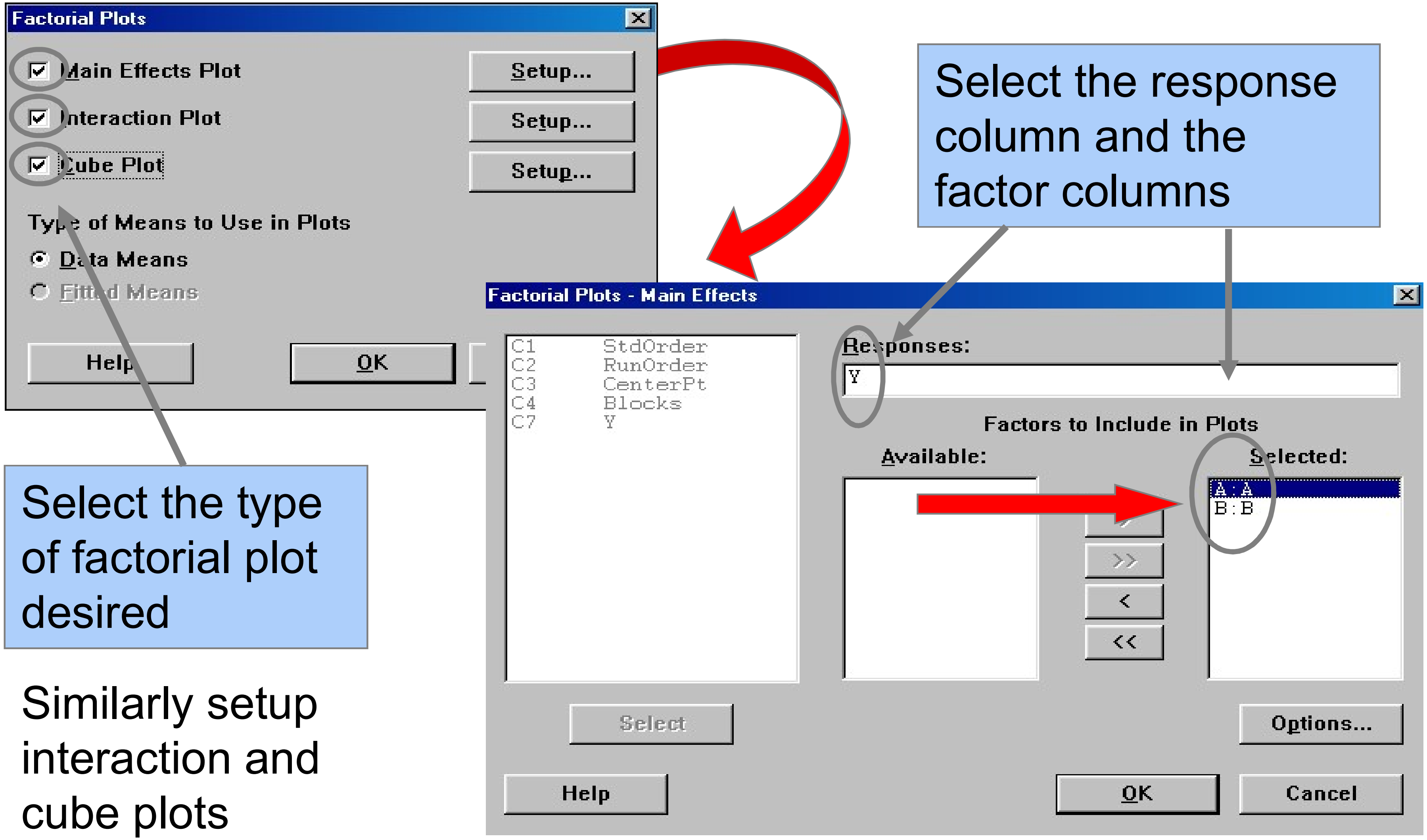
Worksheet 4 ***

	C1	C2	C3	C4	C5	C6	C7
	StdOrder	RunOrder	CenterPt	Blocks	A	B	Y
1	1	1	1	1	-1	-1	48
2	2	2	1	1	1	-1	96
3	3	3	1	1	-1	1	72
4	4	4	1	1	1	1	36
5							

Input the Y response



Factorial Plots in Minitab
Step 1



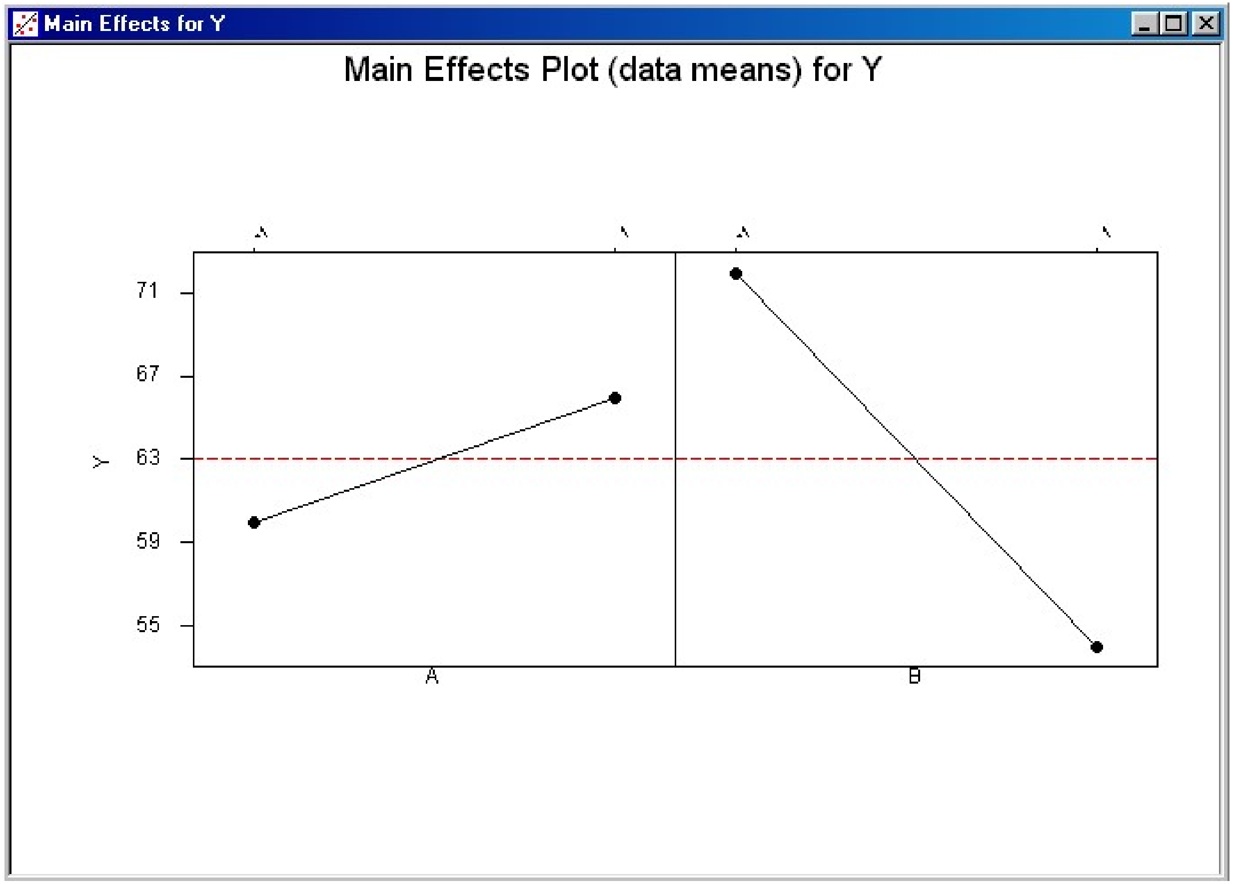
Select the response column and the factor columns

Select the type of factorial plot desired

Similarly setup interaction and cube plots

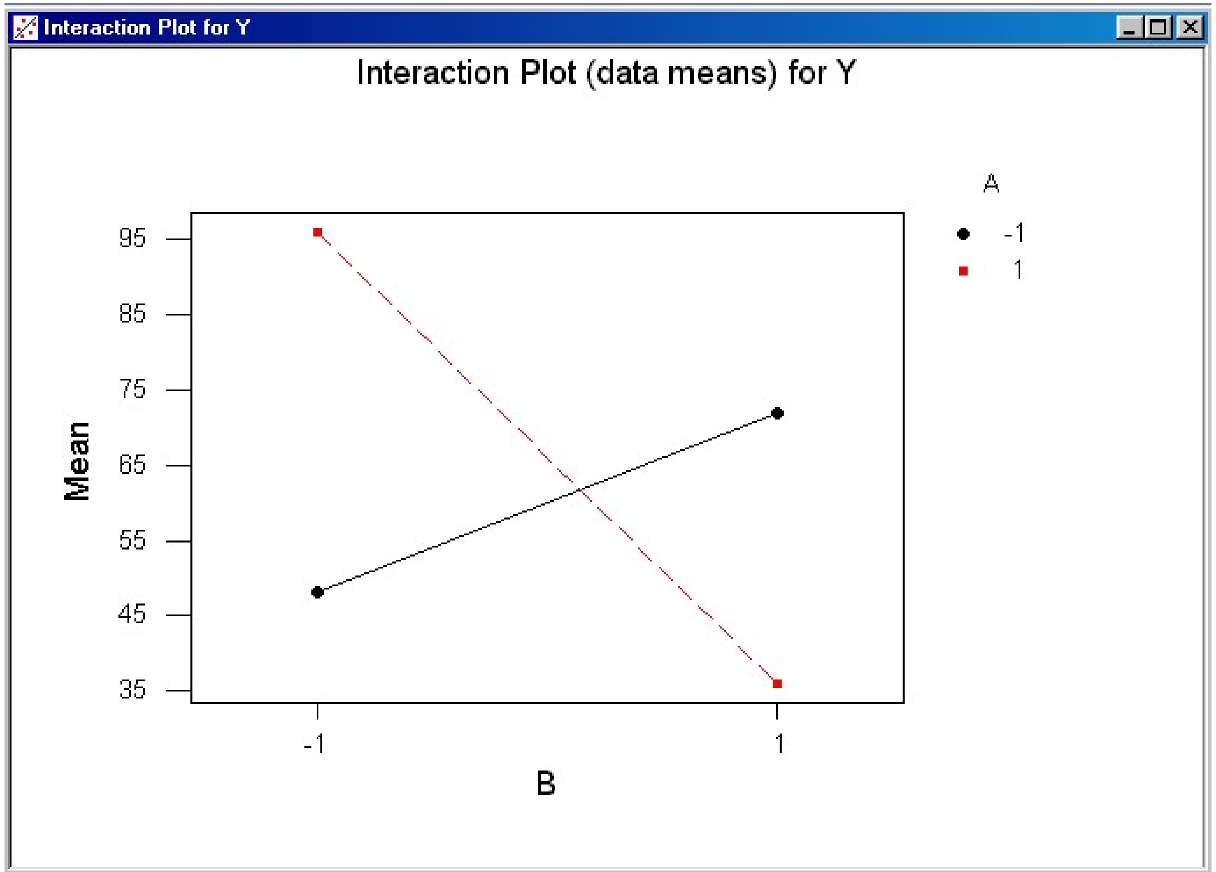
anexas consultancy SE

Main Effects Plot



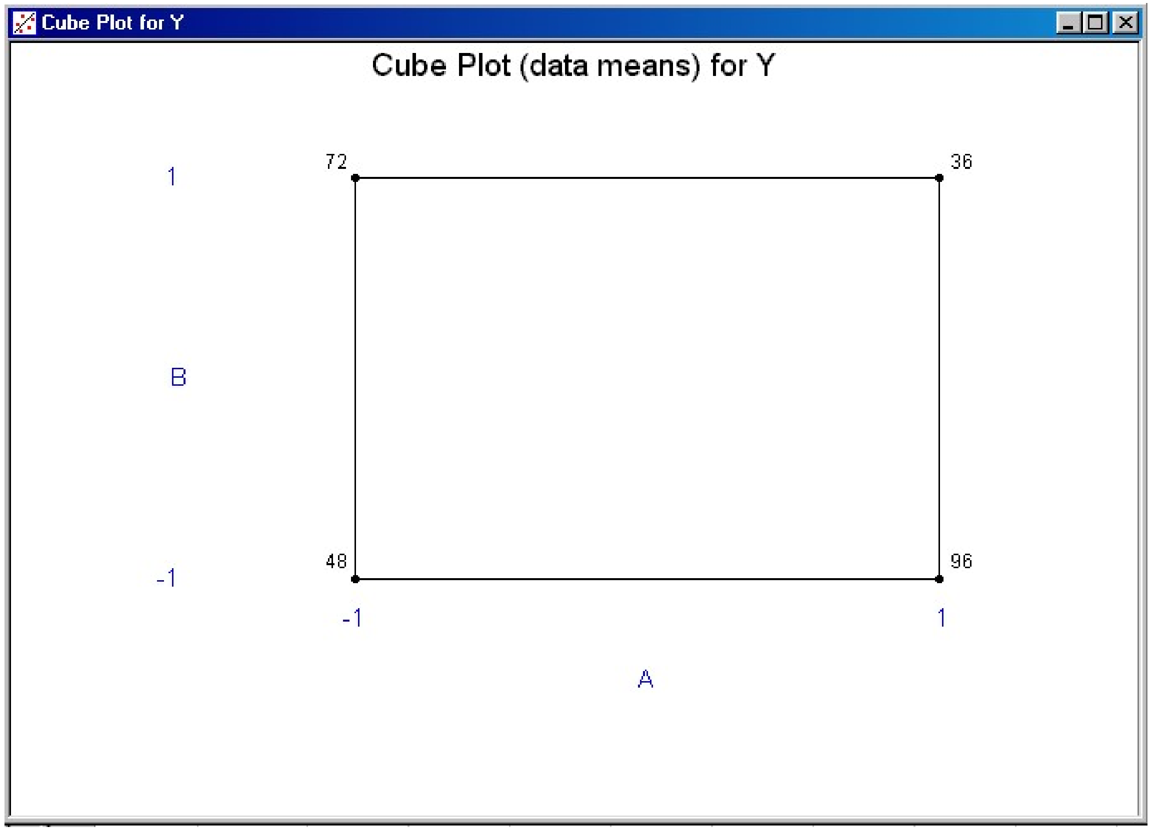
The response is plotted by factor from low level to high level

Interaction Plot



Parallel lines indicate no interaction; the less parallel, the higher the degree of interaction

Cube Plot



The response is plotted on the orthogonal factor axis

The General Linear Model

Why GLM?

- The General Linear Model
- Allows more flexible design
- Allows multiple levels
- Does not require factors to have same number of levels
- Is well suited for business process problems

Setting up a GLM Design

- Account receivables lockup, where payments are withheld, is thought to be caused by four factors
- SPAs (Special Pricing Agreements) -4 categories
- Market sector – 3 demographics
- Sales region – 6 regional centers
- Performance to contract – 3 levels
- Design a DOE to study the problem

anexas consultancy SE

Objectives Review

By the end of this module, the participant should:

- Generate a full factorial design
- Look for factor interactions
- Develop coded orthogonal designs
- Write process prediction equations (models)
- Set factors for process optimization
- Create and analyze designs in Minitab™
- Evaluate residuals
- Develop process models from Minitab™ analysis
- Determine sample size

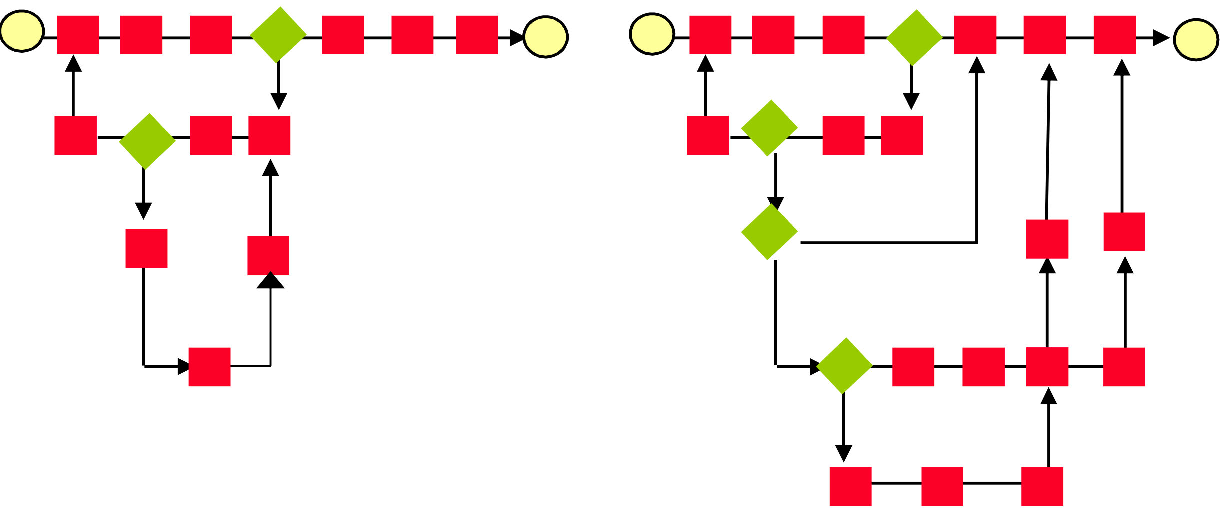
Continuous Improvement

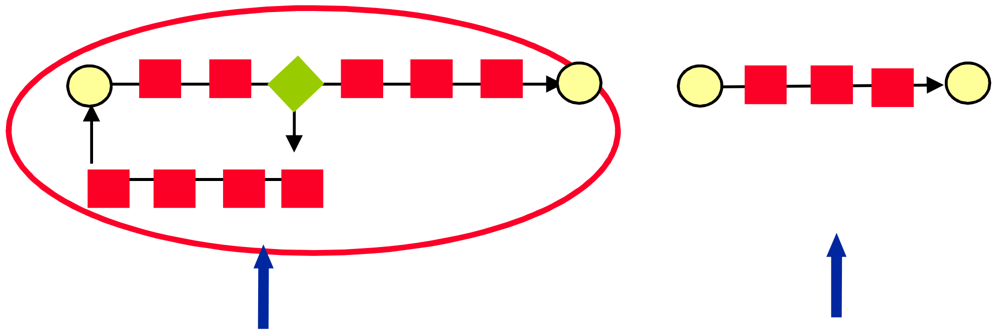
What you *think* It is….

What it *really* is…

What it *should* be…

What it *could* be





IMPROVE :
What we want to put in place in generation 1 that will give us the money to finance generation 2.

IMPROVE :
Generation 2 target.

Benefits of doing a pilot

- Improve the solution that meets customer requirements
- Refine implementation plan
- Lower risk of failure by identifying and fixing possible problems ahead of time
- Confirming expected results and relations between predictive parameters and results (Xs on Y)
- Increase opportunities to receive feedback and buy-in
- Implement the solution earlier and faster for a particular customer segment

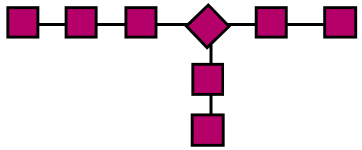
IMPROVE

Purpose : **To determine new improved process design** through idea generation, selection, process design, solution testing , and improvements implementation.

Solutions Refinement

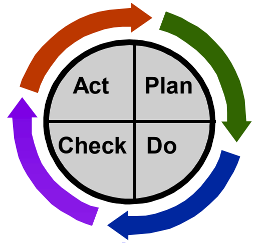
Evaluate potential problems in new process design and improve robustness of this design

New Process



Develop a “should be” process map showing the impact of the solution

Pilot



Pilot the solution on a small scale to increase buy-in and improve overall implementation

Perform cost / benefit analysis of proposed solution

Solution generation and selection

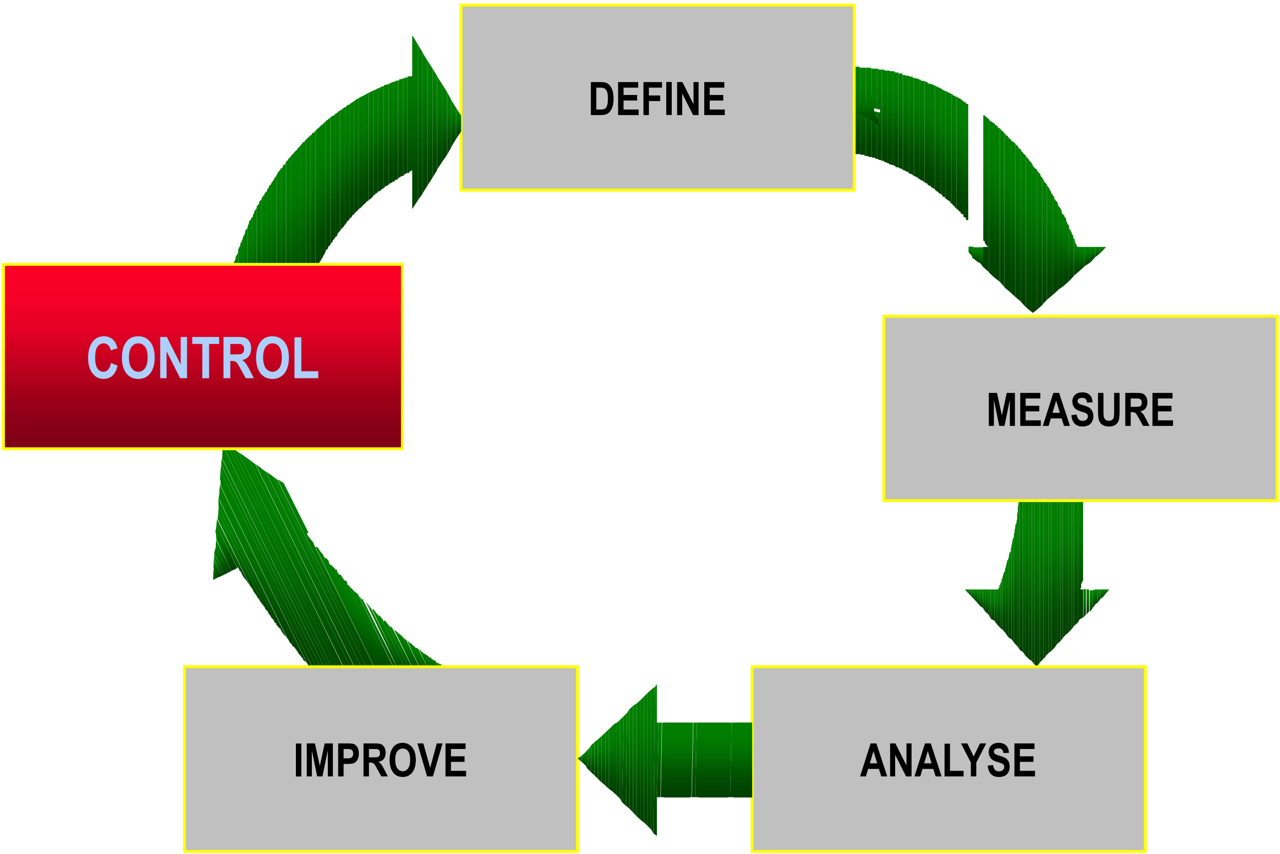
Brainstorming, anti solution, brainwriting, ...

Generate solutions to address the root causes and develop criteria to screen and select solutions (including cost / benefit)

anexas consultancy SE

Module 6: Control Phase

DMAIC : An Improvement Methodology



Control

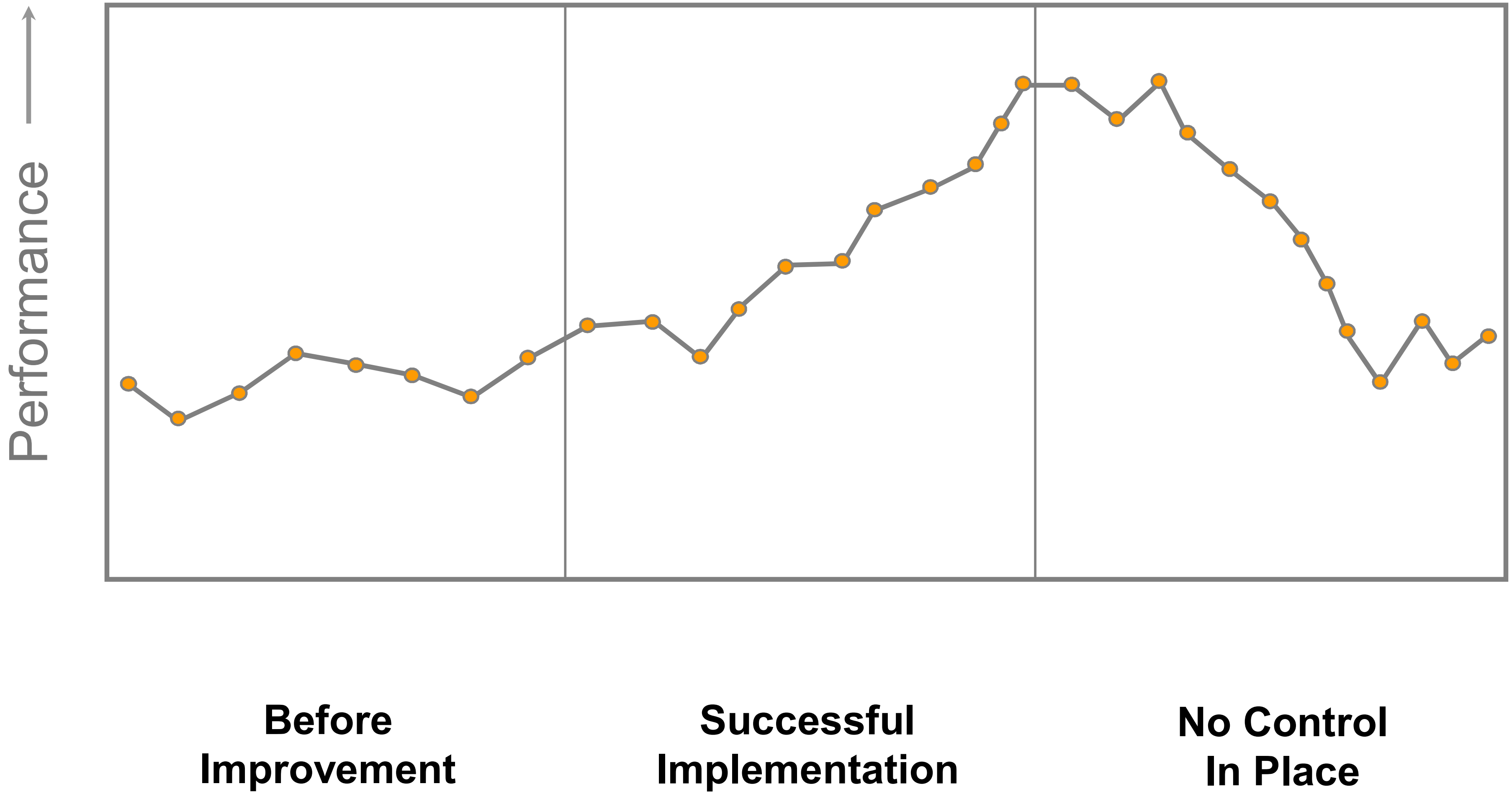
Objective :

- Ensure improvement over time

Steps :

- Create control tools (documentation and dashboard)
- Organise process reviews by Process Owner

Control = ensure gains over time



CONTROL = implement process management

- Process Management Chart
  - process owner’s name
  - process documentation (process mapping, persons involved)
  - customer performance criteria
  - key measures to track, follow and analyse (output, process, input, financials)

- Dashboards
  - graphical display of measurements collected

- Process performance reviews
  - frequency according to process cycle time

- Response plan
  - quick fixing of special causes
  - opportunities for ongoing improvement, i.e. new DMAIC projects

Five S

What Are The Five S's?

- Sorting
  - Selecting or separating
- Simplifying
  - Straighten and store
- Sweeping
  - Scrub and shine
- Standardizing
- Self discipline
  - Systematize

Mistake Proofing
(Poka-Yoke)

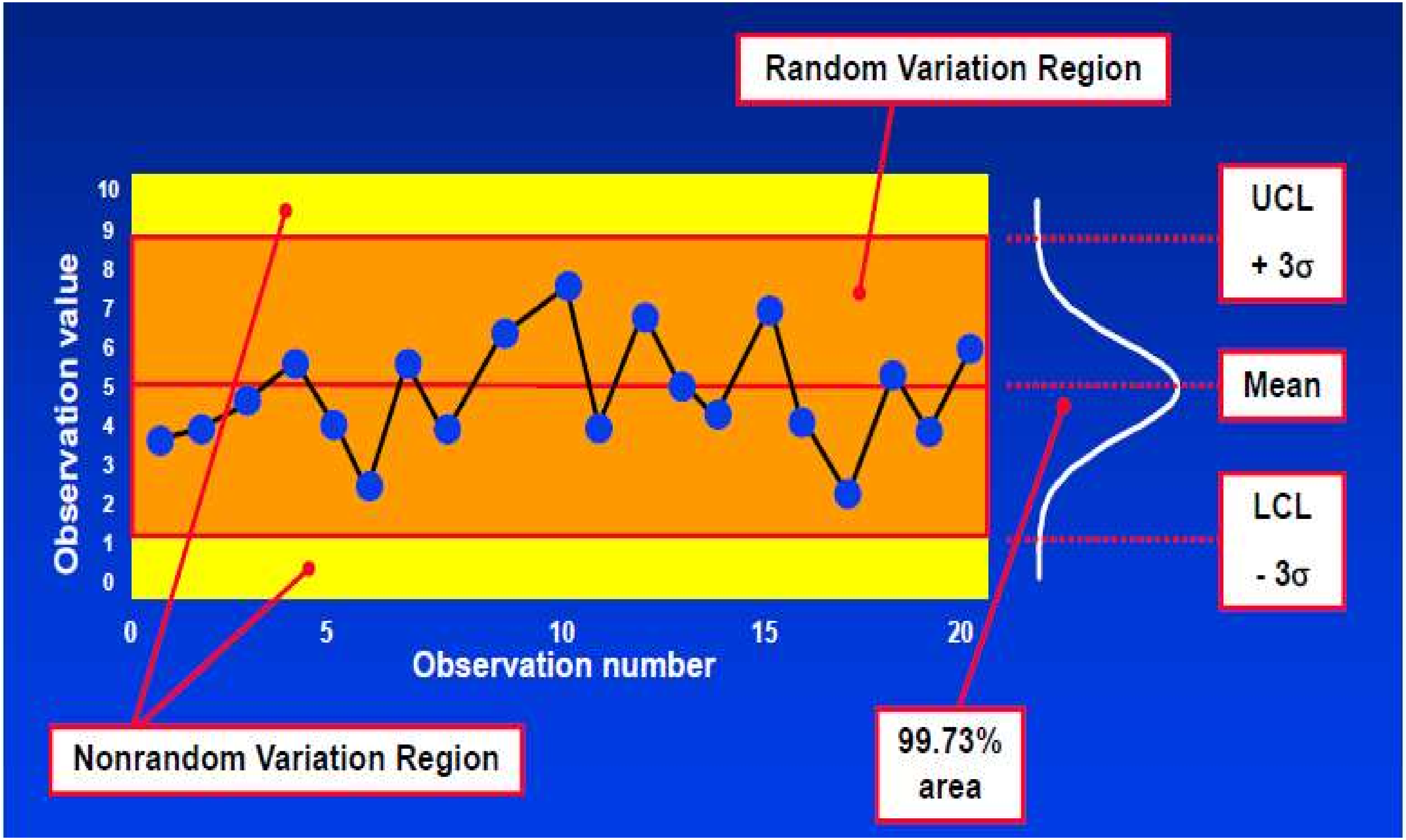
What Is Mistake Proofing (Poka-Yoke)?

- Japanese phrase:
- Yokeru (to avoid), Poka (errors)
- A strategy for preventing errors in processes
- Makes it impossible for defects to pass unnoticed
- Corrects problems as soon as they are detected
- Technique detects defects
- Prevents defects from moving into next area
- Developed by Dr. Shigeo Shingo to achieve zero defects

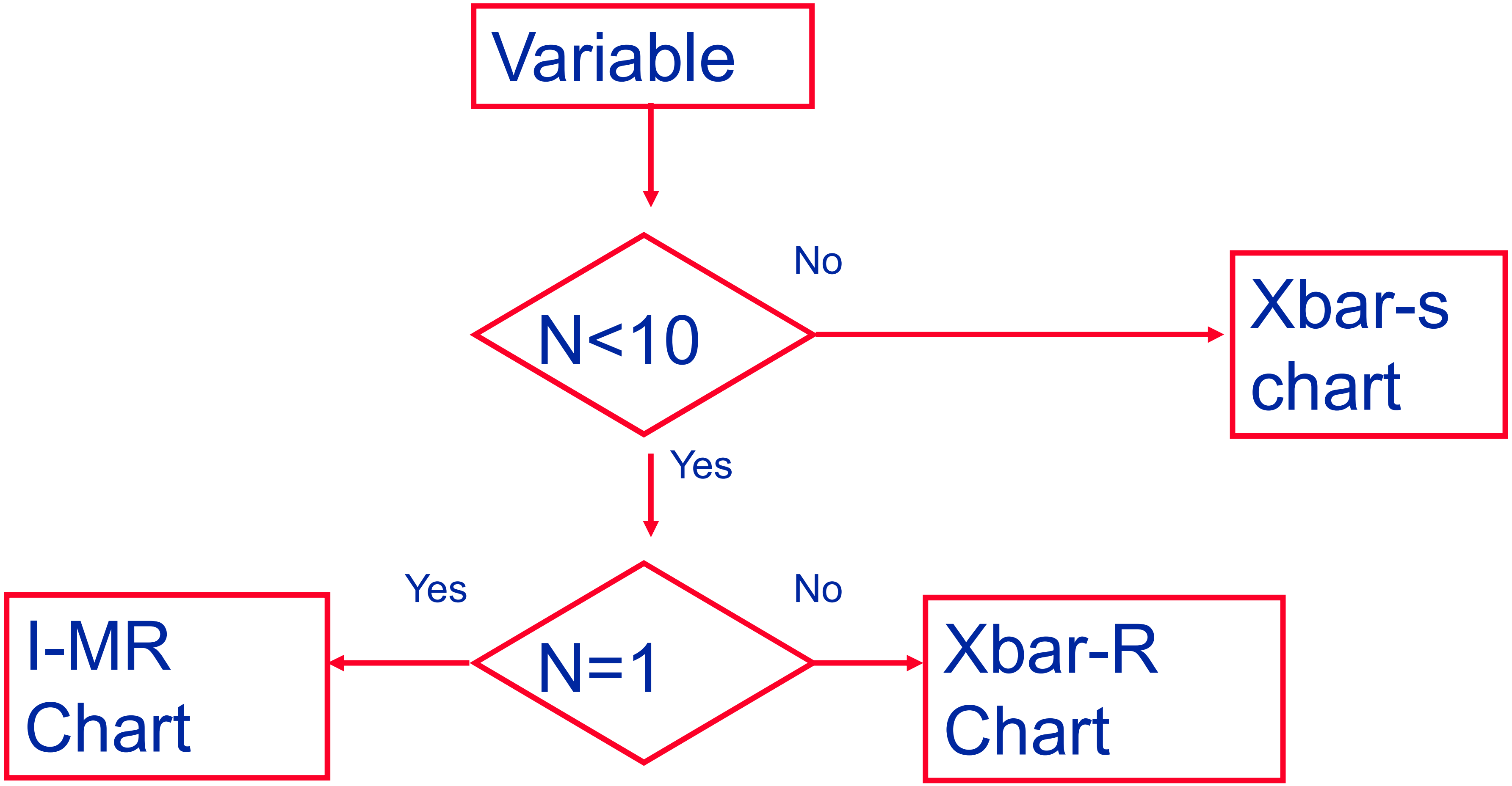
Statistical Process Control for Variables Data (SPC)

Introduction to SPC

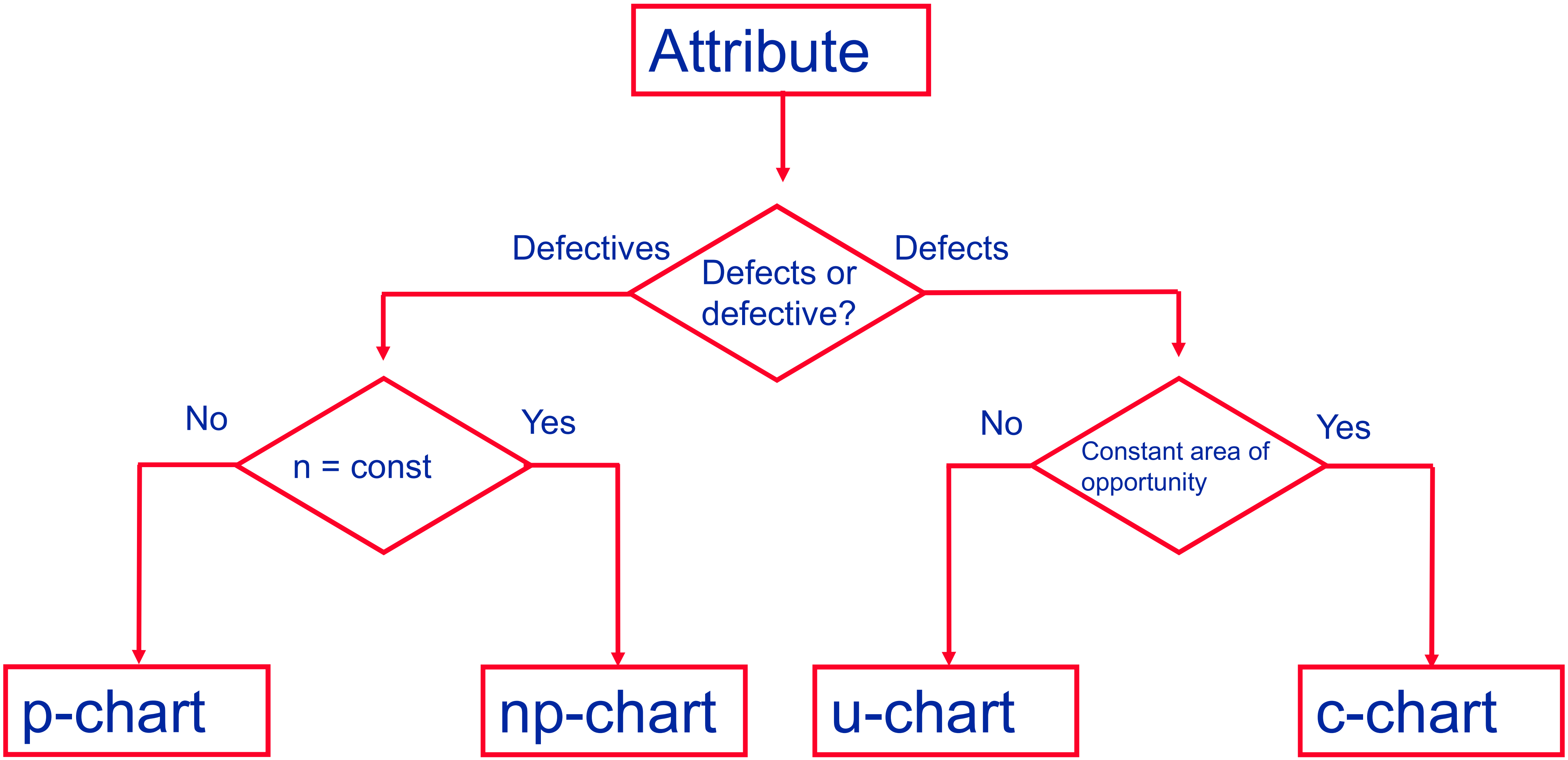
Statistics of a Control Chart



Control Chart Roadmap



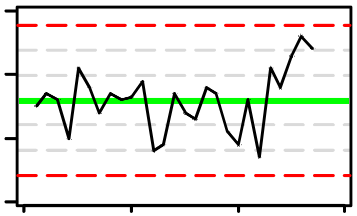
Control Chart Roadmap



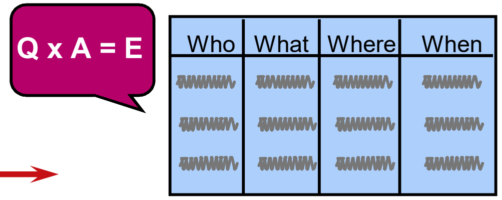
CONTROL

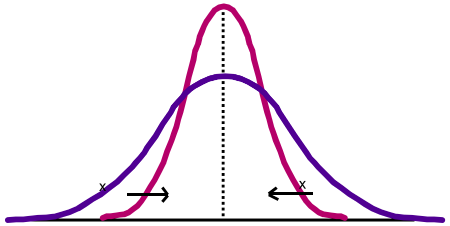
Purpose: **To ensure improvement effectiveness over time** by institutionalisation of the improvement and implementation of ongoing monitoring and reviews.

Monitoring Plan



Implementation Plan





Process Capability

Develop a monitoring plan to insures gains are held over the long term

Develop a full implementation plan including project and change management elements

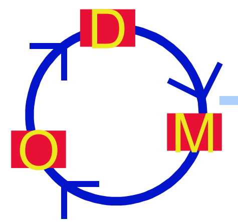
Monitor the process according to plan. Chart data as evidence that process is in control and meeting customer specifications

Documentation / Standardization

Document the process with process maps & procedures to assure the solution becomes part of daily work

Address appropriate changes to broader systems and structures to institutionalise the improvement

Continuous Improvement



- Process ownership to Process Owner (Process Management chart to facilitate transfer)
- Process Owner to held regular process reviews based on dashboards.
- Process Owner to take action when process does not deliver what is expected
- Process has entered Process Management = Define, Measure , Operate.

The DMAIC Storyboard : Six Sigma for Process Improvement

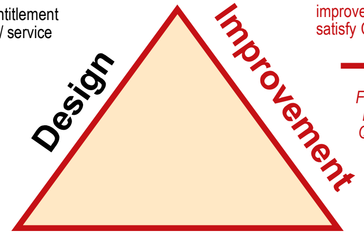
The Three Dimensions Of Process Focus

- Use when:
  - Process does not exist
  - Process is broken
  - Process has reached entitlement
  - Designing new product / service

Optimise Design Across Customer Critical To Quality requirements (CTQs) And Sub process

- Use when incremental improvement is needed to satisfy Customer CTQs

Focus On Root Causes That Drive Customer Critical To Quality requirements (CTQ) Performance



Management

Process Management System

Ongoing Management Of Process

- Use to manage a business process using CTQ measurements

Define

Purpose: **To set set direction for improvement project** by developing a team charter. To defining the customers and their requirements (Critical To Quality = CTQs), mapping the high level business process to be improved.

High Level Map - SIPOC

- Complete high level "as-is" process map, identifying suppliers, inputs, 5-7 high level activities, outputs & customers

Use Survey or Focus Groups?

Voice of Customer (VOC)

- Gather and display data verifying customer requirements (CTQs)

Project Charter

Problem Statement: / Goal: / Business opportunity: / Scope: / Cost Benefit Projection: / Milestones:

- Develop charter to include:
  - Problem statement
  - Goal for improvement
  - Business opportunity
  - Scope of project
  - Milestones for completion
  - Roles

Measure

Purpose: **To measure and understand baseline performance** for the current process by collecting reliable data (quantitative & qualitative).

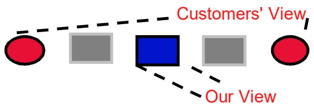
Data Collection

- Develop a data collection plan
  - Operational definition
  - Sampling
  - Measurement System Analysis (MSA including Gage R&R)

Graphical Display

- Display data in graphic form to determine the type of distribution, the metrics to understand variation and set goals for the improvement strategy.
  - Normal Distribution described by Mean and Standard deviation
  - Skewed Distribution described by Q1 (or Q3) and Inter Quartile Range
  - Long tailed distribution described by Median and Span 5-95

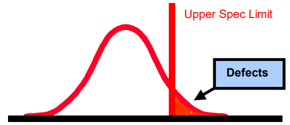
Customer oriented mindset



- Select the measure your customer uses to judge your performance (Key Output Measure Y)
- Plan to collect CONTINUOUS data to measure your Key Output Measure Y

Calculate Process Sigma

Defects "Outside" Spec Limit



- Compute baseline sigma

Analyse

Purpose: **To identify root causes of the problem and opportunities for improvement** by analysing the data and the process.

"As Is" Process Map & Analysis

- Develop an "as is" process map or sub-process map, and analyze it using:
  - Moments of truth (customer contact)
  - Nature of work (value analysis),
  - Flow of work (cycle time)

Root Cause Identification

- Perform cause & effect analysis to identify potential root causes

Root Cause Verification

- Verify root causes through statistical tools or design of experiments

Quantify The Opportunity

Data Stratification / Segmentation

- Stratify / Segment the data to find the vital few areas to focus improvement efforts on.

Review and Finalise Problem Statement

Improve

Purpose: **To determine new improved process design** through idea generation, selection, process design, solution testing , and improvements implementation.

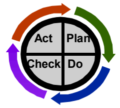
Solutions Refinement

- Evaluate potential problems in new process design and improve robustness of this design

New Process

- Develop a refined "should be" process map showing the impact of the solution

Pilot



- Pilot the solution on a small scale to increase buy-in and improve overall implementation

Solution generation and selection

Brainstorming, anti solution, brainwriting, ...

- Generate solutions to address the root causes
- Develop criteria to screen and select solutions including a Cost / Benefit approach to assess solutions.

Finalised Cost / Benefit Analysis

- Finalise cost / benefit analysis of proposed solution

Control

Purpose: **To ensure improvement effectiveness over time** by institutionalisation of the improvement and implementation of ongoing monitoring and reviews.

Implementation Plan

Q x A = E

- Develop a full implementation plan including project and change management elements

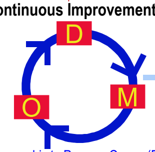
Monitoring Plan

- Develop a monitoring plan to insures gains are held over the long term

Process Capability

- Chart data as evidence that process is in control and meeting customer specifications (variation reduces and process sigma improves)

Continuous Improvement



Documentation / Standardisation

- Document the process with process maps & procedures to assure the solution becomes part of daily work
- Address appropriate changes to broader systems and structures to institutionalize the improvement

- Process ownership to Process Owner (Process Management chart to facilitate transfer)
- Process Owner to held regular process reviews based on dashboards.
- Process Owner to take action when process does not deliver what is expected
- Process has entered Process Management = Define, Measure, Operate

anexas consultancy SE

elt Training

©Anexas