Welcome to Lean and Six Sigma Training

Module 1: Six Sigma Overview

anexas

Roles & Responsibilities

- Owns vision, direction, integration, results
- Leads change

Leadership/ Sr. Executive

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

Yellow Belts

- Supports Black Belts by participating on project teams

Green Belts

All Employees (White Belts)

- Understand vision
- Apply concepts to their job and work area

- Develops deployment and Strategy
- Supports cultural change

Champion / Sponsor

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

Black Belts

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

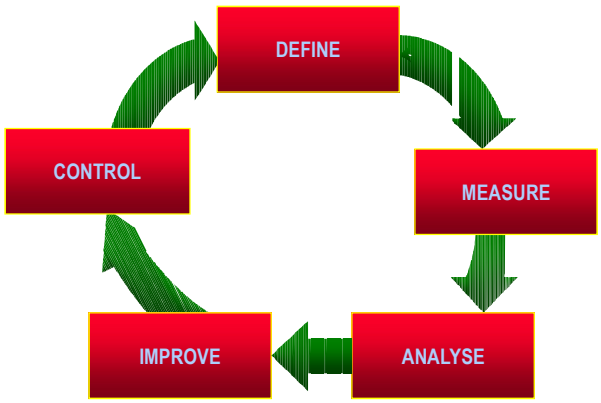
Process Owner

- Trains and coaches Black Belts , Green Belts and leaders

Master Black Belt

anexas

DMAIC : An Improvement Methodology



anexas

DMAIC : An Improvement Methodology

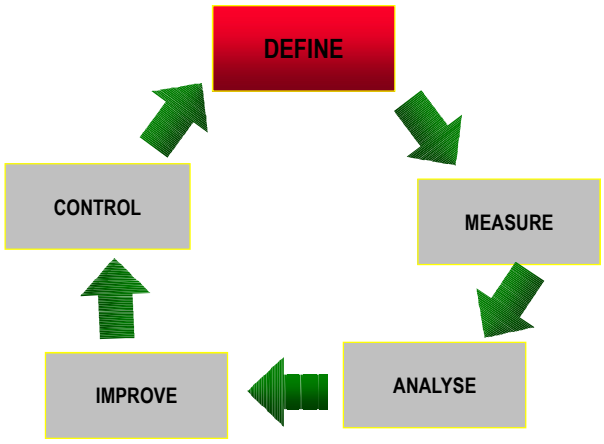
- **DEFINE**: Set direction for improvement
- **MEASURE**: Collect reliable data to understand current process performance
- **ANALYSE**: Identify problem's root causes through process and data analysis
- **IMPROVE**: Determine new improved process design
- **CONTROL**: Ensure improvement effectiveness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

anexas

Module 2: Define Phase

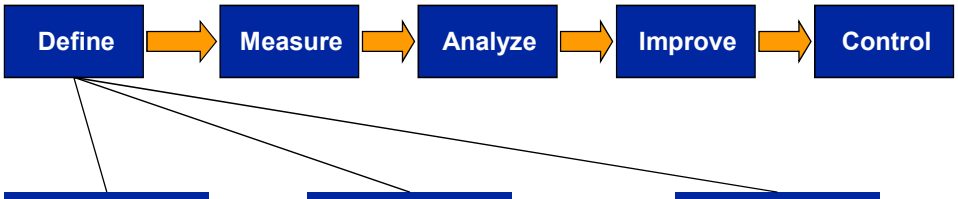
anexas

DMAIC : An Improvement Methodology



anexas

DEFINE Roadmap



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

anexas

DMAIC Project Charter

Project No.:_____

Project Name:	Process :
Resource Plan	**Team Members**
Champion / Sponsor: **Green / Black Belt:** **Functional Managers/Process Owner:** **Coach / Master Black Belt:**	*Text*
Problem Statement	**Scope**
Text	*Text*
Goal Statement	**Customer CTQ's**
Text	*Text*
Estimate Financial Opportunities / Intangible Benefits	**High Level Project Milestone**
Text	*Text*

Validation

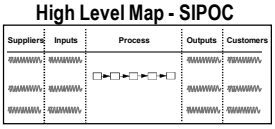
Green / Black Belt	Master Black Belt	Process Owner
CEO	Financial Analyst	Champion / Sponsor

anexas

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.

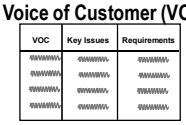
High Level Map - SIPOC



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

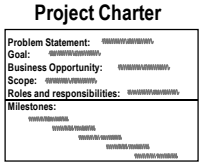
Use Survey or Focus Groups?

Voice of Customer (VOC)



•Gather and display data verifying customer requirements (CTQs)

Project Charter



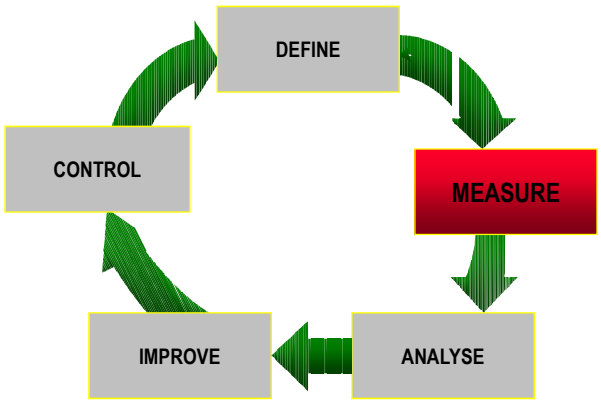
•Develop charter to include:
- Problem statement
- Goal for improvement
- Business opportunity
- Scope of project
- Milestones for completion
- Roles

anexas

Module 3: Measure Phase

anexas

DMAIC : An Improvement Methodology



anexas

Measure

Objective :

- Collect reliable data to understand current process performance

Steps :

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

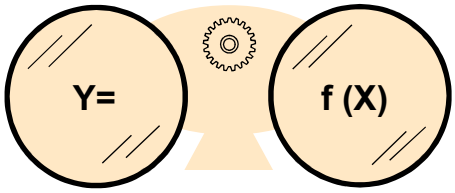
➔ Organize the data collection plan (What ? Why ? When? Who? How? How many ?)

➔ Study process variation

➔ Understand the capability of the process

anexas

Key principles for investigation



Response
- Y
- Dependent
- Output

- Effect
- Symptom
- Monitor

Predictor
- $X_1 ... X_N$
- Independent
- Input-Process variables

- Cause
- Problem
- Control

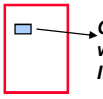
anexas

Compute Process Sigma

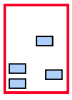
Key Definitions

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)



anexas

Calculate process sigma : formula

Calculate the number of Defects Per Million Opportunities

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

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

anexas CONSULTANCY SE

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 =**

anexas

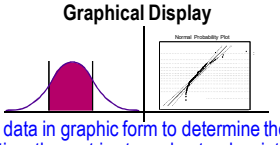
MEASURE

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

Data Collection

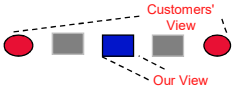
•Develop a data collection plan
- Operational definition
- Sampling

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

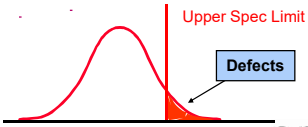
Customer oriented mindset



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

Calculate Process Sigma

Defects "Outside" Spec Limit



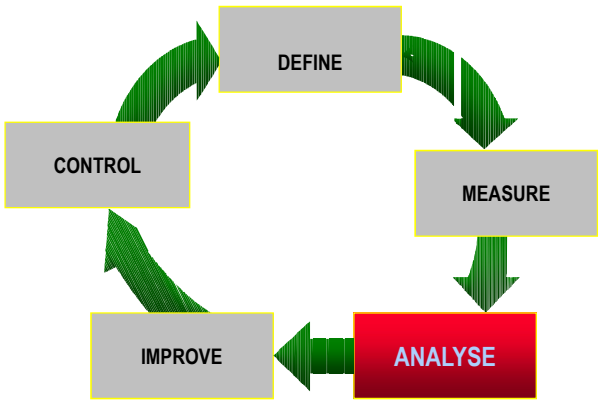
•Compute baseline sigma

anexas

Module 4: Analyse Phase

anexas

DMAIC : An Improvement Methodology



anexas

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

anexas

Introduction to Hypothesis Testing

anexas

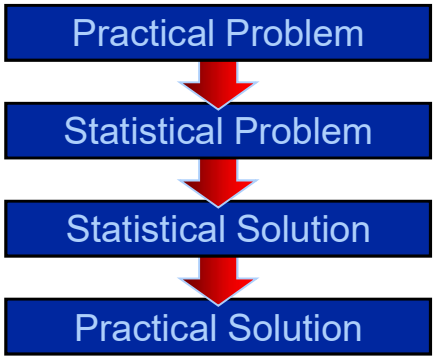
Why Learn Hypothesis Testing?

- 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

anexas

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 _____ ?"



anexas

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

anexas

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

anexas

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

anexas

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

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

anexas

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

anexas

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

anexas

Truth Table: α and β Risk

Truth	Decision: Fail to reject H_o	Decision: 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

anexas

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

anexas

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]

2. State the alternative hypothesis (Ha)

   Atleast 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)

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

4. Collect data

anexas

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

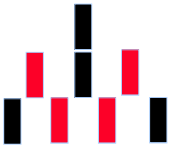
What are Hypothesis Tests of Mean?

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



vs

target value

μ

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

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.

anexas

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**

anexas

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

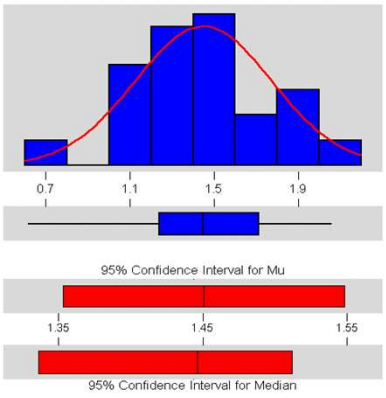
anexas

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



Variable: Cost

Anderson-Darling Normality Test

A-Squared:	0.448
P-Value:	0.267
Mean	1.45055
StDev	0.32465
Variance	0.105400
Skewness	-2.0E-01
Kurtosis	0.169315
N	45
Minimum	0.61651
1st Quartile	1.23956
Median	1.44628
3rd Quartile	1.71317
Maximum	2.05628

95% Confidence Interval for Mu

1.35302 1.54809

95% Confidence Interval for Sigma

0.26877 0.41010

95% Confidence Interval for Median

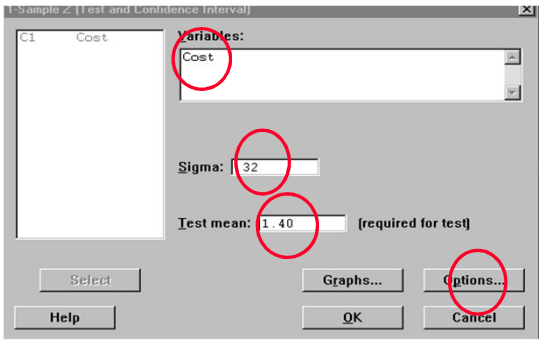
1.33624 1.51171

anexas

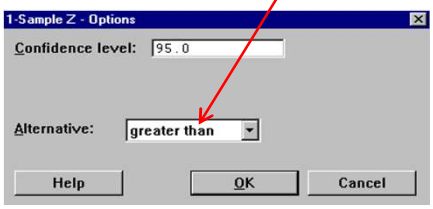
Example: Rising Packing Costs

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

Analysis through Minitab



Choose *greater than* for Alternative
(H_a: $\mu > \$1.40$)



anexas

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

anexas

1 Sample t Test

anexas

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

anexas

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

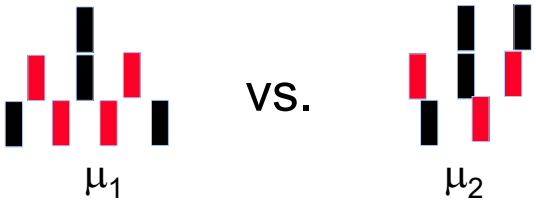
Note: Normality assumptions relaxed when the number of sample observations is large (generally true when sample size >30).

anexas

2 Sample t Test

anexas

Two Sample Comparison



Practical Question
(example)

"Are the two populations statistically different?"

Statistical Question
H_o: $\mu_1 = \mu_2$
H_α: $\mu_1 \neq \mu_2$

anexas

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

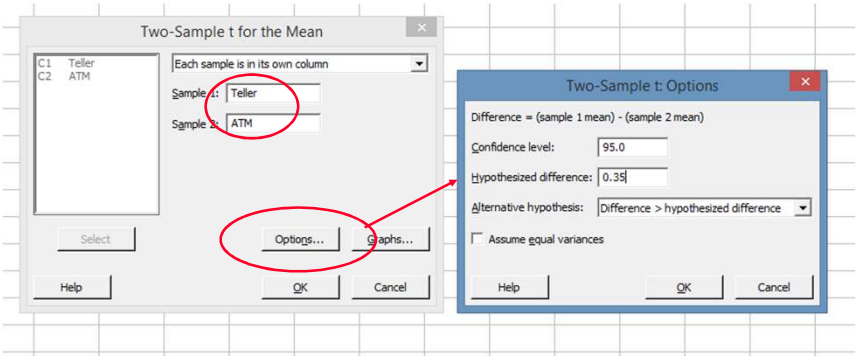
  $\alpha = 0.05$

- **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



anexas

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

anexas

ANOVA
Analysis of Variance

anexas

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

anexas

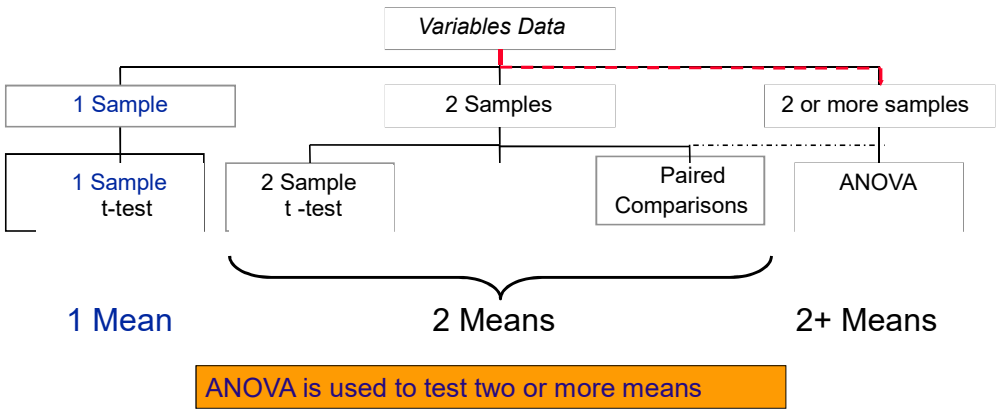
What is ANOVA?

- 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

anexas

When to Use ANOVA

Variables Road Map



ANOVA is used to test two or more means

anexas

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

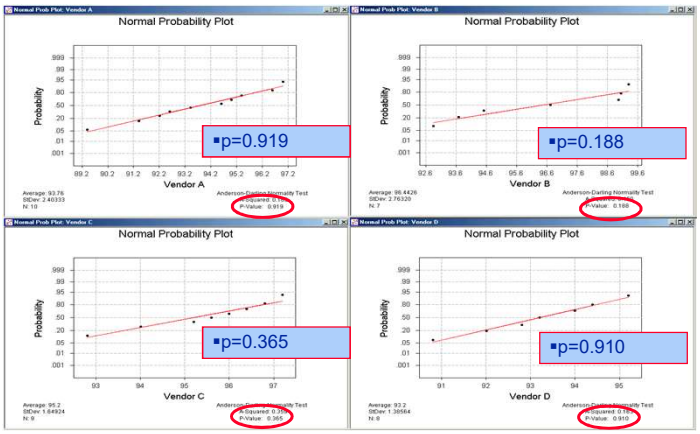
anexas

Prerequisites for ANOVA

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

anexas

Testing Data for Normality

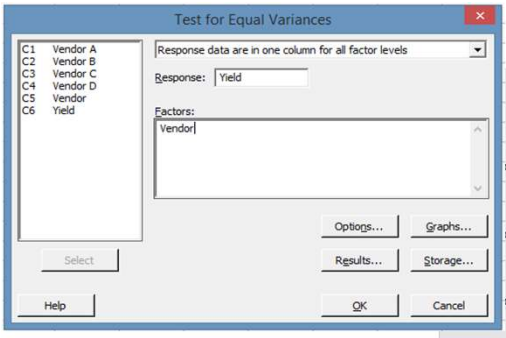


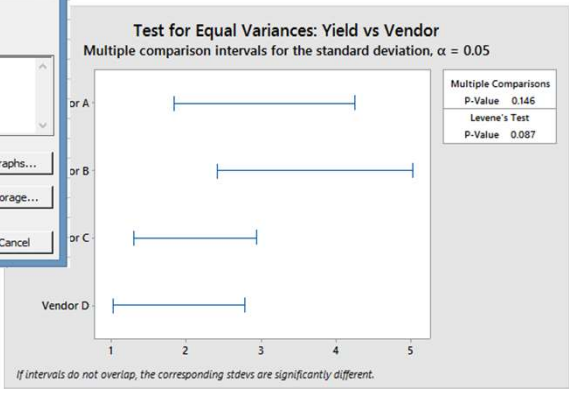
All subgroups have a normal distribution

anexas

Testing Data for Equal Variances

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



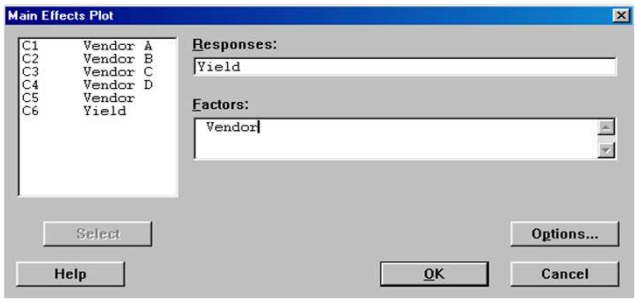


P=0.146 indicates variances are equal

anexas

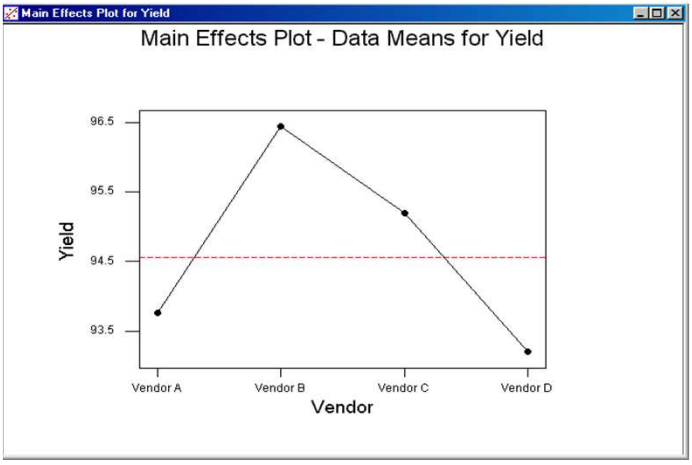
Running a Main Effects Plot

Tool Bar Menu > Stat > ANOVA > Main Effects Plot



anexas

The Main Effects Plot

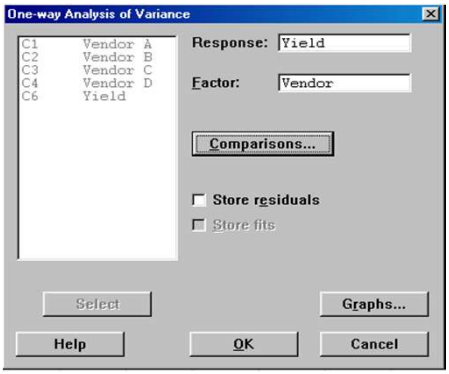


- The plot shows the output vs. the factor

anexas

Running a One Way ANOVA

Tool Bar Menu > Stat > ANOVA > One Way…



anexas

The One Way ANOVA

One-way ANOVA: Yield versus Vendor

```
Analysis of Variance for Yield
Source     DF        SS        MS        F        P
Vendor      3     49.69     16.56     3.74    0.022
Error      30    133.00      4.43
Total      33    182.69

                                 Individual 95% CIs For Mean
                                 Based on Pooled StDev
Level       N      Mean     StDev  --+---------+---------+---------+----
Vendor A   10    93.760     2.403      (------*------)
Vendor B    7    96.443     2.763                (-------*-------)
Vendor C    9    95.200     1.649           (------*------)
Vendor D    8    93.200     1.386  (-------*-------)
                                   --+---------+---------+---------+----
Pooled StDev =    2.106           92.0      94.0      96.0      98.0
```

▪p < 0.05: source is significant!

anexas

Hypothesis Testing of Proportions

anexas

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

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

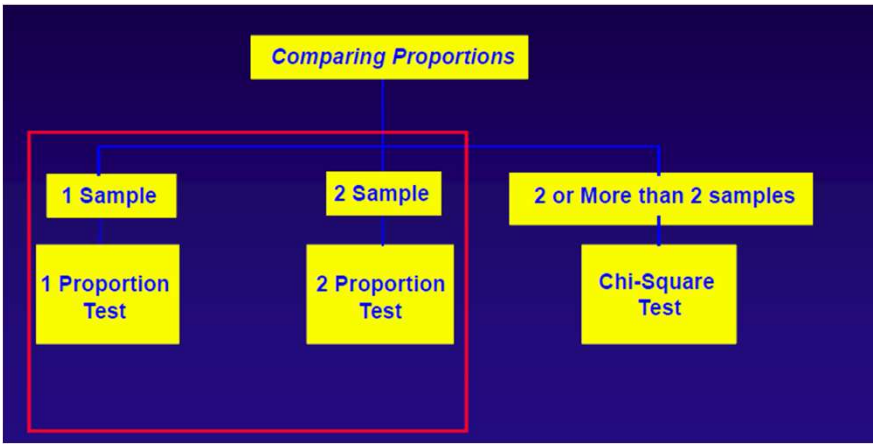
anexas

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

Hypothesis Testing of Proportion - Roadmap

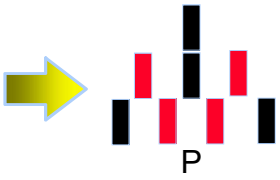


anexas

Comparison of Proportion: 2 Scenarios

1) Single Proportion Comparison

   One population proportion compared to a target value



vs. target value

2) Two Sample Comparison

   Proportions of two independent populations compared to each other

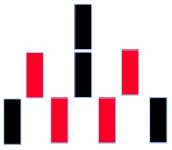
vs.

anexas

1 Proportion Test

anexas

Single Proportion Comparison



P

vs.

target value

Practical Question
(example)

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

Statistical Question

H_0 : P = target value

H_a : P ≠ target value

anexas

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

α = 0.05

What hypothesis test is appropriate?

These hypotheses deal with proportions

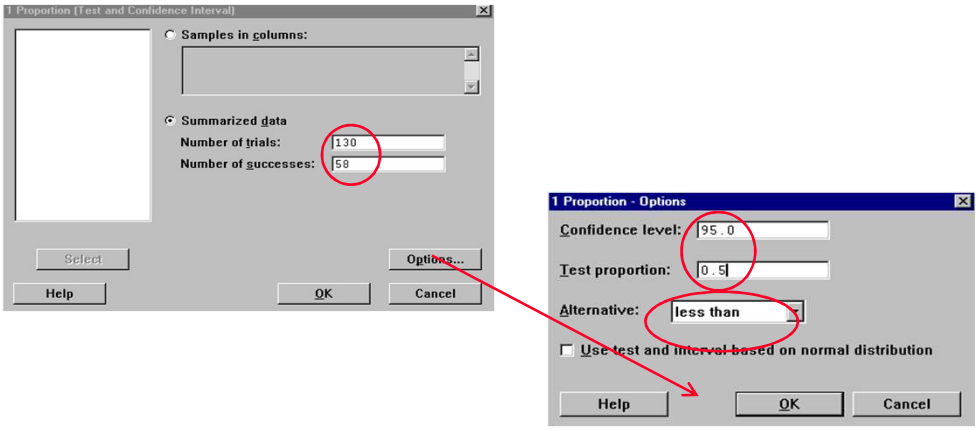
Comparing population proportion against a target proportion using one sample data

Use 1 Proportion Test

Example: IPO Prospectus

Tool Bar Menu > Stat > Basic Statistics > 1 Proportion

Analysis through Minitab



anexas

Example: IPO Prospectus

Test and CI for One Proportion

Test of p = 0.5 vs p < 0.5

Sample	X	N	Sample p	95.0% Upper Bound	Exact 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

2 Proportion Test

anexas

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

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: $P_A - P_B > 0$

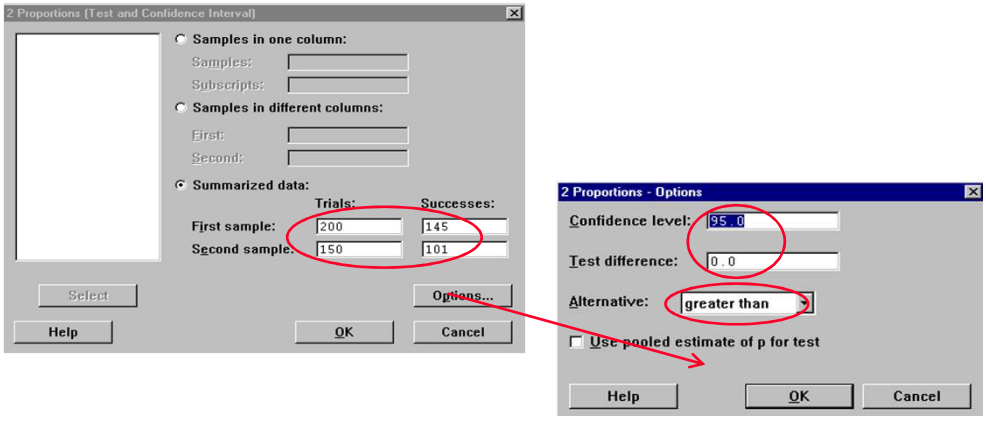
$\alpha = 0.05$

- 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

▪Example: Comparing Medicines

```
Test and CI for Two Proportions
Sample            X       N       Sample p
1               145     200       0.725000
2               101     150       0.673333

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

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

anexas

Hypothesis Testing
Chi-Square Tests

anexas

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)

anexas

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[SM] Projects

A sample of Black Belts[SM] was asked to rate both their six sigma project performance and the average weekly hours spent with the Project Champion[SM] 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

anexas

Example: Black Belt[SM] Projects

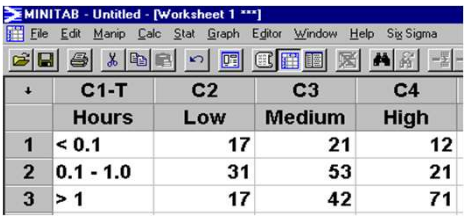
- Practical problem
  - Does the performance of Black Belt[SM] projects depend on time spent with Project Champions[SM]?
- Statistical problem
  - Is there an association between project performance and time spent with Champion[SM]?
  - Null hypothesis: project performance is independent of the time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

anexas

Example: Black Belt[SM] Projects

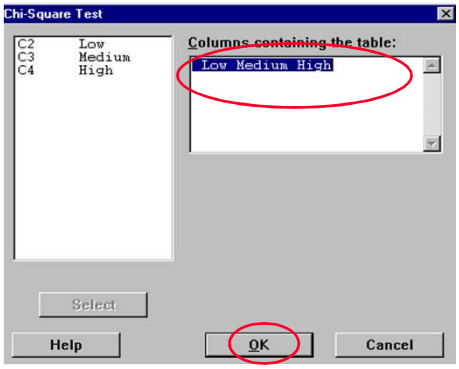
Tool Bar Menu > Stat > Tables > Chi-Square Test

1. Open data file Chi1.mtw



	C1-T	C2	C3	C4
	Hours	Low	Medium	High
1	< 0.1	17	21	12
2	0.1 - 1.0	31	53	21
3	> 1	17	42	71

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



4. Click OK

anexas

Example: Black Belt[SM] Projects

Chi-Square Test: Low, Medium, High

```
Expected counts are printed below observed counts
          Low   Medium     High    Total
    1      17       21       12       50
         11.40    20.35    18.25

    2      31       53       21      105
         23.95    42.74    38.32

    3      17       42       71      130
         29.65    52.91    47.44

Total      65      116      104      285

Chi-Sq =  2.747 +  0.021 +  2.138 +
          2.077 +  2.465 +  7.825 +
          5.396 +  2.250 + 11.702 = 36.622
DF = 4, P-Value = 0.000
```

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

anexas

Example: Black Belt℠ Projects

- Interpretation:
  - p-value = 0.000
  - p-value < α-risk (0.01): reject H_o
  - Infer H_a: sufficient evidence that Black Belt℠ project performance and and time spent with Champion℠ are dependent

anexas

Hypothesis Testing- Correlation and Regression

anexas

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

anexas

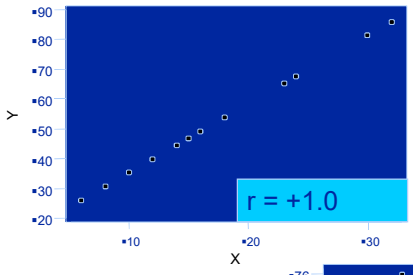
What is Correlation?

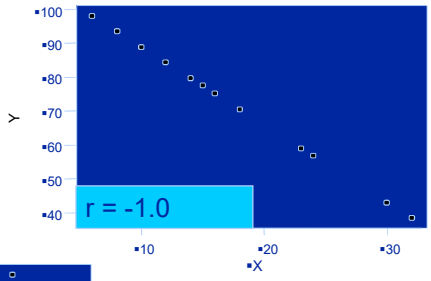
- Correlation is a measure of the strength of association between two quantitative variables
  (Ex: Pressure and Yield)
- Measures the degree of linearity between two variables assumed to be completely independent of each other

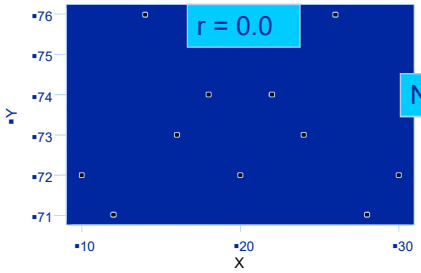
Correlation coefficient or P*earson* correlation coefficient is a way of measuring the strength of correlation

anexas

Correlation Coefficient





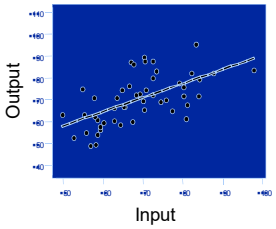


No correlation

anexas

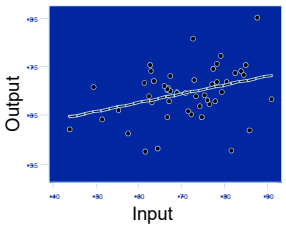
Strength and Direction of "+" Correlation

Moderate positive correlation



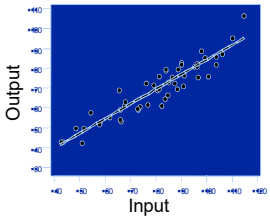
Y=25.7595+0.645418X
R Squared=0.369

Weak positive correlation



Y=56.6537+0.181987X
R Squared=0.115

Strong positive correlation

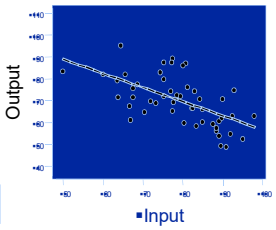


Y=9.77271+0.745022X
R Squared=0.876

anexas

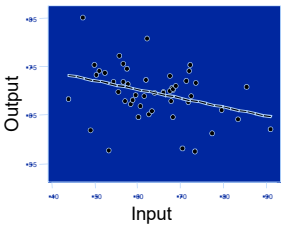
Strength and Direction of "-" Correlation

Moderate negative correlation



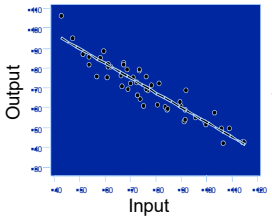
Y=90.3013-0.645418X
R Squared=0.369

Weak negative correlation



Y=74.8524-0.181987X
R Squared=0.115

Strong negative correlation



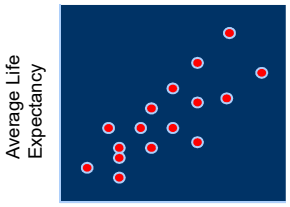
Y=99.1754-0.745022X
R Squared=0.876

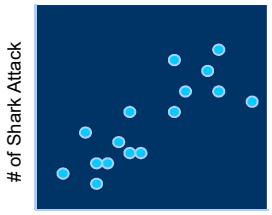
anexas

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

anexas

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

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

anexas

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'

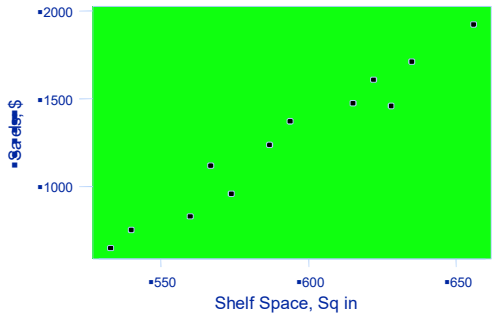
anexas

Example: Cereal Sales

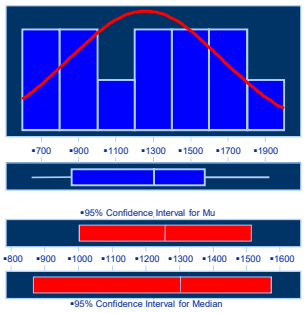
Tool Bar Menu > Stat > Basic Statistics > Display Descriptive Statistics

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

Graph > Plot



Descriptive Statistics



Variable: Sales, $

Anderson-Darling Normality Test

A-Squared:	0.177
P-Value:	0.898
Mean	1258.00
StDev	402.92
Variance	162346
Skewness	-1.8E-02
Kurtosis	-1.04056
N	12
Minimum	651.00
1st Quartile	863.25
Median	1304.00
3rd Quartile	1575.00
Maximum	1924.00

95% Confidence Interval for Mu: 1002.00 – 1514.00

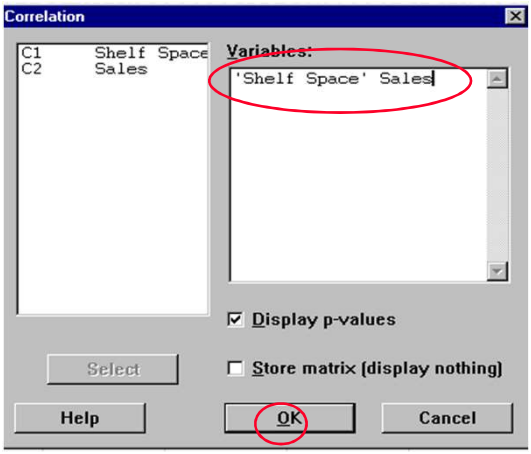
95% Confidence Interval for Sigma: 285.43 – 684.11

95% Confidence Interval for Median: 864.94 – 1573.22

anexas

Example: Cereal Sales

Tool Bar Menu > Stat > Basic Statistics > Correlation



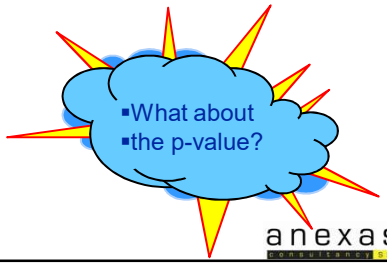
anexas

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



anexas

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

Simple Non-Linear Regression

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

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

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

Simple Linear Regression

Simple Linear Regression

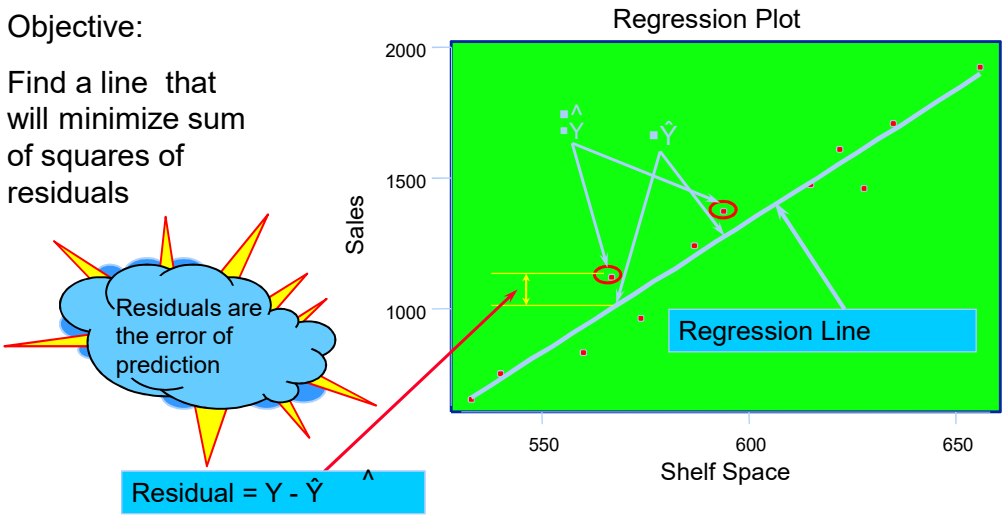
- Use one independent variable (x) to explain the variation in dependent variable (y)
  - Example 1: use shelf space to explain variation sales $
  - Example 2: amount of fertilizer applied to explain the yield of crop
- Method of Least Squares
  - Use the 'Method of Least Squares' to find the best fitting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

anexas

Method of Least Squares

Objective:

Find a line that will minimize sum of squares of residuals



anexas

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

Example: Cereal Sales

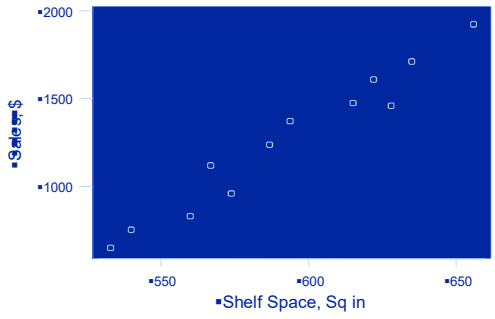
Tool Bar Menu > Stat > Basic Statistics > Display Descriptive Statistics

Practical and Graphical:

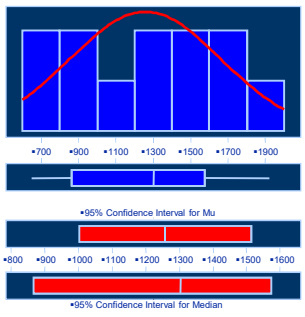
Practical questions about the process?

Plot the data using different techniques

Graph > Plot



Descriptive Statistics



Variable: Sales, $

Anderson-Darling Normality Test

A-Squared:	0.177
P-Value:	0.898
Mean	1258.00
StDev	402.92
Variance	162346
Skewness	-1.8E-02
Kurtosis	-1.04056
N	12
Minimum	651.00
1st Quartile	863.25
Median	1304.00
3rd Quartile	1575.00
Maximum	1924.00

95% Confidence Interval for Mu: 1002.00 – 1514.00

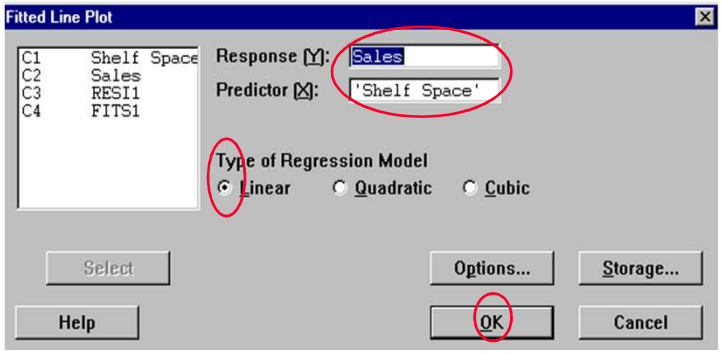
95% Confidence Interval for Sigma: 285.43 – 684.11

95% Confidence Interval for Median: 864.94 – 1573.22

anexas

Example: Cereal Sales

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

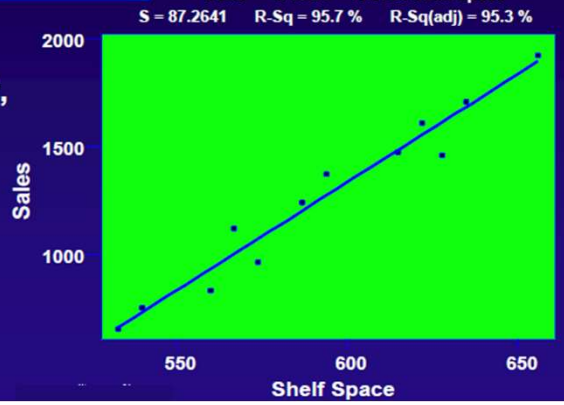


anexas

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?



anexas

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

anexas

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			

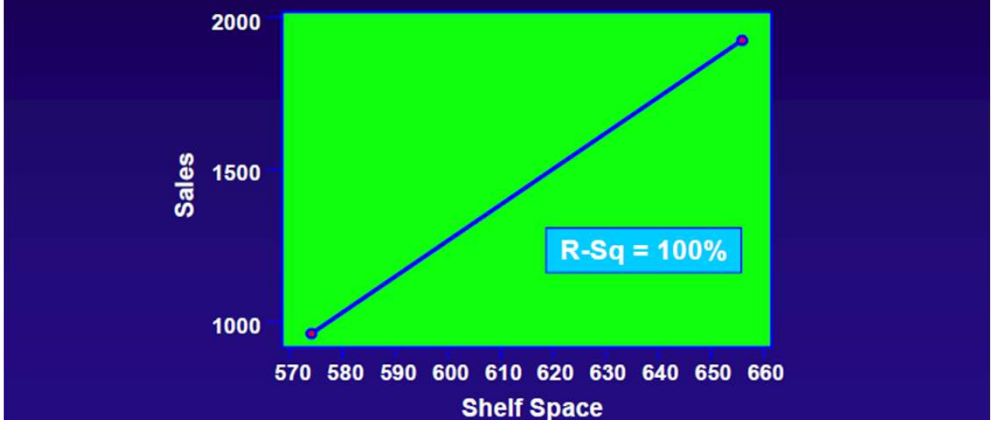
$R^2 = 1709656 / 1785806 = 95.74\%$

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

anexas

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?

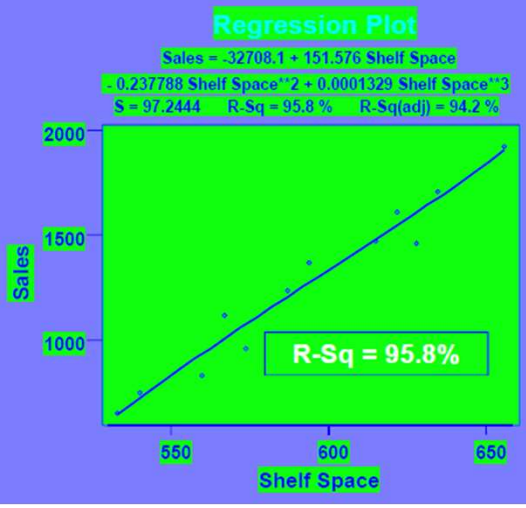


anexas

Example: Cereal Sales

Tool Bar Menu > Stat > Regression > Fitted Line Plot

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



anexas

Example: Cereal Sales

- Which model is better? Linear or Cubic model?

- R-Squared gets bigger as we add more and more terms!
- So should we keep adding terms?

anexas

What is R-Sq (adj)?

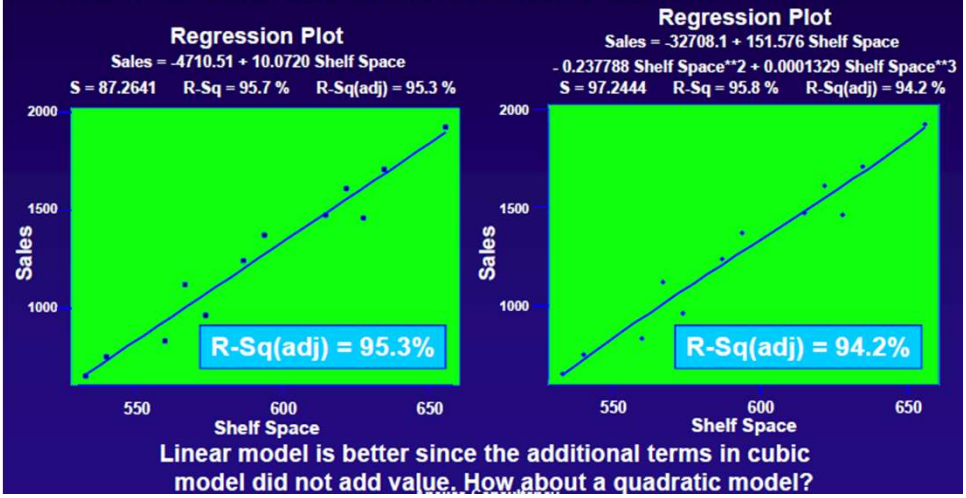
- More realistic measurement and is a modified measure of R-squared
- Takes into account of number of terms in the model and number of data points
- Adj $R^2 = 1 - [SS_{error} / (n-p)] / [SS_{total} / (n-1)] = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$

  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

anexas

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

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?

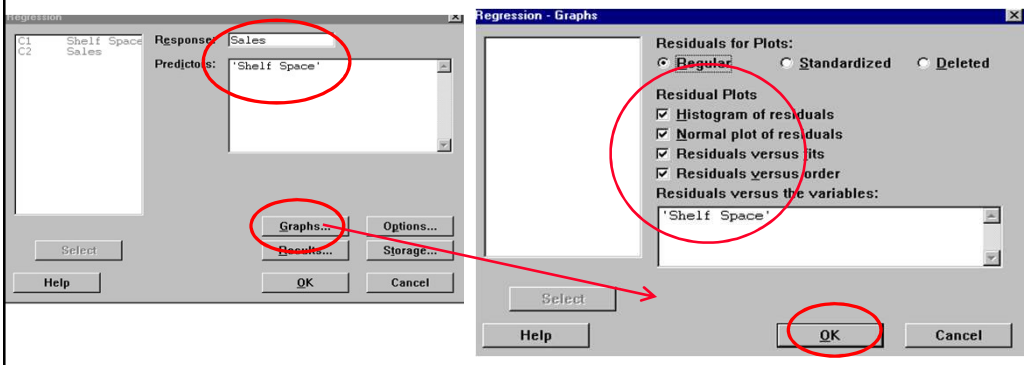
anexas

Checking Assumptions

Tool Bar Menu > Stat > Regression > Regression

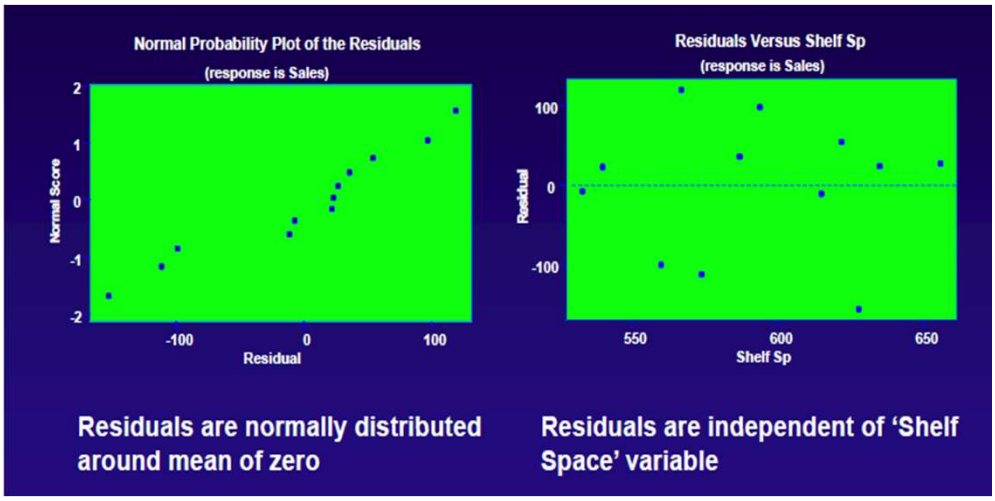
Residuals are error in the fit of regression line

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



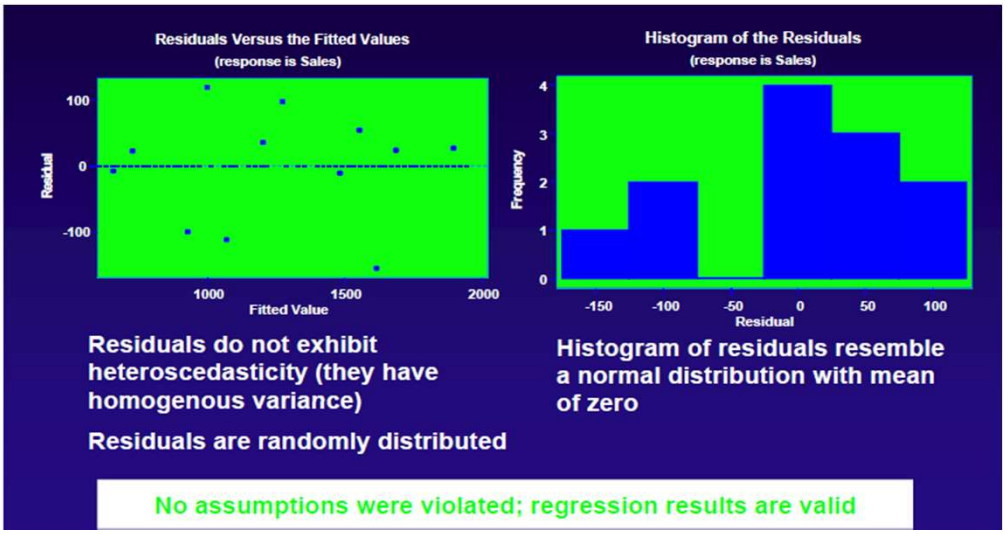
anexas

Assumptions for Regression



anexas

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

anexas

Multiple Regression

anexas

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

$$Y = f(x_1, x_2, x_3, \ldots.)$$

anexas

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

anexas

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

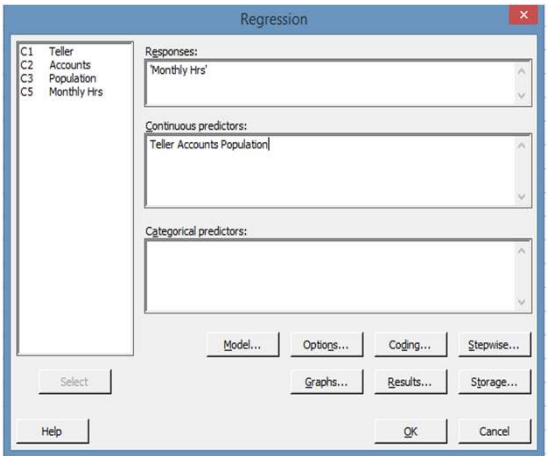
A metric, called variance inflation factor (VIF) calculates the degree of multicollinearity

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

- 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




```
Results for: Banking.MTW

Regression Analysis: Monthly Hrs versus Teller, Accounts, Population

Analysis of Variance

Source        DF   Adj SS   Adj MS  F-Value  P-Value
Regression     3 34184634 11394878  1035.40    0.000
  Teller       1  3332064  3332064   302.77    0.000
  Accounts     1    36039    36039     3.27    0.094
  Population   1  8825347  8825347   801.91    0.000
Error         13   143069    11005
Total         16 34327704


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
104.906  99.58%     99.49%      99.23%


Coefficients

Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant       954      385     2.48    0.028
Teller      0.8250   0.0474    17.40    0.000  6.50
Accounts     0.240    0.133     1.81    0.094  6.54
Population 0.05563  0.00196    28.32    0.000  1.02


Regression Equation

Monthly Hrs = 954 + 0.8250 Teller + 0.240 Accounts + 0.05563 Population
```


anexas

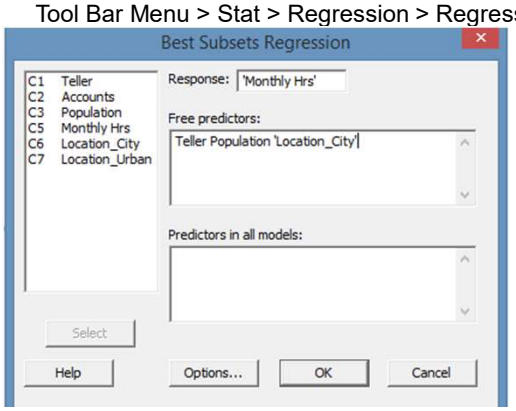
Predictor Variable Selection

- 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

anexas

Best Subsets

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



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

anexas

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)

anexas

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.

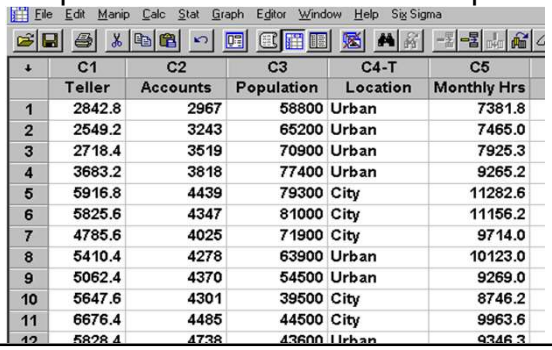
Business Process Example: Bank Labor Hours

- **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**

anexas

Example: Bank Labor Hours

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



↓	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

anexas

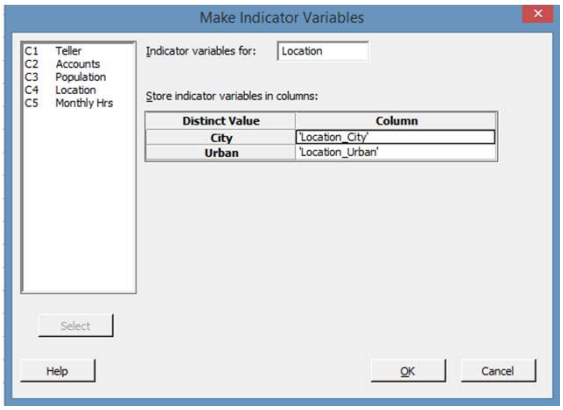
Example: Bank Labor Hours

▪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. Create an indicator variable for “Location”
   - Toolbar>Calc> Make indicator Variables



Example: Bank Labor Hours

- Result should follow the pattern below
- Order 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).

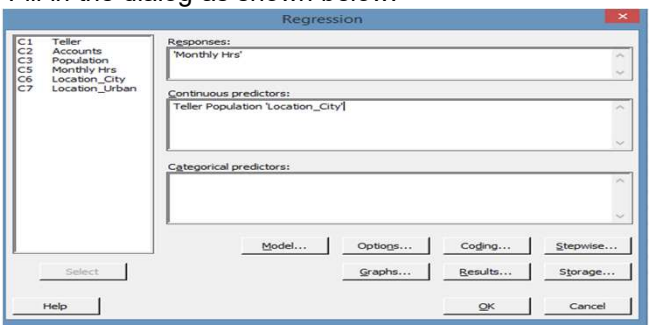
	Location	Urban	City	M
0	Urban	1	0	
0	Urban	1	0	
0	Urban	1	0	
0	Urban	1	0	
0	City	0	1	
0	City	0	1	
0	City	0	1	
0	Urban	1	0	
0	Urban	1	0	
0	City	0	1	
0	City	0	1	
0	Urban	1	0	

Example: Bank Labor Hours

- Create an indicator variable for “Location”
- 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

- Multiple regression steps:
- 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:



anexas

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

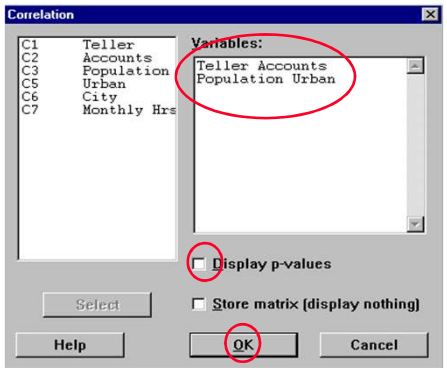
Predictor	Coef	SE Coef	T	P	VIF
Constant	1094.3	147.8	7.41	0.000	
Teller	0.95103	0.02310	41.17	0.000	10.6
Accounts	0.00913	0.05706	0.16	0.876	8.3
Populati	0.0563806	0.0007550	74.68	0.000	1.0
Urban	238.49	27.15	8.78	0.000	1.9

anexas

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



anexas

Example: Bank Labor Hours

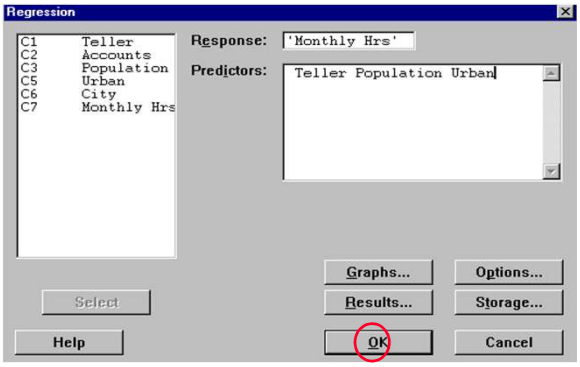
Next step: remove Accounts and check multicollinearity

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

anexas

Example: Bank Labor Hours

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



anexas

Example: Bank Labor Hours

All VIF <5

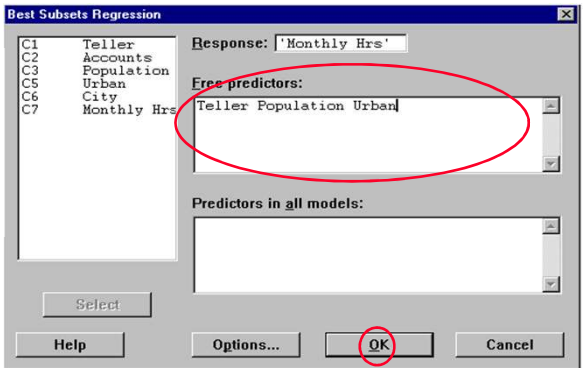
Predictor	Coef	SE Coef	T	P	VIF
Constant	1114.38	74.88	14.88	0.000	
Teller	0.954454	0.008388	113.78	0.000	1.5
Populati	0.0563707	0.0007237	77.89	0.000	1.0
Urban	240.49	23.18	10.38	0.000	1.5

anexas

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

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

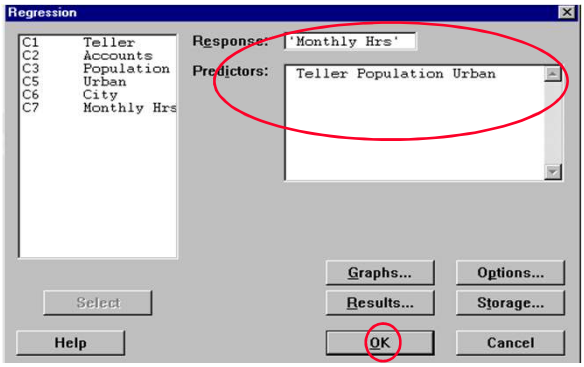
anexas

Example: Bank Labor Hours

- Multiple regression steps:

3. Choose the best candidate and complete regression analysis

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



anexas

Example: Bank Labor Hours

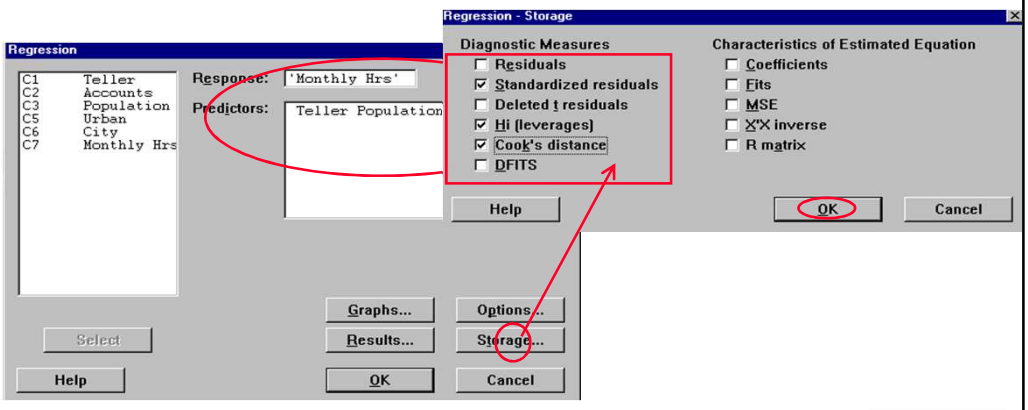
```
The regression equation is
Monthly Hrs = 1114 + 0.954 Teller + 0.0564 Population +
240 Urban

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%
```


anexas

Example: Bank Labor Hours

Multiple regression steps:

4. Perform model diagnostics to identify outliers and unusual observations



anexas

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

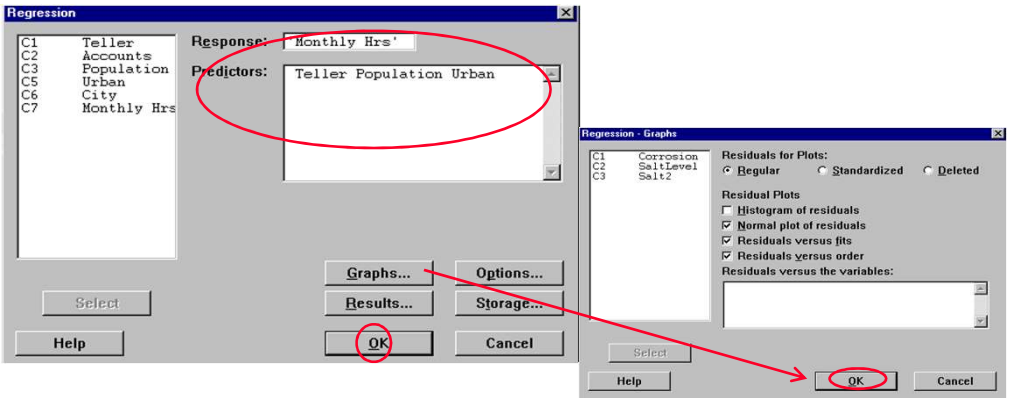
What actions should be taken with the outliers and influential observations, if any ?

anexas

Example: Bank Labor Hours

Multiple regression steps:

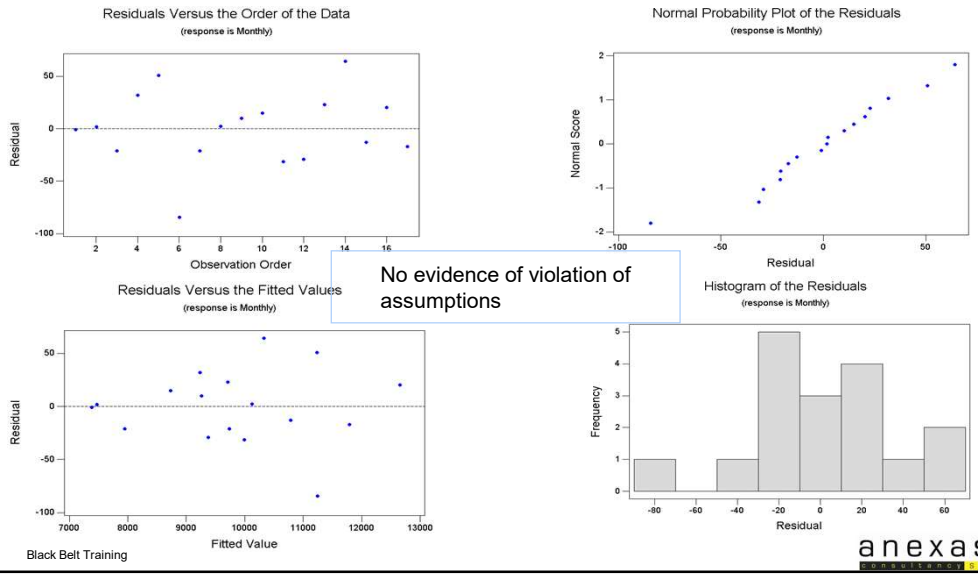
5. Analyze residuals for violation of assumptions



anexas

Example: Bank Labor Hours

Residual Plot Analysis



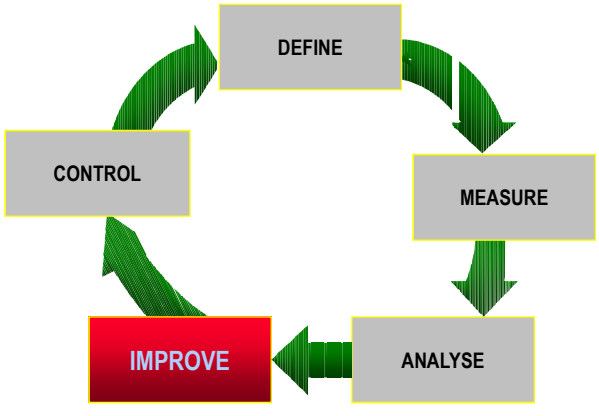
No evidence of violation of assumptions

anexas

Module 5: Improve Phase

anexas

DMAIC : An Improvement Methodology



anexas

Improve

Objective :

Determine new improved process design

Steps :

Generate solutions

Select and test solutions

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

anexas

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

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

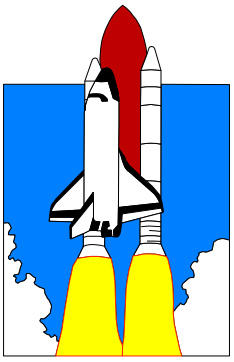
Solution B = development of our own software

Solution C = buy a software and adapt to our needs

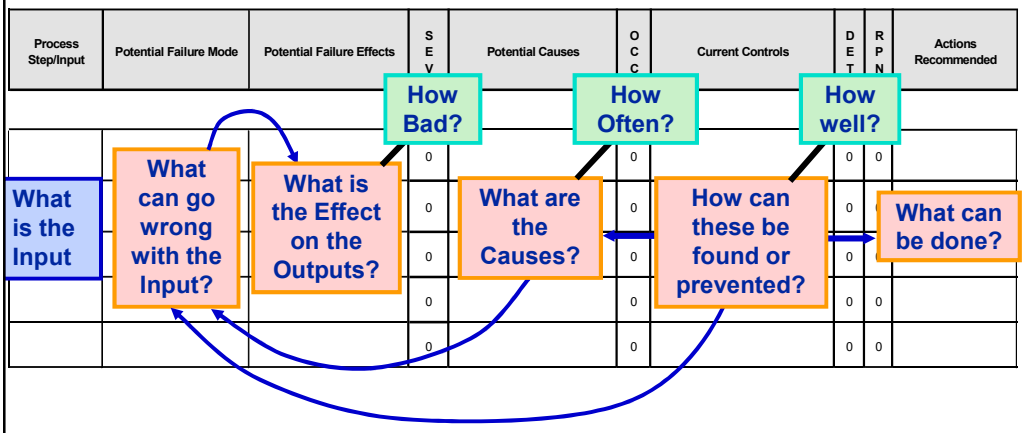
It seems here that solution C is the most satisfying. B also can be considered as an option.

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

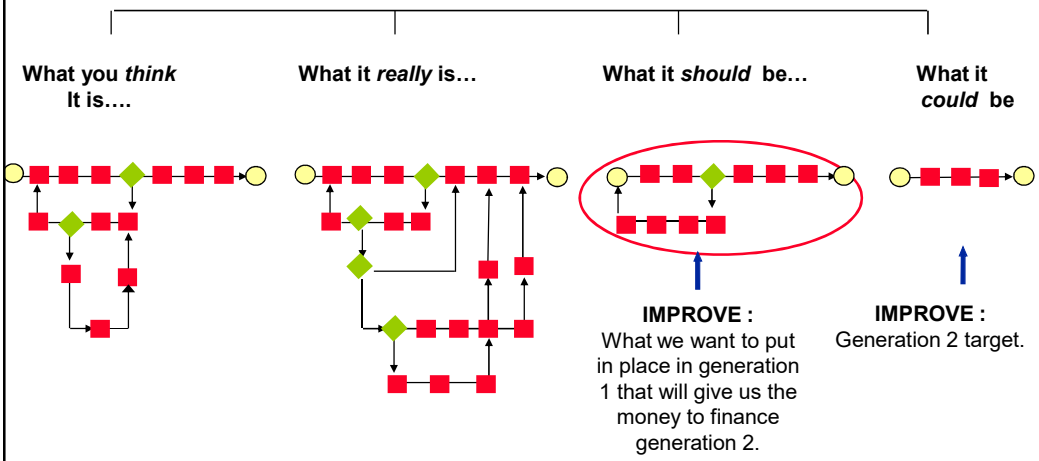


Overview



anexas

Continuous Improvement



anexas

Benefits of doing a pilot

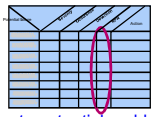
- Improve the solution that meets customer requirements
- Refine implementation plan
- Lower risk of failure by identifying and fixing possible problems ahead of time
- Confirming expected results and relations between predictive parameters and results (Xs on Y)
- Increase opportunities to receive feedback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

anexas

IMPROVE

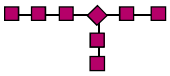
Purpose : **To determine new improved process design** through idea generation, selection, process design, solution testing , and improvements implementation.

Solutions Refinement



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

New Process

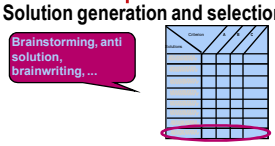


Develop a "should be" process map showing the impact of the solution

Pilot

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

Solution generation and selection



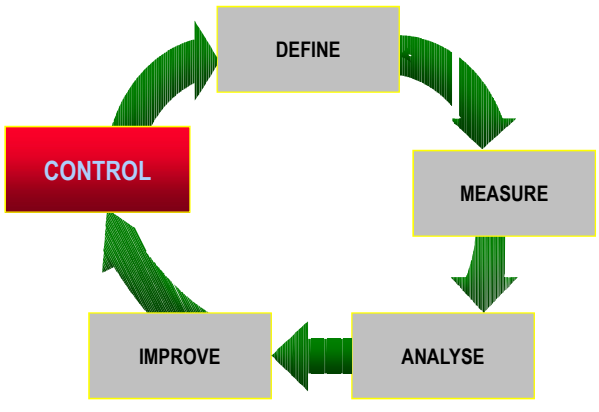
Generate solutions to address the root causes and develop criteria to screen and select solutions (including cost / benefit)

Perform cost / benefit analysis of proposed solution

anexas

Module 6: Control Phase

DMAIC : An Improvement Methodology



Control

Objective :

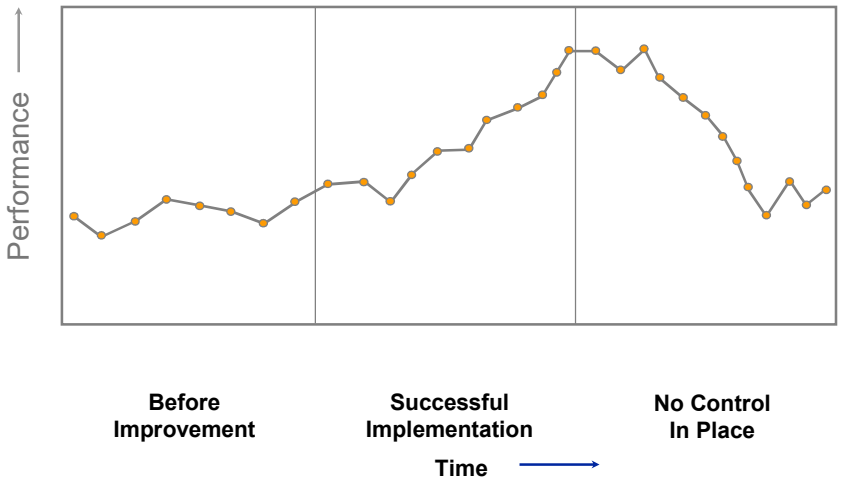
- Ensure improvement over time

Steps :

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

anexas

Control = ensure gains over time



anexas

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

anexas

Mistake Proofing (Poka-Yoke)

anexas

What Is Mistake Proofing (Poka-Yoke)?

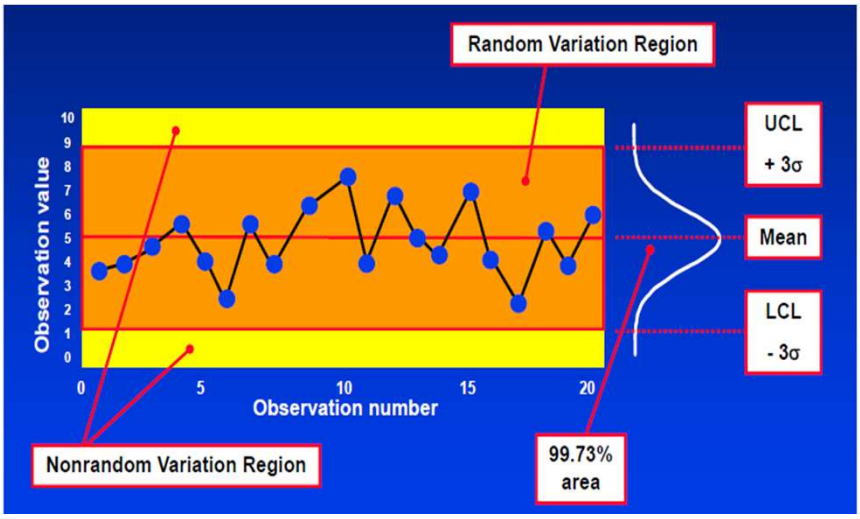
- Japanese phrase:
- Yokeru (to avoid), Poka (errors)
- A strategy for preventing errors in processes
- Makes it impossible for defects to pass unnoticed
- Corrects problems as soon as they are detected
- Technique detects defects
- Prevents defects from moving into next area
- Developed by Dr. Shigeo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

Statistical Process Control for Variables Data (SPC)

Introduction to SPC

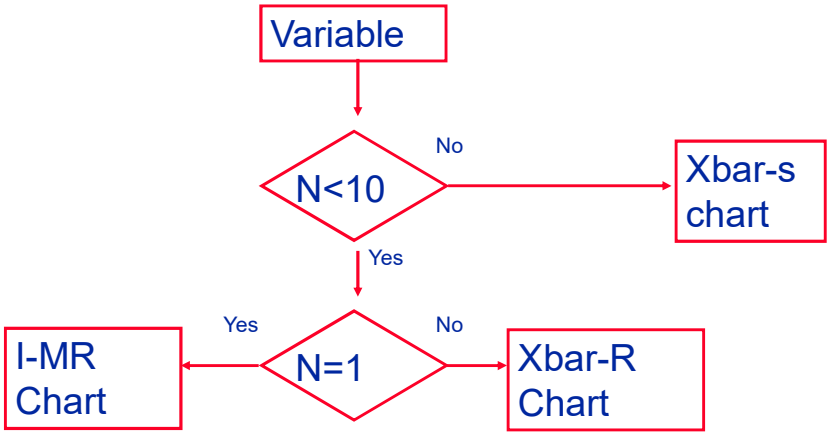
anexas

Statistics of a Control Chart



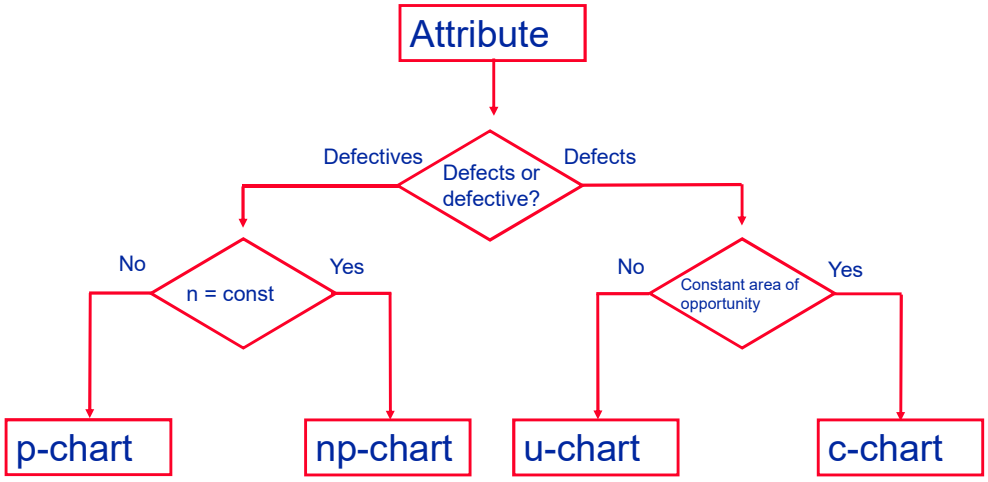
anexas

Control Chart Roadmap



anexas

Control Chart Roadmap

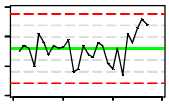


anexas

CONTROL

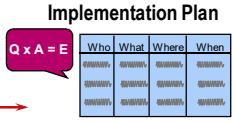
Purpose: **To ensure improvement effectiveness over time** by institutionalisation of the improvement and implementation of ongoing monitoring and reviews.

Monitoring Plan



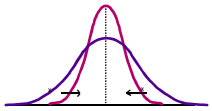
Develop a monitoring plan to insures gains are held over the long term

Implementation Plan



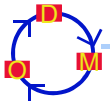
Develop a full implementation plan including project and change management elements

Process Capability



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

Continuous Improvement



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

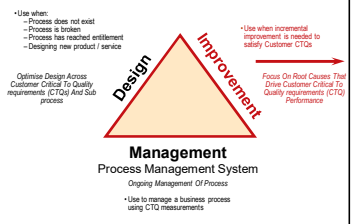
- Process ownership to Process Owner (Process Management chart to facilitate transfer)
- Process Owner to held regular process reviews based on dashboards.
- Process Owner to take action when process does not deliver what is expected
- Process has entered Process Management = Define, Measure , Operate.

Black Belt Training

anexas consultancy SE

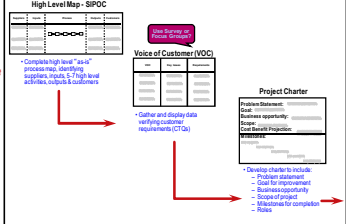
The DMAIC Storyboard : Six Sigma for Process Improvement

The Three Dimensions Of Process Focus



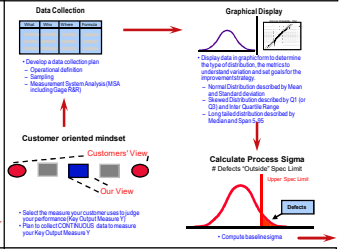
Define

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.



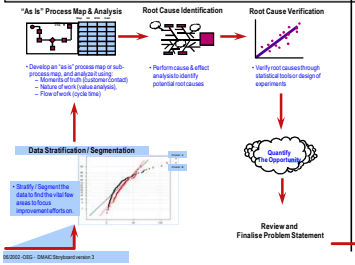
Measure

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



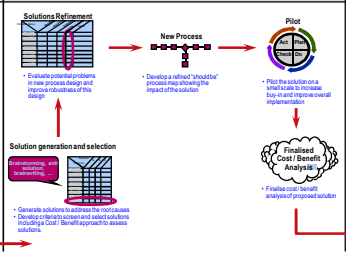
Analyse

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



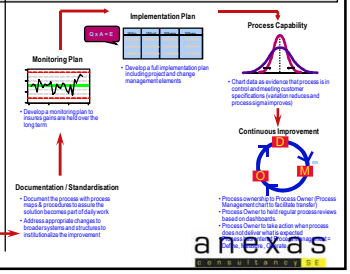
Improve

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



Control

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



anexas consultancy SE