INTRODUCTION TO SIX SIGMA APPLICATIONS

What is Six Sigma?

A Philosophy

- Customer Critical To Quality (CTQ) Criteria
- Breakthrough Improvements
- Fact-driven, Measurement-based, Statistically Analyzed Prioritization
- Controlling the Input & Process Variations Yields a Predictable Product

A Quality Level

- 6σ = 3.4 Defects per Million Opportunities

A Structured Problem-Solving Approach

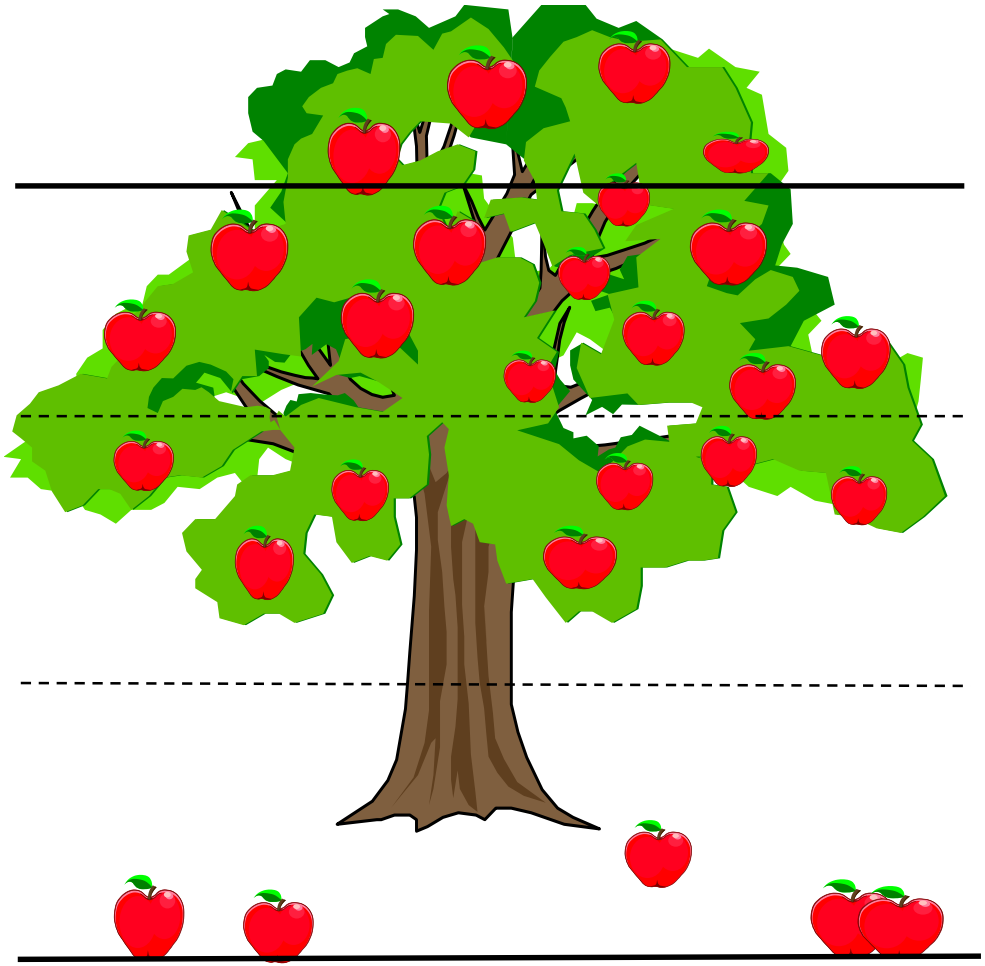
- Phased Project: Measure, Analyze, Improve, Control

A Program

- Dedicated, Trained BlackBelts
- Prioritized Projects
- Teams - Process Participants & Owners

POSITIONING SIX SIGMA

THE FRUIT OF SIX SIGMA



Sweet Fruit
Design for Manufacturability

Process Entitlement

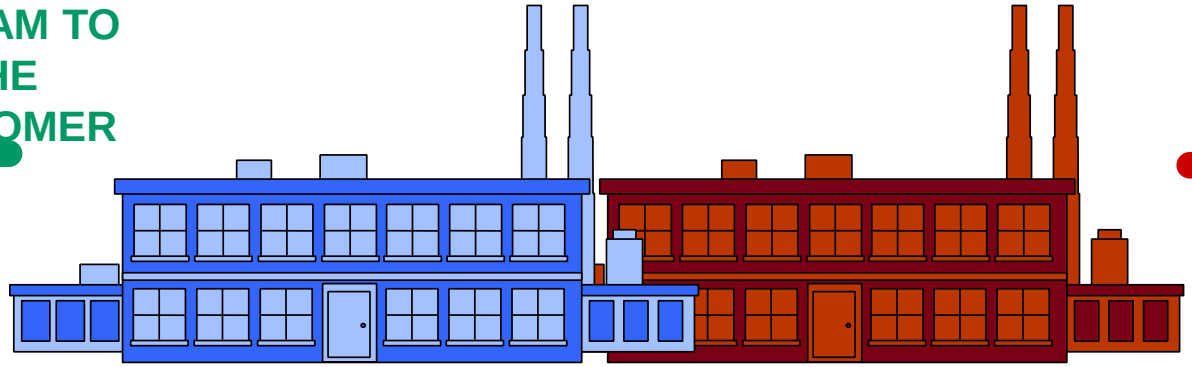
Bulk of Fruit
Process Characterization and Optimization

Low Hanging Fruit
Seven Basic Tools

Ground Fruit
Logic and Intuition

UNLOCKING THE HIDDEN FACTORY

VALUE STREAM TO THE CUSTOMER



WASTE DUE TO INCAPABLE PROCESSES

PROCESSES WHICH PROVIDE PRODUCT VALUE IN THE CUSTOMER'S EYES

- FEATURES OR CHARACTERISTICS THE CUSTOMER WOULD PAY FOR….

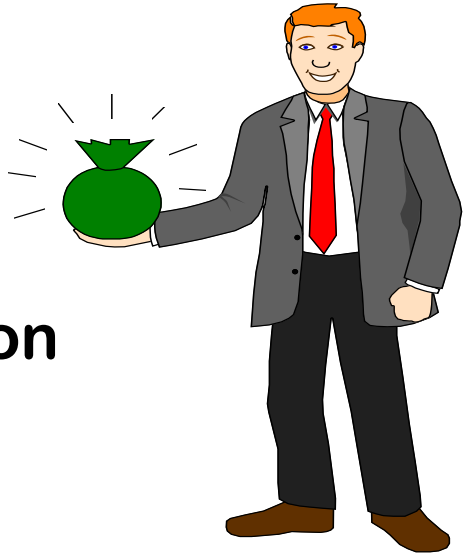
WASTE SCATTERED THROUGHOUT THE VALUE STREAM

- EXCESS INVENTORY
- REWORK
- WAIT TIME
- EXCESS HANDLING
- EXCESS TRAVEL DISTANCES
- TEST AND INSPECTION

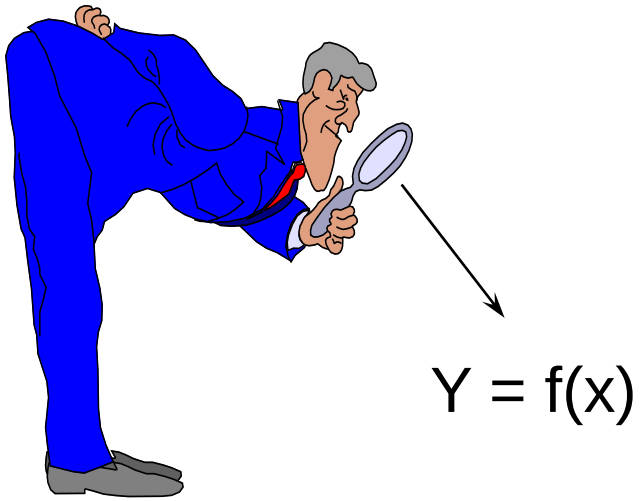
Waste is a significant cost driver and has a major impact on the bottom line...

Common Six Sigma Project Areas

- Manufacturing Defect Reduction
- Cycle Time Reduction
- Cost Reduction
- Inventory Reduction
- Product Development and Introduction
- Labor Reduction
- Increased Utilization of Resources
- Product Sales Improvement
- Capacity Improvements
- Delivery Improvements



The Focus of Six Sigma…..



All critical characteristics (Y) are driven by factors (x) which are “upstream” from the results….

Attempting to manage results (Y) only causes increased costs due to rework, test and inspection…

Understanding and controlling the causative factors (x) is the real key to high quality at low cost...

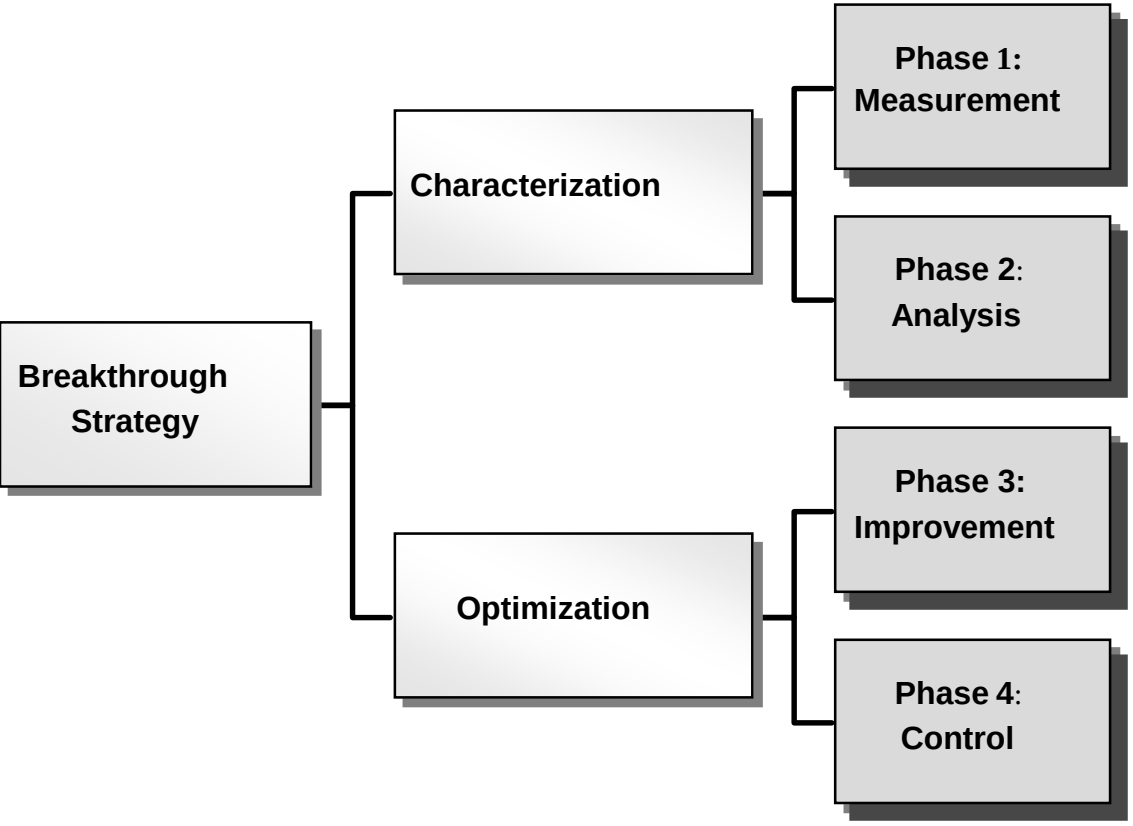
SIX SIGMA COMPARISON

Six Sigma	Traditional
Focus on Prevention	Focus on Firefighting
Low cost/high throughput	High cost/low throughput
Poka Yoke Control Strategies	Reliance on Test and Inspection
Stable/Predictable Processes	Processes based on Random Probability
Proactive	Reactive
Low Failure Rates	High Failure Rates
Focus on Long Term	Focus on Short Term
Efficient	Wasteful
Manage by Metrics and Analysis	Manage by “Seat of the pants”

“SIX SIGMA TAKES US FROM FIXING PRODUCTS SO THEY ARE EXCELLENT, TO FIXING PROCESSES SO THEY PRODUCE EXCELLENT PRODUCTS”

Dr. George Sarney, President, Siebe Control Systems

IMPROVEMENT ROADMAP



Objective

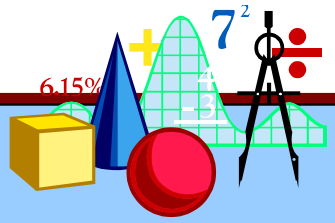
- **Define the problem and verify the primary and secondary measurement systems.**
- **Identify the few factors which are directly influencing the problem.**
- **Determine values for the few contributing factors which resolve the problem.**
- **Determine long term control measures which will ensure that the contributing factors remain controlled.**

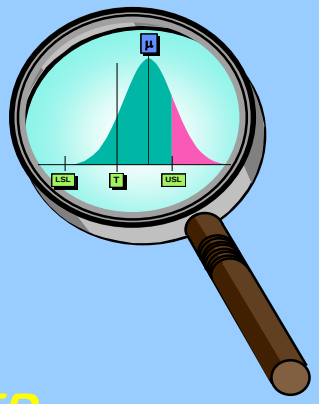
Measurements are critical…

- If we can't accurately measure something, we really don't know much about it.
- If we don't know much about it, we can't control it.
- If we can't control it, we are at the mercy of chance.

WHY STATISTICS?

THE ROLE OF STATISTICS IN SIX SIGMA..





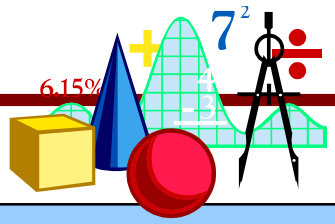
- **WE DON'T KNOW WHAT WE DON'T KNOW**
  - ***IF WE DON'T HAVE DATA, WE DON'T KNOW***
  - ***IF WE DON'T KNOW, WE CAN NOT ACT***
  - ***IF WE CAN NOT ACT, THE RISK IS HIGH***
  - ***IF WE DO KNOW AND ACT, THE RISK IS MANAGED***
  - ***IF WE DO KNOW AND DO NOT ACT, WE DESERVE THE LOSS.***

DR. Mikel J. Harry

- **TO GET DATA WE MUST MEASURE**
- **DATA MUST BE CONVERTED TO INFORMATION**
- **INFORMATION IS DERIVED FROM DATA THROUGH STATISTICS**

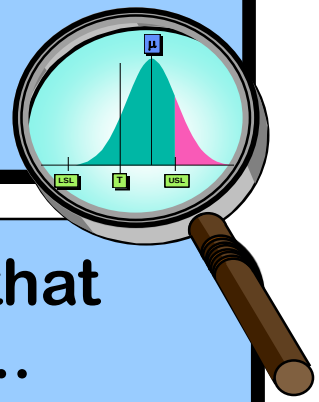
WHY STATISTICS?

THE ROLE OF STATISTICS IN SIX SIGMA..



- Ignorance is not bliss, it is the food of failure and the breeding ground for loss.

DR. Mikel J. Harry



- Years ago a statistician might have claimed that statistics dealt with the processing of data….
- Today's statistician will be more likely to say that statistics is concerned with **decision making in the face of uncertainty**.

Bartlett

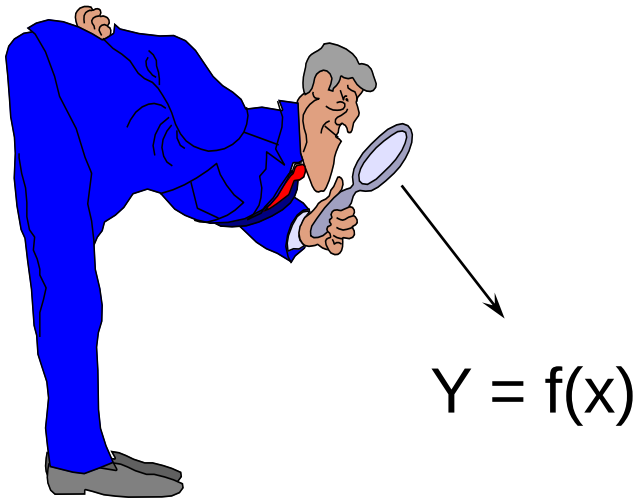
WHAT DOES IT MEAN?

- Sales Receipts
- On Time Delivery
- Process Capacity
- Order Fulfillment Time
- Reduction of Waste
- Product Development Time
- Process Yields
- Scrap Reduction
- Inventory Reduction
- Floor Space Utilization

Random Chance or Certainty….

Which would you choose….?

The Focus of Six Sigma…..



All critical characteristics (Y) are driven by factors (x) which are “downstream” from the results….

Attempting to manage results (Y) only causes increased costs due to rework, test and inspection…

Understanding and controlling the causative factors (x) is the real key to high quality at low cost...

INTRODUCTION TO PROBABILITY DISTRIBUTIONS

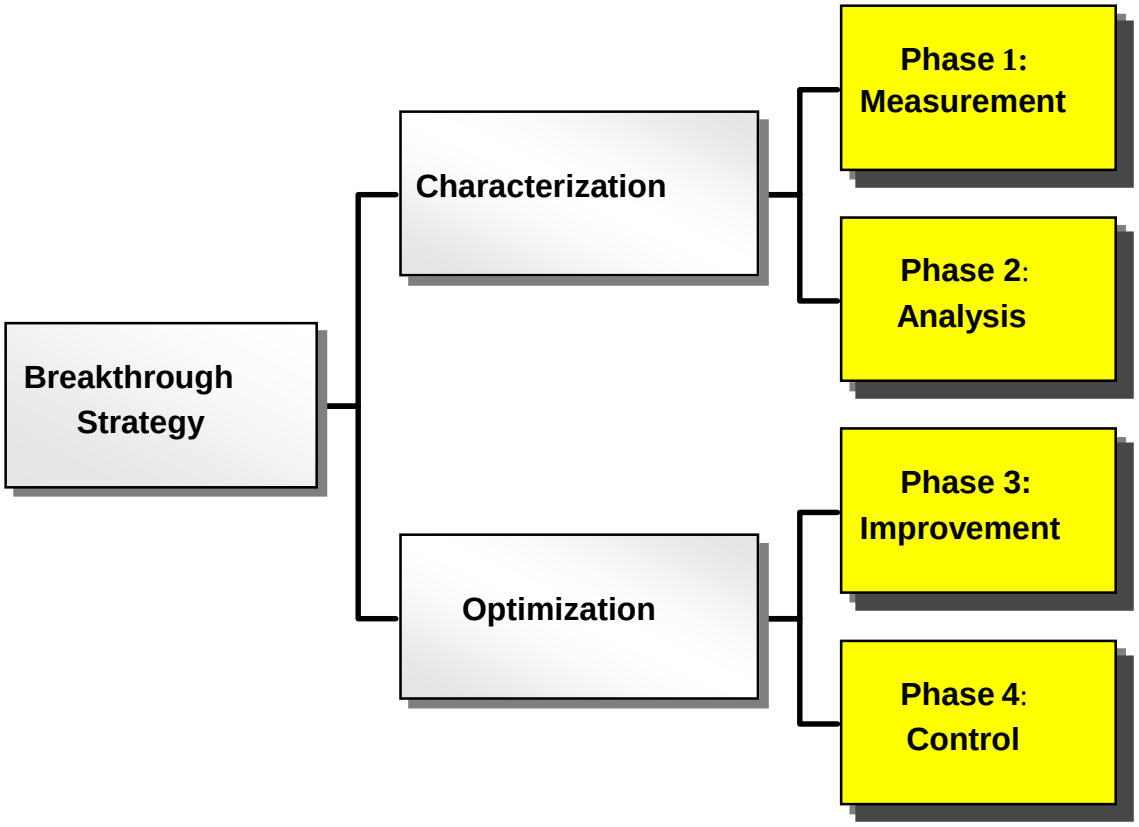
Why do we Care?

An understanding of Probability Distributions is necessary to:

- Understand the concept and use of statistical tools.
- Understand the significance of random variation in everyday measures.
- Understand the impact of significance on the successful resolution of a project.

IMPROVEMENT ROADMAP

Uses of Probability Distributions



Project Uses

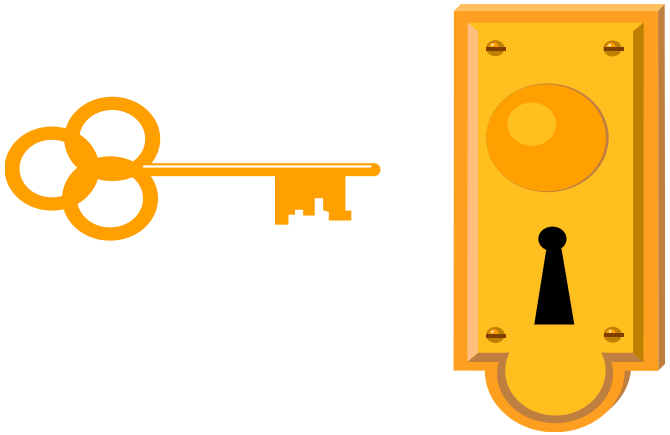
- Establish baseline data characteristics.
- Identify and isolate sources of variation.
- Demonstrate before and after results are not random chance.
- Use the concept of shift & drift to establish project expectations.

KEYS TO SUCCESS

Focus on understanding the concepts

Visualize the concept

Don’t get lost in the math….



Measurements are critical…

- If we can't accurately measure something, we really don't know much about it.
- If we don't know much about it, we can't control it.
- If we can't control it, we are at the mercy of chance.

Types of Measures

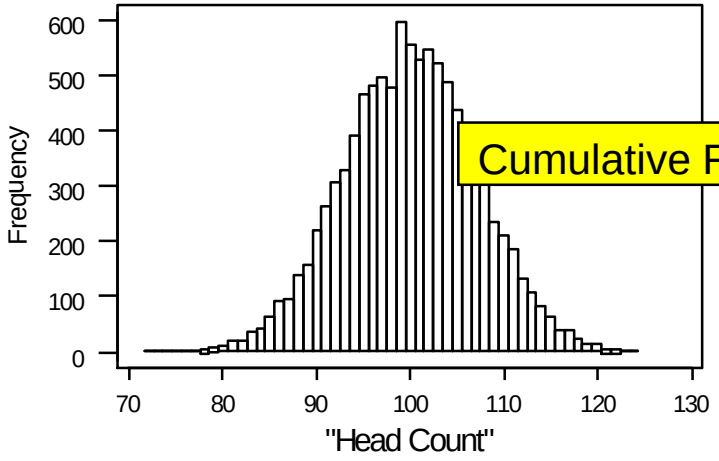
- Measures where the metric is composed of a classification in one of two (or more) categories is called **Attribute data**. This data is usually presented as a "count" or "percent".
  - Good/Bad
  - Yes/No
  - Hit/Miss etc.
- Measures where the metric consists of a number which indicates a precise value is called **Variable data**.
  - Time
  - Miles/Hr

COIN TOSS EXAMPLE

- Take a coin from your pocket and toss it 200 times.
- Keep track of the number of times the coin falls as “heads”.
- When complete, the instructor will ask you for your “head” count.

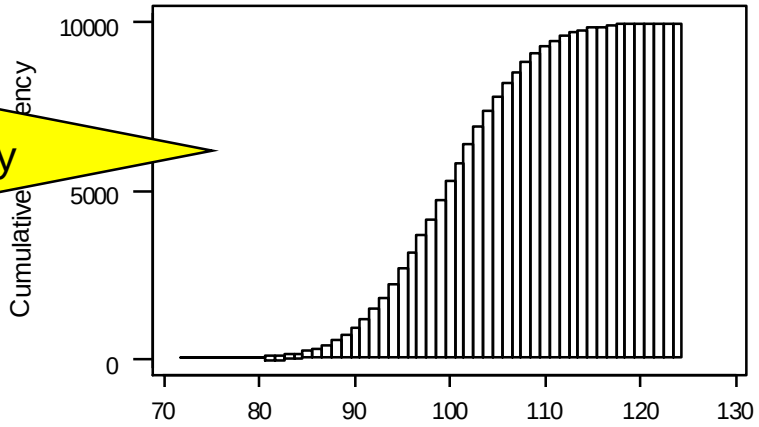
COIN TOSS EXAMPLE

Results from 10,000 people doing a coin toss 200 times.
Count Frequency



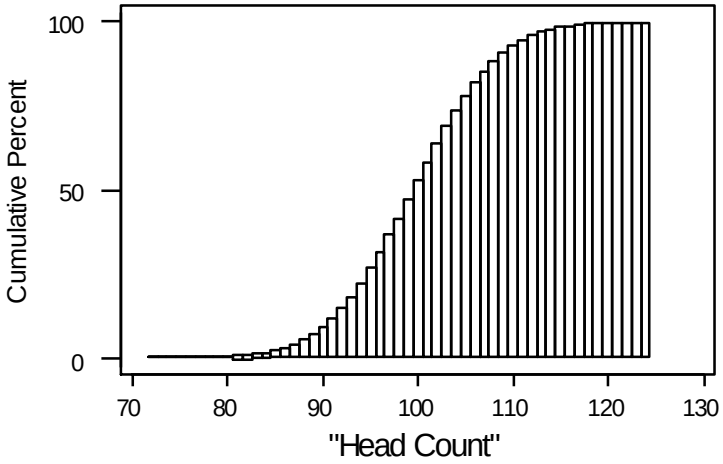
Cumulative Frequency

Results from 10,000 people doing a coin toss 200 times.
Cumulative Count



Cumulative Percent

Results from 10,000 people doing a coin toss 200 times.
Cumulative Percent

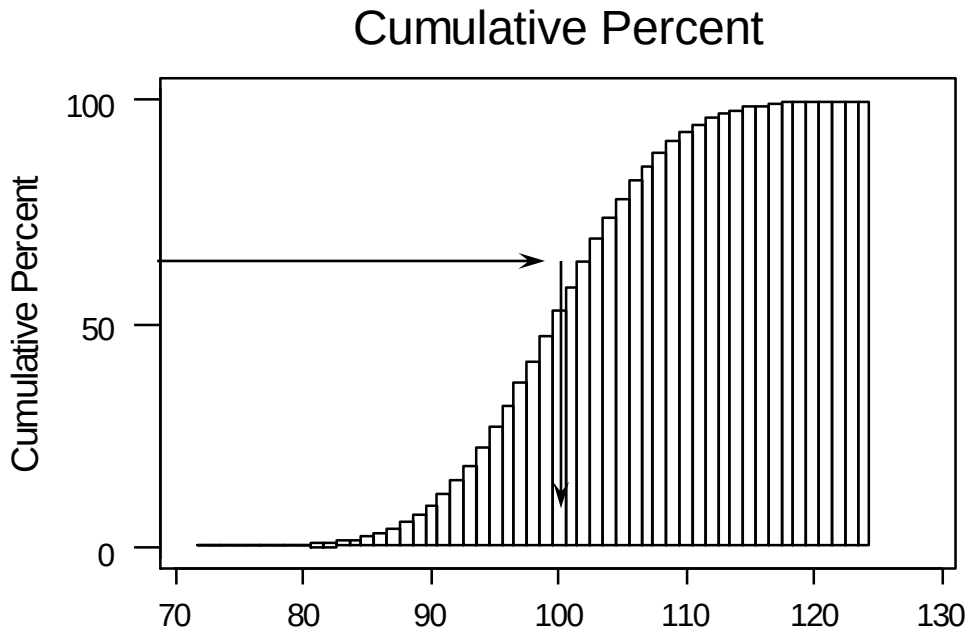


Cumulative count is simply the total frequency count accumulated as you move from left to right until we account for the total population of 10,000 people.

Since we know how many people were in this population (ie 10,000), we can divide each of the cumulative counts by 10,000 to give us a curve with the cumulative percent of population.

COIN TOSS PROBABILITY EXAMPLE

Results from 10,000 people doing a coin toss 200 times



This means that we can now **predict** the change that certain values can occur based on these percentages.

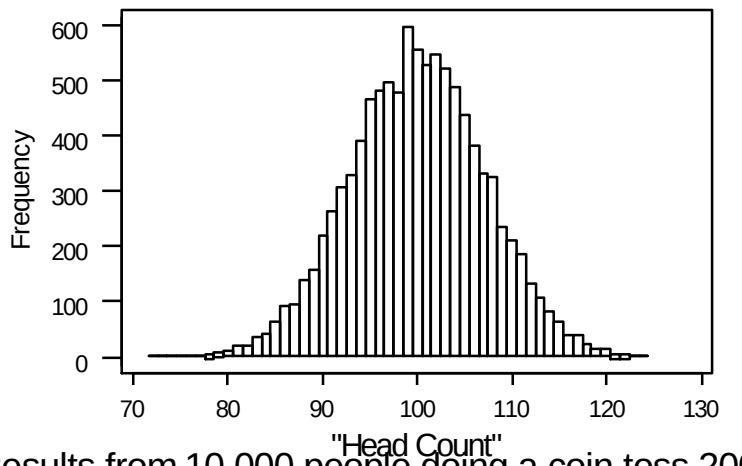
Note here that 50% of the values are less than our expected value of 100.

This means that in a future experiment set up the same way, we would expect 50% of the values to be less than 100.

COIN TOSS EXAMPLE

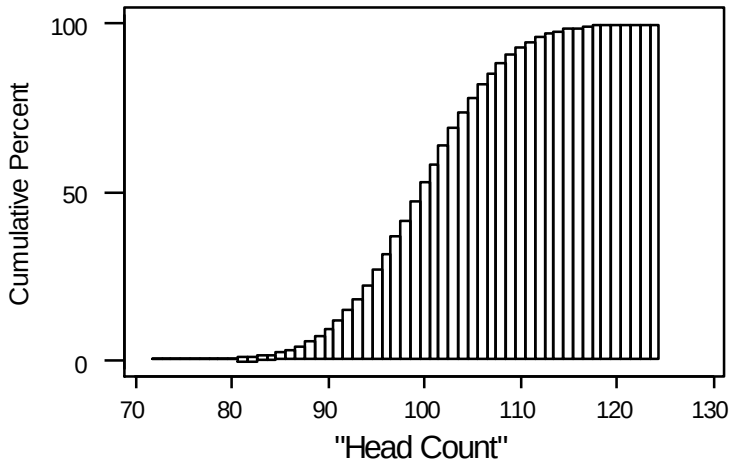
Results from 10,000 people doing a coin toss 200 times.

Count Frequency



Results from 10,000 people doing a coin toss 200 times.

Cumulative Percent



We can now equate a probability to the occurrence of specific values or groups of values.

For example, we can see that the occurrence of a “Head count” of less than 74 or greater than 124 out of 200 tosses is so rare that a single occurrence was not registered out of 10,000 tries.

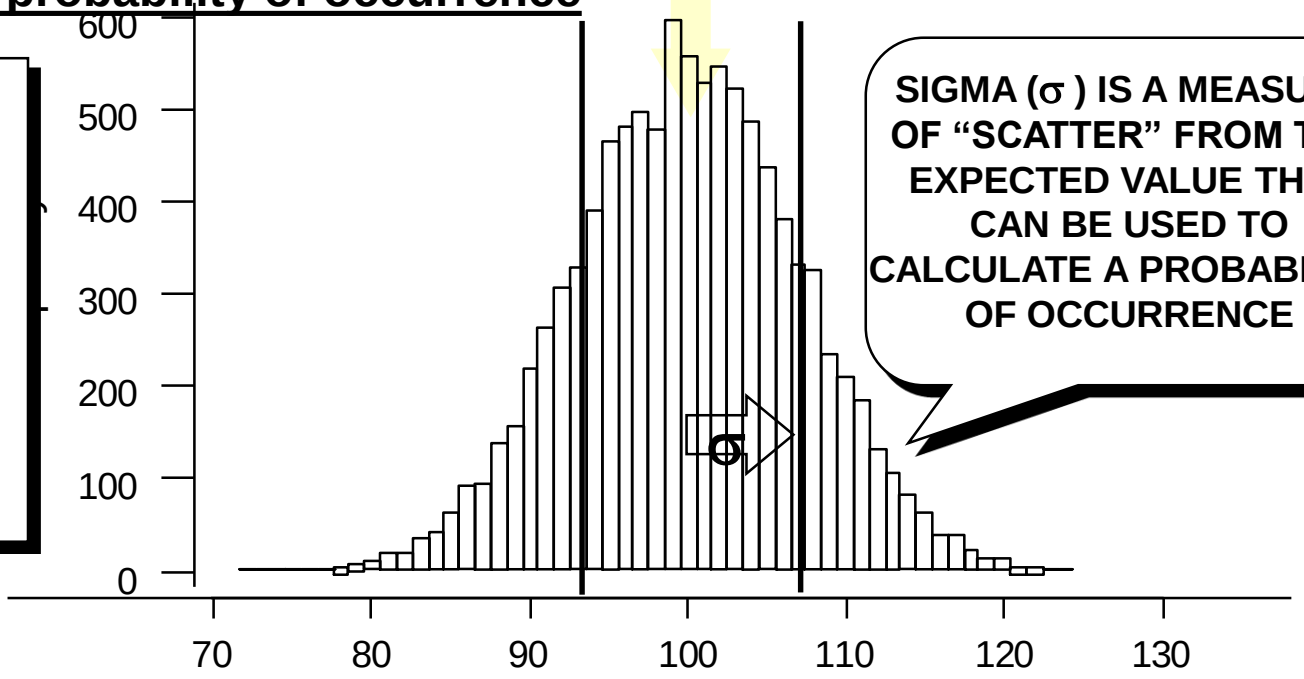
On the other hand, we can see that the chance of getting a count near (or at) 100 is much higher. With the data that we now have, we can actually predict each of these values.

COIN TOSS PROBABILITY DISTRIBUTION

PROCESS CENTERED ON EXPECTED VALUE

% of population = probability of occurrence

If we know where we are in the population we can equate that to a probability value. This is the purpose of the sigma value (normal data).



SIGMA (σ) IS A MEASURE OF "SCATTER" FROM THE EXPECTED VALUE THAT CAN BE USED TO CALCULATE A PROBABILITY OF OCCURRENCE

NUMBER OF HEADS	58	65	72	79	86	93	100	107	114	121	128	135	142
SIGMA VALUE (Z)	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
CUM % OF POPULATION			.003	.135	2.275	15.87	50.0	84.1	97.7	99.86	99.997		

WHAT DOES IT MEAN?

- Common Occurrence
- Rare Event

What are the chances that this "just happened" If they are small, chances are that an external influence is at work that can be used to our benefit….

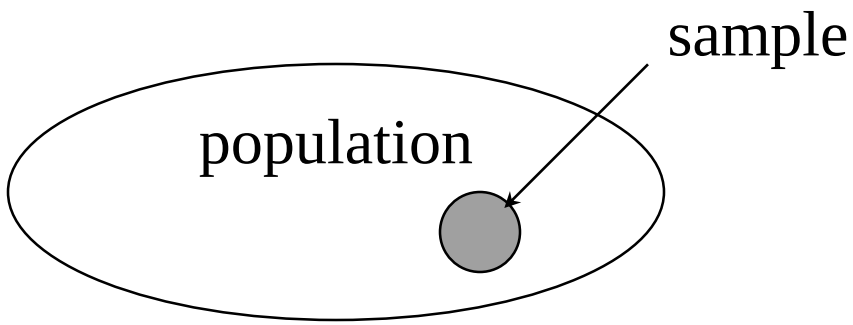
Probability and Statistics

- "the odds of Colorado University winning the national title are 3 to 1"
- "Drew Bledsoe's pass completion percentage for the last 6 games is .58% versus .78% for the first 5 games"
- "The Senator will win the election with 54% of the popular vote with a margin of +/- 3%"

- Probability and Statistics influence our lives daily
- Statistics is the universal lanuage for science
- Statistics is the art of collecting, classifying, presenting, interpreting and analyzing numerical data, as well as making conclusions about the system from which the data was obtained.

Population Vs. Sample (Certainty Vs. Uncertainty)

↖ A **sample** is just a **subset of all possible values**

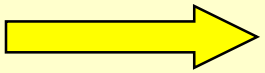


↖ Since the sample does not contain all the possible values, there is some uncertainty about the population. **Hence any statistics, such as mean and standard deviation, are just *estimates* of the true population parameters.**

Descriptive Statistics

Descriptive Statistics is the branch of statistics which most people are familiar. It characterizes and summarizes the most prominent features of a given set of data (means, medians, standard deviations, percentiles, graphs, tables and charts.

Descriptive Statistics describe the elements of a population as a whole or to describe data that represent just a sample of elements from the entire population

Inferential Statistics 

Inferential Statistics

Inferential Statistics is the branch of statistics that deals with drawing conclusions about a population based on information obtained from a sample drawn from that population.

While descriptive statistics has been taught for centuries, inferential statistics is a relatively new phenomenon having its roots in the 20th century.

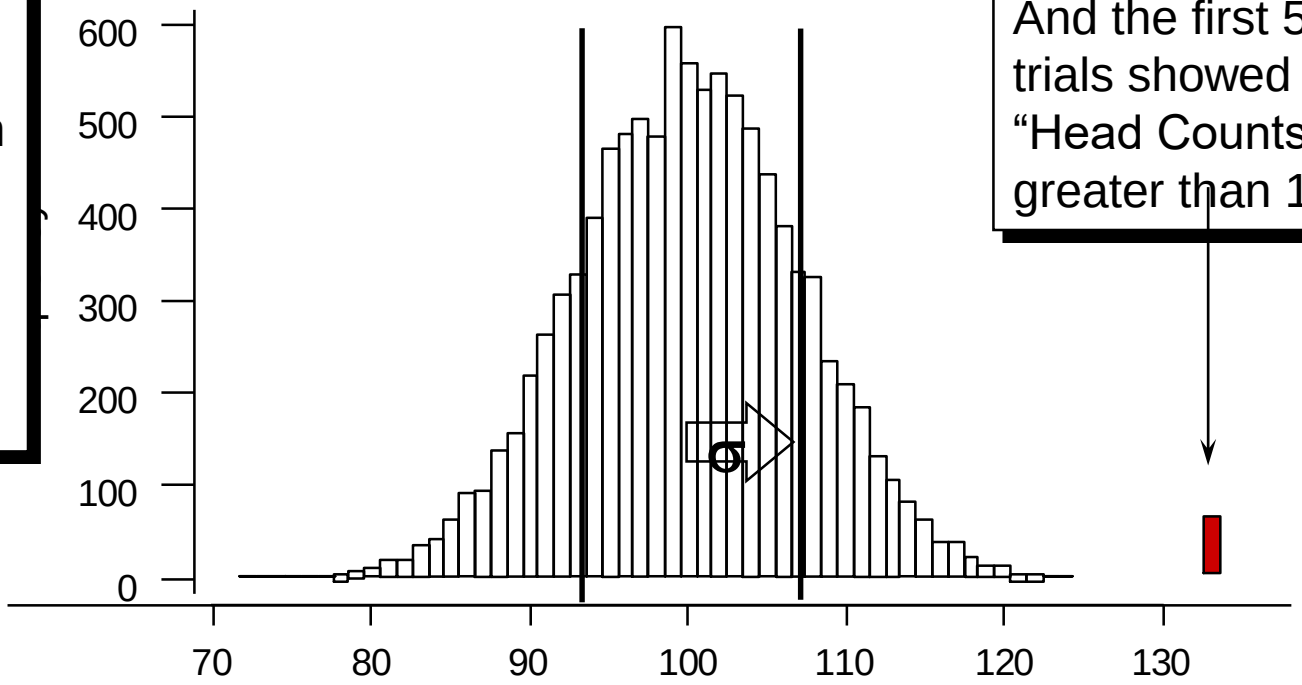
We “infer” something about a population when only information from a sample is known.

Probability is the link between Descriptive and Inferential Statistics

WHAT DOES IT MEAN?

WHAT IF WE MADE A CHANGE TO THE PROCESS?

Chances are very good that the process distribution has changed. In fact, there is a probability greater than 99.999% that it has changed.



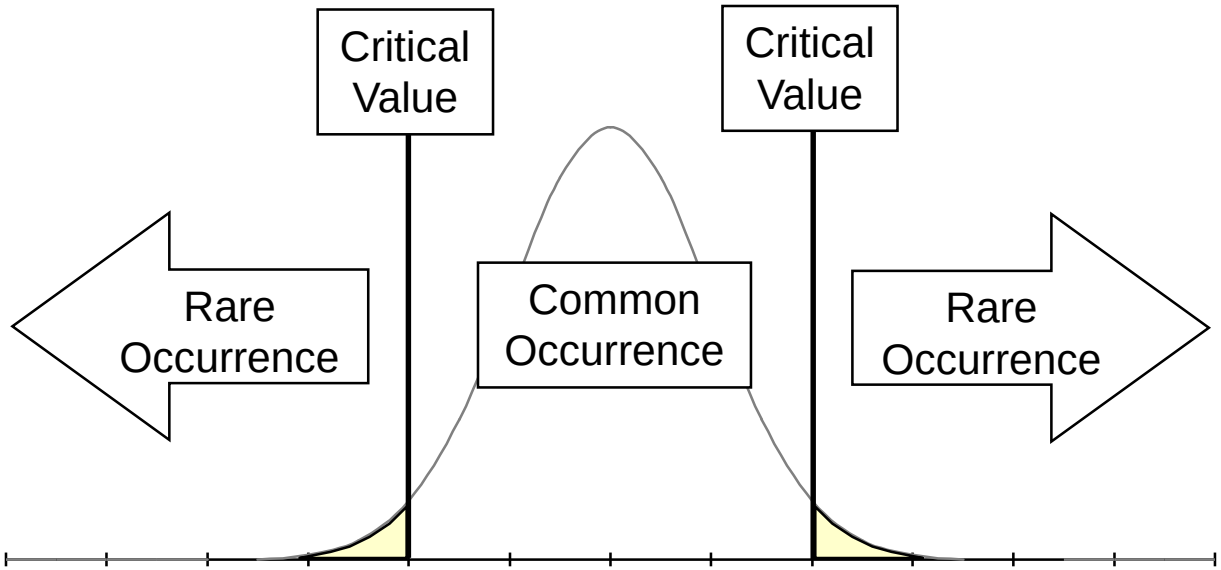
And the first 50 trials showed "Head Counts" greater than 130?

NUMBER OF HEADS	58	65	72	79	86	93	100	107	114	121	128	135	142
SIGMA VALUE (Z)	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
CUM % OF POPULATION			.003	.135	2.275	15.87	50.0	84.1	97.7	99.86	99.997		

USES OF PROBABILITY DISTRIBUTIONS

Primarily these distributions are used to test for significant differences in data sets.

To be classified as significant, the actual measured value must exceed a critical value. The critical value is tabular value determined by the probability distribution and the risk of error. This risk of error is called α risk and indicates the probability of this value occurring naturally. So, an α risk of .05 (5%) means that this critical value will be exceeded by a random occurrence less than 5% of the time.



SO WHAT MAKES A DISTRIBUTION UNIQUE?

CENTRAL TENDENCY

Where a population is located.

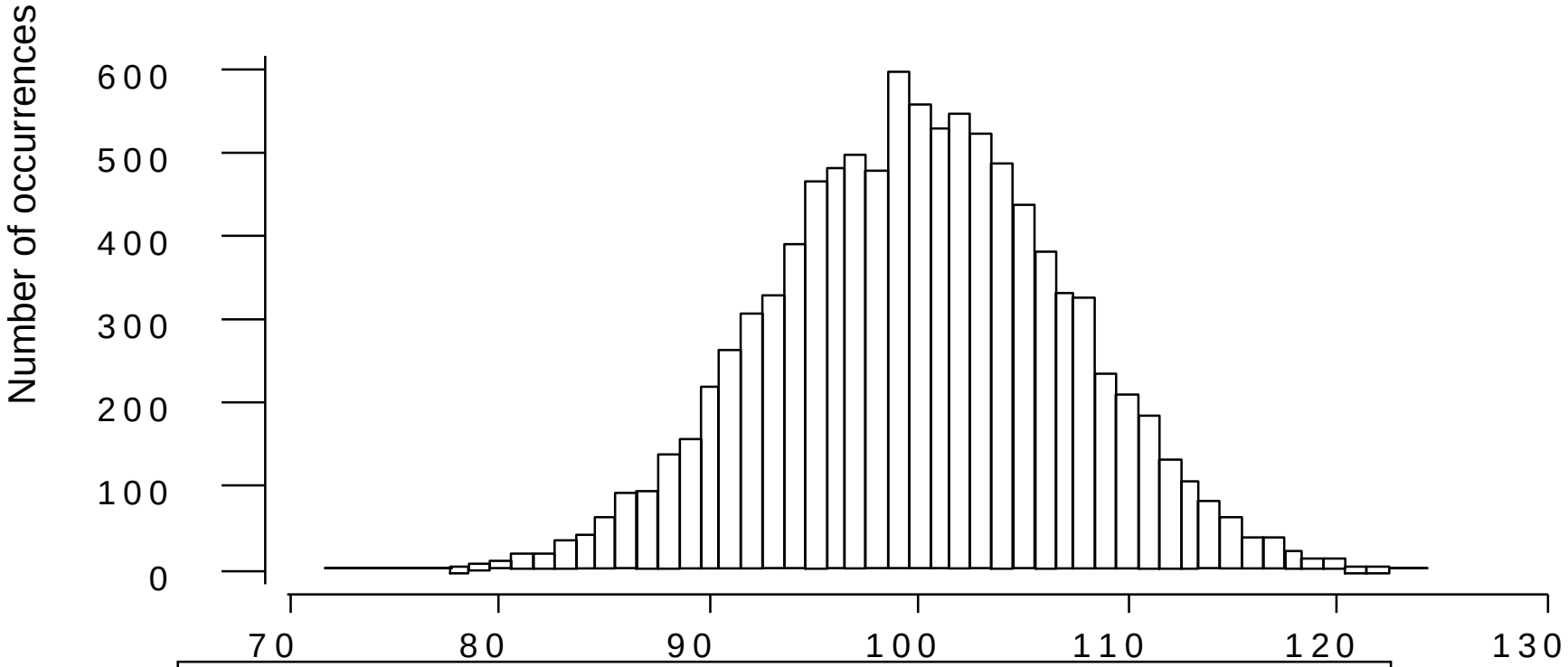
DISPERSION

How wide a population is spread.

DISTRIBUTION FUNCTION

The mathematical formula that best describes the data (we will cover this in detail in the next module).

COIN TOSS CENTRAL TENDENCY



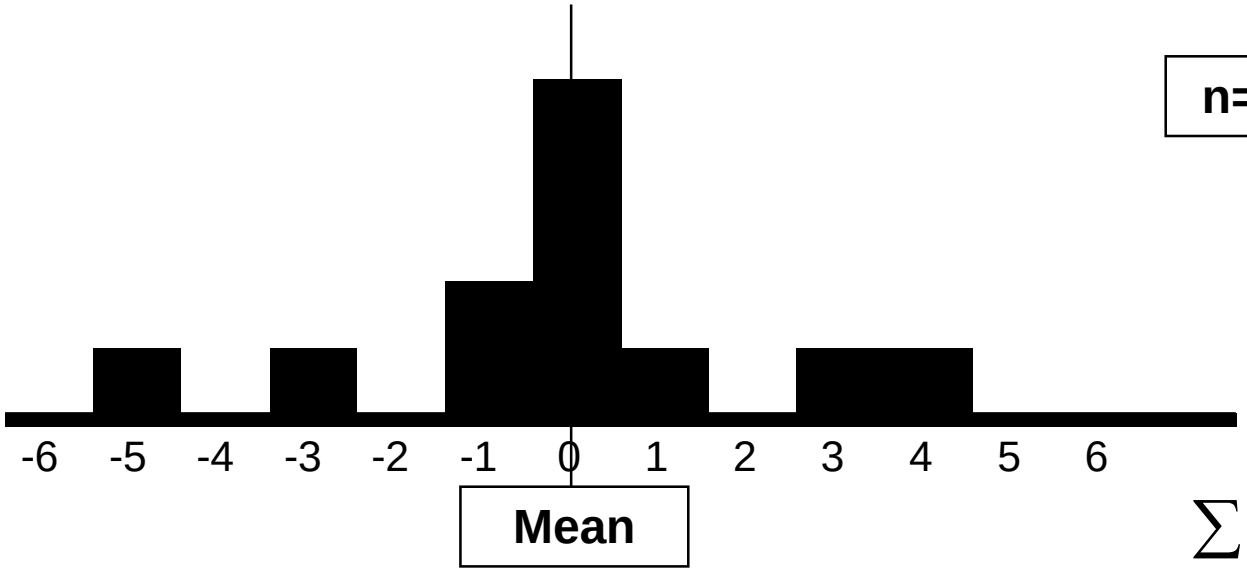
What are some of the ways that we can easily indicate the centering characteristic of the population?

Three measures have historically been used; the **mean**, the **median** and the **mode.**

WHAT IS THE MEAN?

The **mean** has already been used in several earlier modules and is the most common measure of central tendency for a population. The mean is simply the **average value** of the data.

$$mean = \bar{x} = \frac{\sum x_i}{n} = \frac{-2}{12} = -.17$$



ORDERED DATA SET

-5
-3
-1
-1
0
0
0
0
0
1
3
4

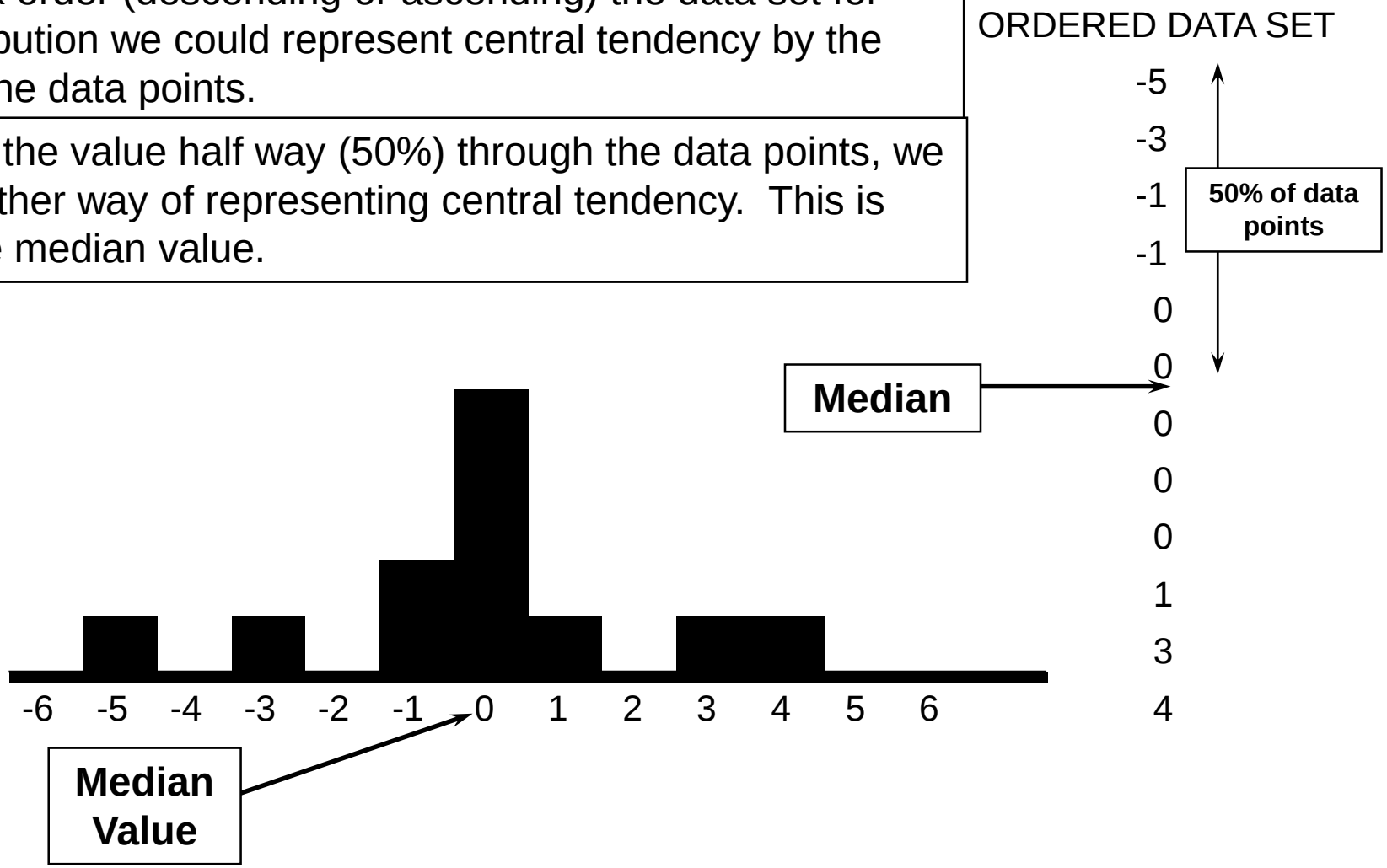
n=12

$$\sum x_i = -2$$

WHAT IS THE MEDIAN?

If we rank order (descending or ascending) the data set for this distribution we could represent central tendency by the order of the data points.

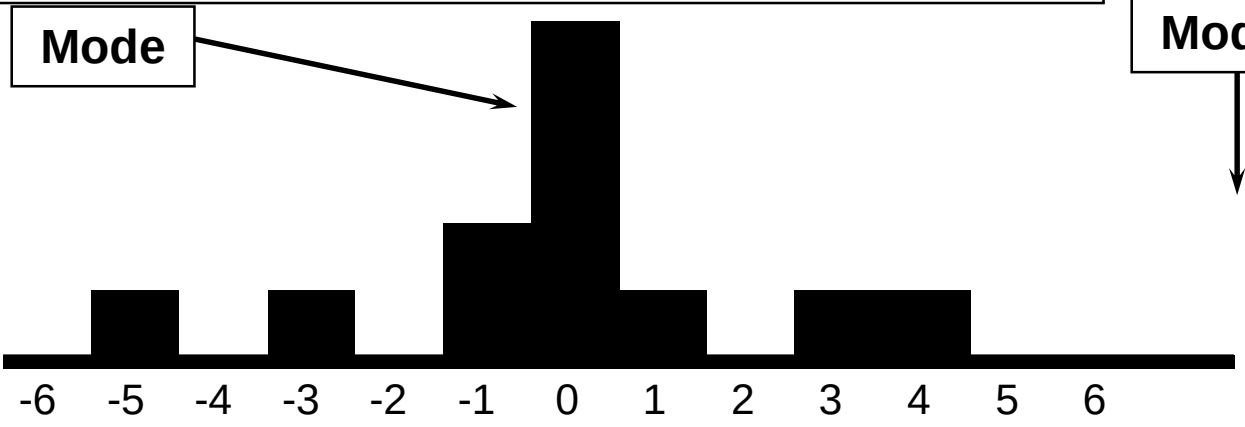
If we find the value half way (50%) through the data points, we have another way of representing central tendency. This is called the median value.



WHAT IS THE MODE?

If we rank order (descending or ascending) the data set for this distribution we find several ways we can represent central tendency.

We find that a single value occurs more often than any other. Since we know that there is a higher chance of this occurrence in the middle of the distribution, we can use this feature as an indicator of central tendency. This is called the **mode.**



ORDERED DATA SET

-5
-3
-1
-1
0
0
0
0
0
1
3
4

Mode

MEASURES OF CENTRAL TENDENCY, SUMMARY

MEAN ($\overline{X}$)

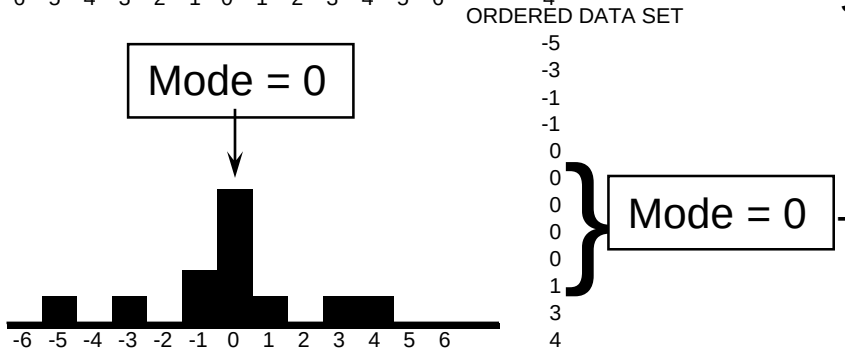
(Otherwise known as the average)

$$\overline{X} = \frac{\sum X_i}{n} = \frac{-2}{12} = .17$$

MEDIAN

(50 percentile data point)

Here the median value falls between two zero values and therefore is zero. If the values were say 2 and 3 instead, the median would be 2.5.



MODE

(Most common value in the data set)

The mode in this case is 0 with 5 occurrences within this data.

SO WHAT'S THE REAL DIFFERENCE?

MEAN

The mean is the most consistently accurate measure of central tendency, but is more difficult to calculate than the other measures.

MEDIAN AND MODE

The median and mode are both very easy to determine. That's the good news….The bad news is that both are more susceptible to bias than the mean.

SO WHAT'S THE BOTTOM LINE?

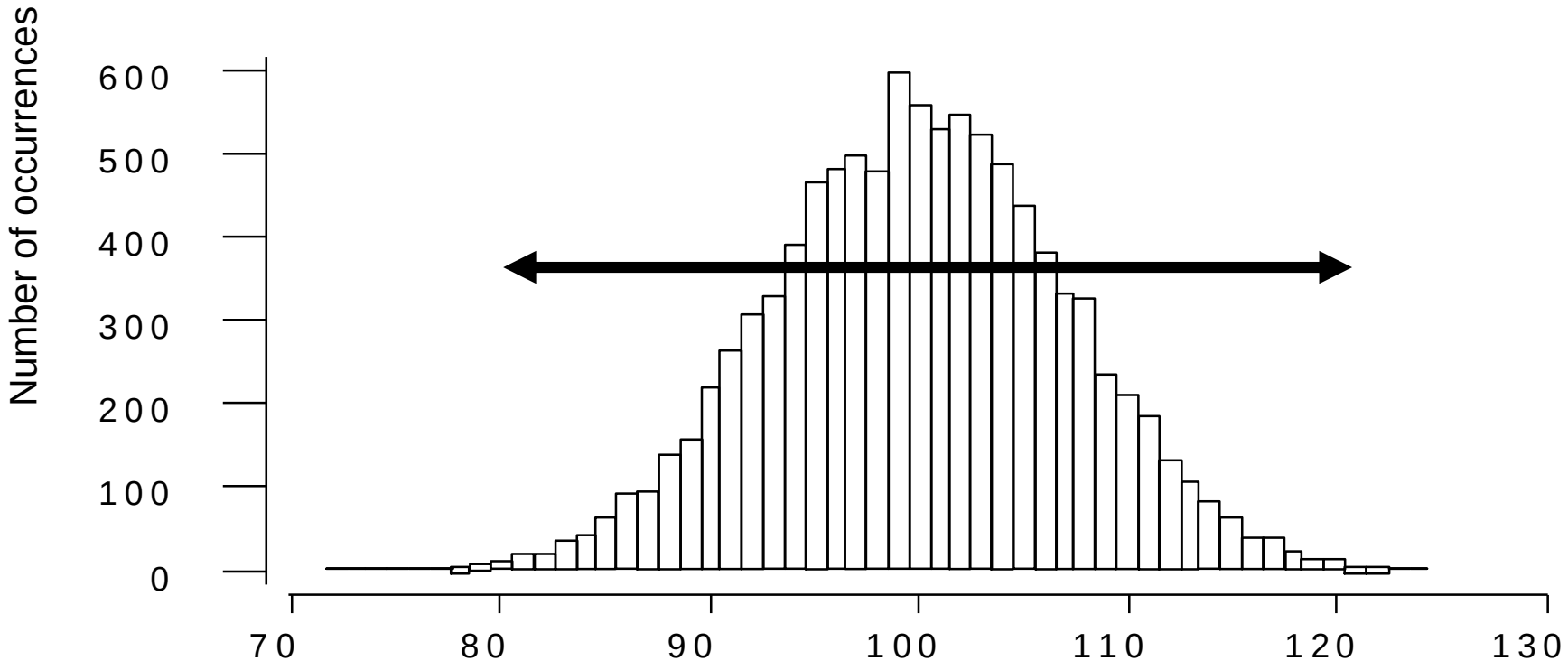
MEAN

Use on all occasions unless a circumstance prohibits its use.

MEDIAN AND MODE

Only use if you cannot use mean.

COIN TOSS POPULATION DISPERSION



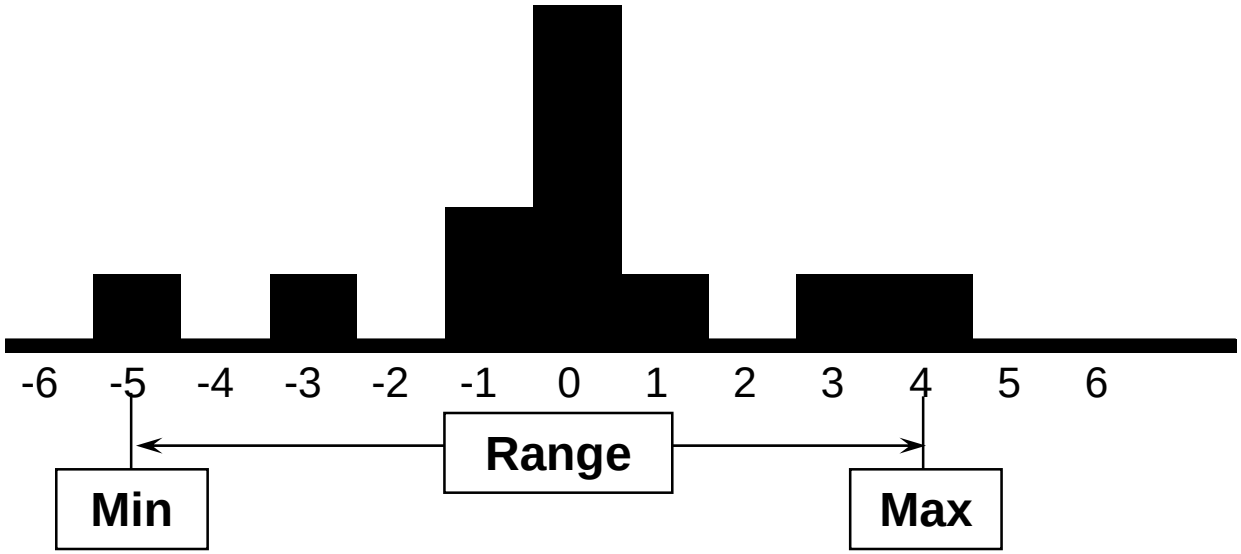
What are some of the ways that we can easily indicate the dispersion (spread) characteristic of the population?

Three measures have historically been used; the **range**, the **standard deviation** and the **variance.**

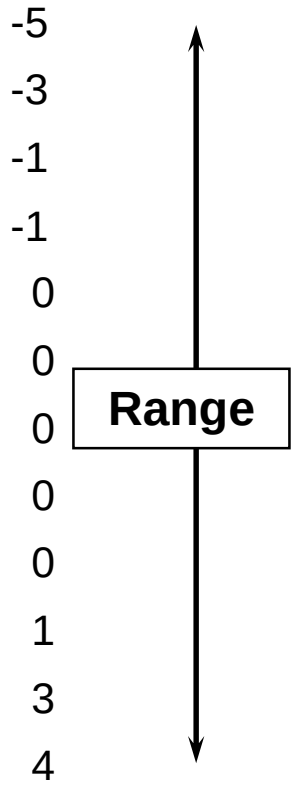
WHAT IS THE RANGE?

The **range** is a very common metric which is easily determined from any ordered sample. To calculate the range simply subtract the minimum value in the sample from the maximum value.

$$Range = x_{MAX} - x_{MIN} = 4 - (-5) = 9$$



ORDERED DATA SET

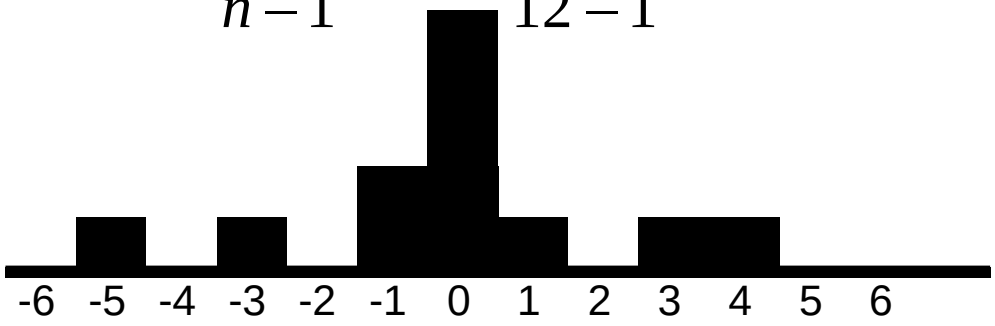


WHAT IS THE VARIANCE/STANDARD DEVIATION?

The **variance (s^2)** is a very robust metric which requires a fair amount of work to determine. The **standard deviation(s)** is the square root of the variance and is the most commonly used measure of dispersion for larger sample sizes.

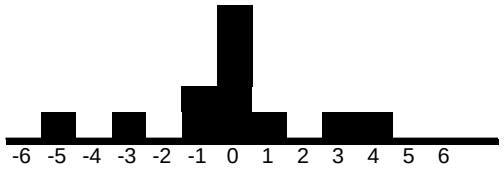
$$\overline{X} = \frac{\sum X_i}{n} = \frac{-2}{12} = -.17$$

$$s^2 = \frac{\sum \left(X_i - \overline{X}\right)^2}{n-1} = \frac{61.67}{12-1} = 5.6$$



DATA SET	$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$
-5	-5-(-.17)=-4.83	(-4.83)2=23.32
-3	-3-(-.17)=-2.83	(-2.83)2=8.01
-1	-1-(-.17)=-.83	(-.83)2=.69
-1	-1-(-.17)=-.83	(-.83)2=.69
0	0-(-.17)=.17	(.17)2=.03
0	0-(-.17)=.17	(.17)2=.03
0	0-(-.17)=.17	(.17)2=.03
0	0-(-.17)=.17	(.17)2=.03
1	0-(-.17)=.17	(.17)2=.03
3	1-(-.17)=1.17	(1.17)2=1.37
4	3-(-.17)=3.17	(3.17)2=10.05
	4-(-.17)=4.17	(4.17)2=17.39
		61.67

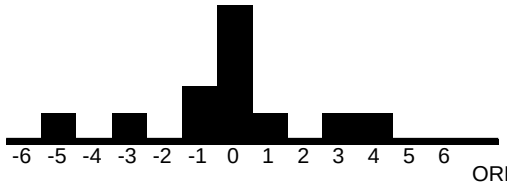
MEASURES OF DISPERSION



ORDERED DATA SET

-5 ← Min=-5
-3
-1
-1
0
0
0
0
0
1
3
4 ← Max=4

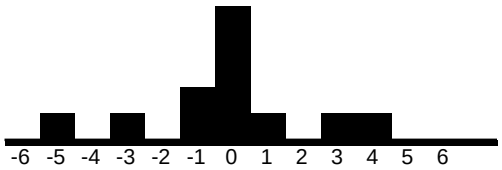
$\overline{X} = \frac{\sum X_i}{n} = \frac{-2}{12} = -.17$



DATA SET	$X_i - \overline{X}$	$(X_i - \overline{X})^2$
-5	-5-(-.17)=-4.83	(-4.83)2=23.32
-3	-3-(-.17)=-2.83	(-2.83)2=8.01
-1	-1-(-.17)=-.83	(-.83)2=.69
-1	-1-(-.17)=-.83	(-.83)2=.69
0	0-(-.17)=.17	(.17)2=.03
0	0-(-.17)=.17	(.17)2=.03
0	0-(-.17)=.17	(.17)2=.03
0	0-(-.17)=.17	(.17)2=.03
1	0-(-.17)=.17	(.17)2=.03
3	1-(-.17)=1.17	(1.17)2=1.37
4	3-(-.17)=3.17	(3.17)2=10.05
	4-(-.17)=4.17	(4.17)2=17.39
		61.67

ORDERED DATA SET

-5
-3
-1
-1
0
0
0
0
0
1
3
4



RANGE (R)

(The maximum data value minus the minimum)

$$R = X_{\max} - X_{\min} = 4 - (-6) = 10$$

VARIANCE (s^2)

(Squared deviations around the center point)

$$s^2 = \frac{\sum (X_i - \overline{X})^2}{n-1} = \frac{61.67}{12-1} = 5.6$$

STANDARD DEVIATION (s)

(Absolute deviation around the center point)

$$s = \sqrt{s^2} = \sqrt{5.6} = 2.37$$

SAMPLE MEAN AND VARIANCE EXAMPLE

$$\hat{\mu} = \overline{X} = \frac{\sum X_i}{N}$$

$$\hat{\sigma}^2 = s^2 = \sum \frac{\left(X_i - \overline{X}\right)^2}{n-1}$$

	X_i	$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$
1	10		
2	15		
3	12		
4	14		
5	10		
6	9		
7	11		
8	12		
9	10		
10	12		
Σ			
$\overline{X}$			
s^2			

SO WHAT'S THE REAL DIFFERENCE?

VARIANCE/ STANDARD DEVIATION

The standard deviation is the most consistently accurate measure of central tendency for a single population. The variance has the added benefit of being additive over multiple populations. Both are difficult and time consuming to calculate.

RANGE

The range is very easy to determine. That's the good news….The bad news is that it is very susceptible to bias.

SO WHAT'S THE BOTTOM LINE?

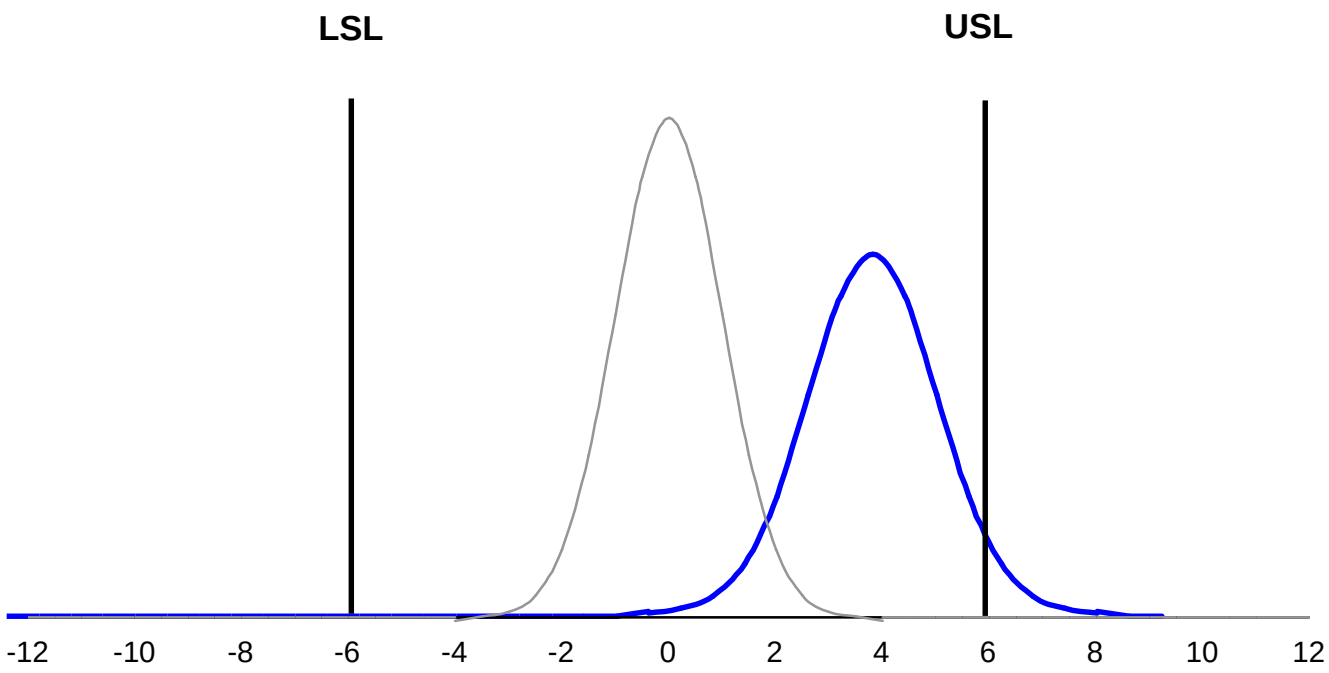
VARIANCE/ STANDARD DEVIATION

Best used when you have enough samples (>10).

RANGE

Good for small samples (10 or less).

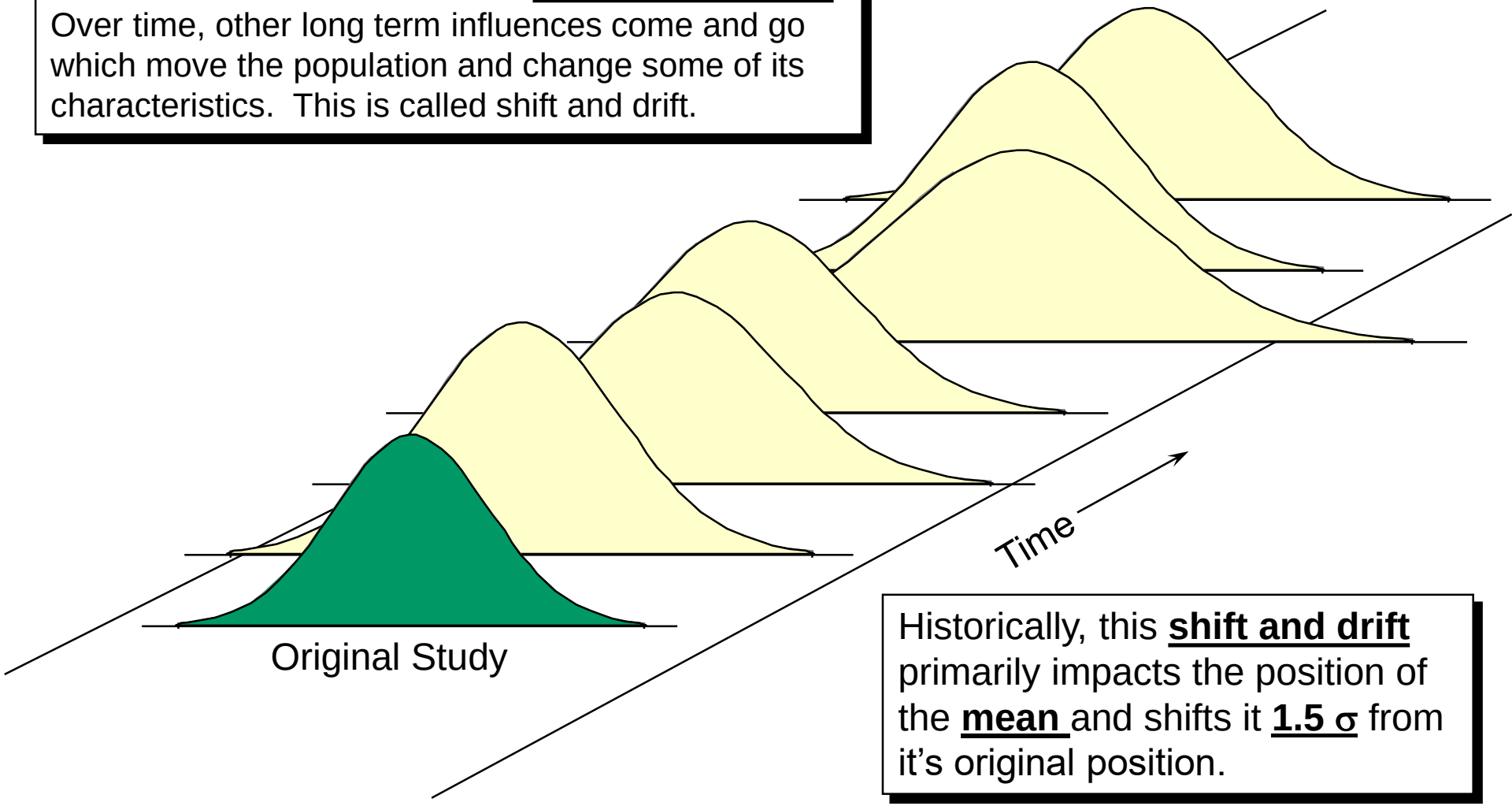
SO WHAT IS THIS SHIFT & DRIFT STUFF...



The project is progressing well and you wrap it up. 6 months later you are surprised to find that the population has taken a shift.

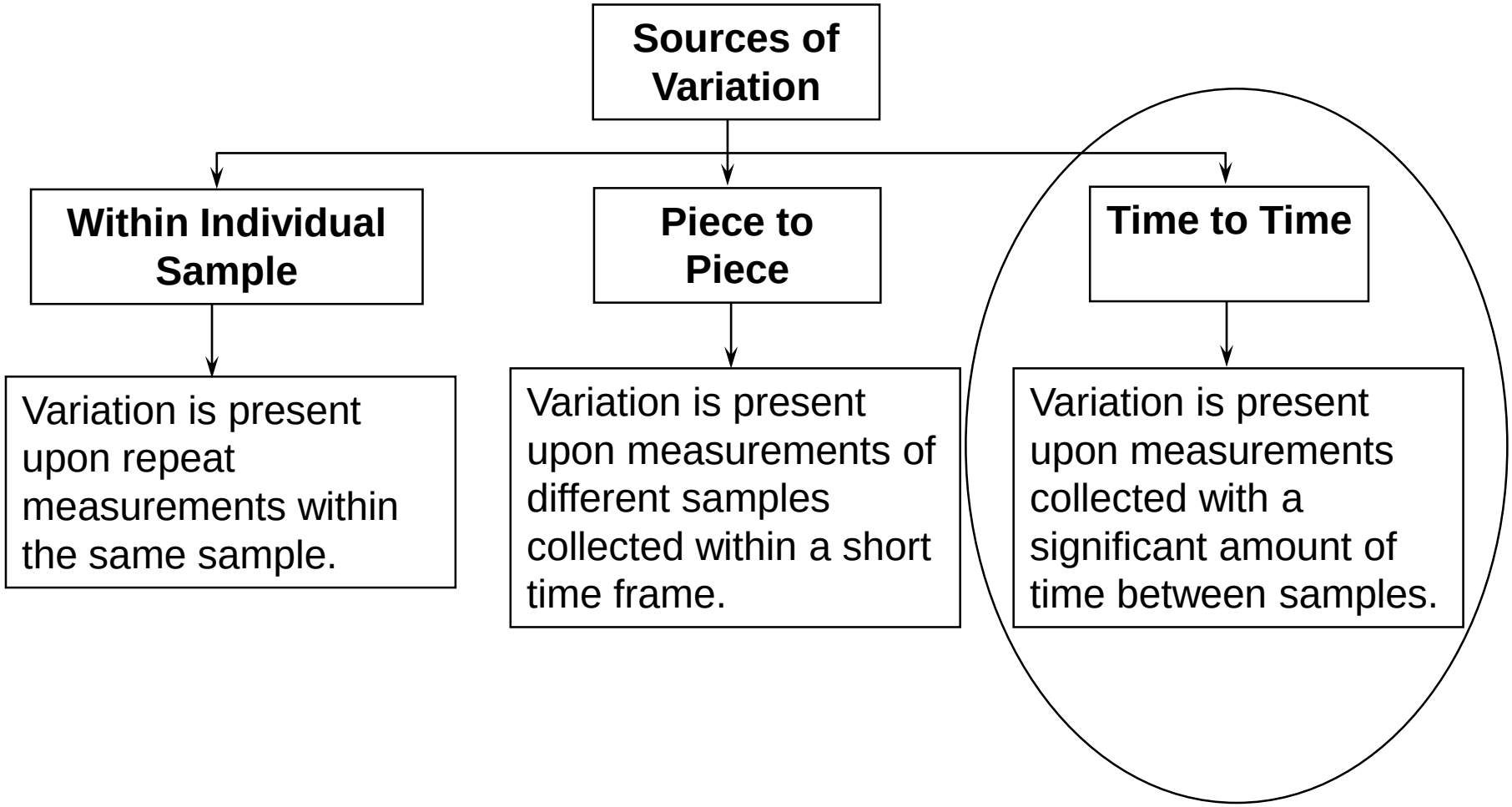
SO WHAT HAPPENED?

All of our work was focused in a **narrow time frame**. Over time, other long term influences come and go which move the population and change some of its characteristics. This is called shift and drift.



Historically, this **shift and drift** primarily impacts the position of the **mean** and shifts it **1.5 σ** from it's original position.

VARIATION FAMILIES



SO WHAT DOES IT MEAN?

To compensate for these long term variations, we must consider two sets of metrics. Short term metrics are those which typically are associated with our work. Long term metrics take the short term metric data and degrade it by an average of 1.5σ.

IMPACT OF 1.5σ SHIFT AND DRIFT

Z	PPM $_{ST}$	C_{pk}	PPM $_{LT}$ (+1.5 σ)
0.0	500,000	0.0	933,193
0.1	460,172	0.0	919,243
0.2	420,740	0.1	903,199
0.3	382,089	0.1	884,930
0.4	344,578	0.1	864,334
0.5	308,538	0.2	841,345
0.6	274,253	0.2	815,940
0.7	241,964	0.2	788,145
0.8	211,855	0.3	758,036
0.9	184,060	0.3	725,747
1.0	158,655	0.3	691,462
1.1	135,666	0.4	655,422
1.2	115,070	0.4	617,911
1.3	96,801	0.4	579,260
1.4	80,757	0.5	539,828
1.5	66,807	0.5	500,000
1.6	54,799	0.5	460,172
1.7	44,565	0.6	420,740

Here, you can see that the impact of this concept is potentially very significant. In the short term, we have driven the defect rate down to 54,800 ppm and can expect to see occasional long term ppm to be as bad as 460,000 ppm.

SHIFT AND DRIFT EXERCISE

We have just completed a project and have presented the following short term metrics:

- Z_{st}=3.5
- PPM_{st}=233
- Cpk_{st}=1.2

Calculate the long term values for each of these metrics.

COMMON PROBABILITY DISTRIBUTIONS AND THEIR USES

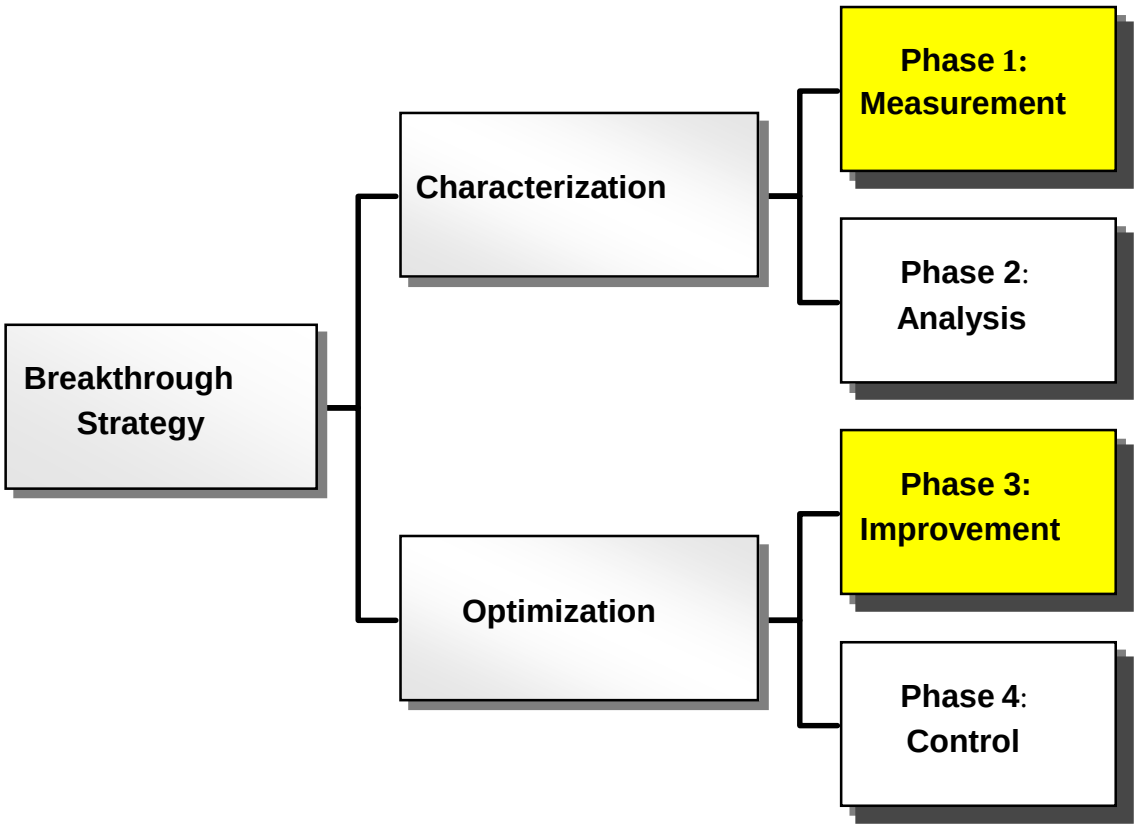
Why do we Care?

Probability distributions are necessary to:

- determine whether an event is significant or due to random chance.
- predict the probability of specific performance given historical characteristics.

IMPROVEMENT ROADMAP

Uses of Probability Distributions



Common Uses

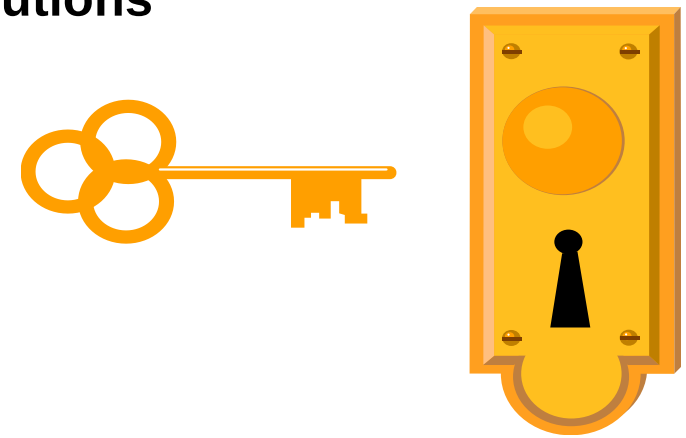
- Baselining Processes
- Verifying Improvements

KEYS TO SUCCESS

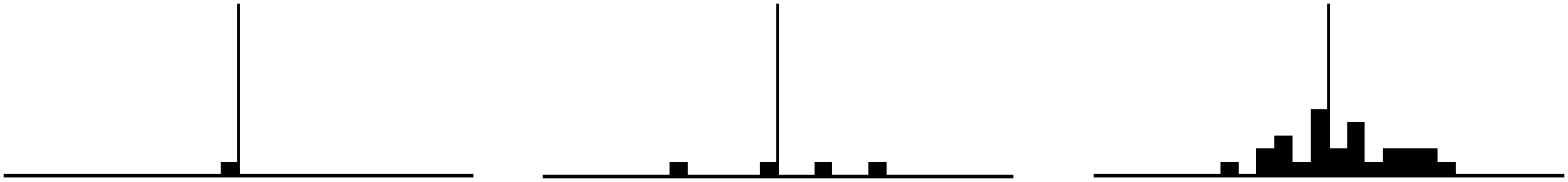
Focus on understanding the use of the distributions

Practice with examples wherever possible

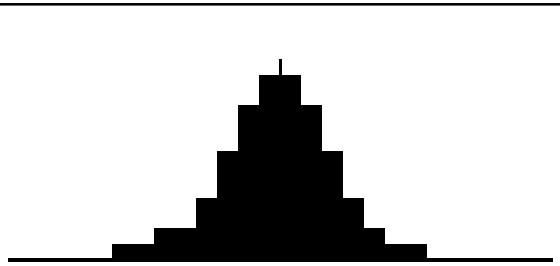
Focus on the use and context of the tool

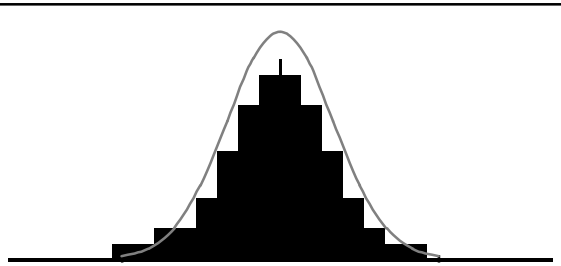


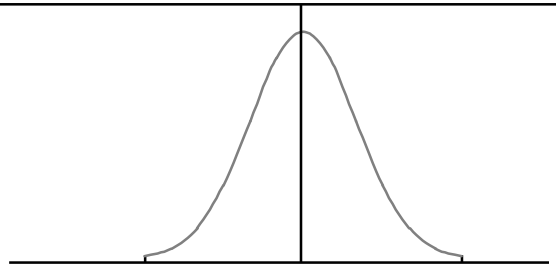
PROBABILITY DISTRIBUTIONS, WHERE DO THEY COME FROM?



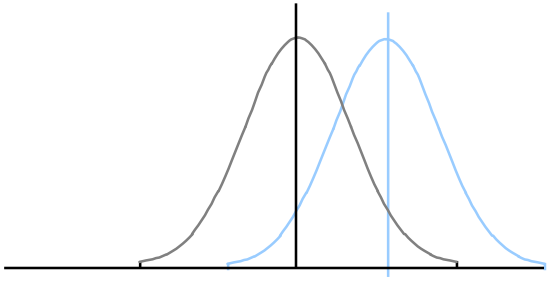
Data points vary, but as the data accumulates, it forms a distribution which occurs naturally.



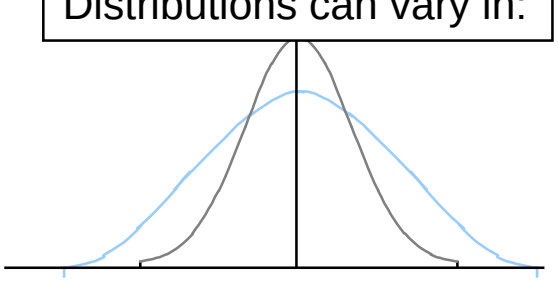




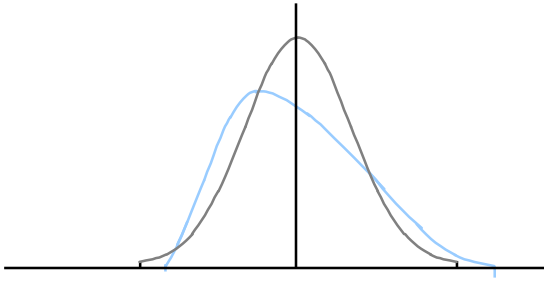
Distributions can vary in:



Location

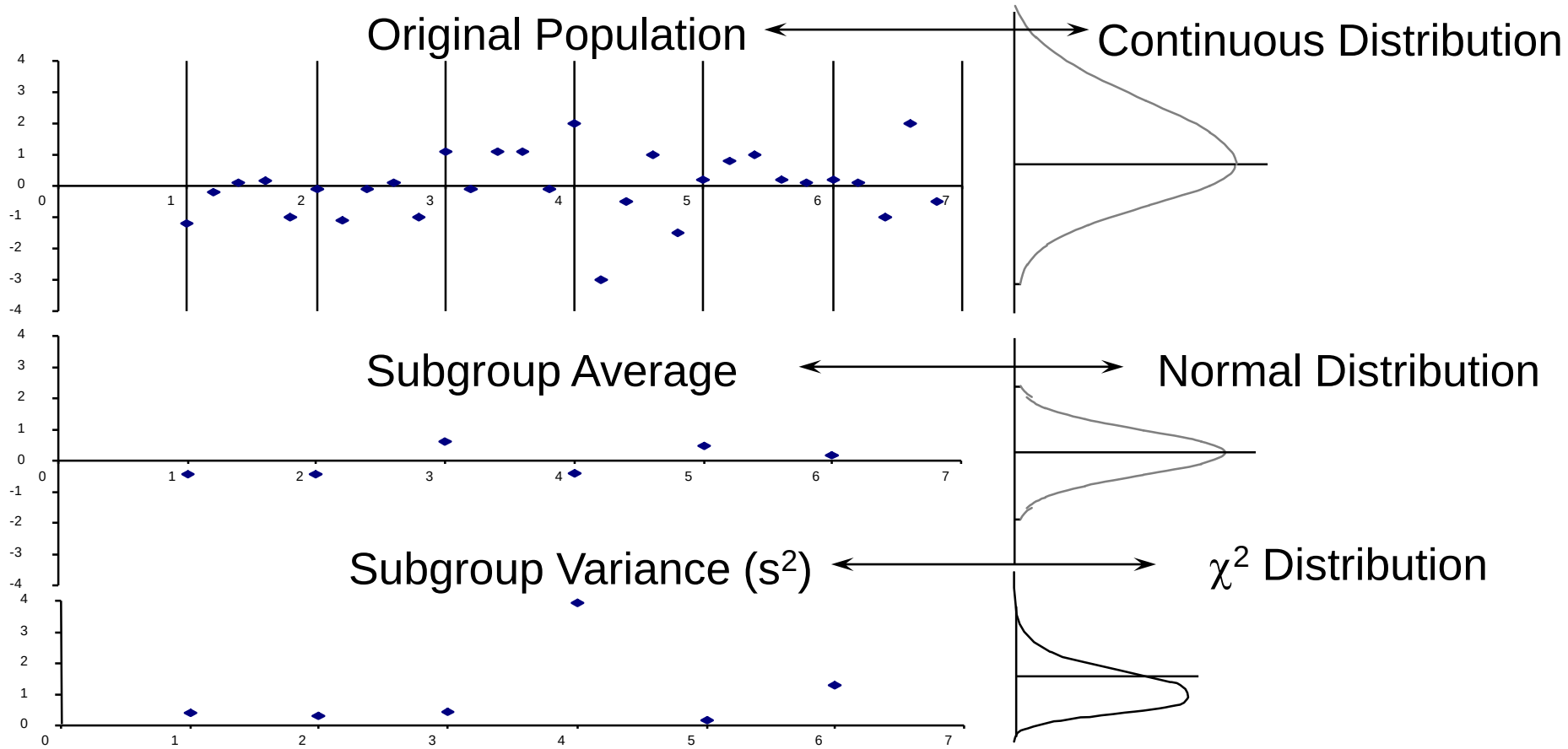


Spread



Shape

COMMON PROBABILITY DISTRIBUTIONS



THE LANGUAGE OF MATH

Symbol	Name	Statistic Meaning	Common Uses
α	Alpha	Significance level	Hypothesis Testing, DOE
χ^2	Chi Square	Probability Distribution	Confidence Intervals, Contingency Tables, Hypothesis Testing
Σ	Sum	Sum of Individual values	Variance Calculations
t	t, Student t	Probability Distribution	Hypothesis Testing, Confidence Interval of the Mean
n	Sample Size	Total size of the Sample Taken	Nearly all Functions
ν	Nu	Degree of Freedom	Probability Distributions, Hypothesis Testing, DOE
β	Beta	Beta Risk	Sample Size Determination
δ	Delta	Difference between population means	Sample Size Determination
Z	Sigma Value	Number of Standard Deviations a value Exists from the Mean	Probability Distributions, Process Capability, Sample Size Determinations

Population and Sample Symbology

Value	Population	Sample
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard Deviation	σ	s
Process Capability	Cp	$\hat{Cp}$
Binomial Mean	$\bar{P}$	$\hat{P}$

THREE PROBABILITY DISTRIBUTIONS

$$t_{CALC} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$Significant = t_{CALC} \geq t_{CRIT}$$

$$F_{calc} = \frac{s_1^2}{s_2^2}$$

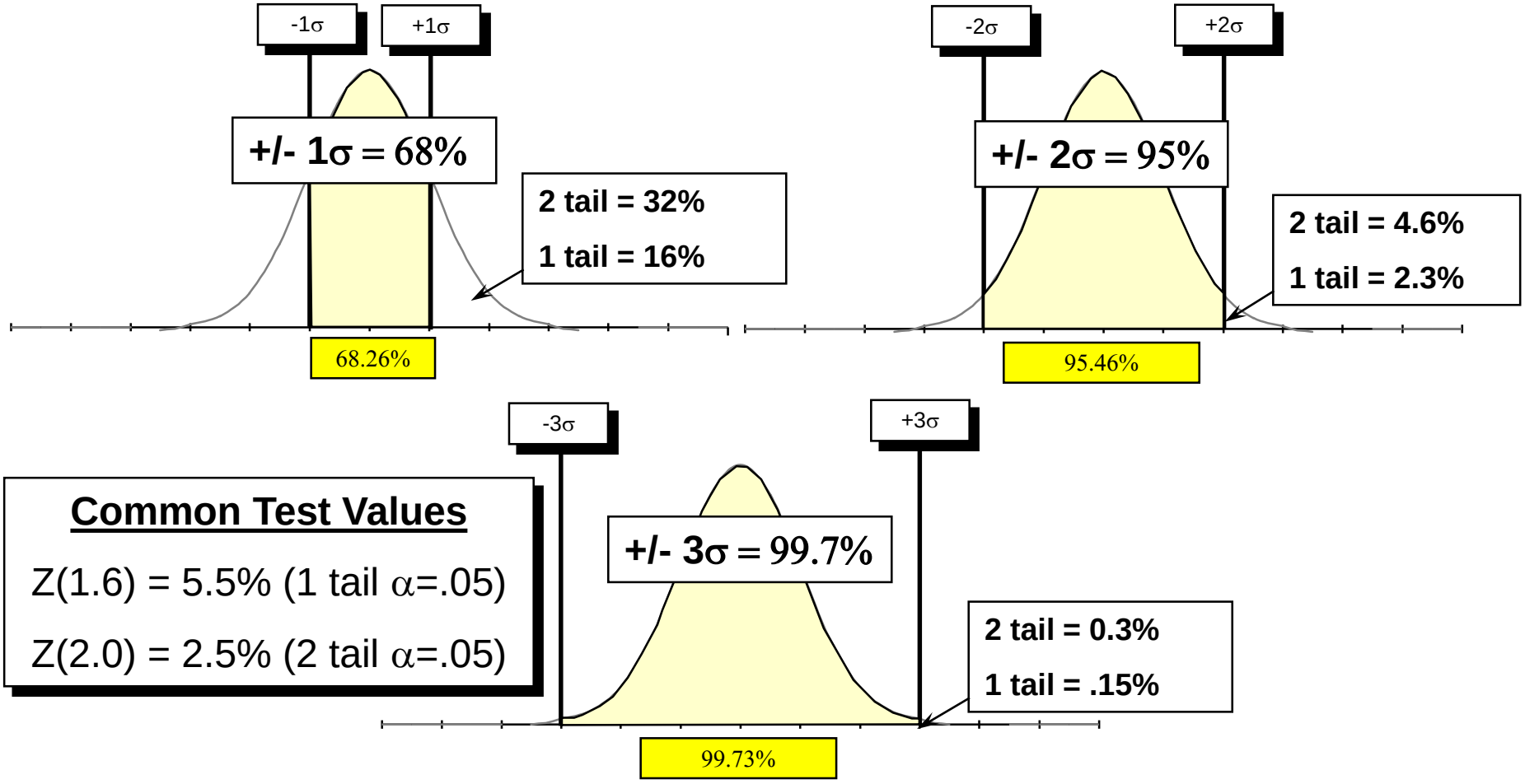
$$Significant = F_{CALC} \geq F_{CRIT}$$

$$\chi^2_{\alpha,df} = \frac{\left(f_e - f_a\right)^2}{f_e}$$

$$Significant = \chi^2_{CALC} \geq \chi^2_{CRIT}$$

Note that in each case, a limit has been established to determine what is random chance verses significant difference. This point is called the critical value. If the calculated value exceeds this critical value, there is very low probability (P<.05) that this is due to random chance.

Z TRANSFORM

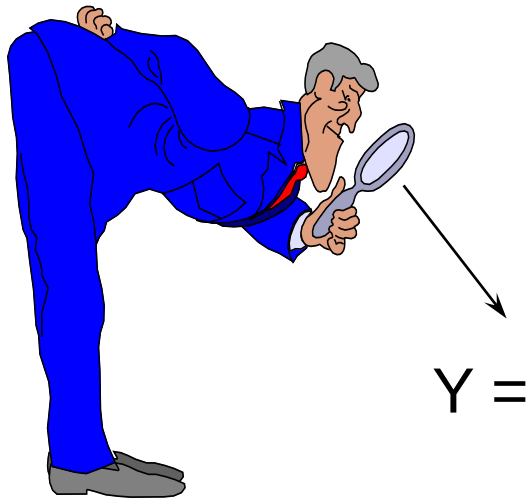


Common Test Values

Z(1.6) = 5.5% (1 tail α=.05)

Z(2.0) = 2.5% (2 tail α=.05)

The Focus of Six Sigma…..



Y = f(x)

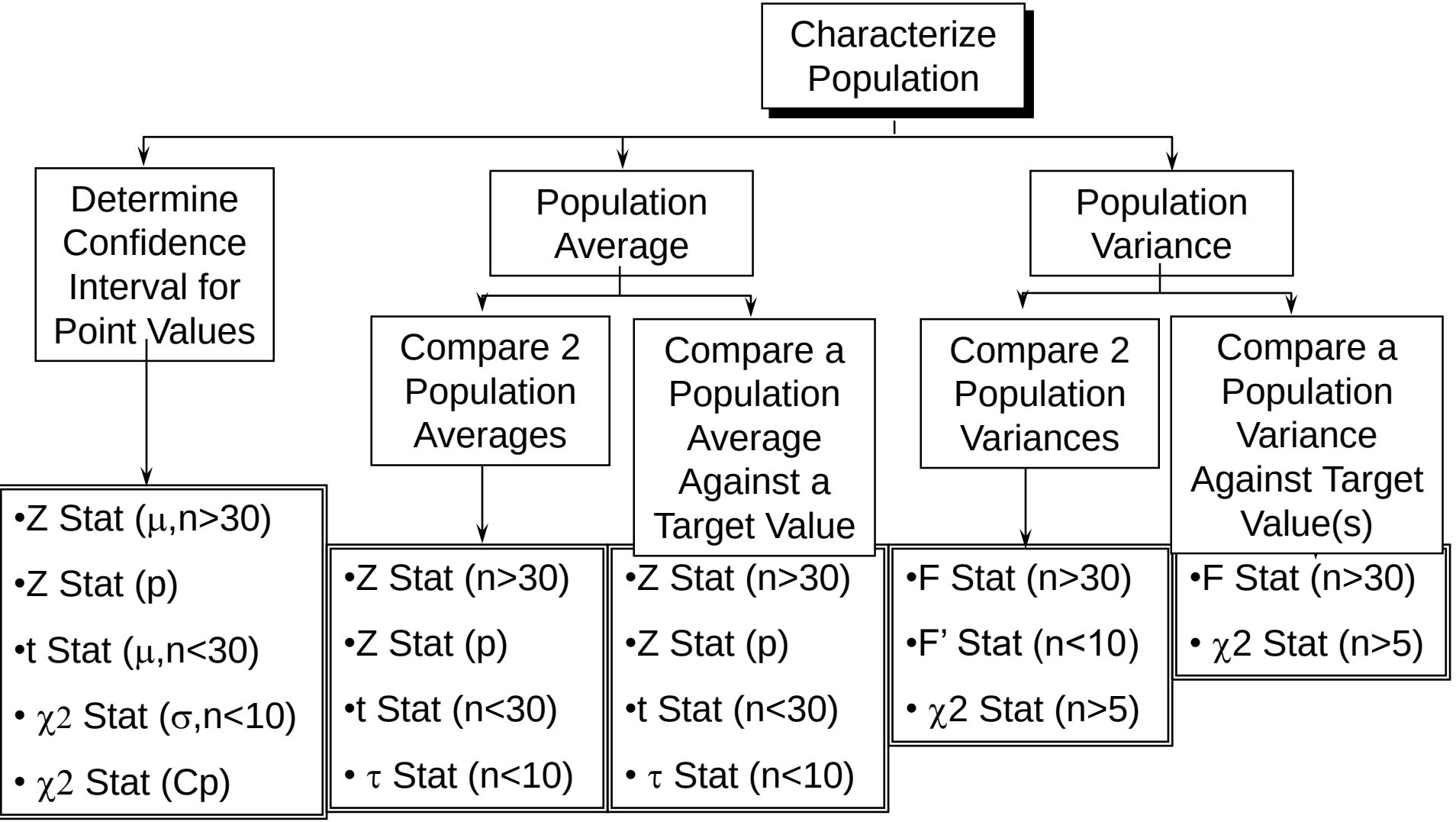
Probability distributions identify sources of causative factors (x). These can be identified and verified by testing which shows their **significant effects against the backdrop of random noise.**

All critical characteristics (Y) are driven by factors (x) which are “downstream” from the results….

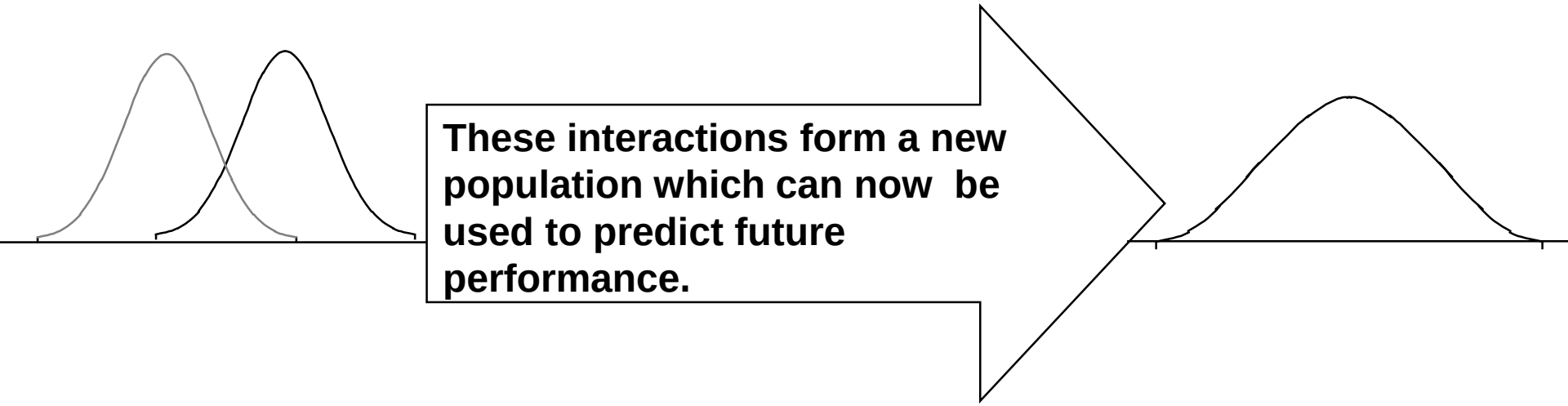
Attempting to manage results (Y) only causes increased costs due to rework, test and inspection…

Understanding and controlling the causative factors (x) is the real key to high quality at low cost...

BUT WHAT DISTRIBUTION SHOULD I USE?

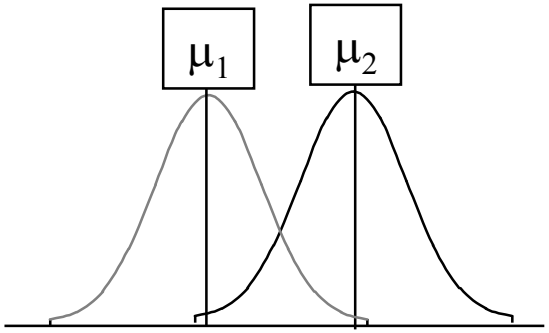


HOW DO POPULATIONS INTERACT?



HOW DO POPULATIONS INTERACT?

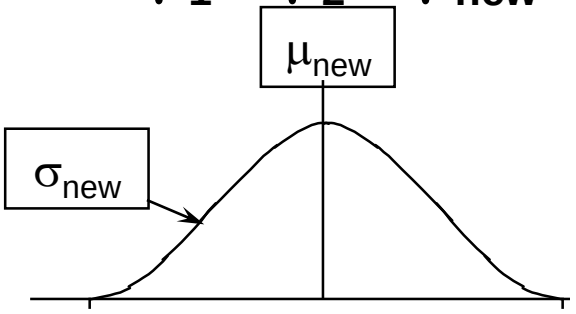
ADDING TWO POPULATIONS

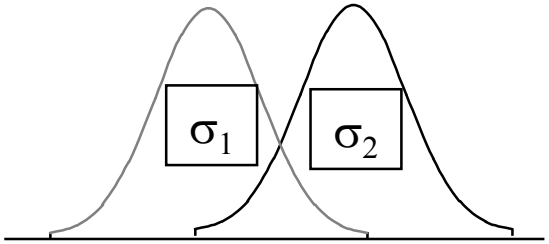


Population means interact in a simple intuitive manner.

Means Add

$$\mu_1 + \mu_2 = \mu_{new}$$





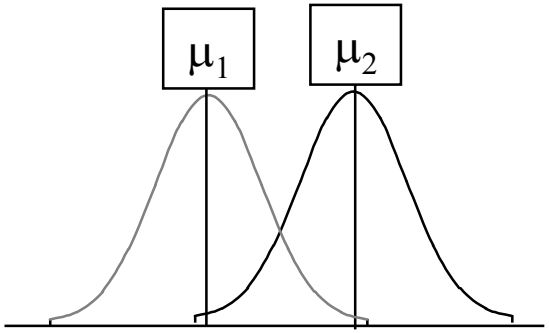
Population dispersions interact in an additive manner

Variations Add

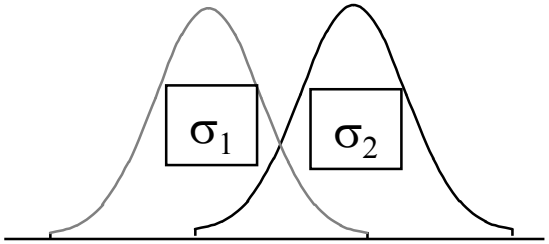
$$\sigma_1^2 + \sigma_2^2 = \sigma_{new}^2$$

HOW DO POPULATIONS INTERACT?

SUBTRACTING TWO POPULATIONS



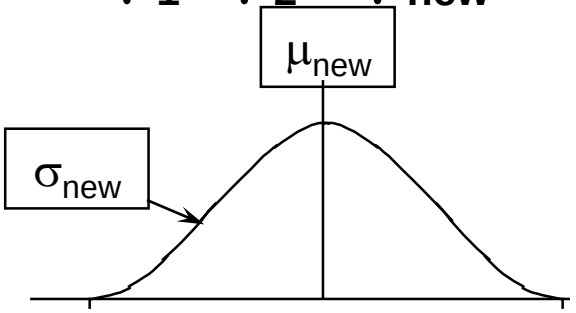
Population means interact in a simple intuitive manner.



Population dispersions interact in an additive manner

Means Subtract

$$\mu_1 - \mu_2 = \mu_{new}$$



Variations Add

$$\sigma_1^2 + \sigma_2^2 = \sigma_{new}^2$$

TRANSACTIONAL EXAMPLE

- **Orders are coming in with the following characteristics:**

$$\overline{X} = \$53{,}000/week$$

$$s = \$8{,}000$$

- **Shipments are going out with the following characteristics:**

$$\overline{X} = \$60{,}000/week$$

$$s = \$5{,}000$$

- **Assuming nothing changes, what percent of the time will shipments exceed orders?**

TRANSACTIONAL EXAMPLE

Orders

$\overline{X} = \$53{,}000$ *in orders/week*

$s = \$8{,}000$

Shipments

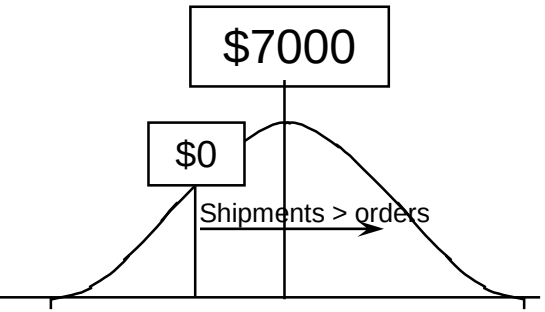
$\overline{X} = \$60{,}000$ *shipped/week*

$s = \$5{,}000$

To solve this problem, we must create a new distribution to model the situation posed in the problem. Since we are looking for shipments to exceed orders, the resulting distribution is created as follows:

$$\overline{X}_{shipments-orders} = \overline{X}_{shipments} - \overline{X}_{orders} = \$60{,}000 - \$53{,}000 = \$7{,}000$$

$$s_{shipments-orders} = \sqrt{s^2_{shipments} + s^2_{orders}} = \sqrt{(5000)^2 + (8000)^2} = \$9434$$

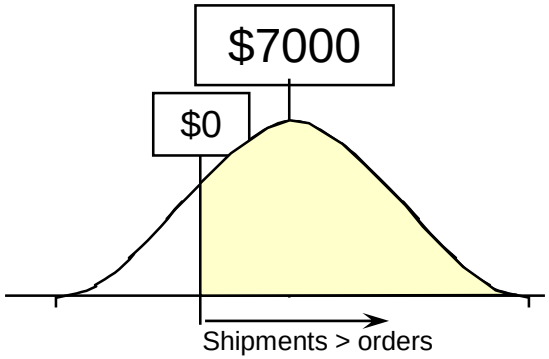


The new distribution looks like this with a mean of $7000 and a standard deviation of $9434. This distribution represents the occurrences of shipments exceeding orders. To answer the original question (shipments>orders) we look for $0 on this new distribution. Any occurrence to the right of this point will represent shipments > orders. So, we need to calculate the percent of the curve that exists to the right of $0.

TRANSACTIONAL EXAMPLE, CONTINUED

$$\overline{X}_{shipments-orders} = \overline{X}_{shipments} - \overline{X}_{orders} = \$60{,}000 - \$53{,}000 = \$7{,}000$$

$$s_{shipments-orders} = \sqrt{s^2_{shipments} + s^2_{orders}} = \sqrt{(5000)^2 + (8000)^2} = \$9434$$



To calculate the percent of the curve to the right of $0 we need to convert the difference between the $0 point and $7000 into sigma intervals. Since we know every $9434 interval from the mean is one sigma, we can calculate this position as follows:

$$\frac{|\mu_0 - \overline{X}|}{s} = \frac{|\$0 - \$7000|}{\$9434} = .74s$$

Look up .74s in the normal table and you will find .77. Therefore, the answer to the original question is that **77% of the time**, shipments will exceed orders.

Now, as a classroom exercise, what percent of the time will shipments exceed orders by $10,000?

MANUFACTURING EXAMPLE

- **2 Blocks are being assembled end to end and significant variation has been found in the overall assembly length.**

- **The blocks have the following dimensions:**

$\overline{X}_1 = 4.00 \text{ inches}$ $\quad$ $\overline{X}_2 = 3.00 \text{ inches}$

$s_1 = .03 \text{ inches}$ $\quad$ $s_2 = .04 \text{ inches}$

- **Determine the overall assembly length and standard deviation.**

CORRELATION ANALYSIS

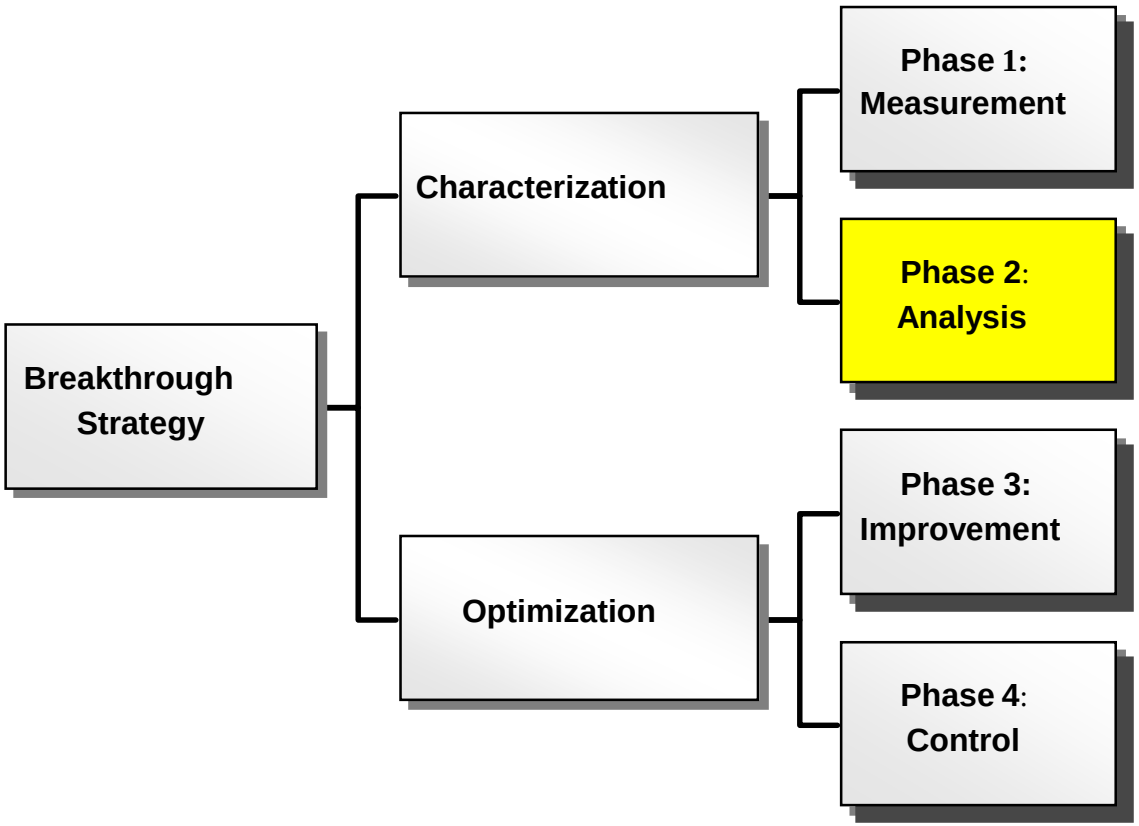
Why do we Care?

Correlation Analysis is necessary to:

- show a relationship between two variables. This also sets the stage for potential cause and effect.

IMPROVEMENT ROADMAP

Uses of Correlation Analysis



Common Uses

- **Determine and quantify the relationship between factors (x) and output characteristics (Y)..**

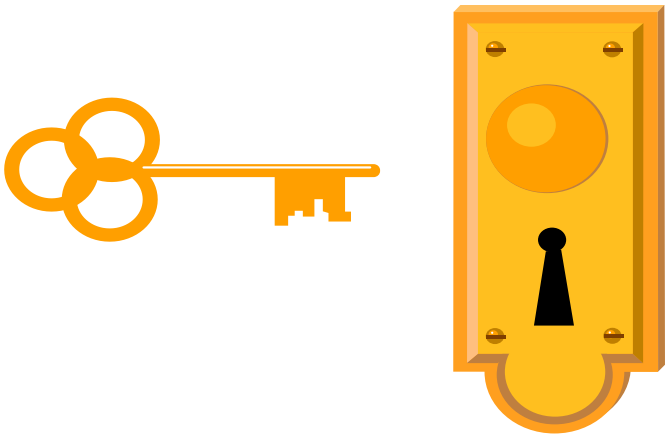
KEYS TO SUCCESS

Always plot the data

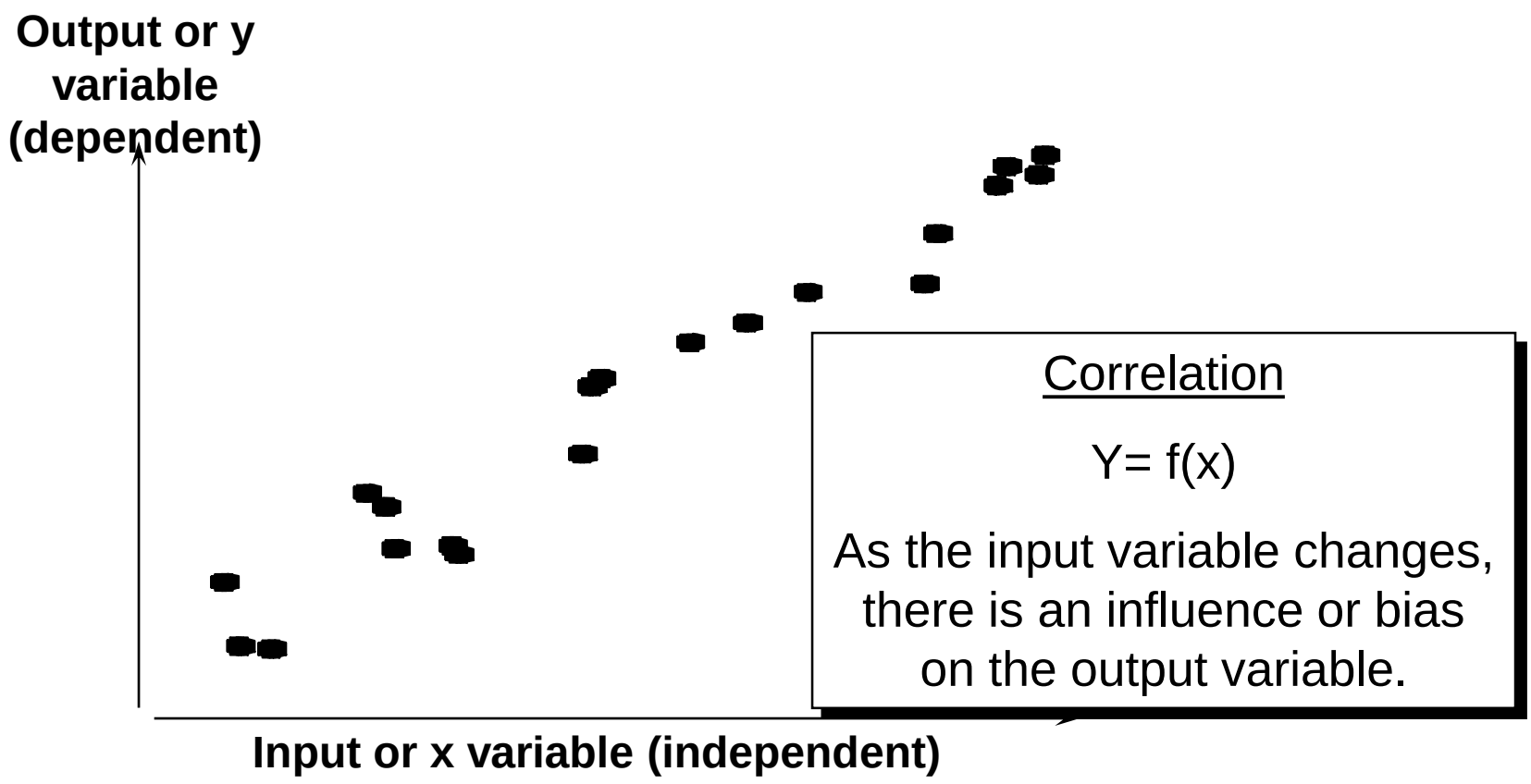
Remember: Correlation does not always imply cause & effect

Use correlation as a follow up to the Fishbone Diagram

Keep it simple and do not let the tool take on a life of its own



WHAT IS CORRELATION?



WHAT IS CORRELATION?

- **A measurable relationship between two variable data characteristics.**

  Not necessarily Cause & Effect (Y=f(x))

- **Correlation requires paired data sets (ie (Y_1,x_1), (Y_2,x_2), etc)**

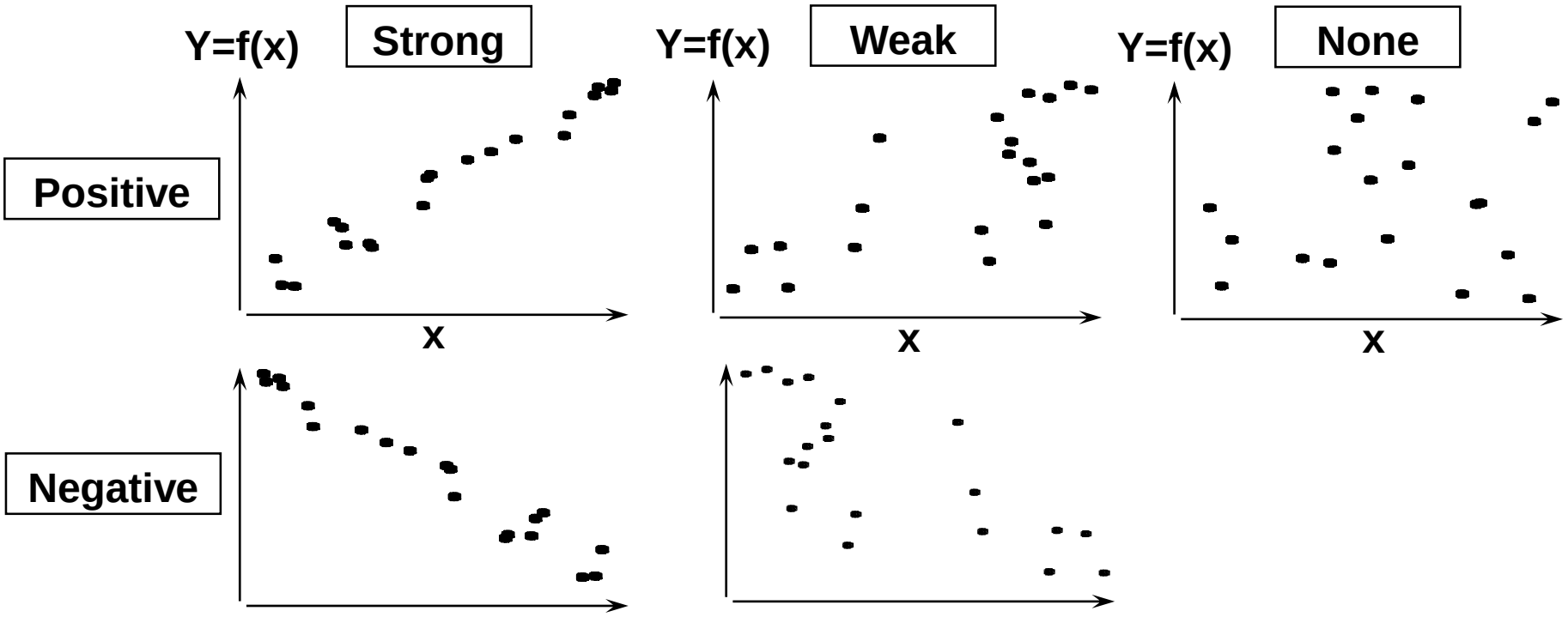
- **The input variable is called the independent variable (x or KPIV) since it is independent of any other constraints**

- **The output variable is called the dependent variable (Y or KPOV) since it is (theoretically) dependent on the value of x.**

- **The coefficient of linear correlation "r" is the measure of the strength of the relationship.**

- **The square of "r" is the percent of the response (Y) which is related to the input (x).**

TYPES OF CORRELATION



CALCULATING "r"

Coefficient of Linear Correlation

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

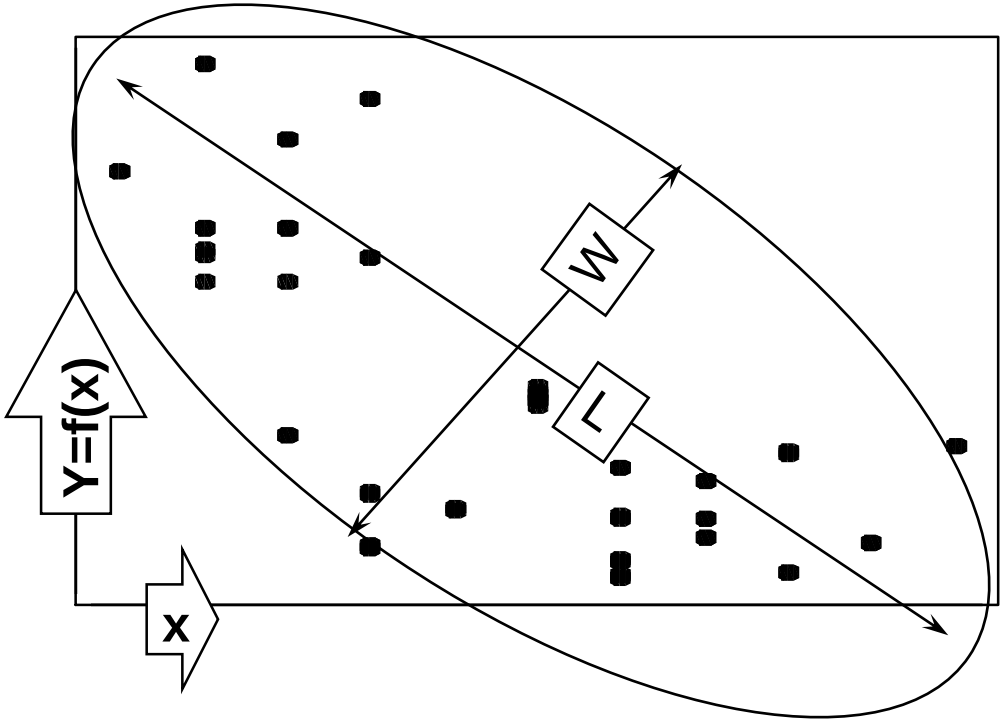
$$r_{CALC} = \frac{s_{xy}}{s_x s_y}$$

- Calculate sample covariance (s_{xy})
- Calculate s_x and s_y for each data set
- Use the calculated values to compute r_{CALC}.
- Add a + for positive correlation and - for a negative correlation.

While this is the most precise method to calculate Pearson's r, there is an easier way to come up with a fairly close approximation...

APPROXIMATING “r”

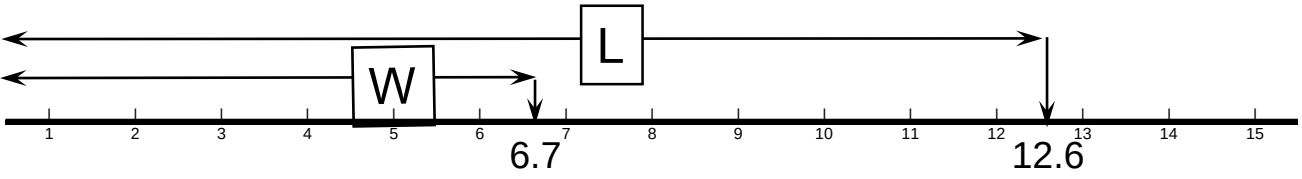
Coefficient of Linear Correlation



- Plot the data on orthogonal axis
- Draw an Oval around the data
- Measure the length and width of the Oval
- Calculate the coefficient of linear correlation (r) based on the formulas below

$$r \approx \pm\left(1-\frac{W}{L}\right)$$

$$r \approx +\left(1-\frac{6.7}{12.6}\right) = +.47$$



+ = positive slope

- = negative slope

HOW DO I KNOW WHEN I HAVE CORRELATION?

Ordered Pairs	r_{CRIT}
5	.88
6	.81
7	.75
8	.71
9	.67
10	.63
15	.51
20	.44
25	.40
30	.36
50	.28
80	.22
100	.20

• The answer should strike a familiar cord at this point… We have confidence (95%) that we have correlation when **$|r_{CALC}| > r_{CRIT}$.**

•Since sample size is a key determinate of r_{CRIT} we need to use a table to determine the correct r_{CRIT} given the number of ordered pairs which comprise the complete data set.

•So, in the preceding example we had 60 ordered pairs of data and we computed a r_{CALC} of -.47. Using the table at the left we determine that the r_{CRIT} value for 60 is .26.

•Comparing $|r_{CALC}| > r_{CRIT}$ we get .47 > .26. Therefore the calculated value exceeds the minimum critical value required for significance.

• Conclusion: We are 95% confident that the observed correlation is significant.

CENTRAL LIMIT THEOREM

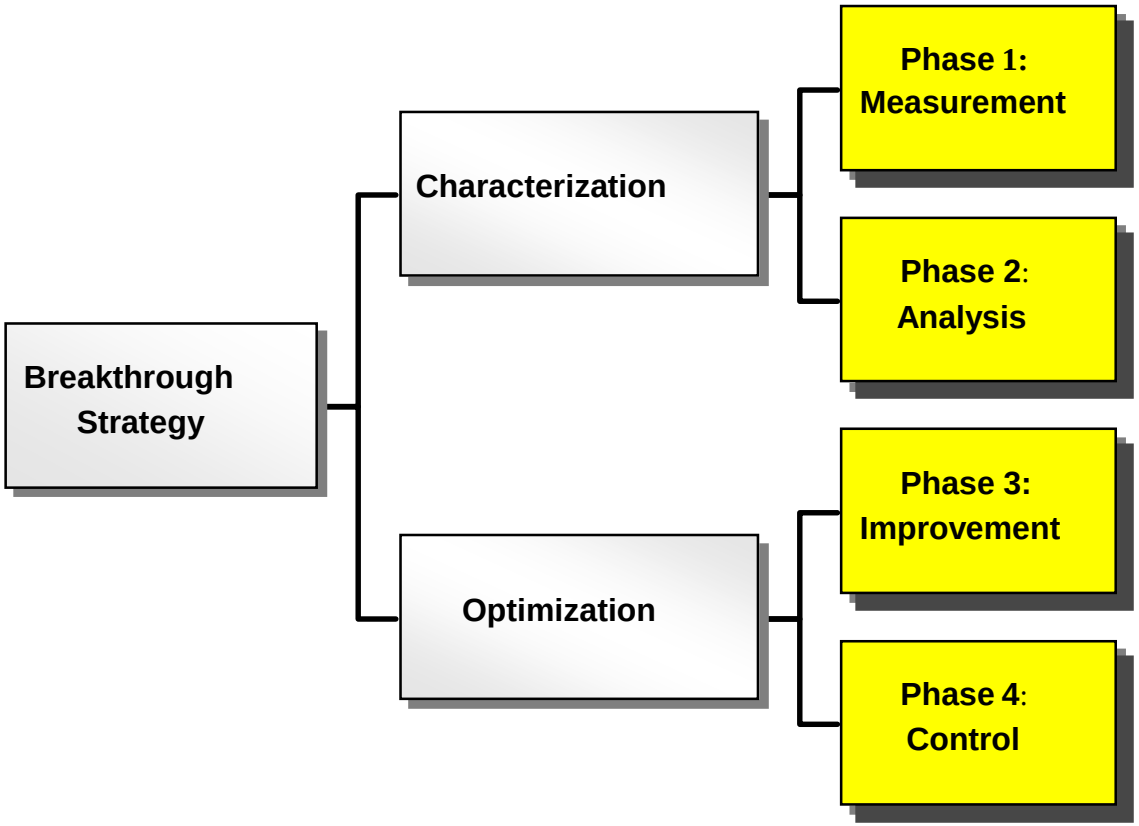
Why do we Care?

The Central Limit Theorem is:

- the key theoretical link between the normal distribution and sampling distributions.
- the means by which almost any sampling distribution, no matter how irregular, can be approximated by a normal distribution if the sample size is large enough.

IMPROVEMENT ROADMAP

Uses of the Central Limit Theorem

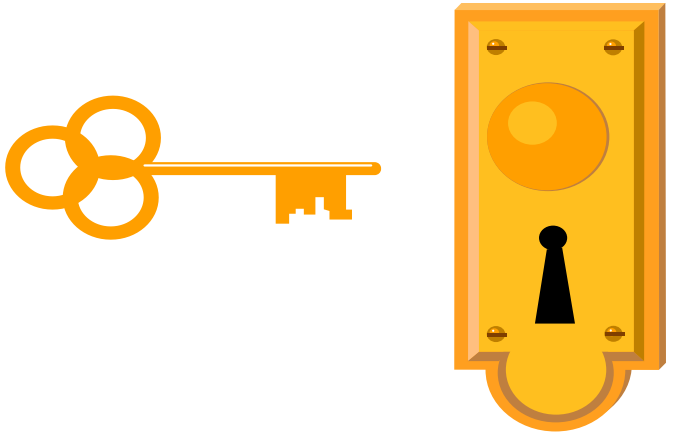


Common Uses

- The Central Limit Theorem underlies all statistic techniques which rely on normality as a fundamental assumption

KEYS TO SUCCESS

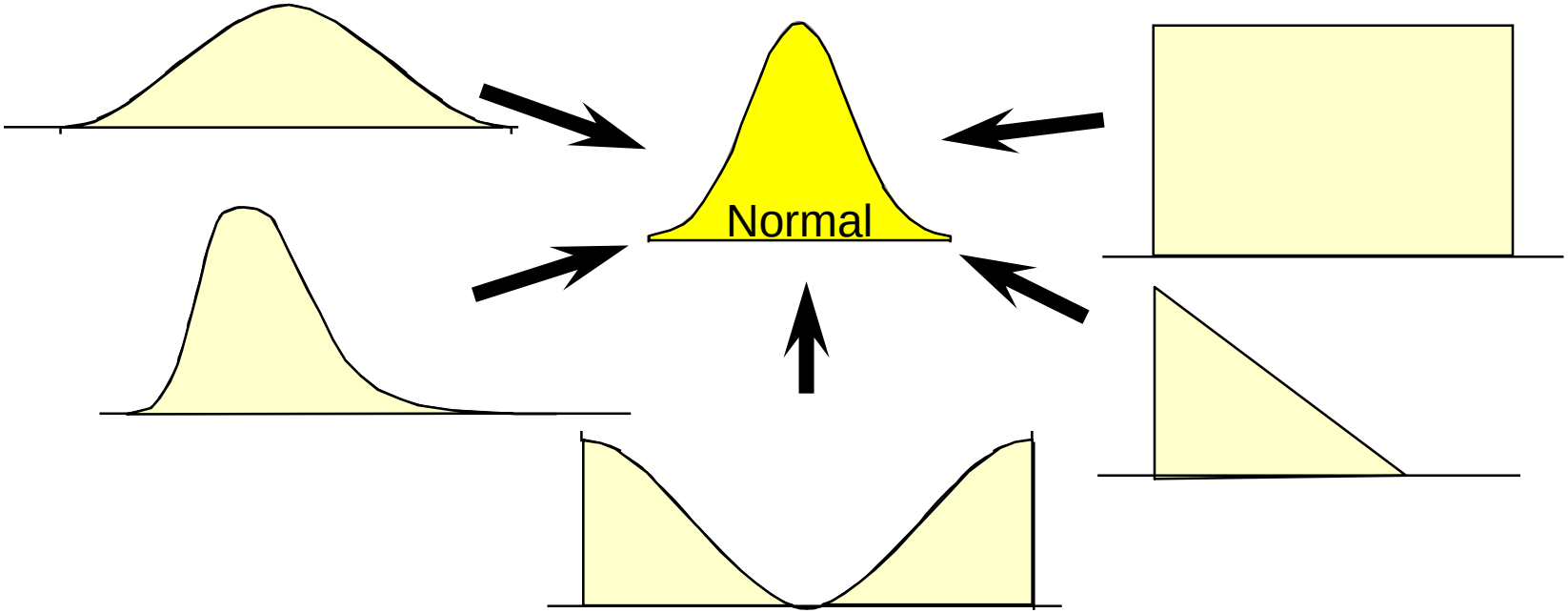
Focus on the practical application of the concept



WHAT IS THE CENTRAL LIMIT THEOREM?

Central Limit Theorem

For almost all populations, the sampling distribution of the mean can be approximated closely by a normal distribution, provided the sample size is sufficiently large.



Why do we Care?

What this means is that no matter what kind of distribution we sample, if the sample size is big enough, the distribution for the mean is approximately normal.

This is the key link that allows us to use much of the inferential statistics we have been working with so far.

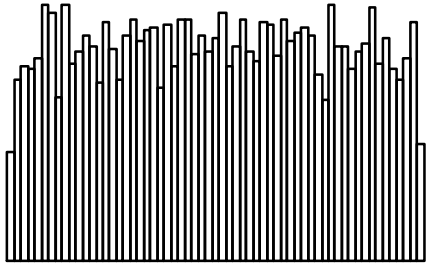
This is the reason that only a few probability distributions (Z, t and χ^2) have such broad application.

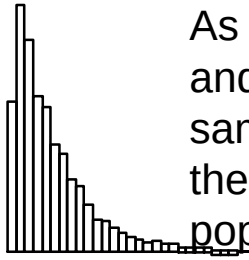
If a random event happens a great many times, the **average** results are likely to be predictable.

Jacob Bernoulli

HOW DOES THIS WORK?

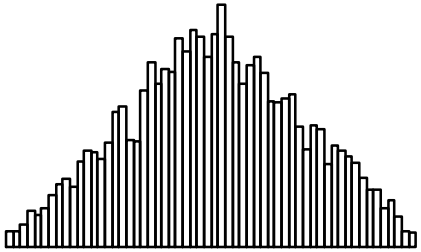
Parent Population

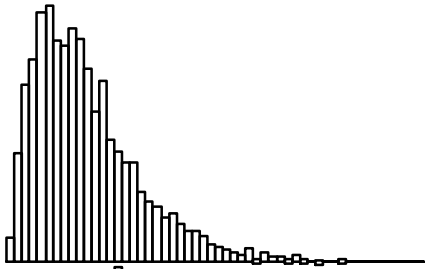




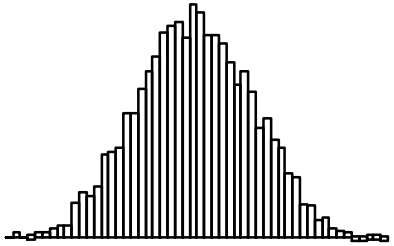
As you average a larger and larger number of samples, you can see how the original sampled population is transformed..

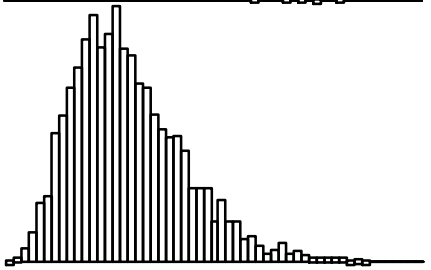
n=2



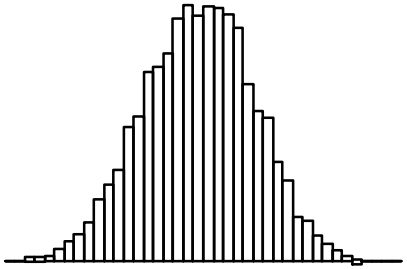


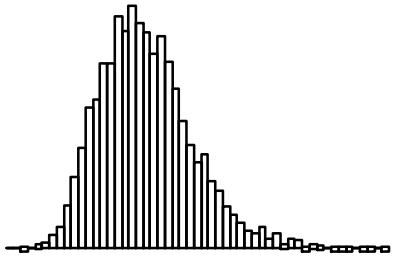
n=5





n=10





ANOTHER PRACTICAL ASPECT

$$s_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

This formula is for the standard error of the mean.

What that means in layman's terms is that this formula is the prime driver of the error term of the mean. Reducing this error term has a direct impact on improving the precision of our estimate of the mean.

The practical aspect of all this is that if you want to improve the precision of any test, increase the sample size.

So, if you want to reduce measurement error (for example) to determine a better estimate of a true value, increase the sample size. The resulting error will be reduced by a factor of $\frac{1}{\sqrt{n}}$. The same goes for any significance testing. Increasing the sample size will reduce the error in a similar manner.

DICE EXERCISE

- Break into 3 teams
  - Team one will be using 2 dice
  - Team two will be using 4 dice
  - Team three will be using 6 dice

- Each team will conduct 100 throws of their dice and record the average of each throw.

- Plot a histogram of the resulting data.

- Each team presents the results in a 10 min report out.

PROCESS CAPABILITY ANALYSIS

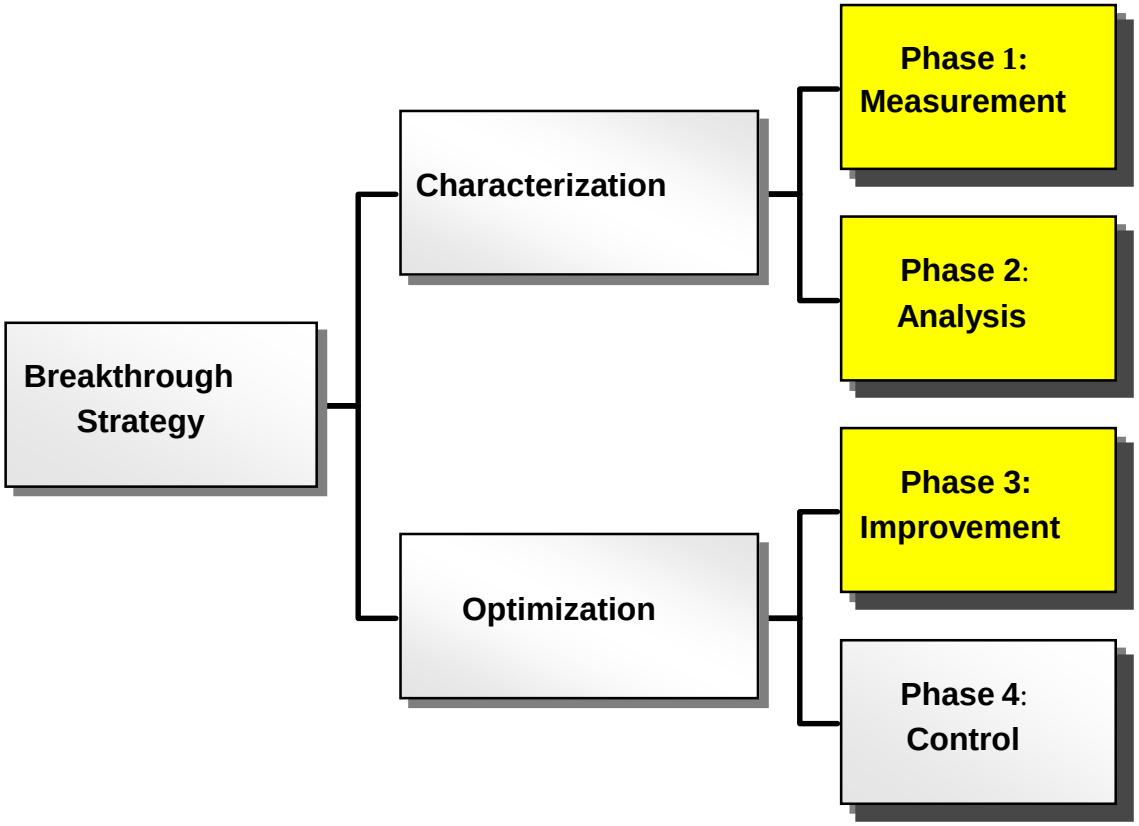
Why do we Care?

Process Capability Analysis is necessary to:

- determine the area of focus which will ensure successful resolution of the project.
- benchmark a process to enable demonstrated levels of improvement after successful resolution of the project.
- demonstrate improvement after successful resolution of the project.

IMPROVEMENT ROADMAP

Uses of Process Capability Analysis



Common Uses

- **Baselining a process primary metric (Y) prior to starting a project.**
- **Characterizing the capability of causitive factors (x).**
- **Characterizing a process primary metric after changes have been implemented to demonstrate the level of improvement.**

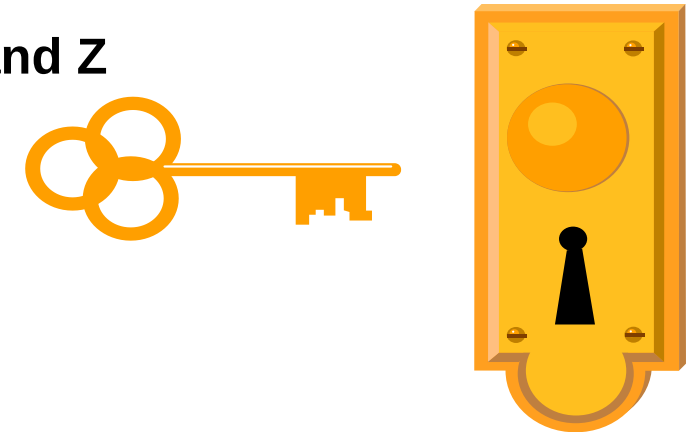
KEYS TO SUCCESS

Must have specification limits - Use process targets if no specs available

Don’t get lost in the math

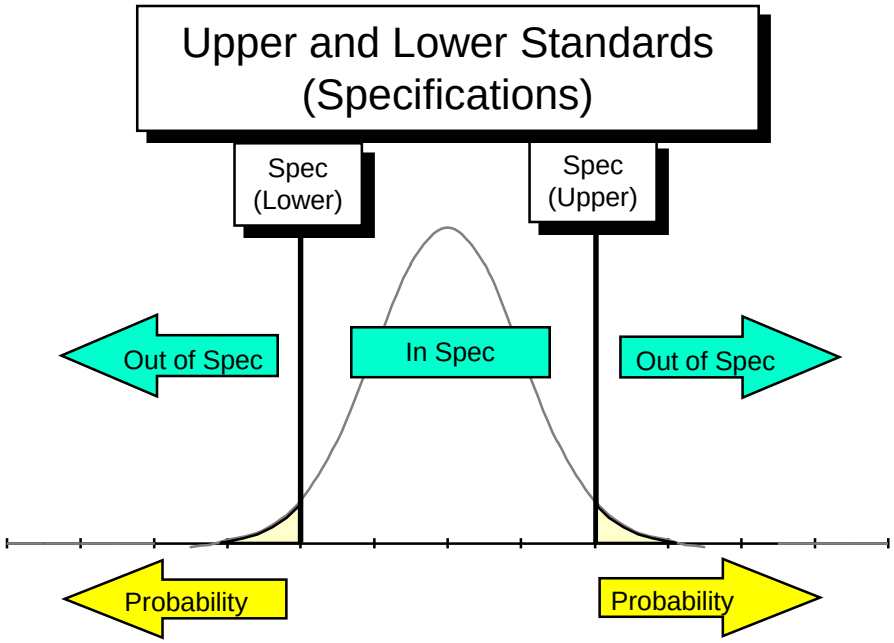
Relate to Z for comparisons (Cpk x 3 = Z)

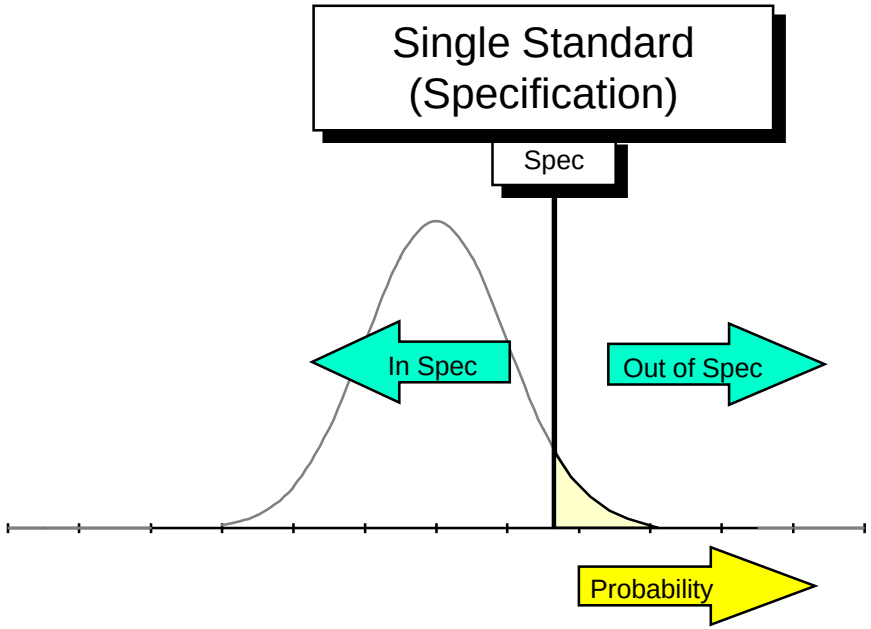
For Attribute data use PPM conversion to Cpk and Z



WHAT IS PROCESS CAPABILITY?

Process capability is simply a measure of how good a metric is performing against and established standard(s). Assuming we have a stable process generating the metric, it also allows us to predict the probability of the metric value being outside of the established standard(s).



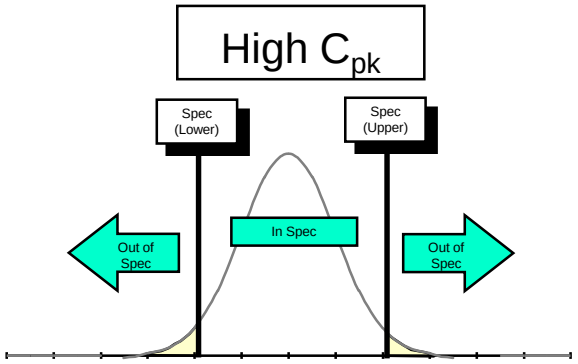


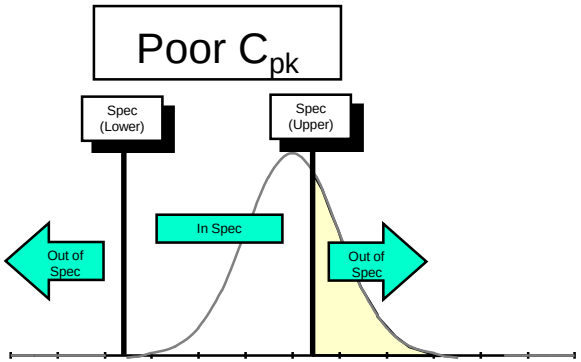
WHAT IS PROCESS CAPABILITY?

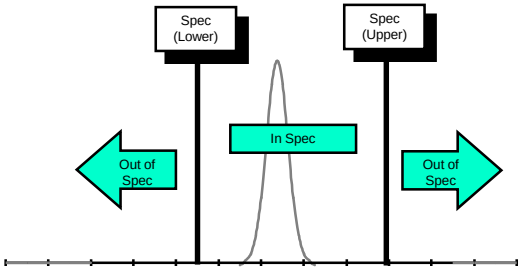
Process capability (C_{pk}) is a function of how the population is centered (|μ-spec|) and the population spread (σ).

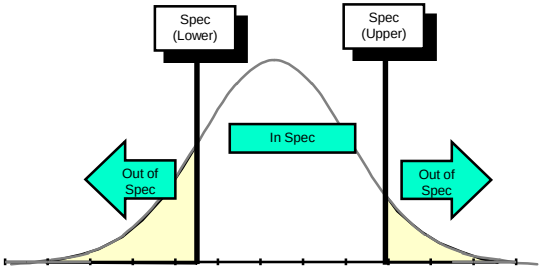
Process Center (|μ-spec|)

Process Spread (σ)

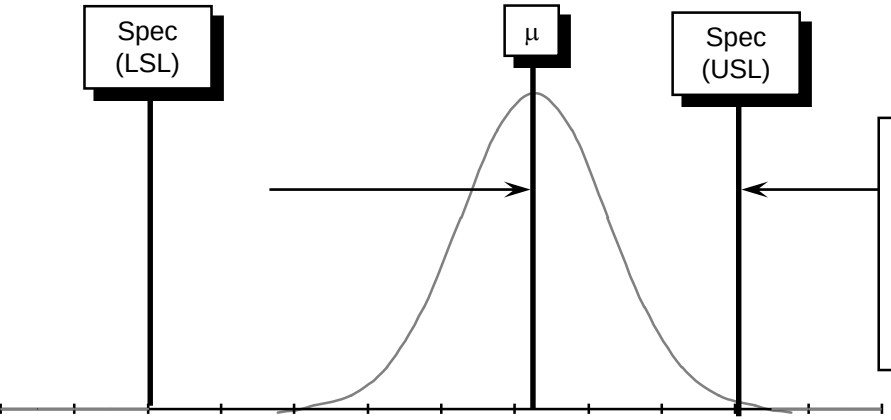








HOW IS PROCESS CAPABILITY CALCULATED



Distance between the population mean and the nearest spec limit (|μ-USL |). This distance divided by 3σ is C_{pk}.

Expressed mathematically, this looks like:

$$C_{PK} = \frac{MIN(\mu - LSL, USL - \mu)}{3\sigma}$$

Note:

LSL = Lower Spec Limit

USL = Upper Spec Limit

PROCESS CAPABILITY EXAMPLE

We want to calculate the process capability for our inventory. The historical average monthly inventory is $250,000 with a standard deviation of $20,000. Our inventory target is $200,000 maximum.

- **Calculation Values:**
  - **Upper Spec value = $200,000 maximum**
  - **No Lower Spec**
  - **μ = historical average = $250,000**
  - **s = $20,000**
- **Calculation:** $C_{PK} = \frac{MIN(\mu - LSL, USL - \mu)}{3\sigma} = \frac{(\$200{,}000 - \$250{,}000)}{3 * \$20{,}000} = -.83$

Answer: C_{pk} = -.83

ATTRIBUTE PROCESS CAPABILITY TRANSFORM

Z	PPM $_{ST}$	C_{pk}	PPM $_{LT}$ (+1.5 σ)
0.0	500,000	0.0	933,193
0.1	460,172	0.0	919,243
0.2	420,740	0.1	903,199
0.3	382,089	0.1	884,930
0.4	344,578	0.1	864,334
0.5	308,538	0.2	841,345
0.6	274,253	0.2	815,940
0.7	241,964	0.2	788,145
0.8	211,855	0.3	758,036
0.9	184,060	0.3	725,747
1.0	158,655	0.3	691,462
1.1	135,666	0.4	655,422
1.2	115,070	0.4	617,911
1.3	96,801	0.4	579,260
1.4	80,757	0.5	539,828
1.5	66,807	0.5	500,000
1.6	54,799	0.5	460,172
1.7	44,565	0.6	420,740
1.8	35,930	0.6	382,089
1.9	28,716	0.6	344,578
2.0	22,750	0.7	308,538
2.1	17,864	0.7	274,253
2.2	13,903	0.7	241,964
2.3	10,724	0.8	211,855
2.4	8,198	0.8	184,060
2.5	6,210	0.8	158,655
2.6	4,661	0.9	135,666
2.7	3,467	0.9	115,070
2.8	2,555	0.9	96,801
2.9	1,866	1.0	80,757
3.0	1,350	1.0	66,807
3.1	968	1.0	54,799
3.2	687	1.1	44,565
3.3	483	1.1	35,930
3.4	337	1.1	28,716
3.5	233	1.2	22,750
3.6	159	1.2	17,864
3.7	108	1.2	13,903
3.8	72.4	1.3	10,724
3.9	48.1	1.3	8,198
4.0	31.7	1.3	6,210

If we take the Cpk formula below

$$C_{PK} = \frac{MIN(\mu - LSL, USL - \mu)}{3\sigma}$$

We find that it bears a striking resemblance to the equation for Z which is:

$$Z_{CALC} = \frac{\mu - \mu_0}{\sigma}$$

with the value μ-μ_0 substituted for MIN(μ-LSL,USL-μ).

Making this substitution, we get :

$$C_{pk} = \frac{1}{3} * \frac{MIN(\mu - LSL, USL - \mu)}{\sigma} = \frac{Z_{MIN(\mu - LSL, USL - \mu)}}{3}$$

We can now use a table similar to the one on the left to transform either Z or the associated PPM to an equivalent C_{pk} value.

So, if we have a process which has a short term PPM=136,666 we find that the equivalent Z=1.1 and Cpk=0.4 from the table.

MULTI-VARI ANALYSIS

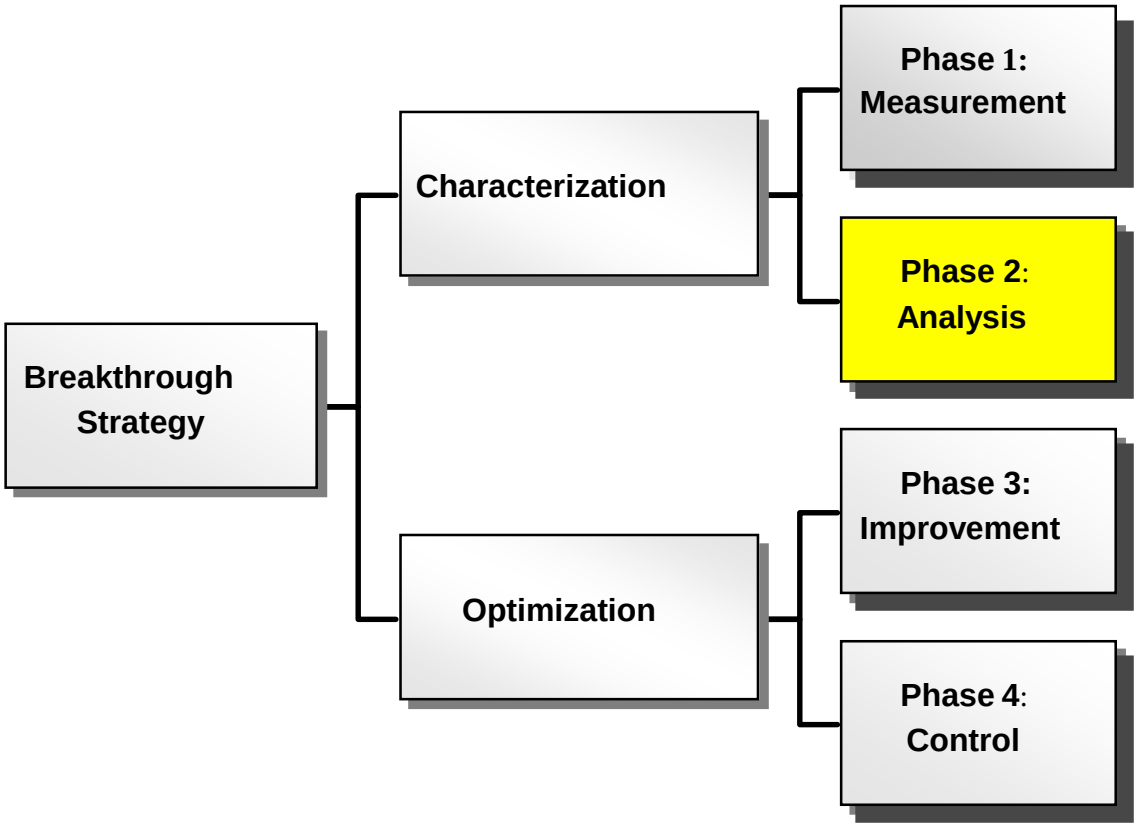
Why do we Care?

Multi-Vari charts are a:

- Simple, yet powerful way to significantly reduce the number of potential factors which could be impacting your primary metric.
- Quick and efficient method to significantly reduce the time and resources required to determine the primary components of variation.

IMPROVEMENT ROADMAP

Uses of Multi-Vari Charts



Common Uses

- **Eliminate a large number of factors from the universe of potential factors.**

KEYS TO SUCCESS

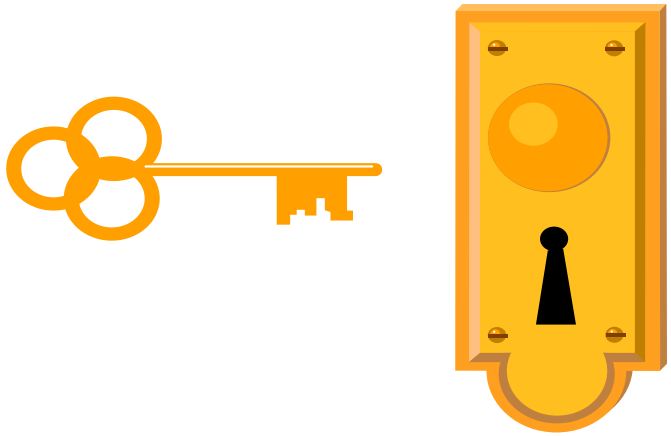
Careful planning before you start

Gather data by systematically sampling the existing process

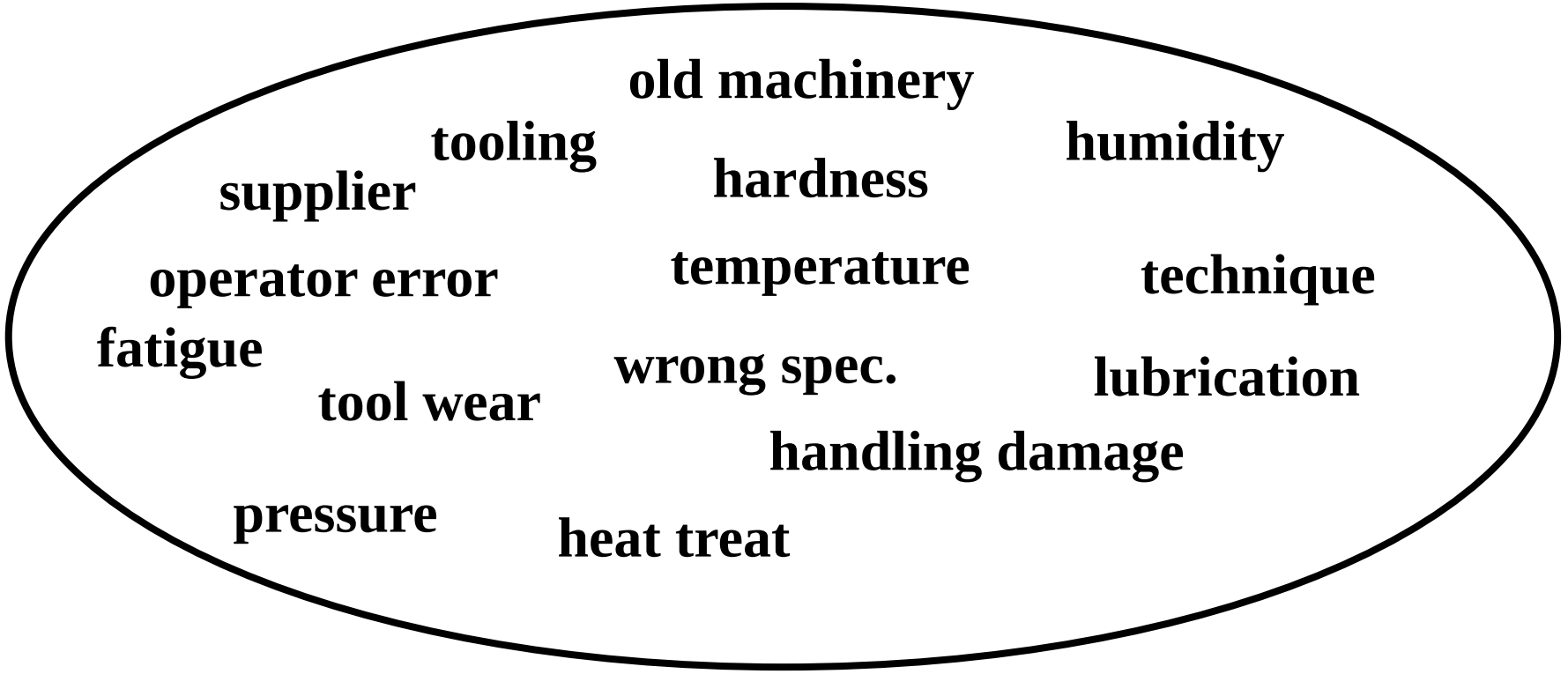
Perform “ad hoc” training on the tool for the team prior to use

Ensure your sampling plan is complete prior to gathering data

Have team members (or yourself) do the sampling to avoid bias



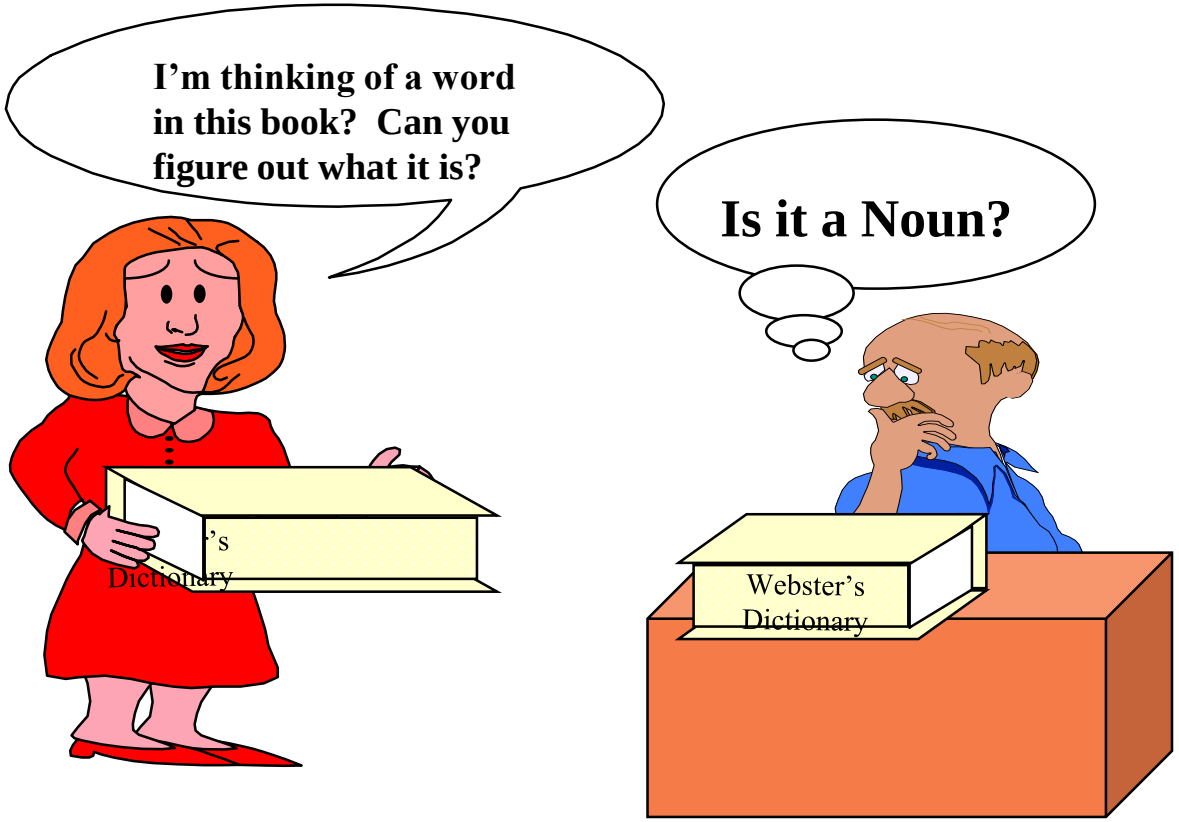
A World of Possible Causes (KPIVs).....



The Goal of the logical search is to narrow down to 5-6 key variables !

REDUCING THE POSSIBILITIES

"…..The Dictionary Game? "



USE A LOGICAL APPROACH TO SEE THE MAJOR SOURCES OF VARIATION

REDUCING THE POSSIBILITIES

How many guesses do you think it will take to find a single word in the text book?

Lets try and see…….

REDUCING THE POSSIBILITIES

How many guesses do you think it will take to find a single word in the text book?

Statistically it should take no more than 17 guesses

2^{17}=2X2X2X2X2X2X2X2X2X2X2X2X2X2X2X2X2= 131,072

Most Unabridged dictionaries have 127,000 words.

<u>Reduction of possibilities can be an extremely powerful technique…..</u>

PLANNING A MULTI-VARI ANALYSIS

- Determine the possible families of variation.
- Determine how you will take the samples.

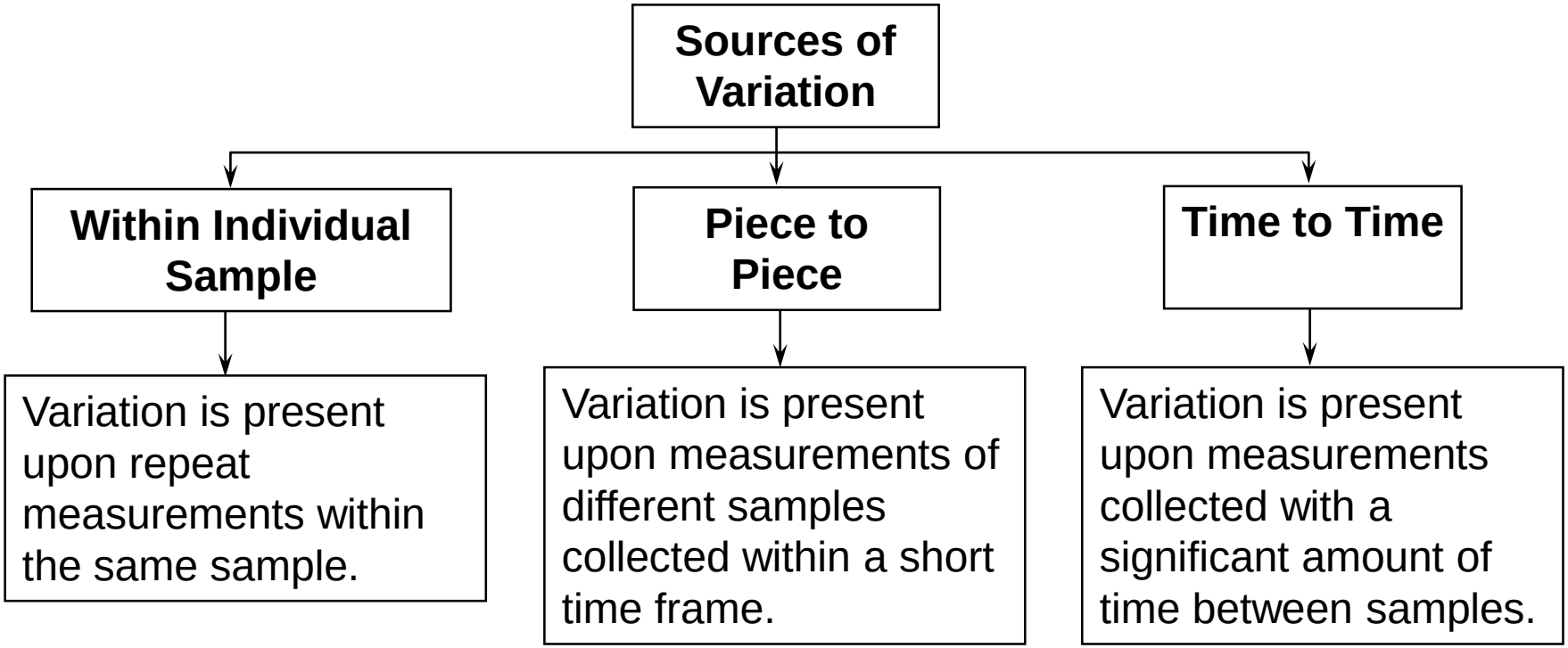
- Take a stratified sample (in order of creation).
- DO NOT take random samples.

- Take a minimum of 3 samples per group and 3 groups.
- The samples must represent the full range of the process.

- Does one sample or do just a few samples stand out?
- There could be a main effect or an interaction at the cause.

MULTI-VARI ANALYSIS, VARIATION FAMILIES

The key is reducing the number of possibilities to a manageable few….

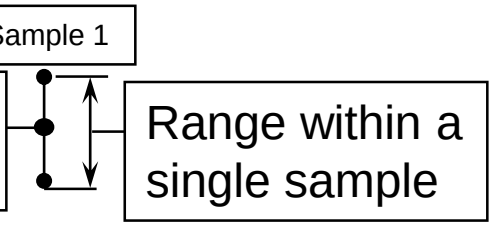


MULTI-VARI ANALYSIS, VARIATION SOURCES

	Within Individual Sample	Piece to Piece	Time to Time
Manufacturing (Machining)	Measurement Accuracy Out of Round Irregularities in Part	Machine fixturing Mold cavity differences	Material Changes Setup Differences Tool Wear Calibration Drift Operator Influence
Transactional (Order Rate)	Measurement Accuracy Line Item Complexity	Customer Differences Order Editor Sales Office Sales Rep	Seasonal Variation Management Changes Economic Shifts Interest Rate

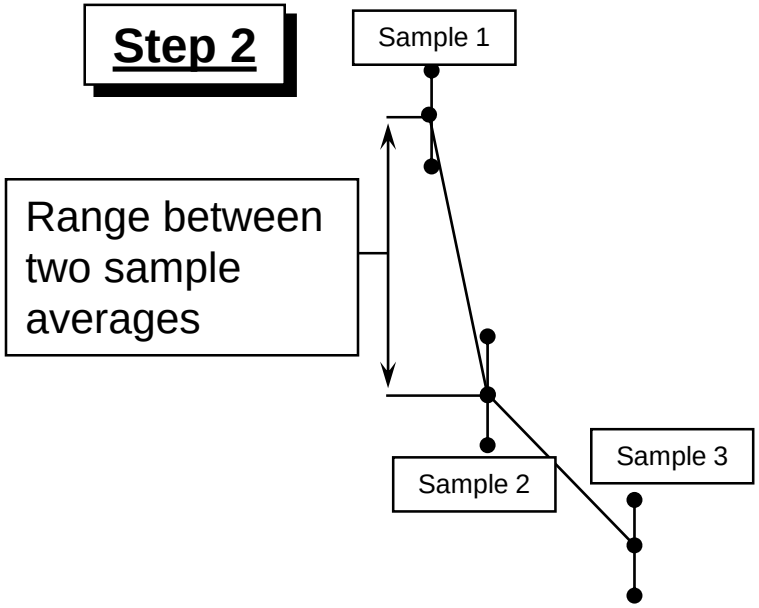
HOW TO DRAW THE CHART

Step 1



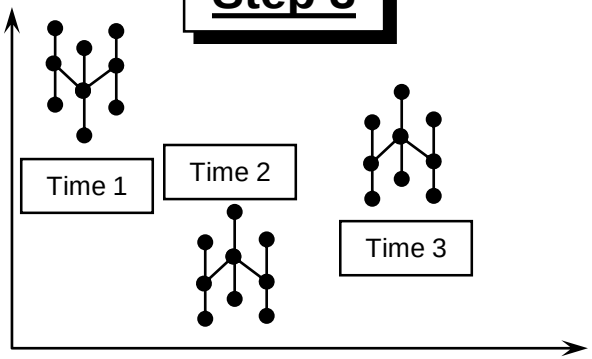
Plot the first sample range with a point for the maximum reading obtained, and a point for the minimum reading. Connect the points and plot a third point at the average of the within sample readings

Step 2



Plot the sample ranges for the remaining “piece to piece” data. Connect the averages of the within sample readings.

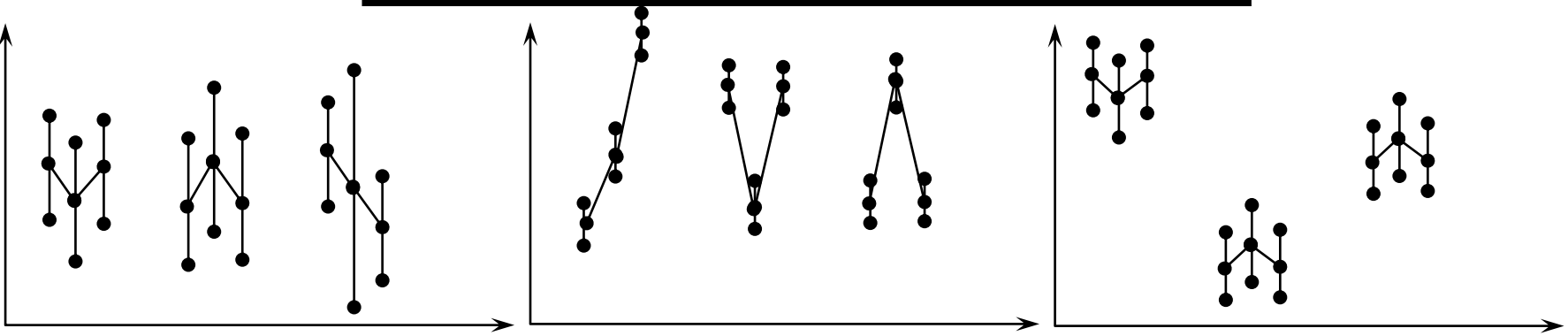
Step 3



Plot the “time to time” groups in the same manner.

READING THE TEA LEAVES….

Common Patterns of Variation



Within Piece

•Characterized by large variation in readings taken of the same single sample, often from different positions within the sample.

Piece to Piece

•Characterized by large variation in readings taken between samples taken within a short time frame.

Time to Time

•Characterized by large variation in readings taken between samples taken in groups with a significant amount of time elapsed between groups.

MULTI-VARI EXERCISE

We have a part dimension which is considered to be impossible to manufacture. A capability study seems to confirm that the process is operating with a C_{pk}=0 (500,000 ppm). You and your team decide to use a Multi-Vari chart to localize the potential sources of variation. You have gathered the following data:

Construct a multi-vari chart of the data and interpret the results.

Sample	Day/ Time	Beginning of Part	Middle of Part	End of Part
1	1/0900	.015	.017	.018
2	1/0905	.010	.012	.015
3	1/0910	.013	.015	.016
4	2/1250	.014	.015	.018
5	2/1255	.009	.012	.017
6	2/1300	.012	.014	.016
7	3/1600	.013	.014	.017
8	3/1605	.010	.013	.015
9	3/1610	.011	.014	.017

SAMPLE SIZE CONSIDERATIONS

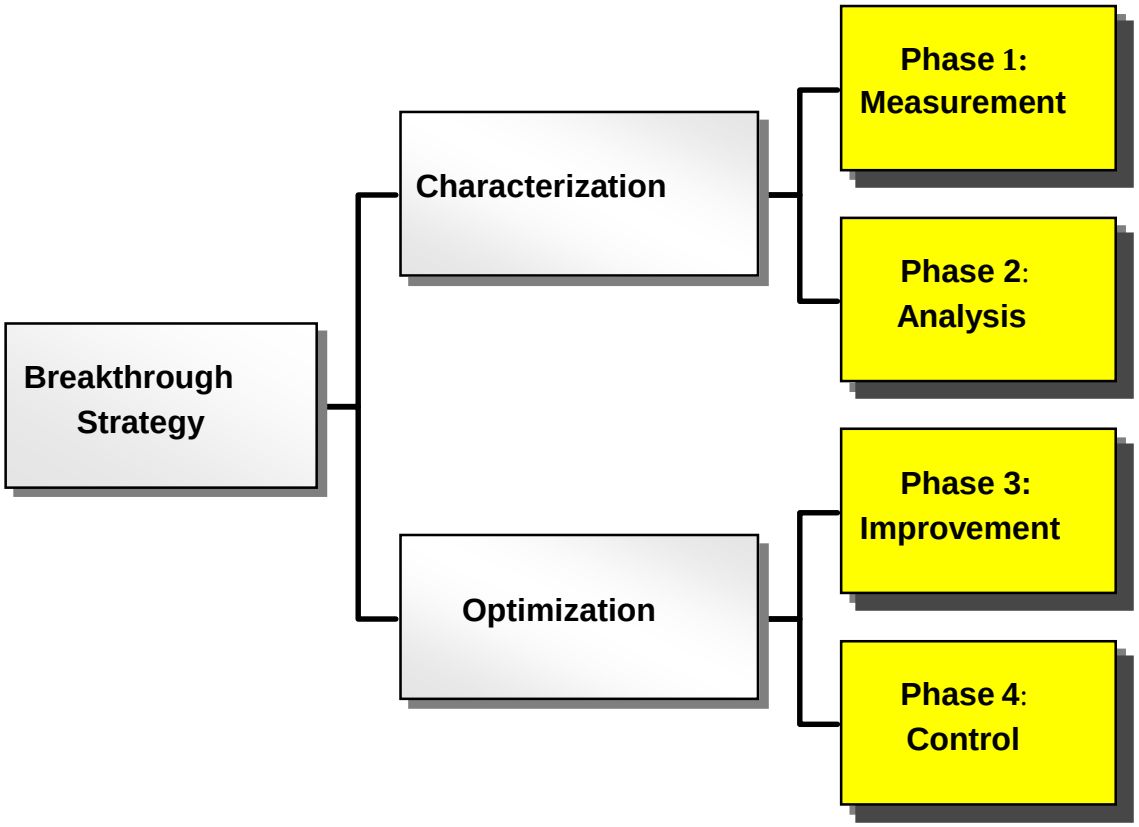
Why do we Care?

<u>The correct sample size is necessary to:</u>

- ensure any tests you design have a high probability of success.
- properly utilize the type of data you have chosen or are limited to working with.

IMPROVEMENT ROADMAP

Uses of Sample Size Considerations



Common Uses

- **Sample Size considerations are used in any situation where a sample is being used to infer a population characteristic.**

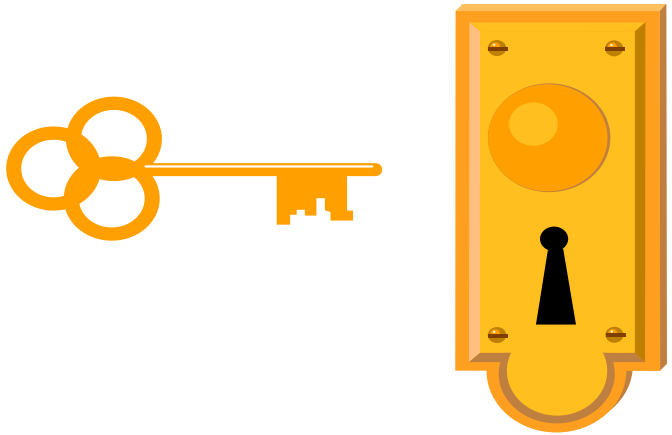
KEYS TO SUCCESS

Use variable data wherever possible

Generally, more samples are better in any study

When there is any doubt, calculate the needed sample size

Use the provided excel spreadsheet to ease sample size calculations



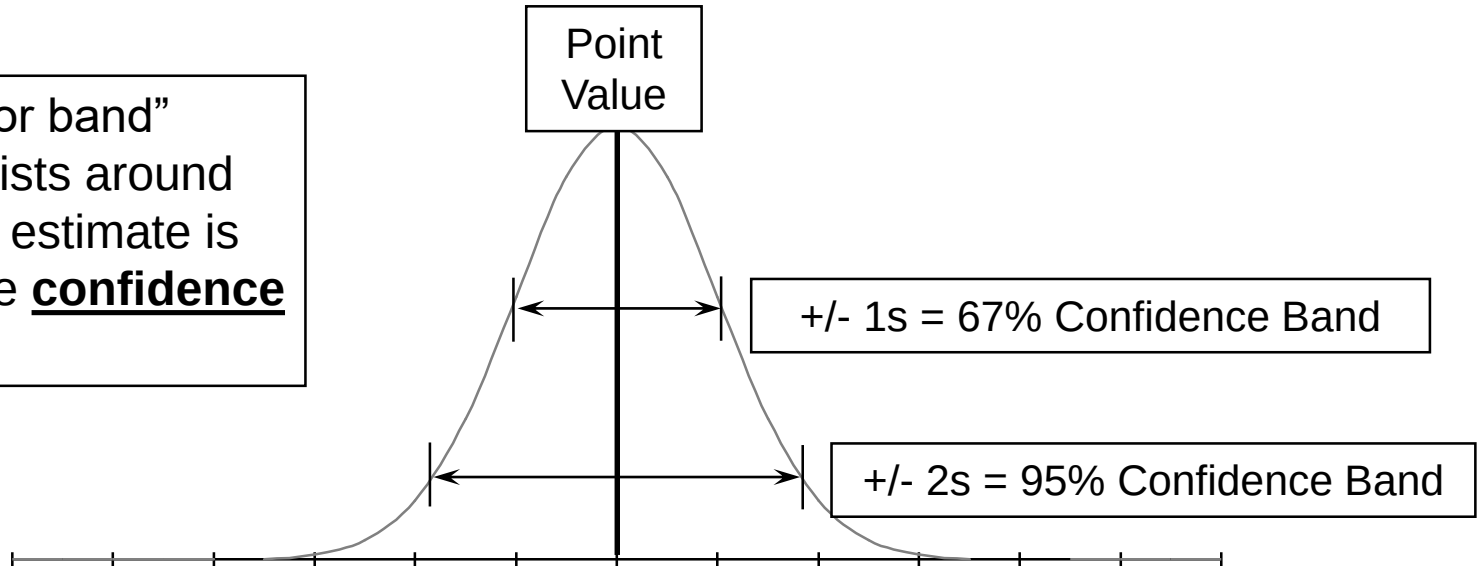
CONFIDENCE INTERVALS

The possibility of error exists in almost every system. This goes for point values as well. While we report a specific value, that value only represents our best estimate from the data at hand. The best way to think about this is to use the form:

true value = point estimate +/- error

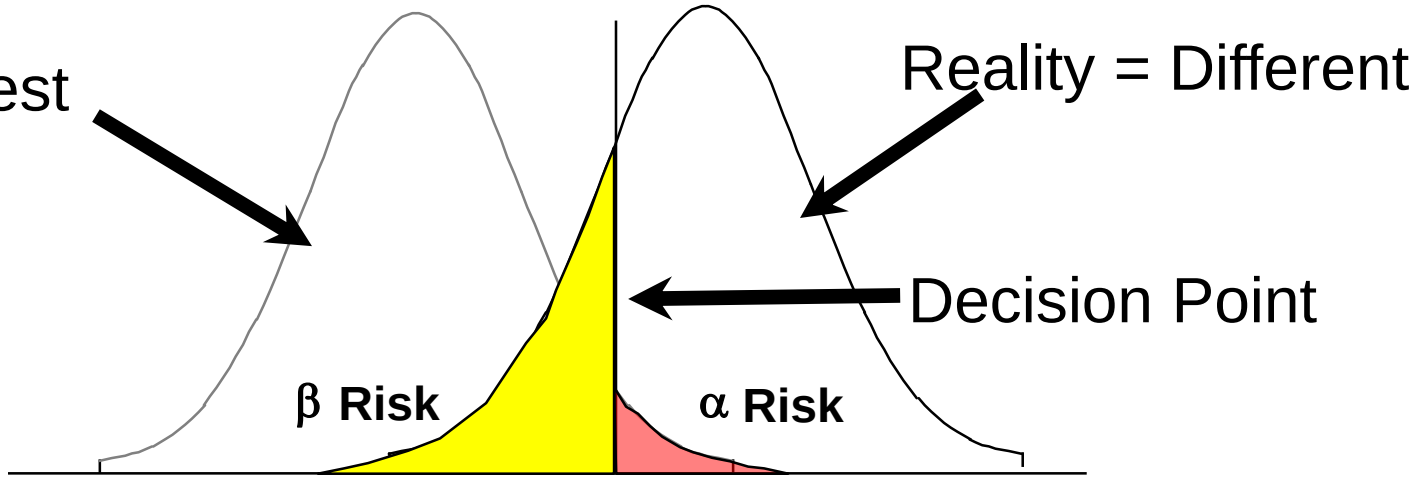
The error around the point value follows one of several common probability distributions. As you have seen so far, we can increase our confidence is to go further and further out on the tails of this distribution.

This "error band" which exists around the point estimate is called the **confidence interval.**



BUT WHAT IF I MAKE THE WRONG DECISION?

Test Decision	Reality: Not different (H_o)	Reality: Different (H_1)
Not different (H_o)	Correct Conclusion	• Type II Error • β risk • Consumer Risk
Different (H_1)	• Type I Error • α Risk • Producer Risk	Correct Conclusion



WHY DO WE CARE IF WE HAVE THE TRUE VALUE?

How confident do you want to be that you have made the right decision?

A person does not feel well and checks into a hospital for tests.

Test Decision	Reality: Not different (H_o)	Reality: Different (H_1)
Not different (H_o)	**Correct Conclusion**	• **Type II Error** • **β risk** • **Consumer Risk**
Different (H_1)	• **Type I Error** • **α Risk** • **Producer Risk**	**Correct Conclusion**

Ho: Patient is not sick

H1: Patient is sick

Error Impact

Type I Error = Treating a patient who is not sick

Type II Error = Not treating a sick patient

HOW ABOUT ANOTHER EXAMPLE?

A change is made to the sales force to save costs. Did it adversely impact the order receipt rate?

Test Decision	Reality: Not different (H_o)	Reality: Different (H_1)
Not different (H_o)	**Correct Conclusion**	• **Type II Error** • **β risk** • **Consumer Risk**
Different (H_1)	• **Type I Error** • **α Risk** • **Producer Risk**	**Correct Conclusion**

Ho: Order rate unchanged

H1: Order rate is different

Error Impact

Type I Error = Unnecessary costs

Type II Error = Long term loss of sales

CONFIDENCE INTERVAL FORMULAS

Mean

$$\overline{X} - t_{a/2,n-1}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + t_{a/2,n-1}\frac{\sigma}{\sqrt{n}}$$

Standard Deviation

$$s\sqrt{\frac{n-1}{\chi^2_{1-a/2}}} \le \sigma \le s\sqrt{\frac{n-1}{\chi^2_{a/2}}}$$

Process Capability

$$Cp\sqrt{\frac{\chi^2_{1-a/2,n-1}}{n-1}} \le Cp \le Cp\sqrt{\frac{\chi^2_{a/2,n-1}}{n-1}}$$

Percent Defective

$$\hat{p} - Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \bar{p} \le \hat{p} + Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

These individual formulas are not critical at this point, but notice that the only opportunity for decreasing the error band (confidence interval) without decreasing the confidence factor, is to increase the sample size.

SAMPLE SIZE EQUATIONS

Mean

$$n = \left(\frac{Z_{a/2}\sigma}{\mu - \overline{X}} \right)^2$$

Allowable error = $\mu - \overline{X}$

(also known as δ)

Standard Deviation

$$n = \left(\frac{\sigma}{s} \right)^2 \chi^2_{a/2} + 1$$

Allowable error = σ/s

Percent Defective

$$n = \hat{p}(1 - \hat{p}) \left(\frac{Z_{\alpha/2}}{E} \right)^2$$

Allowable error = E

SAMPLE SIZE EXAMPLE

We want to estimate the true average weight for a part within 2 pounds. Historically, the part weight has had a standard deviation of 10 pounds. We would like to be 95% confident in the results.

- **Calculation Values:**
  - **Average tells you to use the mean formula**
  - **Significance: α = 5% (95% confident)**
  - **$Z_{\alpha/2} = Z_{.025} = 1.96$**
  - **s=10 pounds**
  - **μ-x = error allowed = 2 pounds**
- **Calculation:** $$n = \left(\frac{Z_{\alpha/2}\sigma}{\mu - \overline{X}} \right)^2 = \left(\frac{1.96 * 10}{2} \right)^2 = 97$$
- **Answer: n=97 Samples**

SAMPLE SIZE EXAMPLE

We want to estimate the true percent defective for a part within 1%. Historically, the part percent defective has been 10%. We would like to be 95% confident in the results.

- **Calculation Values:**
  - **Percent defective tells you to use the percent defect formula**
  - **Significance: α = 5% (95% confident)**
  - **$Z_{\alpha/2} = Z_{.025} = 1.96$**
  - **p = 10% = .1**
  - **E = 1% = .01**
- **Calculation:**

$$n = \hat{p}(1-\hat{p})\left(\frac{Z_{\alpha/2}}{E}\right)^2 = .1(1-.1)\left(\frac{1.96}{.01}\right)^2 = 3458$$

- **Answer: n=3458 Samples**

CONFIDENCE INTERVALS

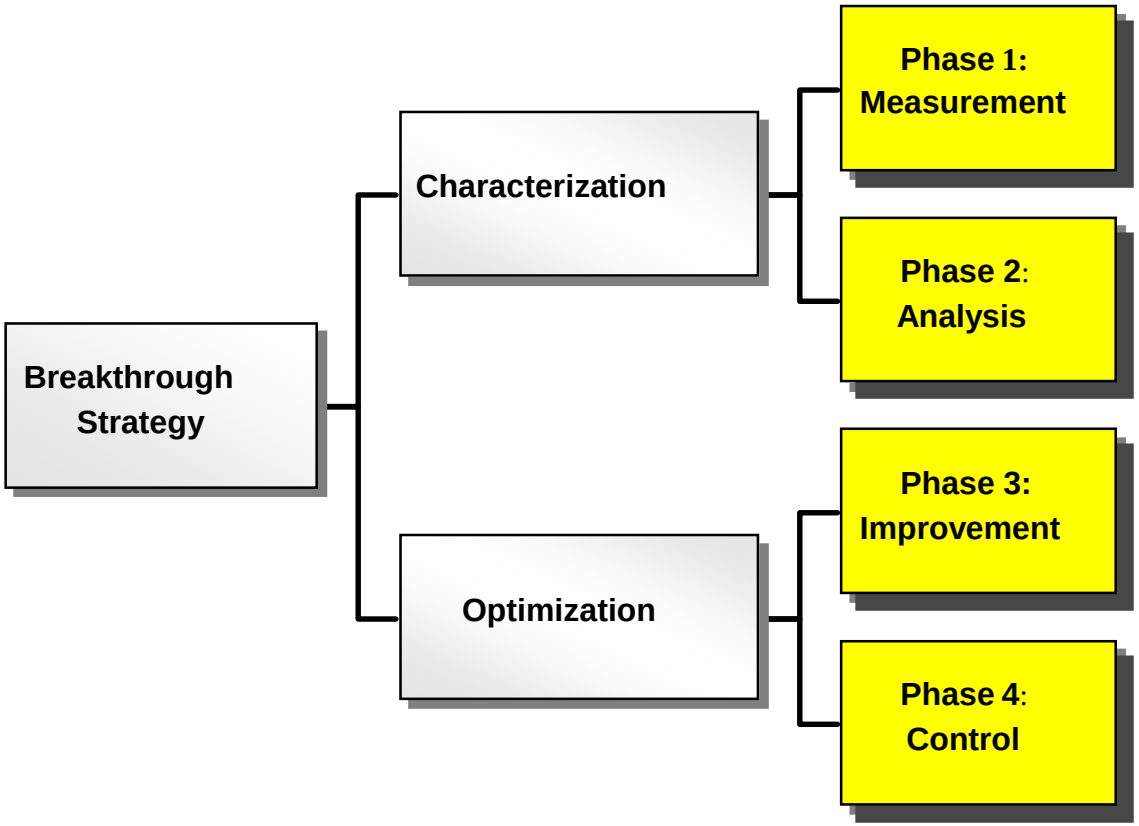
Why do we Care?

Understanding the confidence interval is key to:

- understanding the limitations of quotes in point estimate data.
- being able to quickly and efficiently screen a series of point estimate data for significance.

IMPROVEMENT ROADMAP

Uses of Confidence Intervals



Common Uses

- Used in any situation where data is being evaluated for significance.

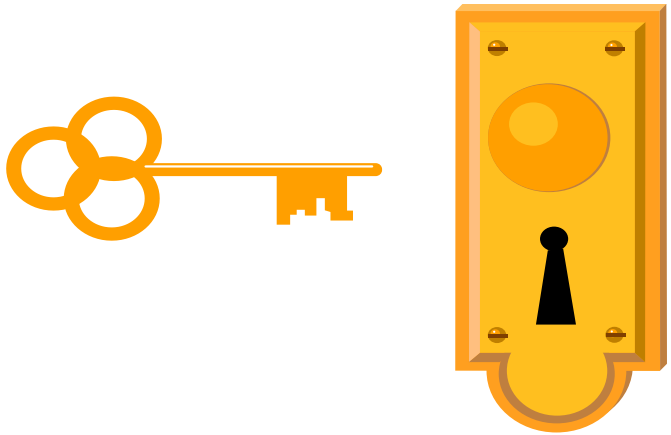
KEYS TO SUCCESS

Use variable data wherever possible

Generally, more samples are better (limited only by cost)

Recalculate confidence intervals frequently

Use an excel spreadsheet to ease calculations



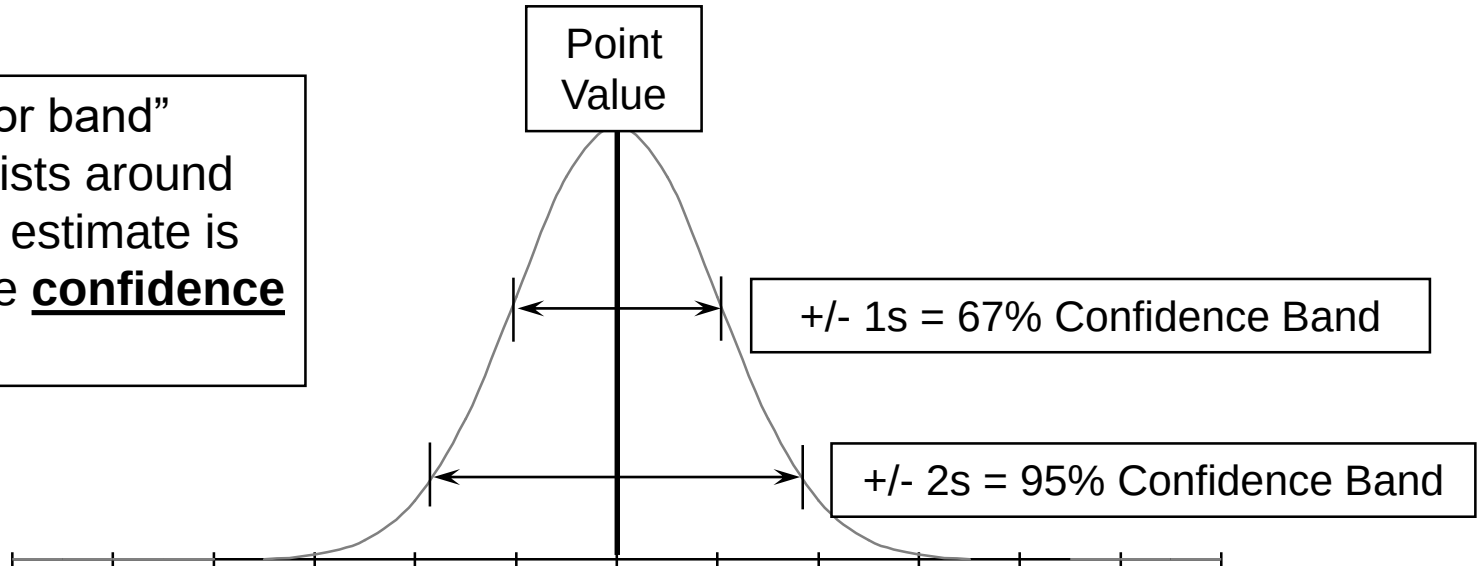
WHAT ARE CONFIDENCE INTERVALS?

The possibility of error exists in almost every system. This goes for point values as well. While we report a specific value, that value only represents our best estimate from the data at hand. The best way to think about this is to use the form:

true value = point estimate +/- error

The error around the point value follows one of several common probability distributions. As you have seen so far, we can increase our confidence is to go further and further out on the tails of this distribution.

This “error band” which exists around the point estimate is called the **confidence interval.**

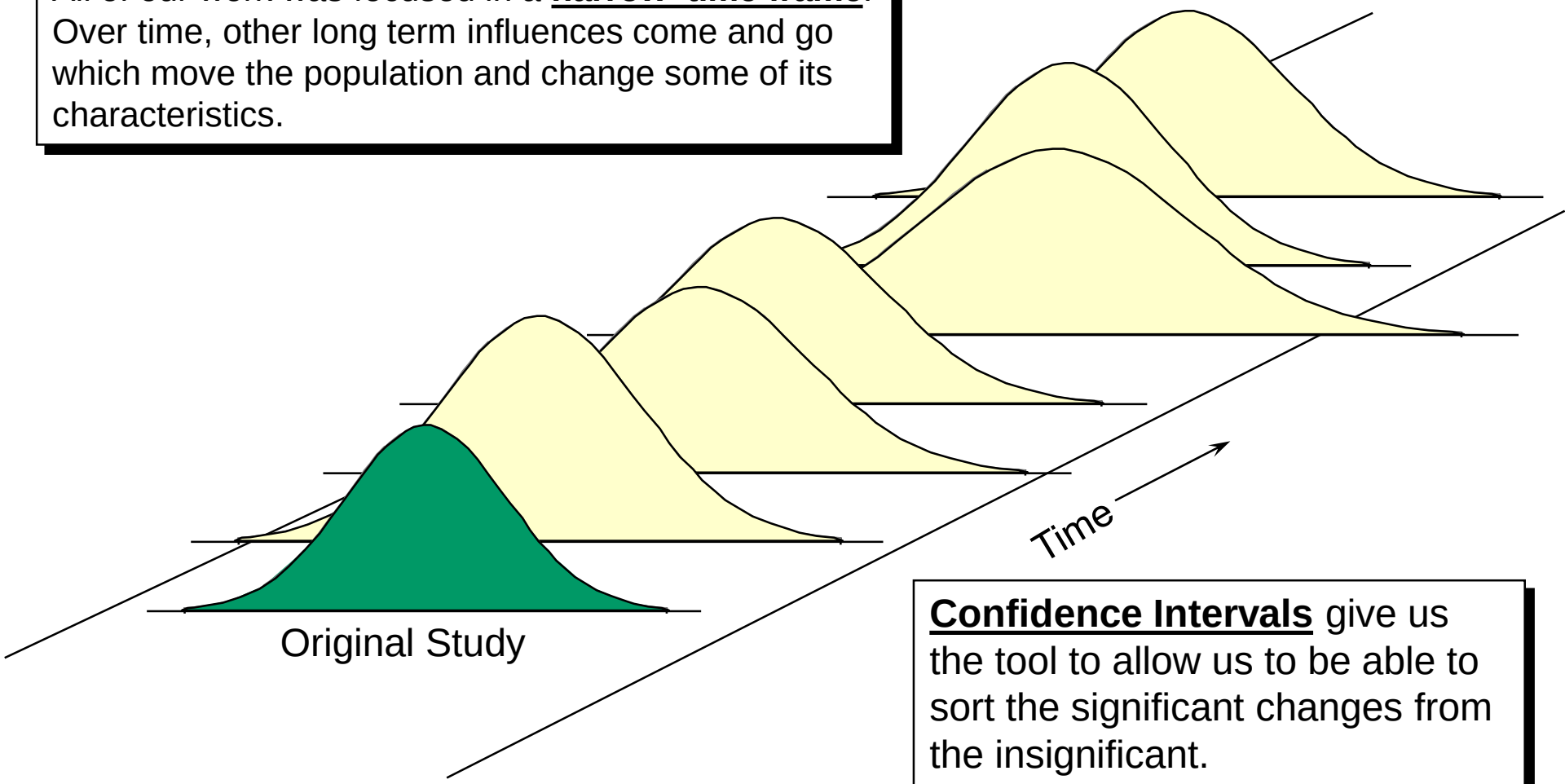


So, what does this do for me?

- The confidence interval establishes a way to test whether or not a significant change has occurred in the sampled population. This concept is called significance or hypothesis testing.
- Being able to tell when a significant change has occurred helps in preventing us from interpreting a significant change from a random event and responding accordingly.

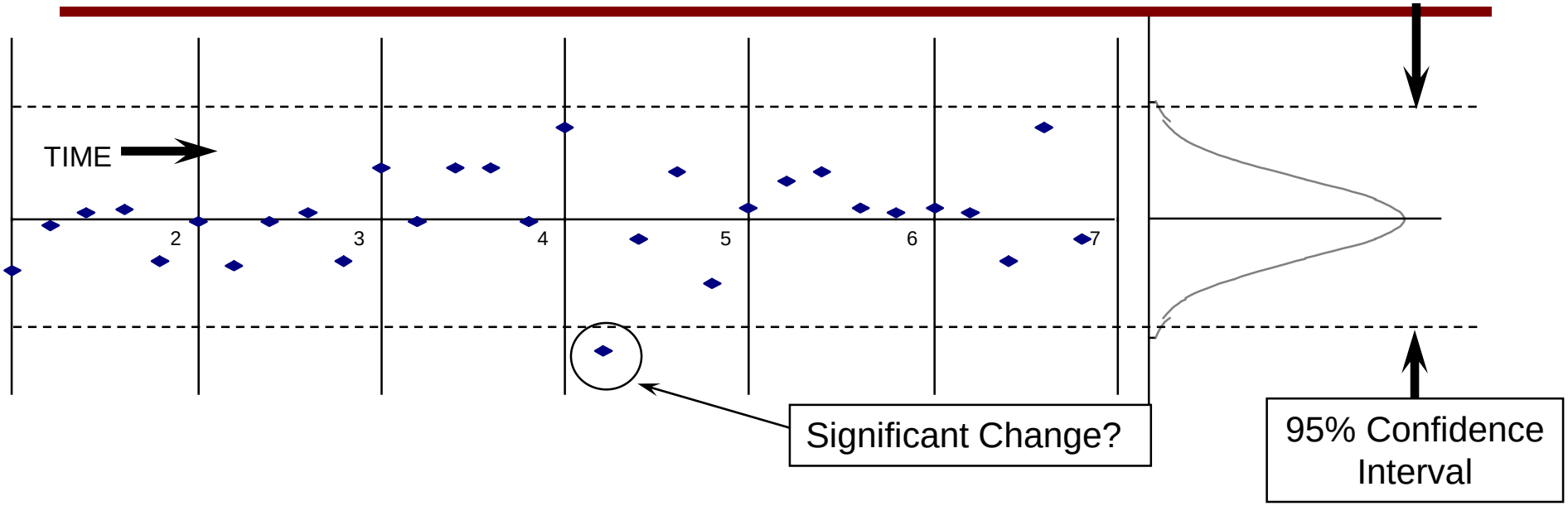
REMEMBER OUR OLD FRIEND SHIFT & DRIFT?

All of our work was focused in a **narrow time frame**. Over time, other long term influences come and go which move the population and change some of its characteristics.



Confidence Intervals give us the tool to allow us to be able to sort the significant changes from the insignificant.

USING CONFIDENCE INTERVALS TO SCREEN DATA



WHAT KIND OF PROBLEM DO YOU HAVE?

- Analysis for a significant change asks the question “What happened to make this significantly different from the rest?”
- Analysis for a series of random events focuses on the process and asks the question “What is designed into this process which causes it to have this characteristic?”.

CONFIDENCE INTERVAL FORMULAS

Mean

$$\overline{X} - t_{a/2,n-1}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{a/2,n-1}\frac{S}{\sqrt{n}}$$

Standard Deviation

$$s\sqrt{\frac{n-1}{\chi^2_{1-a/2}}} \le \sigma \le s\sqrt{\frac{n-1}{\chi^2_{a/2}}}$$

Process Capability

$$Cp\sqrt{\frac{\chi^2_{1-a/2,n-1}}{n-1}} \le Cp \le Cp\sqrt{\frac{\chi^2_{a/2,n-1}}{n-1}}$$

Percent Defective

$$\hat{p} - Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \overline{p} \le \hat{p} + Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

These individual formulas enable us to calculate the confidence interval for many of the most common metrics.

CONFIDENCE INTERVAL EXAMPLE

Over the past 6 months, we have received 10,000 parts from a vendor with an average defect rate of 14,000 dpm. The most recent batch of parts proved to have 23,000 dpm. Should we be concerned? We would like to be 95% confident in the results.

- **Calculation Values:**
  - **Average defect rate of 14,000 ppm = 14,000/1,000,000 = .014**
  - **Significance: α = 5% (95% confident)**
  - **$Z_{\alpha/2} = Z_{.025} = 1.96$**
  - **n=10,000**
  - **Comparison defect rate of 23,000 ppm = .023**
- **Calculation:**

$$\hat{p} - Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \bar{p} \le \hat{p} + Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$.014 - 1.96\sqrt{\frac{.014\,(1-.014)}{10{,}000}} \le \bar{p} \le .014 + 1.96\sqrt{\frac{.014\,(1-.014)}{10{,}000}} \qquad .014 - .0023 \le \bar{p} \le .014 + .0023$$

$$.012 \le \bar{p} \le .016$$

- **Answer: Yes, .023 is significantly outside of the expected 95% confidence interval of .012 to .016.**

CONFIDENCE INTERVAL EXERCISE

We are tracking the gas mileage of our late model ford and find that historically, we have averaged 28 MPG. After a tune up at Billy Bob's auto repair we find that we only got 24 MPG average with a standard deviation of 3 MPG in the next 16 fillups. Should we be concerned? We would like to be 95% confident in the results.

What do you think?

CONTROL CHARTS

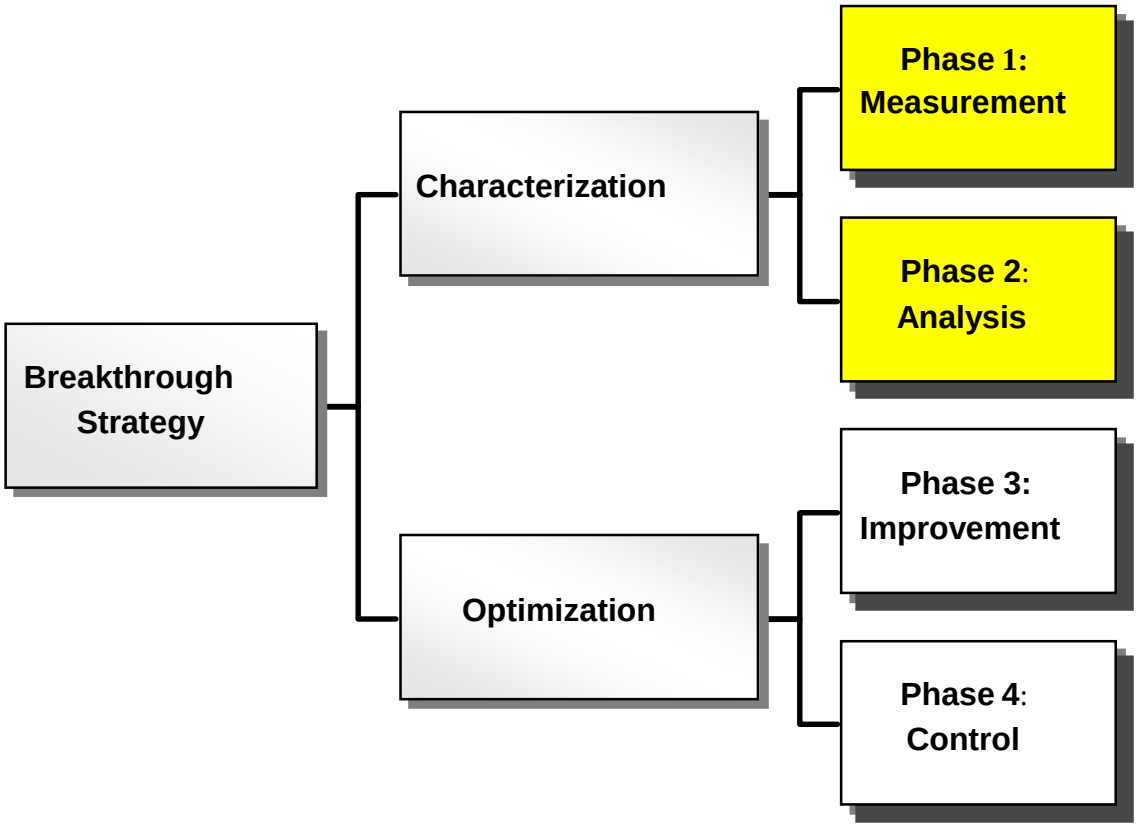
Why do we Care?

Control charts are useful to:

- determine the occurrence of “special cause” situations.
- Utilize the opportunities presented by “special cause” situations” to identify and correct the occurrence of the “special causes” .

IMPROVEMENT ROADMAP

Uses of Control Charts



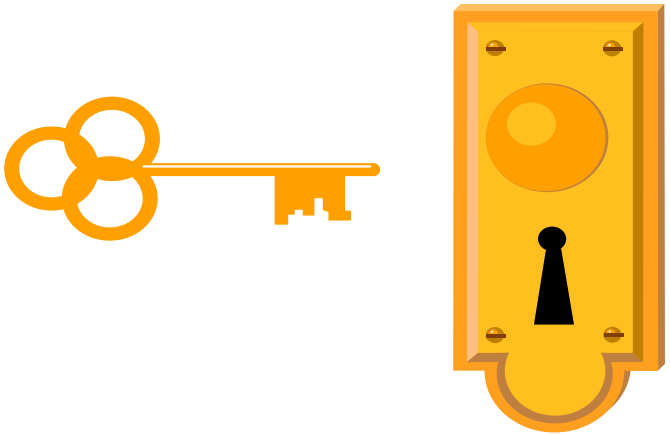
Common Uses

- Control charts can be effectively used to determine “special cause” situations in the Measurement and Analysis phases

KEYS TO SUCCESS

Use control charts on only a few critical output characteristics

Ensure that you have the means to investigate any “special cause”

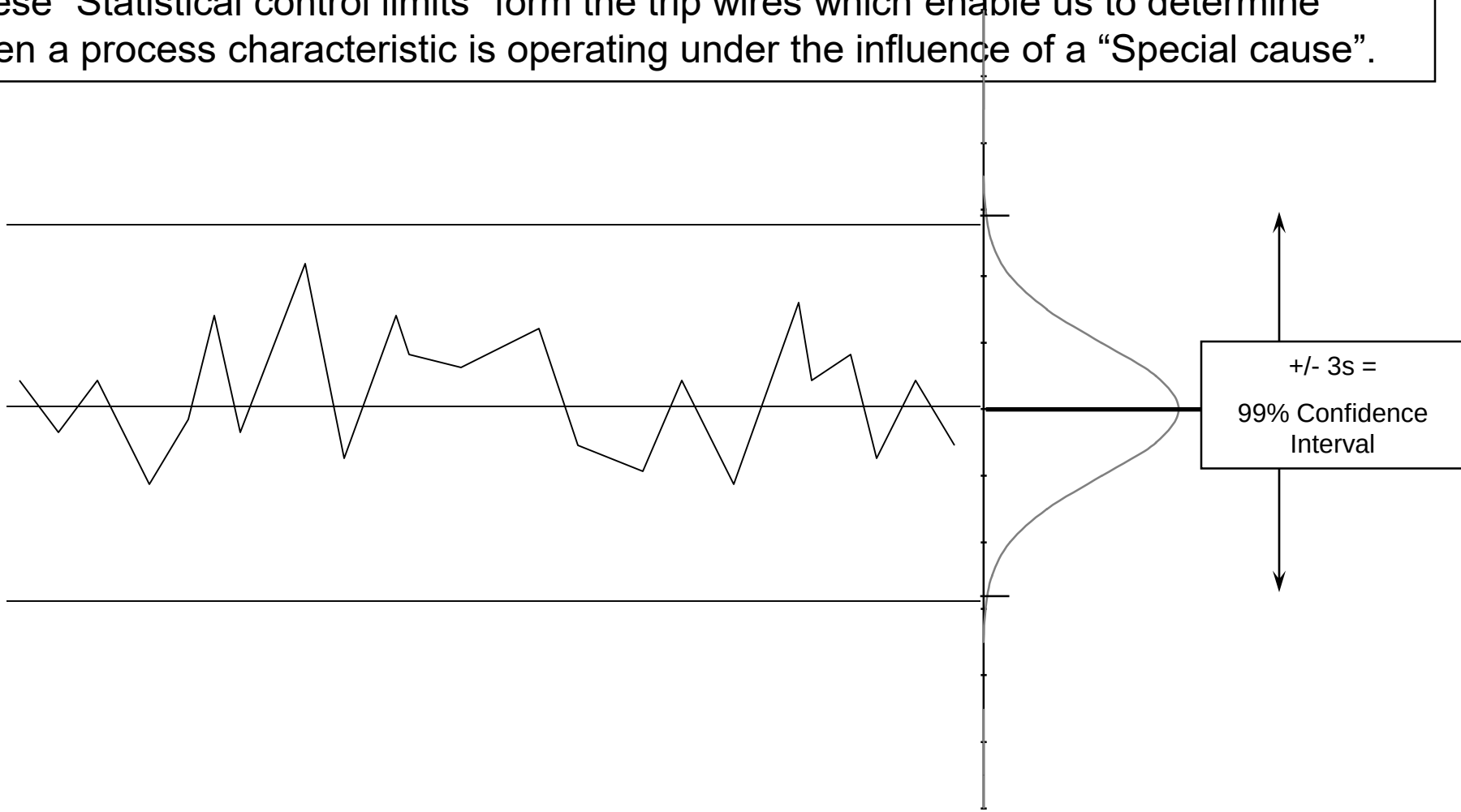


What is a “Special Cause”?

Remember our earlier work with confidence intervals? Any occurrence which falls outside the confidence interval has a low probability of occurring by random chance and therefore is “significantly different”. If we can identify and correct the cause, we have an opportunity to significantly improve the stability of the process. Due to the amount of data involved, control charts have historically used 99% confidence for determining the occurrence of these “special causes”

What is a Control Chart ?

A control chart is simply a run chart with confidence intervals calculated and drawn in. These “Statistical control limits” form the trip wires which enable us to determine when a process characteristic is operating under the influence of a “Special cause”.

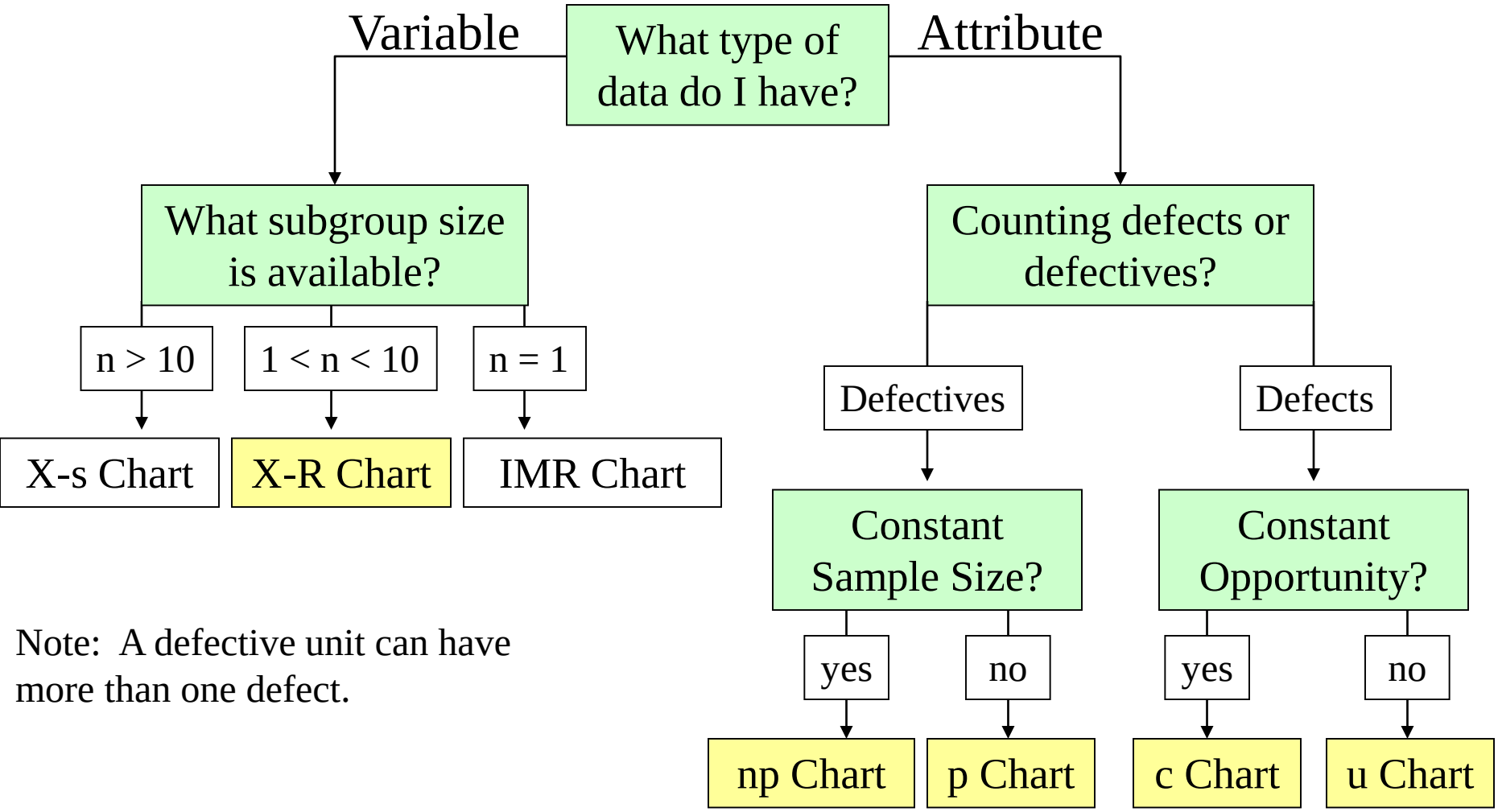


So how do I construct a control chart?

First things first:

- Select the metric to be evaluated
- Select the right control chart for the metric
- Gather enough data to calculate the control limits
- Plot the data on the chart
- Draw the control limits (UCL & LCL) onto the chart.
- Continue the run, investigating and correcting the cause of any “out of control” occurrence.

How do I select the correct chart ?



Note: A defective unit can have more than one defect.

How do I calculate the control limits?

$\overline{X}$ – *R Chart*

For the averages chart:

$$CL = \overline{\overline{X}}$$

$$UCL = \overline{\overline{X}} + A_2 \overline{R}$$

$$LCL = \overline{\overline{X}} - A_2 \overline{R}$$

For the range chart:

$$CL = \overline{R}$$

$$UCL = D_4 \overline{R}$$

$$LCL = D_3 \overline{R}$$

n	D4	D3	A2
2	3.27	0	1.88
3	2.57	0	1.02
4	2.28	0	0.73
5	2.11	0	0.58
6	2.00	0	0.48
7	1.92	0.08	0.42
8	1.86	0.14	0.37
9	1.82	0.18	0.34

$\overline{\overline{X}}$ = average of the subgroup averages

$\overline{R}$ = average of the subgroup range values

A_2 = a constant function of subgroup size (n)

UCL = upper control limit

LCL = lower control limit

How do I calculate the control limits?
p and np Charts

For varied sample size:

$$UCL_p = \overline{P} + 3\frac{\sqrt{\overline{P}\left(1-\overline{P}\right)}}{\sqrt{n}}$$

$$LCL_p = \overline{P} - 3\frac{\sqrt{\overline{P}\left(1-\overline{P}\right)}}{\sqrt{n}}$$

For constant sample size:

$$UCL_{np} = n\overline{P} + 3\sqrt{n\overline{P}\left(1-\overline{P}\right)}$$

$$LCL_{np} = n\overline{P} - 3\sqrt{n\overline{P}\left(1-\overline{P}\right)}$$

Note: P charts have an individually calculated control limit for each point plotted

P = number of rejects in the subgroup/number inspected in subgroup

$\overline{P}$ = total number of rejects/total number inspected

n = number inspected in subgroup

How do I calculate the control limits? c and u Charts

For varied opportunity (u):

$$UCL_u = \bar{U} + 3\frac{\sqrt{\bar{U}}}{\sqrt{n}}$$

$$LCL_u = \bar{U} - 3\frac{\sqrt{\bar{U}}}{\sqrt{n}}$$

For constant opportunity (c):

$$UCL_C = \bar{C} + 3\sqrt{\bar{C}}$$

$$LCL_C = \bar{C} - 3\sqrt{\bar{C}}$$

Note: U charts have an individually calculated control limit for each point plotted

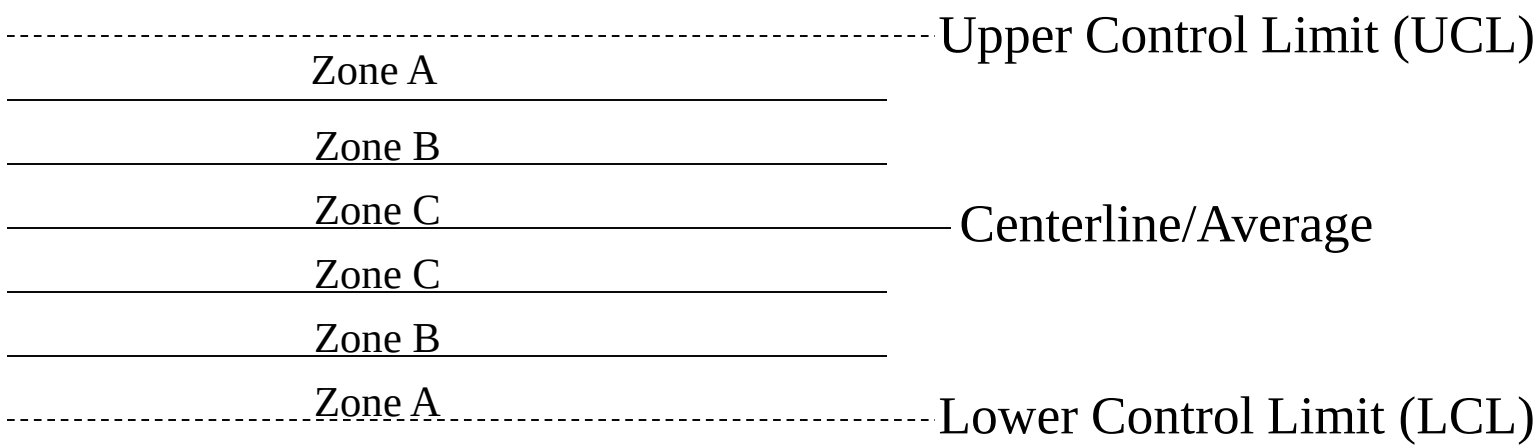
$\bar{C}$ = total number of nonconformities/total number of subgroups

$\bar{U}$ = total number of nonconformities/total units evaluated

n = number evaluated in subgroup

How do I interpret the charts?

- **The process is said to be "out of control" if:**
  - **One or more points fall outside of the control limits**
  - **When you divide the chart into zones as shown and:**
    - **2 out of 3 points on the same side of the centerline in Zone A**
    - **4 out of 5 points on the same side of the centerline in Zone A or B**
    - **9 successive points on one side of the centerline**
    - **6 successive points successively increasing or decreasing**
    - **14 points successively alternating up and down**
    - **15 points in a row within Zone C (above and/or below centerline)**



What do I do when it's "out of control"?

Time to Find and Fix the cause

- Look for patterns in the data
- Analyze the "out of control" occurrence
- Fishbone diagrams and Hypothesis tests are valuable "discovery" tools.

HYPOTHESIS TESTING

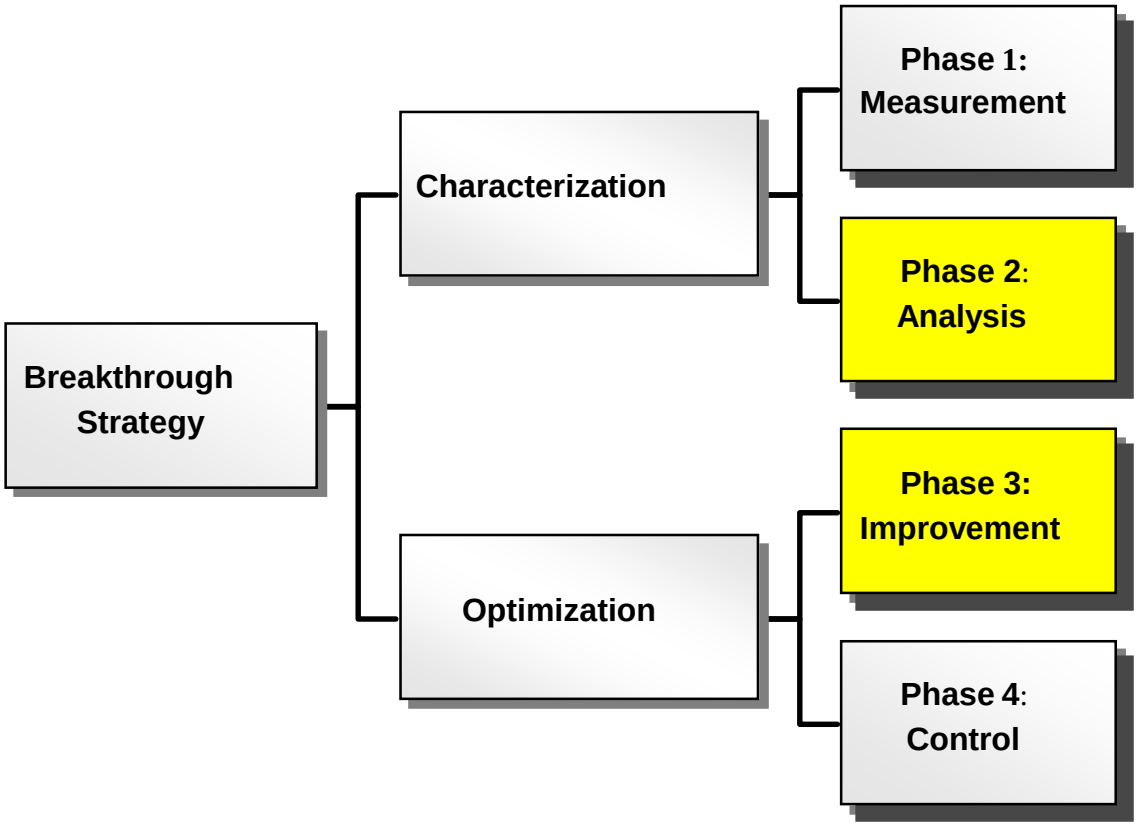
Why do we Care?

Hypothesis testing is necessary to:

- determine when there is a significant difference between two sample populations.
- determine whether there is a significant difference between a sample population and a target value.

IMPROVEMENT ROADMAP

Uses of Hypothesis Testing



Common Uses

- **Confirm sources of variation to determine causative factors (x).**
- **Demonstrate a statistically significant difference between baseline data and data taken after improvements were implemented.**

KEYS TO SUCCESS

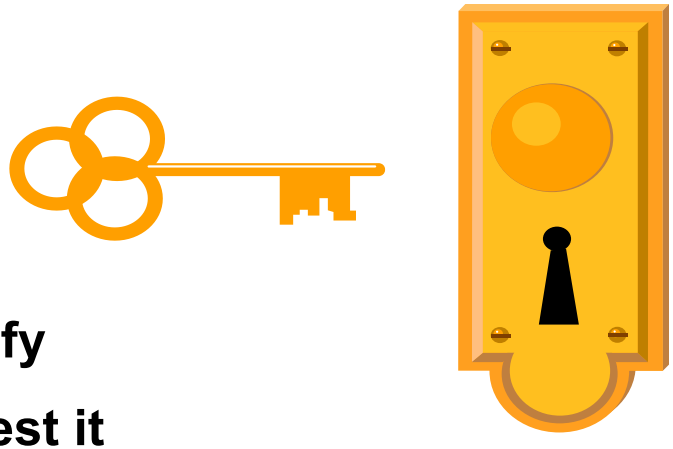
Use hypothesis testing to “explore” the data

Use existing data wherever possible

Use the team’s experience to direct the testing

Trust but verify….hypothesis testing is the verify

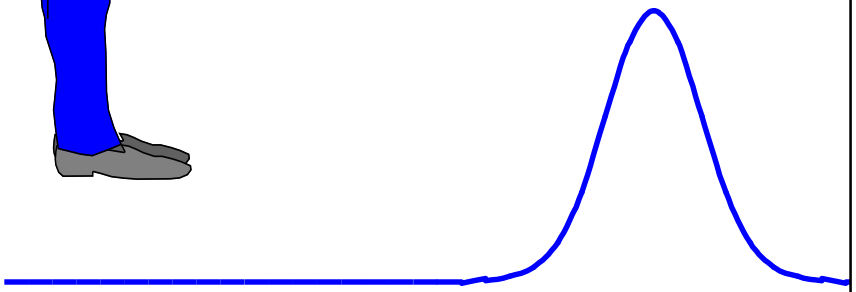
If there’s any doubt, find a way to hypothesis test it



SO WHAT IS HYPOTHESIS TESTING?

The theory of probability is nothing more than good sense confirmed by calculation.

Laplace

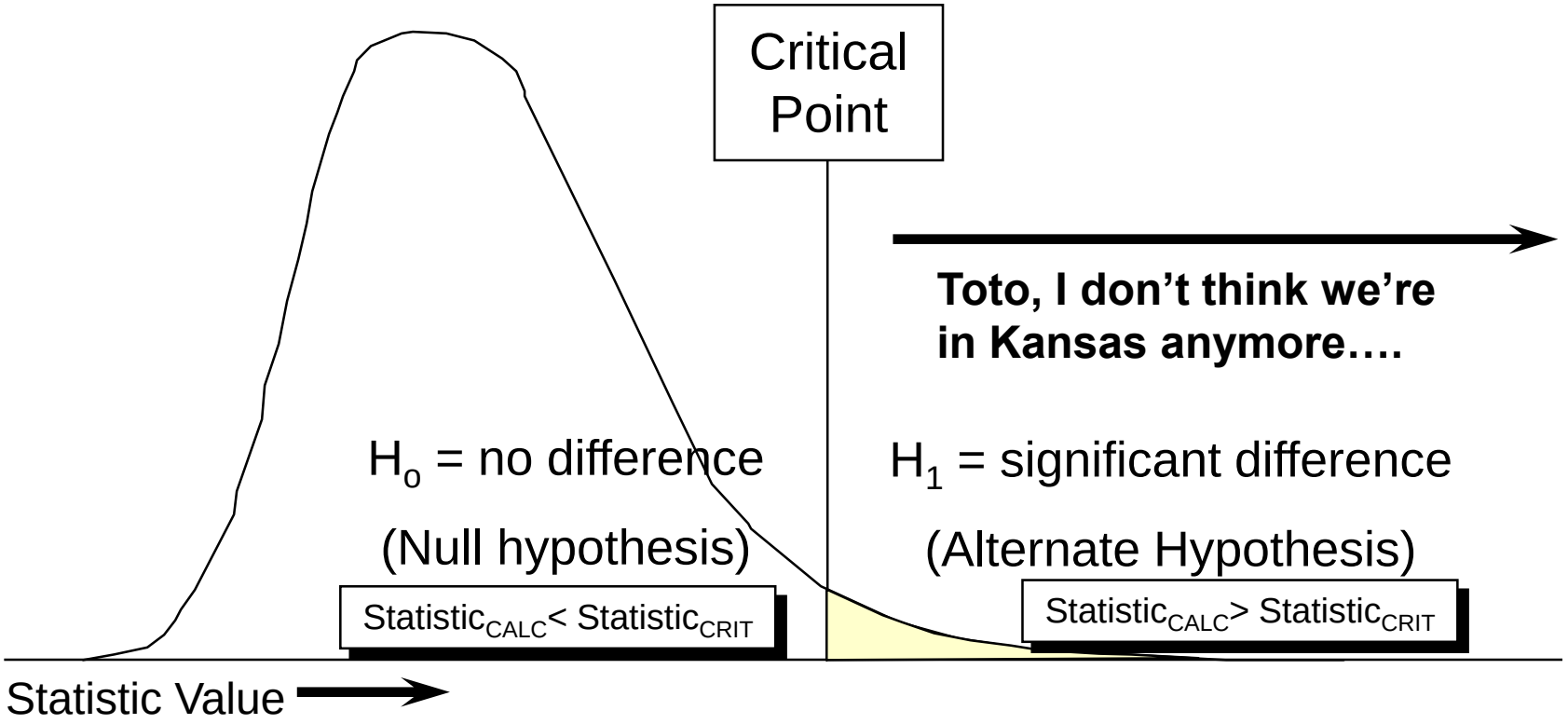


We think we see something…well, we think…err maybe it is… could be….

But, how do we know for sure?

Hypothesis testing is the key by giving us a measure of how confident we can be in our decision.

SO HOW DOES THIS HYPOTHESIS STUFF WORK?



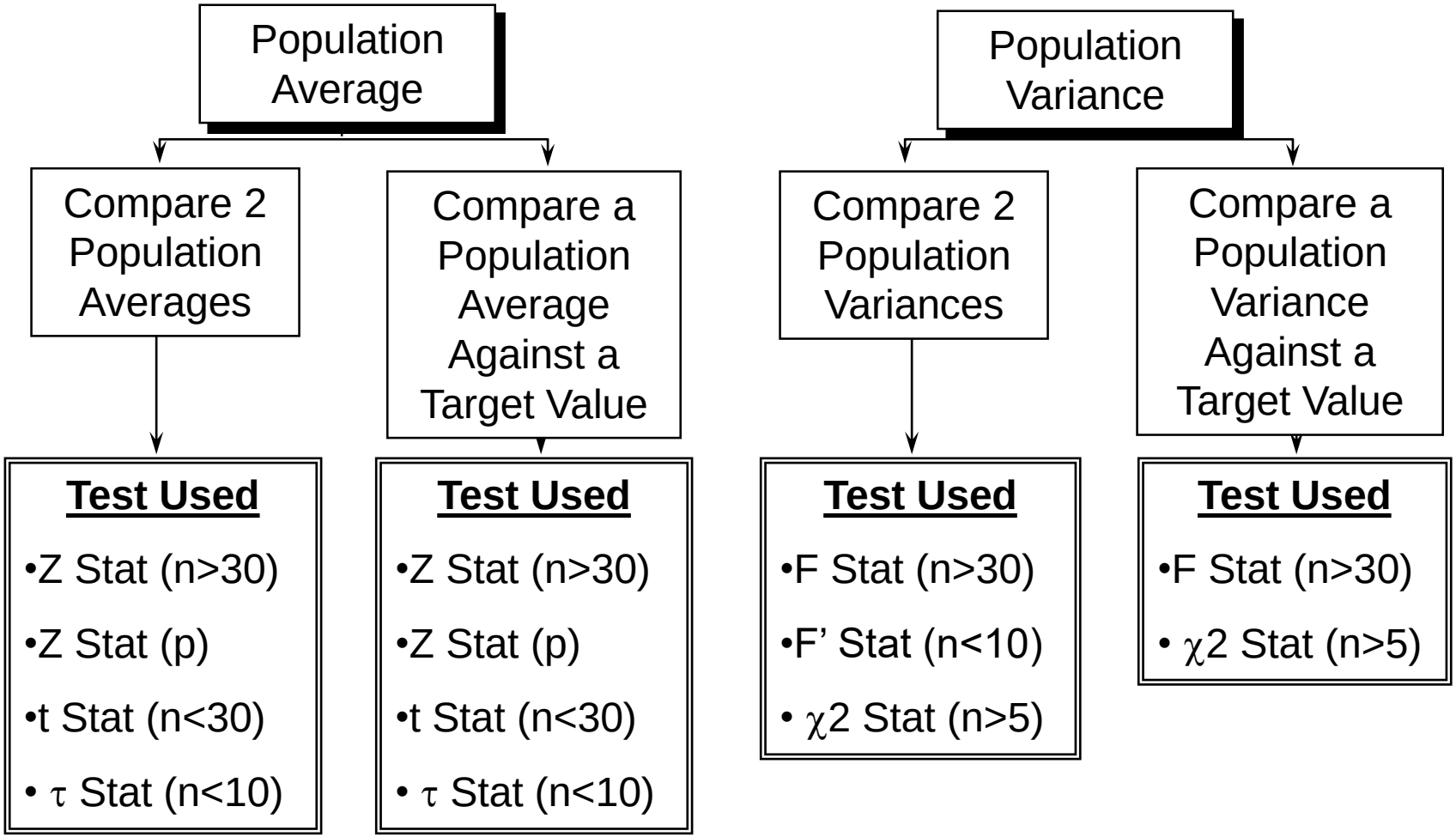
We determine a critical value from a probability table for the statistic. This value is compared with the calculated value we get from our data. If the calculated value exceeds the critical value, the probability of this occurrence happening due to random variation is less than our test α.

SO WHAT IS THIS ‘NULL” HYPOTHESIS?

Mathematicians are like Frenchmen, whatever you say to them they translate into their own language and forth with it is something entirely different. Goethe

Hypothesis	Symbol	How you say it	What it means
Null	**H_o**	**Fail to Reject the Null Hypothesis**	**Data does not support conclusion that there is a significant difference**
Alternative	**H_1**	**Reject the Null Hypothesis**	**Data supports conclusion that there is a significant difference**

HYPOTHESIS TESTING ROADMAP...



HYPOTHESIS TESTING PROCEDURE

- **Determine the hypothesis to be tested (Ho:=, < or >).**
- **Determine whether this is a 1 tail (α) or 2 tail ($\alpha/2$) test.**
- **Determine the α risk for the test (typically .05).**
- **Determine the appropriate test statistic.**
- **Determine the critical value from the appropriate test statistic table.**
- **Gather the data.**
- **Use the data to calculate the actual test statistic.**
- **Compare the calculated value with the critical value.**
- **If the calculated value is larger than the critical value, reject the null hypothesis with confidence of 1-α (ie there is little probability ($p<\alpha$) that this event occurred purely due to random chance) otherwise, accept the null hypothesis (this event occurred due to random chance).**

WHEN DO YOU USE α VS $\alpha/2$?

Many test statistics use α and others use $\alpha/2$ and often it is confusing to tell which one to use. The answer is straightforward when you consider what the test is trying to accomplish.

If there are two bounds (upper and lower), the α probability must be split between each bound. This results in a test statistic of $\alpha/2$.

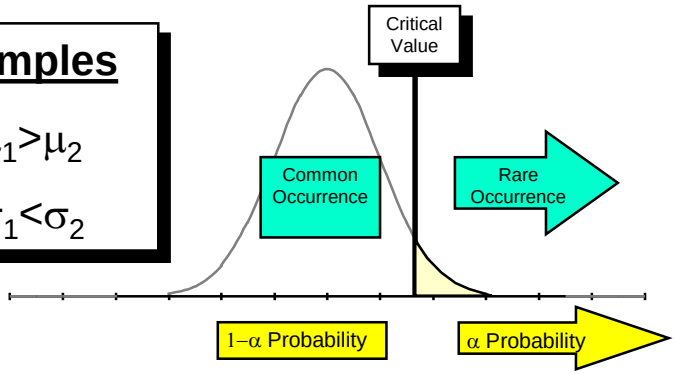
If there is only one direction which is under consideration, the error (probability) is concentrated in that direction. This results in a test statistic of α.

α

Examples

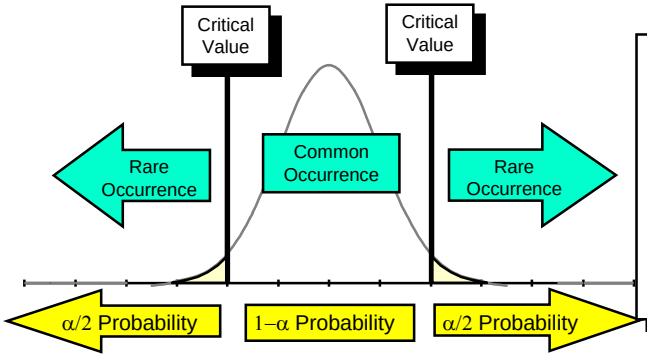
Ho: $\mu_1 > \mu_2$

Ho: $\sigma_1 < \sigma_2$



Test Fails in one direction = α

$\alpha/2$



Examples

Ho: $\mu_1 = \mu_2$

Ho: $\sigma_1 = \sigma_2$

Confidence Int

Test Fails in either direction = $\alpha/2$

Sample Average vs Sample Average

Coming up with the calculated statistic…

$n_1 = n_2 \leq 20$	$n_1 + n_2 \leq 30$	$n_1 + n_2 > 30$
2 Sample Tau	2 Sample t (DF: n_1+n_2-2)	2 Sample Z
$\tau_{dCALC} = \frac{2\left\lvert \overline{X_1} - \overline{X_2} \right\rvert}{R_1 + R_2}$	$t = \frac{\left\lvert \overline{X_1} - \overline{X_2} \right\rvert}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{[(n-1)s_1^2 + (n-1)s_2^2]}{n_1 + n_2 - 2}}}$	$Z_{CALC} = \frac{\left\lvert \overline{X_1} - \overline{X_2} \right\rvert}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_{21}^2}{n_2}}}$

Use these formulas to calculate the actual statistic for comparison with the critical (table) statistic. Note that the only major determinate here is the sample sizes. It should make sense to utilize the simpler tests (2 Sample Tau) wherever possible unless you have a statistical software package available or enjoy the challenge.

Hypothesis Testing Example (2 Sample Tau)

Several changes were made to the sales organization. The weekly number of orders were tracked both before and after the changes. Determine if the samples have equal means with 95% confidence.

- **Ho:** $\mu_1 = \mu_2$
- **Statistic Summary:**
  - $n_1 = n_2 = 5$
  - **Significance:** $\alpha/2$ **= .025 (2 tail)**
  - tau_{crit} **= .613 (From the table for** α **= .025 & n=5)**
- **Calculation:**
  - R_1**=337,** R_2**= 577**
  - X_1**=2868,** X_2**=2896**
  - tau_{CALC}**=2(2868-2896)/(337+577)=|.06|**

$$\tau_{dCALC} = \frac{2\left|\overline{X_1} - \overline{X_2}\right|}{R_1 + R_2}$$

Receipts 1	Receipts 2
3067	3200
2730	2777
2840	2623
2913	3044
2789	2834

- **Test:**
  - H_o: $tau_{CALC} < tau_{CRIT}$
  - H_o: .06<.613 = true? (yes, therefore we will fail to reject the null hypothesis).
- **Conclusion:** Fail to reject the null hypothesis (ie. The data does not support the conclusion that there is a significant difference)

Hypothesis Testing Example (2 Sample t)

Several changes were made to the sales organization. The weekly number of orders were tracked both before and after the changes. Determine if the samples have equal means with 95% confidence.

- **Ho: $\mu_1 = \mu_2$**
- **Statistic Summary:**
  - **$n_1 = n_2 = 5$**
  - **DF=$n_1 + n_2 - 2 = 8$**
  - **Significance: $\alpha/2$ = .025 (2 tail)**
  - **t_{crit} = 2.306 (From the table for α=.025 and 8 DF)**
- **Calculation:**
  - **s_1=130, s_2= 227**
  - **X_1=2868, X_2=2896**
  - **t_{CALC}=(2868-2896)/.63*185=|.24|**
- **Test:**
  - **H_o: $t_{CALC} < t_{CRIT}$**
  - **H_o: .24 < 2.306 = true? (yes, therefore we will fail to reject the null hypothesis).**
- **Conclusion:** **Fail to reject the null hypothesis (ie. The data does not support the conclusion that there is a significant difference**

Receipts 1	Receipts 2
3067	3200
2730	2777
2840	2623
2913	3044
2789	2834

$$t = \frac{\left|\overline{X_1} - \overline{X_2}\right|}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{[(n-1)s_1^2 + (n-1)s_2^2]}{n_1 + n_2 - 2}}}$$

Sample Average vs Target (μ_0)

Coming up with the calculated statistic…

$n \leq 20$	$n \leq 30$	$n > 30$
1 Sample Tau	1 Sample t (DF: n-1)	1 Sample Z
$\tau_{1CALC} = \frac{\lvert\overline{X} - \mu_0\rvert}{R}$	$t_{CALC} = \frac{\lvert\overline{X} - \mu_0\rvert}{\sqrt{\frac{s^2}{n}}}$	$Z_{CALC} = \frac{\lvert\overline{X} - \mu_0\rvert}{\sqrt{\frac{s^2}{n}}}$

Use these formulas to calculate the actual statistic for comparison with the critical (table) statistic. Note that the only major determinate again here is the sample size. Here again, it should make sense to utilize the simpler test (1 Sample Tau) wherever possible unless you have a statistical software package available (minitab) or enjoy the pain.

Sample Variance vs Sample Variance (s^2)

Coming up with the calculated statistic…

$n_1 < 10$, $n_2 < 10$	$n_1 > 30$, $n_2 > 30$
Range Test	F Test (DF_1: n_1-1, DF_2: n_2-1)
$F'_{CALC} = \frac{R_{MAX,n_1}}{R_{MIN,n_2}}$	$F_{calc} = \frac{s^2_{MAX}}{s^2_{MIN}}$

Use these formulas to calculate the actual statistic for comparison with the critical (table) statistic. Note that the only major determinate again here is the sample size. Here again, it should make sense to utilize the simpler test (Range Test) wherever possible unless you have a statistical software package available.

Hypothesis Testing Example (2 Sample Variance)

Several changes were made to the sales organization. The number of receipts was gathered both before and after the changes. Determine if the samples have equal variance with 95% confidence.

- **Ho: $s_1^2 = s_2^2$**
- **Statistic Summary:**
  - **$n_1 = n_2 = 5$**
  - **Significance: $\alpha/2$ = .025 (2 tail)**
  - **F'_{crit} = 3.25 (From the table for n_1, n_2=5)**
- **Calculation:**
  - **R_1=337, R_2= 577**
  - **F'_{CALC}=577/337=1.7**

$$F'_{CALC} = \frac{R_{MAX,n_1}}{R_{MIN,n_2}}$$

Receipts 1	Receipts 2
3067	3200
2730	2777
2840	2623
2913	3044
2789	2834

- **Test:**
  - **H_o: $F'_{CALC} < F'_{CRIT}$**
  - **H_o: 1.7 < 3.25 = true? (yes, therefore we will fail to reject the null hypothesis).**
- **Conclusion: Fail to reject the null hypothesis (ie. can't say there is a significant difference)**

HYPOTHESIS TESTING, PERCENT DEFECTIVE

$n_1 > 30$, $n_2 > 30$

Compare to target (p_0)	Compare two populations (p_1 & p_2)
$Z_{CALC} = \dfrac{\lvert p_1 - p_0 \rvert}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$Z_{CALC} = \dfrac{\lvert p_1 - p_2 \rvert}{\sqrt{\left(\dfrac{n_1 p_1 \_ n_2 p_2}{n_1 + n_2}\right)\left(1 - \dfrac{n_1 p_1 \_ n_2 p_2}{n_1 + n_2}\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

Use these formulas to calculate the actual statistic for comparison with the critical (table) statistic. Note that in both cases the individual samples should be greater than 30.

How about a manufacturing example?

We have a process which we have determined has a critical characteristic which has a target value of 2.53. Any deviation from this value will sub-optimize the resulting product. We want to sample the process to see how close we are to this value with 95% confidence. We gather 20 data points (shown below). Perform a 1 sample t test on the data to see how well we are doing.

2.342	2.749	2.480	3.119
2.187	2.332	1.503	2.808
3.036	2.227	1.891	1.468
2.666	1.858	2.316	2.124
2.814	1.974	2.475	2.470

ANalysis Of VAriance

ANOVA

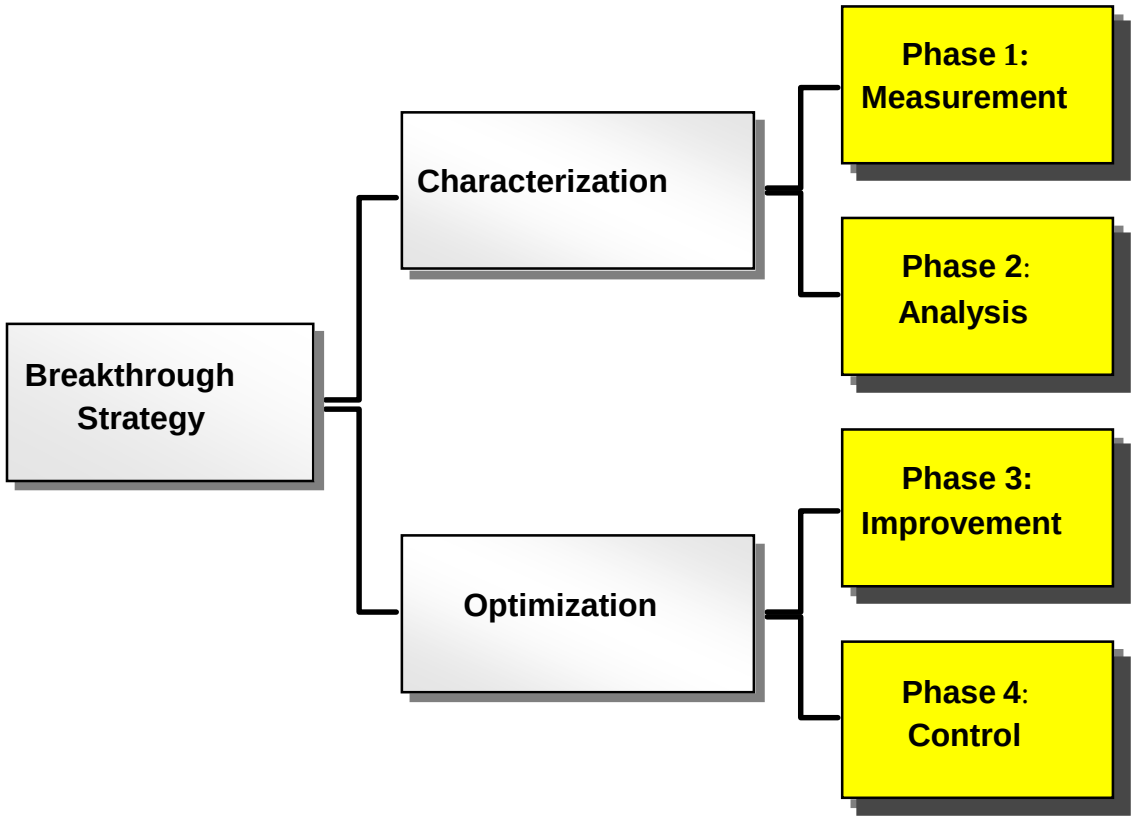
Why do we Care?

Anova is a powerful method for analyzing process variation:

- Used when comparing two or more process means.
- Estimate the relative effect of the input variables on the output variable.

IMPROVEMENT ROADMAP

Uses of Analysis of Variance Methodology---ANOVA



Common Uses

- **Hypothesis Testing**
- **Design of Experiments (DOE)**

KEYS TO SUCCESS

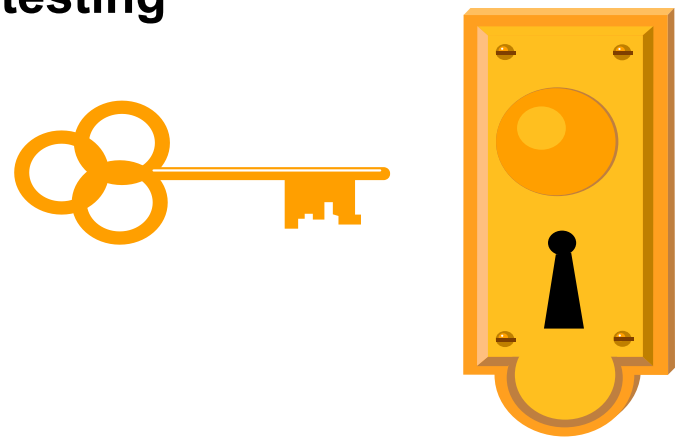
Don’t be afraid of the math

Make sure the assumptions of the method are met

Use validated data & your process experts to identify key variables.

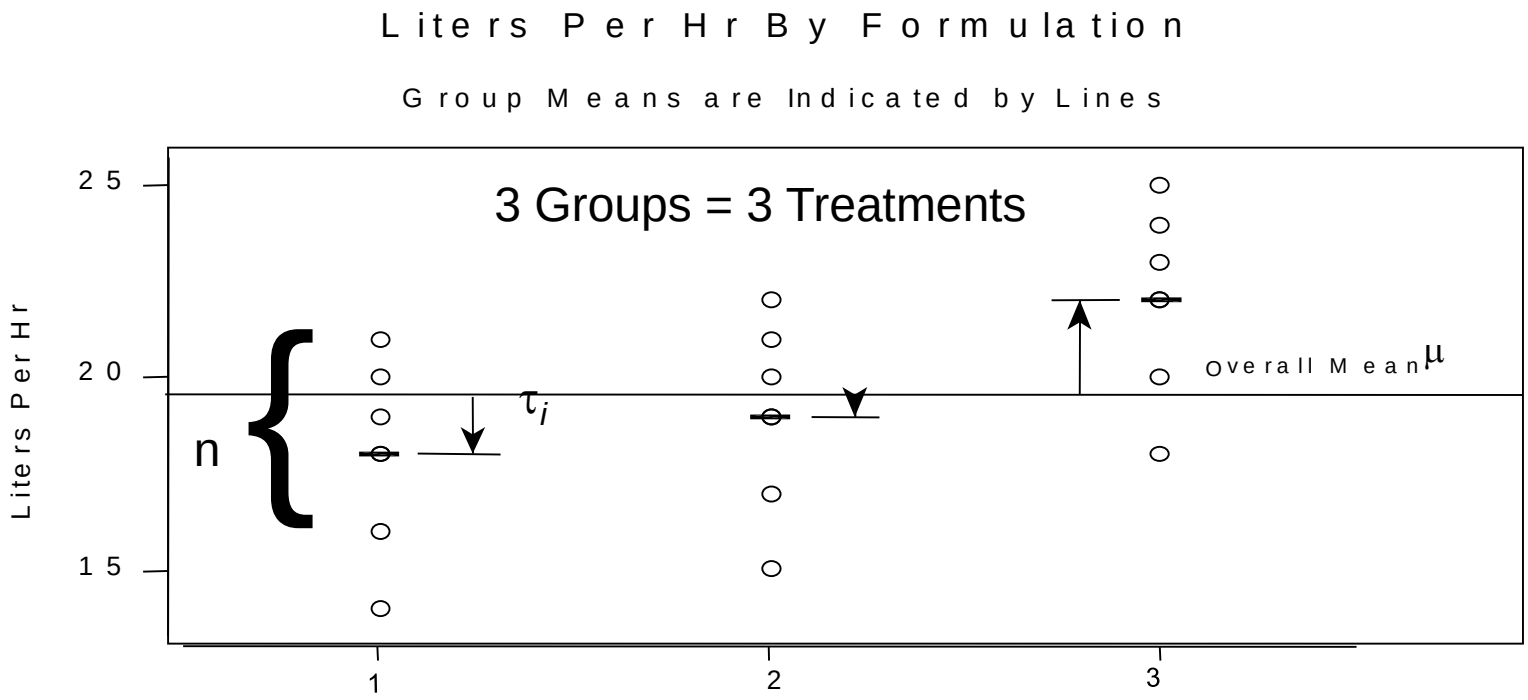
Carefully plan data collection and experiments

Use the team’s experience to direct the testing



Analysis Of Variance -----ANOVA

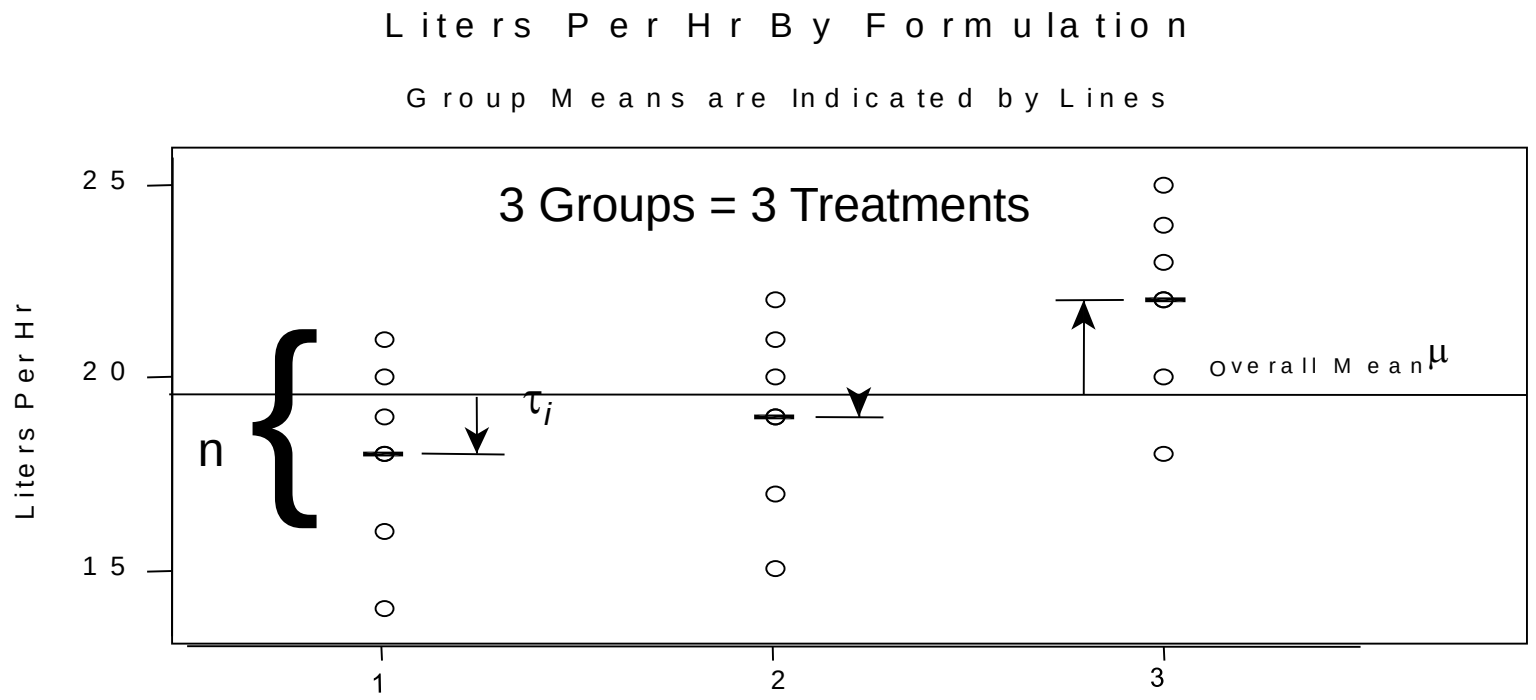
Linear Model



Let's say we run a study where we have three groups which we are evaluating for a significant difference in the means. Each one of these groups is called a "treatment" and represents one unique set of experimental conditions. Within each treatments, we have seven values which are called repetitions.

Analysis Of Variance -----ANOVA

Linear Model



The basic concept of ANOVA is to compare the variation between the treatments with the variation within each of the treatments. If the variation between the treatments is statistically significant when compared with the "noise" of the variation within the treatments, we can reject the null hypothesis.

One Way Anova

The first step in being able to perform this analysis is to compute the "Sum of the Squares" to determine the variation between treatments and within treatments.

Sum of Squares (SS):

$$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}\right)^2 = n\sum_{i=1}^{a}\left(\bar{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2$$

$$SS_{Total} = SS_{Treatments} + SS_{Error}$$

$SS_{Treatments}$ *(Between Treatments)*: **(Variation Between Treatments)**

SS_{Error} *(Within Treatments)*: **(Variation of Noise)**

Note: a = # of treatments (i)
n = # of repetitions within each treatment (j)

One Way Anova

Since we will be comparing two sources of variation, we will use the F test to determine whether there is a significant difference. To use the F test, we need to convert the "Sum of the Squares" to "Mean Squares" by dividing by the Degrees of Freedom (DF).

The Degrees of Freedom is different for each of our sources of variation.

$DF_{between}$ = # of treatments - 1

DF_{within} = (# of treatments)(# of repetitions of each treatment - 1)

DF_{total} = (# of treatments)(# of repetitions of each treatment) - 1

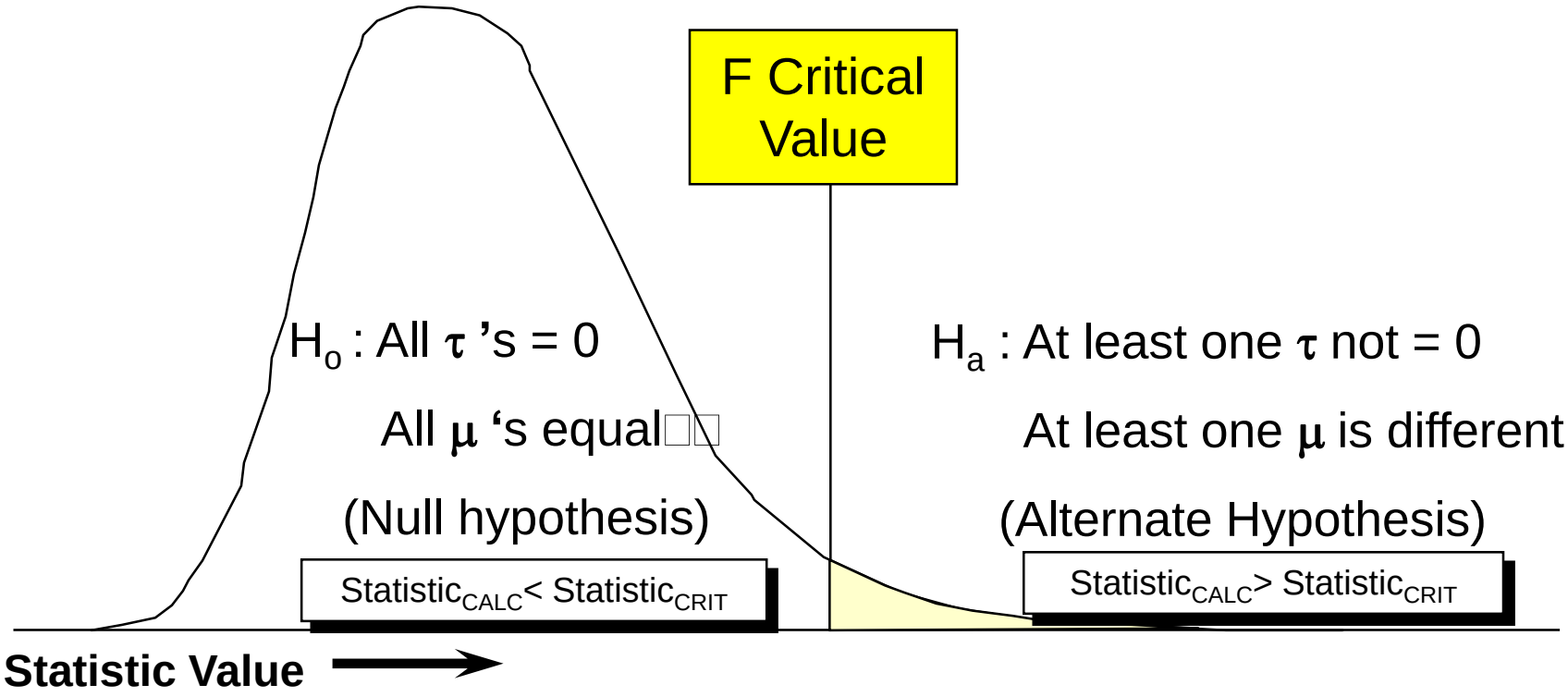
Note that $DF_{total} = DF_{between} + DF_{within}$

$$\text{Mean Square} = \frac{SS}{\text{Degrees of Freedom}}$$

$$F_{Calculated} = \frac{\text{Mean Square}_{\text{Between}}}{\text{Mean Square}_{\text{Within}}} = \frac{MS_{\text{Between}}}{MS_{\text{Within}}}$$

$$F_{\text{Critical}} \Rightarrow a, DF_{Between}, DF_{Within}$$

DETERMINING SIGNIFICANCE



We determine a critical value from a probability table. This value is compared with the calculated value. If the calculated value exceeds the critical value, we will reject the null hypothesis (concluding that a significant difference exists between 2 or more of the treatment means.

Class Example -- Evaluate 3 Fuel Formulations
Is there a difference?

Treatment 1	Treatment 2	Treatment 3
19	20	23
18	15	20
21	21	25
16	19	22
18	17	18
20	22	24
14	19	22

Here we have an example of three different fuel formulations that are expected to show a significant difference in average fuel consumption. Is there a significant difference? Let's use ANOVA to test our hypothesis……..

Class Example -- Evaluate 3 Fuel Formulations
Is there a difference?
Step 1: Calculating the Mean Squares Between

Step #1 = Calculate the Mean Squares Between Value

	Treatment 1	Treatment 2	Treatment 3	Treatment 4
	19	20	23	
	18	15	20	
	21	21	25	
	16	19	22	
	18	17	18	
	20	22	24	
	14	19	22	
Sum Each of the Columns	126	133	154	0
Find the Average of each Column (Ac)	18.0	19.0	22.0	
Calculate the Overall Average (Ao)	19.7	19.7	19.7	19.7
Find the Difference between the Average and Overall Average Squared $(Ac-Ao)^2$	2.8	0.4	5.4	0.0
Find the Sum of Squares Between by adding up the Differences Squared (SSb) and multiplying by the number of samples (replicates) within each treatment.	60.7			
Calculate the Degrees of Freedom Between (DFb=# treatments -1)	2			
Calculate the Mean Squares Between (SSb/DFb)	30.3			

The first step is to come up with the Mean Squares Between. To accomplish this we:

- find the average of each treatment and the overall average
- find the difference between the two values for each treatment and square it
- sum the squares and multiply by the number of replicates in each treatment
- divide this resulting value by the degrees of freedom

Class Example -- Evaluate 3 Fuel Formulations
Is there a difference?
Step 2: Calculating the Mean Squares Within

Step #2 = Calculate the Mean Squares Within Value

Treatment #	Data (Xi)	Treatment Average (Ao)	Difference (Xi-Ao)	Difference $(Xi-Ao)^2$
1	19	18.0	1.0	1.0
1	18	18.0	0.0	0.0
1	21	18.0	3.0	9.0
1	16	18.0	-2.0	4.0
1	18	18.0	0.0	0.0
1	20	18.0	2.0	4.0
1	14	18.0	-4.0	16.0
1	0	18.0	0.0	0.0
1	0	18.0	0.0	0.0
1	0	18.0	0.0	0.0
2	20	19.0	1.0	1.0
2	15	19.0	-4.0	16.0
2	21	19.0	2.0	4.0
2	19	19.0	0.0	0.0
2	17	19.0	-2.0	4.0
2	22	19.0	3.0	9.0
2	19	19.0	0.0	0.0
2	0	19.0	0.0	0.0
2	0	19.0	0.0	0.0
2	0	19.0	0.0	0.0
3	23	22.0	1.0	1.0
3	20	22.0	-2.0	4.0
3	25	22.0	3.0	9.0
3	22	22.0	0.0	0.0
3	18	22.0	-4.0	16.0
3	24	22.0	2.0	4.0
3	22	22.0	0.0	0.0
3	0	22.0	0.0	0.0
3	0	22.0	0.0	0.0
3	0	22.0	0.0	0.0
			Sum of Squares Within (SSw)	102.0
			Degrees of Freedom Within (DFw = (# of treatments)(samples -1)	18
			Mean Squares Within (MSw=SSw/DFw)	5.7

The next step is to come up with the Mean Squares Within. To accomplish this we:

- square the difference between each individual within a treatment and the average of that treatment
- sum the squares for each treatment
- divide this resulting value by the degrees of freedom

Class Example -- Evaluate 3 Fuel Formulations
Is there a difference?
Remaining Steps

Step #3 = Calculate the F value	
Fcalc = Mean Squares Between/Mean Squares Within (MSb/MSw)	5.35

Step #4 = Determine the critical value for F (α/2, DFb,DFw)	
Fcrit = Critical Value from the F table for a=.025, DFb=2, DFw=18	4.56

Step #5 = Compare the calculated F value to the critical value for F	
If Fcalc>Fcrit reject the null hypothesis (significant difference)	TRUE
If Fcalc<Fcrit fail to reject the null hypothesis (data does not support significant difference)	

The remaining steps to complete the analysis are:

- find the calculated F value by dividing the Mean Squares Between by the Mean Squares Within
- Determine the critical value from the F table
- Compare the calculated F value with the critical value determined from the table and draw our conclusions.

CONTINGENCY TABLES (CHI SQUARE)

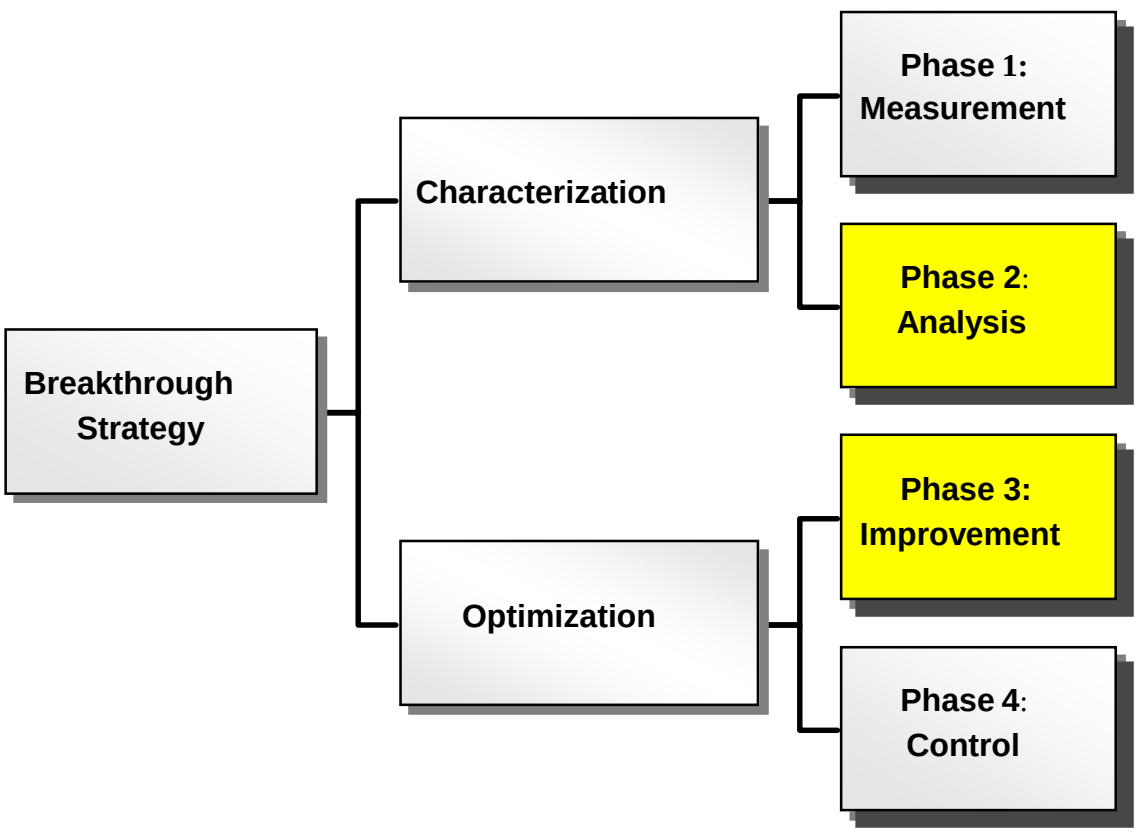
Why do we Care?

<u>Contingency tables are helpful to:</u>

- Perform statistical significance testing on count or attribute data.
- Allow comparison of more than one subset of data to help localize KPIV factors.

IMPROVEMENT ROADMAP

Uses of Contingency Tables



Common Uses

- Confirm sources of variation to determine causative factors (x).
- Demonstrate a statistically significant difference between baseline data and data taken after improvements were implemented.

KEYS TO SUCCESS

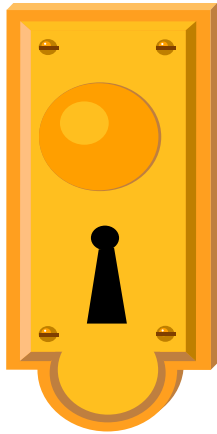
Conduct “ad hoc” training for your team prior to using the tool

Gather data, use the tool to test and then move on.

Use historical data where ever possible

Keep it simple

Ensure that there are a minimum of 5 units in each cell

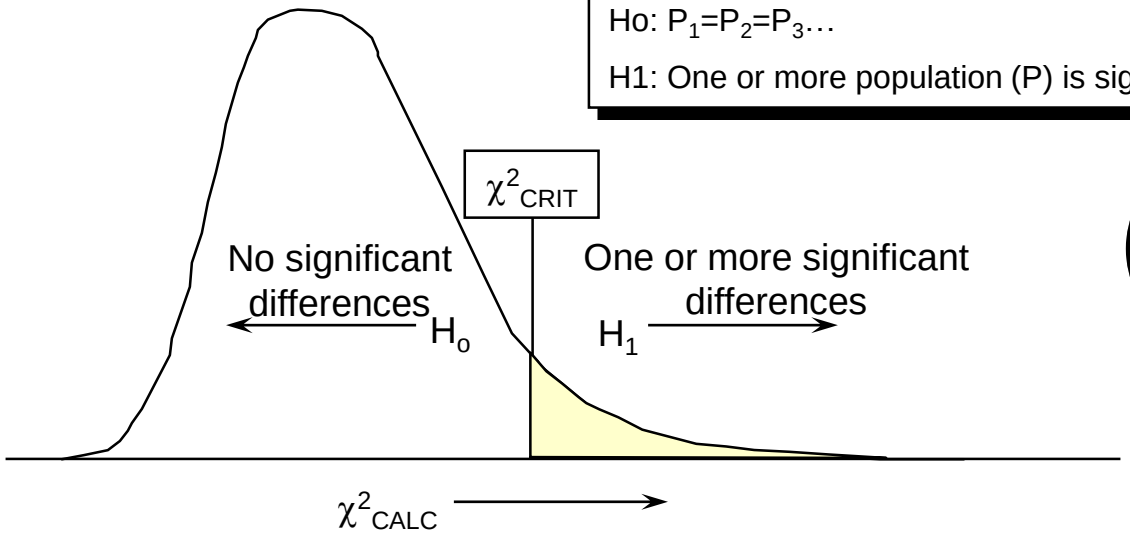


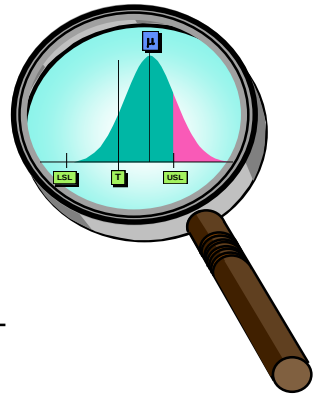
SO WHAT IS A CONTINGENCY TABLE?

A contingency table is just another way of hypothesis testing. Just like the hypothesis testing we have learned so far, we will obtain a "critical value" from a table (χ^2 in this case) and use it as a tripwire for significance. We then use the sample data to calculate a χ^2_{CALC} value. Comparing this calculated value to our critical value, tells us whether the data groups exhibit no significant difference (null hypothesis or H_o) or whether one or more significant differences exist (alternate hypothesis or H_1).

Ho: $P_1=P_2=P_3\ldots$

H1: One or more population (P) is significantly different





So how do you build a Contingency Table?

•Define the hypothesis you want to test. In this example we have 3 vendors from which we have historically purchased parts. We want to eliminate one and would like to see if there is a significant difference in delivery performance. This is stated mathematically as H_o: $P_a=P_b=P_c$. We want to be 95% confident in this result.

•We then construct a table which looks like this example. In this table we have order performance in rows and vendors in columns.

	Vendor A	Vendor B	Vendor C
Orders On Time			
Orders Late			

•We ensure that we include both good and bad situations so that the sum of each column is the total opportunities which have been grouped into good and bad. We then **gather enough data to ensure that each cell will have a count of at least 5**.

• Fill in the data.

	Vendor A	Vendor B	Vendor C
Orders On Time	25	58	12
Orders Late	7	9	5

•Add up the totals for each column and row.

	Vendor A	Vendor B	Vendor C	Total
Orders On Time	25	58	12	95
Orders Late	7	9	5	21
Total	32	67	17	116

Wait, what if I don't have at least 5 in each cell?

Collapse the table

•If we were placing side bets in a number of bars and wondered if there were any nonrandom factors at play. We gather data and construct the following table:

	Bar #1	Bar #2	Bar #3
Won Money	5	7	2
Lost Money	7	9	4

•Since bar #3 does not meet the "5 or more" criteria, we do not have enough data to evaluate that particular cell for Bar #3. This means that we must combine the data with that of another bar to ensure that we have significance. This is referred to as "collapsing" the table. The resulting collapsed table looks like the following:

Bar #3 Collapsed into Bar #2

	Bar #1	Bar #2&3
Won Money	5	9
Lost Money	7	13

•We can now proceed to evaluate our hypothesis. Note that the data between Bar #2 and Bar #3 will be aliased and therefore can not be evaluated separately.

So how do you build a Contingency Table?

•Calculate the percentage of the total contained in each row by dividing the row total by the total for all the categories. For example, the Orders on Time row has a total of 95. The overall total for all the categories is 116. The percentage for this row will be row/total or 95/116 = .82.

	Vendor A	Vendor B	Vendor C	Total	Portion
Orders On Time	25	58	12	95	0.82
Orders Late	7	9	5	21	0.18
Total	32	67	17	116	1.00

•.The row percentage times the column total gives us the expected occurrence for each cell based on the percentages. For example, for the Orders on time row, .82 x32=26 for the first cell and .82 x 67= 55 for the second cell.

	Actual Occurrences				
	Vendor A	Vendor B	Vendor C	Total	Portion
Orders On Time	25	58	12	95	0.82
Orders Late	7	9	5	21	0.18
Total	32	67	17	116	1.00

	Expected Occurrences		
	Vendor A	Vendor B	Vendor C
Orders On Time	26	55	
Orders Late			

So how do you build a Contingency Table?

•Complete the values for the expected occurrences.

	Actual Occurrences				
	Vendor A	Vendor B	Vendor C	Total	Portion
Orders On Time	25	58	12	95	0.82
Orders Late	7	9	5	21	0.18
Column Total	32	67	17	116	1.00

	Expected Occurrences		
	Vendor A	Vendor B	Vendor C
Orders On Time	26	55	14
Orders Late	6	12	3

	Calculations (Expected)		
	Vendor A	Vendor B	Vendor C
Orders On Time	.82x32	.82x67	.82x17
Orders Late	.18x32	.18x67	.18x17

•Now we need to calculate the χ^2 value for the data. This is done using the formula $(a-e)^2/e$ (where a is the actual count and e is the expected count) for each cell. So, the χ^2 value for the first column would be $(25-26)^2/26=.04$. Filling in the remaining χ^2 values we get:

	Actual Occurrences				
	Vendor A	Vendor B	Vendor C	Total	Portion
Orders On Time	25	58	12	95	0.82
Orders Late	7	9	5	21	0.18
Column Total	32	67	17	116	1.00

	Expected Occurrences		
	Vendor A	Vendor B	Vendor C
Orders On Time	26	55	14
Orders Late	6	12	3

	Calculations ($\chi^2=(a-e)^2/e$)		
	Vendor A	Vendor B	Vendor C
Orders On Time	$(25-26)^2/26$	$(58-55)^2/55$	$(12-14)^2/14$
Orders Late	$(7-6)^2/6$	$(9-12)^2/12$	$(5-3)^2/3$

	Calculated χ^2 Values		
	Vendor A	Vendor B	Vendor C
Orders On Time	0.04	0.18	0.27
Orders Late	0.25	0.81	1.20

Now what do I do?

Performing the Analysis

•Determine the critical value from the χ^2 table. To get the value you need 3 pieces of data. The degrees of freedom are obtained by the following equation; DF=(r-1)x(c-1). In our case, we have 3 columns (c) and 2 rows (r) so our DF = (2-1)x(3-1)=1x2=2.

•The second piece of data is the risk. Since we are looking for .95 (95%) confidence (and α risk = 1 - confidence) we know the α risk will be .05.

•In the χ^2 table , we find that the critical value for α = .05 and 2 DF to be 5.99. Therefore, our χ^2_{CRIT} = 5.99

•Our calculated χ^2 value is the sum of the individual cell χ^2 values. For our example this is .04+.18+.27+.25+.81+1.20=2.75. Therefore, our χ^2_{CALC} = 2.75.

•We now have all the pieces to perform our test. Our Ho: is $\chi^2_{CALC} < \chi^2_{CRIT}$. Is this true? Our data shows 2.75<5.99, therefore <u>we fail to reject the null hypothesis that there is no significant difference between the vendor performance in this area.</u>

Contingency Table Exercise

We have a part which is experiencing high scrap. Your team thinks that since it is manufactured over 3 shifts and on 3 different machines, that the scrap could be caused (Y=f(x)) by an off shift workmanship issue or machine capability. Verify with 95% confidence whether either of these hypothesis is supported by the data.

Construct a contingency table of the data and interpret the results for each data set.

	Actual Occurrences		
	Machine 1	Machine 2	Machine 3
Good Parts	100	350	900
Scrap	15	18	23

	Actual Occurrences		
	Shift 1	Shift 2	Shift 3
Good Parts	500	400	450
Scrap	20	19	17

DESIGN OF EXPERIMENTS (DOE)

FUNDAMENTALS

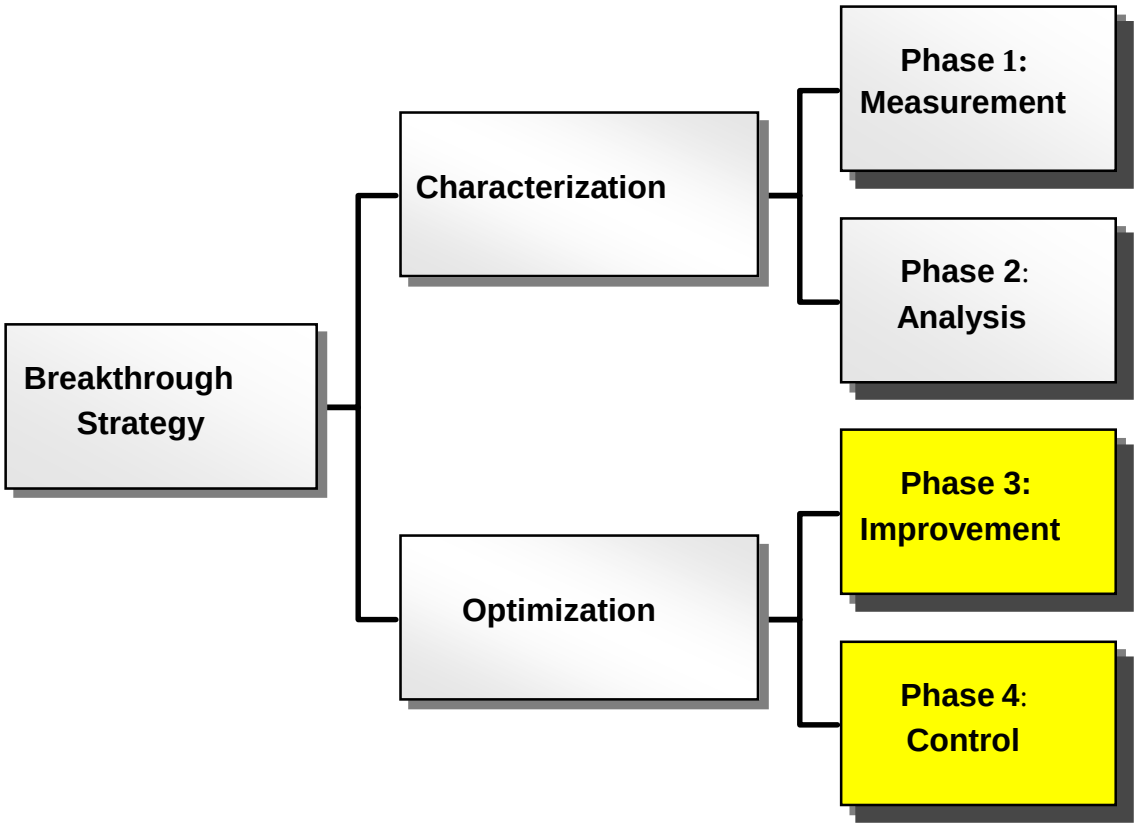
Why do we Care?

Design of Experiments is particularly useful to:

- evaluate interactions between 2 or more KPIVs and their impact on one or more KPOV's.
- optimize values for KPIVs to determine the optimum output from a process.

IMPROVEMENT ROADMAP

Uses of Design of Experiments



- **Verify the relationship between KPIV's and KPOV's by manipulating the KPIV and observing the corresponding KPOV change.**
- **Determine the best KPIV settings to optimize the KPOV output.**

KEYS TO SUCCESS

Keep it simple until you become comfortable with the tool

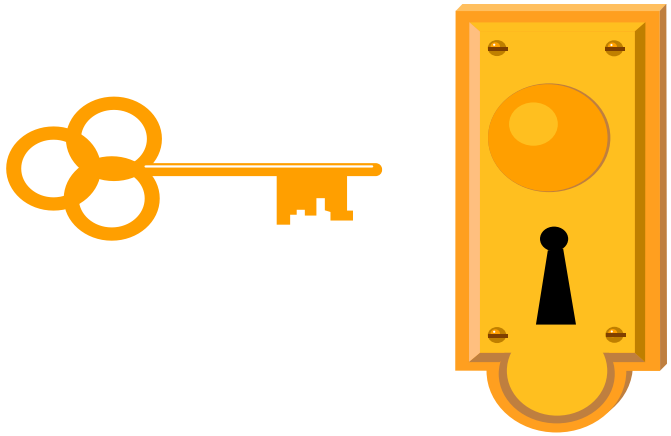
Statistical software helps tremendously with the calculations

Measurement system analysis should be completed on KPIV/KPOV(s)

Even the most clever analysis will not rescue a poorly planned experiment

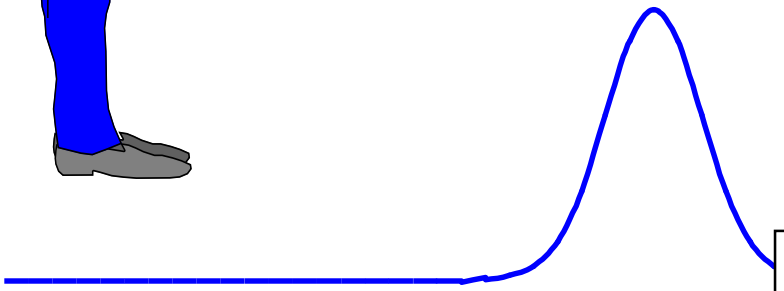
Don’t be afraid to ask for help until you are comfortable with the tool

Ensure a detailed test plan is written and followed



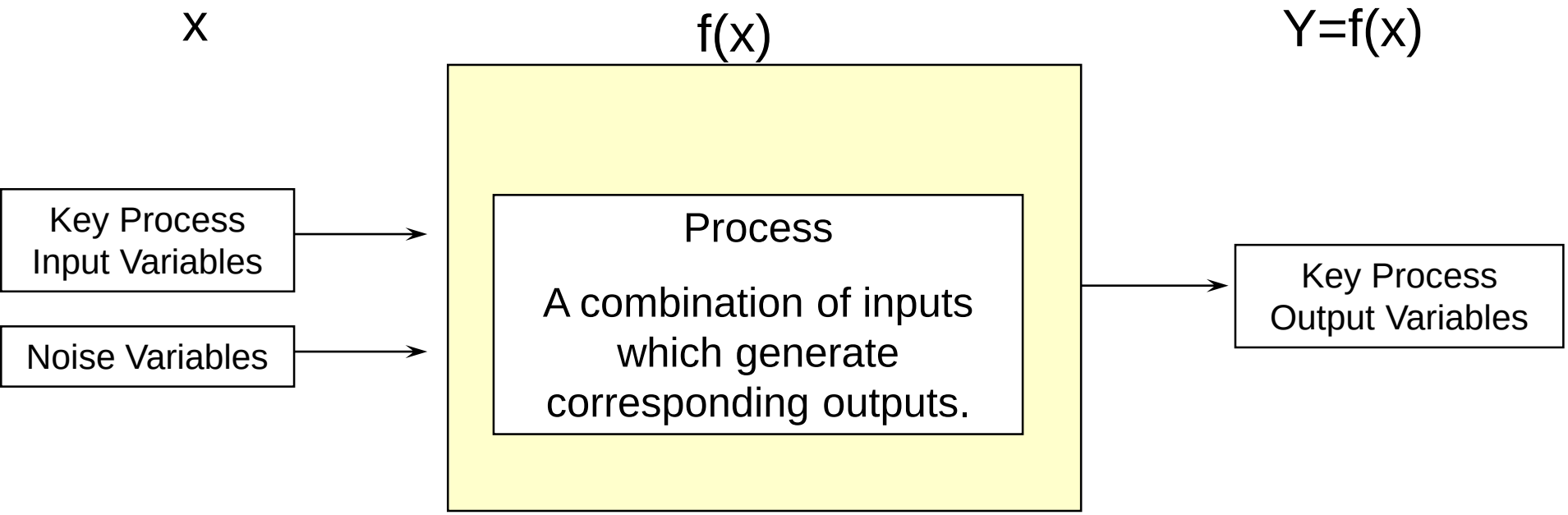
So What Is a Design of Experiment?

…where a mathematical reasoning can be had, it's as great a folly to make use of any other, as to grope for a thing in the dark, when you have a candle standing by you. Arbuthnot



A design of experiment introduces purposeful changes in KPIV's, so that we can methodically observe the corresponding response in the associated KPOV's.

Design of Experiments, Full Factorial

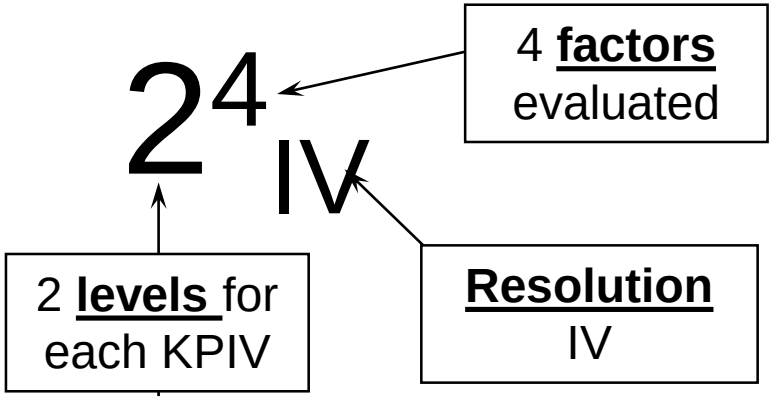


Variables

- Input, Controllable (KPIV)
- Input, Uncontrollable (Noise)
- Output, Controllable (KPOV)

How do you know how much a suspected KPIV actually influences a KPOV? You test it!

Design of Experiments, Terminology



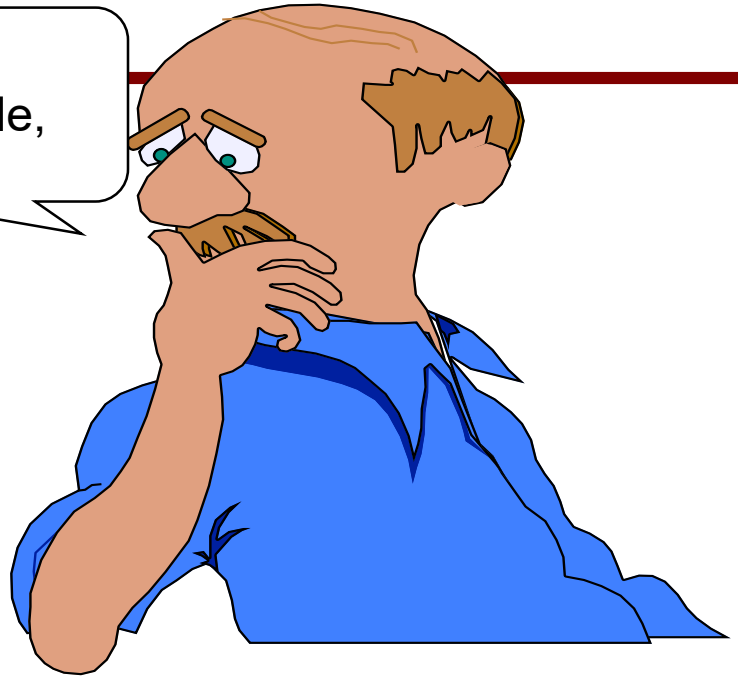
Mathematical objects are sometimes as peculiar as the most exotic beast or bird, and the time spent in examining them may be well employed.

H. Steinhaus

- **Main Effects** - Factors (KPIV) which directly impact output
- **Interactions** - Multiple factors which together have more impact on process output than any factor individually.
- **Factors** - Individual Key Process Input Variables (KPIV)
- **Levels** - Multiple conditions which a factor is set at for experimental purposes
- **Aliasing** - Degree to which an output cannot be clearly associated with an input condition due to test design.
- **Resolution** - Degree of aliasing in an experimental design

DOE Choices, A confusing array...

- Full Factorial
- Taguchi L16
- Half Fraction
- 2 level designs
- 3 level designs
- screening designs
- Response surface designs
- etc...



For the purposes of this training we will teach only full factorial (2^k) designs. This will enable you to get a basic understanding of application and use the tool. In addition, the vast majority of problems commonly encountered in improvement projects can be addressed with this design. If you have any question on whether the design is adequate, consult a statistical expert...

The Yates Algorithm

Determining the number of Treatments

2^3 *Factorial*

Treatment	A	B	C
1	+	+	+
2	+	+	-
3	+	-	+
4	+	-	-
5	-	+	+
6	-	+	-
7	-	-	+
8	-	-	-
	$2^3=8$	$2^2=4$	$2^1=2$	$2^0=1$

One aspect which is critical to the design is that they be "balanced". A balanced design has an equal number of levels represented for each KPIV. We can confirm this in the design on the right by adding up the number of + and - marks in each column. We see that in each case, they equal 4 + and 4- values, therefore the design is balanced.

- **Yates algorithm** is a quick and easy way (honest, trust me) to ensure that we get a balanced design whenever we are building a full factorial DOE. Notice that the number of treatments (unique test mixes of KPIVs) is equal to 2^3 or 8.
- Notice that in the "A factor" column, we have 4 + in a row and then 4 - in a row. This is equal to a group of 2^2 or 4. Also notice that the grouping in the next column is 2^1 or 2 + values and 2 - values repeated until all 8 treatments are accounted for.
- Repeat this pattern for the remaining factors.

The Yates Algorithm

Setting up the Algorithm for Interactions

2^3 *Factorial*

Treatment	A	B	C	AB	AC	BC	ABC
1	+	+	+	+	+	+	+
2	+	+	-	+	-	-	-
3	+	-	+	-	+	-	-
4	+	-	-	-	-	+	+
5	-	+	+	-	-	+	-
6	-	+	-	-	+	-	+
7	-	-	+	+	-	-	+
8	-	-	-	+	+	+	-

Now we can add the columns that reflect the interactions. Remember that the interactions are the main reason we use a DOE over a simple hypothesis test. The DOE is the best tool to study "mix" types of problems.

•You can see from the example above we have added additional columns for each of the ways that we can "mix" the 3 factors which are under study. These are our interactions. The sign that goes into the various treatment boxes for these interactions is the algebraic product of the main effects treatments. For example, treatment 7 for interaction AB is (- x - = +), so we put a plus in the box. So, in these calculations, the following apply:

minus (-) times minus (-) = plus (+)
minus (-) times plus (+) = minus (-)
plus (+) times plus (+) = plus (+)
plus (+) times minus (-) = minus (-)

Yates Algorithm Exercise

We work for a major “Donut & Coffee” chain. We have been tasked to determine what are the most significant factors in making “the most delicious coffee in the world”. In our work we have identified three factors we consider to be significant. These factors are coffee brand (maxwell house vs chock full o nuts), water (spring vs tap) and coffee amount (# of scoops).

Use the Yates algorithm to design the experiment.

So, How do I Conduct a DOE?

- **Select the factors (KPIVs) to be investigated and define the output to be measured (KPOV).**

- **Determine the 2 levels for each factor. Ensure that the levels are as widely spread apart as the process and circumstance allow.**

- **Draw up the design using the Yates algorithm.**

Treatment	A	B	C	AB	AC	BC	ABC
1	+	+	+	+	+	+	+
2	+	+	-	+	-	-	-
3	+	-	+	-	+	-	-
4	+	-	-	-	-	+	+
5	-	+	+	-	-	+	-
6	-	+	-	-	+	-	+
7	-	-	+	+	-	-	+
8	-	-	-	+	+	+	-

So, How do I Conduct a DOE?

- **Determine how many replications or repetitions you want to do. A replication is a complete new run of a treatment and a repetition is more than one sample run as part of a single treatment run.**

- **Randomize the order of the treatments and run each. Place the data for each treatment in a column to the right of your matrix.**

Treatment	A	B	C	AB	AC	BC	ABC	AVG	RUN1	RUN2	RUN3
1	+	+	+	+	+	+	+	18	18		
2	+	+	-	+	-	-	-	12	12		
3	+	-	+	-	+	-	-	6	6		
4	+	-	-	-	-	+	+	9	9		
5	-	+	+	-	-	+	-	3	3		
6	-	+	-	-	+	-	+	3	3		
7	-	-	+	+	-	-	+	4	4		
8	-	-	-	+	+	+	-	8	8		

Analysis of a DOE

- **Calculate the average output for each treatment.**
- **Place the average for each treatment after the sign (+ or -) in each cell.**

Treatment	A	B	C	AB	AC	BC	ABC	AVG	RUN1	RUN2	RUN3
1	+18	+	+	+	+	+	+	18	18		
2	+12	+	-	+	-	-	-	12	12		
3	+6	-	+	-	+	-	-	6	6		
4	+9	-	-	-	-	+	+	9	9		
5	-3	+	+	-	-	+	-	3	3		
6	-3	+	-	-	+	-	+	3	3		
7	-4	-	+	+	-	-	+	4	4		
8	-8	-	-	+	+	+	-	8	8		

Analysis of a DOE

- **Add up the values in each column and put the result under the appropriate column. This is the total estimated effect of the factor or combination of factors.**
- **Divide the total estimated effect of each column by 1/2 the total number of treatments. This is the average estimated effect.**

Treatment	A	B	C	AB	AC	BC	ABC	AVG
1	+18	+18	+18	+18	+18	+18	+18	18
2	+12	+12	-12	+12	-12	-12	-12	12
3	+6	-6	+6	-6	+6	-6	-6	6
4	+9	-9	-9	-9	-9	+9	+9	9
5	-3	+3	+3	-3	-3	+3	-3	3
6	-3	+3	-3	-3	+3	-3	+3	3
7	-4	-4	+4	+4	-4	-4	+4	4
8	-8	-8	-8	+8	+8	+8	-8	8
SUM	27	9	-1	21	7	13	5	63
AVG	6.75	2.25	-0.25	5.25	1.75	3.25	1.25	

Analysis of a DOE

- **These averages represent the average difference between the factor levels represented by the column. So, in the case of factor "A", the average difference in the result output between the + level and the - level is 6.75.**
- **We can now determine the factors (or combination of factors) which have the greatest impact on the output by looking for the magnitude of the respective averages (i.e., ignore the sign).**

Treatment	A	B	C	AB	AC	BC	ABC	AVG
1	+18	+18	+18	+18	+18	+18	+18	18
2	+12	+12	-12	+12	-12	-12	-12	12
3	+6	-6	+6	-6	+6	-6	-6	6
4	+9	-9	-9	-9	-9	+9	+9	9
5	-3	+3	+3	-3	-3	+3	-3	3
6	-3	+3	-3	-3	+3	-3	+3	3
7	-4	-4	+4	+4	-4	-4	+4	4
8	-8	-8	-8	+8	+8	+8	-8	8
SUM	27	9	-1	21	7	13	5	63
AVG	6.75	2.25	-0.25	5.25	1.75	3.25	1.25	

This means that the impact is in the following order:

A	(6.75)
AB	(5.25)
BC	(3.25)
B	(2.25)
AC	(1.75)
ABC	(1.25)
C	(-0.25)

Analysis of a DOE

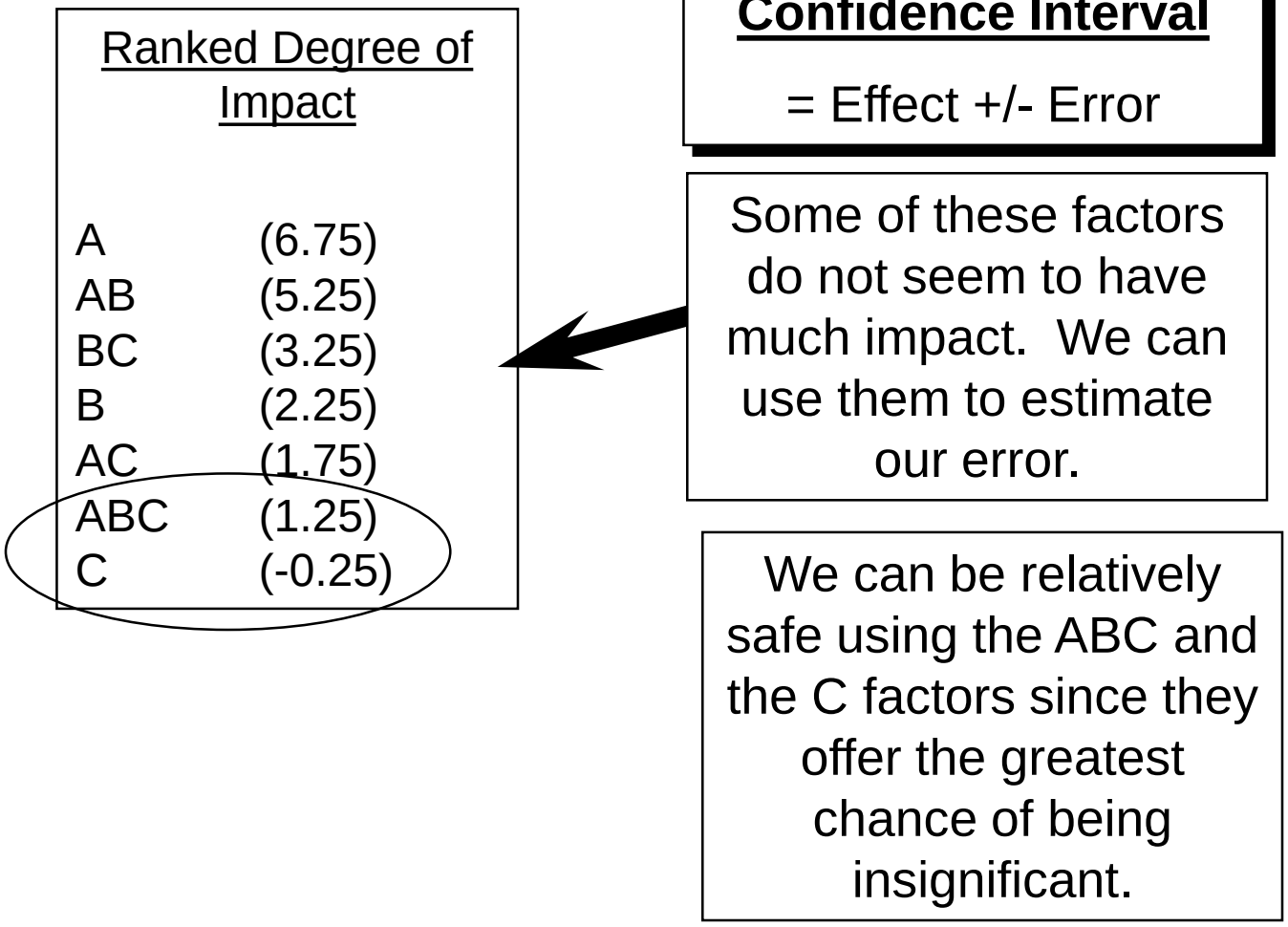
We can see the impact, but how do we know if these results are significant or just random variation?

Ranked Degree of Impact

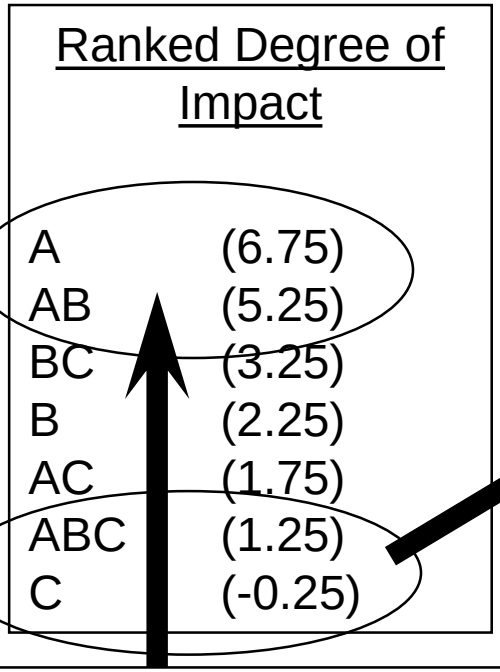
A	(6.75)
AB	(5.25)
BC	(3.25)
B	(2.25)
AC	(1.75)
ABC	(1.25)
C	(-0.25)

What tool do you think would be good to use in this situation?

Confidence Interval for DOE results



Confidence Interval for DOE results



$$Confidence = \pm t_{\alpha/2,DF}\sqrt{\frac{\sum\left(ABC^2 + C^2\right)}{DF}}$$

DF=# of groups used

In this case we are using 2 groups (ABC and C) so our DF=2

For α = .05 and DF =2 we find $t_{\alpha/2,df} = t_{.025,2} = 4.303$

$$Confidence = \pm 4.303\sqrt{\frac{\sum\left(1.25^2 + (-.25)^2\right)}{2}}$$

$$Confidence = \pm(4.303)(.9235)$$

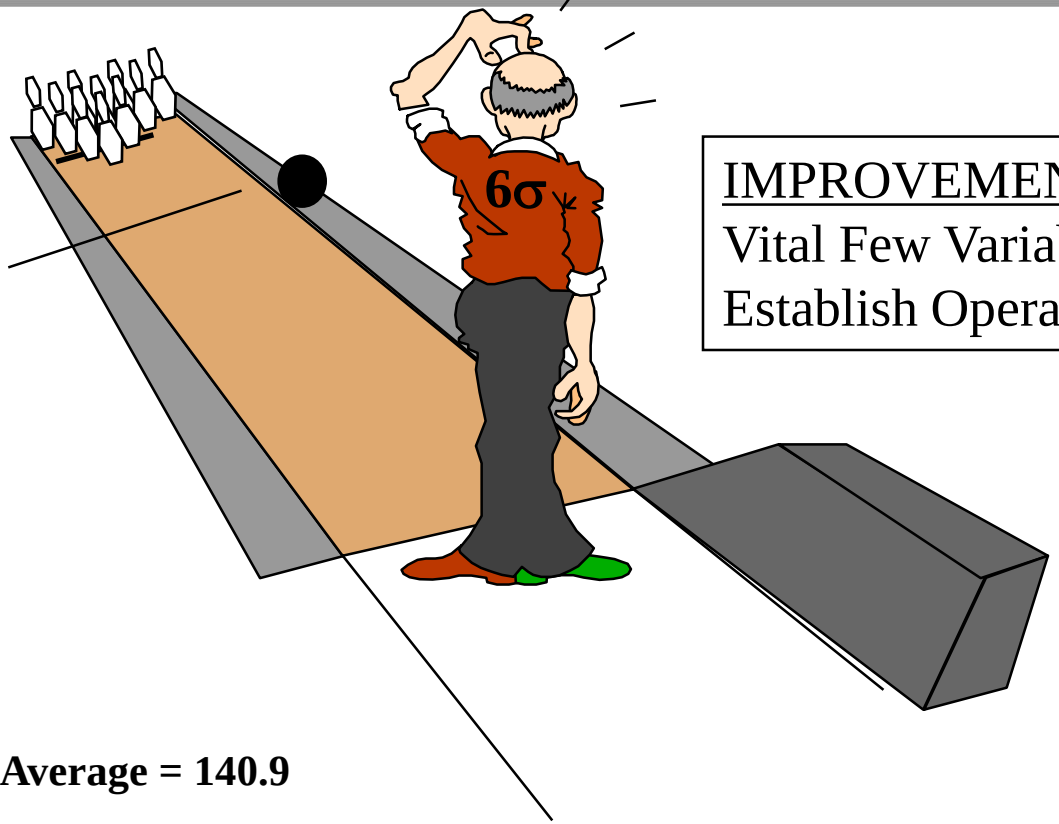
$$Confidence = \pm 3.97$$

Confidence

Since only 2 groups meet or exceed our 95% confidence interval of +/- 3.97. We conclude that they are the only significant treatments.

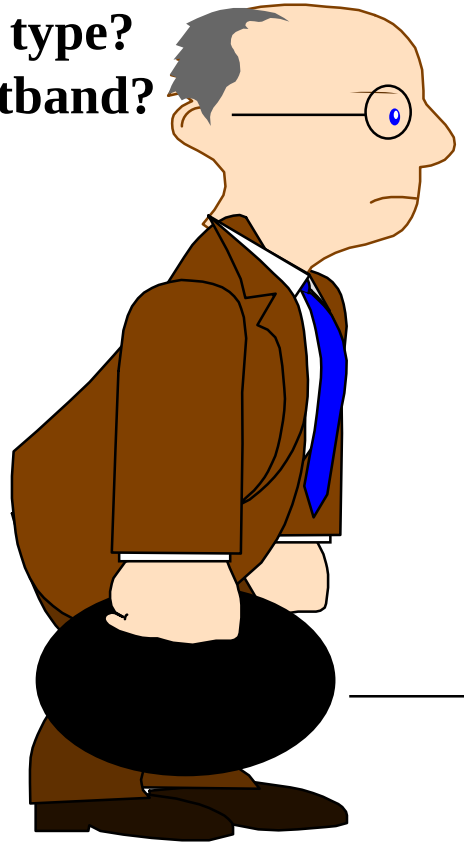
How about another way of looking at a DOE?

What Do I need to do to <u>improve</u> my Game?

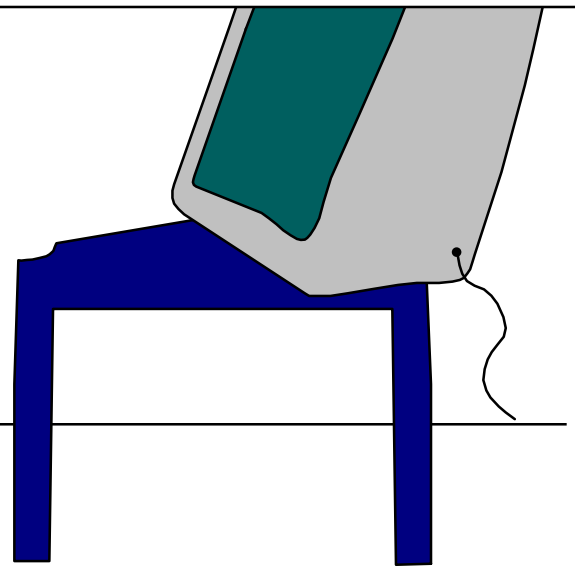


<u>IMPROVEMENT PHASE</u>
Vital Few Variables
Establish Operating Tolerances

How do I know what works for me..... Lane conditions? Ball type? Wristband?



It looks like the **lanes** are in good **condition** today, Mark...
Tim has brought three different **bowling balls** with him but I don't think he will need them all today.
You know he seems to have improved his game ever since he started bowling with that **wristband.**.....



How do I set up the Experiment ?

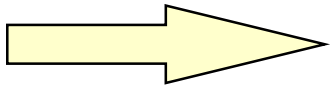
What are all possible Combinations?

(Remember Yates Algorithm?)

	Factor A	Factor B	Factor C
1.	Wristband (+)	hard ball (+)	oily lane (+)
2	Wristband (+)	hard ball (+)	dry lane (-)
3.	Wristband (+)	soft ball (-)	oily lane (+)
4.	Wristband (+)	softball (-)	dry lane (-)
5.	No Wristband(-)	hard ball (+)	oily lane (+)
6.	No Wristband(-)	hard ball (+)	dry lane (-)
7.	No Wristband(-)	soft ball (-)	oily lane (+)
8.	No Wristband(-)	softball (-)	dry lane (-)

A 3 factor, 2 level full factorial DOE would have $2^3=8$ experimental treatments!

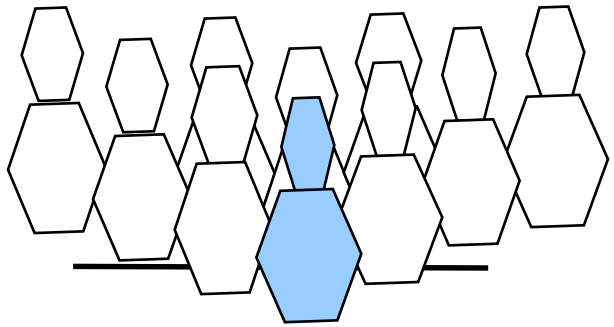
Let's Look at it a different way?



	dry lane	oily lane
hard ball *wristband*	1	2
soft ball	3	4
hard ball *no wristband*	5	6
soft ball	7	8

oily lane
hard bowling ball
wearing a wristband

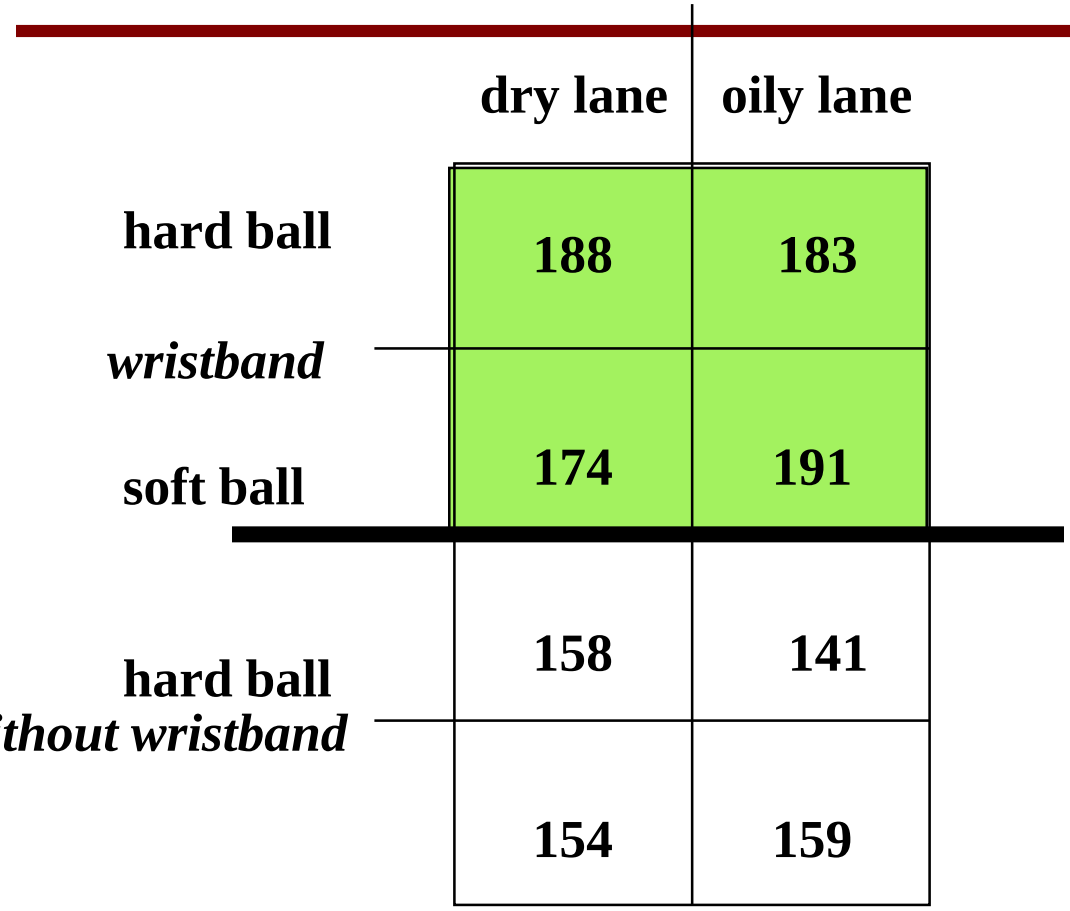
dry lane
hard bowing ball
not wearing wrist band



This is a Full factorial

Let's look at the data!

What about the Wristband?? Did it help me?



Average of "with wristband" scores =184

Higher Scores !!

Average of "without wristband" scores =153

The Wristband appears Better.....
This is called a Main Effect!

What about Ball Type?

	dry lane	oily lane
hard ball *wristband*	188	183
soft ball	174	191
hard ball *no wristband*	158	141
soft ball	154	159

Your best Scores are when:

Dry Lane Hard Ball

OR

Oily Lane Soft Ball

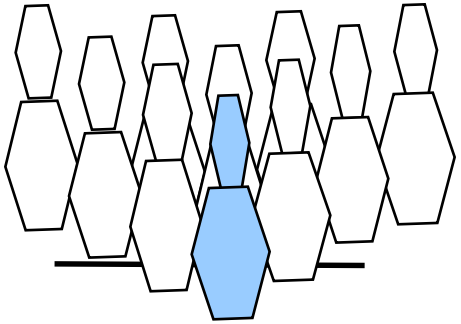
The Ball Type depends on the Lane Condition....
This is called an Interaction!

Where do we go from here?

With Wristband

_____ **and**

When lane is:	**use:**
Dry	**Hard Ball**
Oily	**Soft Ball**

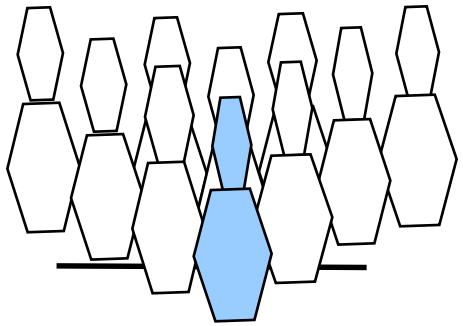


You're on your way to the PBA !!!

Where do we go from here?

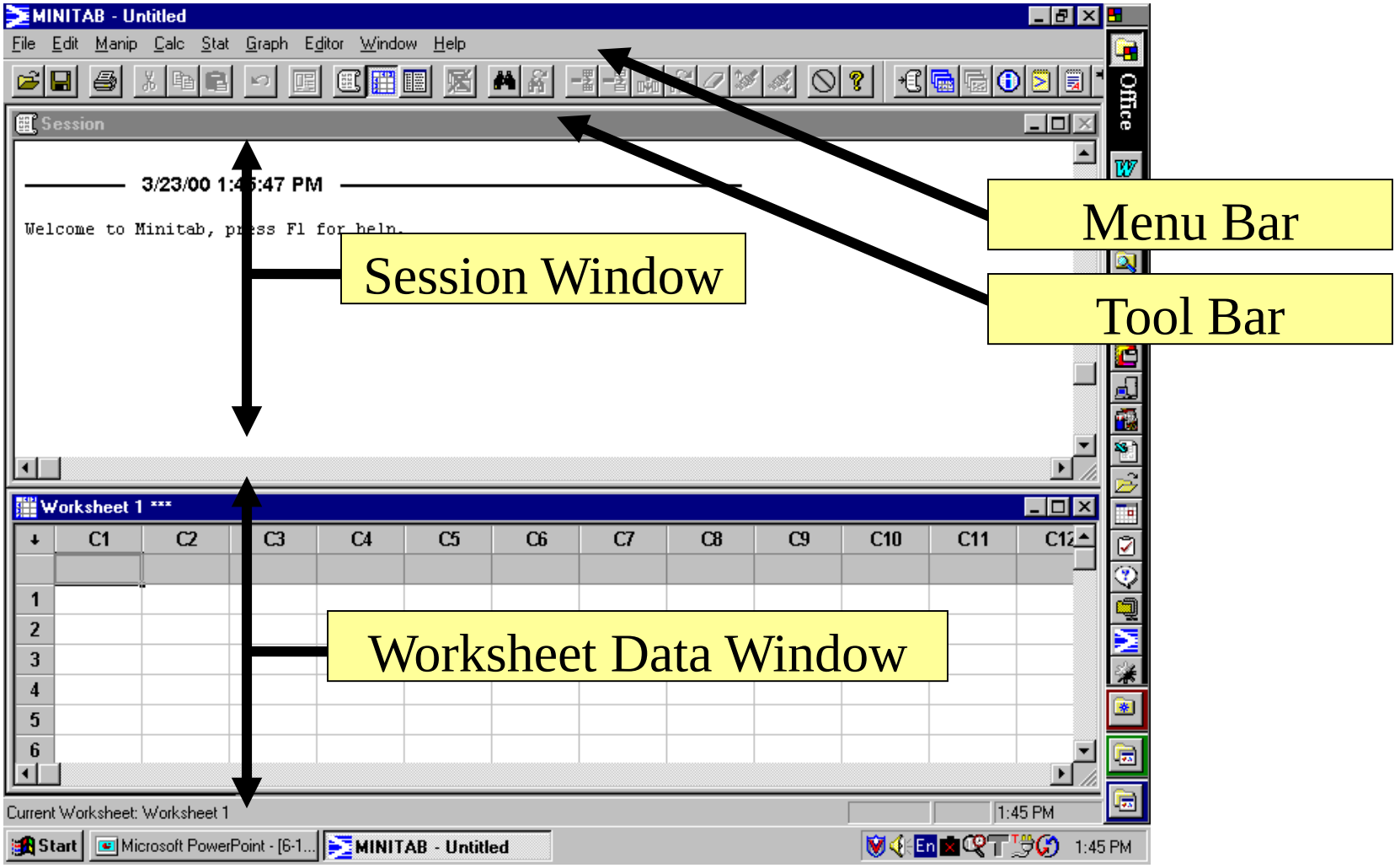
Now, evaluate the results using Yates Algorithm...

What do you think?...

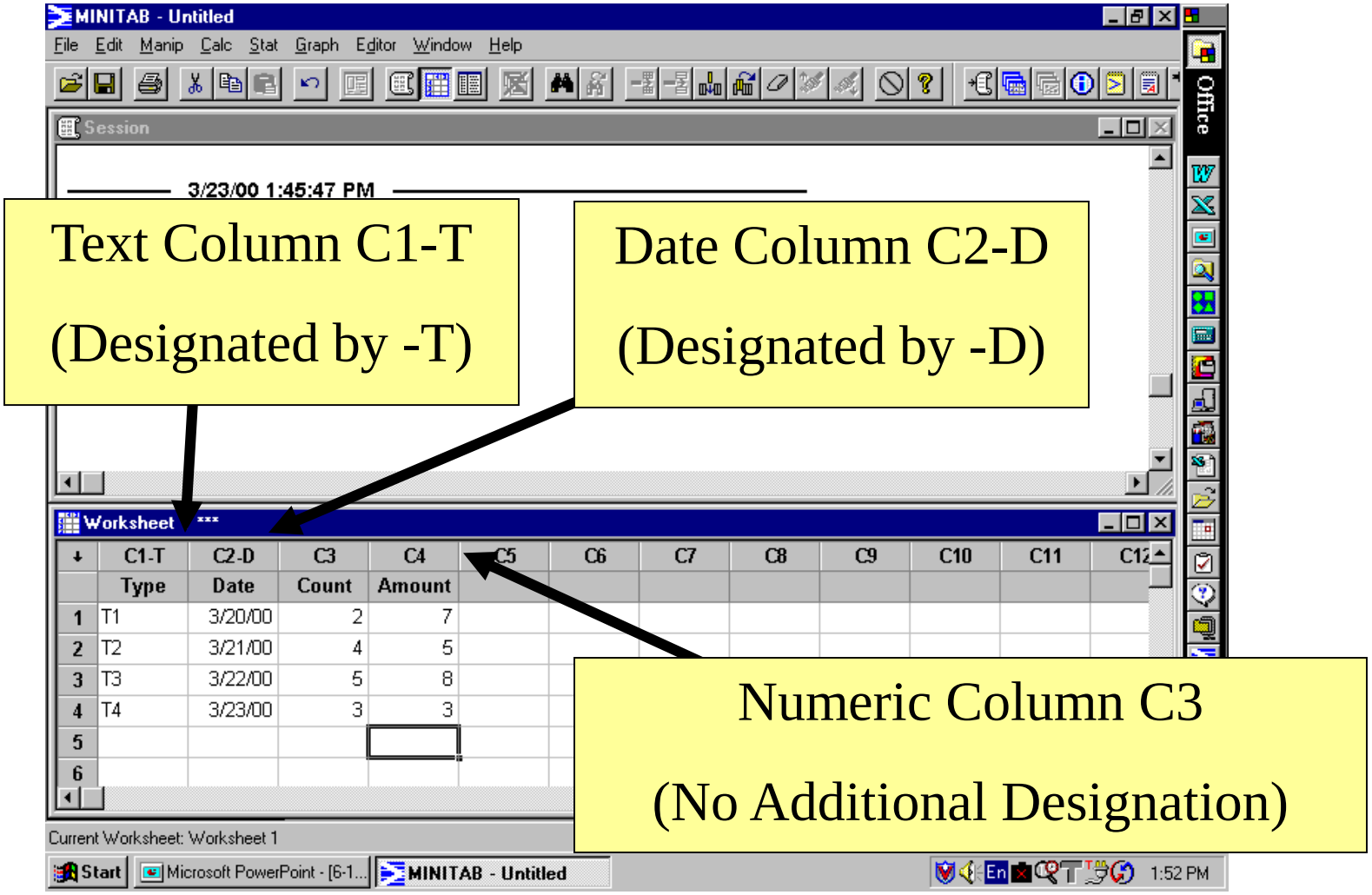


INTRODUCTION TO MINITAB VERSION 13

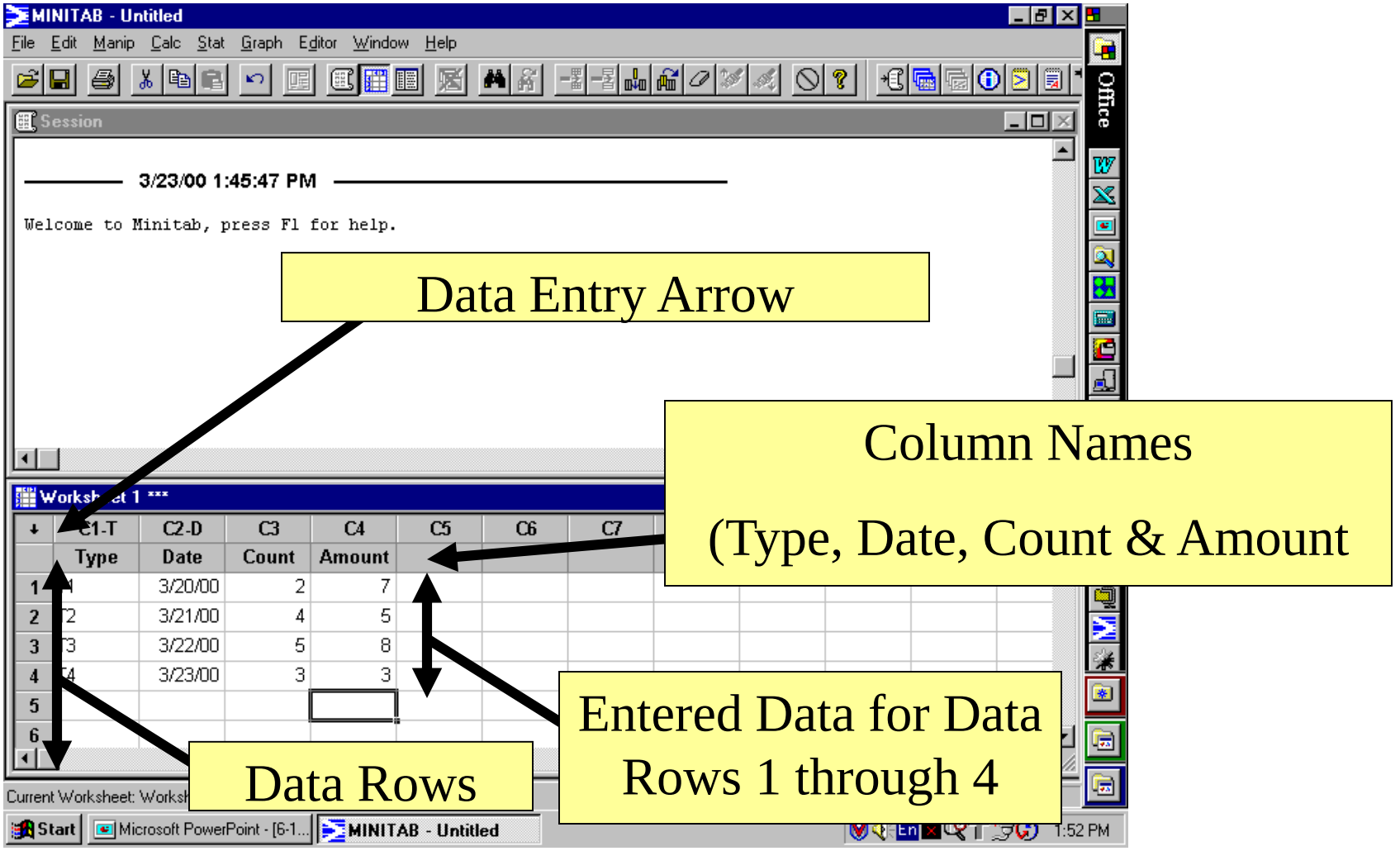
Worksheet Format and Structure



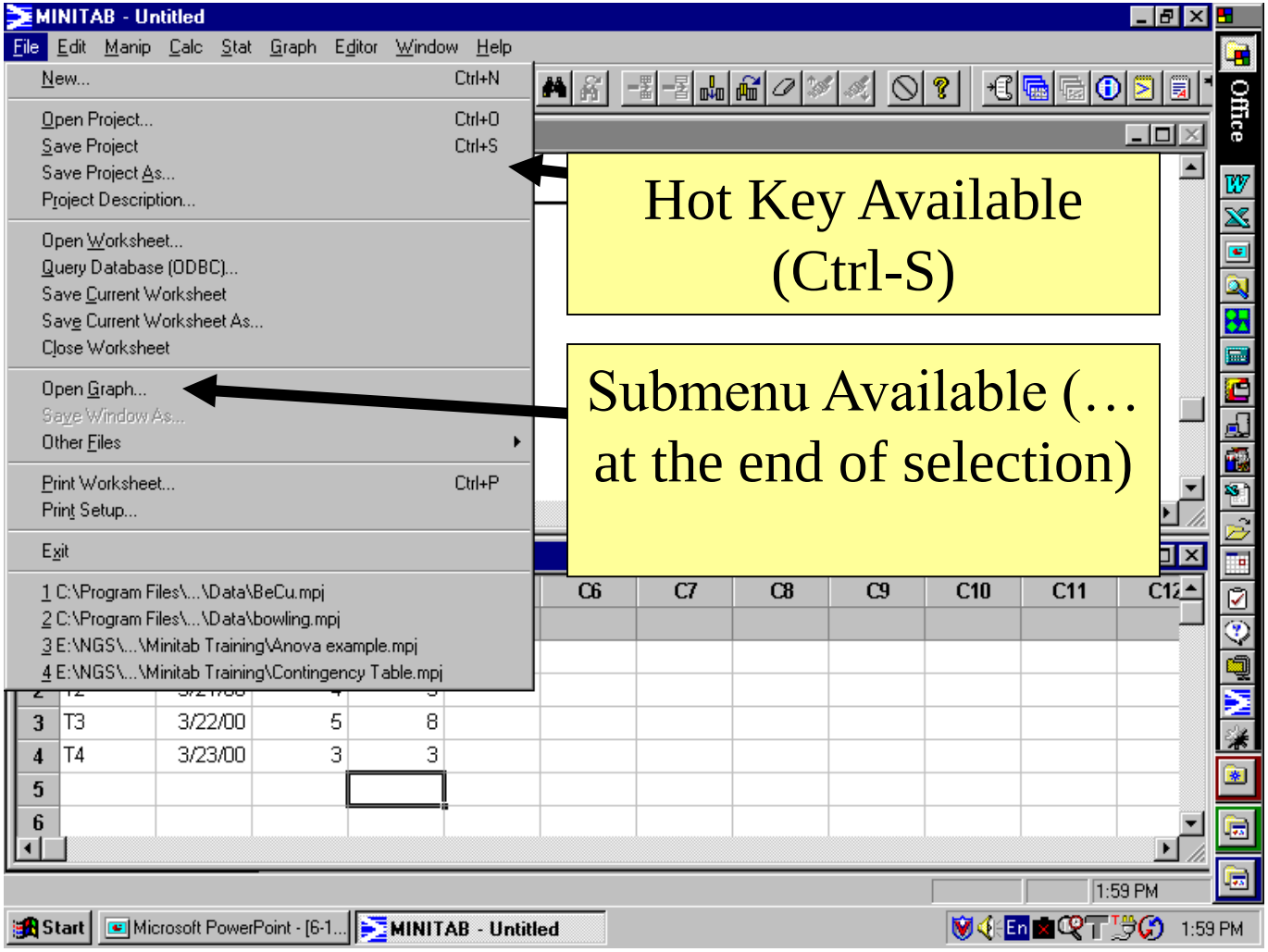
Data Window Column Conventions



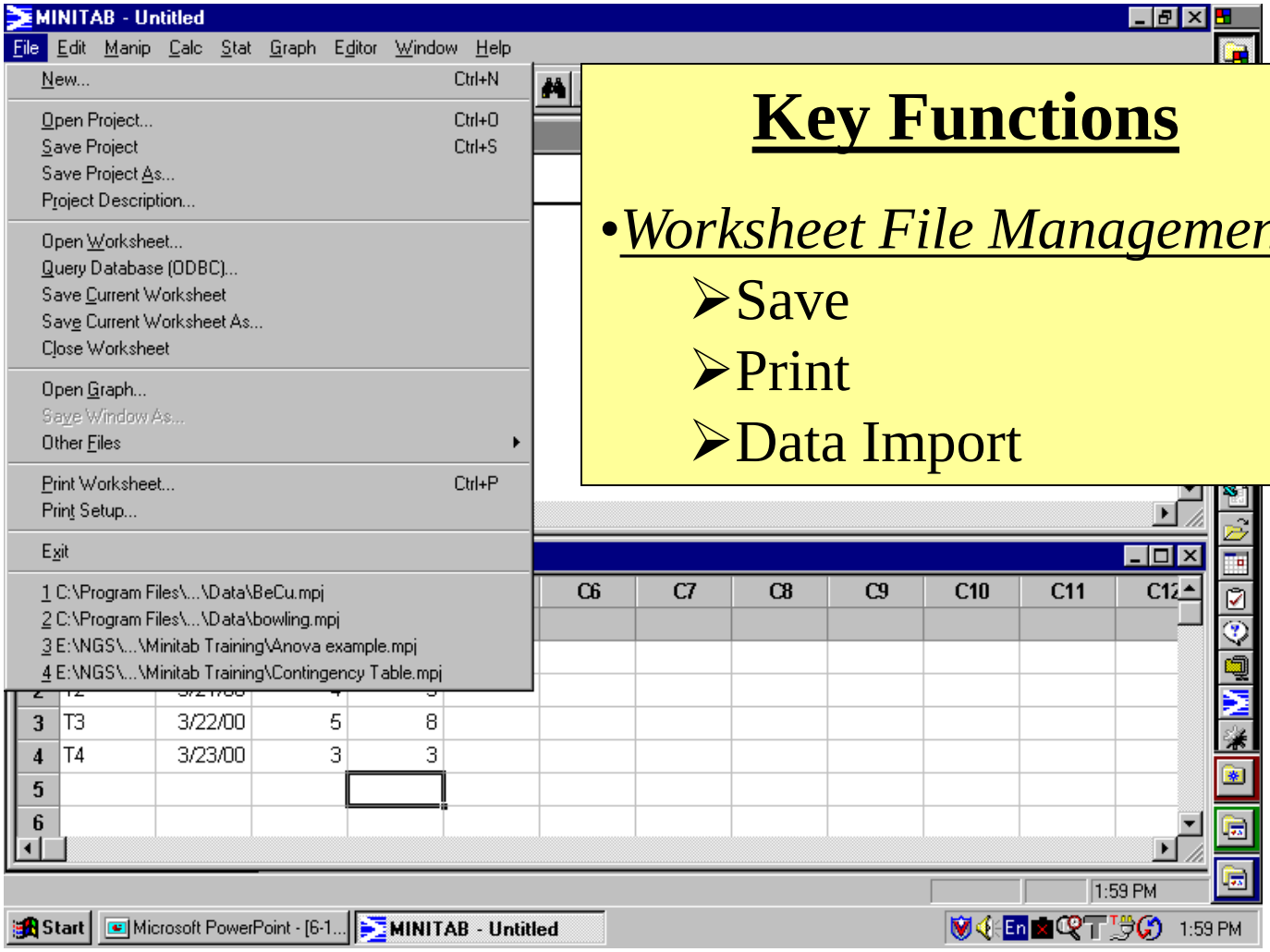
Other Data Window Conventions



Menu Bar - Menu Conventions



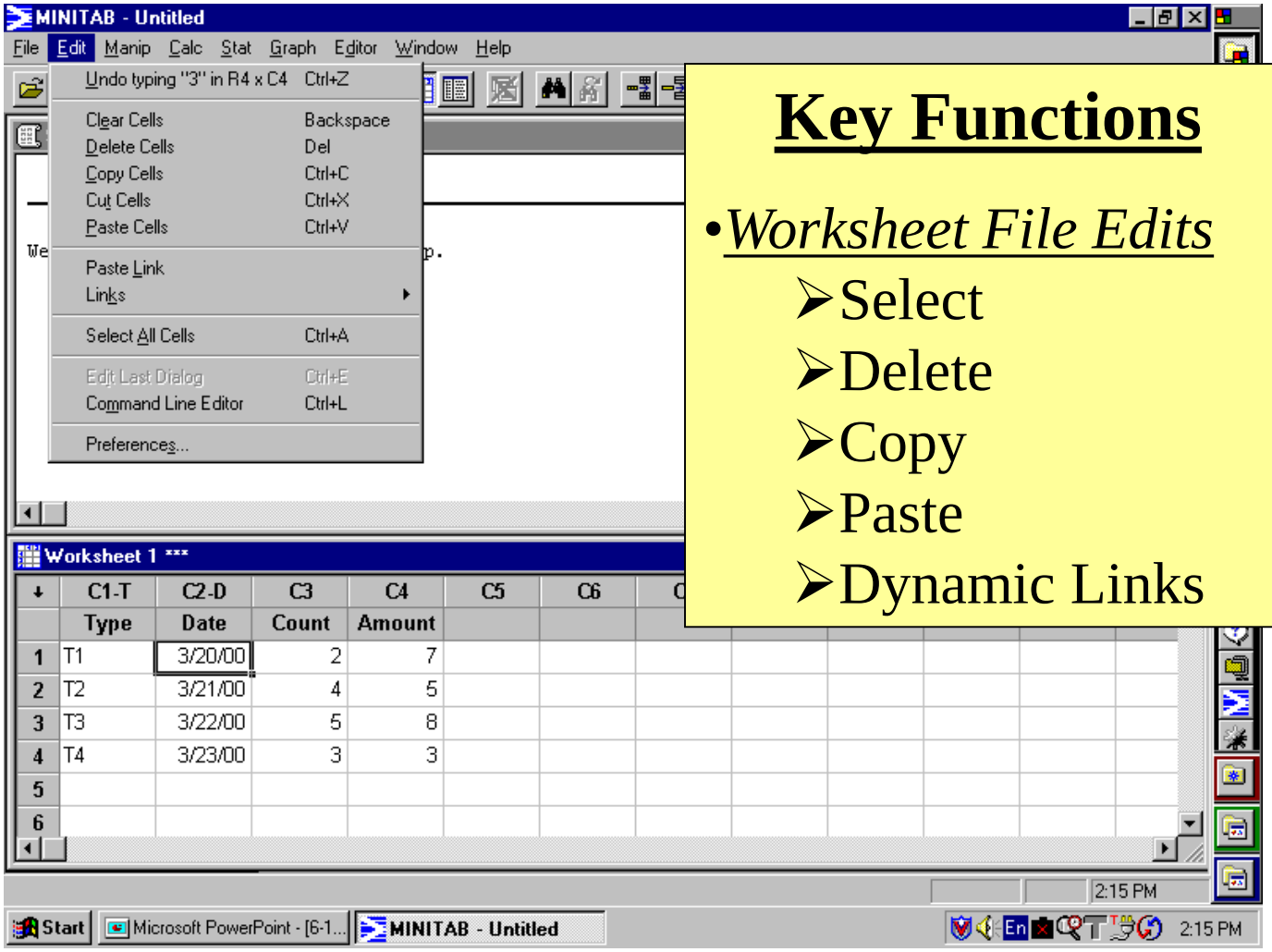
Menu Bar - File Menu



Key Functions

- *Worksheet File Management*
  - Save
  - Print
  - Data Import

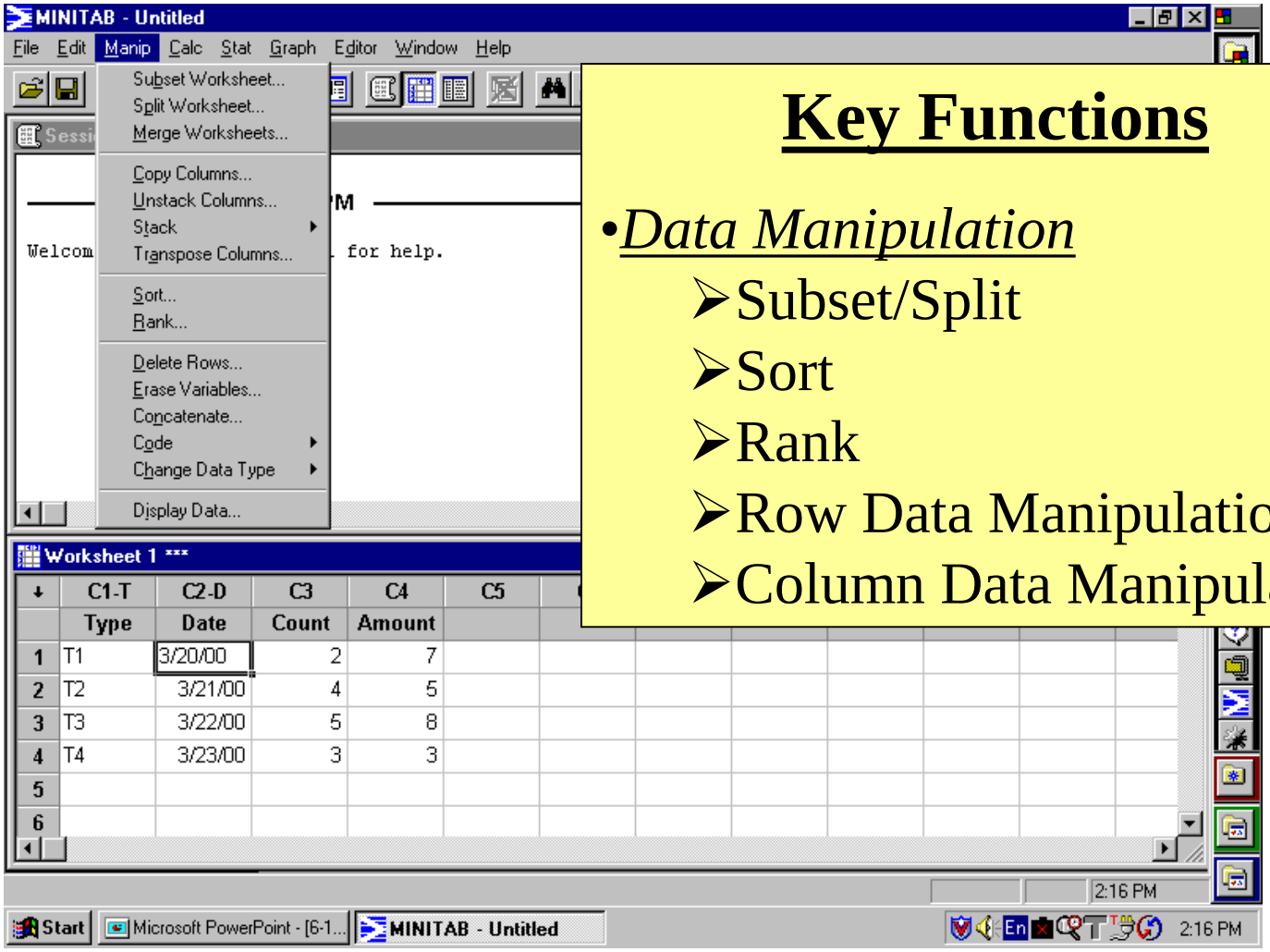
Menu Bar - Edit Menu



Key Functions

- *Worksheet File Edits*
  - Select
  - Delete
  - Copy
  - Paste
  - Dynamic Links

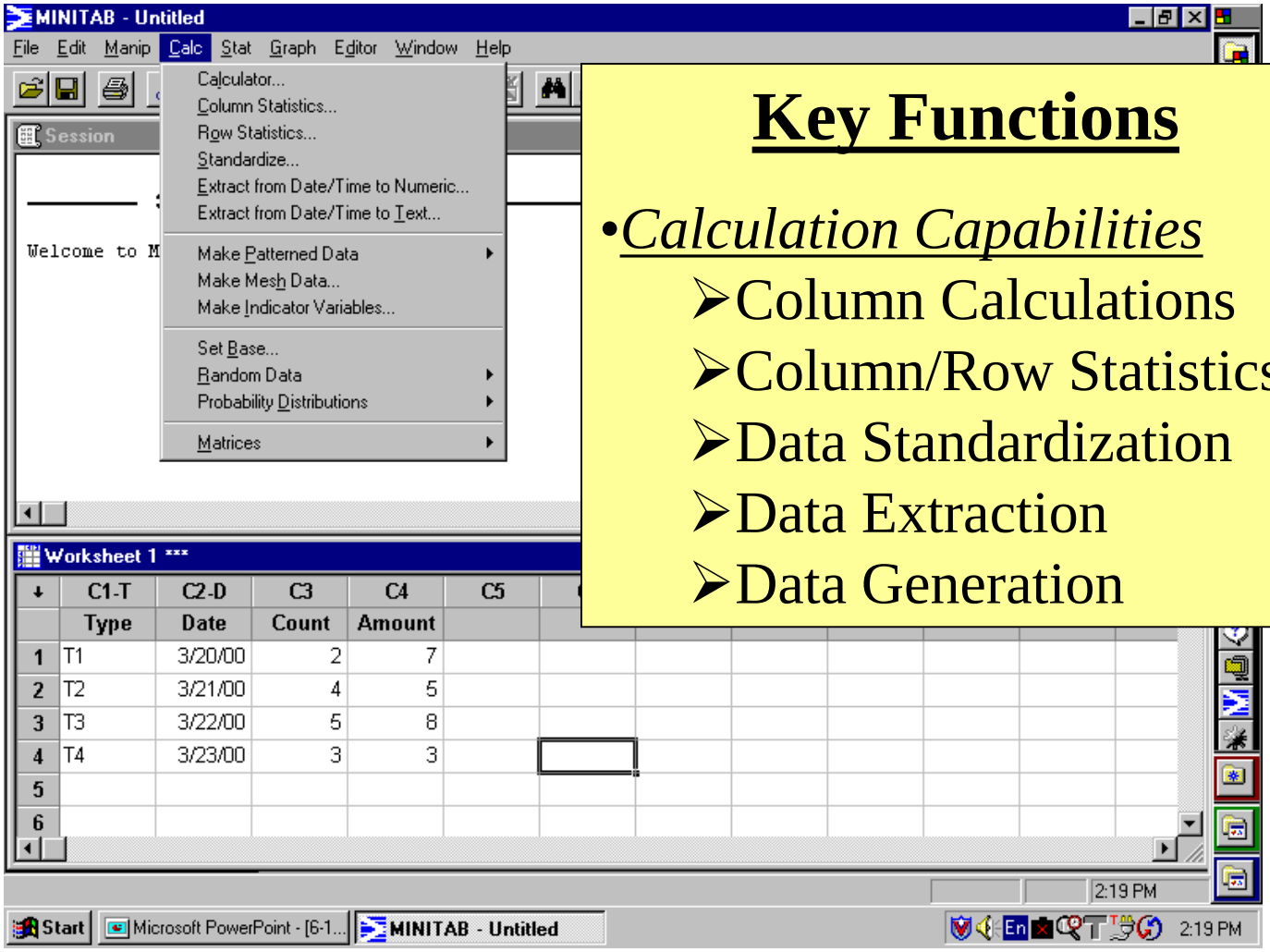
Menu Bar - Manip Menu



Key Functions

- *Data Manipulation*
  - Subset/Split
  - Sort
  - Rank
  - Row Data Manipulation
  - Column Data Manipulation

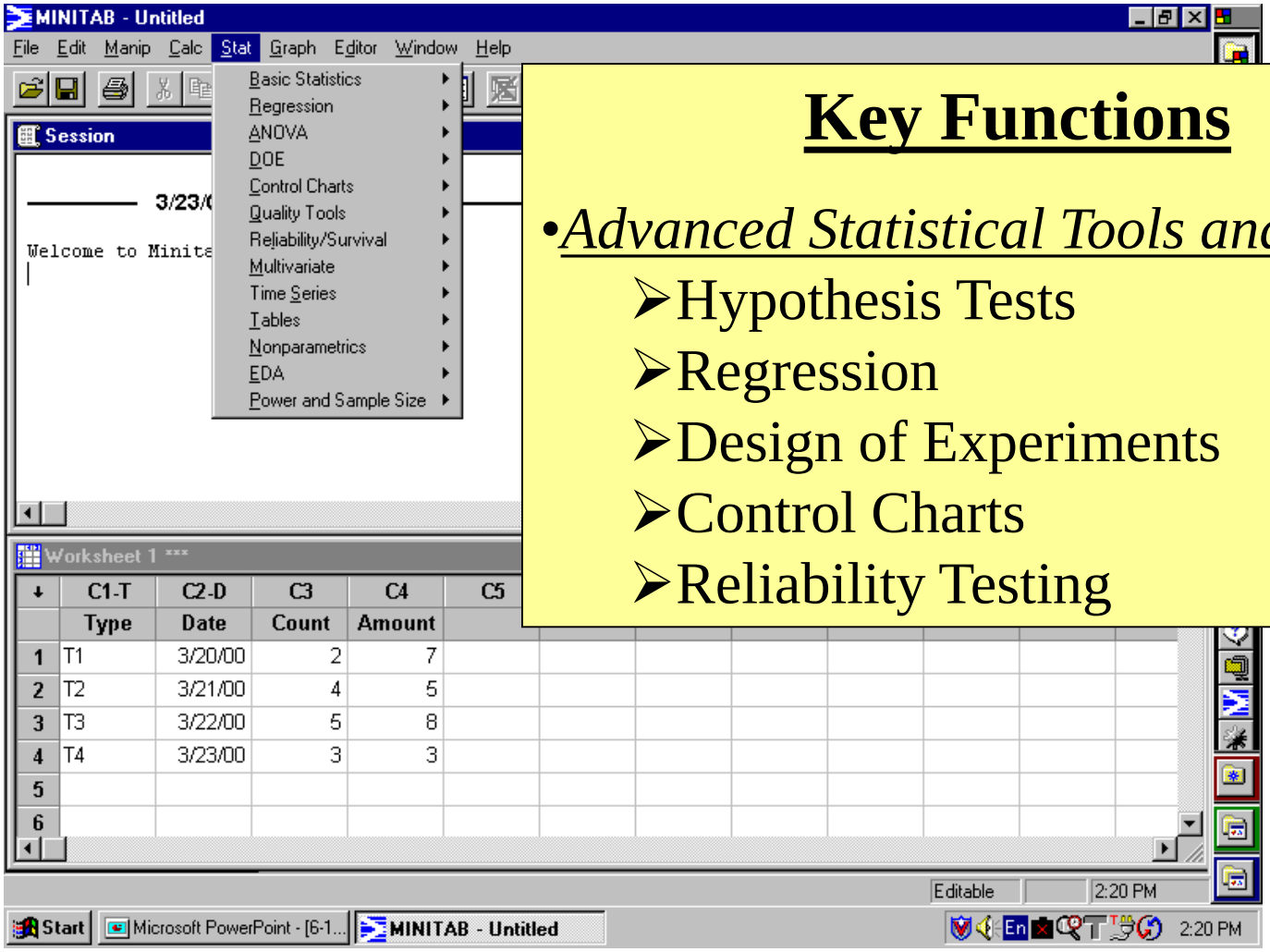
Menu Bar - Calc Menu



Key Functions

- *Calculation Capabilities*
  - Column Calculations
  - Column/Row Statistics
  - Data Standardization
  - Data Extraction
  - Data Generation

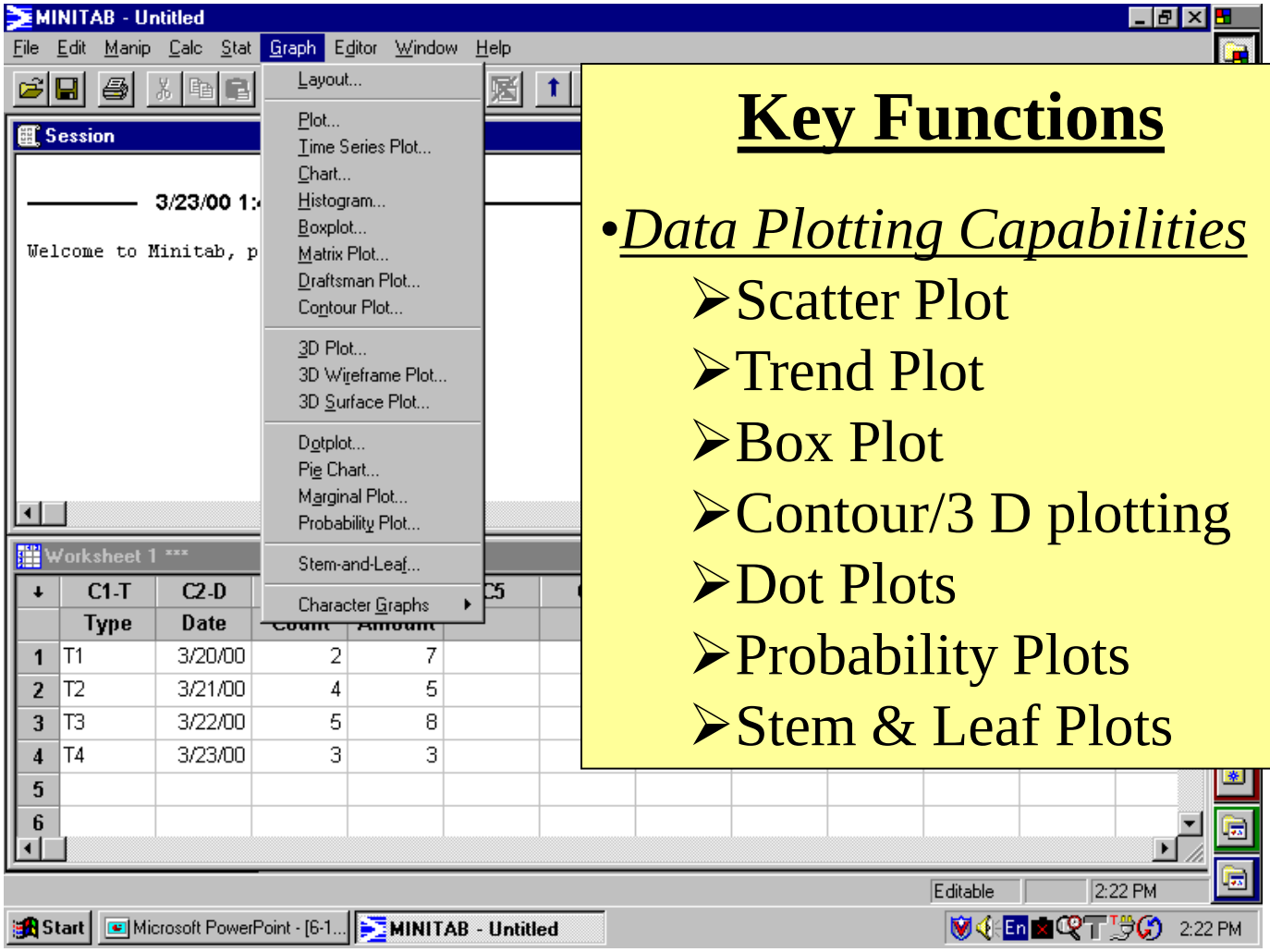
Menu Bar - Stat Menu



Key Functions

- *Advanced Statistical Tools and Graphs*
  - Hypothesis Tests
  - Regression
  - Design of Experiments
  - Control Charts
  - Reliability Testing

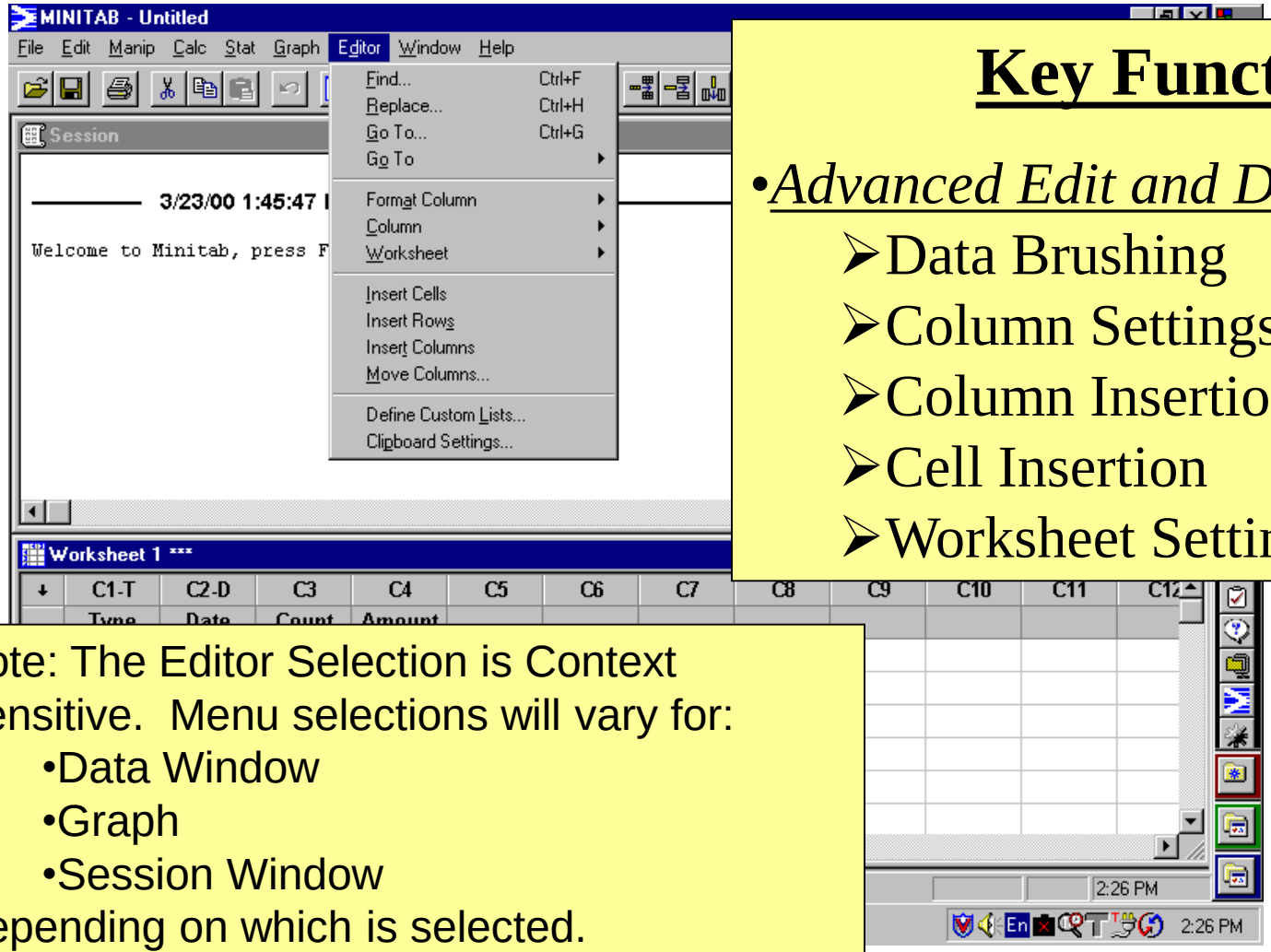
Menu Bar - Graph Menu



Key Functions

- *Data Plotting Capabilities*
  - Scatter Plot
  - Trend Plot
  - Box Plot
  - Contour/3 D plotting
  - Dot Plots
  - Probability Plots
  - Stem & Leaf Plots

Menu Bar - Data Window Editor Menu



Key Functions

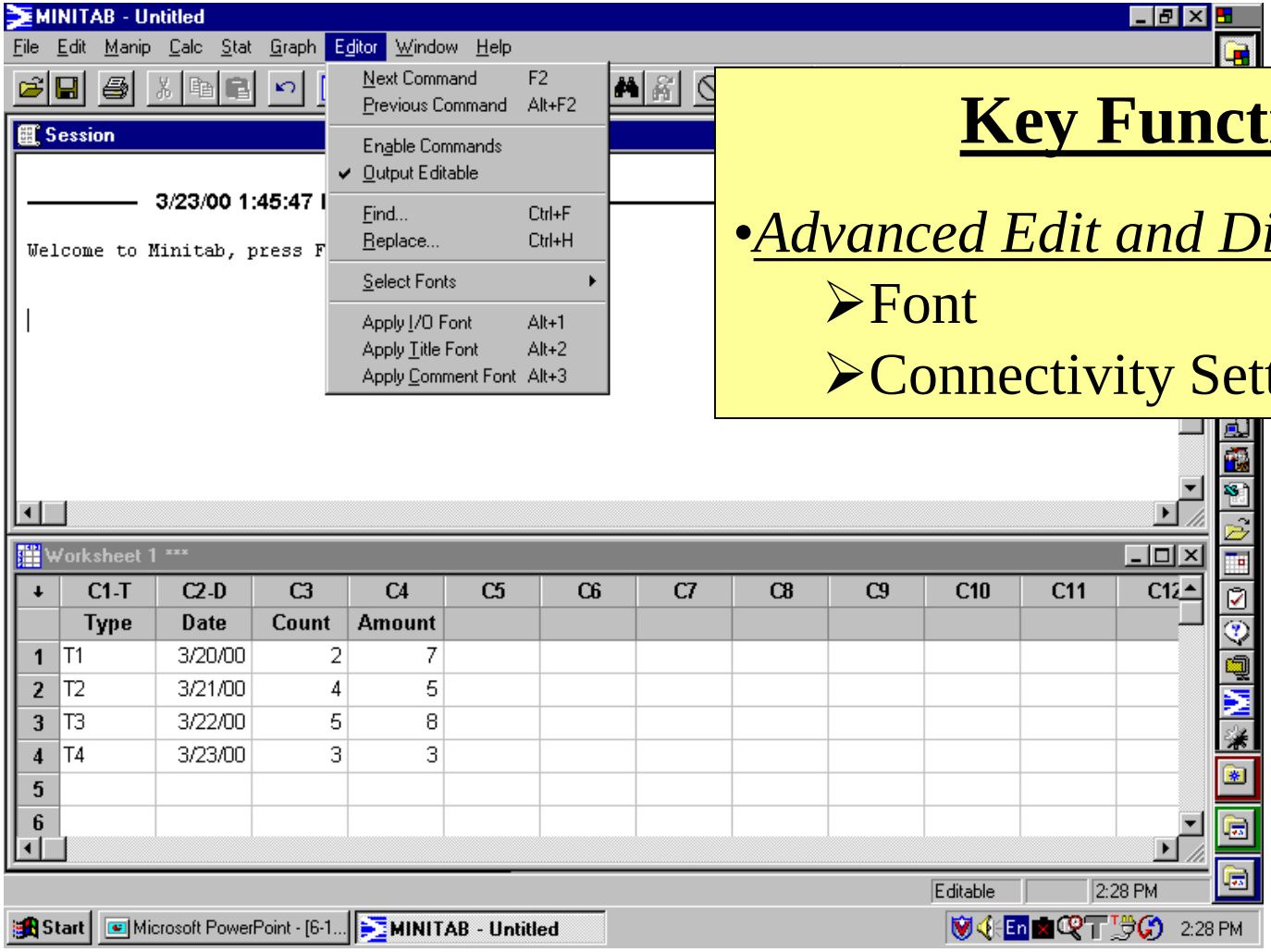
- *Advanced Edit and Display Options*
  - Data Brushing
  - Column Settings
  - Column Insertion/Moves
  - Cell Insertion
  - Worksheet Settings

Note: The Editor Selection is Context Sensitive. Menu selections will vary for:

- Data Window
- Graph
- Session Window

Depending on which is selected.

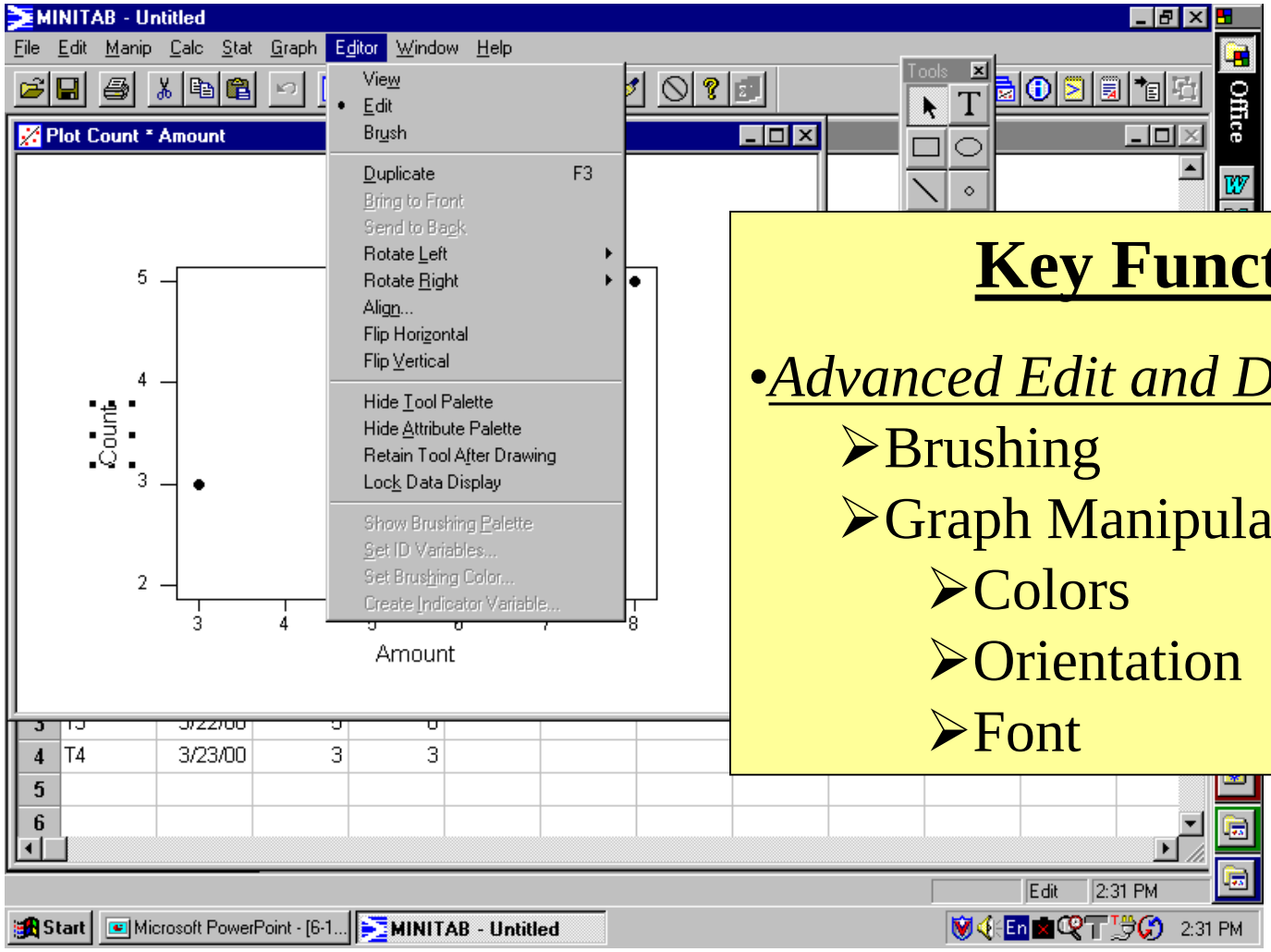
Menu Bar - Session Window Editor Menu



Key Functions

- *Advanced Edit and Display Options*
  - Font
  - Connectivity Settings

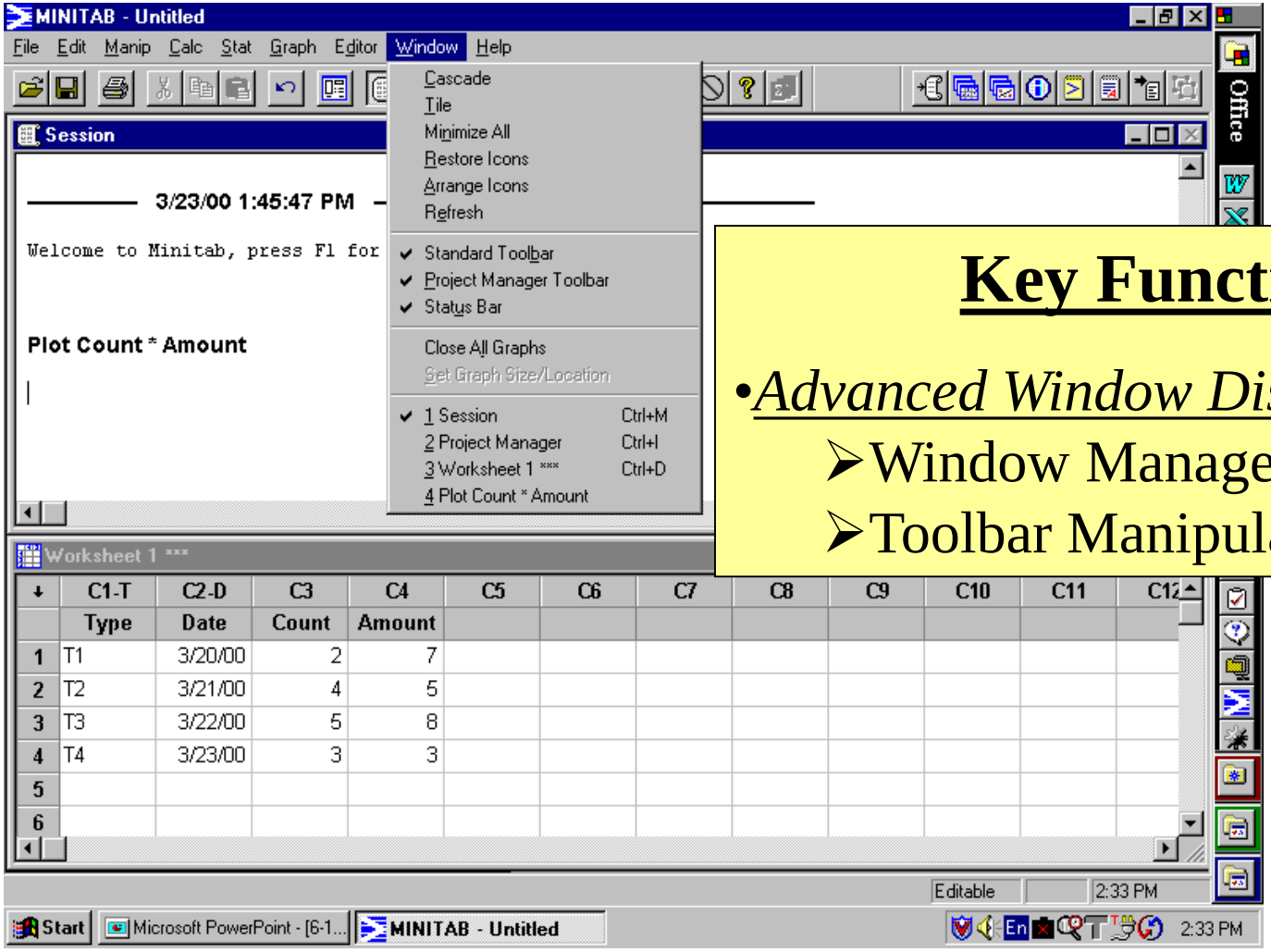
Menu Bar - Graph Window Editor Menu



Key Functions

- *Advanced Edit and Display Options*
  - Brushing
  - Graph Manipulation
    - Colors
    - Orientation
    - Font

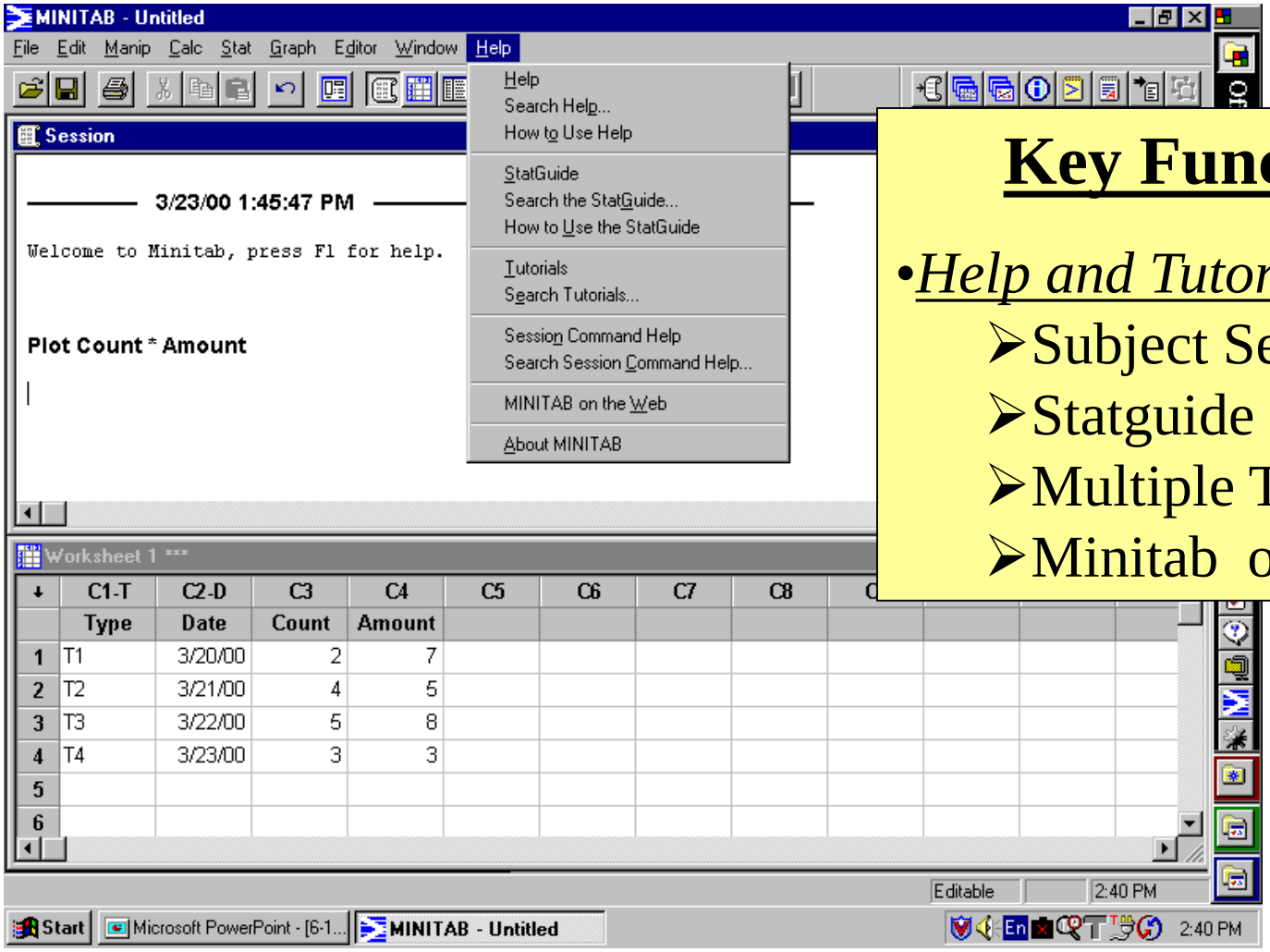
Menu Bar - Window Menu



Key Functions

- *Advanced Window Display Options*
  - Window Management/Display
  - Toolbar Manipulation/Display

Menu Bar - Help Menu

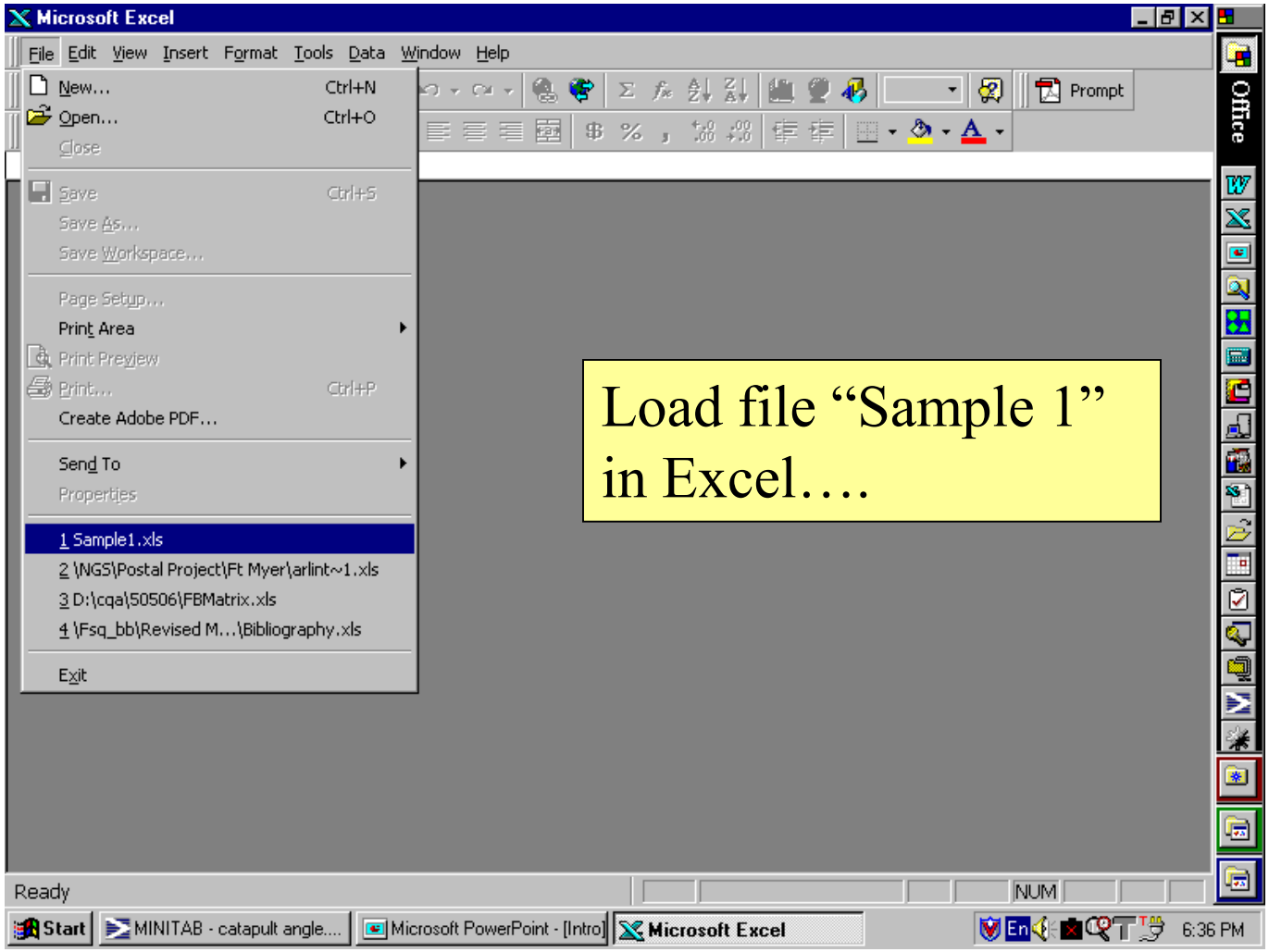


Key Functions

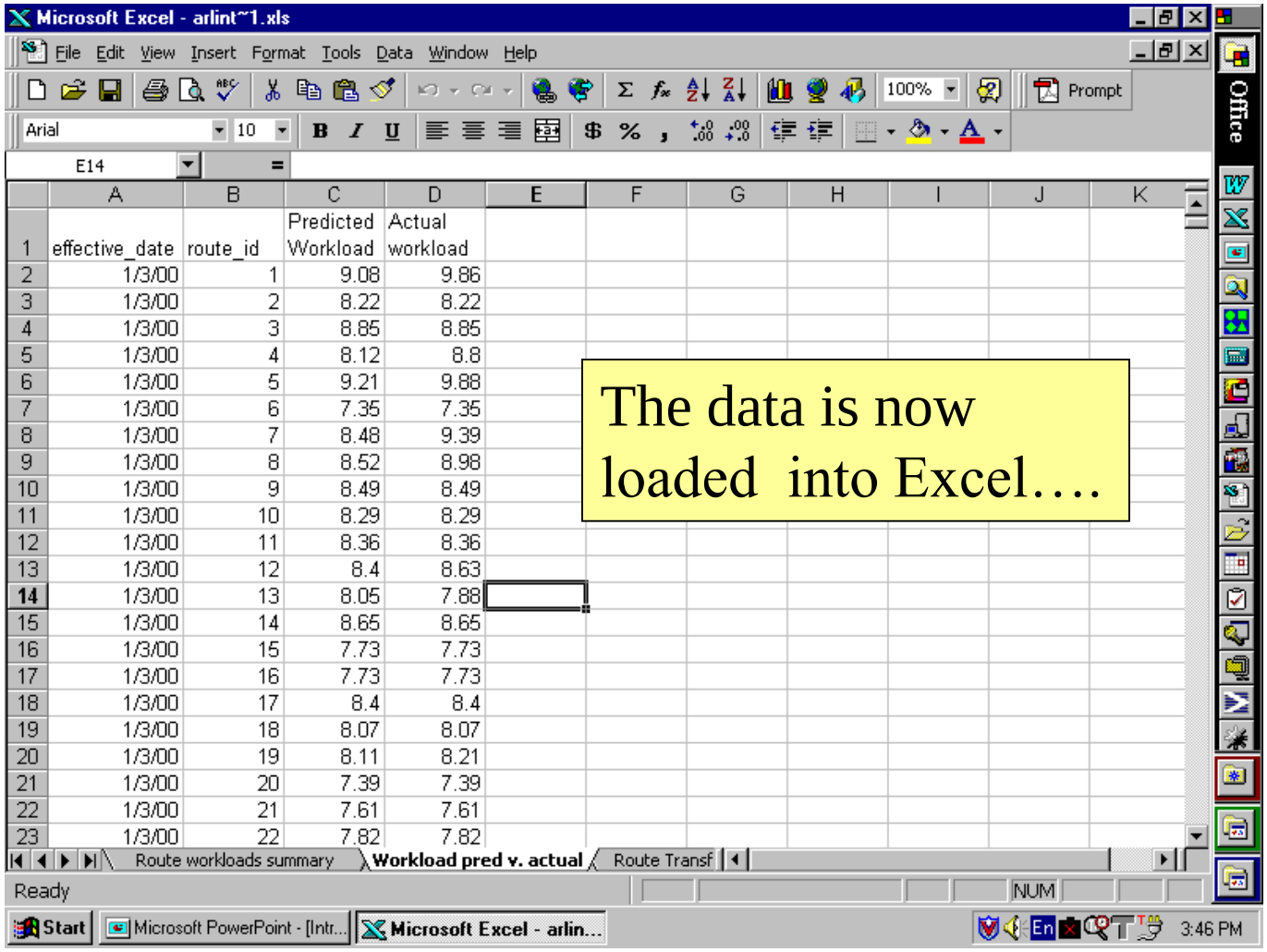
- *Help and Tutorials*
  - Subject Searches
  - Statguide
  - Multiple Tutorials
  - Minitab on the Web

MINITAB INTEROPERABILITY

Starting with Excel...



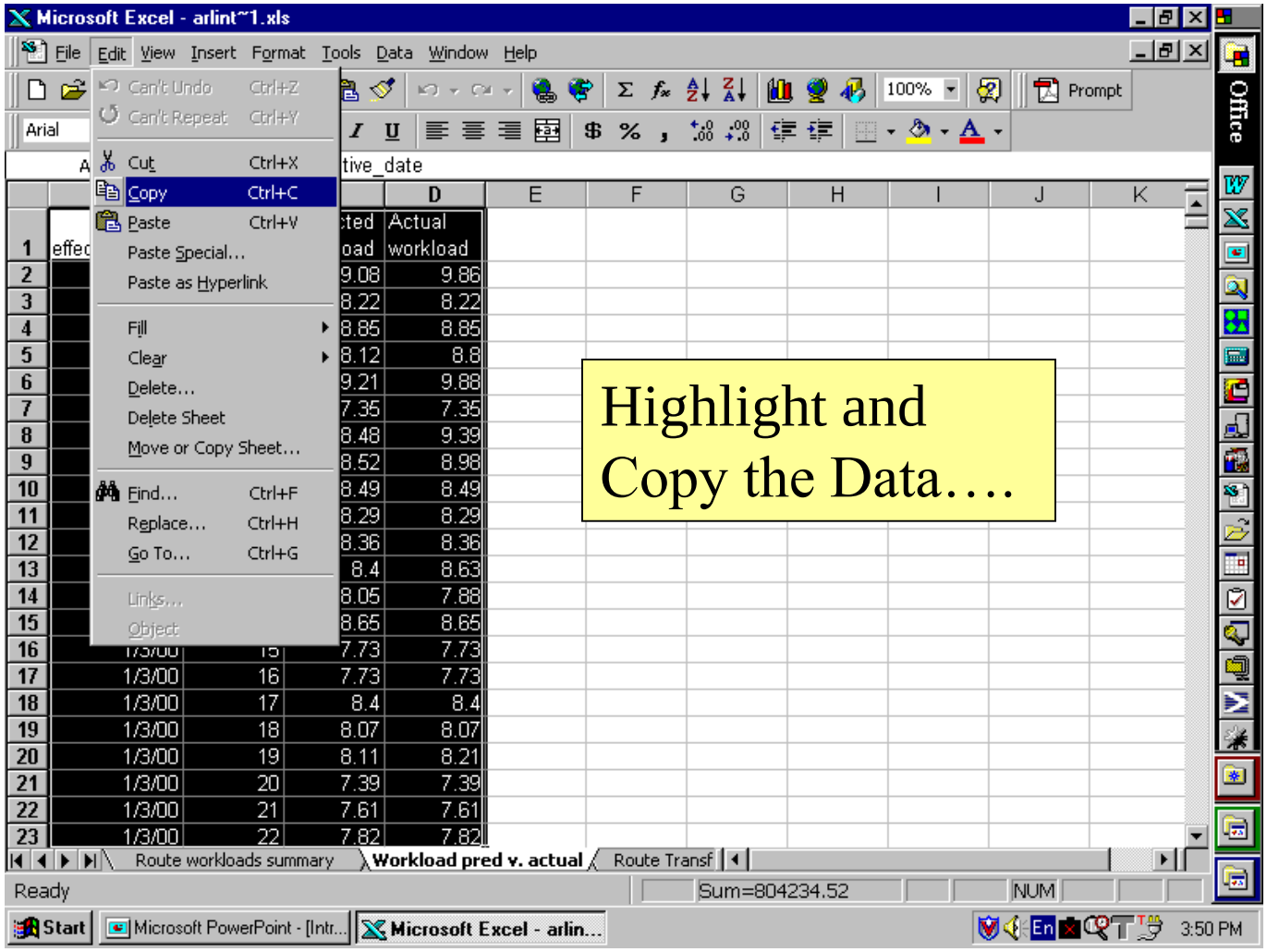
Starting with Excel...



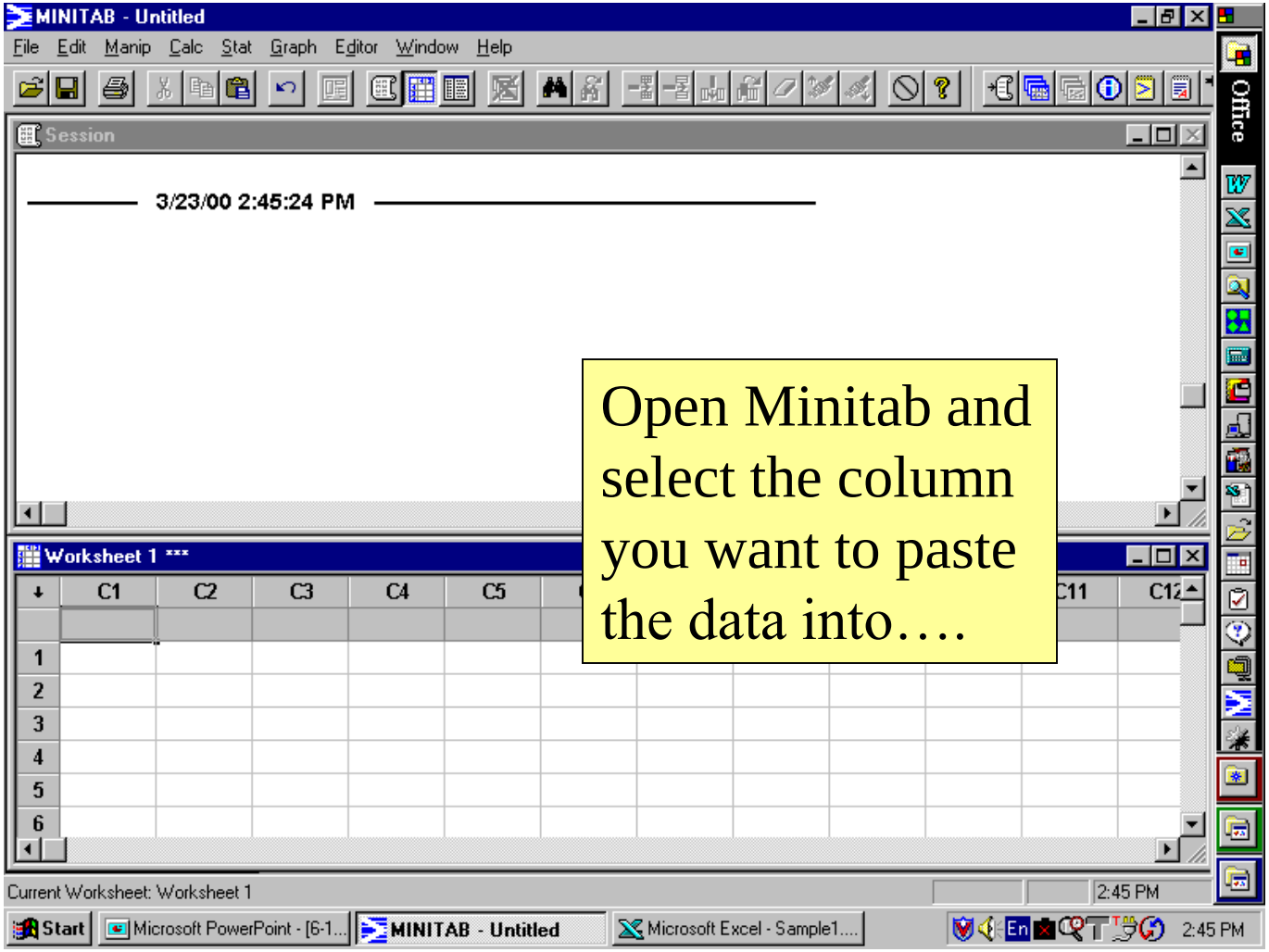
	A	B	C	D	E	F	G	H	I	J	K
1	effective_date	route_id	Predicted Workload	Actual workload							
2	1/3/00	1	9.08	9.86							
3	1/3/00	2	8.22	8.22							
4	1/3/00	3	8.85	8.85							
5	1/3/00	4	8.12	8.8							
6	1/3/00	5	9.21	9.88							
7	1/3/00	6	7.35	7.35							
8	1/3/00	7	8.48	9.39							
9	1/3/00	8	8.52	8.98							
10	1/3/00	9	8.49	8.49							
11	1/3/00	10	8.29	8.29							
12	1/3/00	11	8.36	8.36							
13	1/3/00	12	8.4	8.63							
14	1/3/00	13	8.05	7.88							
15	1/3/00	14	8.65	8.65							
16	1/3/00	15	7.73	7.73							
17	1/3/00	16	7.73	7.73							
18	1/3/00	17	8.4	8.4							
19	1/3/00	18	8.07	8.07							
20	1/3/00	19	8.11	8.21							
21	1/3/00	20	7.39	7.39							
22	1/3/00	21	7.61	7.61							
23	1/3/00	22	7.82	7.82							

The data is now loaded into Excel….

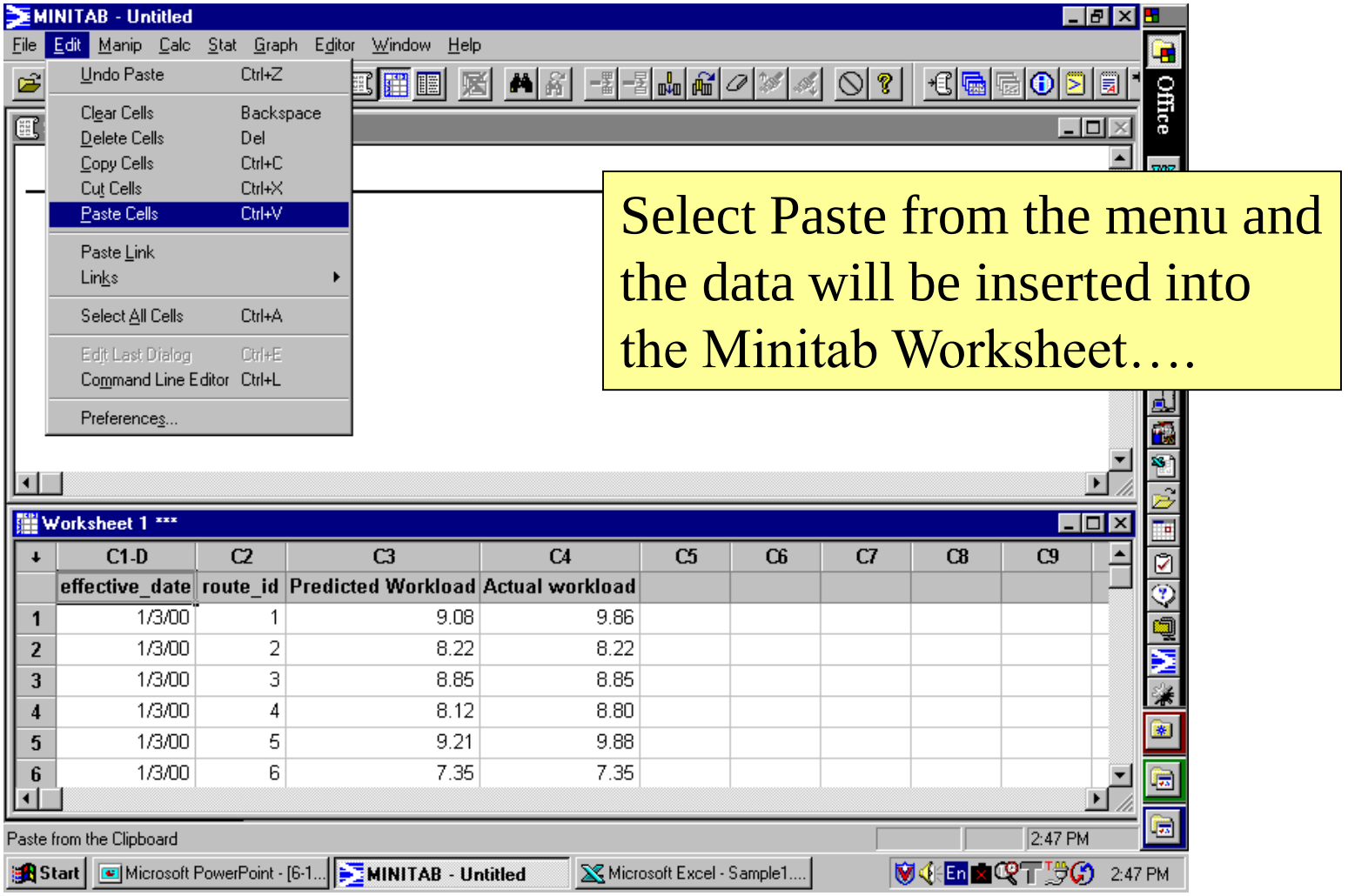
Starting with Excel...



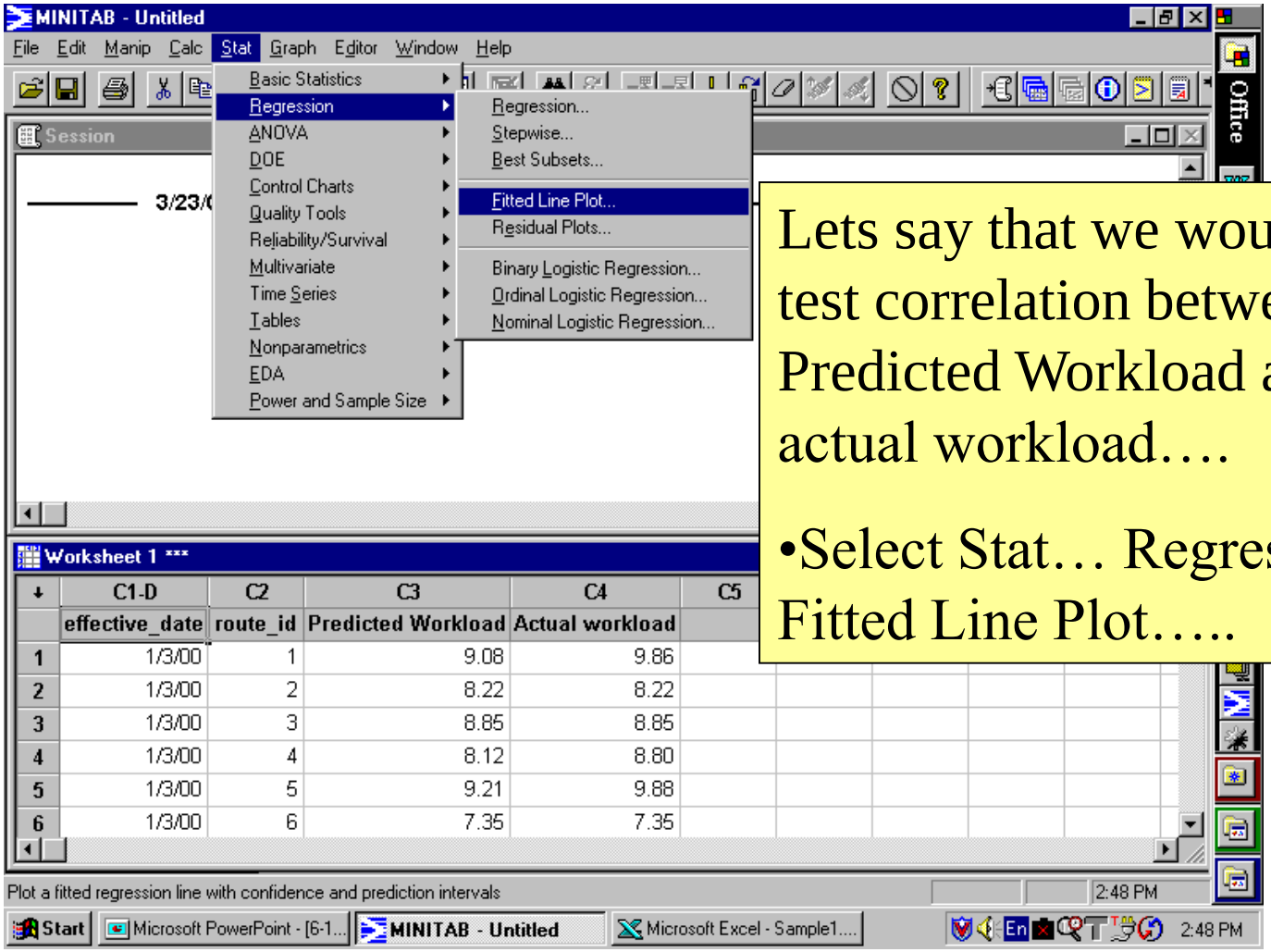
Move to Minitab...



Move to Minitab...



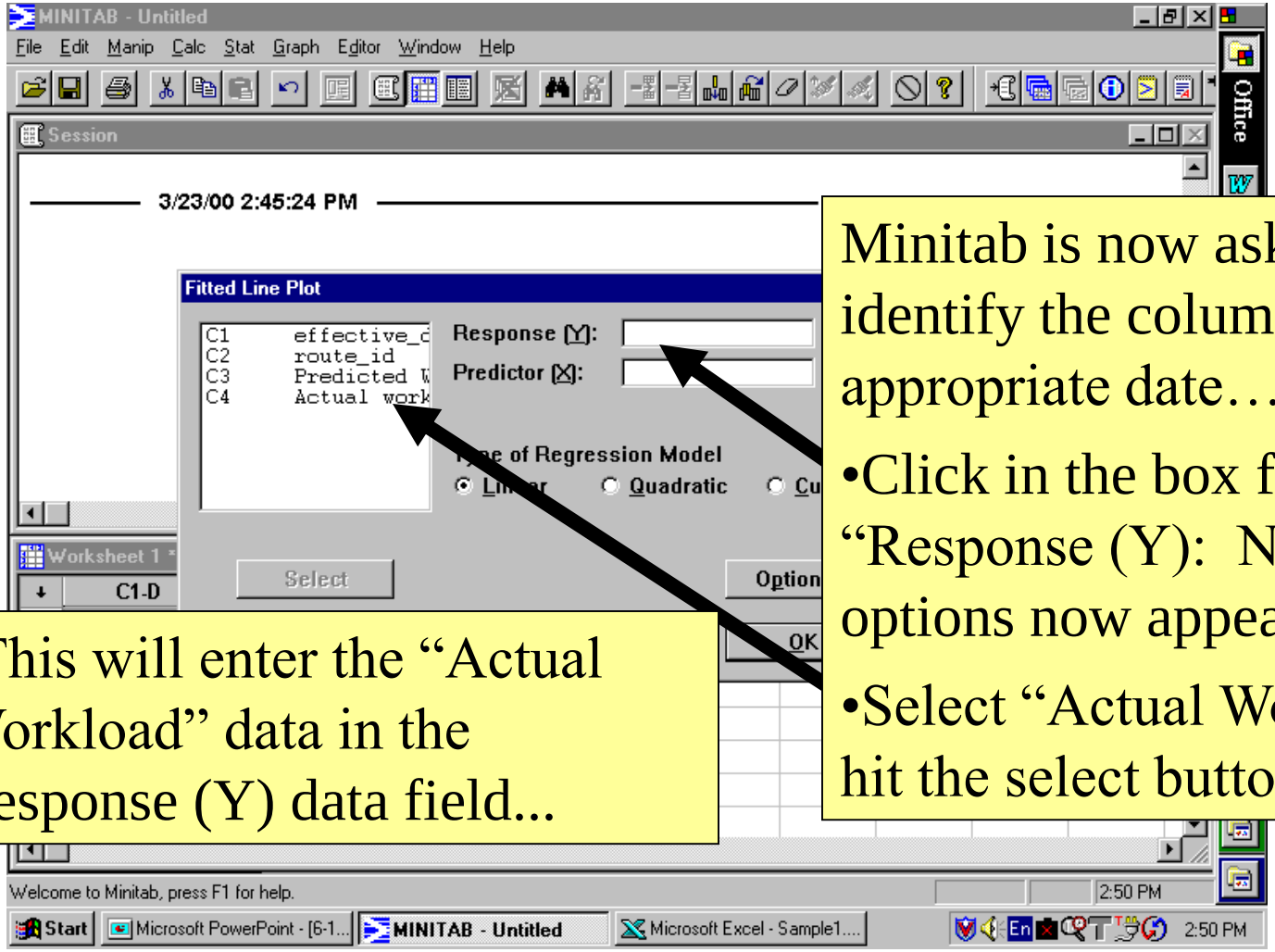
Use Minitab to do the Analysis...



Lets say that we would like to test correlation between the Predicted Workload and the actual workload….

•Select Stat… Regression…. Fitted Line Plot…..

Use Minitab to do the Analysis...

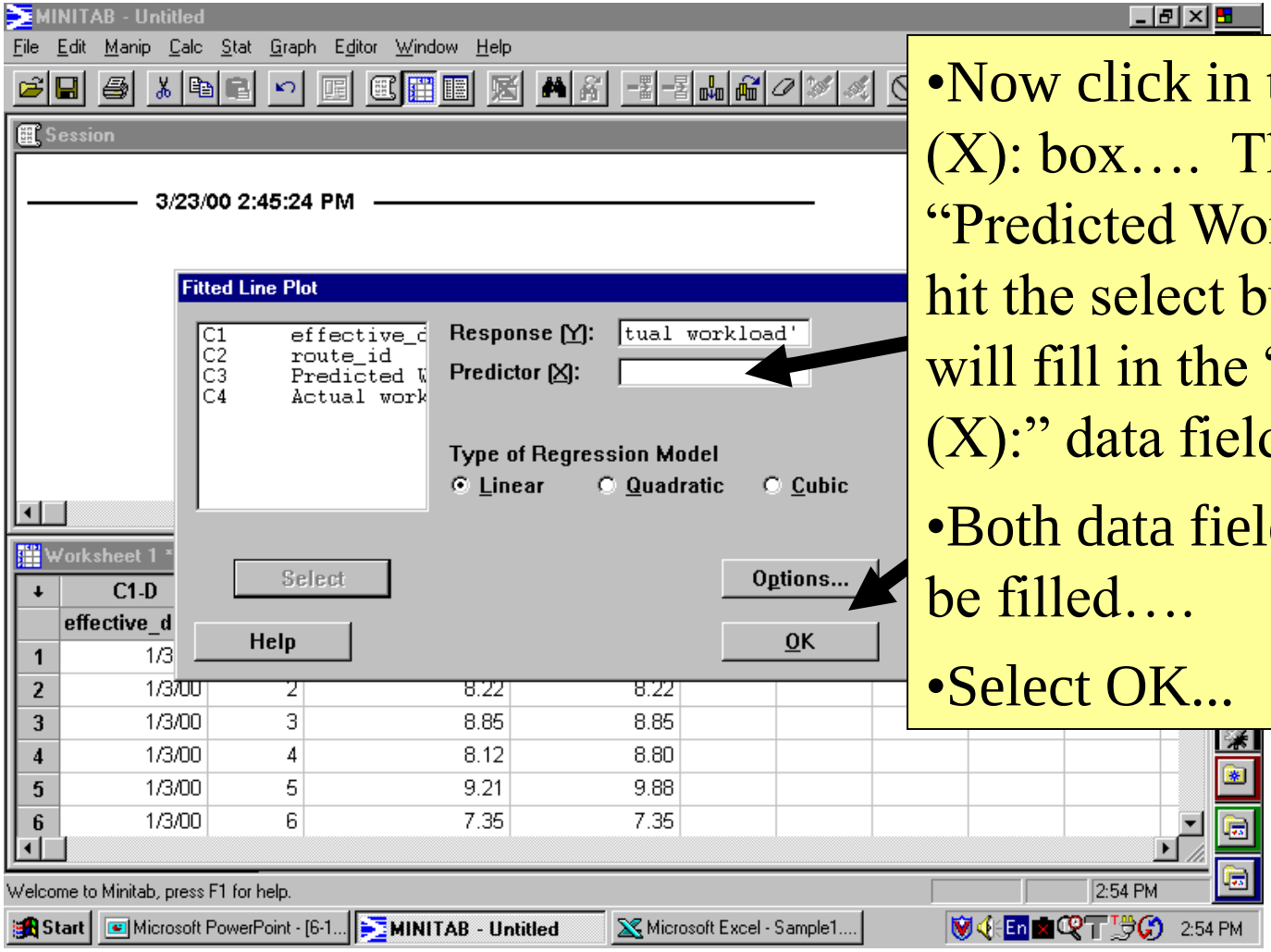


Minitab is now asking for us to identify the columns with the appropriate date….

- Click in the box for "Response (Y): Note that our options now appear in this box.
- Select "Actual Workload" and hit the select button…..

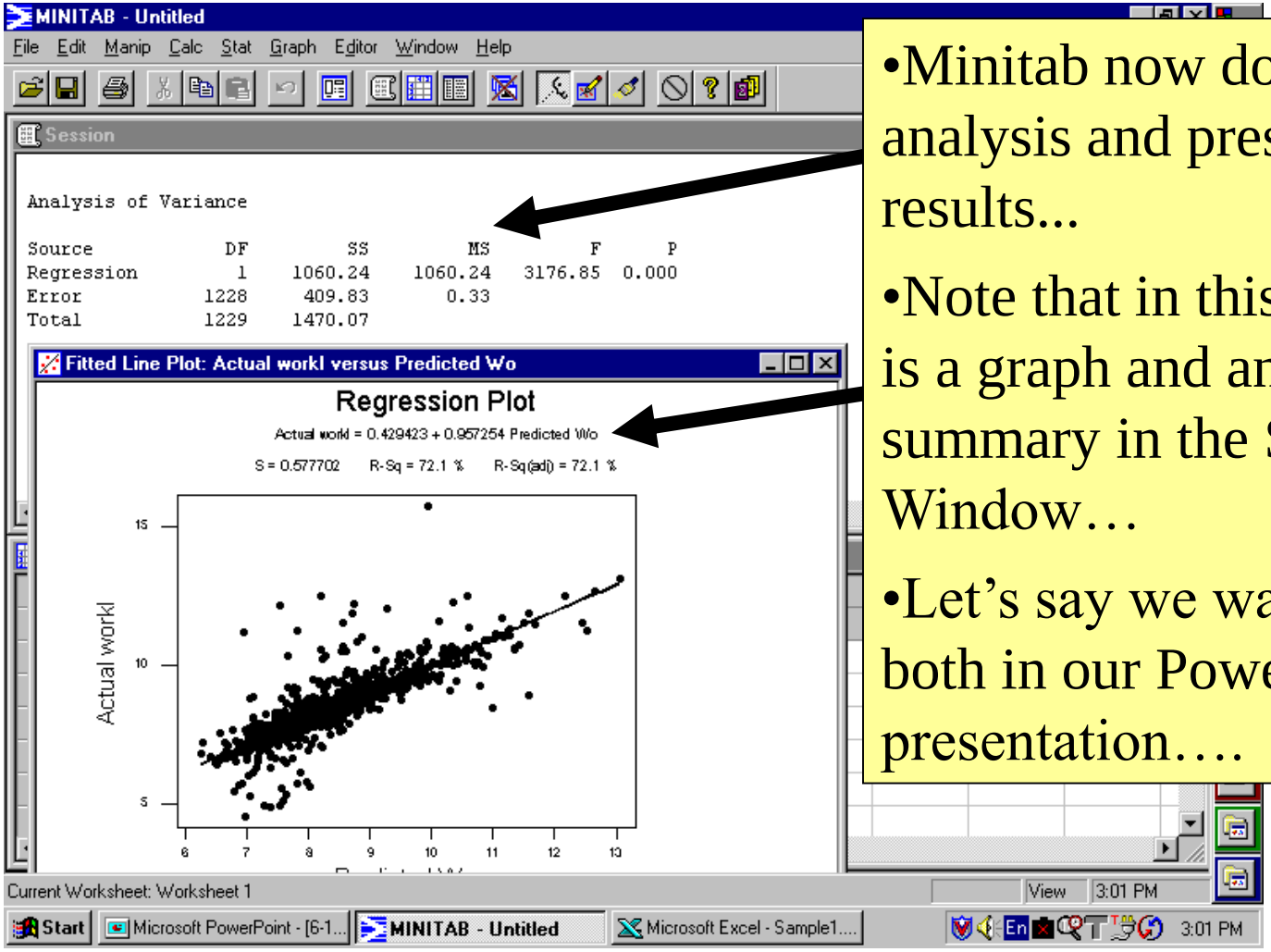
- This will enter the "Actual Workload" data in the Response (Y) data field...

Use Minitab to do the Analysis...



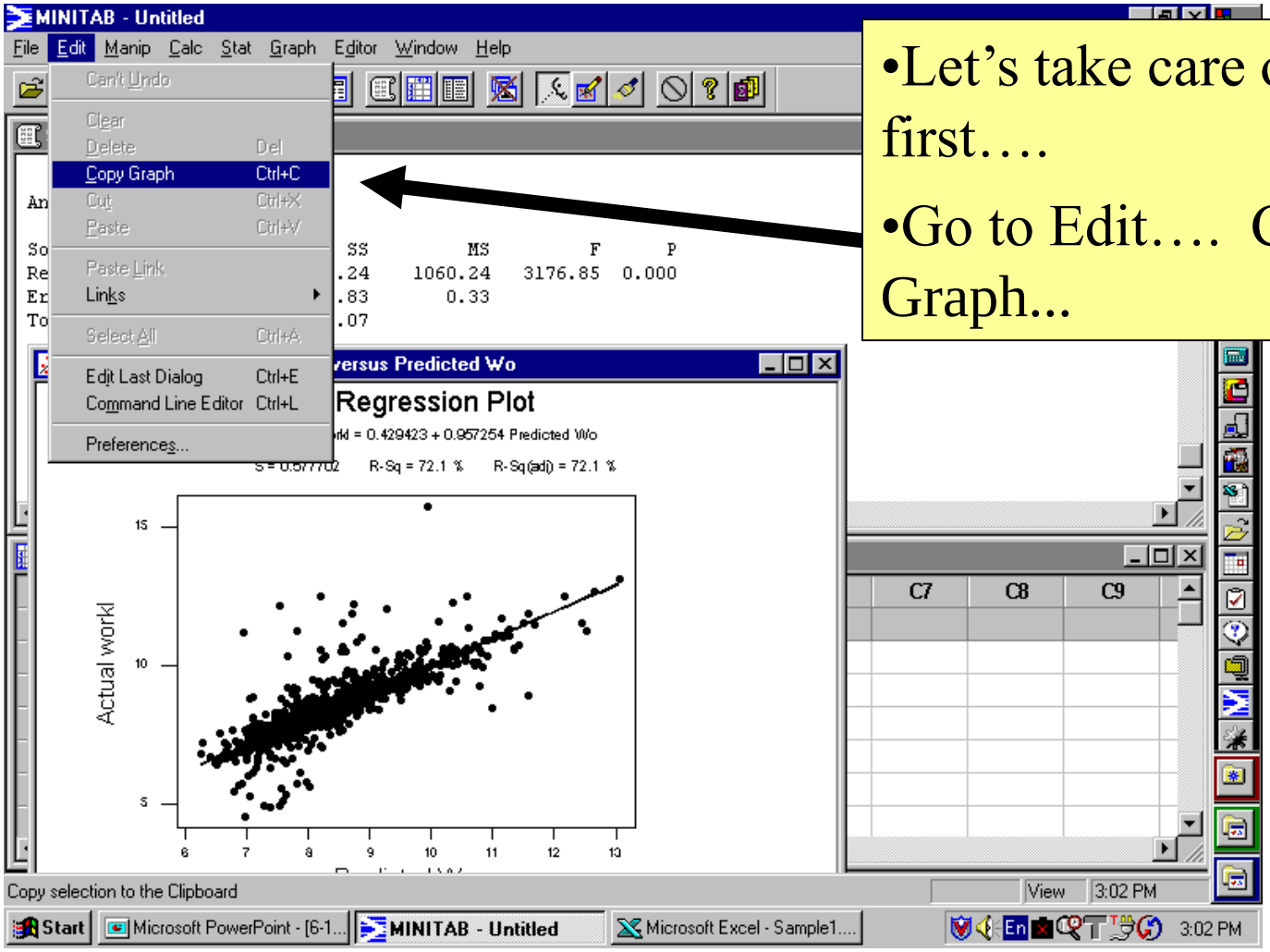
- Now click in the Predictor (X): box…. Then click on "Predicted Workload" and hit the select button… This will fill in the "Predictor (X):" data field...
- Both data fields should now be filled….
- Select OK...

Use Minitab to do the Analysis...



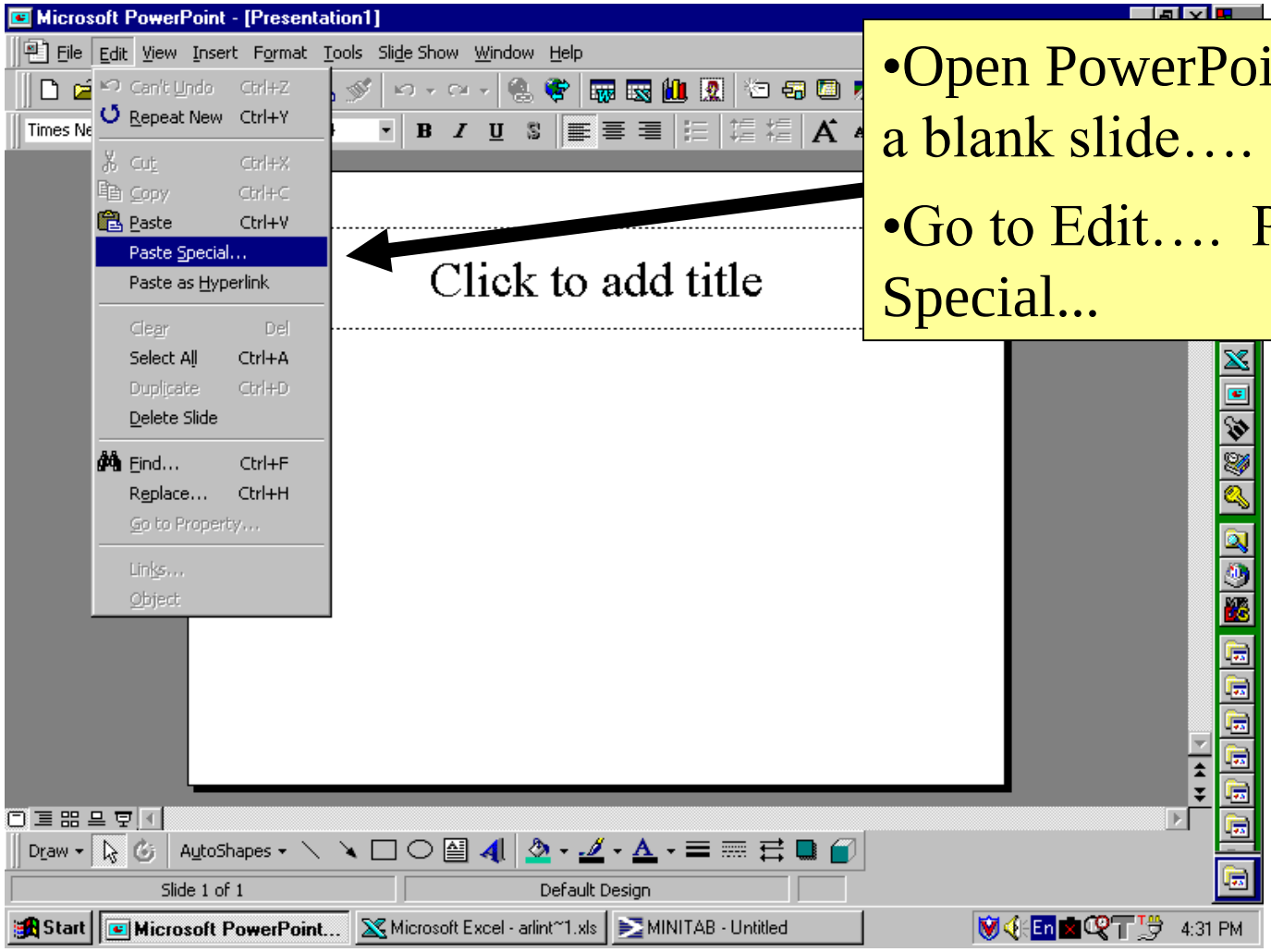
- Minitab now does the analysis and presents the results...
- Note that in this case there is a graph and an analysis summary in the Session Window…
- Let's say we want to use both in our PowerPoint presentation….

Transferring the Analysis...



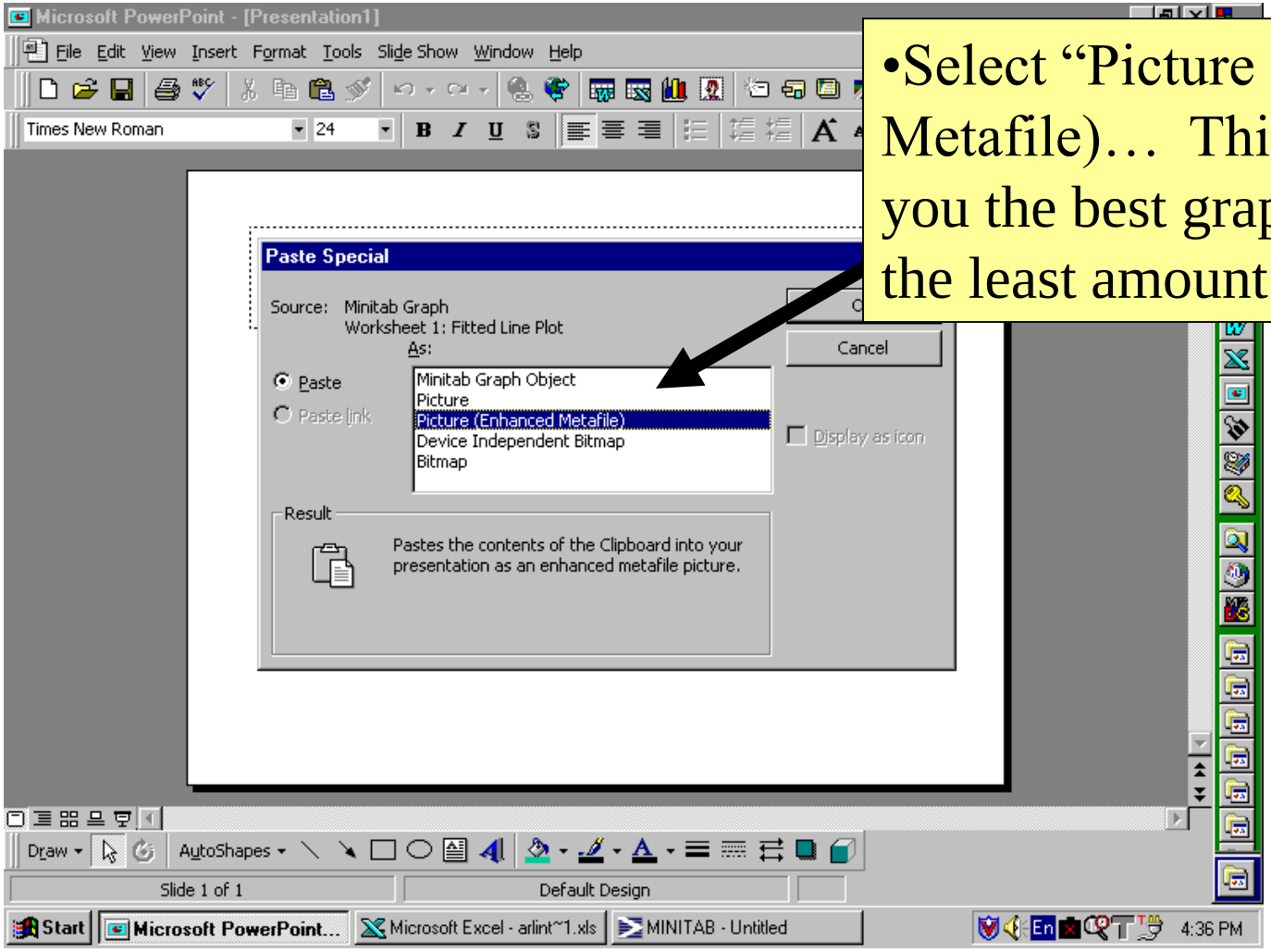
- Let's take care of the graph first….
- Go to Edit…. Copy Graph...

Transferring the Analysis...



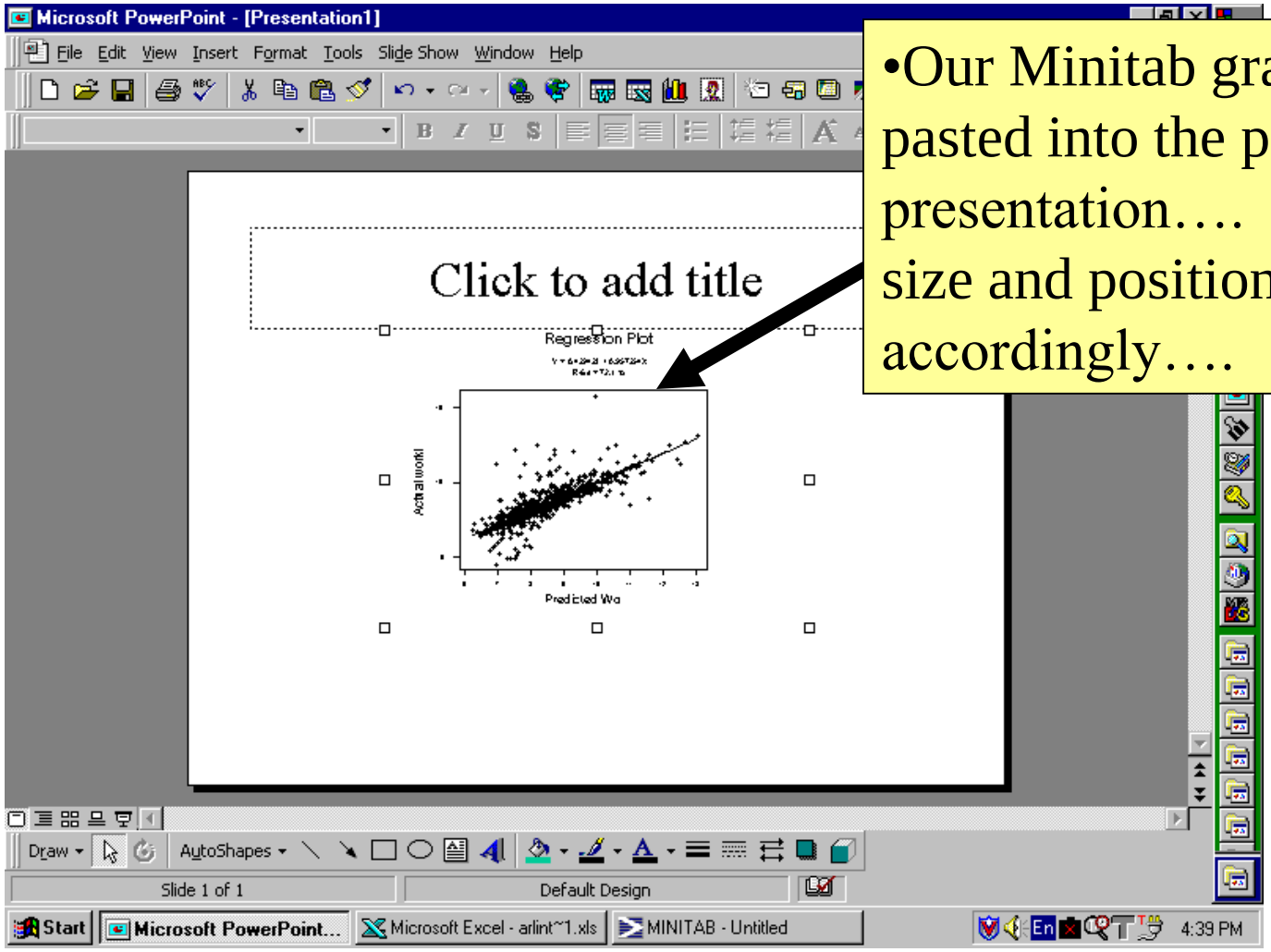
- Open PowerPoint and select a blank slide….
- Go to Edit…. Paste Special...

Transferring the Analysis...



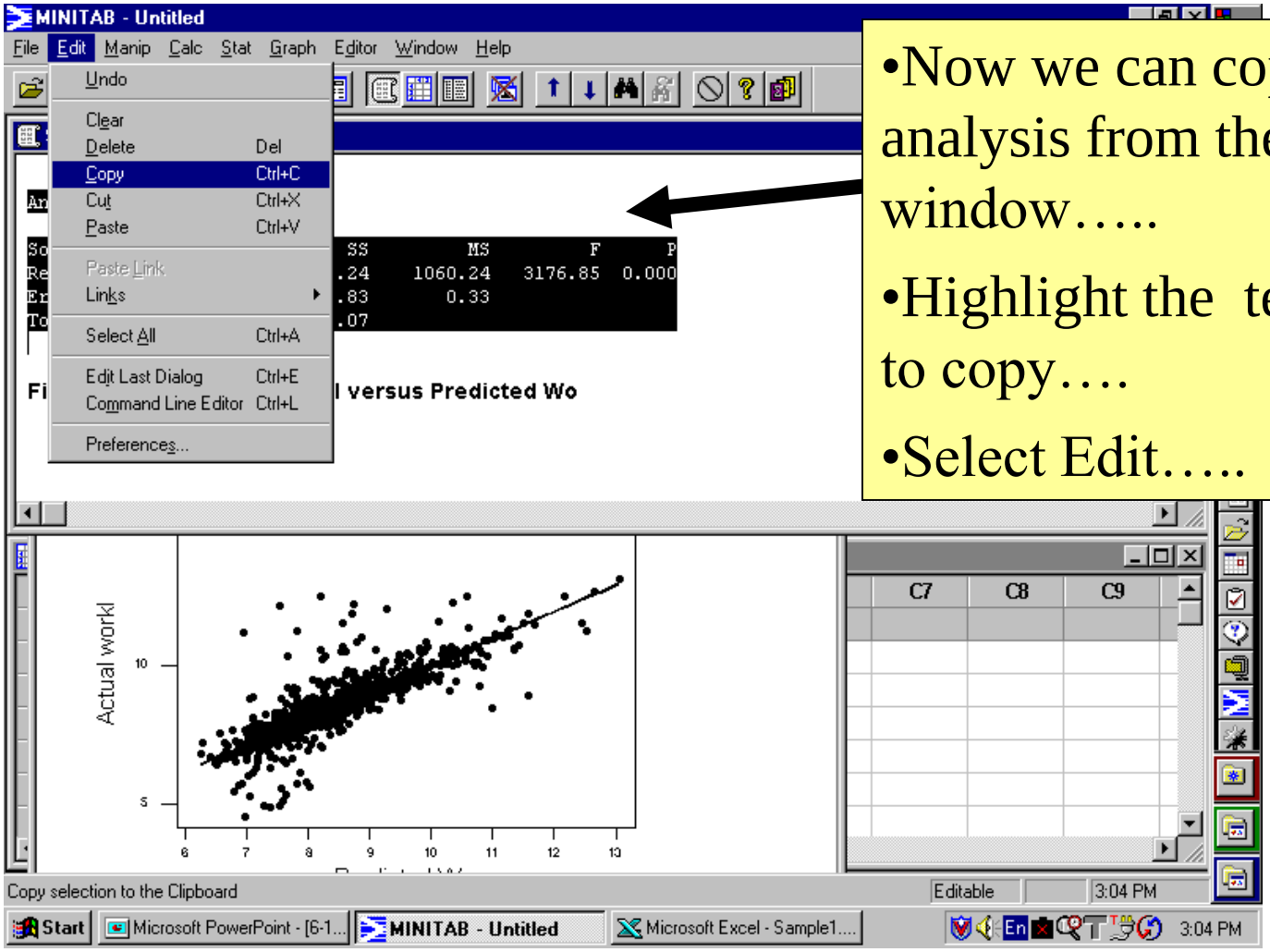
•Select "Picture (Enhanced Metafile)… This will give you the best graphics with the least amount of trouble.

Transferring the Analysis...



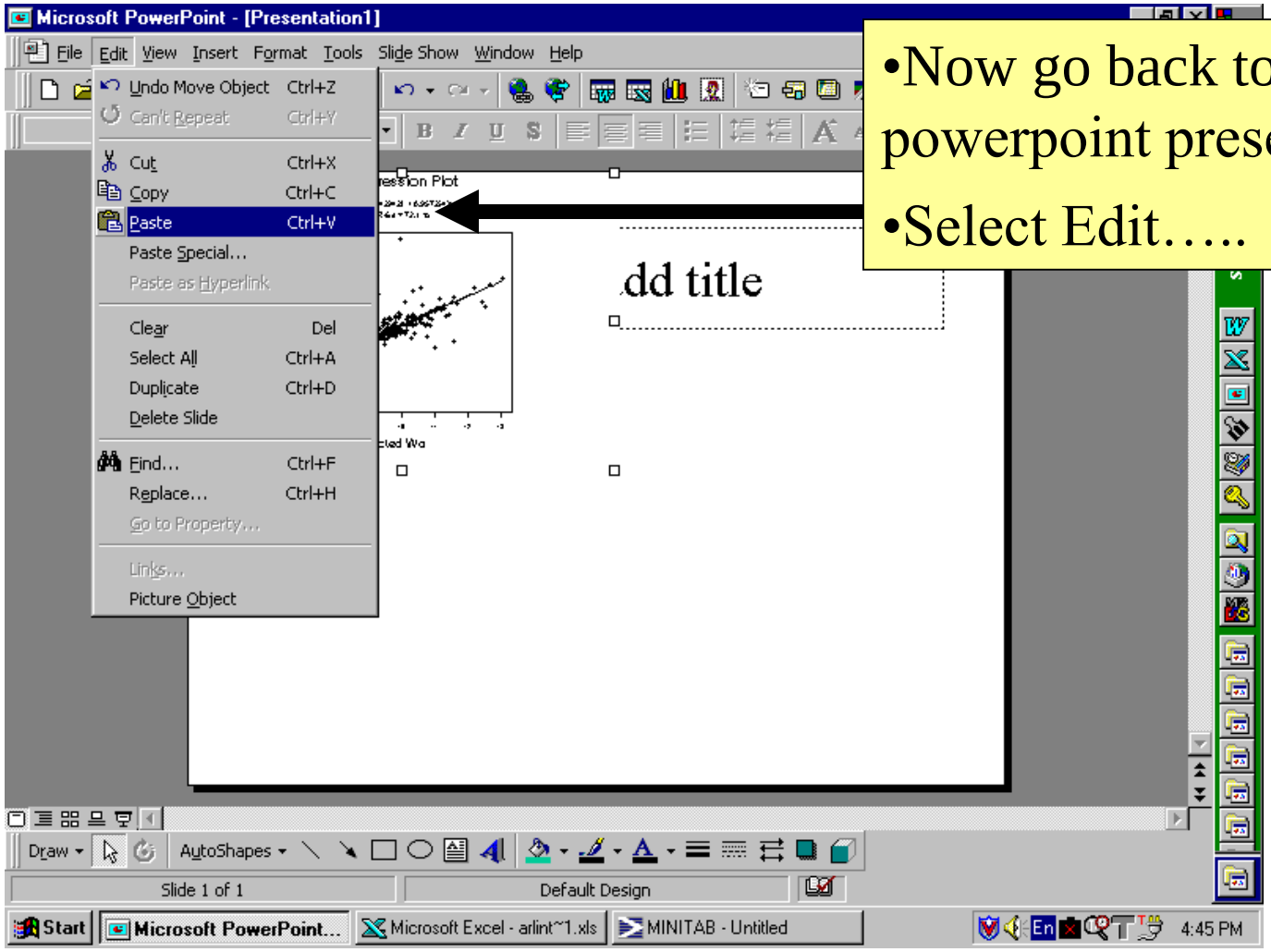
•Our Minitab graph is now pasted into the powerpoint presentation…. We can now size and position it accordingly….

Transferring the Analysis...



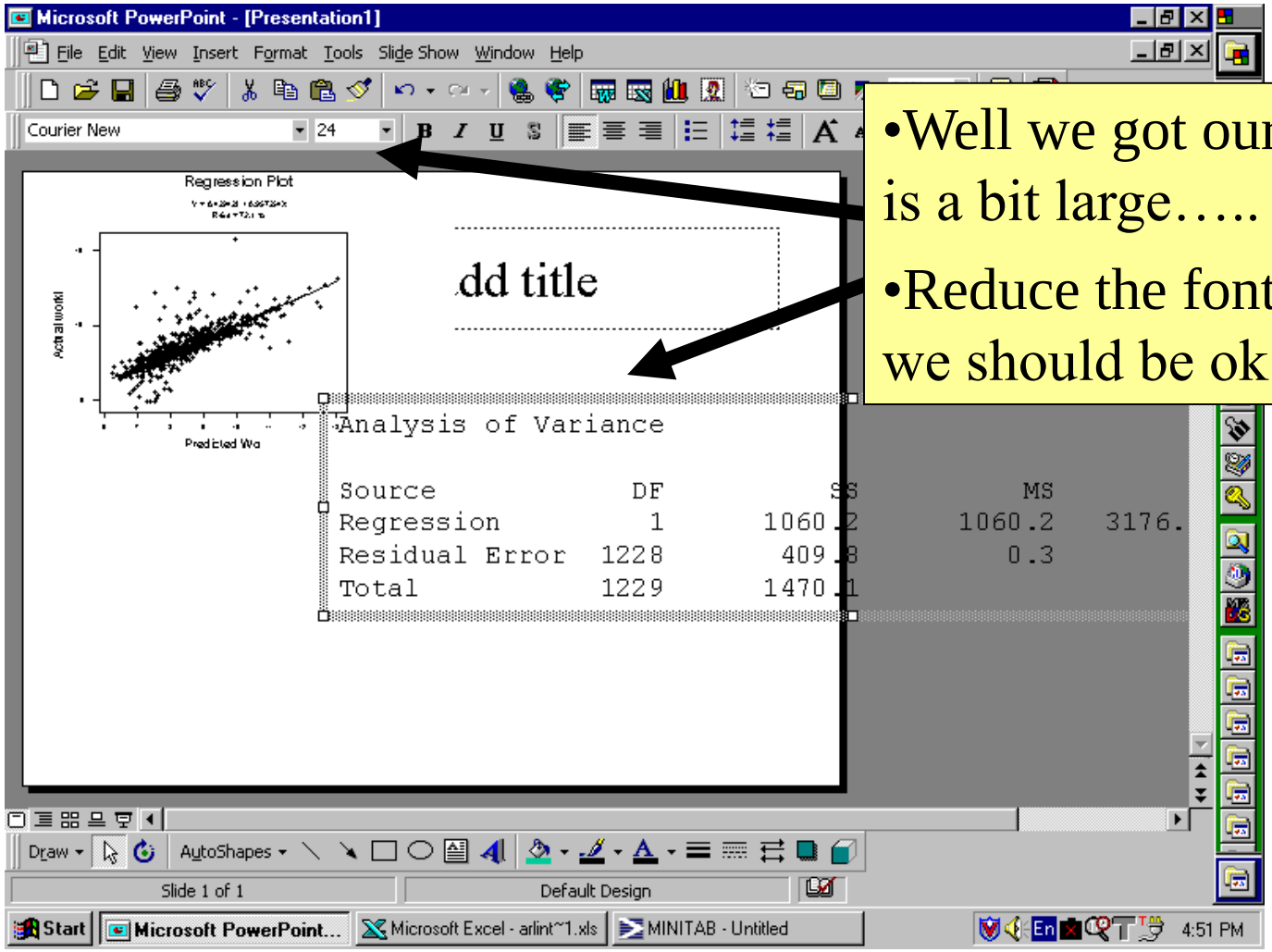
- Now we can copy the analysis from the Session window…..
- Highlight the text you want to copy….
- Select Edit….. Copy…..

Transferring the Analysis...



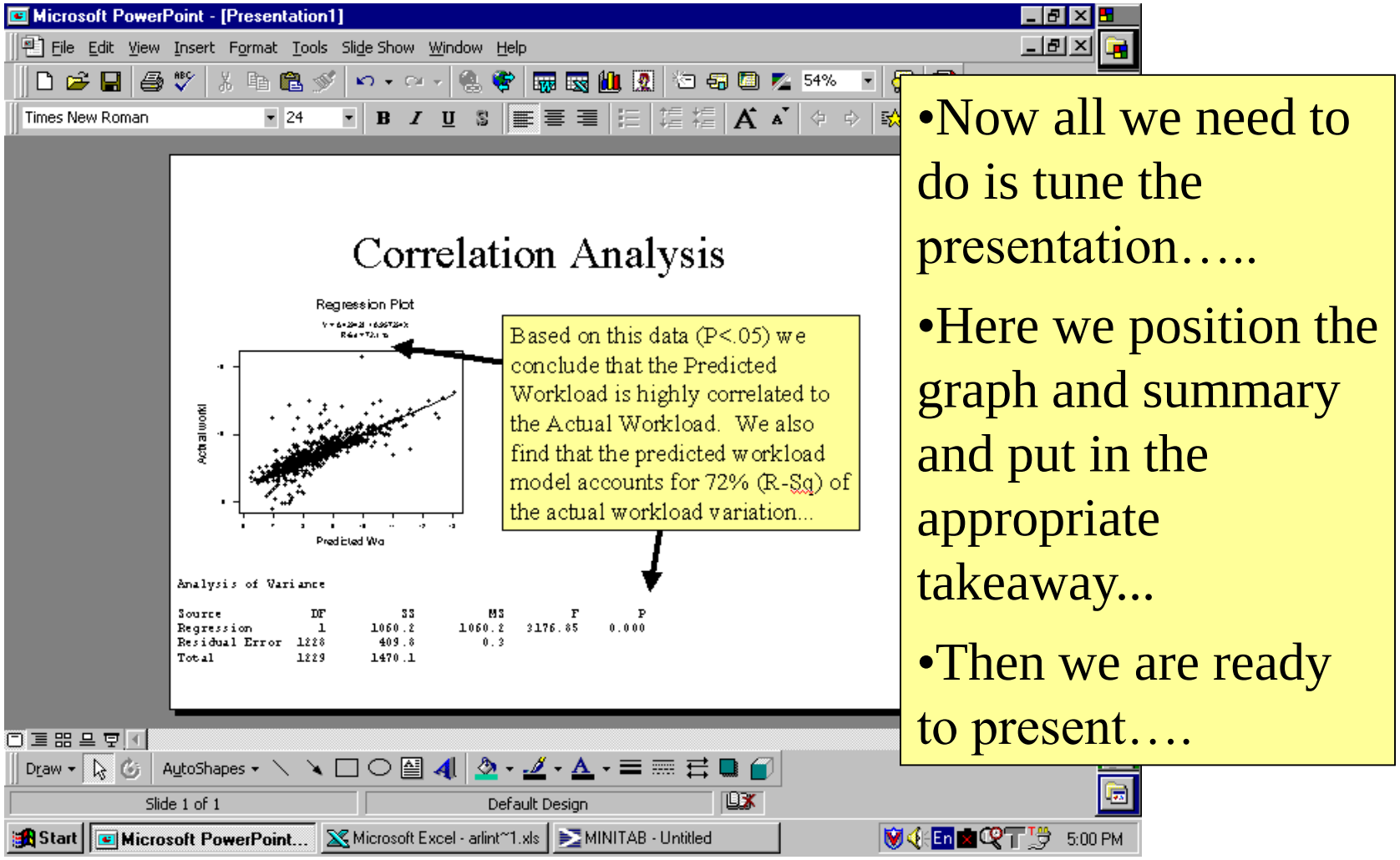
- Now go back to your powerpoint presentation…..
- Select Edit….. Paste…..

Transferring the Analysis...



- Well we got our data, but it is a bit large…..
- Reduce the font to 12 and we should be ok…..

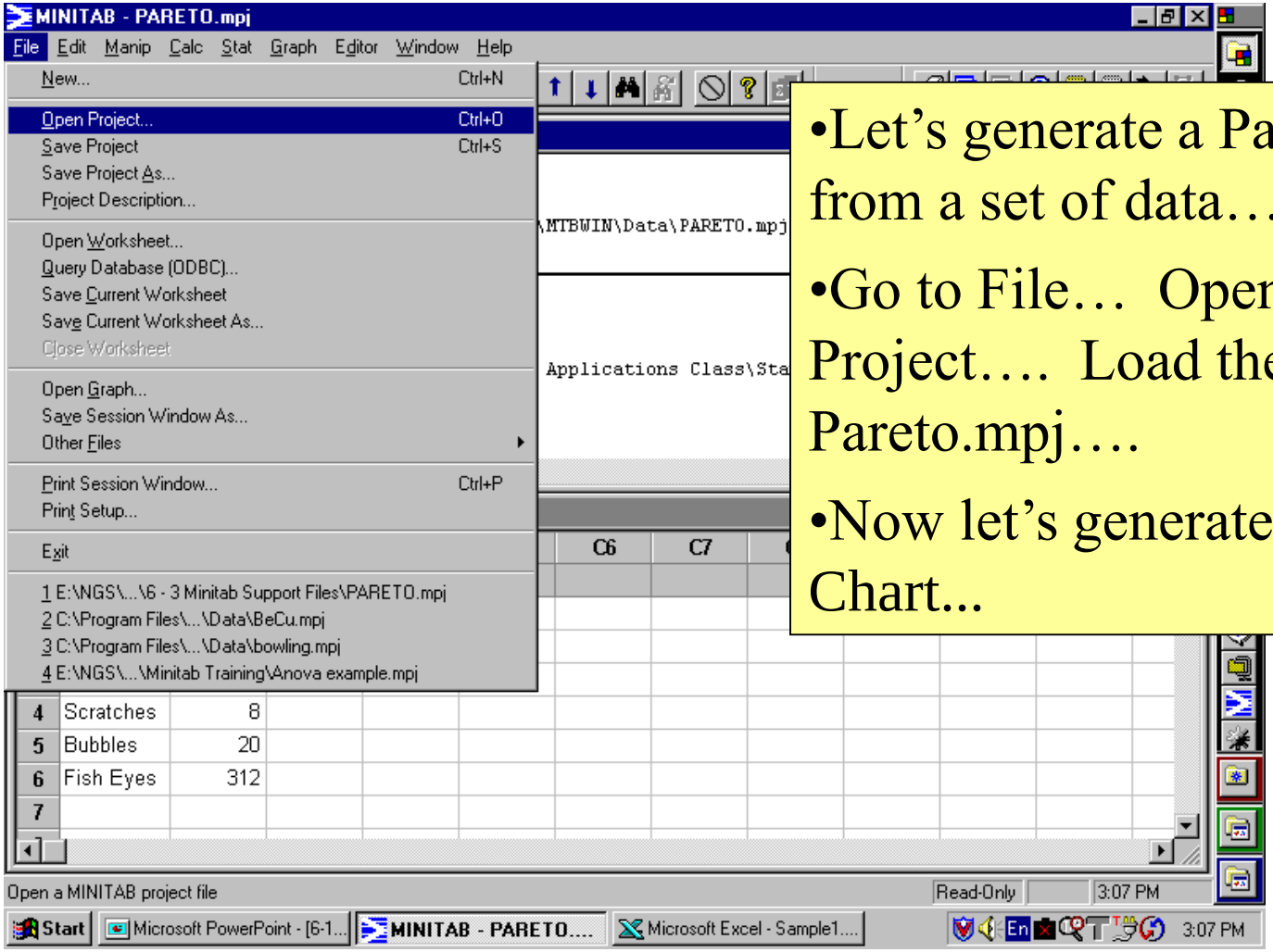
Presenting the results....



- Now all we need to do is tune the presentation…..
- Here we position the graph and summary and put in the appropriate takeaway...
- Then we are ready to present….

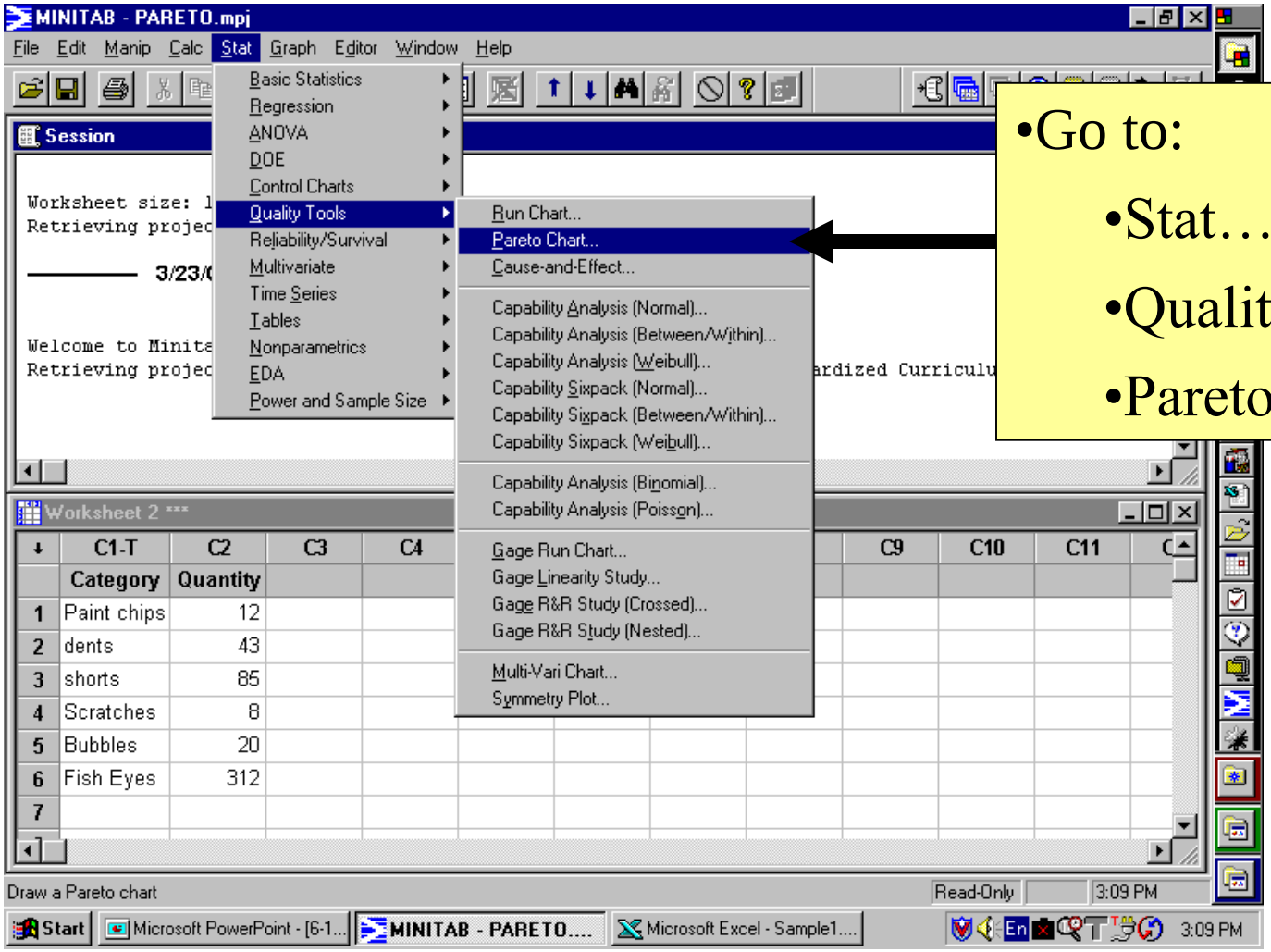
Graphic Capabilities

Pareto Chart....



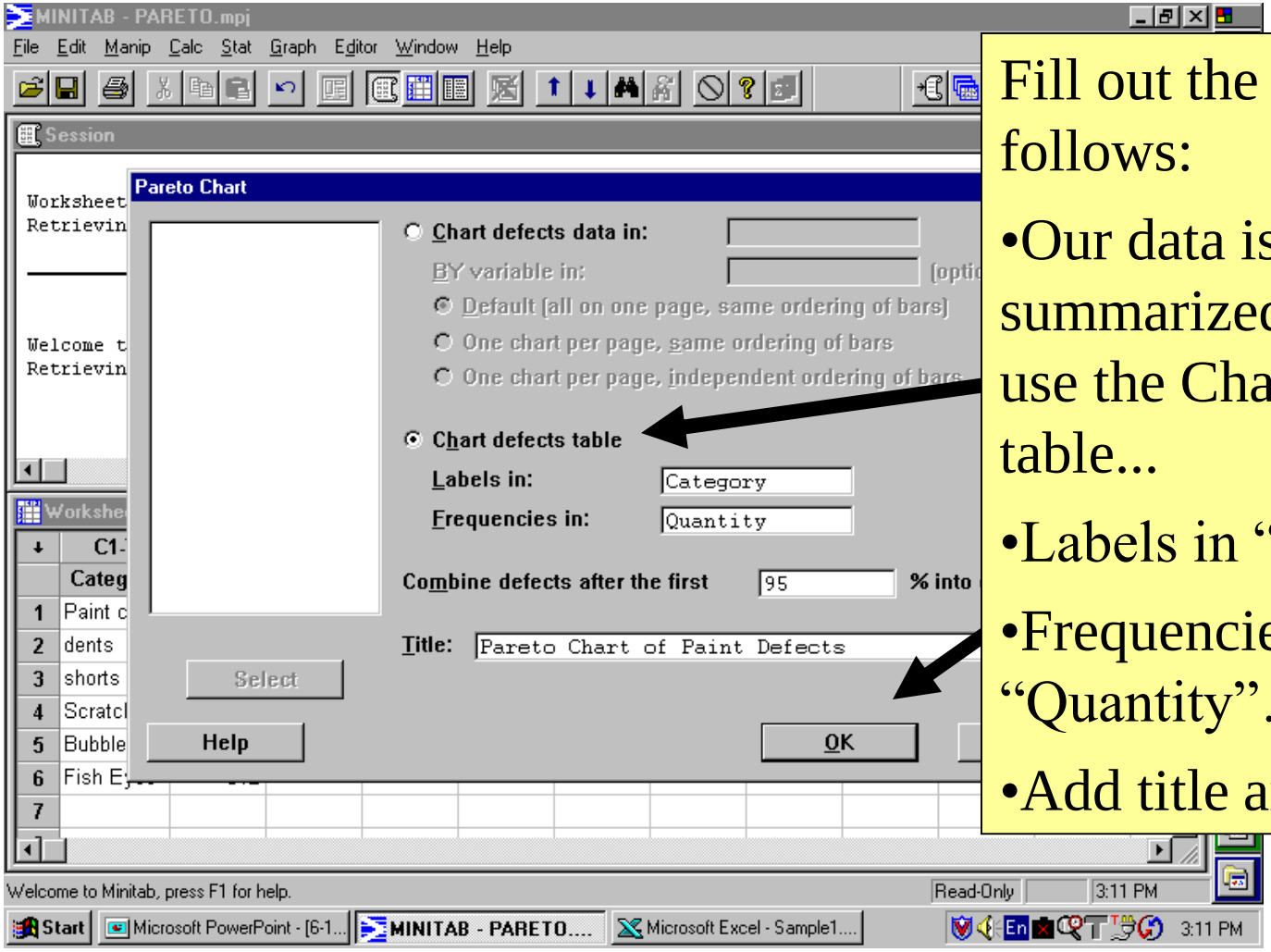
- Let's generate a Pareto Chart from a set of data….
- Go to File… Open Project…. Load the file Pareto.mpj….
- Now let's generate the Pareto Chart...

Pareto Chart....



- Go to:
  - Stat…
  - Quality Tools…
  - Pareto Chart….

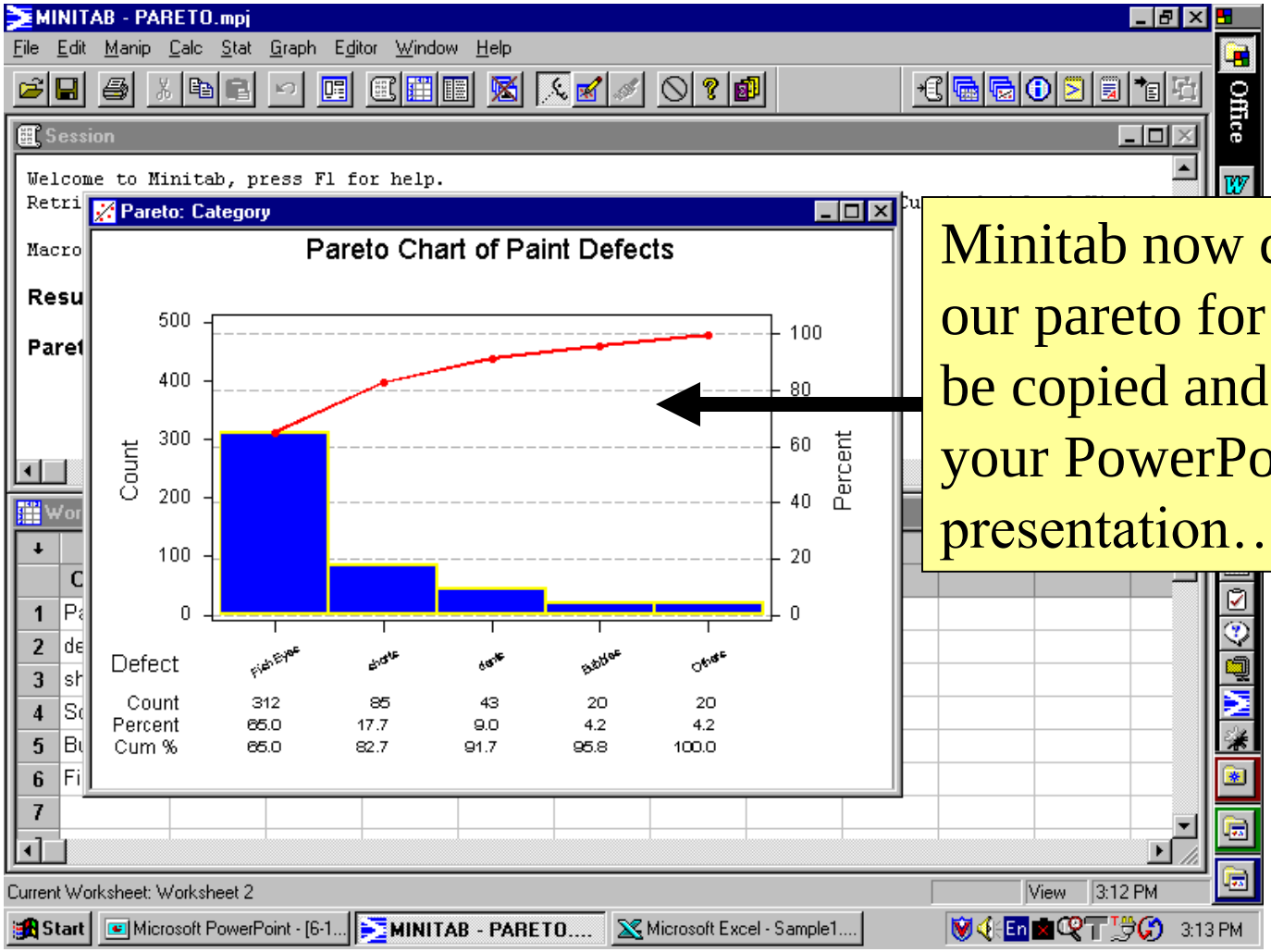
Pareto Chart....



Fill out the screen as follows:

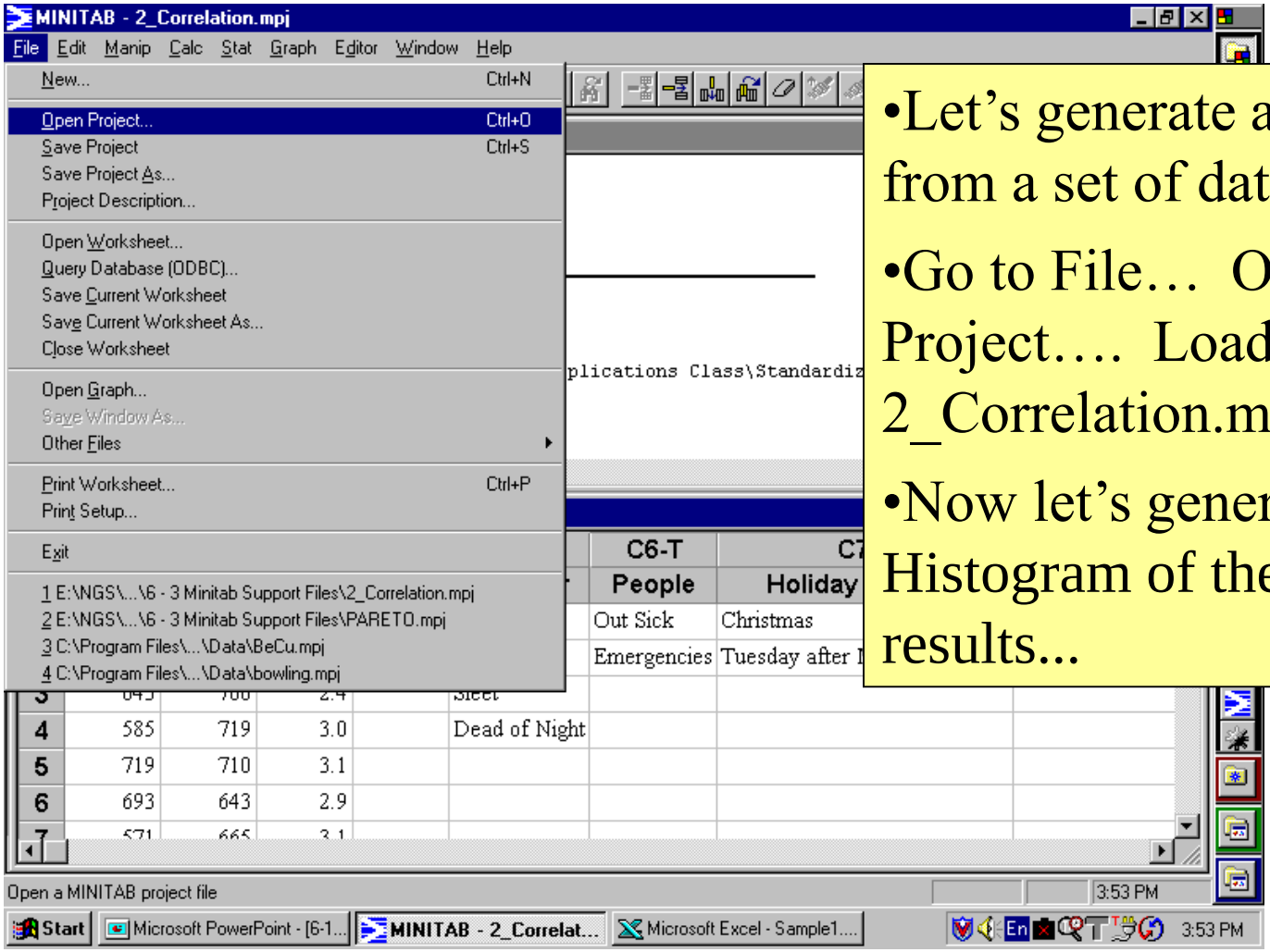
- Our data is already summarized so we will use the Chart Defects table...
- Labels in “Category”…
- Frequencies in “Quantity”….
- Add title and hit OK..

Pareto Chart....



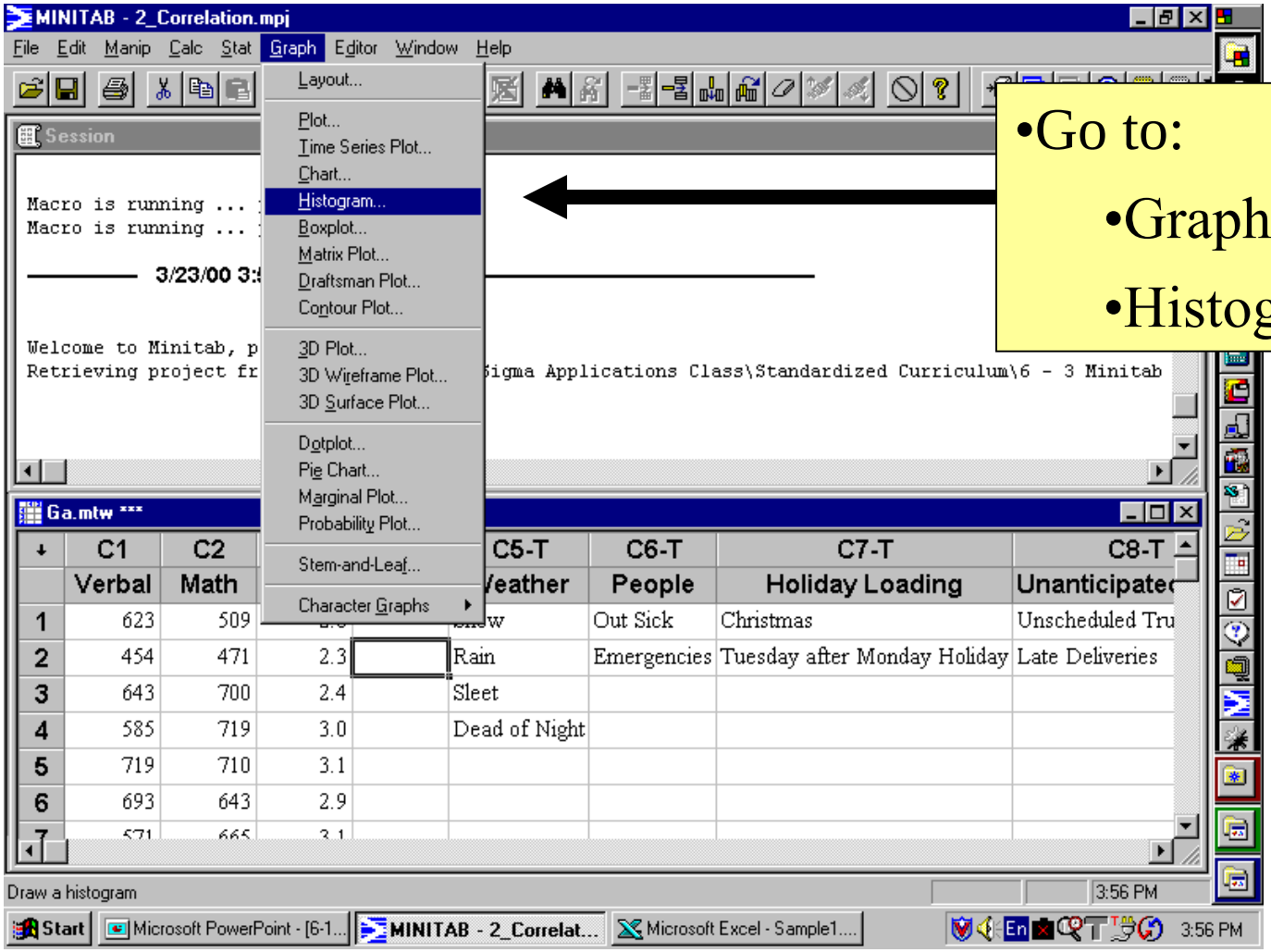
Minitab now completes our pareto for us ready to be copied and pasted into your PowerPoint presentation….

Histogram....



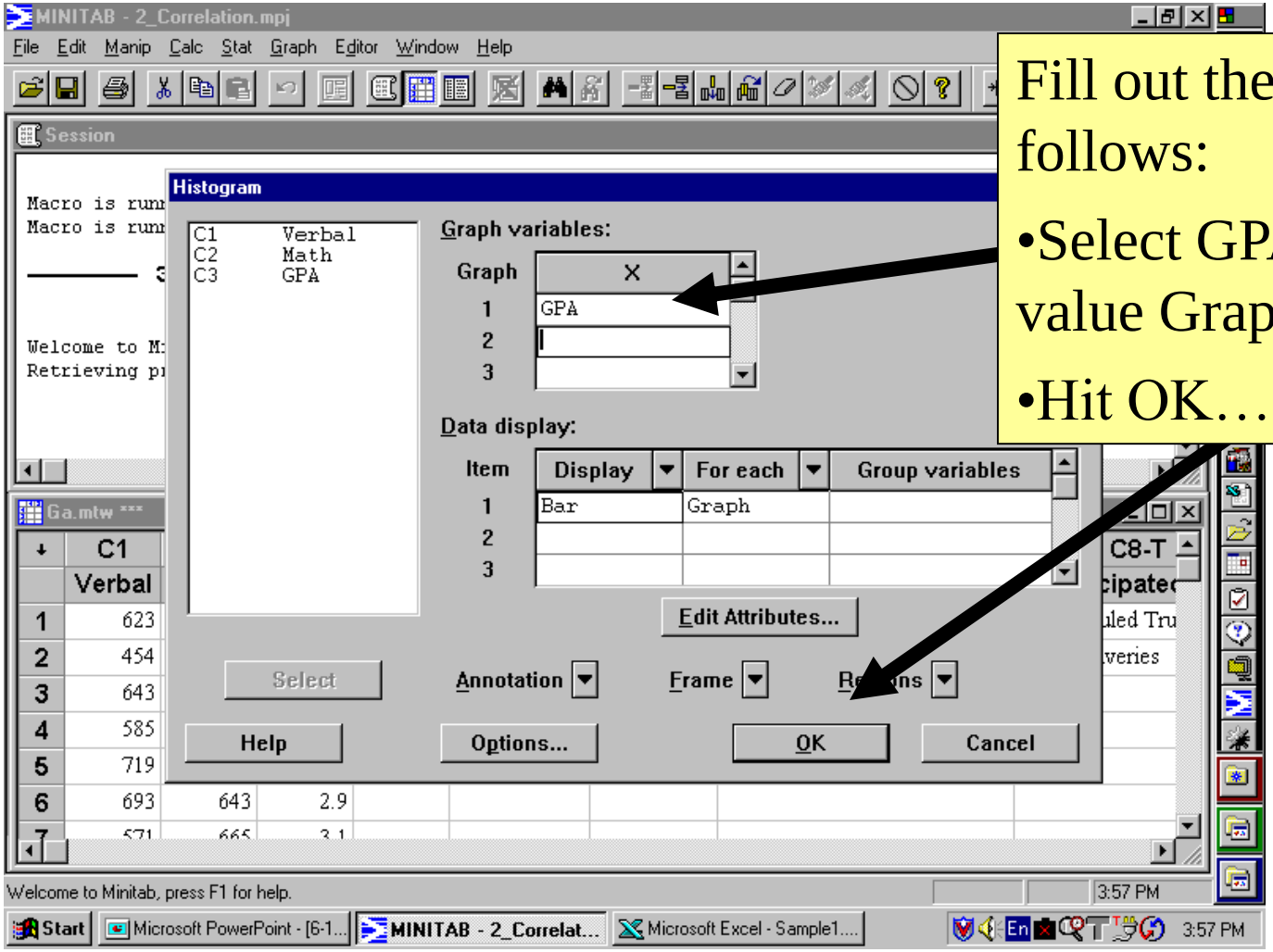
- Let's generate a Histogram from a set of data….
- Go to File… Open Project…. Load the file 2_Correlation.mpj….
- Now let's generate the Histogram of the GPA results...

Histogram....



- Go to:
  - Graph…
  - Histogram…

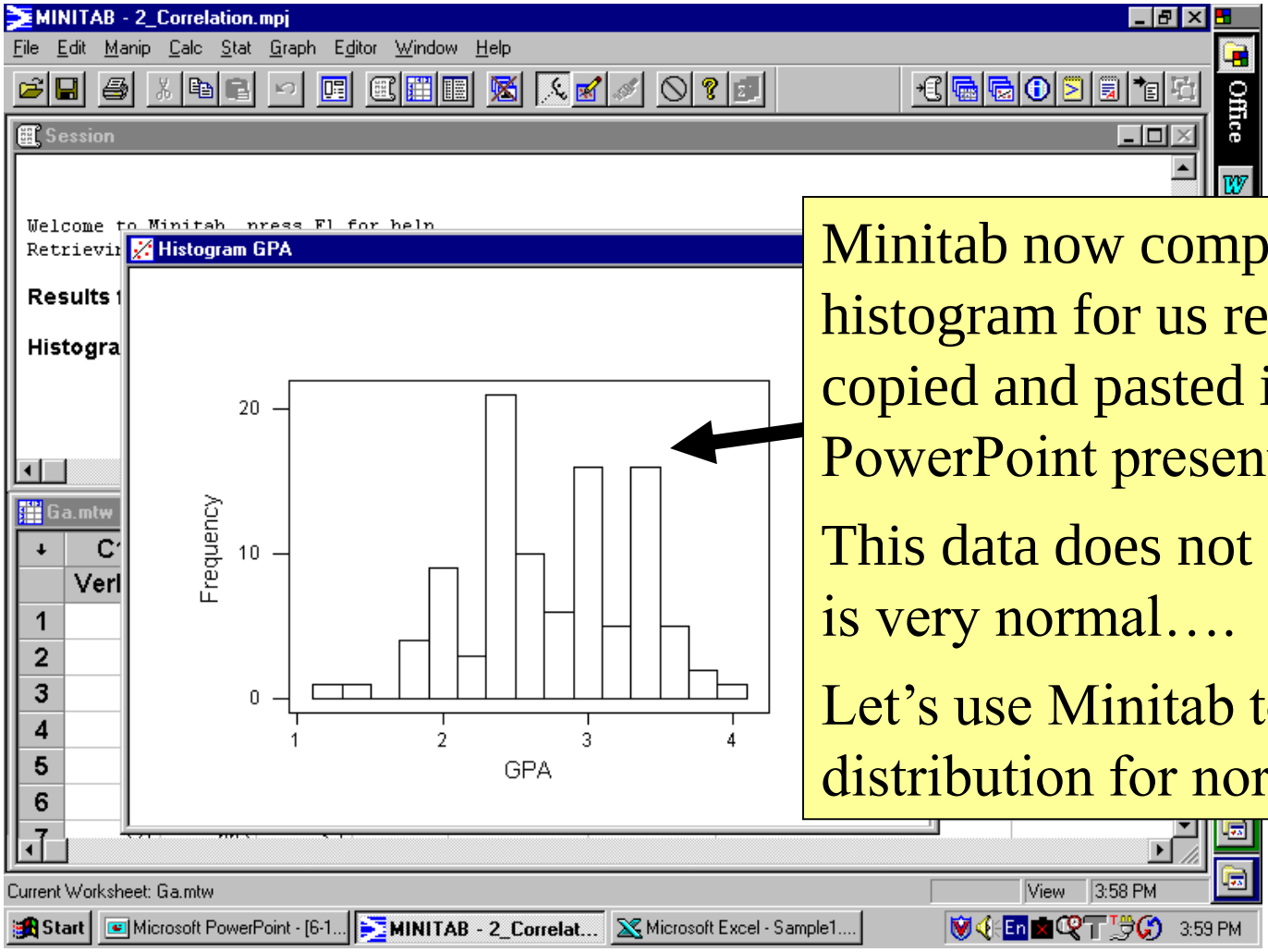
Histogram....



Fill out the screen as follows:

- Select GPA for our X value Graph Variable
- Hit OK…..

Histogram....

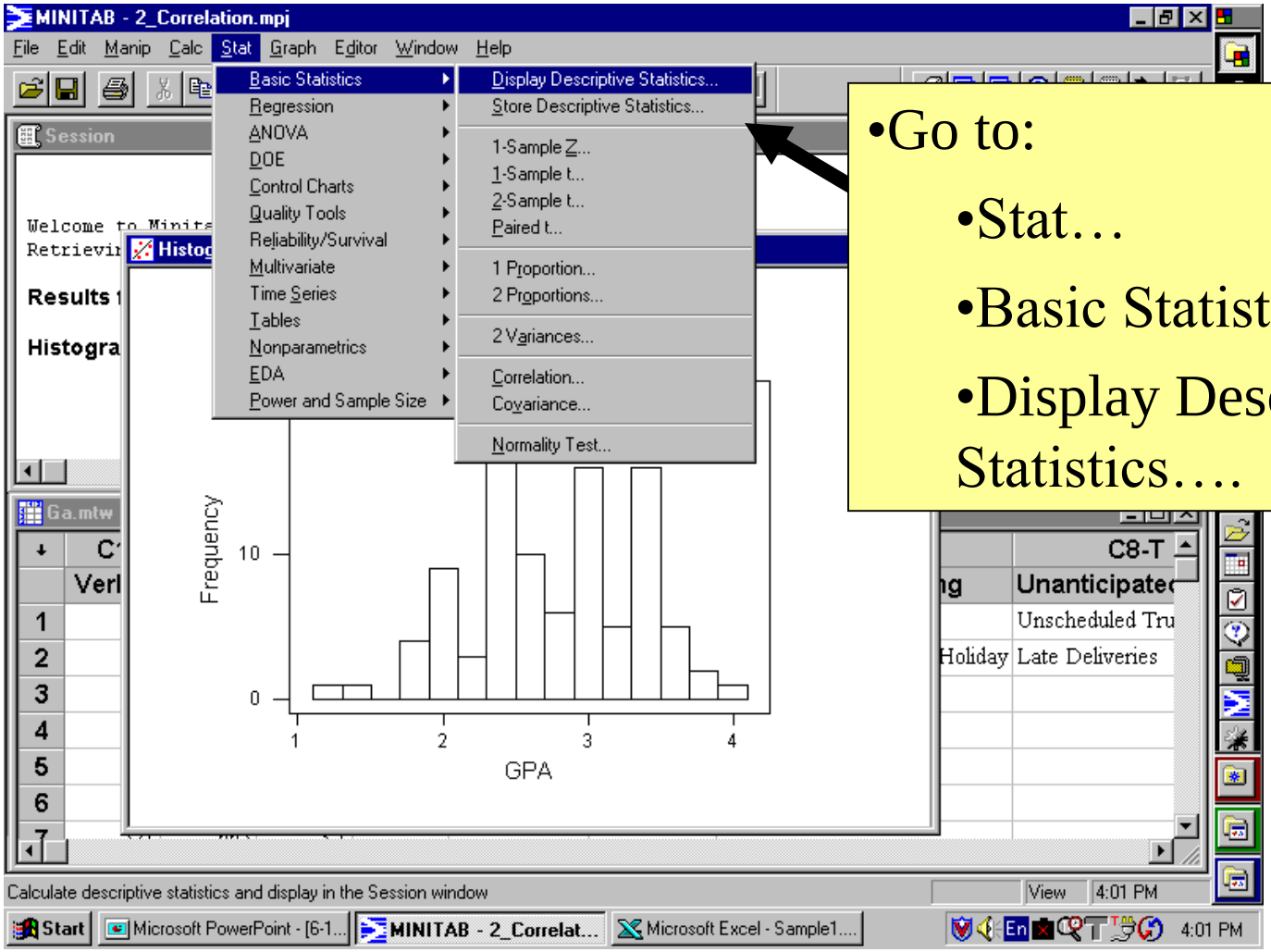


Minitab now completes our histogram for us ready to be copied and pasted into your PowerPoint presentation….

This data does not look like it is very normal….

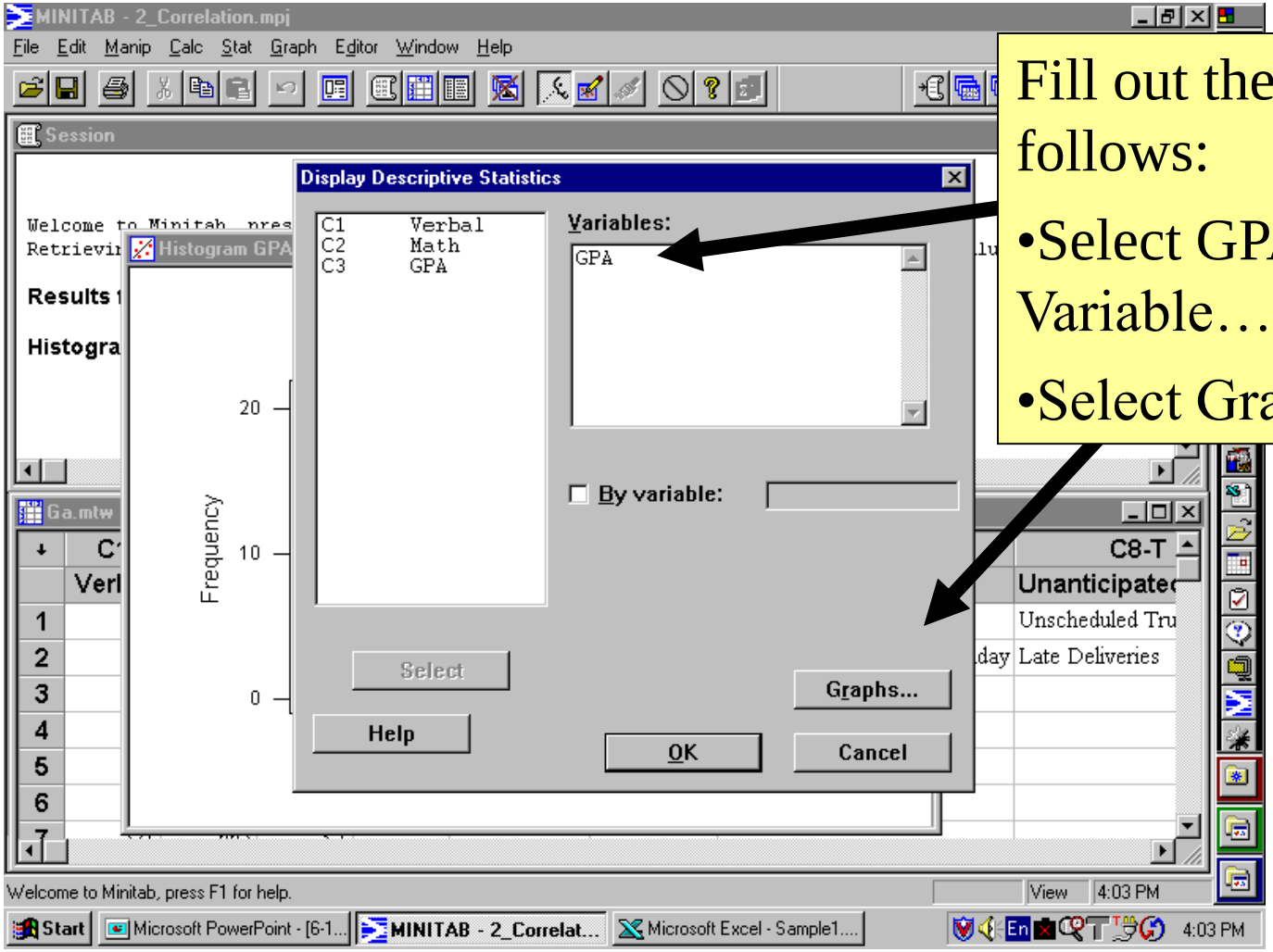
Let's use Minitab to test this distribution for normality……

Histogram....



- Go to:
  - Stat…
  - Basic Statistics…
  - Display Descriptive Statistics….

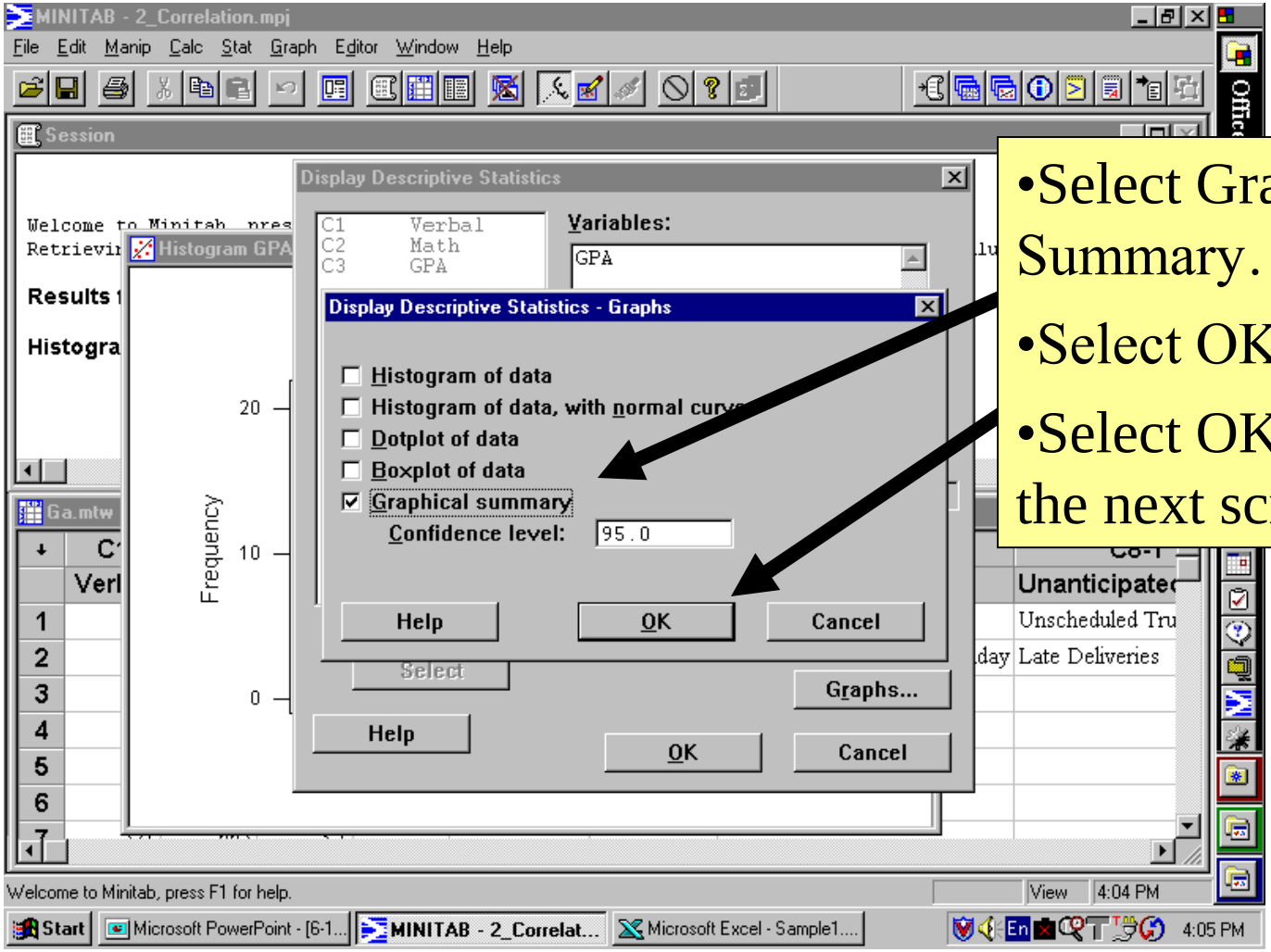
Histogram....



Fill out the screen as follows:

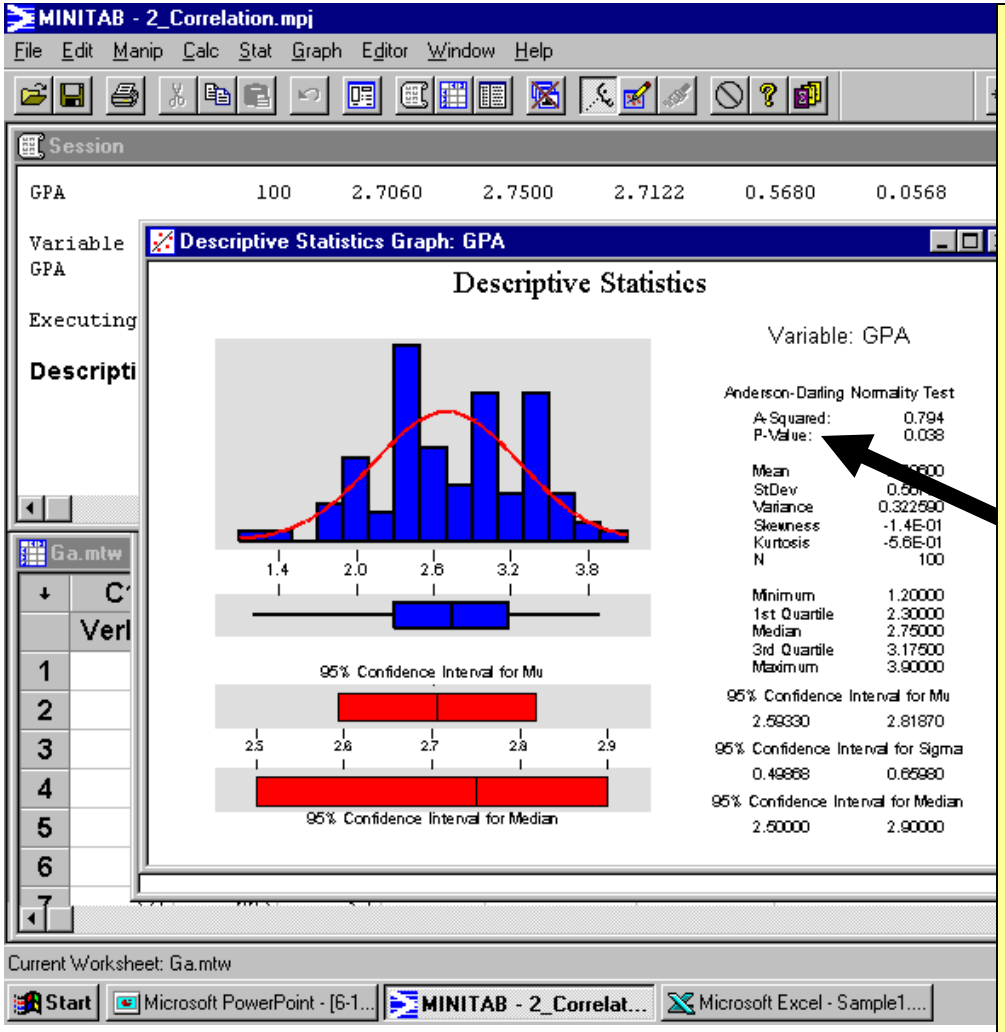
- Select GPA for our Variable….
- Select Graphs…..

Histogram....



- Select Graphical Summary….
- Select OK…..
- Select OK again on the next screen...

Histogram....

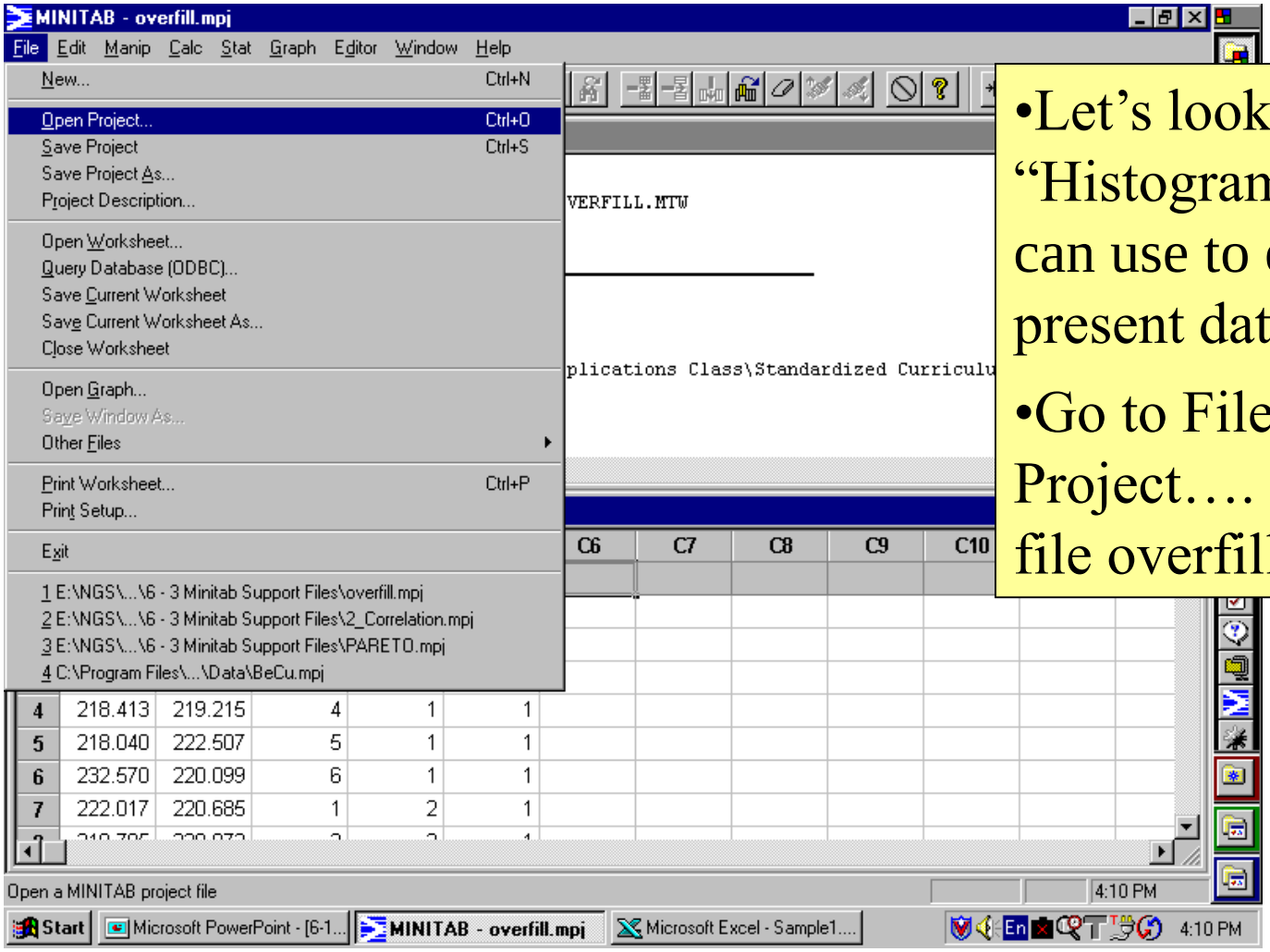


Note that now we not only have our Histogram but a number of other descriptive statistics as well….

This is a great summary slide...

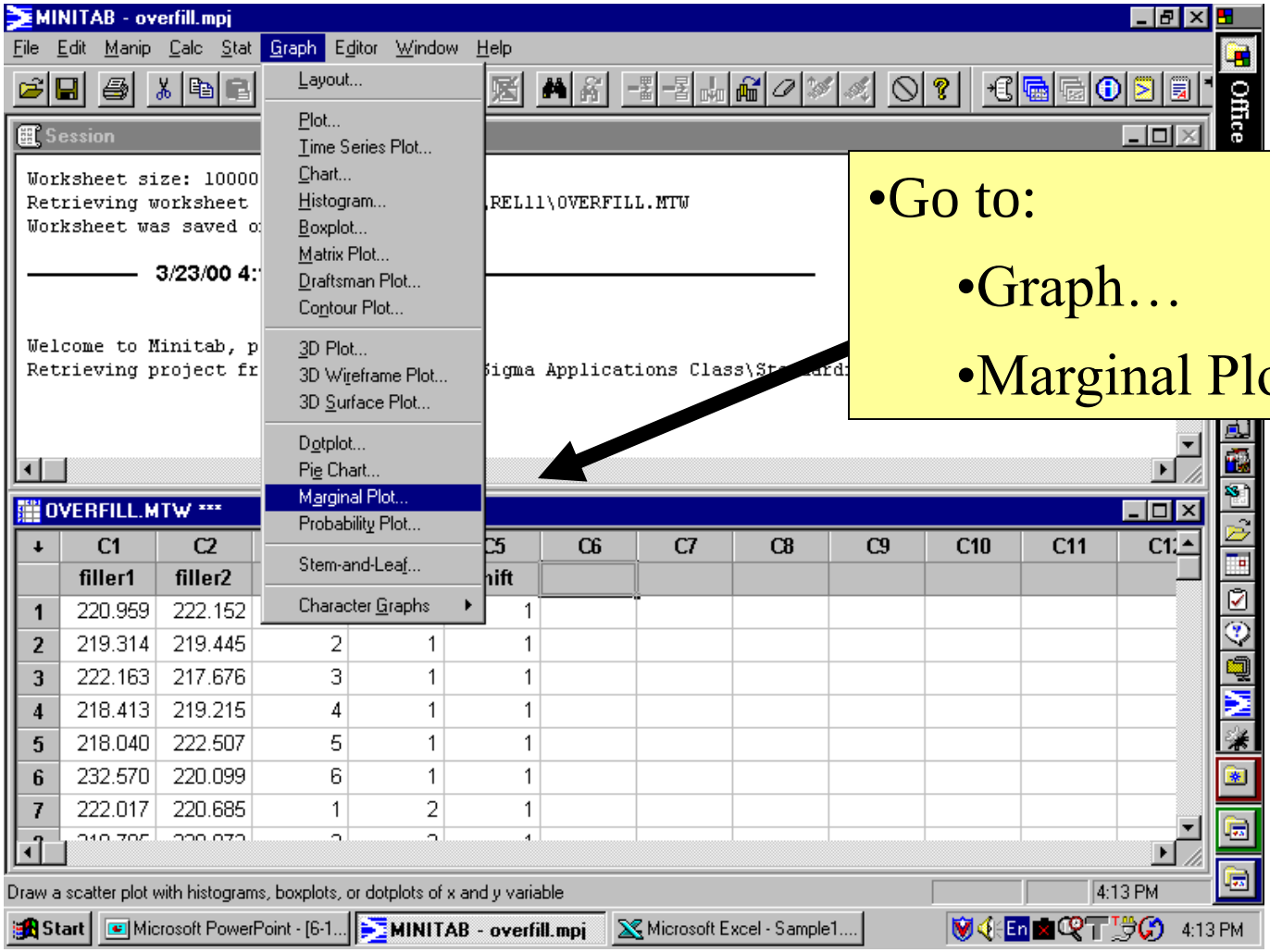
As for the normality question, note that our P value of .038 rejects the null hypothesis (P<.05). So, we conclude with 95% confidence that the data is not normal…..

Histogram....



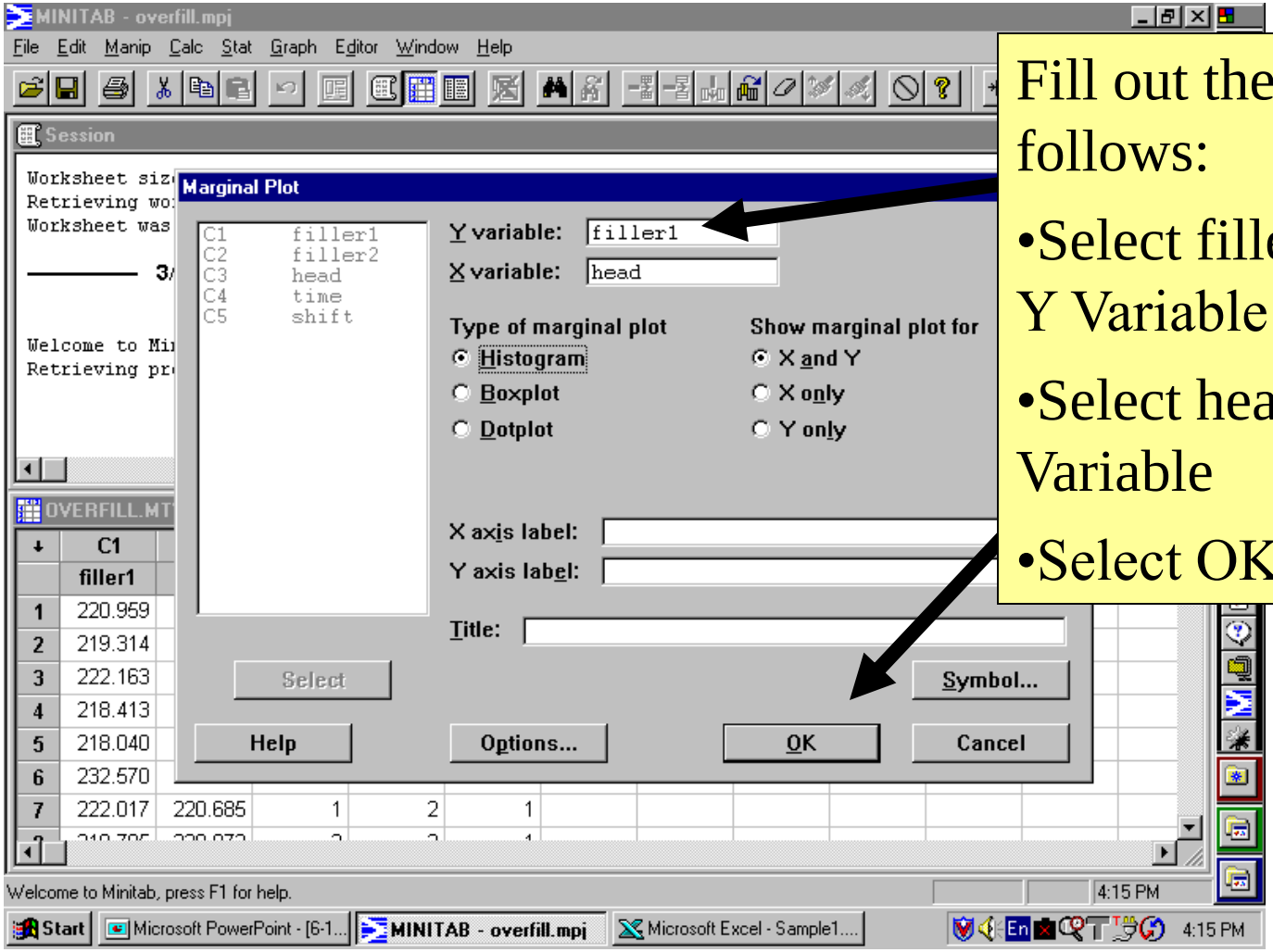
- Let's look at another "Histogram" tool we can use to evaluate and present data….
- Go to File… Open Project…. Load the file overfill.mpj….

Histogram....



- Go to:
  - Graph…
  - Marginal Plot…

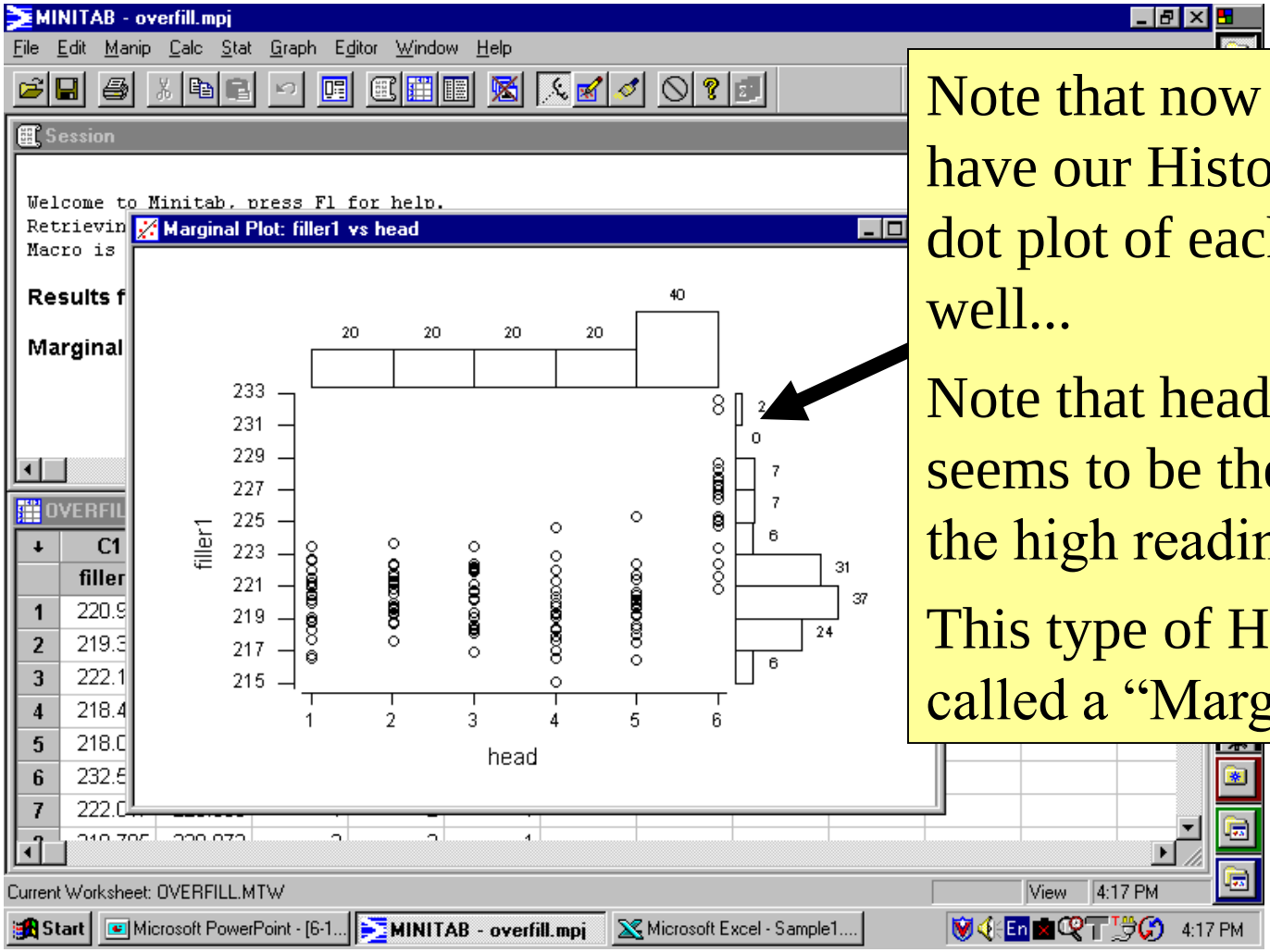
Histogram....



Fill out the screen as follows:

- Select filler 1 for the Y Variable….
- Select head for the X Variable
- Select OK…..

Histogram....

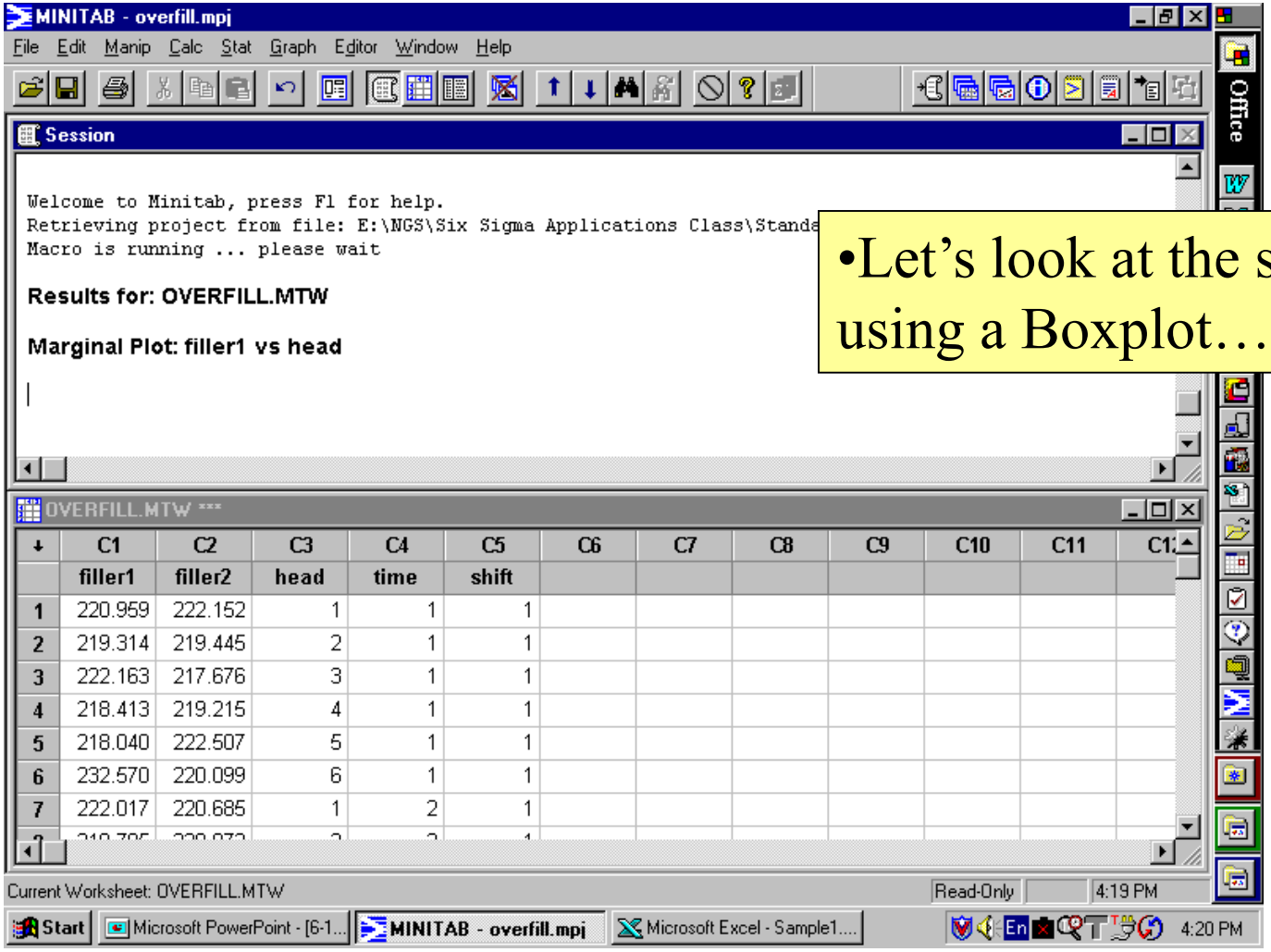


Note that now we not only have our Histogram but a dot plot of each head data as well...

Note that head number 6 seems to be the source of the high readings…..

This type of Histogram is called a “Marginal Plot”..

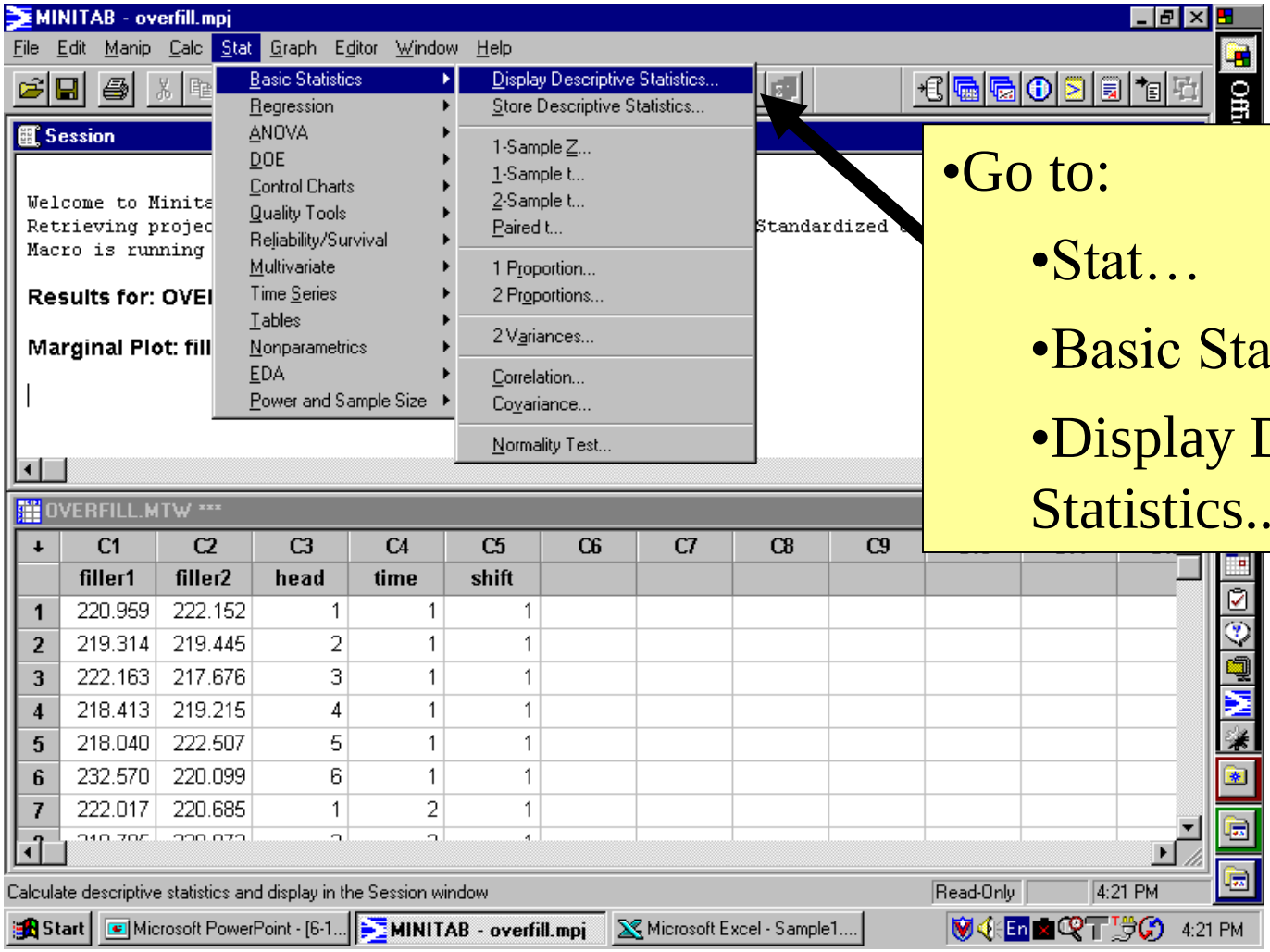
Boxplot....



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
	filler1	filler2	head	time	shift							
1	220.959	222.152	1	1	1							
2	219.314	219.445	2	1	1							
3	222.163	217.676	3	1	1							
4	218.413	219.215	4	1	1							
5	218.040	222.507	5	1	1							
6	232.570	220.099	6	1	1							
7	222.017	220.685	1	2	1							

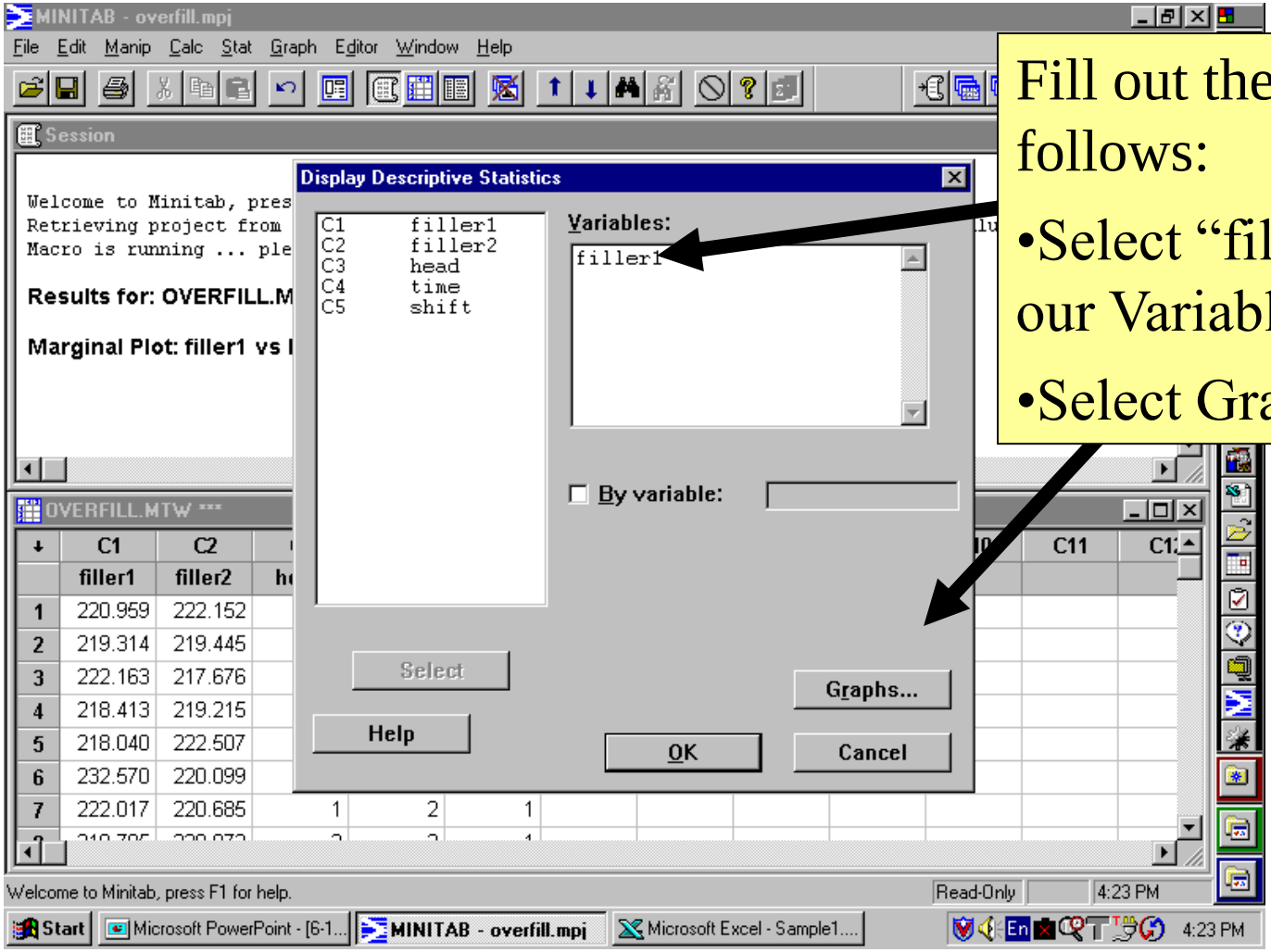
- Let's look at the same data using a Boxplot….

Boxplot....



- Go to:
  - Stat…
  - Basic Statistics…
  - Display Descriptive Statistics...

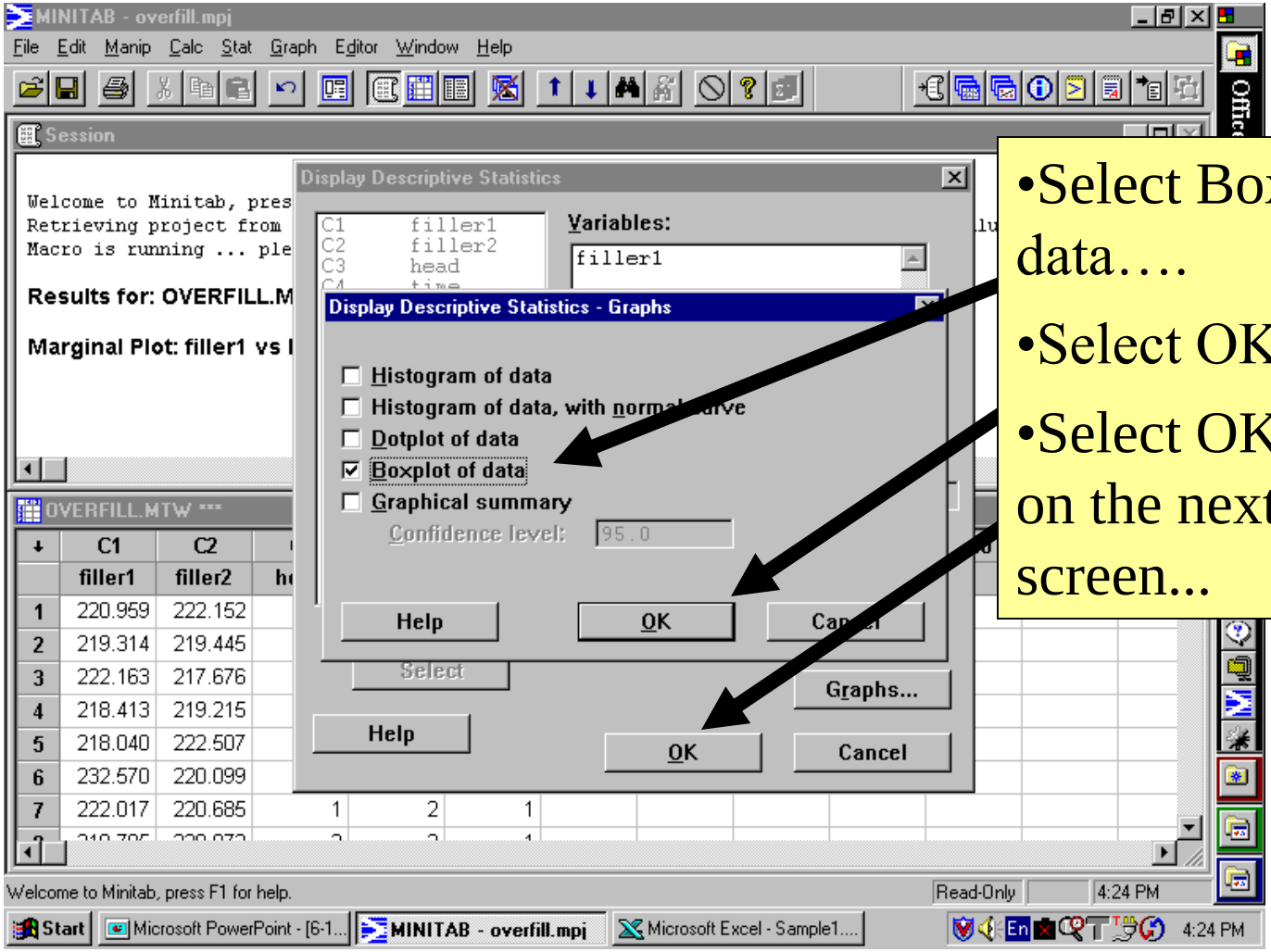
Boxplot....



Fill out the screen as follows:

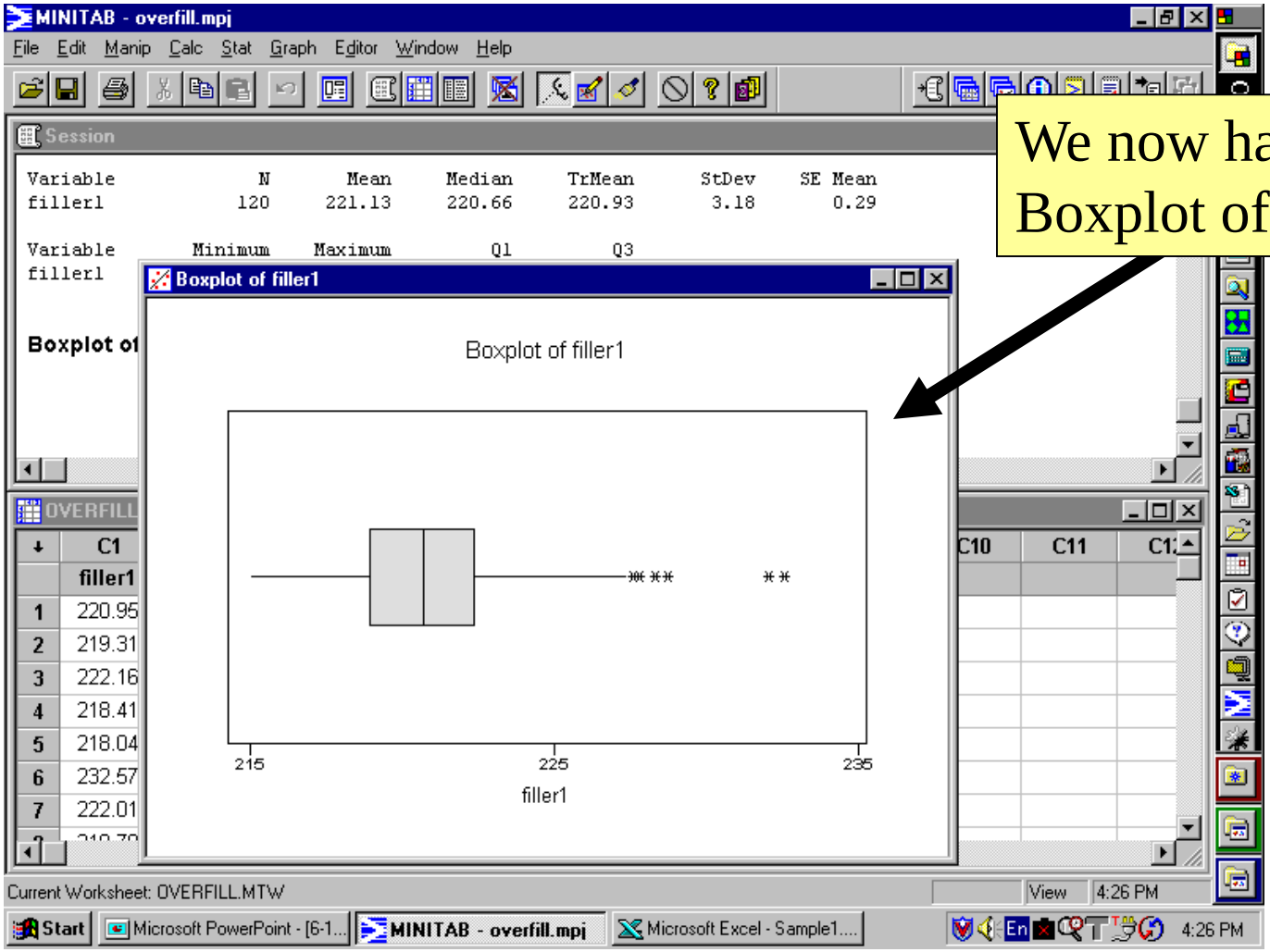
- Select “filler 1” for our Variable….
- Select Graphs…..

Boxplot....



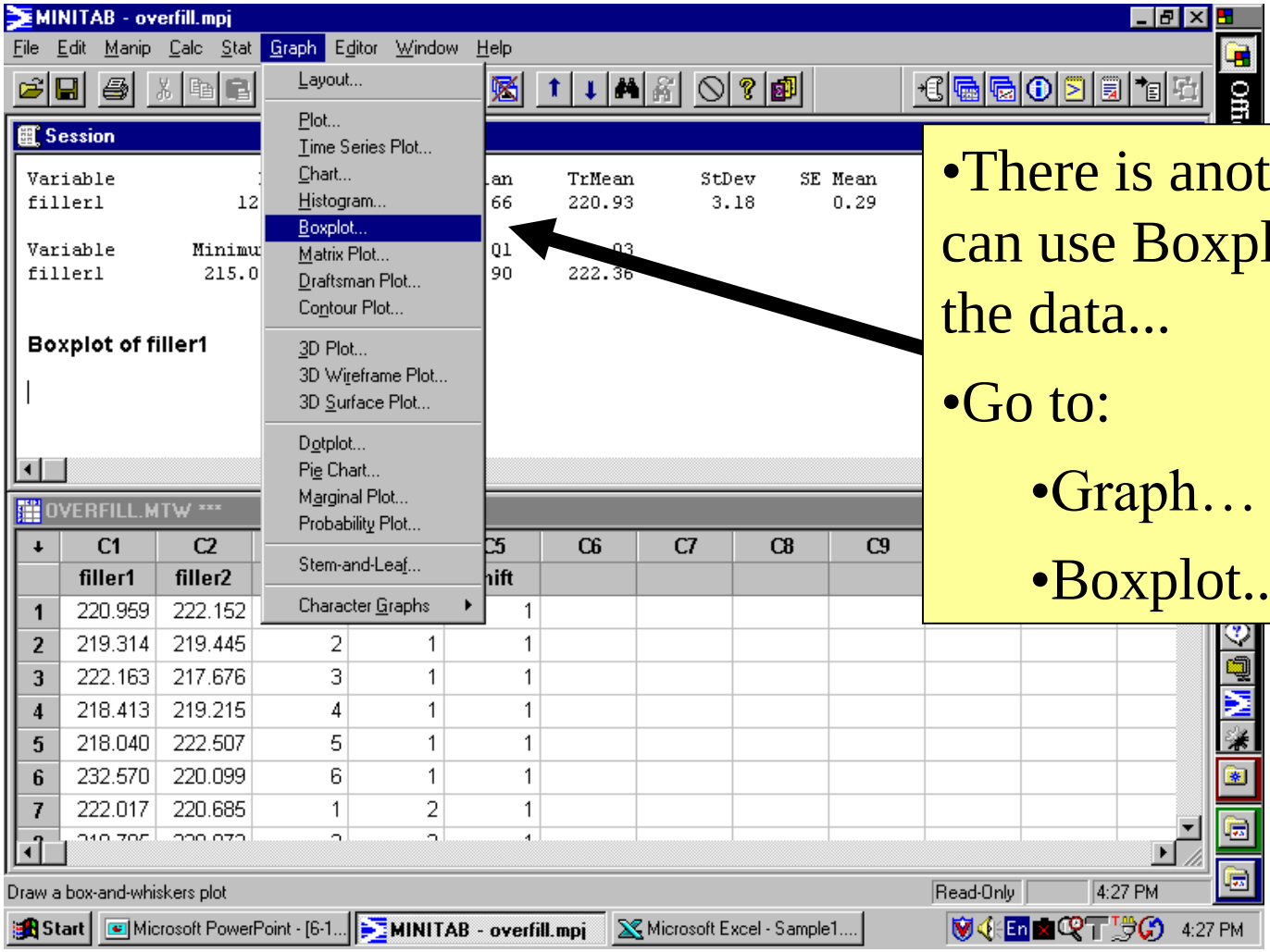
- Select Boxplot of data….
- Select OK…..
- Select OK again on the next screen...

Boxplot....



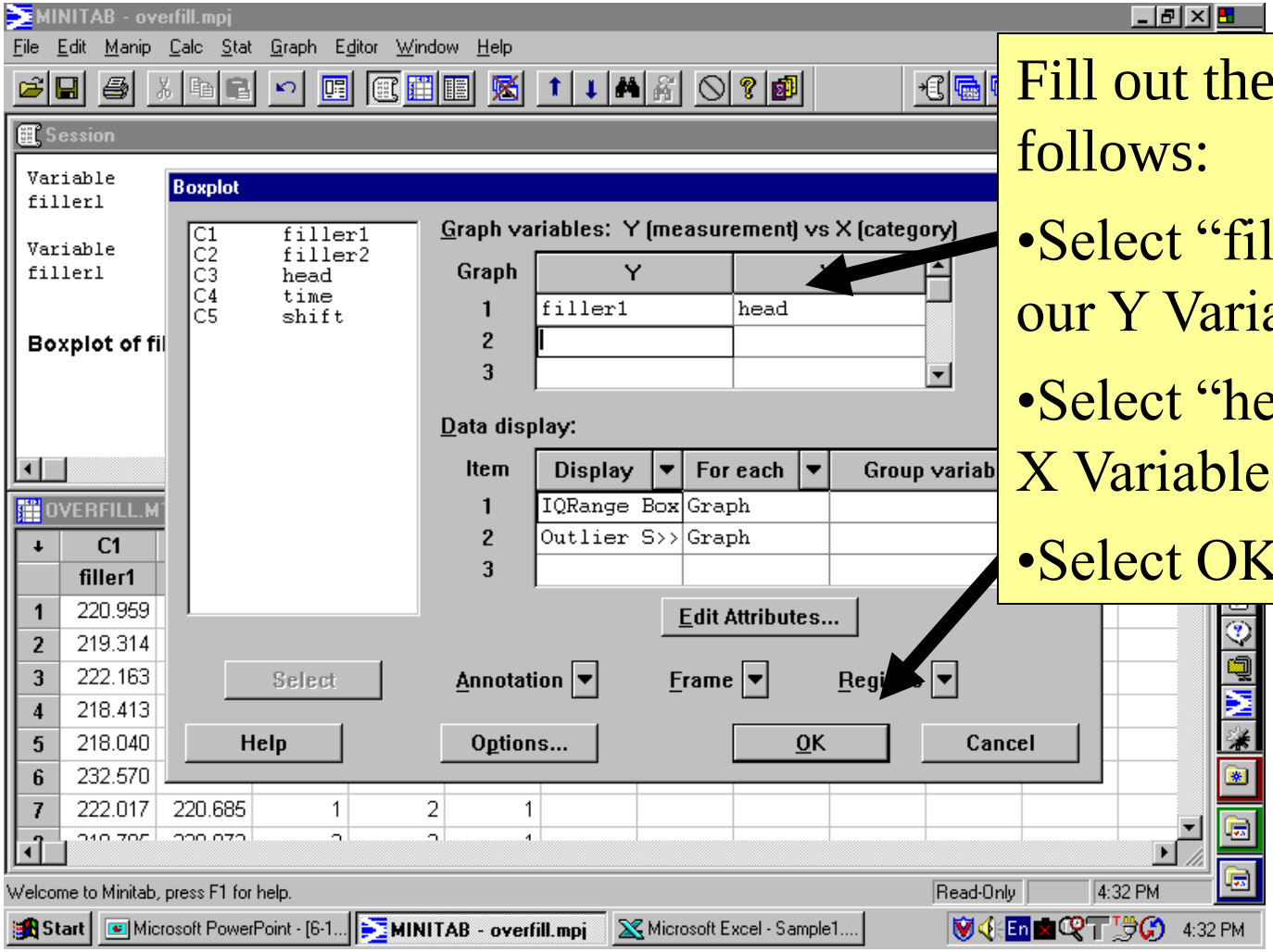
We now have our Boxplot of the data...

Boxplot....



- There is another way we can use Boxplots to view the data...
- Go to:
  - Graph…
  - Boxplot...

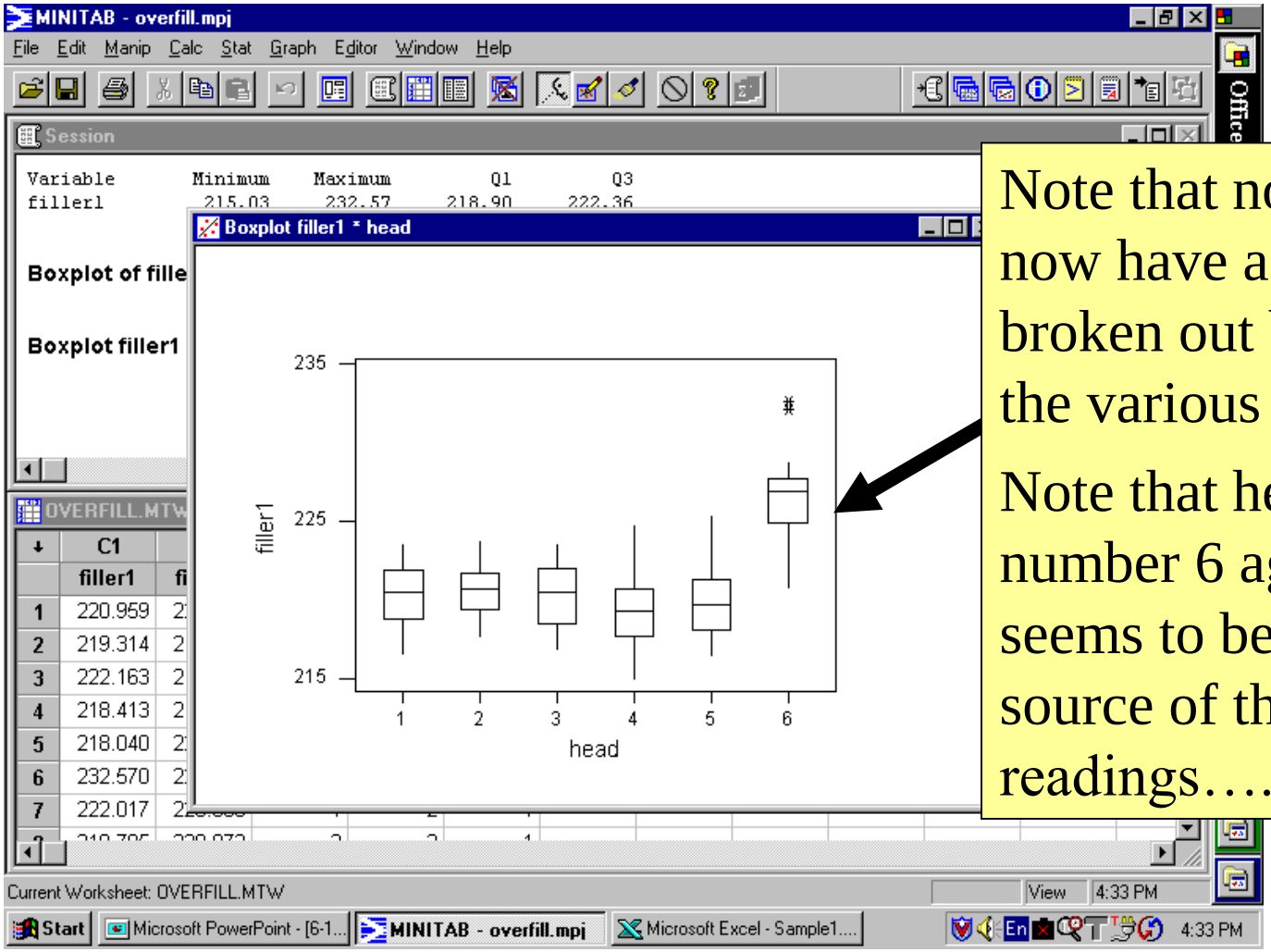
Boxplot....



Fill out the screen as follows:

- Select “filler 1” for our Y Variable….
- Select “head” for our X Variable….
- Select OK…..

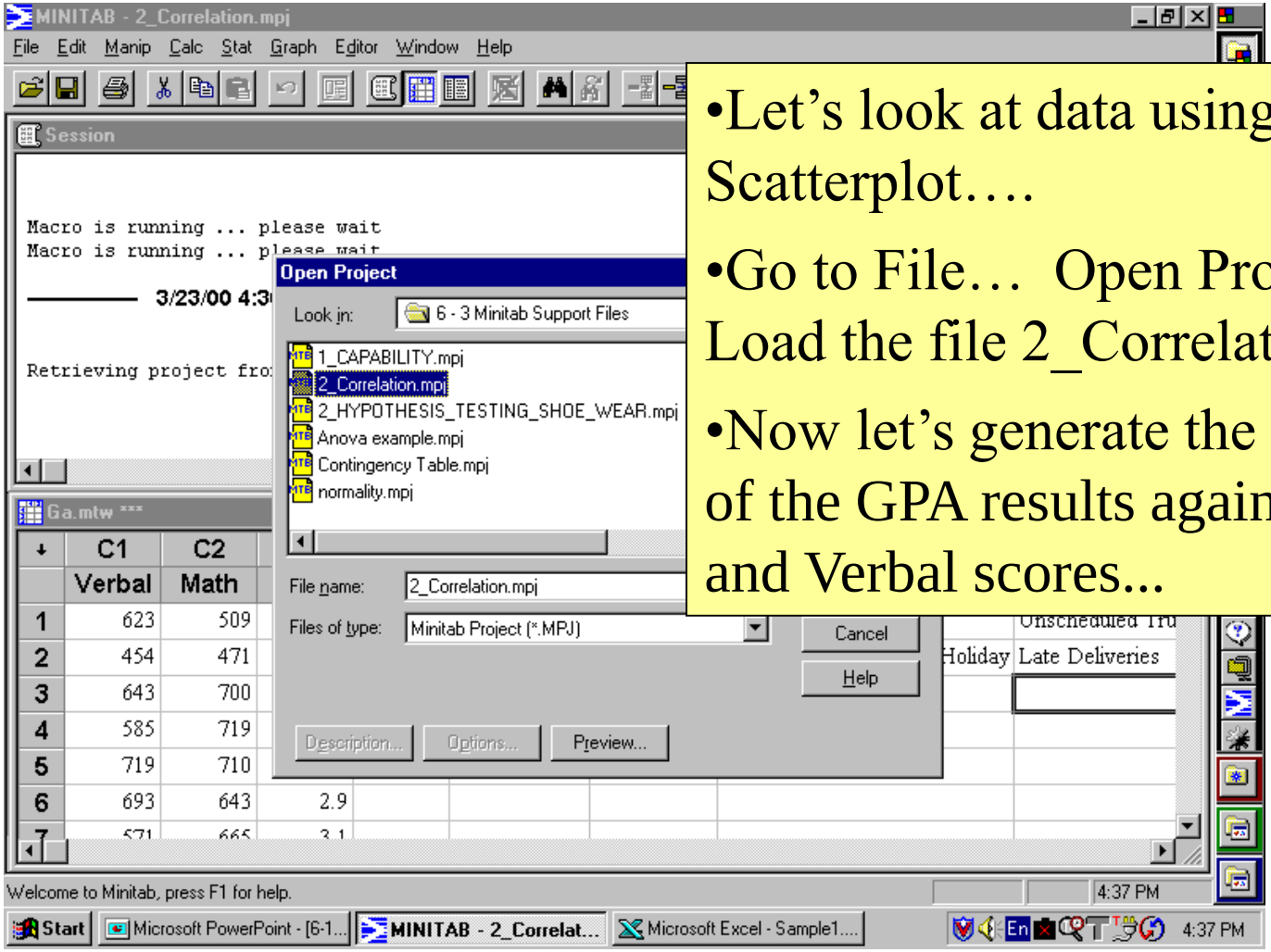
Boxplot....



Note that now we now have a box plot broken out by each of the various heads..

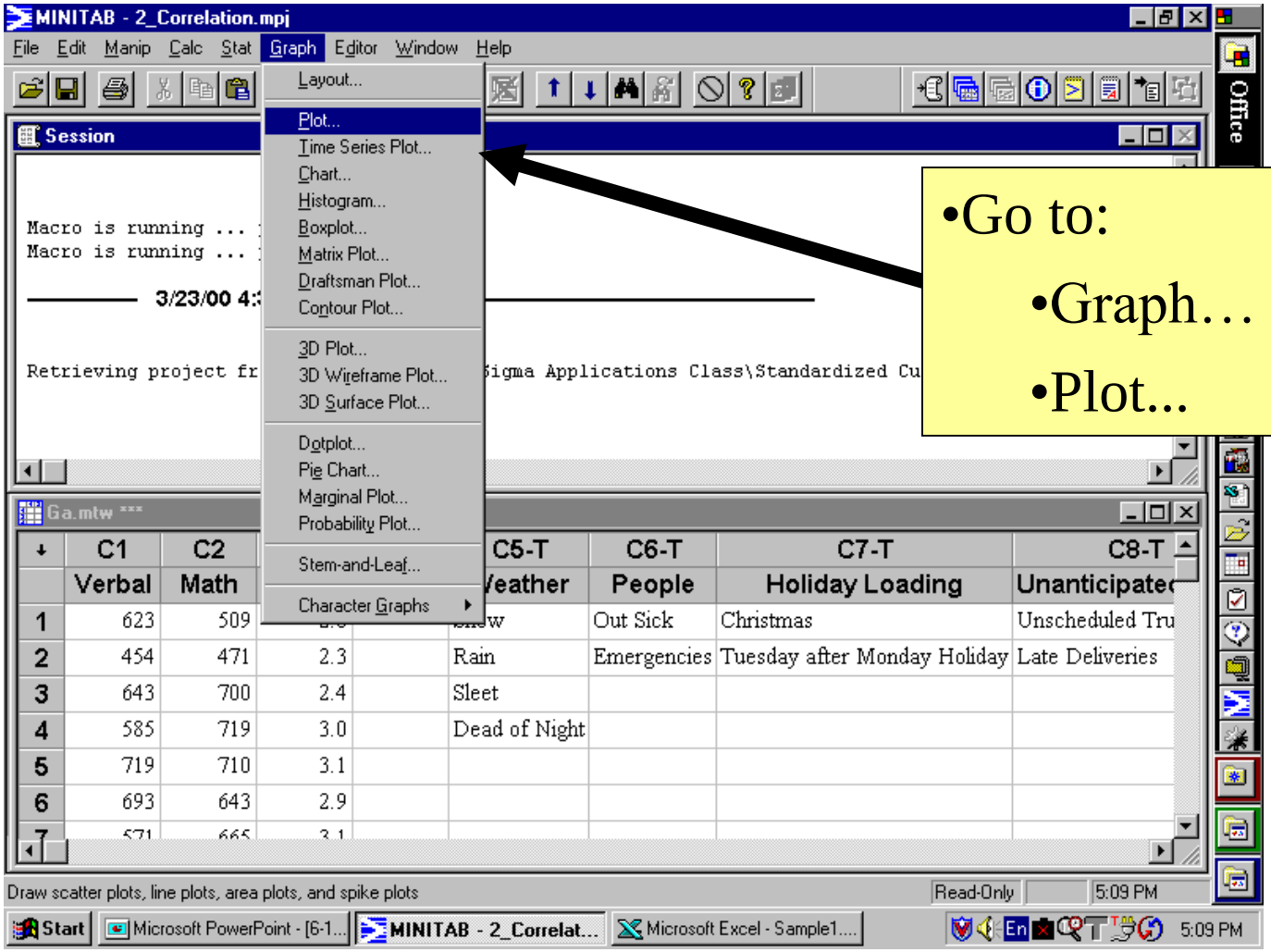
Note that head number 6 again seems to be the source of the high readings…..

Scatter plot....

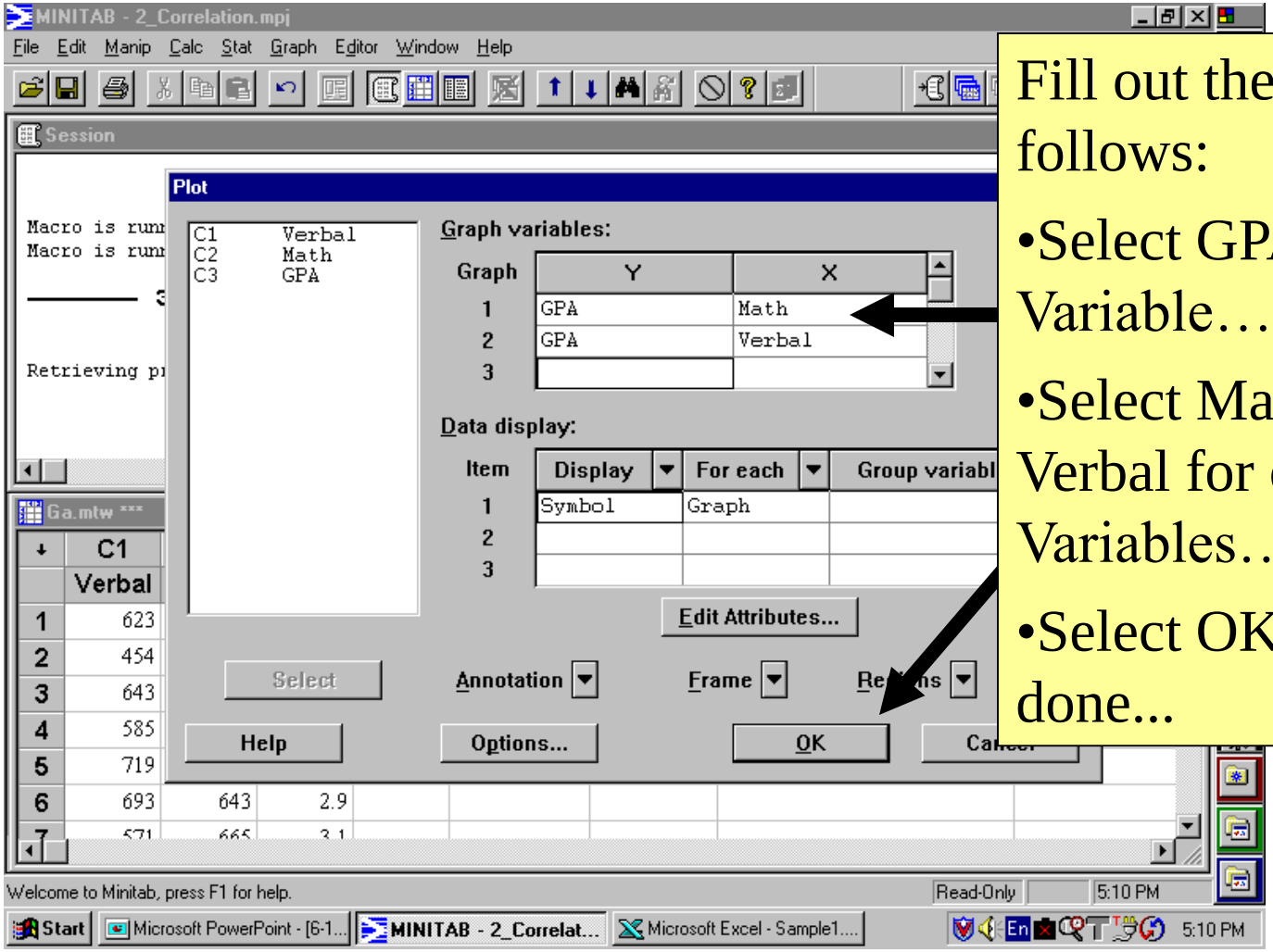


- Let's look at data using a Scatterplot….
- Go to File… Open Project…. Load the file 2_Correlation.mpj….
- Now let's generate the Scatterplot of the GPA results against our Math and Verbal scores...

Scatter plot....



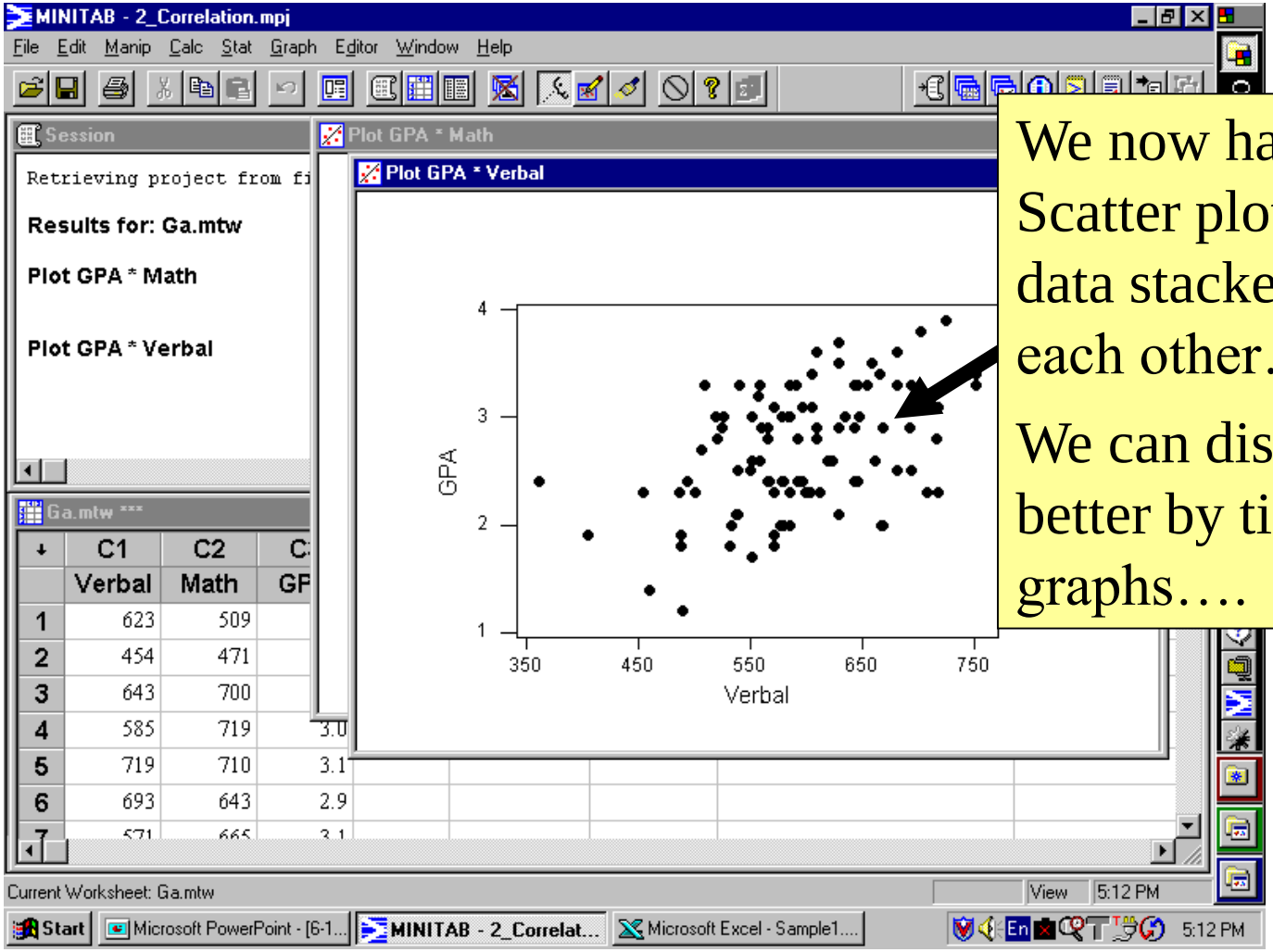
Scatter Plot....



Fill out the screen as follows:

- Select GPA for our Y Variable….
- Select Math and Verbal for our X Variables…..
- Select OK when done...

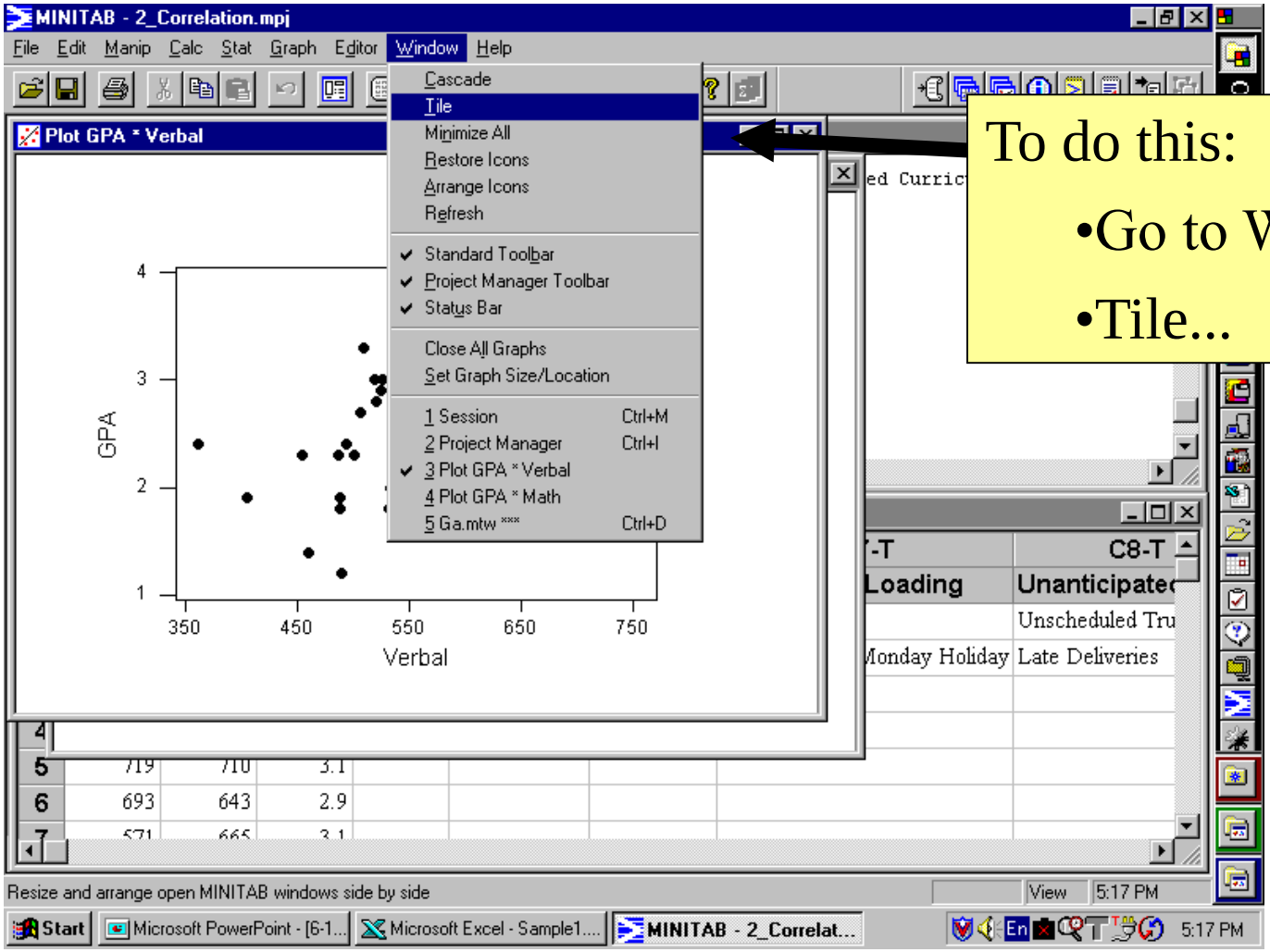
Scatter plot....



We now have two Scatter plots of the data stacked on top of each other…

We can display this better by tiling the graphs….

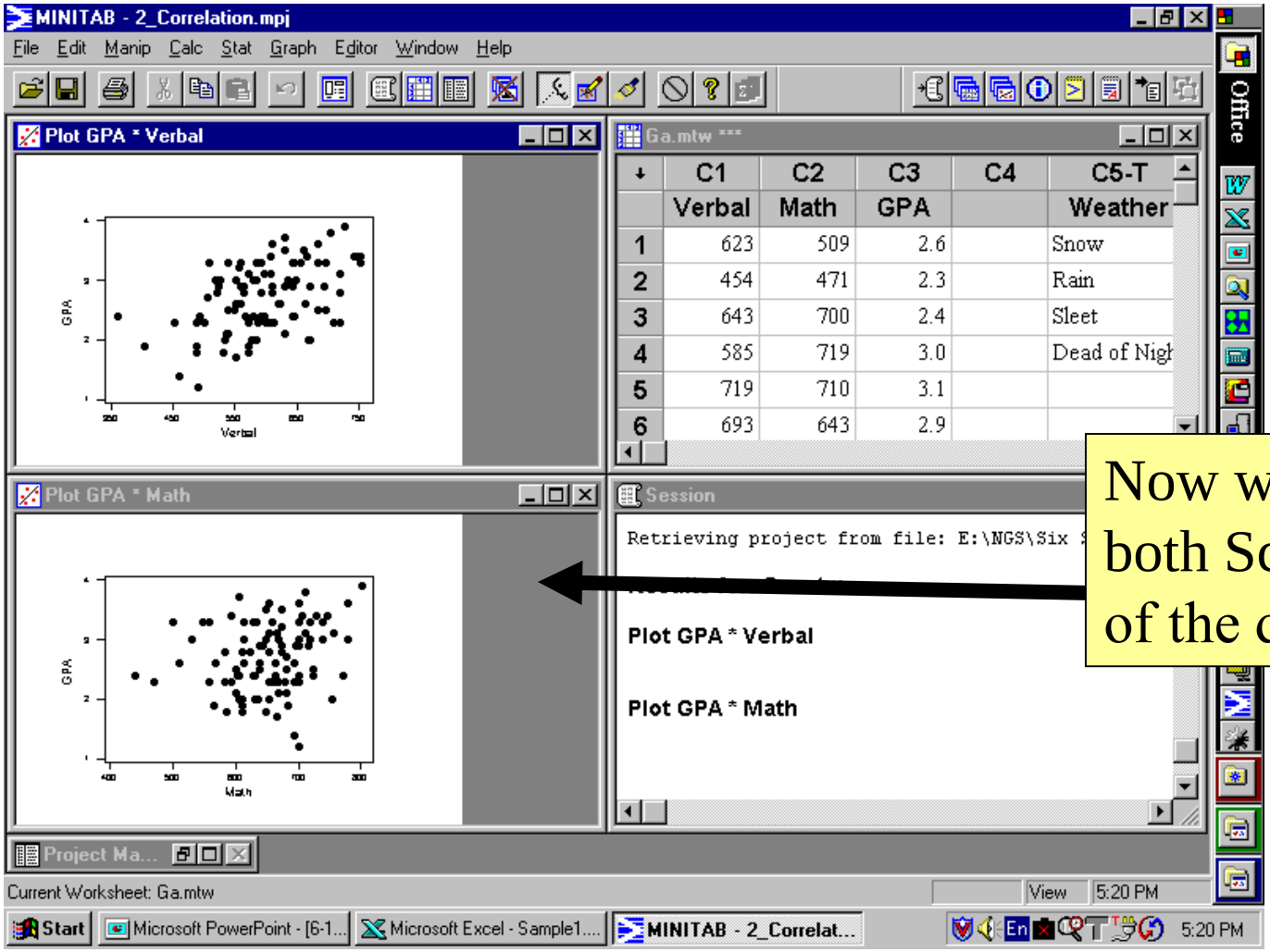
Scatter plot....



To do this:

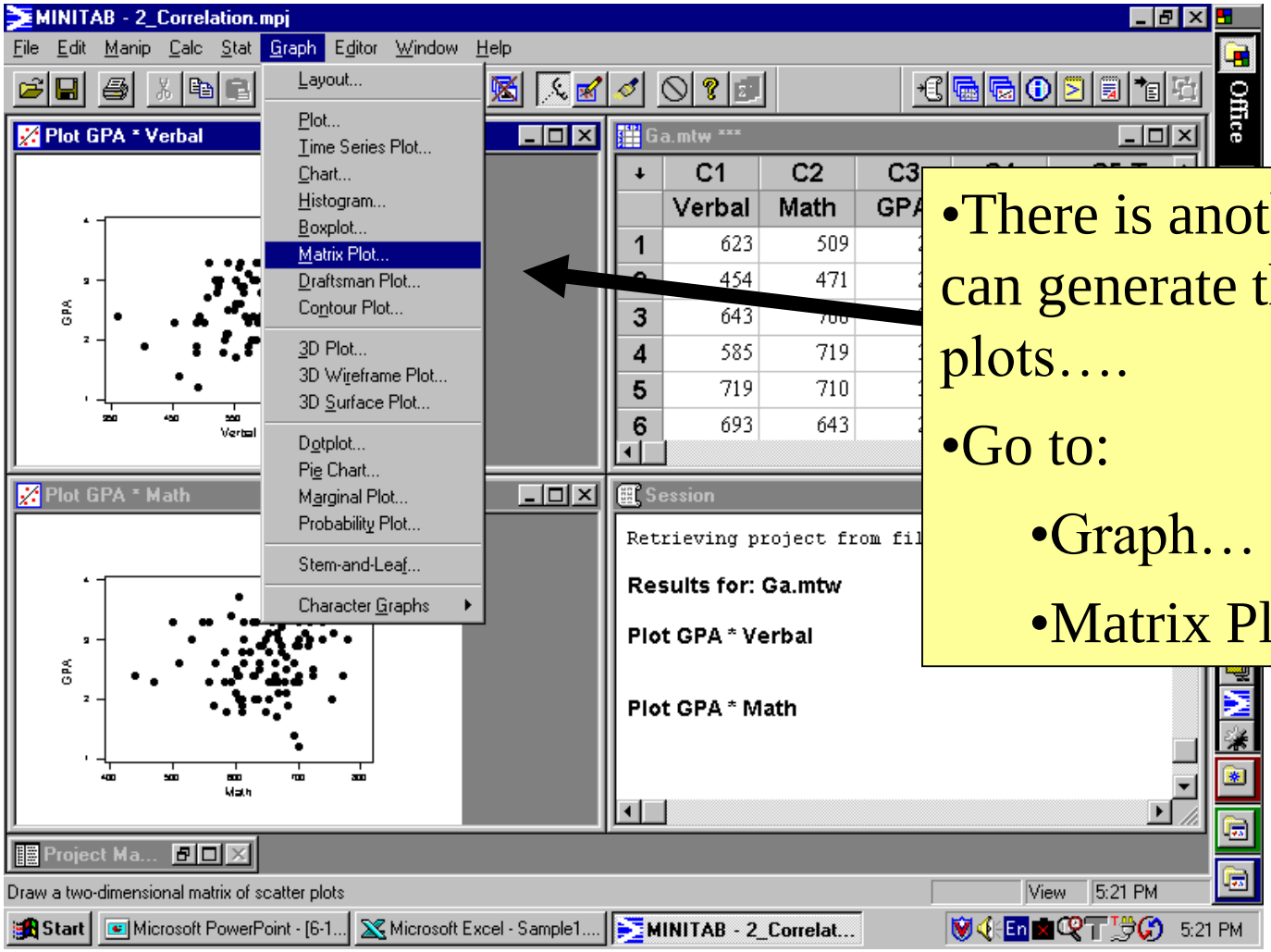
- Go to Window…
- Tile...

Scatter plot....



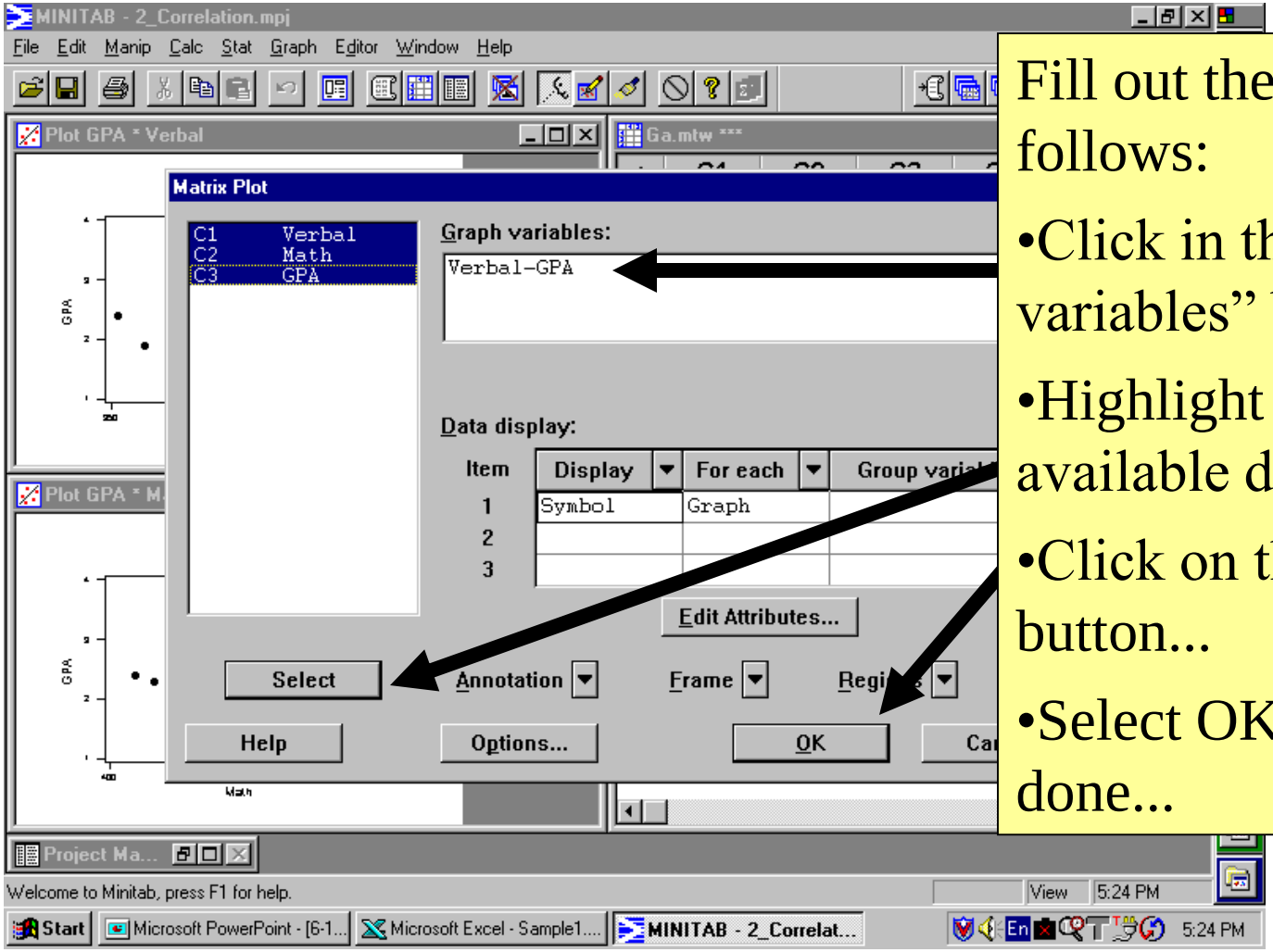
Now we can see both Scatter plots of the data…

Scatter plot....



- There is another way we can generate these scatter plots….
- Go to:
  - Graph…
  - Matrix Plot...

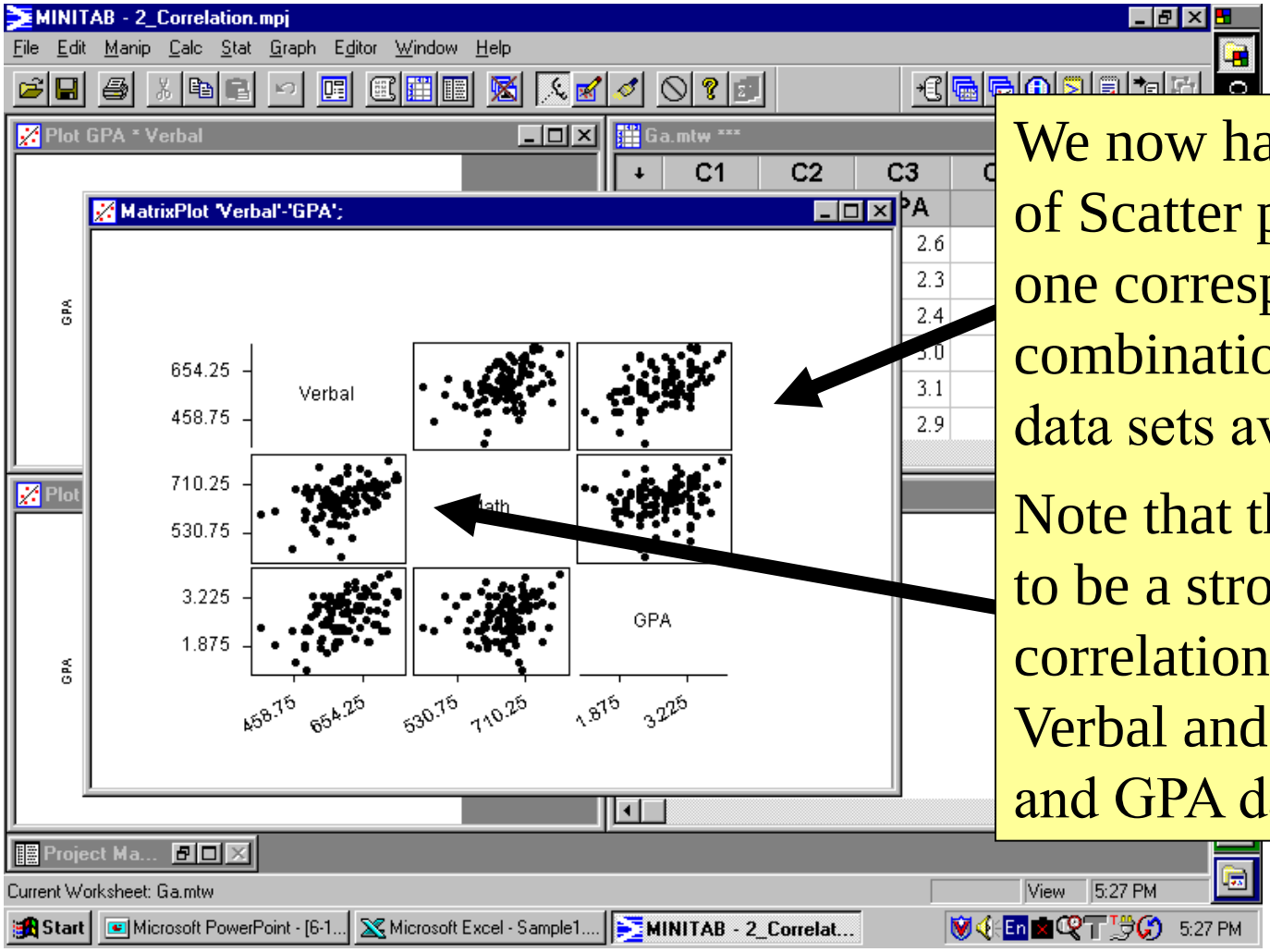
Scatter Plot....



Fill out the screen as follows:

- Click in the "Graph variables" block
- Highlight all three available data sets…
- Click on the "Select" button...
- Select OK when done...

Scatter plot....

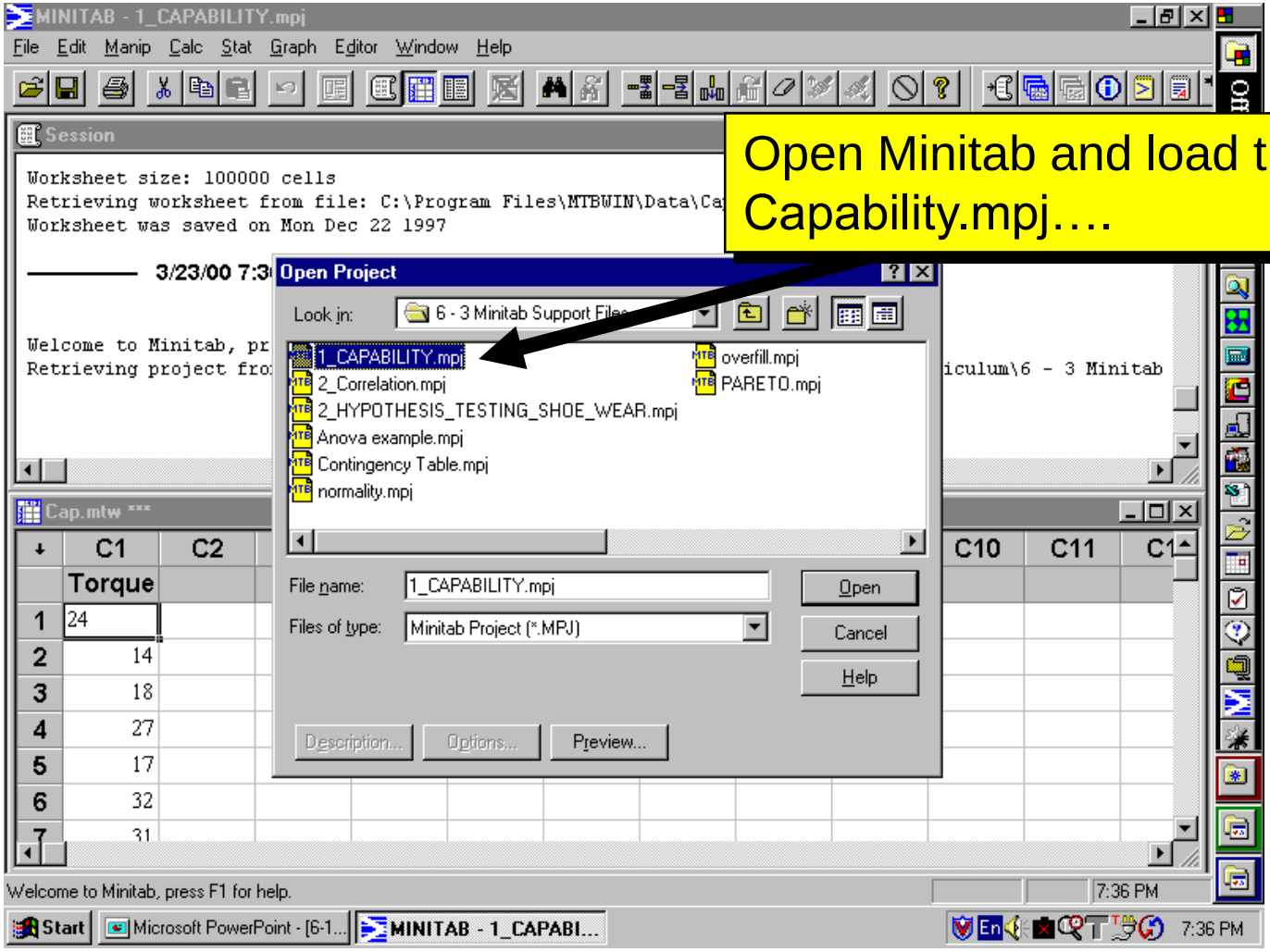


We now have a series of Scatter plots, each one corresponding to a combination of the data sets available…

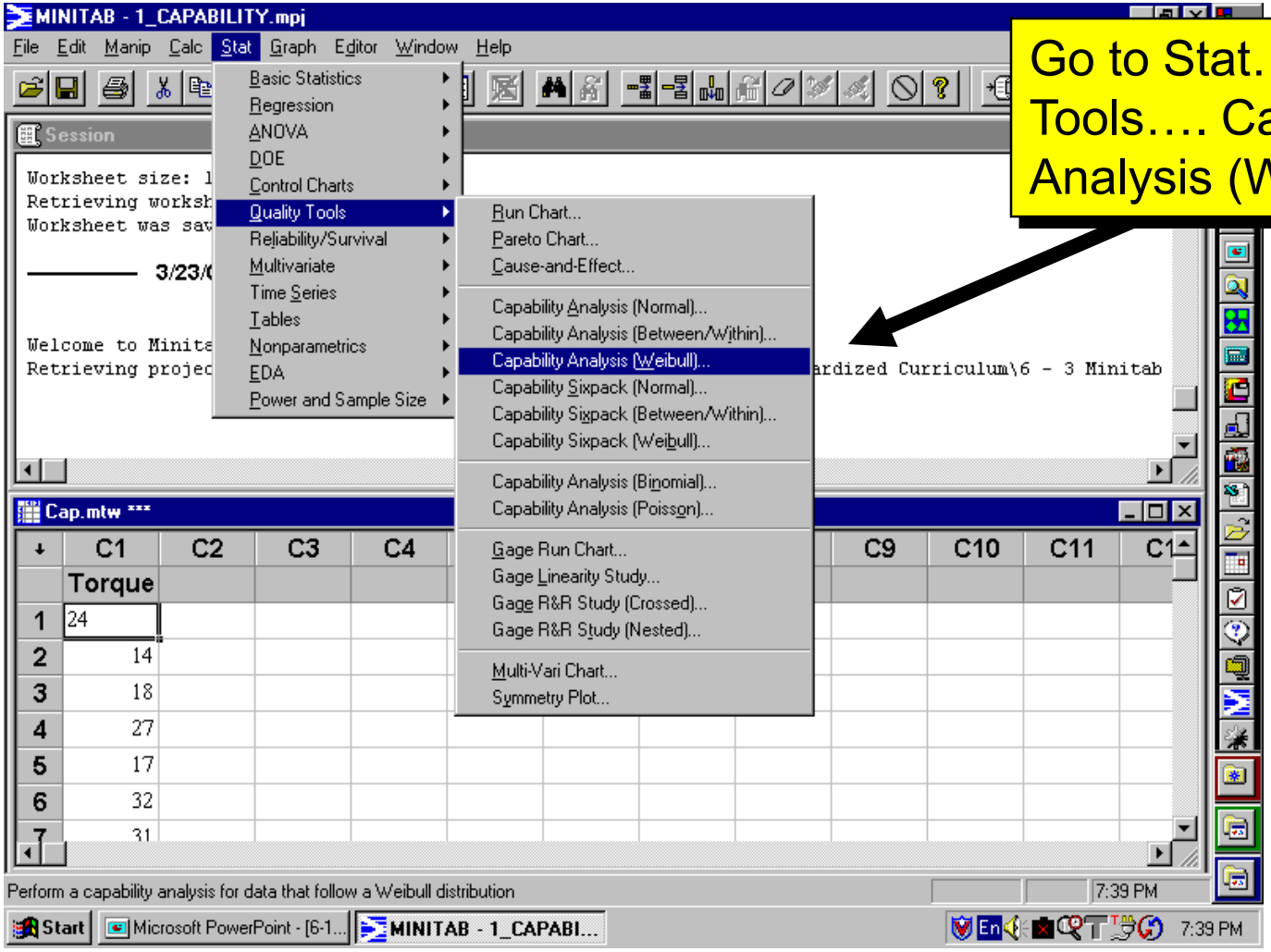
Note that there appears to be a strong correlation between Verbal and both Math and GPA data….

PROCESS CAPABILITY ANALYSIS

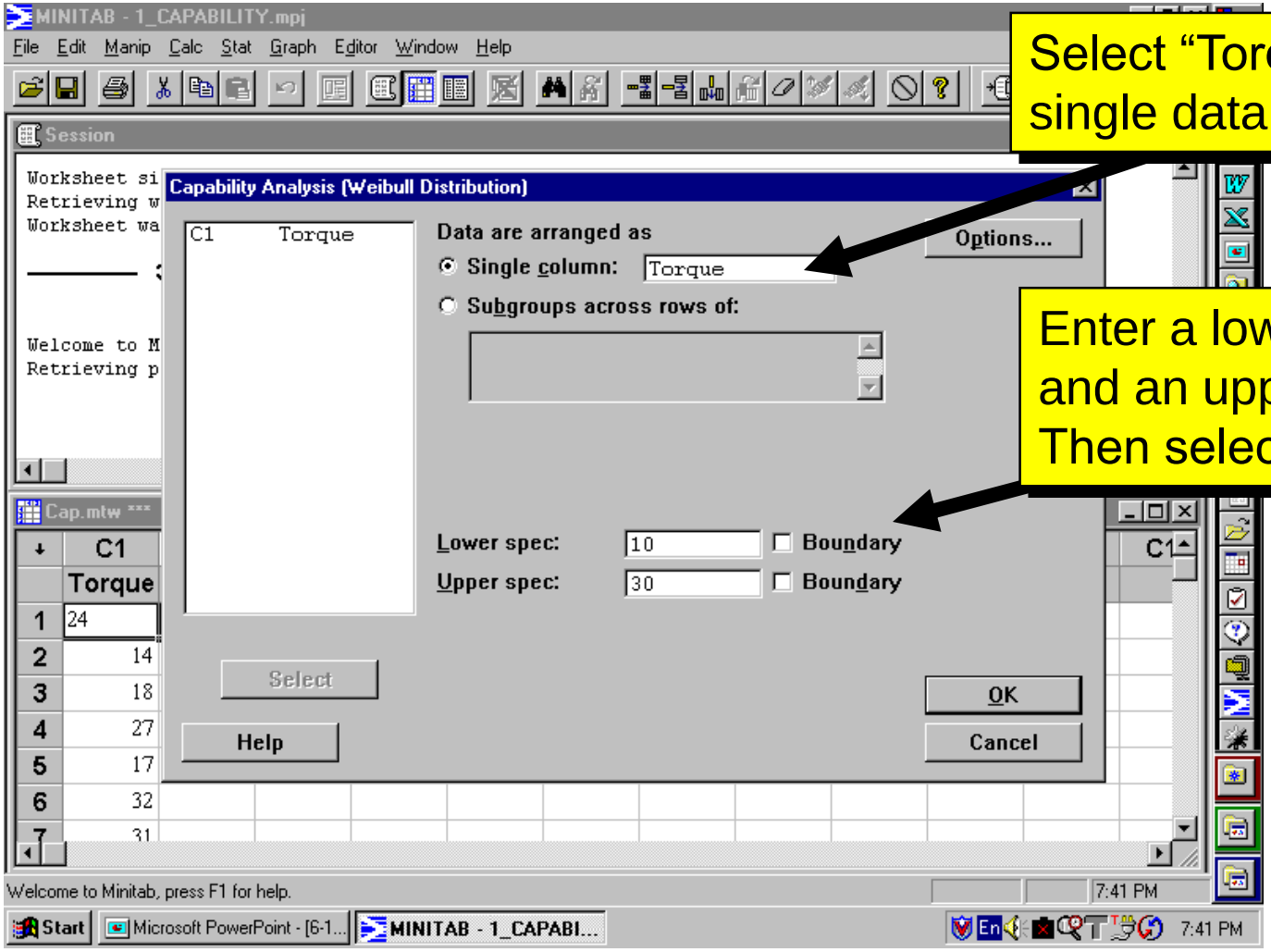
Let's do a process capability study….



SETTING UP THE TEST….



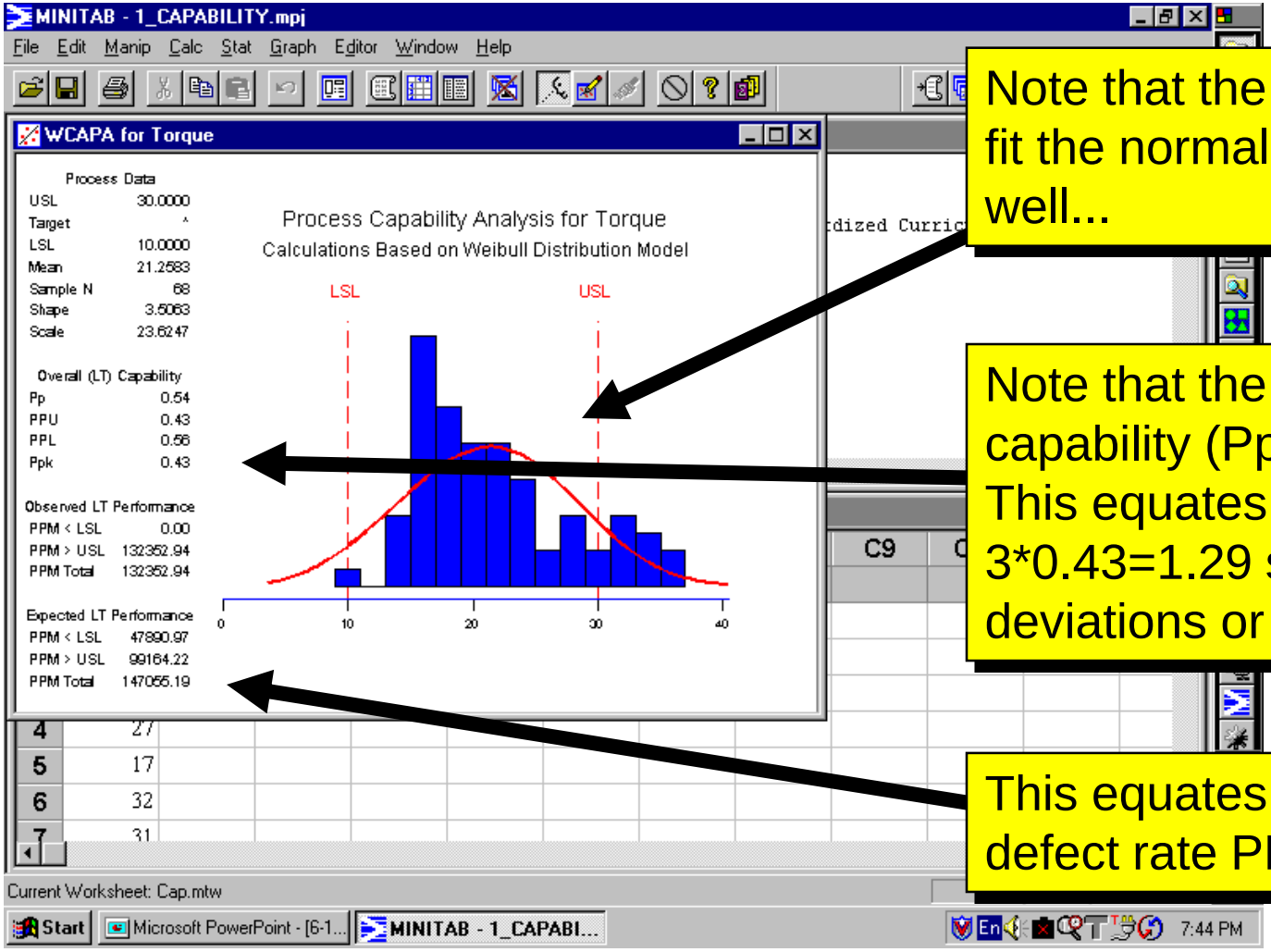
SETTING UP THE TEST….



Select “Torque” for our single data column...

Enter a lower spec of 10 and an upper spec of 30. Then select “OK”….

INTERPRETING THE DATA….



Note that the data does not fit the normal curve very well...

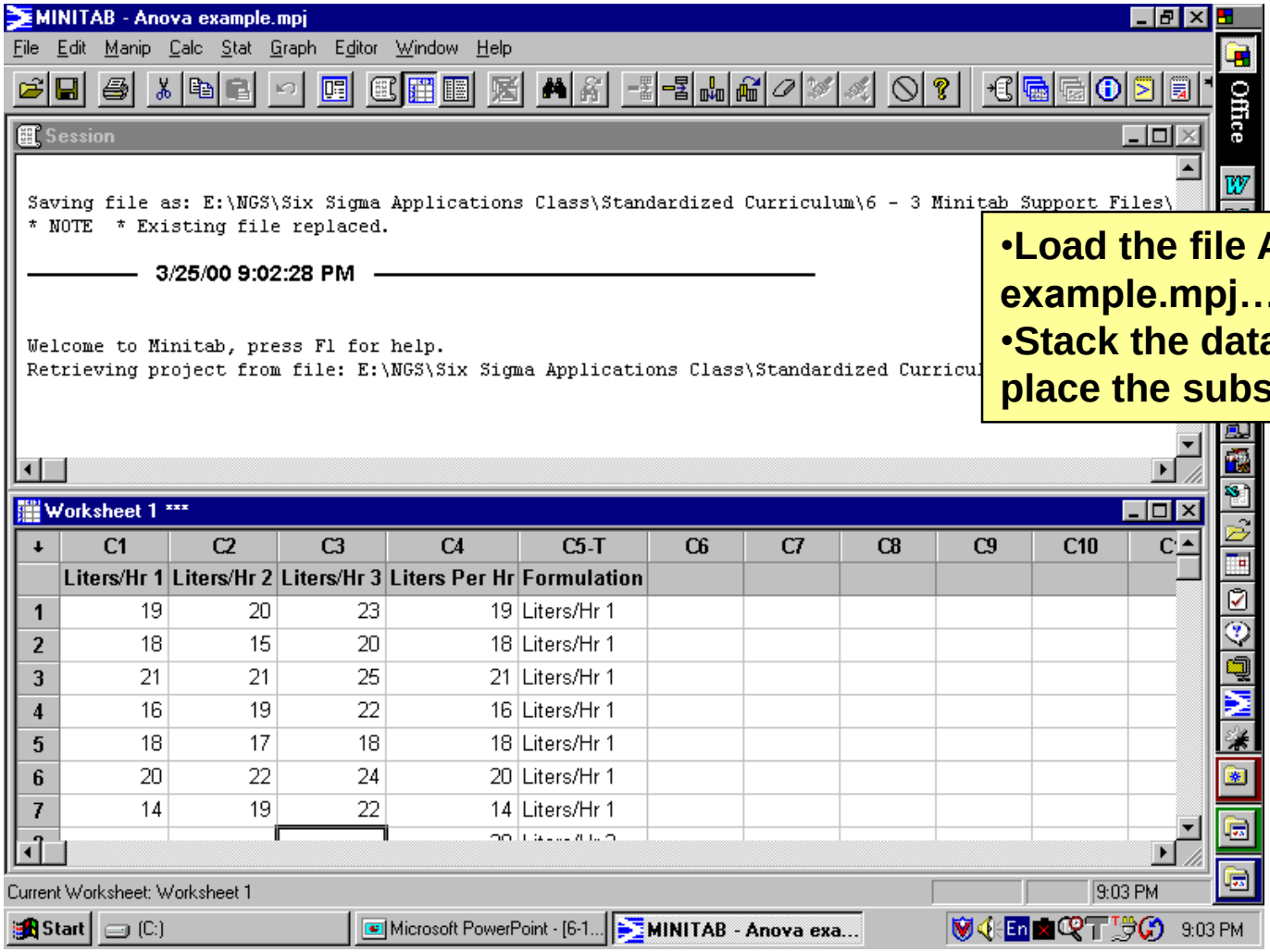
Note that the Long Term capability (Ppk) is 0.43. This equates to a Z value of 3*0.43=1.29 standard deviations or sigma values.

This equates to an expected defect rate PPM of 147,055.

ANalysis Of VAriance

ANOVA

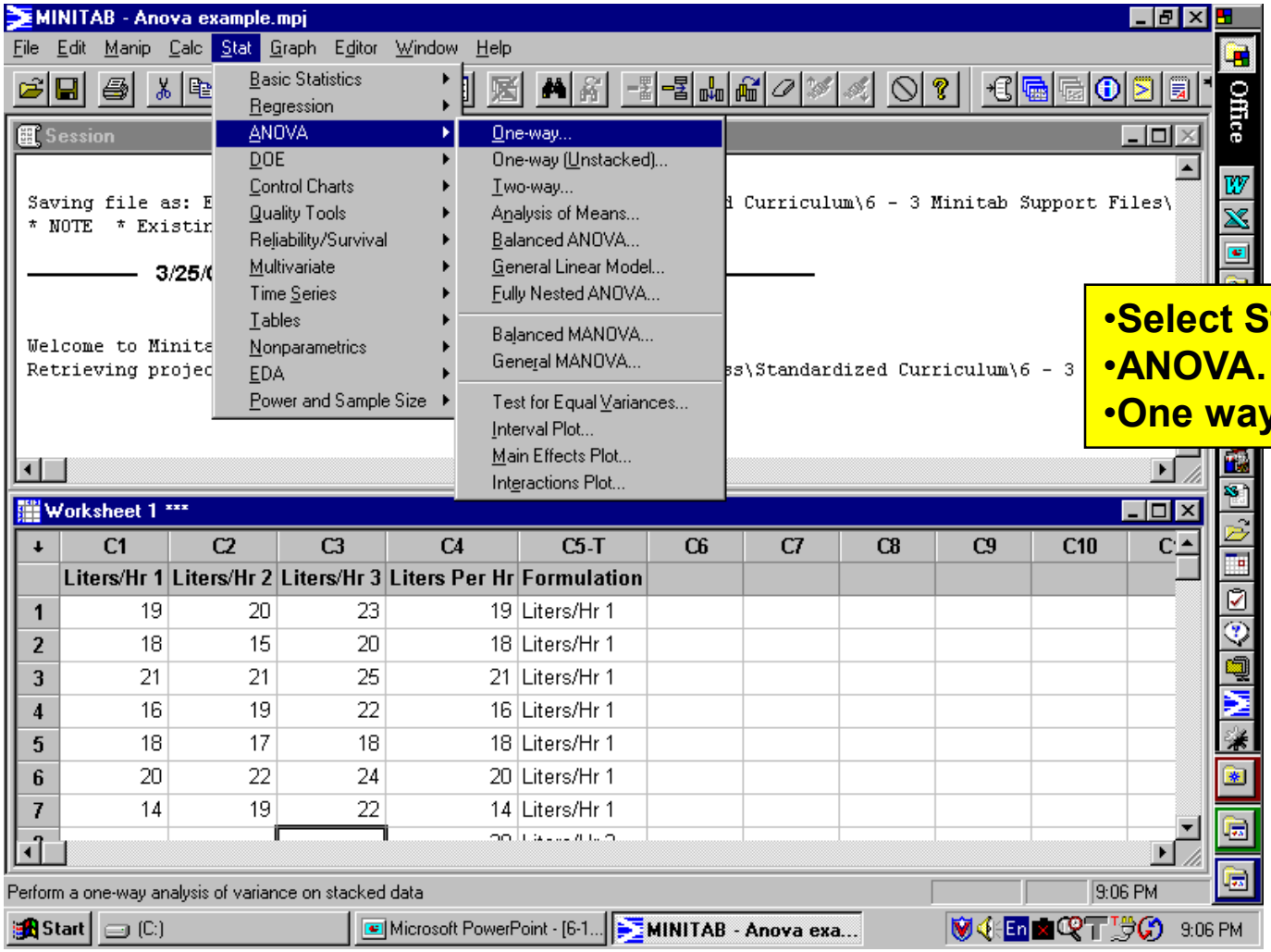
Let's set up the analysis



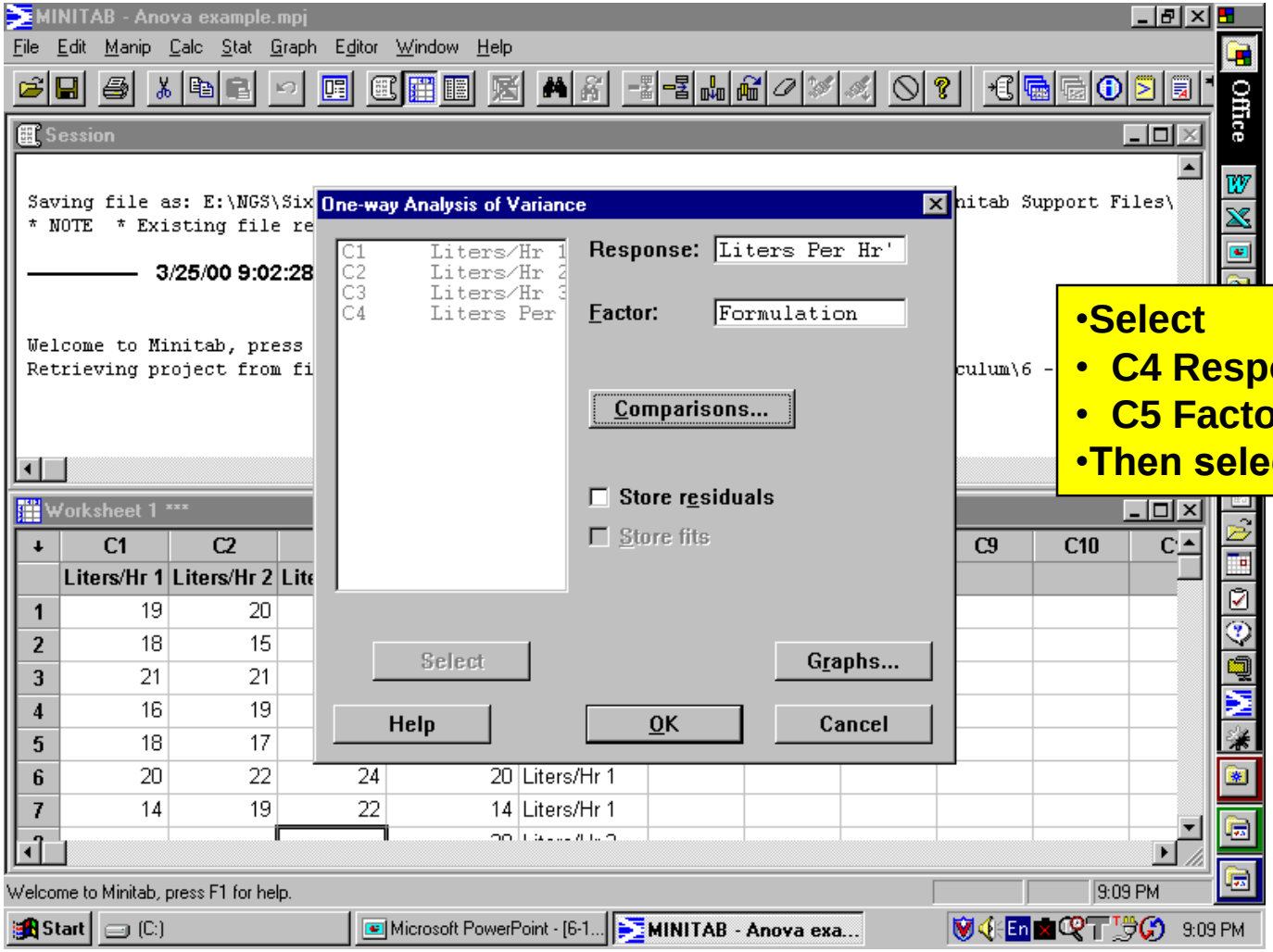
	C1	C2	C3	C4	C5-T	C6	C7	C8	C9	C10	C
	Liters/Hr 1	Liters/Hr 2	Liters/Hr 3	Liters Per Hr	Formulation						
1	19	20	23	19	Liters/Hr 1						
2	18	15	20	18	Liters/Hr 1						
3	21	21	25	21	Liters/Hr 1						
4	16	19	22	16	Liters/Hr 1						
5	18	17	18	18	Liters/Hr 1						
6	20	22	24	20	Liters/Hr 1						
7	14	19	22	14	Liters/Hr 1						

- **Load the file Anova example.mpj…**
- **Stack the data in C4 and place the subscripts in C5**

Set up the analysis….

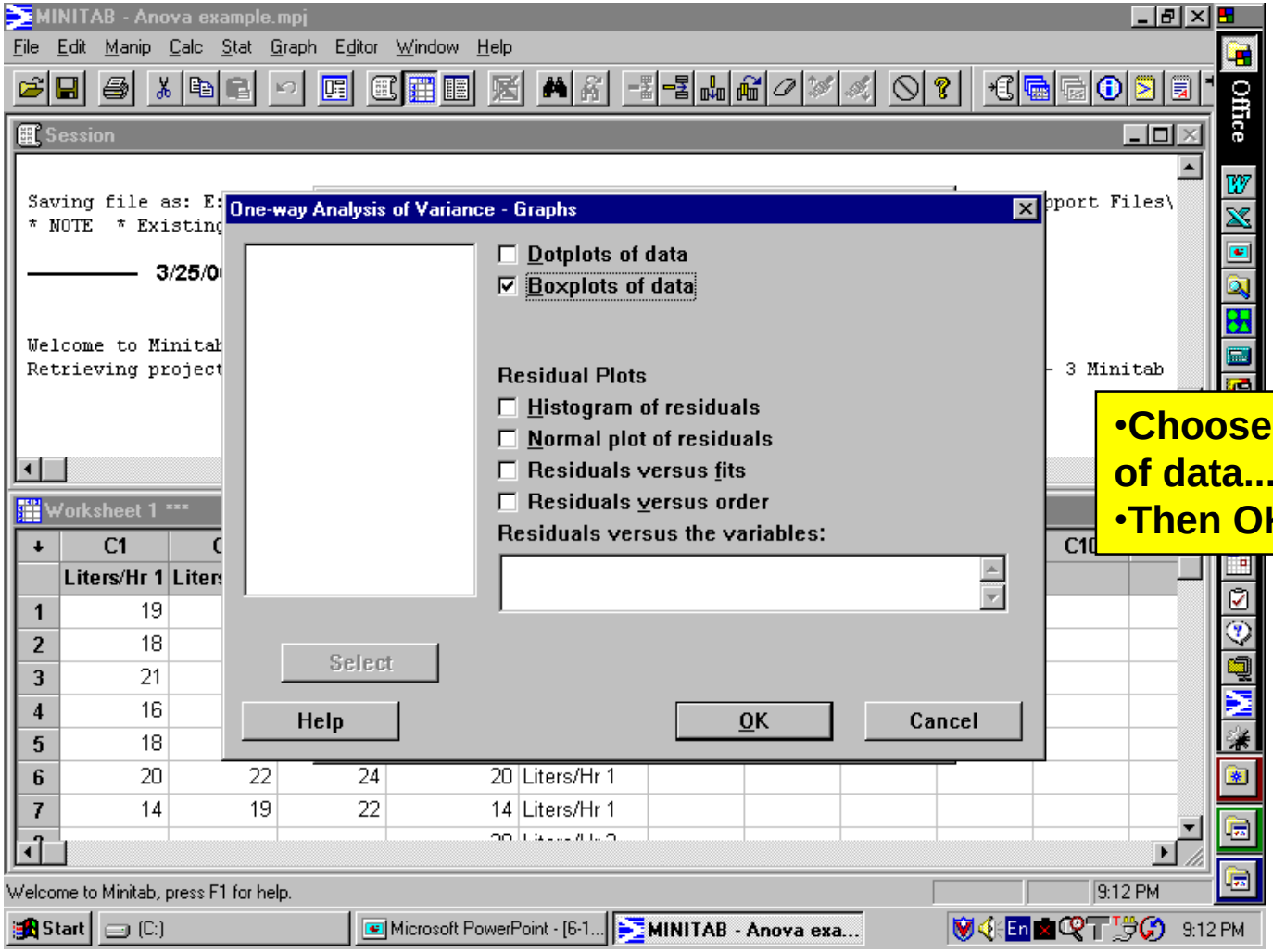


Set up the analysis….



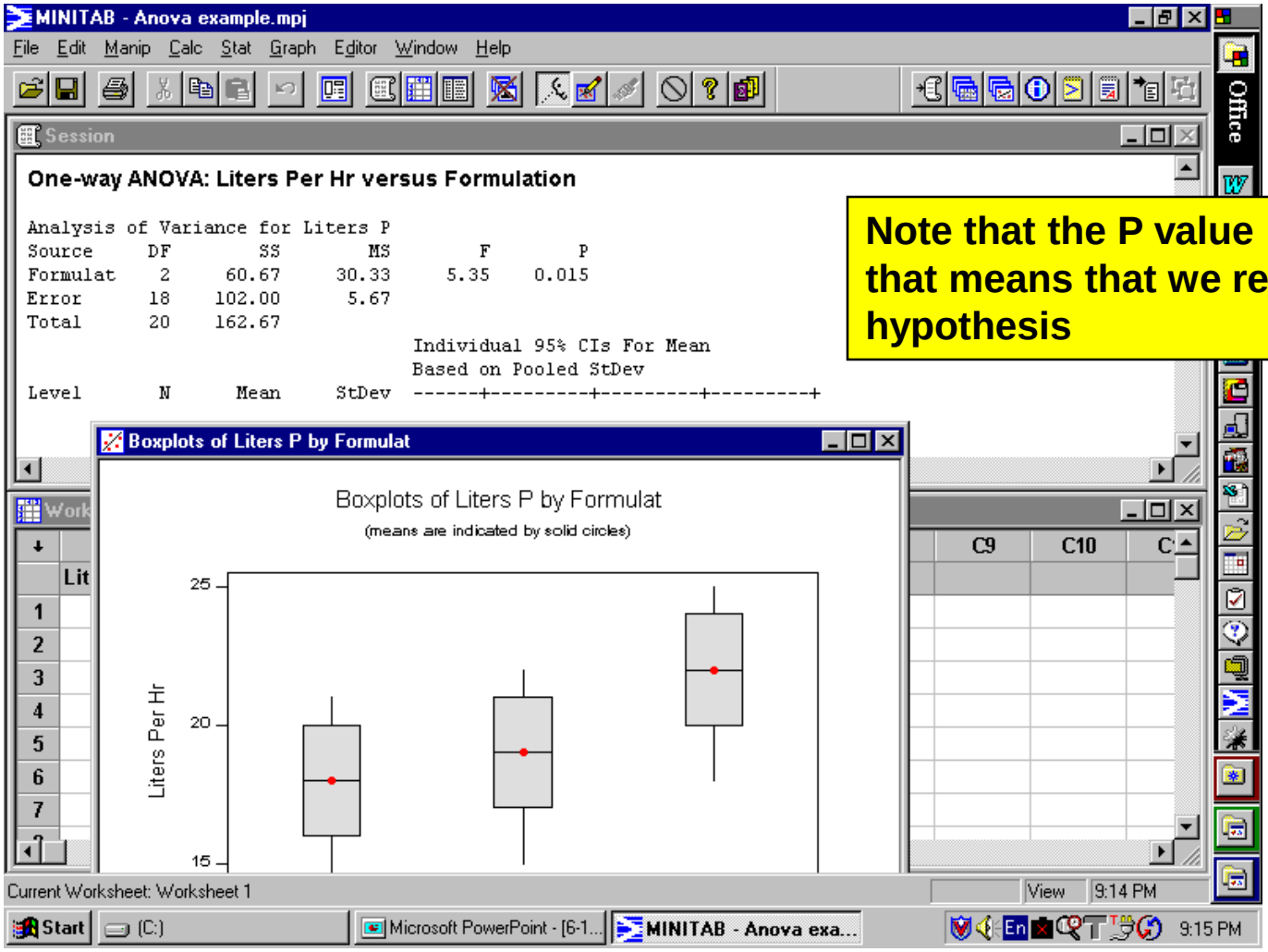
- **Select**
  - **C4 Responses**
  - **C5 Factors**
- **Then select Graphs….**

Set up the analysis….



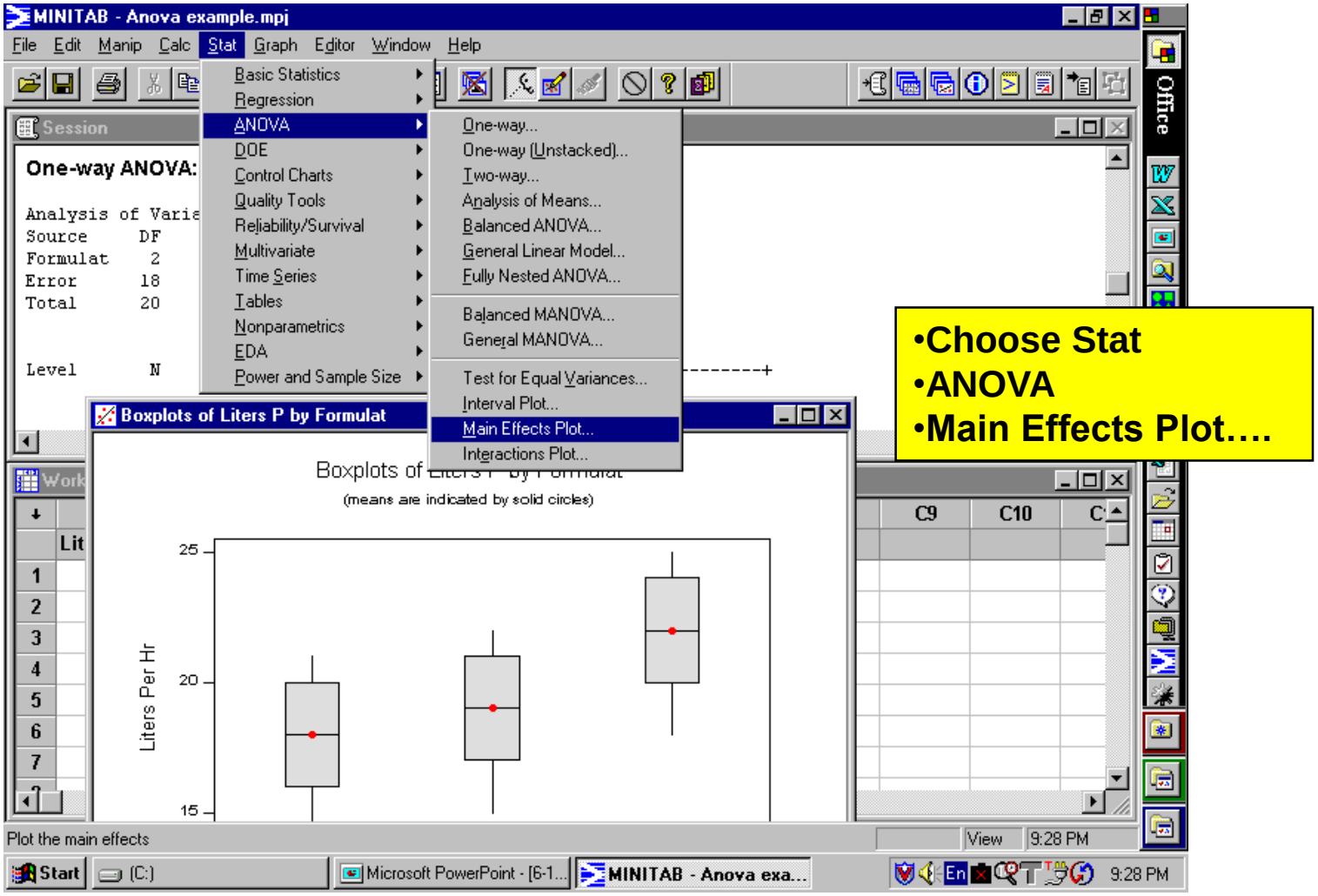
- Choose boxplots of data...
- Then OK

Analyzing the results….

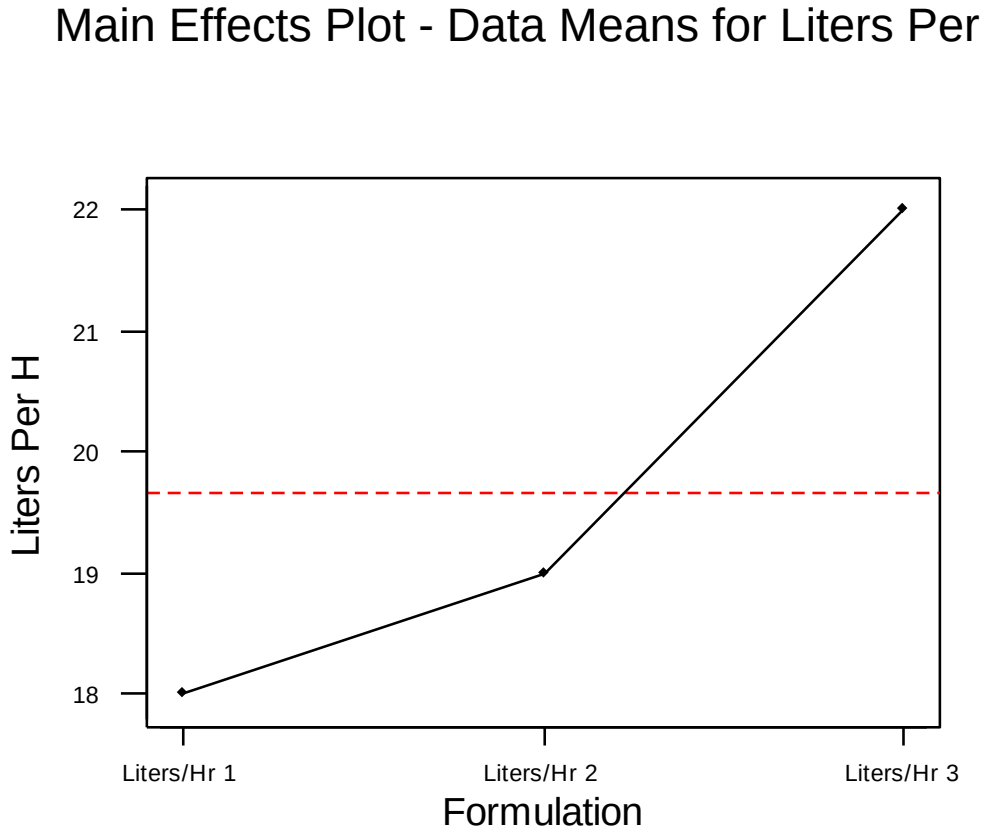


Note that the P value is less than .05 that means that we reject the null hypothesis

Let's Look At Main Effects….



Analyzing Main Effects..



Formulation 1 Has Lowest Fuel Consumption