Live Session 5

1. Welcome/Intro (including polls)
2. Confidence intervals & Sample size – continuous data (means)
3. Confidence intervals & Sample size – discrete data (proportions)
4. Project updates/questions
    - Review Storyboard
    - Sharing tools and applications
5. Assignments for next 2 weeks
6. Wrap up and Feedback

Analyze

Description:

Analyze, describe, and present the data to discover the root cause(s), identify/prioritize critical inputs (x's), determine the inputs impact on the output.

Key Concepts:
Inferential statistics, common distributions, developing a hypothesis, determining the likelihood some event happens based on a sample (calculating probabilities), Using the normal distribution as the "go to" distribution.

Project:
Write a null and alternative hypothesis statement.

Tools:
Hypothesis testing
Chi-square test for independence

Week 3 & 4

Key Concepts:
Collecting sample data, how confidence intervals and sample size are related.

Project:
Utilize the sample size formula.

Tools:
Confidence intervals.

Week 5

Key Concepts:
Determining input's (x) impact on the output (y).

Project:
Use regression to identify relationships between the output (y) and inputs (x's).

Tools:
Correlation
Simple linear regression
Multiple regression
Scatterplot
Trend/ line chart
Pareto chart
Fishbone (cause/effect) diagram

Week 6 &7

Confidence Intervals & Sample Size: Continuous data

Calculate a Point Estimate

Characteristics of the sample are called **statistics,** while characteristics of the population are called **parameters. Statistical inference** consists of methods for estimating and drawing conclusions about parameters, based on the corresponding statistic.

Point estimation is the process of estimating unknown population parameters by known statistics. The value of each sample statistic used as an estimate is called a **point estimate.**

Since a sample is only a small subset of a population, generalizing from a sample to the population carries the risk that the point estimate may not be very accurate. Confidence intervals help provide a means to construct an interval, based on the statistic, that is likely to contain the parameter.

Confidence Interval for the Population Mean

Although we cannot measure how confident we are of a statistic as a point estimate of a parameter, we can use the statistic to find an interval that is likely to contain the parameter.

A **confidence interval** is an estimate of a parameter consisting of an interval of numbers based on a point estimate, together with a **confidence level** specifying the probability that the interval contains the parameter.

Confidence intervals are often reported in the format:

(lower bound, upper bound)

Or

Parameter +/- Margin of Error

Confidence Interval for the Population Mean : Z

Confidence Interval for the Population Mean μ : Z

The Confidence Interval for μ may be constructed only when either of the following two conditions are met:

- The population is normally distributed and σ is known.
- The sample size is large ($n \geq 30$), and the value of σ is known.

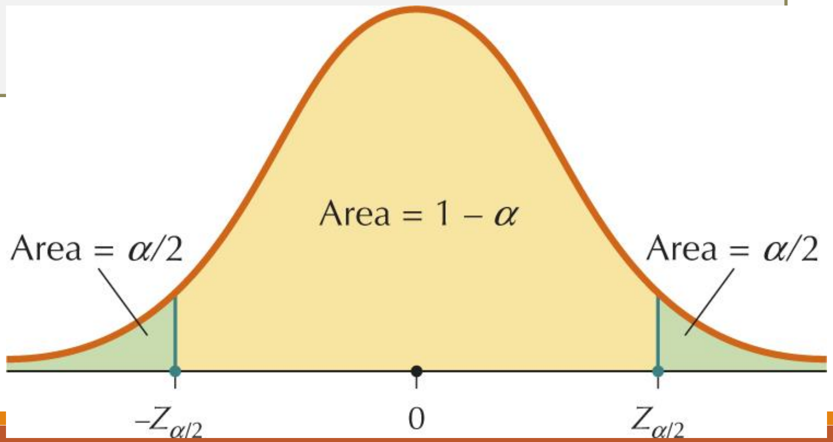
When a random sample of size n is taken from a population, a (100 – α)% confidence interval is given by:

$$\text{lower bound} = \bar{x} - Z_{\alpha/2}(\sigma/\sqrt{n})$$

$$\text{upper bound} = \bar{x} + Z_{\alpha/2}(\sigma/\sqrt{n})$$

The Z interval can also be written as: $\bar{x} \pm Z_{\alpha/2}(\sigma/\sqrt{n})$

α	Confidence Level	$\alpha/2$	$Z_{\alpha/2}$
0.10	90%	0.05	1.645
0.05	95%	0.025	1.96
0.01	99%	0.005	2.576



Confidence Interval for the Population Mean : t

Confidence Interval for the Population Mean μ : t

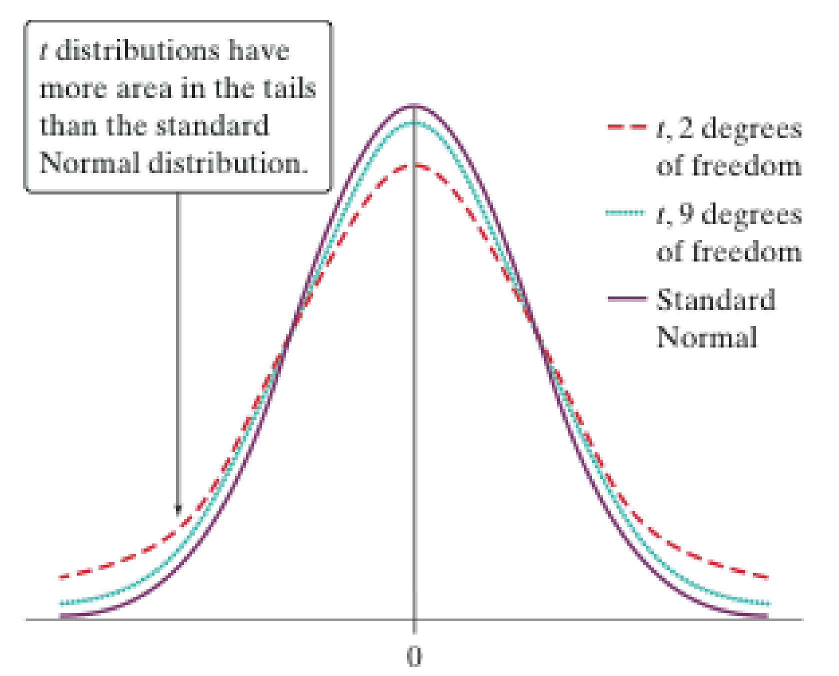
The t interval for μ may be constructed when the following two conditions are met:

- The population is normally distributed.
- The sample size is small (n < 30).

When a random sample of size n is taken from a population, a $100(1 - \alpha)\%$ confidence interval is given by:

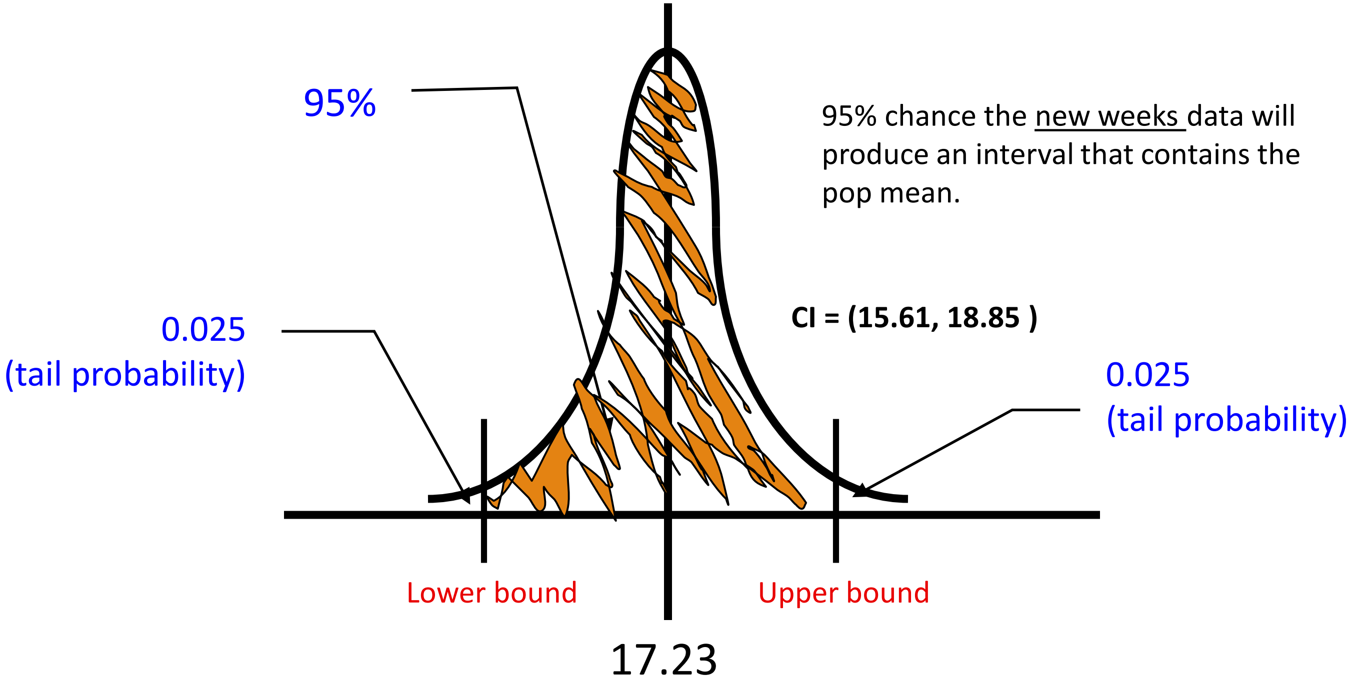
$$\text{lower bound } = \bar{x} - t_{\alpha/2}(s/\sqrt{n})$$
$$\text{upper bound } = \bar{x} + t_{\alpha/2}(s/\sqrt{n})$$

The t interval may also be written as: $\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$



$t_{\alpha/2}$ values can be found in Table D
df = degrees of freedom = n-1

Confidence Interval – Hank's process



- The above distribution is a distribution of sample means, so the probability of the true population mean being within these bounds is 95%.
- In 95% of all xbars (calculated from different samples of this process each having a different xbar) the interval (the shaded region) will include the population mean.
- **We are 95% confident that the population mean lies between this upper and lower bound.**
- Confidence intervals are NOT used to determine an interval at which any one observation of the distribution lies.

Margin of Error

Confidence intervals for the population proportion *p* take the form:

point estimate ± margin of error *E*

“We can estimate μ to within E units with $(1-\alpha)100\%$ confidence.”

The **margin of error** $\boldsymbol{E}$ is a measure of the *precision* of the confidence interval estimate.

For the $\boldsymbol{Z}$ interval, the margin of error takes the form

$$E = Z_{\alpha/2}\,(\sigma/\sqrt{n})$$ =CONFIDENCE.NORM(alpha,standard_dev,size)

For the $\boldsymbol{t}$ interval, the margin of error takes the form

$$E = t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$ =CONFIDENCE.T(alpha,standard_dev,size)

Smaller values of E indicate smaller margin of error, and therefore, greater precision.

The confidence interval is often written as: $\bar{x} \pm E$

Example

The College Board reports that the scores on the 2010 SAT mathematics test were normally distributed. A sample of 25 scores had a mean of 510. Assume the population standard deviation is 100. Construct a 90% confidence interval for the population mean score on the 2010 SAT math test.

$$\text{lower bound} = \bar{x} - Z_{\alpha/2}(\sigma/\sqrt{n})$$
$$= 510 - 1.645(100/\sqrt{25})$$
$$= 477.1$$

$$\text{upper bound} = \bar{x} + Z_{\alpha/2}(\sigma/\sqrt{n})$$
$$= 510 + 1.645(100/\sqrt{25})$$
$$= 542.9$$

We are 90% confident that the population mean SAT score on the 2010 mathematics SAT test lies between 477.1 and 542.9.

Breakout questions:

What is the margin of error?

Can I say that I am 90% confident that all of the math SAT scores for 2010 lie between 477.1 and 542.9? Why or why not?

Sample Size for Estimating μ

A natural question when constructing a confidence interval is "*How large a sample size do I need* to get a tight confidence interval with a high confidence level?"

Sample Size for Estimating the Population Mean
The sample size for a Z interval that estimates μ to within a margin of error E with confidence $(100 - \alpha)\%$ is given by:

$$n = \left(\frac{(Z_{\alpha/2})\sigma}{E} \right)^2$$

Whenever this formula yields a sample size with a decimal, *always round up to the next whole number*.

<u>How are you determining the sample size for your project?</u>

Confidence Intervals & Sample Size: Discrete data

Calculate a Point Estimate

Recall that characteristics of the sample are called **statistics,** while characteristics of the population are called **parameters.** We have dealt with interval estimates of μ, but we may also be interested in the interval estimate for the population proportion of successes, p.

$$\hat{p} = \frac{x}{n} = \frac{\text{number of successes}}{\text{sample size}}$$

is a point estimate of the population proportion p.

Z Interval for p

Confidence Interval for the Population Proportion *p*

The Confidence Interval for *p* may be constructed only when both of the following two conditions are met: *n(p-hat)* ≥ *5* and *n*(1 – *p-hat)* ≥ *5*.

When a random sample of size *n* is taken from a population, a (100 – *α*)% confidence interval is given by:

$$\text{lower bound} = \hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{upper bound} = \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The *Z* interval can also be written as: $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

α	Confidence Level	$\alpha/2$	$Z_{\alpha/2}$
0.10	90%	0.05	1.645
0.05	95%	0.025	1.96
0.01	99%	0.005	2.576

Margin of Error

Confidence intervals for the population proportion *p* take the form:

point estimate ± margin of error *E*

The **margin of error *E*** is a measure of the precision of the confidence interval estimate. For the *Z* interval, the margin of error takes the form:

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

The confidence interval is often written as:

$$\hat{p} \pm E$$

Breakout - Example

There is hardly a day that goes by without some new poll coming out. Especially during election campaigns, polls influence the choice of candidates and the direction of their policies. In October 2004, the Gallup organization polled 1012 American adults, asking them, “Do you think there should or should not be a law that would ban the possession of handguns, except by the police and other authorized persons?” Of the 1012 randomly chosen respondents, 638 said that there should NOT be such a law.

a. Check that the conditions for the Z interval for p have been met.

b. Find and interpret the margin of error E.

c. Construct and interpret a 95% confidence interval for the population proportion of all American adults who think there should not be such a law.

Sample Size for Estimating *p*

A natural question when constructing a confidence interval is "*How large a sample size do I need* to get a tight confidence interval with a high confidence level?"

Sample Size for Estimating the Population Proportion

The sample size for a Z interval that estimates p to within a margin of error E with confidence $100(1 - \alpha)\%$ is given by:

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

Whenever this formula yields a sample size with a decimal, *always round up to the next whole number*.

When *p-hat* is unknown, use $n = \left(\frac{0.5 \cdot Z_{\alpha/2}}{E}\right)^2$

Project Updates and Questions

Project Deliverables

Problem Definition Worksheet

- COMPLETE

Process Improvement Project (includes Storyboard)

- due 4 days after Live Session 10

Name of your project

Process owner: or your Name

Key Dates --->	Team Launch	Define	Measure	Analyze	Improve	Control

DEFINE

MEASURE

ANALYZE

IMPROVE

STORYBOARD TEMPLATE

CONTROL

TEAM MEMBERS

Process Improvement Project – Cycle Time Reduction

Process owner: Dan

Key Dates --->	Team Launch 8/23	Define 9/08	Measure 10/16	Analyze 10/24	Improve 10/31	Control On-Going

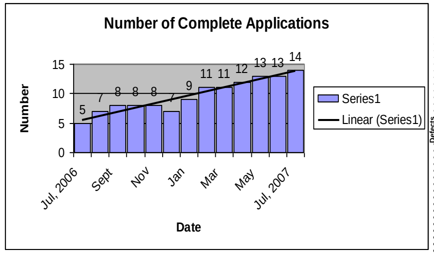
DEFINE

It takes 43 days to process a grant application. Only 8% of applications are being processed within 30 days of receipt. The time to process the application has lead to unhappy applicants and staff who are finding more and more of their daily work time being devoted to "grant administration." The funding levels available to applicants and the number of applications are expected to increase in the near future, which has the potential to compound the problem.

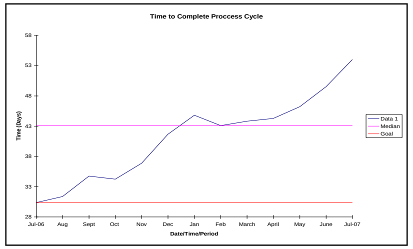
Defects/delays are inherent in the current process. Current SQL is 1.9

MEASURE

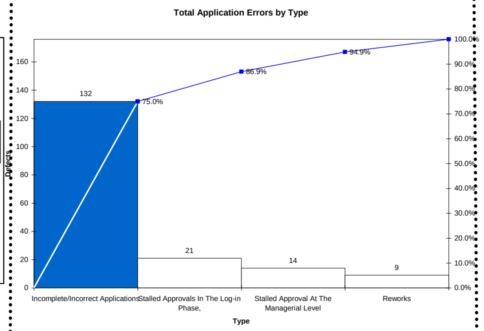
The Number of applications received is increasing.



The time to complete a process cycle is also increasing.



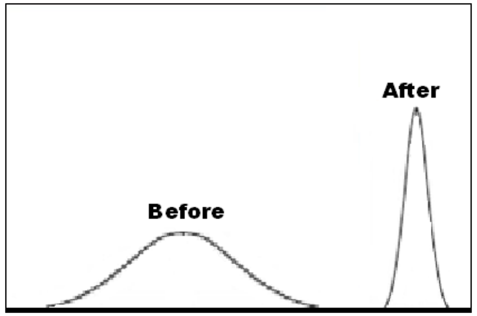
ANALYZE



Problem:
Incomplete and inaccurate applications were identified as the primary factor leading to defects in the process cycle.

Solution:
New Application process incorporating drop down menus

IMPROVE

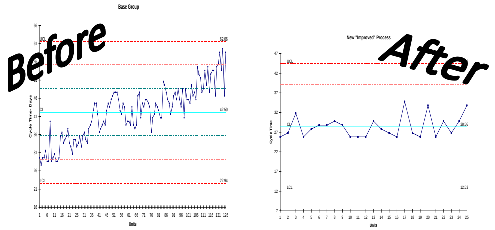


New Application Procedure =

Less Mistakes & Quicker Cycle Time

CONTROL

The defect rate reduced from 93% to 32%



Monthly monitor and review procedure is in place. Out of control signal = action plan.

PROJECT TEAM: *Dan • Mary • Karen • Linda • Peter*

BUSINESS CASE: $54,000 in annual processing costs

Process Improvement Project – Graphing Time Reduction

Mike – MBC 638

Key Dates --->	Team Launch	Define	Measure	Analyze	Improve	Control
	5/11/08	5/19/08	5/26/08	6/6/08	7/4/08	On-Going

DEFINE

- Extensive graphing is required for good data analysis of lab qualification testing
- 350 Engineer hours in the department are spent on repetitive graphing procedures within Excel. This equates to $52,471/year
- A 30% reduction in graphing time could result in a $15,741 annual savings.

= $

Control

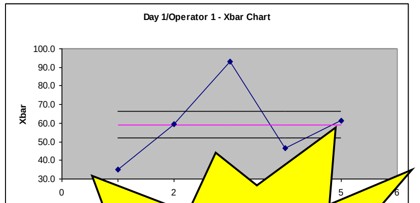
- Maintain Revision Control on original spreadsheet macros
- Provide to Engineering Department
- Survey engineers for usage in 3 months

MEASURE

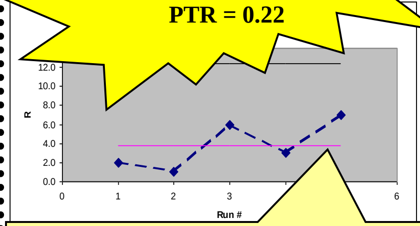
Identified Critical Inputs and Outputs to Measure

Data Stratificatio

Questions about Process	Output	Strati	Measurements
Is y affected by the number of data points recorded in a test?		X1 = ts	Number of data points collected
Is y affected by the total number of columns graphed?	Y = Graphing Time = f (X)	X_2 = Va luded in	Number of different variables measured
Is y affected by the person creating the graphs?		X_3 Operator	Name of engineer
Is y affected by the quality of the graph required?		X_4 = Graph quality	Title, labels, presentation ready
Is y affected by additional graphs required for comparison?		X_5 = Number of graphs	Number of graphs
Is y affected by the total number of columns of data available?		X_6 = Total number of available variables	Total number of available variables
		$Y = f(X_1,X_2,X_3,X_4,X_5,X_6)$	



R Charts in control.
Xbar Charts show ability to measure differences
PTR = 0.22

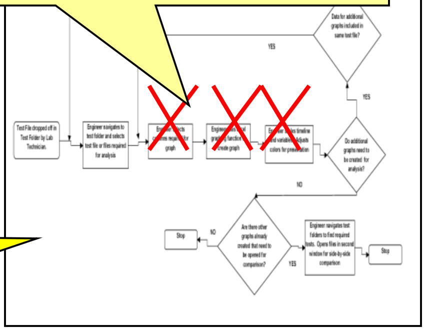


Measurement Systems Analysis using X-bar, R Charts and Precision-To-Total Ratio

ANALYZE

Pareto showed that 80% of graphs consisted of 8 variables or less

Eliminated wasteful, repetitive steps that can be automated with Excel Macros



n=	31
x bar (sec) =	56.7
s =	14.7
1- alpha =	0.95
alpha =	0.05
alpha/2 =	0.025
U=	61.89
L=	51.56

95% Con
51.56 <=
57 +/- 5.1

95% confidence interval that true graphing time is 57 +/- 5.16 seconds

IMPROVE

Before

After

SQL Baseline	
Population Mean =	56.7
Std Deviation =	14.7
X2 (Upper Spec Limit)	60
Z2 =	0.223
P(X>60) =	0.412
P(X is out of spec) =	0.412
DPM =	411655
SQL =	1.72

SQL Improved Process	
Population Mean =	36.2
Std Deviation =	13.5
X2 (Upper Spec Limit)	60
Z2 =	1.762
P(X>60) =	0.039
P(X is out of spec) =	0.039
DPM =	39052
SQL =	3.26

Sigma Quality Level Increased from 1.72 to 3.26

Cost Reduction		
	Current	Improved
Time per graph (sec)	57	36
Time spent graphing/year (hr)	350	224
Annual Cost Reduction	$ 52,471	$ 33,529
% Percent Annual Cost Reduction	36%	

$18,943 annual cost reduction! $

H0: mu >=	40
H1: mu <	40
Acceptable Level of Risk=	10%
New Pro	
New Pro	
P = 2* Z()	0.12
Confidence =	87.9%

Hypothesis test indicates 88% confidence that new process meets the goal of at least a 30% improvement in graphing time

BUSINESS CASE: $18,943 Annual Cost Reduction if Implemented in Engineering Department

Project Status – sharing tools

What tools are you using for:

DEFINE:

MEASURE:

ANALYZE:

Next two weeks

1.Project Next Steps – Measure/Analyze Phases

Measure/Analysis tools
Confirm your sample size
Insights about the problem

2. Coursework BLT's:

5.6 Test Your Knowledge

*5.7 Relate Sample Size to Your Project

6.3 Correlation Video

6.11 Test Your Knowledge: Hand/Foot Exercise

3. Assignments:

Homework #3: *(worth 2 points)*
Three days after live session 5
LaunchPad Assignments
- Chapter **8 Online Quiz** (unlimited attempts)

Upcoming assignment:

Quiz #2 (covers Chapters **3,6,8,9,11.2**)
Three days after live session 6