



Midterm Exam 1



Chapters 1 & 2 - Concepts

Key Terminology

Six Sigma: A data-driven management strategy

Quality: Fitness for use

Two aspects: a) Quality of design - high quality?

b) Quality of conformity - is the high quality consistent?

Characteristics or "dimensions of quality" that affect fitness for use as a product

(i) Performance performs required job (ii) Aesthetics product appearance

(iii) Reliability how often will it fail (iv) Features what the product does

(v) Durability lifetime → will it meet it (vi) Perceived Quality Reputation

(vii) Scrutability easy to repair (viii) Conformance to Standards

Producing a Service... list those dimensions of quality

(i) Responsiveness time to reply to requests

(ii) Professionalism knowledge, timeliness

(iii) Attentiveness caring, personalized attention

Hidden Factors: The part of the plant the customer doesn't see which affects w/bd quality

Quality Engineering: Consists of activities to ensure that ① that quality characteristics are at desired levels, and ② variability is at its minimum

→ **Quality Characteristics:** Important elements of a product/service to the customer

Variability: Describes or quantifies differences within a set of data or parts or c2qc

→ **Statistical Methods:** Analysis, interpretation, and presentation of numerical data

→ **Spec-Reading:** Desired measurements for quality characteristics This is your target value

Nonconformity: Failure to meet spec

Defect: Nonconforming that hurts usc

Variability

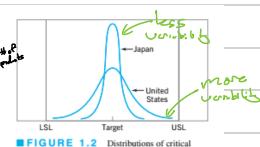
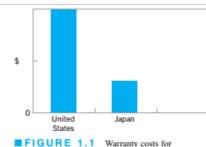
• Quality is inversely proportional to variability

◦ Variability will impact all quality dimensions

• Variability leads to greater cost & waste

example to right compares domestic vs. imported car quality

• Why variability leads to more cost for a company



Quality Costs

Prevention Costs	Internal Failure Costs
Quality planning and engineering	Scrap
New products review	Rework
Product/process design	Restest
Process control	Failure analysis
Burn-in 6-8 hrs covers 20 or test	Downtime
Training & Periodical	Yield losses
Quality data acquisition and analysis	Downgrading (off-specing)
Appraisal Costs	External Failure Costs
Inspection and test of incoming material	Complaint adjustment
Product inspection and test	Returned product/material
Materials and services consumed	Warranty charges
Maintaining accuracy of test equipment	Liability costs
	Indirect costs

- Products found in plant outside of specifications result in higher costs and waste
- Prevention & Appraisal costs are cheaper costs and result in more consistent products and cheaper costs down the line

Internal Failure & External Failure costs >> Prevention & appraisal costs!

Example: Cost of Variance ← Find manufacturing cost per good product

Illustrative Example: We are making 100 mechanical components per day where 75% of the parts conform to specifications, and the other 25% are non-conforming. Of the 25% of nonconforming, 60% can be reworked and the other 40% are scraped. Each part costs \$20 to manufacture and an additional \$4 to be reworked.

$$\frac{\text{cost}}{\text{good}} = \frac{\text{cost to make} (\# \text{made total}) + \text{cost to rework} (\# \text{reworked})}{\# \text{of goods sold}}$$

$$= \frac{\$20(100) + \$4(100)(0.25 \cdot 0.60)}{70 + 15} = \$22.69$$

Now, the Benefit of Improvement ← Find manufacturing cost per good product

Illustrative Example: We improve the previous process such that now 95% of the parts conform to specifications, and the other 5% are non-conforming. Similar to before, of the 5% of nonconforming, 60% can be reworked and the other 40% are scraped. Each part costs \$20 to manufacture and an additional \$4 to be reworked.

$$\frac{\text{cost}}{\text{good}} = \frac{\$20(100) + \$4(100)(0.05)(0.60)}{95 + 3}$$

$$= \$20.53$$

Quality Engineering

Process: A system of inputs and outputs

Example: Doing a process

Example: A hospital is trying to increase the quality of drug administration. To do this, it is considering providing patients with bar-coded wristbands to help guide workers. Your team is charged with studying the effects of bar-coding by carefully watching 250 episodes in which drugs are given to patients without bar coding and 250 episodes with bar coding. Every time a drug is administered you will check the amount, if any, of discrepancy between what was supposed to be given and what was given.

Control Charts: A process monitoring technique which tracks averages in a quality characteristic with time or sample number

- **Control:** Where the characteristic should be; find abnormal sources of variation
- **Control Limit:** Defined by statistics. Alert the user to unusual variability
↳ Different than spec limits

Designed experiments: Discovers very variable factors have a significant influence on quality characteristics in a process

Six Sigma

Broader Definition: Statistical based, data driven, management approach and continuous quality improvement methodology for eliminating defects in a product, process, or service

↳ Reduces variability to a level where defects are unlikely

Three Elements of Six Sigma

(i) **Quality Planning:** Listening to the voice of your customer

(ii) **Quality Assurance:** Establishing a system to prevent quality issues from arising

(iii) **Quality Control & Improvement:** A set of specific steps and tools to ensure products meet requirements and are continuously improved

Name Meaning: Refers to the quality level target such that sigma (σ , or variance σ^2) should be small enough so that 6σ is within the specification limits

Sigma Quality Level: The Probability that product/service is non-defective

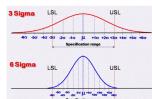
Why is Six Sigma Selected?

Suppose 100 parts: 3σ : $(.9973)^{100} = .7631$ is probability non-defective $\rightarrow 24\%$ defects
 6σ : $(.9999998)^{100} = .99999998$ non-defective $\rightarrow \sim .000002\%$ defects

Some companies use other multiples of σ , depending on what required by company/customers

Example of Six Sigma (will be on HW)

Illustrative Example: Let's say you are making a part that has a radius of 10 mm. Specification limits are between 9 (LSL) and 11 (USL) mm. Before you applied quality improvement, the standard deviation (sigma - a measure of variability) was 1 mm, and after you applied improvements, the standard deviation (sigma) decreased to 0.1. Is the new process 6 sigma quality?



$$\frac{|USL \text{ or } LSL - \text{target}|}{\sigma \text{ (standard deviation)}} = \frac{|10 \text{ mm} - 10 \text{ mm}|}{1 \text{ mm}} = 1 \quad \text{For old 1-sigma level}$$

$$= \frac{|10 \text{ mm} - 10 \text{ mm}|}{0.1 \text{ mm}} = 10 \quad \text{For new 10-sigma level}$$

Lean Manufacturing

Def: Another management philosophy aimed at eliminating waste (e.g., time and materials)

Process Cycle Efficiency (PCE): $(\text{Value added time}) / (\text{Process time})$

$\hookrightarrow \text{Value added time} = \text{time spent making the product more valuable}$

Process Cycle Time (PCT): $(\text{Number of items in progress}) / (\text{Completion rate})$ \hookleftarrow backlog issues as an example

Why is a slow process (low PCE or high PCT) an expensive process?

- Hidden Factory →

- (i) Customers don't like waiting
- (ii) More handling = more personnel
- (iii) More opportunity for damage or loss
- (iv) Inventory has to be higher
- (v) More documentation

DMACC Process

Def: A Project-based 5-step process and problem-solving procedure

$\hookrightarrow \text{Define} \rightarrow \text{Measure} \rightarrow \text{Analyze} \rightarrow \text{Improve} \rightarrow \text{Control}$

Toll gates: Where projects are reviewed to ensure on track

5 stages of DMACC (will be on HW too)

(i) Define Stage

- Main Obj: Identify a great opportunity and verify that it represents a legitimate value opportunity or breakthrough

◦ Breakthrough: Major improvement to a process or product

◦ Value Opportunity: The financial opportunity at stake

- Tools used

◦ SIPOC diagram: High level map of process

\rightarrow Suppliers: Those who provide information, material or other items that are used in the process

\rightarrow Input: The actual information or material provided

\rightarrow Process: The set of steps required to do the work

\rightarrow Output: The product delivered to the customer \rightarrow Other dept. in same company

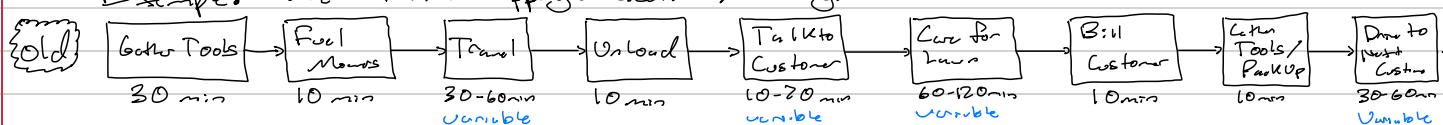
\rightarrow Customers: who buys the product (internal or external)

- Process Mapping: Involves creating flow diagrams for systems

◦ Value stream mapping: A variant of process mapping where engineers focus on steps which could be eliminated or simplified to reduce waste

- Value stream: The min amount of processing steps from raw material to customer

Example: Value Stream Mapping (Lawn Mowing)



Possible Improvements: Using customer's lawnmower, store SOP (home on truck), protocol, personnel/intense, eliminate/reduce customer interactions, optimize route (graph shortest distance)/logistics

- Pareto charting: Plots frequency (# of parts) or cost (time/money) against causes

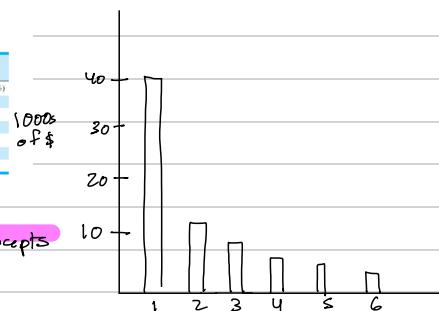
◦ Pareto Rule: 80% of the effects come from 20% of the causes

• Graph Pareto by Hand (Example)

Often in order from largest cost to smallest

Cost or % defects is on y-axis, category is on the x-axis

Monthly Quality-Costs Information for Assembly of Printed Circuit Boards		
Type of Defect	Percentage of Total Defects	Scrap and Rework Costs
1 Insufficient solder	42%	\$37,500.00 (52%)
2 Missing components	20	12,000.00
3 Defective components	15	8,000.00
4 Missing components	10	5,000.00
5 Cold solder joints	7	3,000.00
6 All other causes	5	4,600.00
Total:	100%	\$72,200.00



Can also do this with Lean concepts

Project Charter: Short document
with clear focus to product.

Example to right, but worse across companies

Business Case	Opportunity Statement																					
<ul style="list-style-type: none"> This project supports the business quality goals, namely (a) reduce customer resolution cycle time by x% and (b) improve customer satisfaction by y%. 	<ul style="list-style-type: none"> An opportunity exists to close the gap between our customer expectations and our actual performance by reducing the cycle time of the customer return process. 																					
Goal Statement	Project Scope																					
<ul style="list-style-type: none"> Reduce the overall response cycle time for returned product from our customers by x% year to year. 	<ul style="list-style-type: none"> Overall response cycle time is measured from the receipt of a product return to the time that either the customer has the product replaced or the customer is reimbursed. 																					
Project Plan	Team																					
<table border="1"> <tr> <th>Activity</th> <th>Start</th> <th>End</th> </tr> <tr> <td>Define</td> <td>6/04</td> <td>6/30</td> </tr> <tr> <td>Measure</td> <td>6/18</td> <td>7/04</td> </tr> <tr> <td>Analyze</td> <td>7/15</td> <td>8/09</td> </tr> <tr> <td>Improve</td> <td>8/15</td> <td>9/30</td> </tr> <tr> <td>Control</td> <td>9/15</td> <td>10/30</td> </tr> <tr> <td>Track Benefits</td> <td>11/01</td> <td></td> </tr> </table>	Activity	Start	End	Define	6/04	6/30	Measure	6/18	7/04	Analyze	7/15	8/09	Improve	8/15	9/30	Control	9/15	10/30	Track Benefits	11/01		<ul style="list-style-type: none"> Team Sponsor Team Leader Team Members
Activity	Start	End																				
Define	6/04	6/30																				
Measure	6/18	7/04																				
Analyze	7/15	8/09																				
Improve	8/15	9/30																				
Control	9/15	10/30																				
Track Benefits	11/01																					

(i:) Measure Stage

- Objectives:** Explore & understand the current state of the process
- Tools Used:** Quantify what the current process is performing at
- Measurement Capability:** Assess if the current measurement systems are adequate
- Tools Used:** Charts to understand variability

Histograms, bar charts, run charts, scatter diagrams

(ii:) Analyze Stage

- Objectives:** Use data to begin to determine cause and effect relationships and sources of variation

Common Causes of Variability: Internal Sources (noise)

Assignable Causes of Variability: External Sources

• Tools Used

Control Charts: Used to separate the common and assignable causes of variation

Hypothesis Testing & Confidence Intervals: Used to determine if different conditions produce significant results

Regression Analysis: Allows models of the process to be built

Failure Modes & Effects Analysis (FMEA): A method for identifying and prioritizing sources

For each item, we determine: of variability, or defects in a process or product

(i) Likelihood something will go wrong | (not likely) - 10

(ii) Ability to detect the failure or defect | (likely) - 10

(iii) Severity | (not severe) - 10

Risk Priority #: (i) x (ii) x (iii)

Example

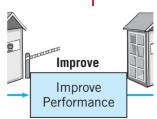
- Severity, 1 (not severe) - 10
- Probability of occurrence, 1 (not likely) - 10
- Probability of detection, 1 (likely) - 10

RPN = A X B X C



RPN = Risk Priority Number

FAILURE MODE & EFFECTS ANALYSIS (FMEA)					Date: 1/1/2018
Process Name: Left Front Seat Belt Install Process Number: SBT 445 Revision: 1.3					
Failure Mode	A) Severity	B) Probability of Occurrence	C) Probability of Detection	Risk Preference Number (RPN)	
1) Select Wrong Color Seat Belt	5	4	3	60	Audited
2) Seat Belt Bolt Not Fully Tightened	9	2	8	144	
3) Trim Cover Clip Misaligned	2	3	4	24	



(i) Improve Stage

Objectives

- Generate and quantify potential solutions
- Evaluate and select final solution
- Verify and gain approval for final solution

• Objectives

Use creative thinking & info gathered previously to make specific suggestions that will lead to desired solution process. Confirmation is important

• Tools Used

- Mistake Proofing (example)
- Optimization
- Pilot Testing
- Conformance Experiments

◦ Documentation

- (i) Solution obtained
- (ii) Alternatives considered
- (iii) Confirmation tests
- (iv) Risk Analysis

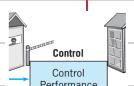
(ii) Control Stage

• Objectives

Complete all remaining work and hand off to improved process so gains in progress are institutionalized

• Tools Used

- Control Charts
- Transition Plans



Objectives

- Develop ongoing process management plans
- Mistake-proof process design and control critical process characteristics
- Develop out-of-control action plans

Chapter 3 - Quantifying Stats in Quality Characteristics

Statistics

Sample Mean: Average Value

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

n is the number of data points in the sample
and $x_1, x_2, x_3, \dots, x_n$ are data points

For this data set: 12, 12, 15, 18, 21, 22, 22, 23, 25, 25, 26, 27, 28, 29, 30, 30, 31

$$\bar{x} = \frac{12+12+15+\dots+31}{17} = \frac{396}{17} = 23$$

Sample Variance:

{Sample var: s^2 }

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Standard Deviation: $s = \sqrt{s^2}$

{Sample std: s }

For this data set: 12, 12, 15, 18, 21, 22, 22, 23, 25, 25, 26, 27, 28, 29, 30, 30, 31

$$s^2 = \frac{(12-23)^2 + (12-23)^2 + \dots + (31-23)^2}{17-1} = 37$$

$$s = \sqrt{37} \approx 6.08$$

MINITAB: Stat > Basic Statistics > Display Descriptive Statistics > In Variables add data

Illustrative Example

Sample 1	Sample 2
$x_1 = 1$	$x_1 = 1$
$x_2 = 3$	$x_2 = 5$
$x_3 = 5$	$x_3 = 9$
$\bar{x} = 3$	$\bar{x} = 5$

Sample 3
$x_1 = 101$
$x_2 = 103$
$x_3 = 105$
$\bar{x} = 103$

Sample 1: $s = 2$

Sample 2: $s = 4$

Sample 3: $s = 2$

Tools to Describe Data Numerically and Visualize Variables

↳ Often easier to arrange sorted (low to high) data

- **Percentiles:** Tells us how much of the data is below a certain point

12, 12, 15, 18, 21, 22, 22, 23, 25, 25, 26, 27, 28, 29, 30, 30, 31

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

↳ Values

↳ Rank (1-indexed)

- **Lower Quartile (Q_1 , or 25th perc.):** Value with rank = $(n+1) \left(\frac{1}{4}\right)$

here: $\frac{17+1}{4} = \text{rank of } Q_1 \rightarrow \frac{18+21}{2} = 19.5$

- **Middle Quartile (Q_2 , or median, or 50th perc.):** Value with rank = $(n+1) \left(\frac{1}{2}\right)$

here: $\frac{17+1}{2} = \text{rank of } Q_2 \rightarrow 12.5$

- **Upper Quartile (Q_3 , or 75th perc.):** Value with rank = $(n+1) \left(\frac{3}{4}\right)$

here: $(17+1) \left(\frac{3}{4}\right) = \text{rank of } Q_3 \rightarrow 28.5$

- **Interquartile Range ($Q_3 - Q_1$):**

here: $Q_3 - Q_1 = 28.5 - 12.5 = 16$

- **Whisker Limits:** Max Limit = $Q_3 + 1.5(Q_3 - Q_1)$

Min Limit = $Q_1 - 1.5(Q_3 - Q_1)$

here: Max Limit = 43, Min Limit = 6

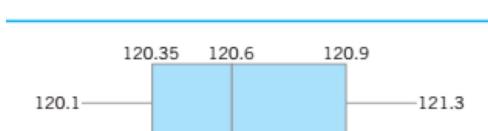
MINITAB: Stat > Basic Statistics > Display Descriptive Statistics > In Variables add data

- **Box Plot Features**

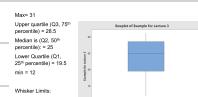
Displays 25th, 50th and 75th quartiles (ends and middle of the box)

- 25th and 75th are the ends of the box
- Median (50th percentile) is a single line through box
- The min and max of data (whiskers)
- Limits of whiskers (check) that the max and min do not go past (points that fall outside this can be outliers, represented as single data points and the new min or max is the next data point within the limit) – limit formula on next slide

TABLE 3.4		
Hole Diameters (in mm) in Wing Leading Edge Ribs		
120.5 120.4 120.7		
120.9 120.2 121.1		
120.3 120.1 120.9		
121.3 120.5 120.8		



here:



Minitab: Graph > Boxplot > Simple > input data column into graph variables

Excel

- Stdev is for sample data (use most often)
- Stdevp is for entire population

• **Histograms:** Plots of frequencies of data occurring w.r.t. the data sorted into bins/intervals

↳ **Bins:** The range that data will be sorted into $\# \text{ bins} = \text{ceil}(\sqrt{n}) \leftarrow \text{round up}$ $\text{bin size} = \frac{\text{max-min}}{\# \text{ bins}}$

- i. Need to determine the bin range for continuous data
- ii. Best if bins are the same size
- iii. Rule of thumb: number of bins = square root of number of observations (round up if not a whole number)

Note: for categorical data or discrete data, each category is its own bin

For this data set: 12, 12, 15, 18, 21, 22, 23, 25, 25, 26, 27, 28, 29, 30, 30, 31

$$\# \text{ of bins} = \sqrt{17} = 4.1 \rightarrow 5$$

$$\text{bin size} = \frac{31-12}{5} = 3.8$$

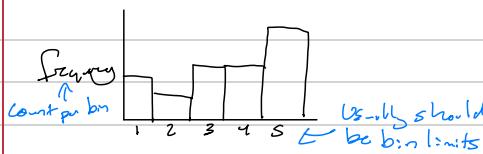
Bin 1: min + bin size $\rightarrow 12 + 3.8 \rightarrow [12, 15.8)$ Inclusive below, Exclusive above

Bin 2: bin 1 end + bin size $\rightarrow 15.8 + 3.8 \rightarrow [15.8, 19.6)$

Bin 3: $[19.6, 25.4)$

Bin 4: $[25.4, 29.2)$

Bin 5: $[29.2, 31]$



MINITAB: Graph > histogram > simple > add your column to graph variables
To change bins: double click on histogram bars (twice), select bins tab, change number of intervals to number of bins (Cut point is how we graphed in class)

↳ Final bin should be inclusive above

Significant Digits / Figures: Numbers of digits contributing to measurement resolution

Rules:

- (i) Non-zero digits always significant

(ii) Any zeros between two significant digits are significant (leading zeros are not)

(iii) Trailing zeros in the decimal portion only are significant (unless indicated by a bar)

Examples:

How many sig. figs in each number?	0.00500 = 3
0.03040 = 4	10 $\bar{2}$ = 4

How would you write this # with the correct sig. figs?	0.899923 (report 2 sig. figs) 0.40 23000.0 (report 3 sig. figs) 23 $\bar{0}$ 00
--	--

Will take away points for tons of sig figs!

Just pay attention to them!

For your sample means: keep as many significant figures as the number with the lowest significant digits in your data set (same for median and quartiles)

For your standard deviation: report to the last place that is significant, but no more

Your Average and SD should have some # of places!

E.g. Sample mean = 4.09 Sample Standard Deviation = 0.00923. What do I report? 4.09 ± 0.01

Probability Distributions Introductory

• **Sample:** Collection of measurements from a larger source or population

• **Population:** All of the elements from a group

• **Probability Distribution:** Mathematical Model that relates value of a variable with its probability of occurrence

• **Continuous Distribution:** The variable is expressed on a continuous scale

• **Discrete Distribution:** The variable can only take on certain values

• **Writing out Probability Statements**

◦ **Discrete:** The probability that x is equal to x_i ?

$$P\{x=x_i\} = p(x_i)$$

• **Calculating mean and variance**

$$\mu = \begin{cases} \int_{-\infty}^{\infty} xf(x) dx, & x \text{ continuous} \\ \sum_{i=1}^{\infty} x_i p(x_i), & x \text{ discrete} \end{cases}$$

◦ **Continuous:** The probability that x is between a and b ?

$$P\{a \leq x \leq b\} = \int_a^b f(x) dx$$

$$\sigma^2 = \begin{cases} \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx, & x \text{ continuous} \\ \sum_{i=1}^{\infty} (x_i - \mu)^2 p(x_i), & x \text{ discrete} \end{cases}$$

• **Median:** 50th percentile

• **Mean:** Central Tendency (Average)

• **Variance:** Spread of the distribution

M, σ^2 (population model parameters)
 \bar{x}, s^2 (sample data notation)

Ray attention to this notation

The Normal Distribution

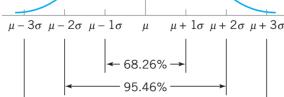
- For a random normal variable x , the probability distribution of x is: $P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

↳ Short hand $N(\mu, \sigma^2)$

○ Continuous, symmetric, "bell-shaped"

•

$$\text{General Formula: } P\{\mu - 3\sigma \leq x \leq \mu + 3\sigma\} = \int_{\mu-3\sigma}^{\mu+3\sigma} f(x) dx$$



In a Probability Band:

$$P\{\mu - \sigma \leq x \leq \mu + \sigma\} = 0.68$$

Due to symmetry:

$$P\{x > \mu\} = P\{x < \mu\} = 0.5$$

Holds for any normally distributed variable 'x'

- Got these properties: Set $\sigma=1$, $\mu=0$ so $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

$$\left\{ \begin{array}{l} a = \mu - 1\sigma = -1 \\ b = \mu + 1\sigma = 1 \end{array} \right.$$

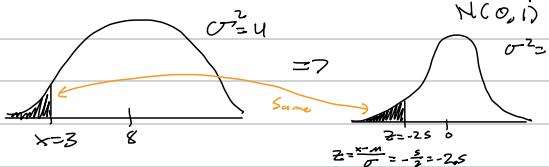
$$\int_{-1}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \text{(using mathematica)} 0.68$$

• Standardization: $N(\mu, \sigma^2) \Rightarrow N(0, 1)$

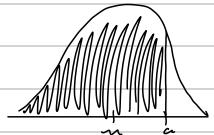
• Standard Normal Distribution: $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

• To standardize any random variable: $z = \frac{x-\mu}{\sigma}$ z is a standard normal variable

• Drawing out normal distributions: Suppose $N(8, 4)$

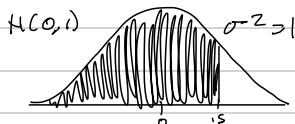


• Cumulative Normal Distribution: The probability that the variable x is less than or equal to some value a



Example: Assume Z is a standard normal random variable, and you want to find the probability that Z is less than or equal to 1.5

$$P\{Z \leq 1.5\} = 0.93$$

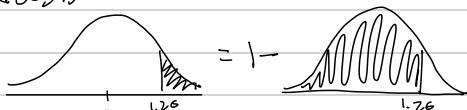


General Probability Demonstration

$$(i) P(Z > 1.26) = 1 - P(Z \leq 1.26)$$

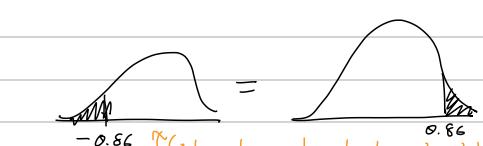
$$(ii) P(Z \leq -0.86) = P(Z > 0.86) = 1 - P(Z \leq 0.86)$$

$N(0, 1)$

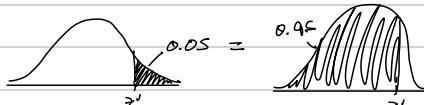


$$= 1 - 0.89616 = 0.1039$$

0.10



$$(iii) P(Z > Z') = 0.05$$



Chapter 4 - Hypothesis Testing

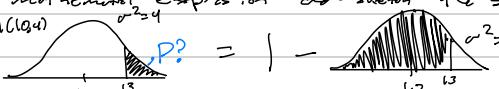
Example: Suppose the current measurements in a strip of wire are assumed to follow a normal distribution with a mean of 10 millamps and a variance of 4 millamps². What is the probability that the measurement will exceed 13 millamps?

1) Convert the word problem to mathematical expression and sketch the situation.

$$\mu = 10$$

$$P(x \geq 13)$$

$$\sigma^2 = 4$$

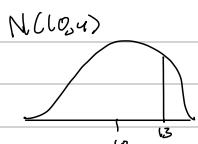


$$P? = 1 -$$

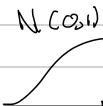
$$P(x \leq 13)$$

3) Convert sketch and formula to standard normal (if using tables in back of book) and a cumulative normal distribution (so it is compatible with the table in the book and Minitab)

Minitab: Calc > Probability dist. > Normal



=>



$$z = \frac{x - \mu}{\sigma} = \frac{13 - 10}{2} = 1.5$$

$$1 - P(z \leq 1.5) = 0.07$$

$$\rightarrow P(z \geq 1.5)$$

→ What about the integral?

$$P(z \geq 1.5) = \int_{1.5}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \quad \text{or} \quad P(x \geq 13) = \int_{13}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Checking for normality

→ How do we know if the normal distribution is a reasonable model for our data?

- **Probability Plotting:** Graphical method for determining whether sample data conform to a hypothesized distribution based on subjective visual examination.

- **Constructing a normal prob. plot by hand:** using $z_i = \frac{x_i - \bar{x}}{s}$

How-to (Steps)

1) Ranking data and ordering it $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$ where $x_{(i)}$ is the ranked value, $n =$ total number of values

2) Calculating cumulative frequency $F_i = i/n$ for each data point

3) Find a z_i value for each rank using the cumulative frequency value is the probability

Find z_{16} s.t. $P(z \leq z_{16}) = \frac{16-0.5}{17} = 0.25$

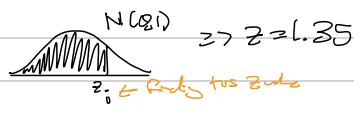
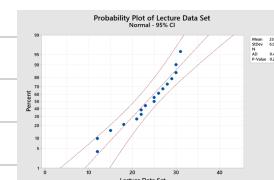
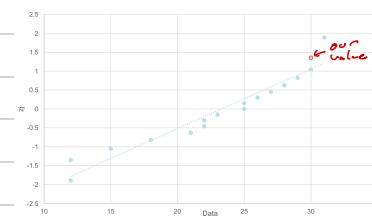
4) Plot the actual data (x -axis) vs. corresponding z_i (y -axis)

Normal Probability Plot in Minitab - Minitab: Graph > probability plot > graph variables, choose column with desired data

Note: each value is plotted versus the cumulative frequency with the y -axis is transformed so that the fitted distribution forms a straight line if normally distributed.

$$C. \text{ Freq } \left(\frac{i-0.5}{n} \right)$$

$$\text{For rank 16: } \frac{16-0.5}{17} = 0.91 \rightarrow z_{16}: P(z \leq z_{16}) = 0.91$$



$z_16 > Rely to the Rule$

$P(z > 1.35) < \alpha$

means not normal

In literature,

$\alpha = 0.05$

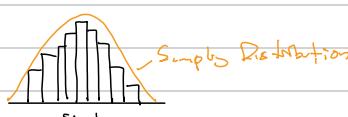
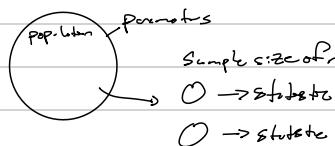
or $\alpha = 0.01$

Statistics Basic

- **Statistical Inference:** Drawing conclusions about a population based on sample from the population (of them selected randomly)

- **Statistics:** Any function of sample data that doesn't contain unknown parameters

- **Sampling Distribution:** The probability distribution of statistics



Suppose x_1, x_2, \dots, x_n are normally distributed random variable w/ given μ and σ^2

→ If x_1, x_2, \dots, x_n is a random sample of size n , the distribution of the sample mean is:

Normally distributed with mean of μ and covariance of $\frac{\sigma^2}{n}$

$$M_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

$$N(M, \frac{\sigma^2}{n}) \text{ shorthand}$$

- **Central Limit Theorem:** Implies that the sample mean will be approximately normally distributed for large sample sizes, regardless of the distribution you are sampling from.

→ **LARGE Sample size:** Some suggest $n > 20$ (non-normal), $n = 3-4$ (closer to normal)

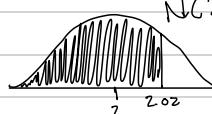
• Example

Electrodes are manufactured with a population mean thickness of 2.0 mm and a standard deviation of 0.048 mm. Find the probability that a random sample of $n = 16$ electrodes will have a sample mean diameter less than or equal to 2.02 mm? (keep 2 sig figs)

↳ Central Limit Theorem Applies

$$\bar{X} \Rightarrow N(\mu, \frac{\sigma^2}{n}) \quad n=16, \mu=2.0 \text{ mm}, \sigma=0.048 \text{ mm}, \sigma^2=0.0023$$

$$P(\bar{X} \leq 2.02)$$



$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

New!
Formulae

$$= \frac{2.02 - 2.0}{0.048/\sqrt{16}} = \frac{0.02}{0.012} = 1.66$$

$$P(Z \leq 1.66) = 0.45$$

Hypothesis Testing

- **Parameters:** What probability distributions we described with

e.g. normal dist. is described with μ, σ

- **Statistical Hypothesis:** A statement about values of parameters of a probability distribution

- **Null Hypothesis:** States no statistical difference exists (H_0)

- **Alternative Hypothesis:** Contradicts to null hypothesis (H_1)

Example: Suppose we think the mean inside diameter of bearing is 1.500 in

$$H_0: \mu = 1.500 \quad H_1: \mu \neq 1.500 \quad \text{two-sided}$$

$$\left\{ \begin{array}{l} \mu < 1.500 \\ \mu > 1.500 \end{array} \right. \quad \text{one-sided}$$

- Two outcomes of a hypothesis test:

(i) Reject H_0 or (ii) Fail to reject H_0

- **Test Statistic:** Used to test a hypothesis and is computed from a random sample of the population

- **Sampling Distribution (Null Distribution):** The sampling distribution of the test statistic assuming that H_0 (the null hypothesis) is true

- **Critical Region:** Sets values of the test statistic that lead to the rejection of the null hypothesis

↳ Can either fail to reject or reject H_0 , critical region defines these cutoffs

- Two types of errors

- (i) **Type I errors:** If the null hypothesis is rejected when it is true

↳ Defn: $\alpha = P\{\text{type I error}\} = P\{\text{reject } H_0 | H_0 \text{ is true}\}$ This is specified

- (ii) **Type II errors:** If the null hypothesis is not rejected when it is false

↳ Defn: $\beta = P\{\text{type II error}\} = P\{\text{fail to reject } H_0 | H_0 \text{ is false}\}$ Hard to control

○ Type II errors are uncontrollable & unknown, so we try to minimize + avoid it

- **Conclusions:** Must be worded very clearly

- **Rejection of Null:** A strong conclusion because it controls

○ to be small

- **Failing to Reject:** Weak conclusion because it has to control β

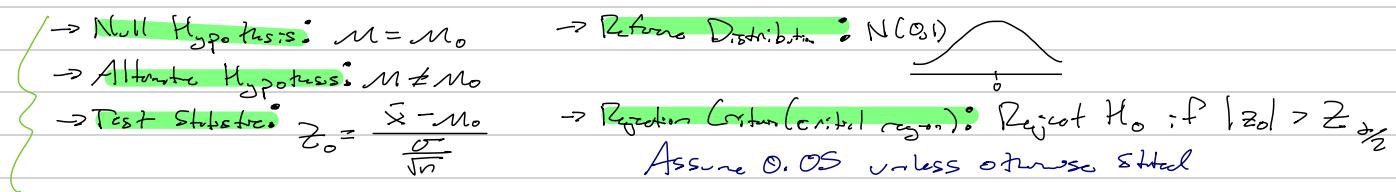
↳ to be small

The Claim is the	The Decision is	The Conclusion is
Null Hypothesis	Fail to Reject the Null	The evidence is <i>not sufficient</i> to reject the claim.
	Reject the Null	The evidence is <i>sufficient</i> to reject the claim.
Alternative Hypothesis	Fail to Reject the Null	There is <i>insufficient</i> evidence to support the claim.
	Reject the Null	There is <i>sufficient</i> evidence to support the claim.

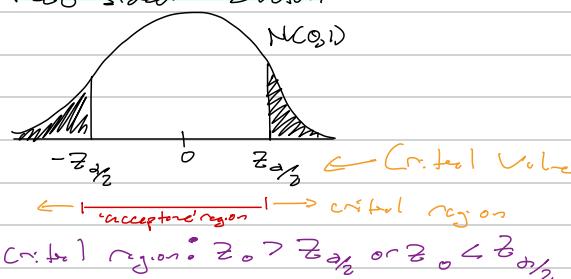
We never say accept the null hypothesis

One Sample Z-test & Inference on the Mean of a Population, Variance Known

Suppose x is a random variable with unknown mean μ and a known variance σ^2 and we wish to test the hypothesis that the population mean is equal to a standard value (H_0).



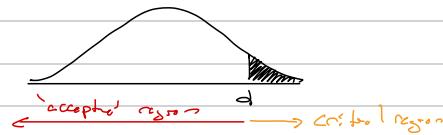
Two-sided Sketch



One-sided Test

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$



General Procedure C for this class!

- Parameters of Interest: From the problem context, identify known/unknown parameters of interest, the test you're going to do and identify type of problem.
- Null Hypothesis: State the null hypothesis H_0 .
- Alternative Hypothesis: State the alternative hypothesis H_1 .
- Test Statistic: Define an appropriate test statistic & reference distribution.
- Reject H_0 : If state rejection criteria for the null hypothesis. Should be drawn on plot w/ test statistic.
- Computations: Compute necessary sample quantities, substitute those into eqn for test statistic, and compute test value. Compare to rejection criteria.
- Draw Conclusion: Decide whether or not H_0 should be rejected and report that in the problem context. Phrase appropriately!

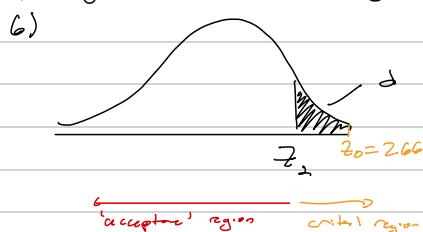
Example

Example: You are testing the response time of a distributed computer system. The system manager wants to know whether the mean response time to a specific command exceeds 75 ms. Previous experience has determined the standard deviation is 8 ms. Use $\alpha = 0.05$. A test is conducted executing the command 25 times, and the mean response time is 79.25 ms.

4) Test Statistic: $Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{79.25 - 75}{8/\sqrt{25}} = 2.66$

Reference Dist: $N(0,1)$

5) Rejection criterion: $|Z_{\text{reject}}| > Z_{\alpha} = Z_{0.05}$



1) Parameters of Interest: μ , inferring mean of population, known variance mean response time

↳ Test: One-sample z-test

2) $H_0: \mu = 75 \text{ ms}$

3) $H_1: \mu > 75 \text{ ms}$ (one-sided)

7) Conclusion: $Z_0 > Z_{\alpha}$

Reject the null Hypothesis. Conclude that the mean response time exceeds 75 ms.

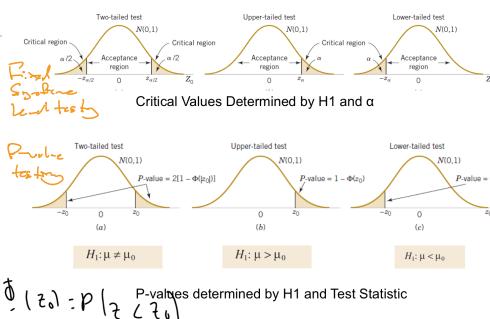
$P(Z < Z_{\alpha}) = 0.95$

$Z_{\alpha} = 1.65$ critical value

How do I do this in Minitab? Minitab>Stat>Basic Stats>OneSampleZ – click summarized data, but you could also highlight a column of data, options will allow us to set the alpha value and the alternative hypothesis

Summary of Z-tests

Testing Hypotheses on the Mean, Variance Known (Z-Tests)		
Null hypothesis: $H_0: \mu = \mu_0$	Test statistic: $Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$	Alternative Hypothesis
$H_1: \mu \neq \mu_0$	Probability above $ z_0 $ and probability below $- z_0 $: $P = 2(1 - \Phi(z_0))$	Rejection Criterion for Fixed-Level Tests
$H_1: \mu > \mu_0$	Probability above z_0 : $P = 1 - \Phi(z_0)$	
$H_1: \mu < \mu_0$	Probability below z_0 : $P = \Phi(z_0)$	
		Two-tailed test: $N(0,1)$
		Upper-tailed test: $N(0,1)$
		Lower-tailed test: $N(0,1)$
		$\alpha/2$
		Acceptance region: $z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
		Critical region: $ z_0 > z_{\alpha/2}$
		$\alpha/2$
		Two-tailed test: $N(0,1)$
		Upper-tailed test: $N(0,1)$
		Lower-tailed test: $N(0,1)$
		$\alpha/2$
		Acceptance region: $z_0 > z_{\alpha/2}$
		Critical region: $z_0 > z_{\alpha}$
		α
		Acceptance region: $z_0 < -z_{\alpha}$
		Critical region: $z_0 < -z_{\alpha}$
		α
		Two-tailed test: $N(0,1)$
		Upper-tailed test: $N(0,1)$
		Lower-tailed test: $N(0,1)$
		α
		Acceptance region: $z_0 > z_{\alpha}$
		Critical region: $z_0 > z_{\alpha/2}$
		$\alpha/2$
		Acceptance region: $z_0 < -z_{\alpha/2}$
		Critical region: $z_0 < -z_{\alpha}$
		α
		Two-tailed test: $N(0,1)$
		Upper-tailed test: $N(0,1)$
		Lower-tailed test: $N(0,1)$
		α
		Acceptance region: $z_0 > z_{\alpha/2}$
		Critical region: $z_0 > z_{\alpha}$
		α
		Acceptance region: $z_0 < -z_{\alpha/2}$
		Critical region: $z_0 < -z_{\alpha}$
		α



Hypothesis Testing Basics

- Making statements about population probability distribution parameters (e.g. mean or variance)
- Testing the statements using a test statistic – calculated from a randomly selected sample
- Determining criteria for rejection of a null hypothesis - defined by Type I error level and H_1
- Remembering when making conclusions that the test is designed such that rejection of null is strong, not rejecting null – not so strong
- General: so many different hypothesis tests!!

P-values

R-value: The probability of obtaining a sample outcome at least as extreme as the one obtained given the value stated if the null hypothesis is true
 \Rightarrow Reporting the compatibility of the data with the null hypothesis

Rejection Criteria: Reject H_0 if the p-value $< \alpha$

Back to our example, but now with p-values!

The response time of a distributed computer system is a CTQ. The system manager wants to know whether the mean response time to a specific command exceeds 75 ms. Previous experience has determined the standard deviation is 8.0 ms. A test is conducted executing the command 25 times, and the mean response time is 79.25 ms. Compute the P-value appropriate for this situation.

Rejection Criteria: Reject H_0 if p-value $< \alpha$



Z-test Assumptions

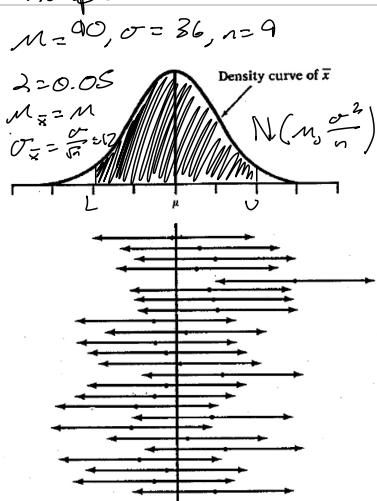
- Cts data
- Normal dist. D.t.
- Random Sample
- Popultn Std dev is known

Confidence Intervals

The interval between two statistics that includes the true value of the parameter with some probability
 $\rightarrow 100(1-\alpha)\%$ confidence interval

$$\text{m: } P(L \leq m \leq U) = 1 - \alpha \quad \text{where the statistics } \begin{cases} L = \text{Lower confidence limit} \\ U = \text{Upper confidence limit} \end{cases}$$

Example



Find L and U s.t.

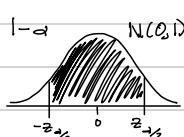
$$P(L \leq \bar{x} \leq U) = 0.95$$

$$\rightarrow L = 66.5, U = 113.5$$

Assuming L & U are equivalent from the sample mean

$$\rightarrow L, U \text{ are } 23.5 \text{ units from } m$$

Gently



$$1 - \alpha$$

$$P(-z_{\alpha/2} \leq \frac{\bar{x} - m}{\sigma / \sqrt{n}} \leq z_{\alpha/2})$$

$$\rightarrow P(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

CI includes

L U

$$\text{One side: } m \leq \bar{x} + z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{x} - z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}} \leq m$$

Recall

- $Z_{\alpha/2}$ is value of the $N(0,1)$ s.t. $P(Z \geq Z_{\alpha/2}) = \frac{\alpha}{2}$
- Z_{α} is value of the

Minitab: Stat > Basic Stats > One sample Z, click summarized data and no hypothesis test, click options to select the % of the CI and if it is two sided or one sided

Actual Example/Application

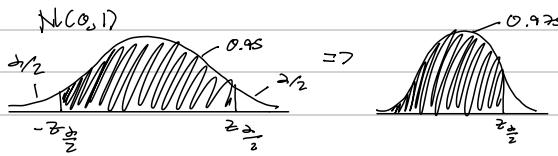
We pull a sample of 20 electrodes from an assembly line of a process with known variance in electrode thickness to be 0.64 mm². The average thickness of the sample is 7.9 mm. Construct a 95% confidence interval.

$$n = 20$$

$$\sigma^2 = 0.64 \text{ mm}^2 \rightarrow \sigma = 0.8$$

$$\bar{x} = 7.9 \text{ mm}, s = 0.08$$

$$N(0, 1)$$



$$P\left\{\bar{x} - \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

$$L = 7.9 - 1.96 \frac{0.8}{\sqrt{20}} = 7.5$$

$$U = 7.9 + 1.96 \frac{0.8}{\sqrt{20}} = 8.3$$

Often our result would be said like this:

~~"We are 95% confident that the mean of the population is between 7.5 and 8.3."~~

~~Not this!~~

Perhaps a better description would be:

"In repeated sampling, this method will produce intervals that capture the true population mean about 95% of the time."

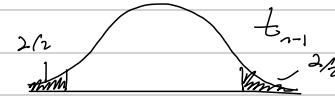
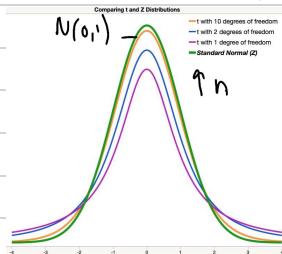
Use this front bearing exam!"

One Sample t-test

- Inference on mean distribution, unknown variance.
- Suppose x is a normal random variable with unknown mean μ and unknown variance σ^2 .
- Null Hypothesis: $H_0: \mu = \mu_0$
- Alternative Hypothesis: $H_1: \mu \neq \mu_0$
- Test Statistic: $T_0 = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$
- Rejection Criterion: $|T_0| > t_{\alpha/2, n-1}$

Two-tailed degrees of freedom

Rejection Criterion: $|T_0| > t_{\alpha/2, n-1}$



For the t-distribution:

- Is a continuous probability distribution of the z-score when the estimated standard deviation is used in the denominator rather than the true standard deviation.
- It is a type of normal distribution, except it accounts for the fact that as sample size increases, variability decreases.
- The t-distribution, like the normal distribution, is bell-shaped and symmetric about zero, but it has heavier tails, which means it tends to produce values that fall farther from its mean.
- As the sample size increases, the t-distribution becomes more similar to a normal distribution. The shape depends on n !! And this is why you MUST write down n each time you draw out a t-distribution.

Summary

Testing Hypotheses on the Mean of a Normal Distribution, Variance Unknown

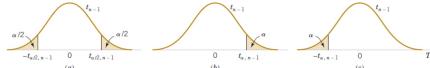
Null hypothesis: $H_0: \mu = \mu_0$

Test statistic: $T_0 = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$

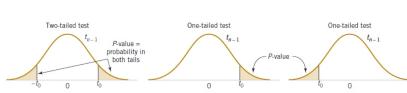
t dist with $n-1$ D.F

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Probability above $ T_0 $ and probability below $- T_0 $	$T_0 > t_{\alpha/2, n-1}$ or $T_0 < -t_{\alpha/2, n-1}$
$H_1: \mu > \mu_0$	Probability above T_0	$T_0 > t_{\alpha, n-1}$
$H_1: \mu < \mu_0$	Probability below T_0	$T_0 < -t_{\alpha, n-1}$

Fixed Significance Test by



P-Value



t-test Assumes

- Data are discrete.
- Data are normally distributed.
- Data are random samples.

Example

Inference on the mean of a distribution, unknown variance

Rubber can be added to asphalt to reduce road noise. The viscosity (cP) of 15 specimens of asphalt is measured. To be suitable for use, the mean should be equal to 3200 cP. Test the hypothesis using alpha=0.05.

Using this data
 $n=15$

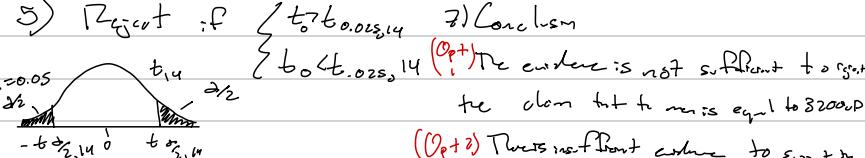
cP
3193
3124
3153
3145
3093
3466
3355
3079
3182
3227
3256
3332
3204
3282
3170

- Inference on mean, unknown variance, one-sample
- One-sample t-test

$$H_0: \mu = 3200 \text{ cP} = \mu_0 \quad H_1: \mu \neq 3200 \text{ cP}$$

$$T_0 = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}, \quad \bar{x} = 3193 + 3124 + \dots + 3170 = 3217$$

$$\text{Plug & Chug} = 0.64 \quad S = \sqrt{\frac{(3193 - 3217)^2 + \dots + (3170 - 3217)^2}{15 - 1}} = 106$$



$$t_{\alpha/2} = 2.145$$

men not bad equal to 3200 cP.

(Op+2) The curve is not sufficient to reject the claim that the mean is equal to 3200 cP.

CI for One-sample t-test

Suppose x is a normal random variable with unknown mean μ and unknown variance σ^2 . From a sample of n observations, the sample mean \bar{x} and sample variance s^2 are computed. Then:

$$\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

Is a 100(1-alpha)% two-sided CI

where $t_{\alpha/2, n-1}$ is the value of the t distribution with $n-1$ degrees of freedom such that $P(t_{n-1} \geq t_{\alpha/2, n-1}) = \alpha/2$

$$-\infty < \mu < \bar{x} + t_{\alpha, n-1} \frac{s}{\sqrt{n}}$$

Is a 100(1-alpha)% upper confidence bound

where $t_{\alpha, n-1}$ is the percentage point of the t distribution with $n-1$ degrees of freedom such that $P(t_{n-1} \geq t_{\alpha, n-1}) = \alpha$

$$\bar{x} - t_{\alpha, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq 0$$

Is a 100(1-alpha)% lower confidence bound

where $t_{\alpha, n-1}$ is the percentage point of the t distribution with $n-1$ degrees of freedom such that $P(t_{n-1} \geq t_{\alpha, n-1}) = \alpha$

→ For our previous example

$$\bar{x} = 3217$$

$$s = 106$$

$$n = 15$$

$$L = 3217 - 2.145 \frac{106}{\sqrt{15}} = 3158$$

$$U = 3217 + 2.145 \frac{106}{\sqrt{15}} = 3276$$

$$t_{0.025, 14} = 2.145$$

$$3158 \leq \mu \leq 3276$$

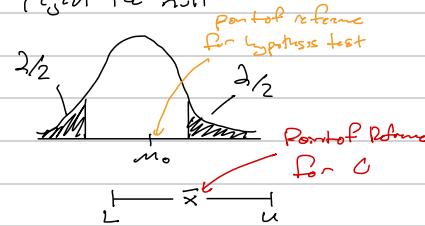
CI Relationship w/ Hypothesis Test

- Both do make inferences about a population parameter (but in different ways)
- A CI provides a range of values likely containing the true population parameter based on a set of data
- A hypothesis test determines whether a specific claim about the population parameter is supported by the data
- Difference in emphasis of the point of reference

Minitab: Stat>Basic Stat>1-sample t-test, samples in columns (drop down), double click data, check hypothesis test select type and in options select alpha value

Illustration

If the value in your H_0 is in your CI, you likely cannot reject the null



Inference or decisions on Normal Data (Chi-square test)

Suppose we wish to test the hypothesis that the variance of a normal population (σ^2) equals a specified value, say σ_0^2 , or that the standard deviation of the population (σ) is equal to σ_0 . Let x_1, x_2, \dots, x_n be a random sample of n observations from this population to test:

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_a: \sigma^2 \neq \sigma_0^2$$

$$\text{Test Statistic: } \chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \quad s = \text{sample variance with } n \text{ observations}$$

Reference Distribution: Chi-Squared Distribution

Useful for many controls
about the variability of your process

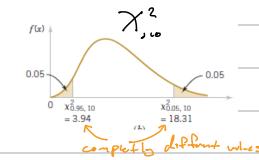
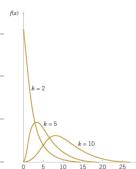
Which Hypothesis should I do?

Test	# of Populations	Variance Known?	Parameter of interest?	Reference Dist.
One-sample Z-test	1	Yes	Mean	Standard normal
Two-sample Z-test	2	Yes	Difference between means	Standard normal
One-sample t-test	1	No	Mean	t-distribution
Two-sample (pooled) t-test (equal variance)	2	No	Means of two populations	t-distribution
Two-sample t-test (unequal variance)	2	No	Means of two populations	t-distribution
Paired t-test	2 (simplifies to 1)	No	Difference between paired means	t-distribution
1 Variance test	1	No	Variance of one population	chi-squared distribution

If both mean and variance of populations are known, it's a probability problem

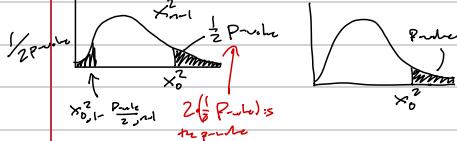
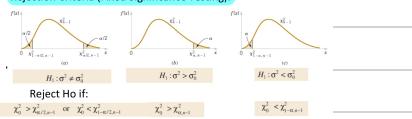
Chi-squared Distribution

- χ^2 has a chi-squared distribution if true null is true
- Very Features:
 - (i) Depends on $n-1$ degrees of freedom
 - (ii) Not symmetric
 - (iii) Not centered at 0
 - (iv) Total Area under curve is still 1!



Rejections

Rejection Criteria (Fixed Significance Testing):



Summary

Null hypothesis: $H_0: \sigma^2 = \sigma_0^2$
Test statistic: $\chi^2_0 = \frac{(n-1)s^2}{\sigma_0^2}$

Alternative Hypothesis

$H_1: \sigma^2 \neq \sigma_0^2$	$\chi^2_0 > \chi^2_{\alpha/2, n-1}$ or $\chi^2_0 < \chi^2_{1-\alpha/2, n-1}$
$H_1: \sigma^2 > \sigma_0^2$	$\chi^2_0 > \chi^2_{\alpha, n-1}$
$H_1: \sigma^2 < \sigma_0^2$	$\chi^2_0 < \chi^2_{1-\alpha, n-1}$

Ref. Dist: Chi-Square

Check Normality! It is sensitive to this

Example

Inference on the variance of a distribution

Rubber can be added to asphalt to reduce road noise. The viscosity (cP) of 15 specimens of asphalt is measured. s is calculated to be 106. Let's say we are interested in knowing if the variance less than 30,000.

Use $\alpha = 0.05$

1) Parameter of interest: $\sigma^2 \rightarrow$ 1-variance test

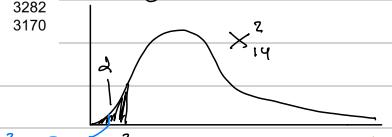
2) $H_0: \sigma^2 = 30,000 \text{ cP}^2$

3) $H_a: \sigma^2 < 30,000 \text{ cP}^2$

$$4) \chi^2_0 = \frac{(n-1)s^2}{\sigma^2} = \frac{(15-1)(106)^2}{30,000} = 5.2$$

Reference Distribution: χ^2_{14}

5) Rejection Criteria: Reject H_0 if $\chi^2_0 < \chi^2_{1-\alpha, n-1}$

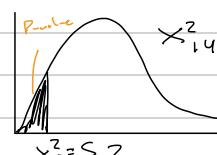


6) Computation: $5.2 < 6.57$

7) Conclusion: Reject H_0 . We have evidence to reject the claim that the variance is 30,000 cP². Having evidence to accept the hypothesis that variance is less than 30,000 cP². We accept H_1 . (or accept H_1 , but not accept H_0)

purple

Reject H_0 : purple < 2



$$\chi^2_0 = 5.2$$

$$\chi^2_{1-\alpha, 14} = 6.57$$

χ^2_0 lies between

0.490 and 0.975

\rightarrow purple is below 1-tuse

$(0.01 < p < 0.025) \rightarrow$ both < 2 so reject H_0 !

Fixed-Significance Level

Confidence Intervals for Inference on the Variance (χ^2)

Suppose x is a normal random variable with unknown mean μ and unknown variance σ^2 . From a sample of n observations, the sample mean \bar{x} and sample variance s^2 are computed. Then:

(100- α)/2. two-sided CI

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}$$

Where $\chi^2_{\alpha/2, n-1}$ denotes the percentage point of the chi-square distribution such that $P(\chi^2_{n-1} \geq \chi^2_{\alpha/2, n-1}) = \alpha/2$ and $\chi^2_{1-\alpha/2, n-1}$ denotes the percentage point of the chi-square distribution such that $P(\chi^2_{n-1} \geq \chi^2_{1-\alpha/2, n-1}) = 1 - \alpha/2$

(100- α)/2. upper bound

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha, n-1}}$$

Where $\chi^2_{1-\alpha, n-1}$ denotes the percentage point of the chi-square distribution such that $P(\chi^2_{n-1} \geq \chi^2_{1-\alpha, n-1}) = 1 - \alpha$

(100- α)/2. lower bound

$$\frac{(n-1)s^2}{\chi^2_{\alpha, n-1}} \leq \sigma^2$$

Where $\chi^2_{\alpha, n-1}$ denotes the percentage point of the chi-square distribution such that $P(\chi^2_{n-1} \geq \chi^2_{\alpha, n-1}) = \alpha$

How do I do this in Minitab? Stat > basic stat > 1 variance

For our example (Upper bound)

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha, n-1}} \rightarrow \sigma^2 \leq \frac{(15-1)(106)^2}{\chi^2_{0.05, 14}}$$

$$\chi^2_{1-\alpha, n-1} = \chi^2_{0.05, 14} = 6.57$$

$$\Rightarrow \sigma^2 \leq \frac{(14)(106)^2}{6.57} = 234.42 \text{ cP}^2$$

6-sigma Example/Application

Inference on the variance of a distribution

Rubber can be added to asphalt to reduce road noise. The viscosity (cP) of 15 specimens of asphalt is measured. The asphalt specification limit is 1100 cP away from the mean in either direction – what is the sigma level of this process given that we know $\sigma = 173$ (from our hypothesis test) and we assume the mean of the population is on target?

$$\text{Sigma level} = \frac{\text{Spec limit} - \text{target}}{\sigma} = \frac{1100}{173} = 6.35 \rightarrow \text{at least } 6.35$$

Data

3193
3124
3153
3145
3093
3466
3355
3079
3182
3227
3256
3332
3204
3282
3170

1-variance Test Assumptions

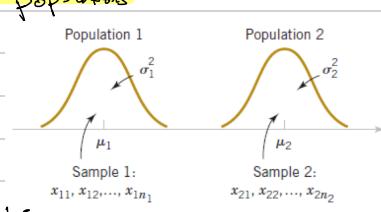
1. The data are continuous.

2. The data are randomly sampled.

3. The data are normally distributed.

Inference on means of two normally distributed populations

Let's say population 1 has a mean of μ_1 and a variance of σ_1^2 and population 2 has a mean of μ_2 and a variance of σ_2^2 and inferences are based on two random samples of size n_1 and n_2 respectively.



Want to know: is there a difference between two populations

Two-Sample Z-test

Statistical inference on the difference in means $\mu_1 - \mu_2$ of the populations below with known variances of σ_1^2 and σ_2^2

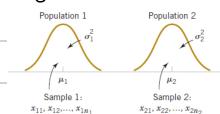
$$\text{Null: } H_0: \mu_1 - \mu_2 = \Delta_0$$

Refers Dist: $N(0,1)$

$$\text{Alt: } H_1: \mu_1 - \mu_2 \neq \Delta_0 \quad (\text{two-tail})$$

Rejection Criteria: Same as one-sample Z-test

$$\text{Test statistic: } Z_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$



Summary

$$\text{Null hypothesis: } H_0: \mu_1 - \mu_2 = \Delta_0$$

$$\text{Ref: } N(0,1)$$

$$\text{Test statistic: } Z_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$(10-2)$$

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probability above $ z_0 $ and probability below $- z_0 $, $P = 2[1 - \Phi(z_0)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probability above z_0 , $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probability below z_0 , $P = \Phi(z_0)$	$z_0 < -z_\alpha$

Assumptions

- (1) $X_{11}, X_{12}, \dots, X_{1n_1}$ is a random sample from population 1.
- (2) $X_{21}, X_{22}, \dots, X_{2n_2}$ is a random sample from population 2.
- (3) The two populations represented by X_1 and X_2 are independent.
- (4) Both populations are normal.

- (5) The variances of the populations are known

Samples from two populations are independent
if the samples selected from one population have no relationship with the samples selected from the other population.

Example

Paint Drying Time A product developer is interested in reducing the drying time of a primer paint.

Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample average drying times are $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

$$H_0: \mu_1 - \mu_2 = 0 \quad (n_1 = n_2)$$

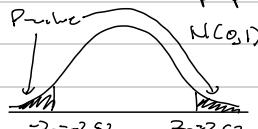
$$H_1: \mu_1 - \mu_2 \neq 0 \quad (n_1 \neq n_2)$$

$$Z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{121 - 112}{\sqrt{\frac{8^2}{10} + \frac{8^2}{12}}} = 2.52$$

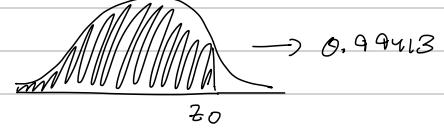
$$P\text{-value: Reject } H_0: \text{if } p\text{-value} < \alpha = 0.05$$

$$P\text{-value} = 0.012 < \alpha = 0.05$$

Reject H_0 and accept H_1 , and conclude there is a difference btwn the drying time of the two



$$p\text{-value} = 2(1 - 0.9943) = 0.012$$



Confidence Intervals (do they cross zero?)

Confidence Intervals (do they cross zero!?)

100(1- α)% two-sided confidence interval on $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $Z_{\alpha/2}$ is the value of the $N(0,1)$ such that $P(z \geq Z_{\alpha/2}) = \alpha/2$

$$\mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + Z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\bar{x}_1 - \bar{x}_2 - Z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2$$

where Z_α is the value of the $N(0,1)$ such that $P(z \geq Z_\alpha) = \alpha$

Two-Sample t-test (variance unknown)

Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ \leftarrow unknown var, but assumed true

Suppose we have two independent populations with unknown means μ_1 and μ_2 and equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and we wish to know if the means are different:

H₀: $\mu_1 - \mu_2 = \Delta_0$ ($\mu_1 = \mu_2$)

H_a: $\mu_1 - \mu_2 \neq \Delta_0$ ($\mu_1 \neq \mu_2$)

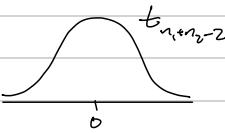
$$\text{Test Statistic: } T_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Since $n_1 = n_2$

Pooling Sample Variance

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

Weighted average of two sample variances



Rejection Criteria: See below exercise

t-dist with n_1+n_2-2 degrees of freedom

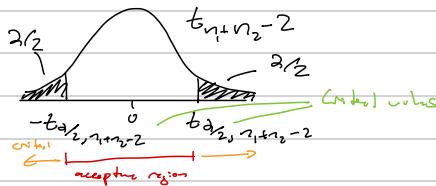
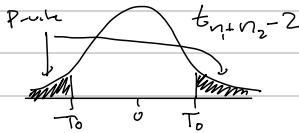
Exercise: Sketch out fixed rejection criteria & p-value for two-sided H_a

P-value:

Reject H₀: P p-value < α

Fixed:

Reject: P $T_0 > t_{\alpha/2, n_1+n_2-2}$ or $T_0 < -t_{\alpha/2, n_1+n_2-2}$



Example

Yield from a Catalyst Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently used; but catalyst 2 is acceptable. Because catalyst 2 is cheaper, it should be adopted, if it does not change the process yield. A test is run in the pilot plant

Observation Number	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75
	$\bar{x}_1 = 92.255$	$\bar{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

1) Interested in difference b/w 2 means, specifically yield of catalyst 1 and catalyst 2, variance unknown \rightarrow two-sample t-test
 \hookrightarrow Assume equal variances \leftarrow rule of thumb later

2) $H_0: \mu_1 - \mu_2 = 0$

3) $H_a: \mu_1 - \mu_2 \neq 0$

$$4) T_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{92.255 - 92.733}{2.7 \sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.35$$

$$5) S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{7 \cdot 2.39^2 + 7 \cdot 2.98^2}{8+8-2} = 2.3$$

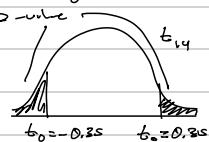
$$\rightarrow S_p = \sqrt{S_p^2} = \sqrt{2.3} = 2.3$$

Checking in Minitab: stat -> basic stat -> 2 sample t -> input columns and other info (CLICK ASSUME EQUAL VARIANCE UNDER OPTIONS!)

5) Reject: P p-value < → 0.05

$$6) t_0 = -0.35$$

7) p-value > α



Fail to reject H₀, redundant new evidence to say the yield from catalyst 1 is different from catalyst 2.

0.5 < p-value < 0.8

Two Sample t-test (variances unknown & unequal)

Case 1: $\sigma_1^2 \neq \sigma_2^2$

Suppose we have two independent populations with unknown means μ_1 and μ_2 and unequal variances $\sigma_1^2 \neq \sigma_2^2$ and we wish to test if the means are different:

Example calculation: $s_1 = 7.63$
 $s_2 = 15.3$
 $n_1 = n_2 = 10$

$H_0: H_a$ are same as 2 sample t-test w/ equal variance
 $T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ Reference Df for t-dist with \checkmark
 degrees of freedom
 $v = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2$
 If v is not an integer, round down

Summary: two-sample t-test (unequal variance)

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$	Ref.: t-dist with
Test statistic: $T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	v dof (10-14)
Alternative Hypotheses	P-Value
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probability above $ t_0 $ and probability below $- t_0 $
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probability above t_0
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probability below t_0
	Rejection Criterion for Fixed-Level Tests
	$t_0 > t_{\alpha/2, n_1+n_2-2}$ OR $t_0 < -t_{\alpha/2, n_1+n_2-2}$
	$t_0 > t_{\alpha, n_1+n_2-2}$ $t_0 < -t_{\alpha, n_1+n_2-2}$

Note: replace n_1+n_2 with v in rejection criteria!

Checking in Minitab: stat -> basic stat ->
 2 sample t -> input columns and in options deselect assume equal variance

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

If v is not an integer, round down to the nearest integer.

Assumptions

- (1) $X_{11}, X_{12}, \dots, X_{1n_1}$ is a random sample from population 1.
 - (2) $X_{21}, X_{22}, \dots, X_{2n_2}$ is a random sample from population 2.
 - (3) The two populations represented by X_1 and X_2 are independent.
 - (4) Both populations are normal.
- (5) The variances of the populations are unequal

If you're in doubt, assume unequal

How do I know which t-test to use?

- 1) Are measurements being made on the exact same object? (Paired)
- 2) If the ratio of the larger variance to the smaller variance is less than 4 then we can assume the variances are approximately equal (use pooled t-test)
 OR you could run a two-variance test (will learn about in the future!)

- Remember variance is standard deviation squared!!
- This rule of thumb can vary widely

↳ Use this for exam!

Confidence Intervals

Equal Variance Two-sided

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Unequal Variance Two-sided

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Paired t-test

A special case of the two-sample t-tests when observations of the two populations are observed in pairs where each pair of observations is taken under homogenous conditions

To compare differences, we define D_j :

$$D_j = X_{1j} - X_{2j}, j = 1, 2, \dots, n. \quad n = \text{# of observations}$$

where D_j 's are assumed to be normally distributed with mean μ_D and variance σ_D^2

n is the number of paired observations (j is the paired observation number)

1 and 2 represents the paired data we are comparing

Confidence Interval

$$H_0: \mu_D = \Delta_0$$

$$\bar{d} - t_{\alpha/2, n-1} s_D / \sqrt{n} \leq \mu_D \leq \bar{d} + t_{\alpha/2, n-1} s_D / \sqrt{n}$$

$$H_a: \mu_D \neq \Delta_0$$

$$\text{Test Statistic: } T_0 = \frac{\bar{D} - \Delta_0}{S_D / \sqrt{n}}$$

where $\bar{D} = \text{sample average of } n \text{ differences}$

$S_D = \text{sample standard deviation of } n \text{ differences}$

Reference Dist: t-distr with $n-1$ degrees of freedom

Rejection Criteria: Same as one sample t-test

Summary

Null hypothesis: $H_0: \mu_D = \Delta_0$

Ref: t-dist w/ $n-1$ dof

$$\text{Test statistic: } T_0 = \frac{\bar{D} - \Delta_0}{S_D / \sqrt{n}} \quad (10-24)$$

Alternative Hypotheses

P-Value

Rejection Criterion for Fixed-Level Tests

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu_D \neq \Delta_0$	Probability above $ t_0 $ and probability below $- t_0 $	$t_0 > t_{\alpha/2, n-1}$ or $t_0 < -t_{\alpha/2, n-1}$
$H_1: \mu_D > \Delta_0$	Probability above t_0	$t_0 > t_{\alpha, n-1}$
$H_1: \mu_D < \Delta_0$	Probability below t_0	$t_0 < -t_{\alpha, n-1}$

Assumptions:

- (1) Differences are continuous
- (2) Differences are normally distributed
- (3) Differences are obtained from a random sample
- (4) Differences are obtained from homogenous pairs

Minitab: stat -> basic stat-> paired t

or input differences data into one sample t-test

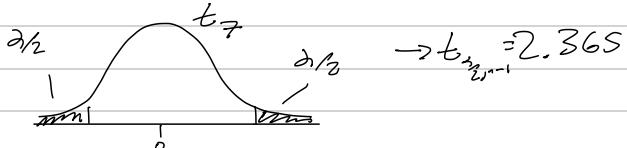
Example

TABLE 4.3 Paired Tensile Strength Data for Example 4.11 (n=8)			
Specimen	Machine 1	Machine 2	Difference
1	74	78	-4
2	76	79	-3
3	74	75	-1
4	69	66	3
5	58	63	-5
6	71	70	1
7	66	66	0
8	65	67	-2

Got \bar{D}
and s_D
Frontier
of Differ.

Do these machines yield the same tensile strength values on the same sample?

- (i) In testing the difference in mean tensile strength measured by two devices, same sample \rightarrow paired diff \Rightarrow paired t-test
- (ii) $H_0: \mu_D = 0$ (iii) $\mu_D \neq 0$
- (iv) $t_0 = \frac{\bar{D}}{s_D / \sqrt{n}} = \frac{-1.38}{2.67 / \sqrt{8}} = -1.46$
- (v) Reject $H_0: t_0 > t_{\alpha/2, n-1}$ or $t_0 < -t_{\alpha/2, n-1}$



Mean difference: -1.38

Standard deviation of differences: 2.67

Alpha = 0.05

Conclusion: F: I to reject null (H_0), we do not have sufficient evidence to reject the claim that the machines measure differently

End of Exam 1
Motivation

Exam 1 Summary

Which Hypotheses do I use?

Test	# of Populations	Variance Known?	Parameter of interest?	Reference Dist.
One-sample Z-test	1	Yes	Mean	Standard normal
Two-sample Z-test	2	Yes	Difference between means	Standard normal
One-sample t-test	1	No	Mean	t-distribution
Two-sample (pooled) t-test (equal variance)	2	No	Means of two populations	t-distribution
Two-sample t-test (unequal variance)	2	No	Means of two populations	t-distribution
Paired t-test	2 (simplifies to 1)	No	Difference between paired means	t-distribution
1 Variance test	1	No	Variance of one population	chi-squared distribution

How to Represent your data

- 1) Data are often represented as the mean \pm the standard deviation and you should tell people how many repeats are in your sample (n) (write this in your caption as "Data are represented as the mean \pm the standard deviation with $n=X$ ".)
- 2) Statistical significance (when the null is rejected) is represented by a symbol (usually $*$), accompanied by the alpha level, and statistical test with appropriate detail. E.g., ** represents p-value < 0.05 compared to X , using a X-sided, X test"
- 1) Describe the test! E.g. two-sided or one sided? One sample or two sample? T=t-test or z=z-test? If t-test, is it assuming equal variance, if it is a z-test what is the variance?

Test 1

- Open book open note and class materials – you can have a calculator, but no access to internet (PdL notes are ok!)
- BRING YOUR STATISTICAL TABLES!!! (They will not be provided for the exam, but you will need them)
- You will be asked to sketch out the distributions and label things
- How can I do well?
 - Make sure you know how to do the homework – review the correct answers if something went wrong
 - Do the practice exam (to be posted later)
- What is going to be on the exam?
 - Anything in lectures 1-10 and HW1-5 is fair game, and the types of problems you see in the practice exam
 - Note: the test can't cover every single type of problem and you won't see the exact same problems in the practice exam on Test 1!

You can check hypothesis tests with your calculator just know what you're doing, and show all your work!

Topics

(Not necessarily comprehensive but good things to study)

- vocabulary and concepts regarding statistical process control, six sigma, lean manufacturing, and the DMAIC process
- tools to represent data, and check assumptions (e.g., normality)
- probability problems, properties of probability distributions
- correctly performing a hypothesis testing procedure, picking the correct test and alternative hypothesis (see list of tests)
- how to correctly phrase conclusions in a hypothesis test
- how to correctly phrase what the p-value is
- how to correctly phrase what confidence intervals are representing, and calculate them
- how to sketch hypothesis tests on a reference (sampling) distribution
- concepts surrounding what type I and type II error are, and why they allow us to make conclusions in a hypothesis test
- concepts around the central limit theorem, and sampling distributions vs. population distributions

Present conclusions

ECHE 313 Practice Exam 1

Solutions on canvas

Write your name:

Write the following statement, and sign your name acknowledging your agreement:

"I neither received nor gave external assistance on this exam"

Don't pay attention to this question

The exam layout is such that all problems add up to 100 points. This point total is meant to guide you in time allocation for this test, will be adjusted such that this exam is 25% of your total class grade. You may use your class materials, class notes, homework, book, and Minitab or Excel to aid in the completion of the exam. The exam is designed to be completed in 1hr and 15 min, but you will have between 10 AM 2/27 to 10 AM 2/28 to complete the exam. All of the necessary tables are provided in the back of your book. Please show your work and all steps necessary to arrive at your answer. Note that it is expected you do this entire test by hand, using a calculator and the tables found in the back of your book. So while you are allowed to use Minitab if you wish to check your answers, it should not replace the work you have done by hand using the tables for your answers. Where applicable, it is best to write out the general formulas you want to use first, then write in your substituted values so it is clear. In addition, for this exam, we can use the rule of thumb that if the sample variance ratio (larger to smaller) is lower than 4, you can assume equal variance for a t-test (if needed). Also, when reporting numbers, keep the same number of decimals as the data you are working with unless there are specific directions on how many sig figs to report.

Good luck!

Premise: You are a process engineer working for a lean six-sigma pharmaceutical company that makes advanced protein-based biologics. Unless otherwise stated, your company uses $\alpha=0.01$.

- 1) (12 points) You are finishing up a project where you focused on increasing process cycle efficiency and are ready to hand over the improved process.
 - a. (2 points) What step of the DMAIC process are you in?

Control Stage

- b. (2 points) Name at least one tool that is used in this stage.

Transition Plan Control Chart

- c. (4 points) Explain what process cycle efficiency is and how it contributes to lean manufacturing.

$$PCB = \frac{\text{Value added time}}{\text{Process Time}}$$

See Key

- d. (4 points) Explain why a slow process is an expensive process with at least 3 specific reasons.

See Key

- 2) (39 points) Your next project is to determine if the bioreactor time can be reduced by getting a newer bioreactor system. You are able to demo a new unit and track the reactor time on 5 randomly selected batches in both the old and new bioreactors. You obtain the following data:

The following table shows the time it took in minutes for the batch to reach the designated amount of bacteria:

$$\bar{x}_1 = 262, s_1 = \sqrt{132}$$

$$\bar{x}_2 = 149, s_2 = \sqrt{94}$$

Old Bioreactor Time (min)	New Bioreactor Time (min)
245	165
270	141
262	151
275	142
260	147

- a. (3 points) List the name of the hypothesis test you would use

Two sample t-test (pooled) assuming equal variance
 Use rule of thumb

- b. (3 points) Name the parameter of interest in the context of the problem statement

Mean Bioreactor time of the old and new (say time difference)

- c. (2 points) Write out the correct null hypothesis for this test

$$H_0: \mu_1 - \mu_2 = 0$$

- d. (2 points) Write out the correct alternative hypothesis for this test

$H_a: \mu_1 - \mu_2 \neq 0$ two-sided Can be interpreted as one-sided

- e. (7 points) Write the formula of the test statistic and calculate (report 3 sig figs)

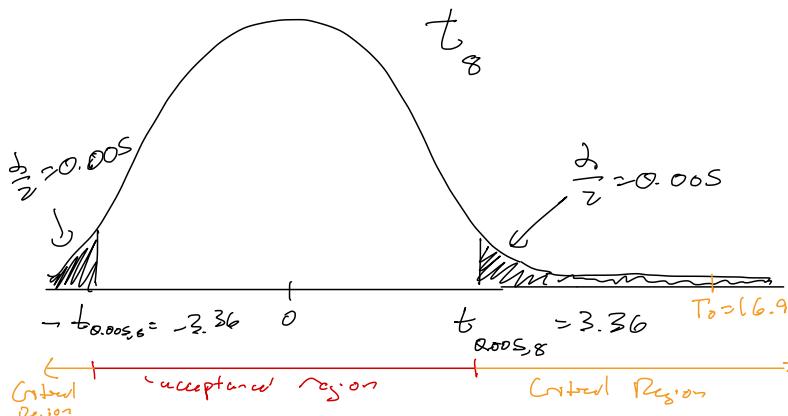
$$T_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{262 - 149}{10.62 \sqrt{\frac{1}{5} + \frac{1}{5}}} = 16.9$$

\uparrow
 $S_p = 10.62$

- f. (4 points) Write out the correct rejection criteria for fixed significance level testing and the critical value (3 sig figs)

$$\text{Reject } H_0 : \begin{cases} t_0 > t_{\alpha/2, n_1 + n_2 - 2} \\ t_0 < t_{\alpha/2, n_1 + n_2 - 2} \end{cases} \quad t_{0.005, 8} = \pm 3.36$$

- g. (9 points) Sketch out your reference distribution making sure to fully define the distribution you are sketching on with the correct label, drawing it with the correct shape and where zero is. In addition, label 1) acceptance region, 2) the rejection region (also known as the critical region), 3) α , 4) the critical value and 5) the test statistic.



- h. (3 points) State your conclusion and remember to word your conclusion appropriately in the context of the problem statement.

$$T_0 < t_{0.005, 8}$$

H_0 Reject H_0 , Accept H_A , Context \rightarrow Morous news lower!

See key

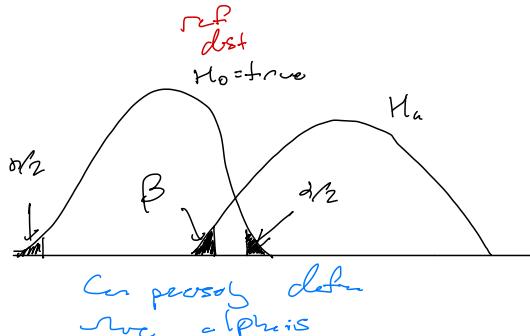
- i. (3 points) Explain what plot you would construct to check if the new bioreactor data are normally distributed (what feature(s) of the plot would you check?)

Normal probability plot, see F:ts linear (looks distinct non-linear shape)

- j. (3 points) Define in words what (α) is (what does it represent, what is its definition?) and explain why it helps us make strong conclusions.

α is Type I error =

See Key



- 3) (32 points) To determine if the new bioreactor is suitable for the process, the team also needs to know if the new bio reactor produces a yield that is greater than 810 mg/L of product. Twelve random samples are taken from batches in the piolet run:

Yield (mg/L)
807
825
816
805
815
820
822
817
823
817
810
812

$n \approx 12$

Sample Ave. 815.75

Sample Standard Deviation: 6.31

- a. (2 points) List the *name* of the hypothesis test you would use

One sample t-test

- b. (3 points) Name the parameter of interest in the context of the problem statement

Mean process yield

- c. (2 points) Write out the correct null hypothesis for this test

$H_0: \mu = 810 \text{ mg/L}$

- d. (2 points) Write out the correct alternative hypothesis for this test

$$H_a: \mu > \mu_0 = 810 \text{ mg/L}$$

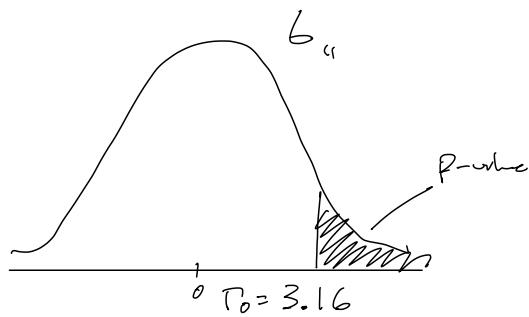
- e. (3 points) Write the formula of the test statistic and calculate (report 3 sig figs)

$$T_0 = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{815.78 - 810}{\frac{6.31}{\sqrt{12}}} = 3.16$$

- f. (2 points) Write out the correct rejection criteria using the p-value method

Reject if $p\text{-value} < \alpha = 0.01$

- g. (6 points) Sketch out your reference distribution making sure to fully define the distribution you are sketching on with the correct label, drawing it with the correct shape and where zero is. In addition, label 1) the test statistic and 2) the p-value.



- h. (2 points) Find the p-value using the tables in the back of your book

- i. (2 points) Explain what the p-value is in words (i.e. what is the definition of a p-value?)

See
Very

- j. (3 points) State your conclusion and remember to word your conclusion appropriately in the context of the problem statement.

Reject

- k. (2 points) You calculate your 99% confidence interval and show it to your boss. Explain what the 99% confidence interval really represents generally, in words (Hint: do not just say it means you are 99% confident that the mean lies in the CI range). You do not need to actually calculate the CI to answer this question correctly.

- l. (3 points) If the manufacturer told you the standard deviation on the yield coming from these new units is known to be 4.00 mg/L, and you wanted to know if the yield was *different* than 810 mg/L, name the hypothesis test you would conduct, and if your alternative hypothesis would be 1 or 2 sided.

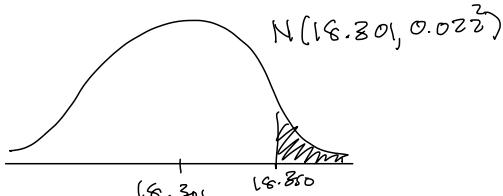
One-sample z-test, and z-sid test due to df P-value

- 4) (17 points) Your boss was so impressed with the projects you have been doing that she asked you to help the manufacturing floor with some questions regarding their packaging unit, which boxes up the biologics with instructional materials. Weight measurements are taken on each and every product automatically, thus it is known that the population mean μ is 18.301 g and the population standard deviation is σ is 0.022 g and is normally distributed. New weight restrictions are being instated and the manufacturing floor wants to know the probability that a product will have a weight above a new limit of 18.350 g.

*Probability
problem!
Not hypothesis
Test!*

- a. (11 points) Write out the probability equation you wish to solve, including the integral (but you do not need to solve the integral just set it up). Also include a probability distribution sketch with the 1) probability you wish to find shaded in 2) the correct shape, 3) correct label of the distribution, 4) where 18.350g is and 5) where zero is.

$$P(x > 18.350) = \int_{18.301}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2} dx$$



- b. (6 points) Solve the probability problem using the *tables in the back of the book*. Just like in class, sketches can help solve these problems, especially when using the tables. Report an appropriate number of sig figs based on the numbers given in the problem statement.

$$Z = \frac{16.380 - 16.38}{0.022} = 223$$

$$\rightarrow P-value \approx 0.013$$

Midterm Exam 2



Chapter 4 Continued - Test 2 Material

Inference on Two populations' variances

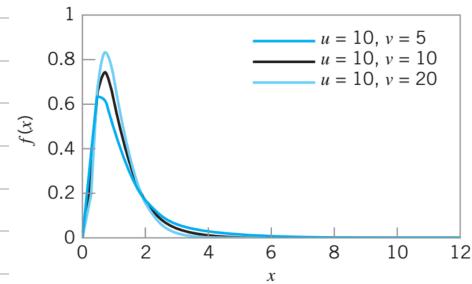
→ Two independent, normally distributed populations

$$\text{Null: } H_0: \sigma_1^2 = \sigma_2^2$$

$$\text{Alt: } H_a: \sigma_1^2 \neq \sigma_2^2 \text{ tested}$$

$$\text{Test Stat: } F_0 = \frac{S_1^2}{S_2^2} \leftarrow \text{DoF}_1 = n_1 - 1$$

Sensitive to normality – should check!



Reference distribution: F-distribution

(i) The curve is not symmetric (skewed right)

(ii) DF quantiles for each set of DoFs

(iii) F statistic ≥ 0

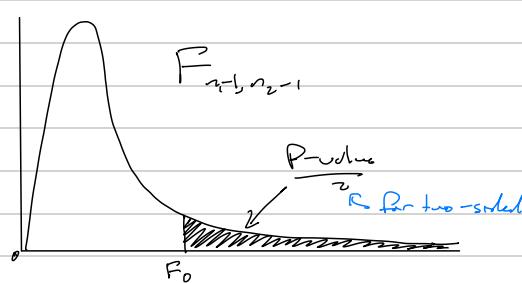
(iv) As DoF increase curve gets more narrow

$$\text{Note: } F_{0.99, v_1, v_2} = 1 / F_{0.01, v_2, v_1}$$

Rejection Criterion: (two-sided)

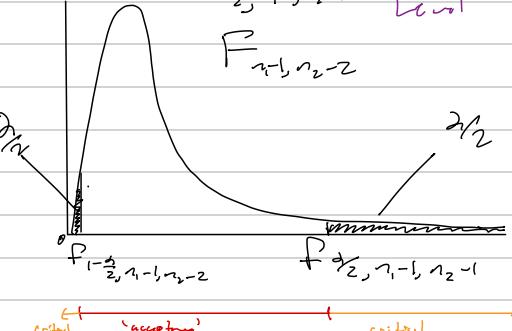
$F_{\alpha/2}$

Practical < 2



$$F_0 \geq F_{\alpha/2, n_1-1, n_2-1}$$

$$\text{or } F_0 \leq F_{1-\alpha/2, n_1-1, n_2-1}$$



Fixed
Sig. F.
Level

Summary

Null hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

$$\text{Test statistic: } F_0 = \frac{S_1^2}{S_2^2}$$

(10-31)

Alternative Hypotheses

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

$$H_1: \sigma_1^2 < \sigma_2^2$$

Rejection Criterion

$$f_0 > f_{\alpha/2, n_1-1, n_2-1} \text{ or } f_0 < f_{1-\alpha/2, n_1-1, n_2-1}$$

$$f_0 > f_{\alpha, n_1-1, n_2-1}$$

$$f_0 < f_{1-\alpha, n_1-1, n_2-1}$$

Assumptions

The data are normally distributed

Samples are independent

Sensitive to normality!

Example

Example 10-13 Semiconductor Etch Variability Oxide layers on semiconductor wafers are etched in a mixture of gases to achieve the proper thickness. The variability in the thickness of these oxide layers is a critical characteristic of the wafer, and low variability is desirable for subsequent processing steps. Two different mixtures of gases are being studied to determine whether one is superior in reducing the variability of the oxide thickness. Sixteen wafers are etched in each gas. The sample standard deviations of oxide thickness are $s_1 = 1.96$ angstroms and $s_2 = 2.13$ angstroms, respectively. Is there any evidence to indicate that either gas is preferable? Use a fixed-level test with $\alpha = 0.05$.

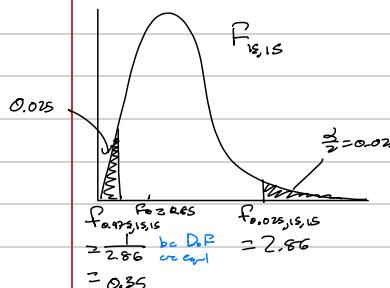
1) Parameters of interest: Variances oxide thickness for two mixtures of gases, unknown variances

∴ Two-variance test

$$2) H_0: \sigma_1^2 = \sigma_2^2 \quad 4) F_0 = \frac{s_1^2}{s_2^2} = \frac{1.96^2}{2.13^2} = 0.85$$

$$3) H_a: \sigma_1^2 \neq \sigma_2^2$$

$$5) \text{Reject } H_0: F_0 > F_{0.025, 15, 15} \text{ or } F_0 < F_{0.975, 15, 15}$$



Fail to reject H_0 , we don't have enough evidence to reject the claim the variances of the oxide layer produced by the different wafers are different.

Single Factor Analysis of Variance

• Analysis of Variance: Used for comparing means when there are more than two levels in a single factor

→ Factor = Entity being tested

→ Level = Settings within a factor

• Replicates: # of repeats taken at each level

• Randomization: Randomly order data such that nuisance variables are eliminated

Factor levels	Tensile Strength of Paper (psi)						Totals	Averages
	Observations							
Hardwood Concentration (%)	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

of R...

Factors = 1

Levels = 4

Replicates ≥ 6

Defining and Development of ANOVA

Typical Data for a Single-Factor Experiment

Treatment	Observations						Totals	Averages
1	y_{11}	y_{12}	...	y_{1n}	$\bar{y}_{1\cdot}$			
2	y_{21}	y_{22}	...	y_{2n}	$\bar{y}_{2\cdot}$			
...		
...		
a	y_{a1}	y_{a2}	...	y_{an}	$\bar{y}_{a\cdot}$			
					$\bar{y}_{\cdot\cdot}$			

$a = \# \text{ of treatments}$ (y_{ij}^{obs} vs $\bar{y}_{ij}^{\text{mean}}$)

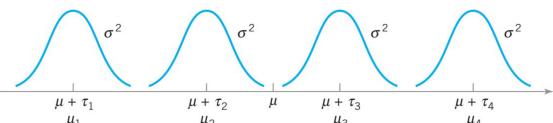
$y_{ij} = j^{\text{th}} \text{ observation of } i^{\text{th}} \text{ treatment}$

$n = \# \text{ of observations at each level}$

Linear Statistical Model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \begin{cases} i=1,2,\dots,a \\ j=1,2,\dots,n \end{cases}$$

where $\tau_i = i^{\text{th}} \text{ treatment effect}$
 $\varepsilon_{ij} = \text{random error}$



Hypothesis Testing for Equality of Treatment Effects

Null: $H_0: \tau_1 = \tau_2 = \tau_3 = \dots = \tau_a = 0$

Alt: $H_a: \tau_i \neq 0 \text{ for at least one } i$

Test Statistic: $F_o = \frac{SS_{\text{treatments}}/a}{SS_E/(n-a)}$ where $a = \# \text{ of treatments}$
 $n = \# \text{ of replicates per each treatment level}$

Refined Distribution: F-distribution with $(a-1)$ [top] and $a(n-1)$ [bottom] DoF

Rejection Criteria: $H_0: F_o > F_{\alpha/2, a-1, a(n-1)}$ This is always one-sided

Calculations for ANOVA

The Analysis of Variance for a Single-Factor Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_o
Treatments	$SS_{\text{Treatments}}$	$a - 1$	$MS_{\text{Treatments}}$	$\frac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n - 1)$	MS_E	
Total	SS_T	$an - 1$		

$$SS_T = SS_{\text{Treatments}} + SS_E$$

$$F_o = \frac{MS_{\text{treat}}}{MS_E} = \frac{SS_{\text{treat}}/a}{SS_E/(n-a)}$$

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \text{total sum of squares}$$

$$SS_{\text{Treatments}} = n \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \text{treatment sum of squares}$$

$n \times a$ squared terms

$a : \# \text{ of squared terms}$

$$\bar{y}_i = \frac{1}{n} \sum_{j=1}^n y_{ij} \quad \bar{y}_{i\cdot} = y_i/n \quad i = 1, 2, \dots, a$$

$$\bar{y}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^n y_{ij} \quad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N$$

Deriving & Development of ANOVA

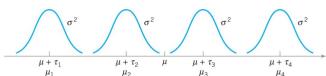
a = number of treatments

y_{ij} = the jth observation of the ith treatment

n = number of observations for each level

Linear statistical model: $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ $\begin{cases} i=1,2,\dots,a \\ j=1,2,\dots,n \end{cases}$

Where: y_{ij} = observation
 μ = overall mean (common to all treatments)
 τ_i = the ith treatment effect
 ϵ_{ij} = random error



Hypotheses testing the equality of treatment effects

Null and Alternative Hypothesis:

$$H_0: \tau_1 = \tau_2 = \dots = \tau_a = 0 \quad H_1: \tau_i \neq 0 \text{ for at least one } i$$

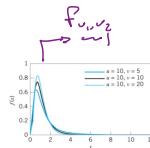
Test Statistic: (reference distribution is the F-distribution)

$$F_0 = \frac{SS_{\text{Treatments}} / (a-1)}{SS_E / [a(n-1)]} = \frac{MS_{\text{Treatments}}}{MS_E}$$

DOF

$$f_0 > f_{0.05, a-1, a(n-1)}$$

Sign. 1
n.s.



Calculations for ANOVA (same n)

The Analysis of Variance for a Single-Factor Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F ₀
Treatments	SS _{Treatments}	a - 1	MS _{Treatments}	$\frac{MS_{\text{Treatments}}}{MS_E}$
Error	SS _E	a(n - 1)	MS _E	
Total	SS _T	an - 1		

$$SS_T = SS_{\text{Treatments}} + SS_E$$

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 \text{ total sum of squares}$$

$$SS_{\text{Treatments}} = n \sum_{i=1}^a (\bar{y}_i - \bar{y}_{..})^2 \text{ treatment sum of squares}$$

$$\bar{y}_i = \sum_{j=1}^n y_{ij} \quad \bar{y}_{..} = \bar{y}_i / n \quad i = 1, 2, \dots, a$$

$$\bar{y}_{..} = \sum_{i=1}^a \sum_{j=1}^n y_{ij} / n \quad \text{Total number of observations}$$

General

i = treatment

j = observation

b*nr* = a var sc

dots = total

✓ = replicates

a = # of treatments

y_{ij} = jth observation of ith treatment

\bar{y}_i = grand mean of all observations

$\bar{y}_{..}$ = total of all observations

\bar{y}_{ij} = average of observations @ treatment i

What does the calculation look like?

Tensile Strength of Paper (psi)

Factor	Observations						$\bar{y}_{..}$	\bar{y}_{ij}	$SS_T = SS_{\text{Treatments}} + SS_E$
	1	2	3	4	5	6			
1	5	7	8	15	11	9	10	60	10.00
2	10	12	17	13	18	19	15	94	15.67
3	15	14	18	19	17	16	18	102	17.00
4	20	19	25	22	23	18	20	127	21.17
							383	15.96	

Treatment Level
a=4

$$SS_T = SS_{\text{Treatments}} + SS_E$$

$$SS_E = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = (7-15.96)^2 + (8-15.96)^2 + \dots + (20-15.96)^2 = 513$$

$$SS_{\text{Treatments}} = n \sum_{i=1}^a (\bar{y}_i - \bar{y}_{..})^2 = 6[(10-15.96)^2 + (15.67-15.96)^2 + (17-15.96)^2 + (21.17-15.96)^2] = 382.8$$

$$\Rightarrow SS_B = 513 - 382.8 = 130.2$$

Hypothesis Test? Does hardwood concentration impact the tensile strength of paper? $\alpha=0.01$

1) Parameter of interest: Multiple pop. means of tensile strength \rightarrow ANOVA!

2) $H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$

3) $\exists i \in \{1, 2, 3, 4\}$ s.t. $\tau_i \neq 0$

$$4) F_0 = \frac{\frac{SS_{\text{Treatments}}}{(a-1)}}{\frac{SS_E}{(a-1)}} = \frac{\frac{130.2}{3}}{\frac{382.8}{20}} = 19.6$$

5) Reject: if $F_0 > F_{0.01, 3, 20}$

$$F_{0.01, 3, 20} \approx 4.94 \text{ from charts}$$



Minitab: Stat -> ANOVA -> one-way

⇒ Conclusions: Reject H_0 as $F_0 > F_{0.01, 3, 20}$

At least one treatment has an impact on the tensile strength of paper. More generally,

Hardwood Concentration impacts paper tensile strength!

What do you do when the null is rejected?

Multiple Comparison Methods - (post-hoc tests) is performed to find which means are different after ANOVA has been performed and H_0 is rejected.

Fisher's Least Significant Difference (LSD) Test:

$$\text{Declare Significantly Different if: } |\bar{y}_{ij} - \bar{y}_{jk}| > LSD$$

$$\rightarrow LSD = t_{\alpha/2, n(n-1)} \sqrt{\frac{2MSE}{n}}$$

Example

Observations								
Hardwood Concentration (%)	1	2	3	4	5	6	Totals	Averages
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

Minitab: Stat -> ANOVA -> one-way (click the comparisons button – error rate)

$$S_{us.10} : |10 - 15.67| = 5.67^*$$

$$S_{us.15} : |10 - 17| = 7^*$$

$$S_{us.20} : |10 - 21.17| = 11.17^*$$

$$10 \text{ vs } 15 : |10 - 17| = 7$$

$$10 \text{ vs. } 20 : |10 - 21.17| = 11.17$$

$$15 \text{ vs. } 20 : |15 - 21.17| = 6.17$$

$$LSD = t_{0.005, 20} \sqrt{\frac{2MSE}{n}}$$

$$= 2.845 \sqrt{2 \left(\frac{6.5}{6} \right)}$$

$$= 4.19$$

20 : C ^{10 vs 20 is equal}
 15 : C ^{10 vs 15 is equal}
 10 : B ^{10 vs 20 is equal}
 S : A ^{Different}

Assumptions

- (i) Normal
- (ii) No hidden variables

Summary

Declare significantly different if: $|\bar{y}_{ij} - \bar{y}_{jk}| > LSD$

where LSD, the least significant difference, is

$$LSD = t_{\alpha/2, n(n-1)} \sqrt{\frac{2MS_E}{n}}$$

Better notes on LSD test:

- Do this ONLY after ANOVA null hypothesis is rejected
- Find differences between all of the treatments
- Compare differences to the LSD to see if they are significantly different

Mini tab might produce an output like this:
 If a pair has the same letter, they are not significantly different

Factor	N	Mean	Grouping
20%	6	21.17	A
15%	6	17.00	A B
10%	6	15.67	B
5%	6	10.00	C

Fisher's LSD Test

Declare significantly different if: $|\bar{y}_i - \bar{y}_j| > LSD$
where LSD, the least significant difference, is

$$LSD = t_{\alpha/2,a(n-1)} \sqrt{\frac{2MS_E}{n}}$$

Better notes on LSD test:

- Do this after ANOVA null hypothesis is rejected
- Find differences between all of the treatments
- Compare differences to the LSD to see if they are significantly different

Minitab might produce an output like this:

If a pair has the same letter, they are not significantly different

Sometimes funny overlap occurs – for example, 15% is not different than 20% or 10% but, 10% and 20% are different

Factor	N	Mean	Grouping
20%	6	21.17 A	
15%	6	17.000 A B	
10%	6	15.67 B	
5%	6	10.00 C	

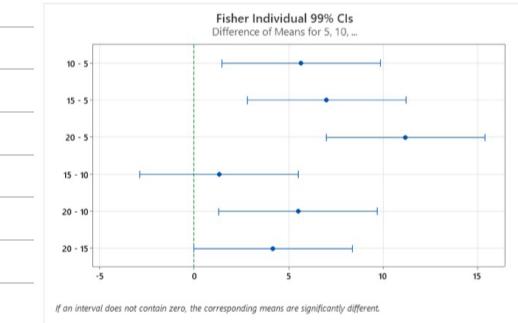
Confidence Intervals

A $100(1 - \alpha)\%$ confidence interval on the mean of the i th treatment μ_i is

$$\bar{y}_{i \cdot} - t_{\alpha/2,a(n-1)} \sqrt{\frac{MS_E}{n}} \leq \mu_i \leq \bar{y}_{i \cdot} + t_{\alpha/2,a(n-1)} \sqrt{\frac{MS_E}{n}}$$

A $100(1 - \alpha)$ percent confidence interval on the difference in two treatment means $\mu_i - \mu_j$ is

$$\bar{y}_{i \cdot} - \bar{y}_{j \cdot} - t_{\alpha/2,a(n-1)} \sqrt{\frac{2MS_E}{n}} \leq \mu_i - \mu_j \leq \bar{y}_{i \cdot} - \bar{y}_{j \cdot} + t_{\alpha/2,a(n-1)} \sqrt{\frac{2MS_E}{n}}$$



ANOVA Assumptions

- (i) Errors (and observations) are normally distributed
- (ii) Each treatment has the same variance (equal variance assumption)
- (iii) Errors and observations are independently distributed

Residual

The difference between an observation (y_{ij}) and its estimated (or fitted) value from the statistical model being studied

Normal Probability Plot of Residuals

Minitab

Can ask Minitab to plot a normal probability plot after pasting in residuals into one column, or use ANOVA function, and select graphs, then select – normal probability plot of residuals

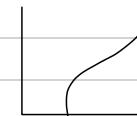
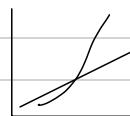
(i) Rank data

(ii) Calc cumulative Frequency

(iii) Find Z_{ij} 's

(iv) Plot residuals vs. Z_{ij} 's

Looking for straight line



Checking equal Variance at each Factor Level

→ Minitab does not automatically generate

Look for: even residuals, if not...

Rule of thumb we used for t-tests applies, or, you could check using your 2 variance test!

Can generate in Excel or Minitab: Residuals in one column, factor level in another, using scatter plot

Checking Equal Variance as a function of predicted value

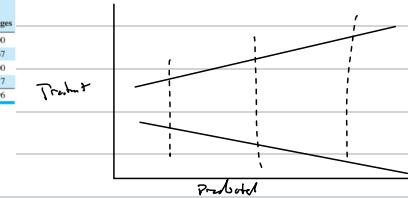
Residuals plotted as a function of predicted value

Minitab: use ANOVA function, and select graphs, then select - residuals vs. fits

Can also generate in Excel or Minitab: Residuals in one column, predicted value in the other, using scatter plot

What if there is a pattern?

- Data transformations
- Residual Analysis



Checking Independence

No formal plot... but:

- 1) Addressed via experimental design (randomize!)
- 2) Can plot residuals versus any suspected or nuisance variables
- 1) Example: plot residuals versus run order (run order, temp, pH...)

When is my data not independent?

- (1) repeated measurements are taken on the same subject → paired
- (2) residuals or observations are correlated with a nuisance variable

Simple Linear Regression

- Use simple linear regression for building an empirical model
- Understand how the method of least squares is used to estimate the parameters in a linear regression model
- Analyze residuals to determine whether the regression model is an adequate fit to the data or whether assumptions are violated
- Test statistical hypotheses and construct confidence intervals on regression model parameters
- Apply the correlation model
- Use simple transformations to achieve a linear regression model

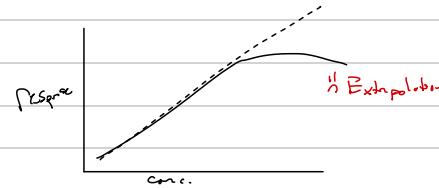
Regression Modeling

- **Regression Modeling:** A mathematical model fit to a set of data relating a response variable (y) to an independent variable (x_1, x_2, \dots, x_n)
- **Empirical Models:** Models where the structure is determined by the observed relationship among experimental data, but not necessarily mechanistically relevant
- **Simple Linear Regression Model:** There is only one indep. or predictor variable (x) and the relationship with response (y) is assumed to be linear

$$y = \beta_0 + \beta_1 x + \epsilon$$

y = expected response
 β_0 = intercept
 β_1 = slope

x = predictor variable
 ϵ = random error term

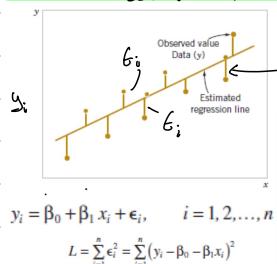


- **Pitfalls:** Empirical Models

→ Correlation does not equal causation!!!

→ Valid in the range of original data (be careful with extrapolation)

- How to estimate the coefficients in simple linear regression



Method of Least Squares
 Minimize the sum of squares of vertical deviations

The least squares estimates of the intercept and slope in the simple linear regression model are

$$\text{Avg of all response values} \rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(\sum_{i=1}^n x_i)}{\sum_{i=1}^n x_i^2} S_{xy} \quad (11-8)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

Fitted or Estimated Regression Line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Predicted Value

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$$

Each Ldt. Each Pair Satisfies this

Example: Fit a simple linear regression line to these data.

TABLE • 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level $x(\%)$	Purity $y(\%)$
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

$$n=20$$

$$\sum x_i = 23.42 \quad \sum y_i = 1843.21$$

$$\sum x_i^2 = 24.28 \quad \sum y_i^2 = 170044$$

The least squares estimates of the intercept and slope in the simple linear regression model are

$$\hat{y}_0 = \bar{y} - \hat{B}_1 \bar{x} \quad (11-7)$$

$$\hat{B}_1 = \frac{\sum_{i=1}^n y_i - \left(\frac{\sum_{i=1}^n y_i}{n} \right) \left(\frac{\sum_{i=1}^n x_i}{n} \right)}{\sum_{i=1}^n x_i^2 - \left(\frac{\sum_{i=1}^n x_i}{n} \right)^2} \quad (11-8)$$

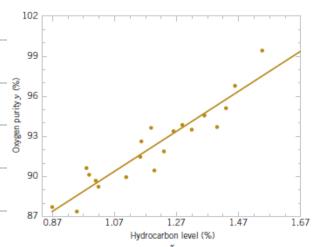
where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

$$S_{xx} = 24.28 - \frac{(23.42)^2}{20} = 0.68$$

$$S_{xy} = 2214 - \frac{23.42(1843.21)}{20} = 10.18$$

$$\tilde{B}_1 = \frac{10.18}{0.68} = 14.95$$

$$\tilde{B}_0 = \bar{y} - \tilde{B}_1 \bar{x} = 92.1605 - (14.95)(1.196) \\ = 74.28$$



$$\hat{y} = 74.28 + 14.95x$$

Minitab: Stat > Regression -> Regression -> Fit Regression Model

Minitab: Stat > Regression -> Fitted Line -> choose x, y and linear

ANOVA for Linear Regression

- Null & Alternative Hypotheses:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

- Test Statistic

$$F_0 = \frac{SS_R / 1}{SS_E / n-2} = \frac{MS_R}{MS_E}$$

Signal
noise

- Refined Distribution: F-distr

↳ Want to use: $F_{1, n-2}$

- Rejection Criteria:

$$\text{Reject } F_0 \quad F_0 > F_{\alpha, 1, n-2}$$

or $p\text{-value} < \alpha$

Calculating the Error

$$\text{Result: } S_{ST} = SS_R + SS_E$$

$$\bullet \text{Residual: } e_i = y_i - \hat{y}_i$$

Actual vs predicted

$$\bullet \text{Error Sum of Squares: } SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\bullet \text{Mean Squared Error: } SS_E / n-2$$

$$\bullet \text{Sum of Squares Total: } SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\bullet \text{Regression Sum of Squares: } SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Minitab: Stat>Regression>Fit Regression Model (options sets the alpha)

Example

TABLE 11-1 Oxygen and Hydrocarbon Levels		
Observation Number	Hydrocarbon Level x_i (%)	Purity y_i (%)
1	1.02	90.45
2	1.15	91.43
3	1.29	93.74
4	1.46	96.78
5	1.36	94.45
6	0.87	87.59
7	1.23	91.77
8	1.55	99.42
9	1.40	93.65
10	1.19	93.54
11	1.15	92.52
12	0.98	90.56
13	1.01	93.42
14	1.11	89.35
15	1.20	90.39
16	1.26	93.25
17	1.32	93.41
18	1.43	94.98
19	0.95	87.33

1) Parameter of Interest: Slope of Linear Model; one regression \rightarrow simple linear regression

2) $H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

3) $F_0 = SSe_{\text{R}} / 1$

$SSe_{\text{E}} / n-2$

4) Reject if $F_0 > F_{0.05, 1, 18}$

$\bar{y} = 92.16$

$\hat{y} = \beta_0 + \beta_1 x_i \Rightarrow \hat{y} = 74.28 + 14.947x$

$n=1 \Rightarrow \hat{y} = 74.28 + 14.947(0.49) = 89.08$

: all data points

$n=20 \Rightarrow \hat{y} = 74.28 + 14.947(0.48) = 88.48$

$SS_{\text{E}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (89.08 - 92.16)^2 + \dots + (88.48 - 92.16)^2 = 152.17$

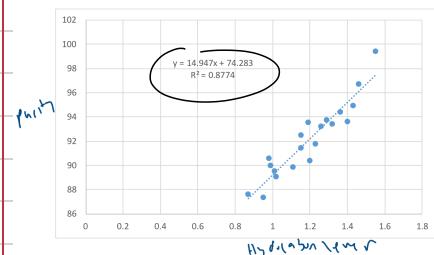
$SS_{\text{R}} = \sum_{i=1}^n (y_i - \bar{y})^2 = (90.01 - 92.16)^2 + \dots + (82.33 - 92.16)^2 = 173.38$

$SS_{\text{E}} = SS_{\text{R}} - SSe_{\text{E}} = 21.2$

$\Rightarrow F_0 = \frac{SS_{\text{R}} / 1}{SS_{\text{E}} / (n-2)} = 129 \quad \left. \begin{array}{l} 129 > 4.41 \rightarrow \text{Reject } H_0 \\ \text{There is a significant linear relationship between hydrocarbon level and purity} \end{array} \right.$

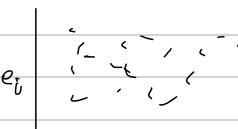
$F_{0.05, 1, 18} = 4.41 \text{ from Table}$

$R^2 = \frac{SS_{\text{R}}}{SS_{\text{T}}} = 1 - \frac{SS_{\text{E}}}{SS_{\text{T}}} = \frac{152}{173.38} = 0.88$



R^2 represents how much error is accounted for by the model

1



Checking Assumptions

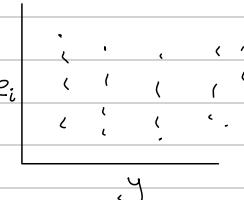
- Errors (and observations) are normally distributed
- Errors (and observations) are independently distributed
- Variance is constant

$e_i = y_i - \hat{y}_i \quad i = 1, 2, \dots, n$
residuals
constant vs.
estimated vs.

Linear?
Normality
Plot residuals
on Normal Probability Plot

Path?
Independent Variance
Plot residuals vs.
Random or
non-random variables

Path? or Spur
Constant Variance
① Plot residuals
vs. predicted (constant)
② Plot vs. independent
variable (x_i)



Confidence Intervals around Coefficients

Under the assumption that the observations are normally and independently distributed, a $100(1 - \alpha)\%$ confidence interval on the slope β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad (11-29)$$

Similarly, a $100(1 - \alpha)\%$ confidence interval on the intercept β_0 is

$$\hat{\beta}_0 - t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \quad (11-30)$$

How to write and interpret:

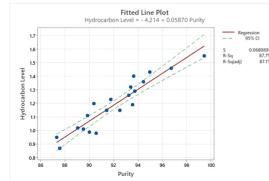
In repeated samples the intervals calculated capture the true slope or intercept (1- $\alpha\%$) of the true

Notes on CIs around Response

- CIs (confidence intervals) around a response: give a range of values associated with the true response for a given x
- You can construct confidence limits of your regression line by repeating the CIs for different values of x

How to Write?

CIs: In repeated experiments, there 95% chance that the CI on the responses calculated contain the true responses



How to generate?

To get the bounds on a graph: Minitab: Stat > Regression -> Fitted Line -> choose x, y and linear (Options tab: display CIs)

You can also use the "predict" function to have Minitab give you the CIs for any given x

Minitab: Regression -> Regression -> Fit Regression Model (select Results button)

Need to choose "expanded tables" from dropdown menu

Confidence Intervals around the Response

A $100(1 - \alpha)\%$ confidence interval on the mean response at the value of $x = x_0$, say $\mu_{Y|x_0}$, is given by

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} \leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} \quad (11-31)$$

where $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is computed from the fitted regression model.

(Nicely worth) ANOVA for Linear Regression

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \beta_j \neq 0 \text{ for at least one } j$$

Test Statistic:

$$F_0 = \frac{SS_E / 1}{SS_E / (n-2)} \frac{MS_E}{MS_E}$$

Reference Distribution:

F-distribution
Write as: $F_{1, n-2}$

Rejection Criteria: Reject H_0 if $f_0 > f_{\alpha, 1, n-2}$.

Regression on Transformed Variables

Intrinsically Linear Models: Models that can be transformed into linear equations

Transformation is useful when: Relationship isn't linear and/or when residuals show a pattern

Examples

$$Y = \beta_0 e^{\beta_1 x} \epsilon \rightarrow \ln Y = \ln \beta_0 + \beta_1 x + \ln \epsilon \rightarrow Y = \beta_0 + \beta_1 \left(\frac{1}{x}\right) + \epsilon \rightarrow Y = \beta_0 + \beta_1 z + \epsilon$$

(i) All statistical analyses should be done and interpreted on transformed variables

(ii) Back calculations can be done to report values

Example

Find a statistically satisfactory ($\alpha = 0.05$) model for:

Observation Number	Wind Velocity (m/s)	DC Output (Amp)	Observation Number	Wind Velocity (m/s)	DC Output (Amp)
1	5.00	1.562	14	5.80	1.527
2	6.00	1.522	15	5.80	1.527
3	3.40	1.027	16	3.60	1.137
4	3.00	0.989	17	3.20	1.050
5	0.00	2.236	18	6.80	2.112
6	0.70	2.384	19	7.00	1.860
7	9.00	2.768	20	5.40	1.769
8	3.05	0.958	21	9.10	2.303
9	4.15	2.168	22	9.20	2.319
10	6.20	1.866	23	4.10	1.390
11	2.90	0.603	24	3.95	1.348
12	6.85	1.498	25	3.85	1.348
13	4.60	1.562			

Final Model: $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right) + \epsilon$

Minitab

Transform: Calc > Calculator (input your function)

For Residuals: Run regression using "Fit Regression Model" - click storage button and select residuals

Scatterplot: Graph>Scatterplot> select x and y data

Steps

- Plot data out
- Fit data with a model we think could work (try linear)
- Analyze the residuals (ANOVA, R2 - plot residuals, normality)
- Transform the data if there is a problem and regress data again
- Check the results again (ANOVA, R2, residuals, normality)
- Determine if model is better
- Write out final model if: ANOVA is significant, and assumptions are met

We will rely on Minitab a lot more!

1) Scatter Plot doesn't look linear

2-3) Tryed linear \rightarrow Residuals vs Fit results is clear pattern

\hookrightarrow Equal variance assumption violated, try a diff. model

4) Try reciprocal of data

5-6) Results look good now! If they don't, just try a diff. model

7) $y = \text{DC output}$ } $y = 2.974 - 6.488 \left(\frac{1}{x}\right)$ Always specify variables

$x = \text{wind velocity}$

Conclusions: The transformed model for DC output and wind velocity is significant.

Common Transformations

Method	Math Operation	Good for:	Bad for:
Log	$\ln(x)$ $\log_{10}(x)$	Right skewed data $\log(x)$ is especially good at handling higher order powers of 10 (e.g., 100, 1000, etc.)	Zero values Negative values
Square root	\sqrt{x}	Right skewed data	Negative values
Square	x^2	Left skewed data	Negative values
Cube root	$x^{1/3}$	Right skewed data Negative values	Not as effective at normalizing as log
Reciprocal	$1/x$	Making small values bigger and big values smaller	Zero values Negative values



Multiple Linear Regression

Multiple Regression Model: A linear regression that contains more than one regressor (X)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

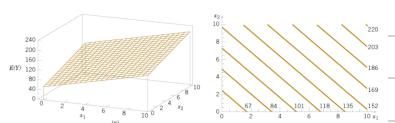


FIGURE 12.1 (a) The regression plane for the model $E(Y) = 50 + 10x_1 + 7x_2$. (b) The contour plot.

Nonlinear Surfaces

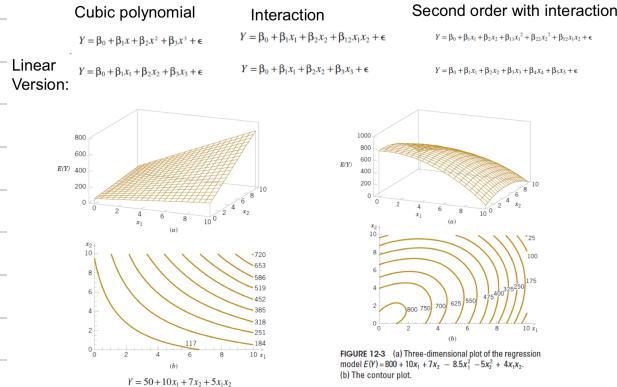


FIGURE 12.3 (a) Three-dimensional plot of the regression model $E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 + 4x_1x_2$; (b) The contour plot.

Estimated Coefficients and Variance

Coefficients: Found by method of least squares

\hookrightarrow Residuals: $e_i = y_i - \hat{y}_i$ square of these minimized

Since concepts simple
just a lot more terms

Variance from different sources

• Variance from Random Error: SS_E from calculation of residuals

$$SS_T = SS_R + SS_E \quad \hookrightarrow SS_E = \sum e_i^2$$

• Variance from Regression: SS_R = variance from regression

• Variance from individual regression terms: $SS_p(\beta_k) =$ variance coming from specific regression term

• Partial F-test: Components

Hypothesis Testing

Recall

$n = \# \text{ of data points}$

$k = \# \text{ of regression terms}$

$m = \# \text{ of parameters}$

$r = \# \text{ of residuals}$

$t = \# \text{ of terms}$

$\nu = n - m - k$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

$\nu = \# \text{ of degrees of freedom}$

$\nu = \# \text{ of errors}$

$\nu = \# \text{ of residuals}$

Example

Find a statistically satisfactory (alpha = 0.05) model for:

Brake Horsepower (y) rpm	Road Octane Number	Compression
225	2000	90
212	1800	94
229	2400	88
222	1900	91
219	1800	96
278	2500	96
246	3000	94
237	3200	90
233	2800	88
224	3400	86
223	1800	90
230	2500	89
		104

Method

Plot: Graph > Scatter plot > plot response against all X's

Fit Model: Stat > Regression > Fit Regression Model (add Response, and Predictors by double click) (Determining model by button "Model" to add or subtract terms)

Contour Plots: Stat - regression > regression > contour or surface plot

1) Plot the data

2) Fit a multiple linear regression model. Check the ANOVA results, R2adj, residuals, normality

3) Based on results, consider adding or removing regressors

4) If model is changed – repeat 2-3

5) Generate final model and contour plots

Final Result

Regression Equation

Brake Horsepower = $-266.0 + 0.01071 \text{ rpm} + 3.135 \text{ Road Octane Number}$

↓ is low → S is R + U

Goal: simplest model which significantly describes the data

Control Charts

- Statistical Process Control:** A collection of problem-solving tools useful in achieving process stability and reducing variability.

- The Magnificent Seven
 - Histogram or stem-and-leaf plot
 - Check sheet
 - Pareto chart
 - Cause-and-effect diagram
 - Defect concentration diagram
 - Scatter diagram
 - Control chart

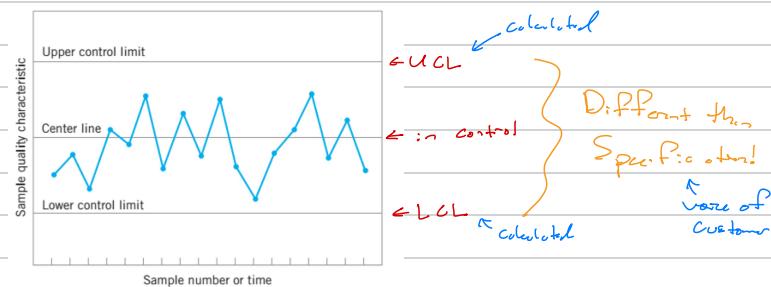
Showalter's Control Charts & Types of Variability

- Chance causes of variation:** Inherent variability in a process or background noise - a process operating with only chance causes is in statistical control.
- Assignable Causes of Variation:** Sources of variability that are not due to chance - a process operating outside of assignable ranges is an out-of-control process.
- Sources:** Improperly adjusted/controlled machines
Operator errors
Defective raw material *
- Control Charts:** Online process monitoring technique

Control Chart Uses

- Monitoring:** Using the control chart to determine if the process is in control, and quickly detect any assignable causes.
- Estimating:** Using the control chart to estimate process parameters (mean or variance).
- Improving:** Using the control chart to improve process parameters.

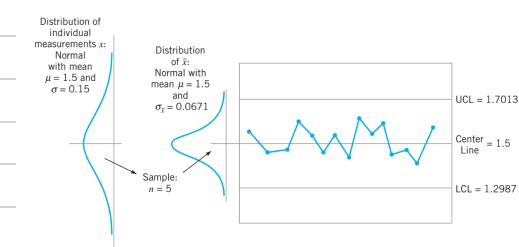
Features of a Control Chart



Relationship to Hypothesis Testing

Null Hypothesis: $M = M_0$ in statistical control

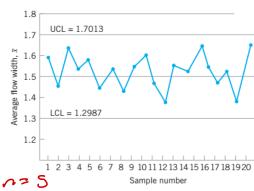
Alternative: $M \neq M_0$



Type I error: Rejecting H_0 when H_0 is true; concluding the process is out of control when it is not.

Type II error: Not rejecting H_0 when it is false; assuming a process is in control when it really is out of control.

Example



Control chart: uses Sample average to monitor a process mean

$MoLS$
 $\sigma = 0.15$
 $100(1-\alpha)\%$ of Sample means should fall between control limits

Three Sigma Control limits: Define $Z_{\alpha/2} = 3 \Rightarrow \alpha = 0.002 \Rightarrow$

$$\text{Control limits: } \bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

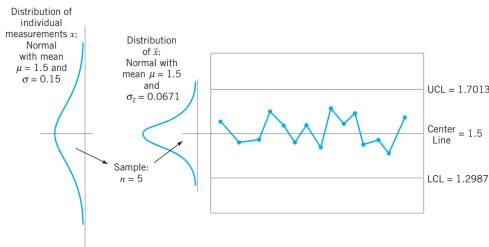
(Known values)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.15}{\sqrt{5}} = 0.0671$$

$$UCL = \bar{x} + 3(0.0671) = 1.7013$$

$$LCL = \bar{x} - 3(0.0671) = 1.2987$$

General Model



w = quality characteristic of interest

M_w = mean of w

σ_w = std dev of w

L = distance from centerline expressed in std dev

$$UCL = M_w + L \sigma_w$$

$$Center = M_w$$

$$LCL = M_w - L \sigma_w$$

Design of a Control Chart

Solving Control Limits?

- Set alpha (Europe)

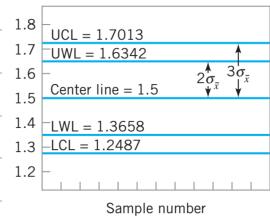
- Set Sigma Level (U_3)

U_3 is pretty standard compromise
balancing type I and II error

Two Limit Control Charts

- Action limits - outer limits that trigger a search for assignable cause

- Warning limits - inner limits which raise alarms



Average Run Length (ARL) - the average # of points that must be plotted before a point will indicate an out-of-control condition

$$ARL = \frac{1}{p}; p = \alpha, \text{ probability the point exceeds control limit}$$

Average Run Signal (ARS) - the average amount of time between out-of-control signals

$$ARS = ARL \times (\text{time between samples})$$

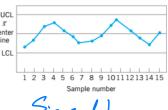
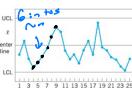
Analysis of Patterns A process can be out of control even if all points are within limits!

Run: A series of observations of the same type

Run Length: How many points are in the series

Pattern: Non-random behavior

Zone Rules: Criteria to help identify patterns



Dotsy Patterns and Sensitizing Rules

Standard Action Signal:	1. One or more points outside of the control limits 2. Two of three consecutive points outside the two-sigma warning limits but still inside the control limits 3. Four or five consecutive points beyond the one-sigma limits 4. A run of eight consecutive points on one side of the center line 5. Six points in a row steadily increasing or decreasing 6. Fifteen points in a row in zone C (both above and below the center line) 7. Eight points in a row alternating up and down 8. Eight points in a row on both sides of the center line with none in zone C 9. An unusual or nonrandom pattern in the data 10. One or more points near a warning or control limit
Sensitizing Rules:	Western Electric Rules
If yes, investigate more, get more data Raises alarm	Out of control if yes

Sensitizing Rules: Criteria to Increase Sensitivity



⇒ Out of Control example.

X-Chart: Average of samples

R-Chart: the difference between mean values of a sample

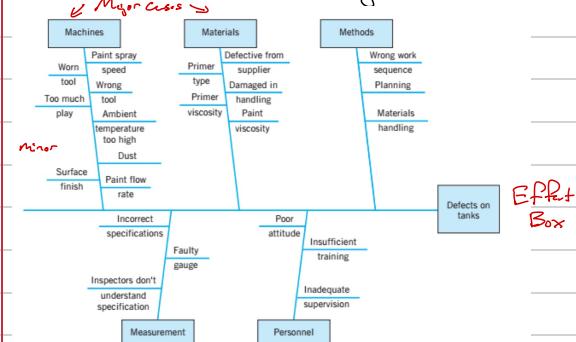
Out of Control Action Plans

OCAP: A flowchart or text-based description of the sequence of actions that will take place following an out-of-control event

Checkpoints: Potential assignable causes

Terminos: Action taken to Resolve

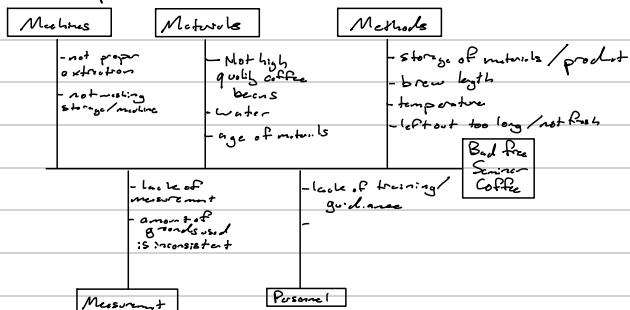
Cause and Effect Diagram



How to construct a C&E diagram

- Define the problem or effect to be analyzed.
- Form the team to perform the analysis. Often the team will uncover potential causes through brainstorming.
- Draw the effect box and the center line.
- Specify the major potential cause categories and join them as boxes connected to the center line.
- Identify the possible causes and classify them into the categories in step 4. Create new categories, if necessary.
- Rank order the causes to identify those that seem most likely to impact the problem.
- Take corrective action.

Example: Free Seminar Coffee



Test 2 Review

Topics

General:

- how to correctly phrase conclusions in a hypothesis test
- how to correctly phrase what the p-value is
- how to correctly phrase what confidence intervals are representing
- how to sketch hypothesis tests on a reference (sampling) distribution

Inferring the variance of two populations:

- how to conduct the hypothesis test
- how to draw conclusions

ANOVA:

- how to conduct the test (by hand and interpret Minitab) and draw conclusions
- what situations ANOVA is used in (when to pick it as the statistical test)
- what, generally, the test statistic represents
- what the reference distribution is as related to the test statistic and when the test statistic follows the reference distribution
- know how to check assumptions and what assumptions are made
- how and when to perform a posthoc test
- how and when to transform data

Linear Regression:

- know how to perform simple linear regression by hand (how to perform the hypothesis test – won't ask the regression terms by hand)
- how to set up equations and hypothesis for multiple and simple linear regression
- know how to interpret results from Minitab output on simple and multiple linear regression
- explain what R^2 is (+adjusted), what it can and cannot be used for, and how to calculate by hand for simple and multiple linear regression
- know how to check assumptions and what assumptions are made
- how and when to transform data
- describe what the confidence intervals on the response and coefficients are

Control Charts:

- Be able to explain what impact wide or narrow control limits have on type I and type II error, and be able to define type I and type II errors for control charts
- Be able to explain in words and sketches why it is important to use control charts for both process mean and variability
- Calculate control limits and warning limits for \bar{x} and R charts
- Be able to identify and out of control process
- ARL calculation and concept

ECHE 313 Exam 2

Write your name: _____

Write the following statement, and sign your name acknowledging your agreement:

"I neither received nor gave external assistance on this exam"

The exam layout is such that all problems add up to 100 points. This point total is meant to guide you in time allocation for this test, will be adjusted such that this exam is 25% of your total class grade. You may use your class materials, class notes, homework, book, and Minitab to aid in the completion of the exam, but you are not allowed to use any outside resources. The exam is designed to be completed in 1 hr and 15 min, but you will have between 10 AM 4/9 to 10 AM 4/10 to complete the exam. Note: Once you start the exam, you have 4 hours to complete the test. All of the necessary tables are provided in the back of your book. Please show your work and all steps necessary to arrive at your answer. It is expected you do this entire test by hand, using a calculator and the tables found in the back of your book. So while you are allowed to use Minitab if you wish to check your answers, it should not replace the work you have done by hand using the tables for your answers. Where applicable, it is best to write out the general formulas you want to use first, then write in your subbed in values so it is clear. Also, when reporting numbers, keep the same number of decimals as the data you are working with unless there are specific directions on how many sig figs to report. Good luck!

Premise: You work as a process engineer for a company that makes eclipse glasses. What a great time to be in this business! These glasses work by filtering light, reducing light exposure to the eye. Unless otherwise stated, your company uses $\alpha=0.05$.

Problem 1 (67 points total) Your team wants to change the concentration of materials that are used to block light in your product to save money. Material B is particularly expensive and is typically set at 5.00 wt% in the material. You want to know if the concentration of material B has an impact on the % transmittance of light, so you vary material B and test the amount of light that is transmitted through the material, and get the following data:

$\alpha = 0.05$

Material B (wt%)	% Transmittance			Ave.	Standard Deviation
	Replicate 1	Replicate 2	Replicate 3		
(typical) 5.00	0.0023	0.0023	0.0024	0.0023	0.000058
4.00	0.0023	0.0025	0.0028	0.0025	0.000252
3.00	0.0027	0.0024	0.0021	0.0024	0.000300
2.00	0.0032	0.0038	0.0035	0.0035	0.000300
1.00	0.0041	0.0039	0.0045	0.0042	0.000306

$$\bar{y}_{..} = \frac{0.0023 + 0.0025 + 0.0024 + 0.0035 + 0.0048}{5} = 0.00298$$

- a. (1 points) List the *name* of the hypothesis test you would use

ANOVA

- b. (2 points) Name the parameter of interest in the context of the problem statement

Multiple means, specifically $\mu_1, \mu_2, \mu_3, \mu_4$ at different levels of material B.

- c. (2 points) Write out the correct null hypothesis for this test

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$$

- d. (2 points) Write out the correct alternative hypothesis for this test

$$H_a: \mu_i \neq 0 \text{ for some } i \in \{1, 2, 3, 4\}$$

- e. (6 points) Write the formula of the test statistic, and other formulas needed to fully calculate the test statistic from the data above, and calculate its value given that SS_E is 0.00000068

$$F_0 = \frac{\sum_{i=1}^n (\bar{y}_{i..} - \bar{y}_{..})^2 / (a-1)}{\sum_{i=1}^n s_{i..}^2 / (a(n-1))}$$

$$SS_{\text{treatments}} = n \sum_{i=1}^n (\bar{y}_{i..} - \bar{y}_{..})^2$$

$$= 3 \left[(0.0028 - 0.00298)^2 + (0.0025 - 0.00298)^2 + (0.0024 - 0.00298)^2 + (0.0035 - 0.00298)^2 + (0.0042 - 0.00298)^2 \right]$$

$$= 8.364 \times 10^{-6} \quad \text{If you use calc/excel, you get to } 0.0000079$$

$$F_0 = \frac{(8.364 \times 10^{-6}) / (5-1)}{(0.00000068) / (5(5-1))} = 30.75$$

$$F_0 = 30.75 \quad F_0 = 29 \text{ if using exact value } 8$$

- f. (4 points) Describe what the test statistic represents (specifically the numerator and denominator of the test statistic as they relate to sources of variation) and what it means if the test statistic is very large versus very small

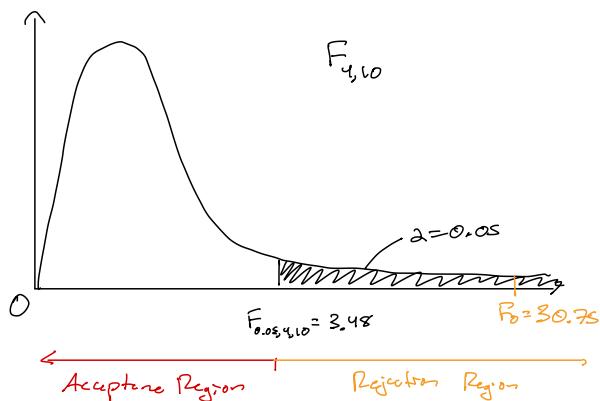
F_0 is effectively $\frac{\text{Var}_{\text{treatments}}}{\text{Var}_{\text{error}}}$. The numerator describes variance from treatments and the denominator describes variance from random error. If F_0 is very large variance is coming from treatments (but hypothesis test determines significance). If it is small, variance is comparable to random error. DF has an impact on the value.

- g. (4 points) Define your general rejection criteria using fixed significance level testing, sub in values and write out the critical value

$$\text{Reject } H_0 : F_0 > F_{\alpha, n-1, n(n-1)}$$

$$\Rightarrow F_0 > F_{0.05, 4, 10} = 3.48$$

- h. (10 points) Sketch out the reference distribution for this problem and label: the type of reference distribution used and any defining characteristics (shape, DOF if applicable, and where zero is), the critical value(s), the test statistic, alpha, the acceptance region, and the rejection region.



- i. (3 points) Perform any remaining calculations and state your conclusions. Remember to state them in the context of the problem statement, and phrase them appropriately

$$F_0 = 30.75 > F_{\text{crit}} = 3.48 \rightarrow \text{Reject } H_0$$

Reject the null and accept the alternative. We have significant evidence to indicate that the wt% of material B impacts % transmission.

- j. (7 points) Describe when it is appropriate to do post-hoc testing, and assuming it is appropriate in this case, conduct a Fisher's post hoc test and identify which groups are significantly different than the normal process setting (Material B at 5%).

A post-hoc test is only appropriate once the ANOVA H_0 is rejected!

Determine significantly different: $f: |\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > LSD$

$$S_{us. 4} = |0.0023 - 0.0028| = 0.0005$$

$$LSD = t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{2MSE}{n}}$$

$$S_{us. 3} = |0.0023 - 0.0021| = 0.0002$$

$$= t_{0.025, 10} \sqrt{\frac{2(0.0000063)}{3}}$$

$$S_{us. 2} = |0.0023 - 0.0028| = 0.0005$$

$$= 2.228 (2.129 \times 10^{-4})$$

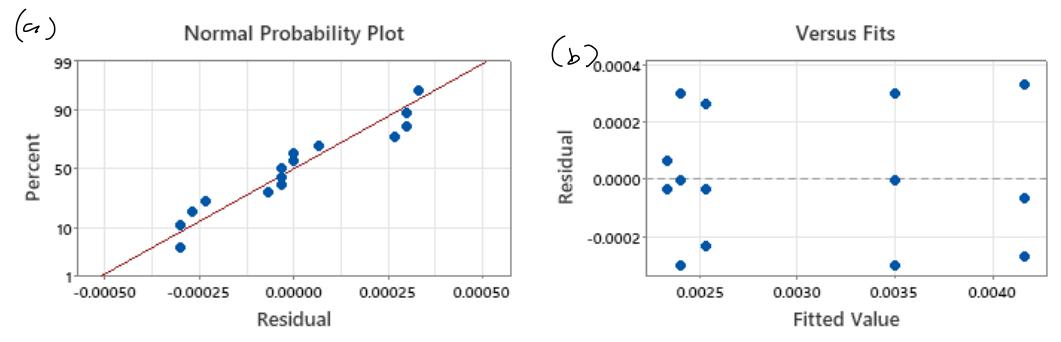
$$S_{us. 1} = |0.0023 - 0.0042| = 0.0019$$

$$= 4.74 \times 10^{-4}$$

$$= 0.000474$$

1 and 2 wtr% are significantly different than S-wtr% (norm)

- k. (4 points) If these are the residual plots:



1. State what assumptions are being checked for each plot

(a) Errors and observations are normally distributed

(b) Constant Variance

2. Describe what you should do if the assumptions are violated

Transform the data and perform ANOVA again.

- I. You are concerned that the variance of the data for 5 wt% treatment is different than the other treatments.
- (1 points) What is the name of the hypothesis test I should perform if I want to know if the variance from 5% treatment and 4% treatment are different?

Z -variance test!

- (2 points) Write out the correct null and alternative hypotheses for this test

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{where } 1 = 5\text{wt\%}, \text{ and}$$

$$H_a: \sigma_1^2 \neq \sigma_2^2 \quad 2 = 4\text{wt\%}$$

- (5 points) Calculate the test statistic (report 3 significant figures)

Write out the general formula:

Sub in values and calculate:

$$F_0 = \frac{s_1^2}{s_2^2}$$

$$F_0 = \frac{0.000058^2}{0.000252^2} = 0.052973$$

- (8 points) 1. Write out the rejection critical using p-value method and 2. determine the p-value from your test statistic (report 2 sig figs in your p-value). 3. Show the p-value graphically. In your sketch, you should have the reference distribution labeled with any defining characteristics (shape, zero, and degrees of freedom if applicable), the test statistic, and the p-value).

(i) $R_{\text{cut}}: F_{p-\text{val}} < 2 = 0.05$

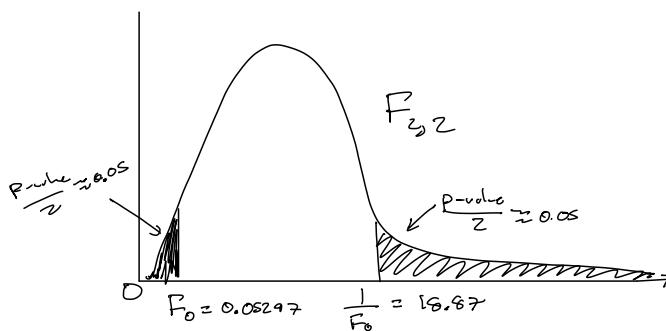
(ii) $F_{z_1 z_2} (F_{n_1-1, n_2-1})$

$$19.00 = F_{0.05, 2, 2}$$

$$\therefore p\text{-val} \approx 2(0.05) = 0.10$$

$$\Rightarrow p\text{-val} \approx 0.10$$

(iii)



5. (3 points) State your conclusions. Remember to state them in the context of the problem statement, and phrase them appropriately. Also use your conclusion to answer: is there cause for concern?

$P-value = 0.1 > 0.05 \Rightarrow$ Fail to reject H_0 . We do not have enough evidence to say that the variance at 5 wt% is different than 4 wt%. No cause for concern as the implies constant variance is still valid.

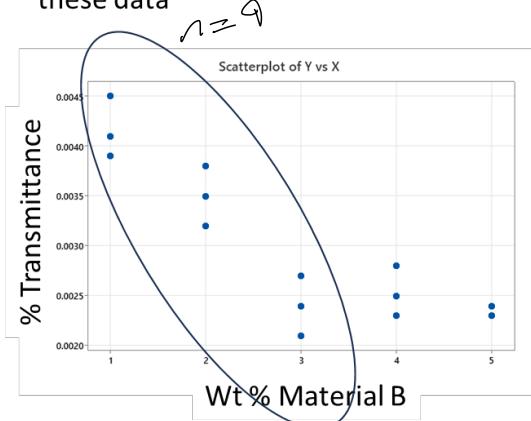
- m. (3 points) Provide and answer to your boss who was wondering: "Why is the rejection criteria for ANOVA one sided, whereas sometimes an F-test can be one- or two-sided?"

ANOVA is looking to see if a "signal" is larger than "noise". We therefore only care if the variance of the "signal" is larger than random error, thus it is one sided. If using an F-test we are determining if two variances are different, thus we need a two-sided test.

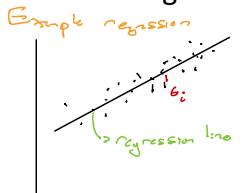
- Problem 2.** (22 points total) You suspect that the Material B wt% (from 1 to 3 wt% - predictor variable) is linearly correlated to % Transmittance (response variable). You want to perform a hypothesis test to see if the linear correlation is significant.

Performing linear regression on these data

$k = \# \text{ of regression lines}$
 $n = \# \text{ of data points}$



- a. (3 points) You perform the liner regression and your estimated coefficient for the slope is -0.00088 and the intercept is 0.0051. What is the name of the method that is used to estimate these coefficients, and describe generally how that method works. (Hint: I am NOT looking for how to calculate S_{xx} or S_{xy} here). You can use a diagram and general formulas for your answer if needed.



The method of least squares is used to estimate the coefficients in linear regression. In this method, the vertical deviations from data and the regression line (e_i 's) are minimized. These are the residuals, and the model aims to minimize $\sum_{i=1}^n e_i^2$.

- b. This is your Minitab output from the simple linear regression:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	0.00000468167	0.000468167	51.16	0.000
Error	7	0.00000064056	0.000000914		
Total	8				

- (1) (2 points) State the degrees of freedom for the $SS_{\text{Regression}}$ and the SS_{Error}

$$SS_{\text{Regression}} : \nu_C = 1$$

$$SS_{\text{Error}} : n - \nu_C - 1 = n - 2 = 7$$

- (2) (2 points) Write out the formula used and calculate the SS_{Total} using information given in the table above

$$\begin{aligned} SS_{\text{Total}} &= SS_R + SS_E \\ &= 0.00000468167 + 0.00000064056 \\ &= 0.0000053 \end{aligned}$$

- (3) (2 points) Write out the formula for and calculate R^2

$$\frac{SS_R}{SS_T} = \frac{0.00000468167}{0.0000053} = 0.88$$

- (4) (3 points) State your conclusions. Remember to state them in the context of the problem statement, and phrase them appropriately. Assume that no issues were noted with the residual plots.

$P_{\text{value}} = 0.2 > 0.05 \rightarrow P_{\text{reject }} H_0 \text{ and accept } H_a$. We have significant evidence to reject the claim that there is no linear relationship between material B from 1-3% and % transmittance. The wt% of material B from 1-3% is linearly correlated with % transmittance.

- c. (10 points) You think there might be another variable that could be useful in your model (wt% of another material, Material A). You add in this variable and run multiple linear regression to get the following output:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	0.000005	0.000002	26.21	0.001
Wt% B	1	0.000002	0.000002	19.92	0.004
Wt% A	1	0.000000	0.000000	1.03	0.349
Error	6	0.000001	0.000000		
Total	8	0.000005			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0003018	89.73%	86.30%	76.89%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.005844	0.000760	7.69	0.000	
X	-0.001100	0.000246	-4.46	0.004	4.00
X2	-0.000289	0.000285	-1.02	0.349	4.00

- (1) (3 points) The R^2 increased after adding the new variable. Should we keep the variable in the model? Why or why not?

No. R^2 is not a valid tool for determining the significance of a variable in regression. In addition, the partial F-test shows P-value > 0.05 = α for wt% A, so there isn't enough evidence to say it has a significant relationship with % transmittance.

- (2) (3 points) R^2 adjusted before adding Material A as a predictor was 86%. Explain why R^2 adjusted did not go up when adding the new variable. Use the formula in your answer.

$$R_{adj}^2 = \frac{SS_R/k}{SS_T/(n-1)}$$

R_{adj}^2 includes degrees of freedom which prevents k^2 from increasing simply by adding more regression terms.

- (3) (4 points) State your conclusions appropriately in the context of the problem statement and write out the final model

Fail to reject H_0 . We don't have enough evidence to say wt% A has a significant relationship with % transmittance.

The final model is: $y = 0.0051 - 0.00088x$

$\begin{cases} y = \% \text{ transmittance} \\ x = \text{wt\% material B} \end{cases}$

Problem 3. (11 points total) Your company wants to establish an Xbar control chart on the process for component C wt%. There are 22 preliminary samples, each of size 8, on the internal diameter of the seal. The summary data (in mm) are as follows:

Sample #	Range (R)	Average (Xbar)
1	5	15
2	9	11
3	8	11
4	4	13
5	9	13
6	5	10
7	10	14
8	5	14
9	9	12
10	8	13
11	8	15
12	8	11
13	8	15
14	8	12
15	5	13
16	8	14
17	7	14
18	4	11
19	9	12
20	8	15
21	8	14
22	5	11
Sum	158	283

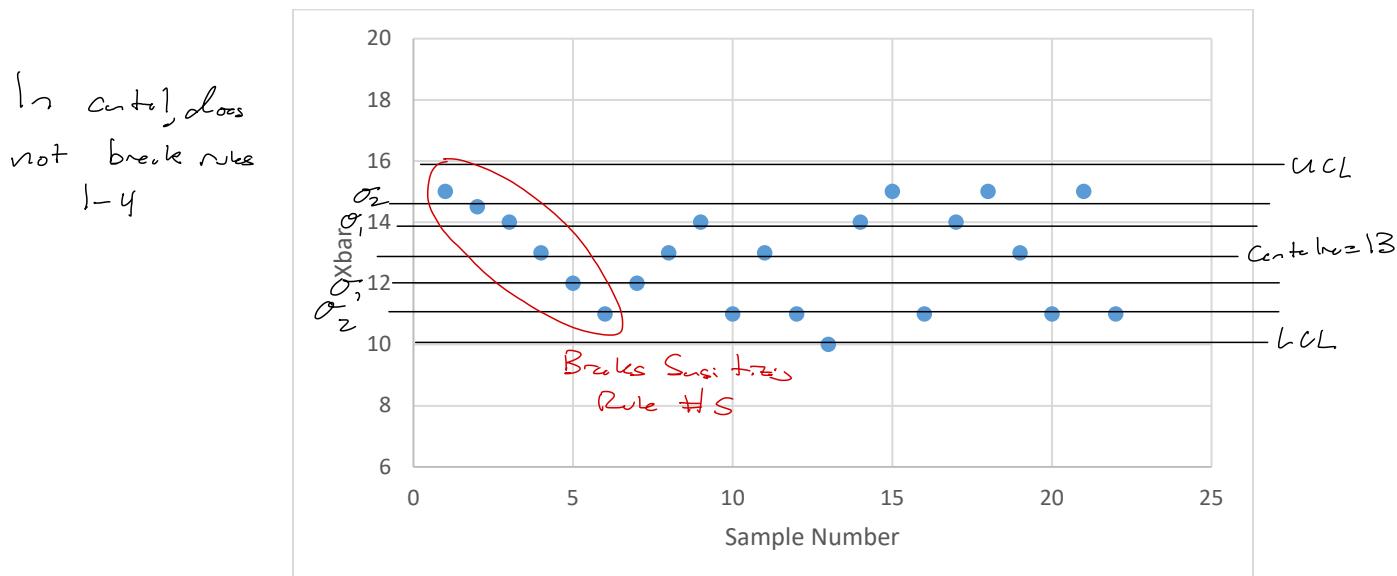
$$\bar{R} = \frac{158}{22} = 7.18 \quad \bar{\bar{x}} = \frac{283}{22} = 13.0$$

- a. (8 Points) Write out the general formulas for and calculate the centerline as well as the three sigma control limits that should be used on the Xbar and R control chart

$$X_{bar} \left\{ \begin{array}{l} UCL = \bar{x} + A_2 \bar{R} = 13.0 + 0.373(7.18) = 15.7 \approx 16 \\ Cen-tre = \bar{x} = 13.0 \approx 13 \\ LCL = \bar{x} - A_2 \bar{R} = 13.0 - 0.373(7.18) = 10.3 \approx 10 \end{array} \right.$$

$$R \left\{ \begin{array}{l} UCL = D_4 \bar{R} = 1.864(7.18) = 13.4 \approx 13 \\ Cen-tre = \bar{R} = 7.18 \\ LCL = D_3 \bar{R} = 0.136(7.18) = 0.98 \approx 1 \end{array} \right.$$

- b. (3 points) In the next month, you take an additional 20 samples in a similar manner as before. Sketch in your sigma control limits (UCL and LCL) and centerline from part A, as well as the 1 sigma and 2 sigma warning limits on this graph of your data. After analyzing this plot, what is your recommendation?



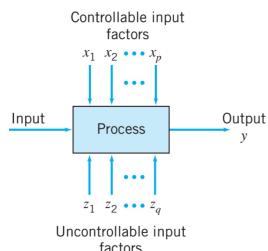
Process should be investigated as sampling #5 is broken

Final Exam



Design of Experiments

- Improve Stage!
- Improvements are Project Driven
- Also useful for 6 sigma (product design)



Designed Experiment: A series of tests in which purposeful changes are made to input variables of a process so we may observe and identify changes in output

- Objectives can be:
- (1) Determine which variables are most influential on y
 - (2) Determine what to set influential x 's so y is near required/optimal
 - (3) Determine what to set x 's so variability in y is small
 - (4) Determine where to set x 's so impact of z 's are minimal

- Uses:
- (1) Identify variables to investigate in out-of-control process
 - (2) Determine optimum settings (stability or cost)
 - (3) Identify optimum specification limits

Guidelines for Designing Experiments

- Pre-experimental planning
1. Recognition of and statement of the problem
 2. Choice of factors and levels
 3. Selection of the response variable
 4. Choice of experimental design
 5. Performing the experiment
 6. Data analysis
 7. Conclusions and recommendations
 8. Follow up runs confirming testing
 9. Iterations

Factors: Controllable inputs to be varied

Levels: Specific levels at which runs are made

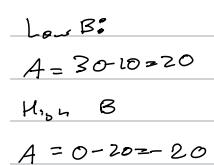
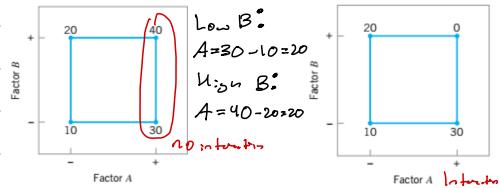
Response Variable: Should provide useful info about process (help achieve goal)

Factorial Experiments

Factorial Experiment: All possible combinations of the levels of the factors investigated

Main Effect: The change in response produced by a change in the level of a primary factor

Interaction: Occurs when the difference in response between the levels of one factor is not the same at all levels of the others



Run	A	B	Main Effect A:
1	+	+	$A = \bar{y}_4 + \bar{y}_4$
2	+	-	$= \frac{30+40}{2} - \frac{20+10}{2}$
3	-	+	
4	-	-	$= 15$

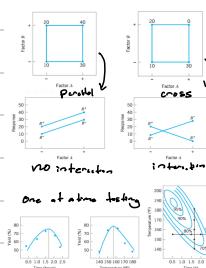
Importance of Interactions

Interactions can mask effects:

$$A = \frac{30-0}{2} - \frac{10+20}{2} = 0$$

Problems with one-at-a-time testing:

- inefficient
- no assured you will get optimum results



2-Factor Experiments (ANOVA)

		Factor B			Factors: A & B (2-factor)	
		1	2	...		
Factor A	1	$y_{11}, y_{12}, \dots, y_{1n}$	$y_{121}, y_{122}, \dots, y_{12n}$	\dots	y_{1m}, y_{10m}	
	2	$y_{21}, y_{22}, \dots, y_{2n}$	$y_{221}, y_{222}, \dots, y_{22n}$	\dots	y_{2m}, y_{20m}	
:	:	:	:	:	:	
0	$y_{01}, y_{02}, \dots, y_{0n}$	$y_{021}, y_{022}, \dots, y_{02n}$	\dots	y_{0m}, y_{00m}	\dots, y_{0n}	

Factors: A & B (2-factor)

Levels of Factor A = a

Levels of Factor B = b

Replicated n times

Effect of interaction

between A and B

normal dist. random error

$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$

$i = 1, 2, \dots, a$

$j = 1, 2, \dots, b$

$k = 1, 2, \dots, n$

Overall mean effect

Effect of Factor A

in level of Factor A

Effect of Factor B

in level of Factor B

Hypothesis Testing

Null & Alternative Hypothesis test for A, B, and AB:

$$H_0: \tau_i = 0 \text{ for all } i$$

$$H_0: \beta_j = 0 \text{ for all } j$$

$$H_0: (\tau_B)_{ij} = 0 \text{ for all } (ij)$$

$$H_a: \tau_i \neq 0 \text{ for any } i$$

$$H_a: \beta_j \neq 0 \text{ for any } j$$

$$H_a: (\tau_B)_{ij} \neq 0 \text{ for any } ij$$

Test Statistic: Reference distribution: F-distr.

The ANOVA Table for a Two-Factor Factorial, Fixed Effects Model			
Source of Variation	Degrees of Freedom	Sum of Squares	F _a
A	SS _A	a-1	$MS_A = \frac{SS_A}{a-1}$
B	SS _B	b-1	$MS_B = \frac{SS_B}{b-1}$
Interaction	SS _{AB}	(a-1)(b-1)	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$
Error	SS _E	ab(n-1)	$MS_E = \frac{SS_E}{ab(n-1)}$
Total	SS _T	abn-1	

Rejection Criteria:

$$\text{Reject if: } F_0 > F_{\alpha}$$

Total observations of
i-th level of factor A

$$y_{i..} = \sum_{j=1}^b \sum_{k=1}^n y_{ijk}$$

$$\bar{y}_{i..} = \frac{y_{i..}}{bn} \quad i = 1, 2, \dots, a - \text{row average}$$

$$\bar{y}_{..j} = \frac{\sum_{i=1}^a y_{i..}}{an} \quad j = 1, 2, \dots, b - \text{column average}$$

$$\bar{y}_{ij.} = \frac{y_{ij.}}{n} \quad i = 1, 2, \dots, a$$

$$\bar{y}_{...} = \frac{\sum_{i=1}^a \sum_{j=1}^b y_{ij.}}{abn} \quad \text{Cell average}$$

Grand total: $y_{...} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n y_{ijk}$

$$\bar{y}_{...} = \frac{y_{...}}{abn} \quad \text{Grand Mean}$$

$$\left\{ \begin{array}{l} SS_T = SS_A + SS_B + SS_{AB} + SS_E \\ SS_T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{...})^2 \\ SS_A = bn \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 \end{array} \right.$$

$$\left\{ \begin{array}{l} SS_B = an \sum_j (\bar{y}_{..j} - \bar{y}_{...})^2 \\ SS_{AB} = n \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{..j} + \bar{y}_{...})^2 \\ SS_E = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij.})^2 \end{array} \right.$$

Review - 2 Factor Experiments

Data for a Two-Factor Factorial Design

Factor B				
	1	2	...	b
1	Y ₁₁₁ , Y ₁₁₂ , ... Y _{11n}	Y ₁₂₁ , Y ₁₂₂ , ... Y _{12n}	...	Y _{1b1} , Y _{1b2} , ... Y _{1bn}
2	Y ₂₁₁ , Y ₂₁₂ , ... Y _{21n}	Y ₂₂₁ , Y ₂₂₂ , ... Y _{22n}	...	Y _{2b1} , Y _{2b2} , ... Y _{2bn}
:	:	:	:	:
d	Y _{d11} , Y _{d12} , ... Y _{dn}	Y _{d21} , Y _{d22} , ... Y _{dn}	...	Y _{db1} , Y _{db2} , ... Y _{dn}

Factors : A and B

Levels of factor A = a

Levels of factor B = b

Replicated n times

Y = response, y_{ijk}

Number of observations = abn

Null and Alternative Hypothesis Tests for A, B and AB:

$$H_0: \tau_i = 0 \text{ (for all } i\text{), } H_1: \tau_i \neq 0 \text{ (for some } i\text{)}$$

$$H_0: \beta_j = 0 \text{ (for all } j\text{), } H_1: \beta_j \neq 0 \text{ (for some } j\text{)}$$

$$H_0: (\tau_B)_{ij} = 0 \text{ (for all } ij\text{), } H_1: (\tau_B)_{ij} \neq 0 \text{ (for some } ij\text{)}$$

Test Statistic: (reference distribution is the F-distribution)

TABLE 13.3 The ANOVA Table for a Two-Factor Factorial Fixed Effects Model

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F _a
A	SS _A	a-1	$MS_A = \frac{SS_A}{a-1}$	$F_A = \frac{MS_A}{MS_E}$
B	SS _B	b-1	$MS_B = \frac{SS_B}{b-1}$	$F_B = \frac{MS_B}{MS_E}$
Interaction	SS _{AB}	(a-1)(b-1)	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$F_{AB} = \frac{MS_{AB}}{MS_E}$
Error	SS _E	ab(n-1)	$MS_E = \frac{SS_E}{ab(n-1)}$	
Total	SS _T	abn-1		

Rejection Criteria:

Reject if $F_0 > F_{\alpha}$, DOF numerator, DOF denominator

Data for a Two-Factor Factorial Design

Factor B				
	1	2	...	b
1	Y ₁₁₁ , Y ₁₁₂ , ... Y _{11n}	Y ₁₂₁ , Y ₁₂₂ , ... Y _{12n}	...	Y _{1b1} , Y _{1b2} , ... Y _{1bn}
2	Y ₂₁₁ , Y ₂₁₂ , ... Y _{21n}	Y ₂₂₁ , Y ₂₂₂ , ... Y _{22n}	...	Y _{2b1} , Y _{2b2} , ... Y _{2bn}
:	:	:	:	:
d	Y _{d11} , Y _{d12} , ... Y _{dn}	Y _{d21} , Y _{d22} , ... Y _{dn}	...	Y _{db1} , Y _{db2} , ... Y _{dn}

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

$$SS_T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{...})^2 = bn \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2 + an \sum_{j=1}^b (\bar{y}_{..j} - \bar{y}_{...})^2 + n \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{..j} + \bar{y}_{...})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (y_{ijk} - \bar{y}_{ij.})^2$$

1) Parameters of Interest: Main effects and interactions of method of application and primer type

$$2) H_0: \tau_1 = \tau_2 = \tau_3 = 0$$

$$\beta_1 = \beta_2 = 0$$

$$(\tau_B)_{11} = (\tau_B)_{12} = (\tau_B)_{21} = (\tau_B)_{22} = (\tau_B)_{31} = (\tau_B)_{32} = 0$$

$$3) H_0: \tau_i \neq 0 \text{ for any } i: (i=1-3)$$

$$\beta_j \neq 0 \text{ for any } j: (j=1-2)$$

$$(\tau_B)_{ij} \neq 0 \text{ for any } ij: (i=1-3, j=1-2)$$

4) Test Stats - See Summary

⇒ Rejection Criteria: $\text{Reject if: } F_0 > F_{\alpha}$, DOF numerator, DOF denominator

6) Calculations

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

$$\bar{y}_{...} = \frac{89.8}{3} = 4.9889$$

$$(S)(2)(3)$$

$$SS_A = bn \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 = 2(3) [(4.0 - 4.9889)^2 + (5.4 - 4.9889)^2 + (4.5 - 4.9889)^2] = 4.5881$$

$$SS_B = an \sum_j (\bar{y}_{..j} - \bar{y}_{...})^2 = 3(3) [(4.667 - 4.9889)^2 + (5.511 - 4.9889)^2] = 4.90889$$

$$SS_{AB} = n \sum_{i,j} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{..j} + \bar{y}_{...})^2$$

$$= 3 [(4.266 - 4.7827 - 4.667 + 4.9889)^2 + \dots] = 0.2411$$

$$SS_E = \sum_{i,j,k} (y_{ijk} - \bar{y}_{ij.})^2 = (4.0 - 4.266)^2 + (4.5 - 4.266)^2 + \dots = 0.98667$$

$$all \ points$$

all points

	Calcs	F ₀	F _{crit} :
A	$F_0 = \frac{SS_A / a-1}{SS_B / ab(a-1)} = \frac{0.5811 / 3-1}{0.98667 / 3(2)(3-1)} = 27.858$	F _{0.05, 2, 12} = 3.89	✓ yes
B	$F_0 = \frac{SS_B / b-1}{SS_B / ab(a-1)} = \frac{4.90829 / 2-1}{0.98667 / 12} = 59.7027$	F _{0.05, 1, 12} = 4.75	✓ yes
AB	$F_0 = \frac{SS_{AB} / (a-1)(b-1)}{SS_B / ab(a-1)} = \frac{0.2411 / 2(1)}{0.98667 / 12} = 1.4667$	F _{0.05, 2, 12} = 3.89	X no

TABLE 13.3 The ANOVA Table for a Two-Factor Factorial, Fixed Effects Model									
Source of Variation		Sum of Squares		Degree of Freedom		Mean Square		F ₀	
A		SS _A	a-1	M _{SAA} = $\frac{SS_A}{a-1}$		F _{0A} = $\frac{M_{SAA}}{M_{SAB}}$		F _{0A}	$\frac{M_{SAA}}{M_{SAB}}$
B		SS _B	b-1	M _{SBB} = $\frac{SS_B}{b-1}$		F _{0B} = $\frac{M_{SBB}}{M_{SAB}}$		F _{0B}	$\frac{M_{SBB}}{M_{SAB}}$
Interaction		SS _{AB}	(a-1)(b-1)	M _{SAB} = $\frac{SS_{AB}}{(a-1)(b-1)}$		M _{SAB}		F _{0AB}	$\frac{M_{SAB}}{M_{SAB}}$
Error		SS _e	ab(a-1)	M _{Se} = $\frac{SS_e}{ab(a-1)}$		M _{Se}		F _{0e}	$\frac{M_{Se}}{M_{Se}}$
Total		SS _T	ab(a-1)						

7) Conclusions (assuming assumptions are valid)

We conclude that primer type & application method have a significant impact on adhesion force. There isn't enough evidence to say there is an interaction between application method and primer type.

8) Plots - Main effects & Interaction

9) Residual Analysis (assumptions)

10) Model generation

Plots & Minitab and Excel

Interaction Plots: Will plot average response at each combination of factor settings vs. factor setting

Main Effects Plots: Will plot average response at each

Excel

- 1) Set up your data table
- 2) Calculate cell averages, row and column averages and grand averages
- 3) Calculate sum of squares terms (use the space)
 - a. SST and SSE – a term for each data point
 - b. SSA and SSB – a term for each row or column average
 - c. SSAB – a term for each averaged cell
- 4) Calculate F₀ for each term (A, B, AB) and compare with critical values
- 5) Plot interaction and main effects plots
 - a. Interaction plots – cell averages at each level
 - b. Main effects plots – column or row averages at each level

Minitab: See Document on Course

Checking Assumptions

TABLE 13.6 Residuals for the Aircraft Primer Paint Experiment									
Primer Type	Application Method								
1	Dipping	Spraying							
2	-0.26, 0.23, 0.03	0.10, -0.40, 0.30							
3	0.30, -0.40, 0.10	-0.26, 0.03, 0.23							
	-0.03, -0.13, 0.16	0.34, -0.17, -0.23							

$$\begin{aligned} e_{ijk} &= y_{ijk} - \hat{y}_{ijk} \quad \text{predicted} \\ r_{ikl} &= y_{ikl} - \hat{y}_{ikl} \quad \text{calculated} \end{aligned}$$

Assumptions and Plots:

- 1) The errors are normally distributed (plot residuals on normal probability plot)
- 2) The errors are independently distributed (plot residuals versus run order)
- 3) Each factor level has the same variance (residuals versus factor levels, both A and B levels in this case, and predicted values)

Checking Assumptions - Cont

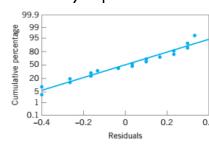


FIGURE 13.13 Normal probability plot of the residuals from Example 13.5.

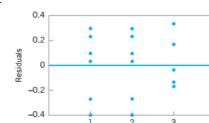


FIGURE 13.14 Plot of residuals versus primer type.

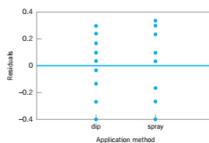


FIGURE 13.15 Plot of residuals versus application method.

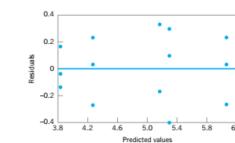


FIGURE 13.16 Plot of residuals versus predicted values.

Model Generation

- 1) Conduct the DOE analysis outlined for the full model (ANOVA, residuals and plots)
- 2) Based on results, consider removing terms from the model that are above the alpha level in the ANOVA table
- 3) Repeat 2-3 till all of the factors are significant
- 4) Generate final model and contour plots (if applicable)

Process is called backward elimination (automatically done via):

Stat->DOE-> Factorial > Analyze Factorial Design > Select "stepwise" button > Select "backward elimination"

$$y_{ijk} = \mu + \tau_i + \beta_j + (\beta\beta)_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \\ k = 1, 2, \dots, n \end{cases}$$

$$y_{ijk} = \mu + \tau_i + \beta_j + (\beta\beta)_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \\ k = 1, 2, \dots, n \end{cases}$$

$$Y = \text{grand mean} + (\text{grand mean} - \text{average at } t1) t1 + (\text{grand mean} - \text{average at } t2) t2 + (\text{grand mean} - \text{average at } t3) t3 + \dots \quad (\text{Note: } t1, t2, B1, \dots \text{ And so on... these are 0 or 1 depending on the conditions})$$

$$(\text{grand mean} - \text{average at } B1) B1 + (\text{grand mean} - \text{average at } B2) B2 + \dots$$

If interactions were present, it would be each combination of factor levels, with coefficients given by: (grand mean + cell average - row average A - column average B)

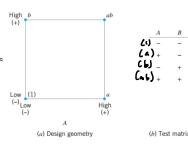
The 2^k Experiment

2^k Factorial Design - A designed experiment with k factors at 2 levels

- Forms the basis for many other DoEs
- Simplified Analyses
- # of runs = $n = n^{k+1}$ $k = \# \text{ of factors}$

e.g. If an experiment has 4 factors, replicated twice, how many runs are there

$$\# \text{ of runs} = 2(2^4) = 2 \cdot 16 = 32$$



- Each corner is a run
- Lower case letters = high setting
- no letter = low setting
- (1) = all low settings

Let C_1, a, b , and ab = totals of all n observations taken at the design points

→ Main Effect of A :

→ Interaction AB

$$A = \bar{y}_{A+} - \bar{y}_{A-}$$

$$= \frac{a+ab}{2n} - \frac{b+ab}{2n} = \frac{1}{2n} (a + ab - b - C_1)$$

contrast

$$AB = \frac{1}{2n} (ab + C_1) - a - b$$

→ Main Effect of B :

$$B = \bar{y}_{B+} - \bar{y}_{B-}$$

$$= \frac{b+ab}{2n} - \frac{a+C_1}{2n} = \frac{1}{2n} (b + ab - a - C_1)$$

Contrasts in 2^k Factorial Design

TABLE 13.7
Signs for Effects in the 2^2 Design

Run	Factorial Effect			
	I	A	B	AB
1 (1)	+	-	-	+
2 a	-	+	-	-
3 b	+	-	+	-
4 ab	+	+	+	+

Contrast: The total of all n observations at each combination of factors, levels, multiplied by the corresponding sign of that factor or interaction

$$\text{Effect} = \frac{\text{Contrast}}{n2^{k-1}}$$

C Contrast $A = a + ab - b - C_1$

$$\text{Contrast } AB = ab + C_1 - a - b$$

total

Sum of squares for this example:

Error Dof	$(\text{Total Dof} - \text{all main effects})/(\text{Interaction Dof})$
$\text{SS}_A = \frac{[a + ab - b - C_1]^2}{4n}$	$\text{SS}_B = \frac{[b + ab - a - C_1]^2}{4n}$
$\text{SS}_{AB} = \frac{[ab + C_1 - a - b]^2}{4n}$	

(13.15)

2^k experiment: total Dof = $n2^{k-1}$

$$SS_T = \sum_{\text{all points}} (\text{obs} - \bar{y}_{\text{grand mean}})^2$$

Dof: $A > 1$

$B > 1$

$AB > 1$

... (all main effects/intactions)

$$SS_E = \text{can be same as}$$

before if g is here

replicated

Example

A router is used to cut registration notches in printed circuit boards. The average notch dimension is satisfactory, and the process is in statistical control (see the \bar{x} and R control charts in Figure 13.18), but there is too much variability in the process. This excess variability leads to problems in board

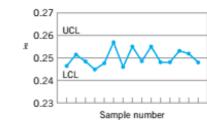
assembly. The components are inserted into the board using automatic equipment, and the variability in notch dimension causes improper board registration. As a result, the auto-insertion equipment does not work properly. How would you improve this process?

Since the process is in statistical control, the quality improvement team assigned to this project decided to use a designed experiment to study the process. The team considered two factors: bit size (A) and speed (B). Two levels were chosen for each factor (bit size A at $\frac{1}{8}$ and $\frac{3}{8}$ and speed B at 40 rpm and 80 rpm), and a 2^2 design was set up. Since variation in notch dimension was difficult to measure directly, the team decided to measure it indirectly. Sixteen test boards were instrumented with accelerometers that allowed vibration in the (X , Y , Z) coordinate axes to be measured. The resultant vector of these three components was used as the response variable. Since vibration at the surface of the board when it is cut is directly related to variability in notch dimension, reducing vibration levels will also reduce the variability in notch dimension.

Four boards were tested at each of the four runs in the experiment, and the resulting data are shown in Table 13.8.

■ TABLE 13.8
Data from the Router Experiment

AB	Factors		Vibration	Total	
	Run	A	B		
+	1	(1)	-	18.2 18.9 12.9 14.4	64.4
-	2	a	+	27.2 24.0 22.4 22.5	96.1
-	3	b	-	15.9 14.5 15.1 14.2	59.7
+	4	ab	+	41.0 43.9 36.3 39.9	161.1



1) Parameters of Interest: Mean ERPs and interactions of bit size (A) and speed or vibration (B)

2) $H_0: A=0, B=0, AB=0$

3) $H_1: A \neq 0, B \neq 0, AB \neq 0$

$$\sum_{k=2}^{n=4} Dof = A \times B = 4 \times 3 = 12$$

$$Total = n^2 k^2 = 4 \times 4 = 16$$

$$Error = 12$$

Reject: $F_0 \Rightarrow F_1$ $F_2 \Rightarrow Dof$ Dof

Top: SS_A Bottom: SS_B

Dof_F Dof_E

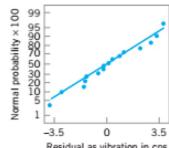
■ TABLE 13.9
Analysis of Variance for the Router Experiment

Sources of Variation	Sums of Squares	Degrees of Freedom	Mean Square	F ₀	P-value
Bit size (A)	1,107.226	1	1,107.226	185.25	1.17×10^{-4}
Speed (B)	227.256	1	227.256	38.03	4.82×10^{-2}
AB	303.301	1	303.631	50.80	1.20×10^{-4}
Error	71,723	15	5,977		
Total	1,709.836	15			

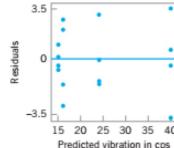
Model & Results

■ TABLE 13.8
Data from the Router Experiment

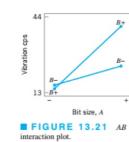
Run	Factors		Vibration	Total
	A	B		
1	(1)	-	18.2 18.9 12.9 14.4	64.4
2	a	+	27.2 24.0 22.4 22.5	96.1
3	b	-	15.9 14.5 15.1 14.2	59.7
4	ab	+	41.0 43.9 36.3 39.9	161.1



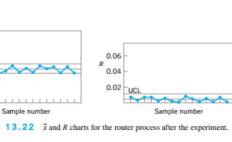
■ FIGURE 13.19
Normal probability plot, Example 13.6.



■ FIGURE 13.20
Plot of residuals versus \hat{y} , Example 13.6.

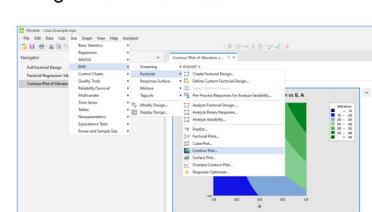
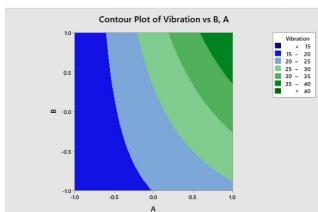


■ FIGURE 13.21
AB interaction plot.



■ FIGURE 13.22
 \bar{x} and R charts for the router process after the experiment.

Concern: if speed (B) is too slow will slow down production – this tells us that we can have high B if we have low A



Minitab: The directions for this are similar to the directions before, except you should select a 2-level factorial in the menu. You can also generate a contour plot with the model to help you determine the optimum settings.

Effects:

$$A = \frac{1}{2n} (a+b-b-CI) = \frac{1}{2(4)} (46.1 + 161.1 - 59.7 - 64.4) = 133.1$$

$$= \frac{133.1}{8} = 16.64$$

$$B = \frac{1}{2n} (b-a-a-CI) = \frac{1}{2(4)} (59.7 + 161.1 - 46.1 - 64.4) = \frac{60.8}{8} = 7.64$$

$$AB = \frac{1}{2n} (ab + (1) - a - b) = \frac{1}{8} (161.1 + 64.4 - 96.1 - 59.7) = \frac{69.7}{8} = 8.71$$

$$SS_A = \frac{133.1^2}{4(4)} = 1107 \quad SS_{AB} = \frac{69.7^2}{4(4)} = 304$$

$$SS_B = \frac{60.8^2}{4(4)} = 227 \quad SS_E = SS_T - SS_A - SS_B - SS_{AB} = 170 - 1107 - 227 - 304 = 74.7$$

Conclusions

Both bit size and speed have a significant impact on vibration, and there is a significant interaction between bit size and speed.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{12} x_1 x_2 + \epsilon$$

y = response variable

x_i 's = coded variables (-1, 1)

= interaction terms are products

B terms = coefficients for model $\rightarrow \frac{\text{Effect}}{2}$

β_0 = grand average

For our above example:

$$y = 23.83 + \left(\frac{16.64}{2}\right)x_1 + \left(\frac{7.64}{2}\right)x_2 + \left(\frac{8.71}{2}\right)x_1 x_2$$

β_0 = grand mean

β_1 = main effect A

β_2 = main effect B

effect P interaction AB

$$\epsilon_{ij} = y_{ij} - \hat{y}_{ij}$$

residual = data point - predicted value

concern (small bit b low speed) $\rightarrow (-1)(-1)$

$$\text{Estimate } \hat{y} = 23.83 + \frac{16.64}{2}(-1) + \frac{7.64}{2}(-1) + \left(\frac{8.71}{2}\right)(-1)(-1) = 16.64$$

Summary