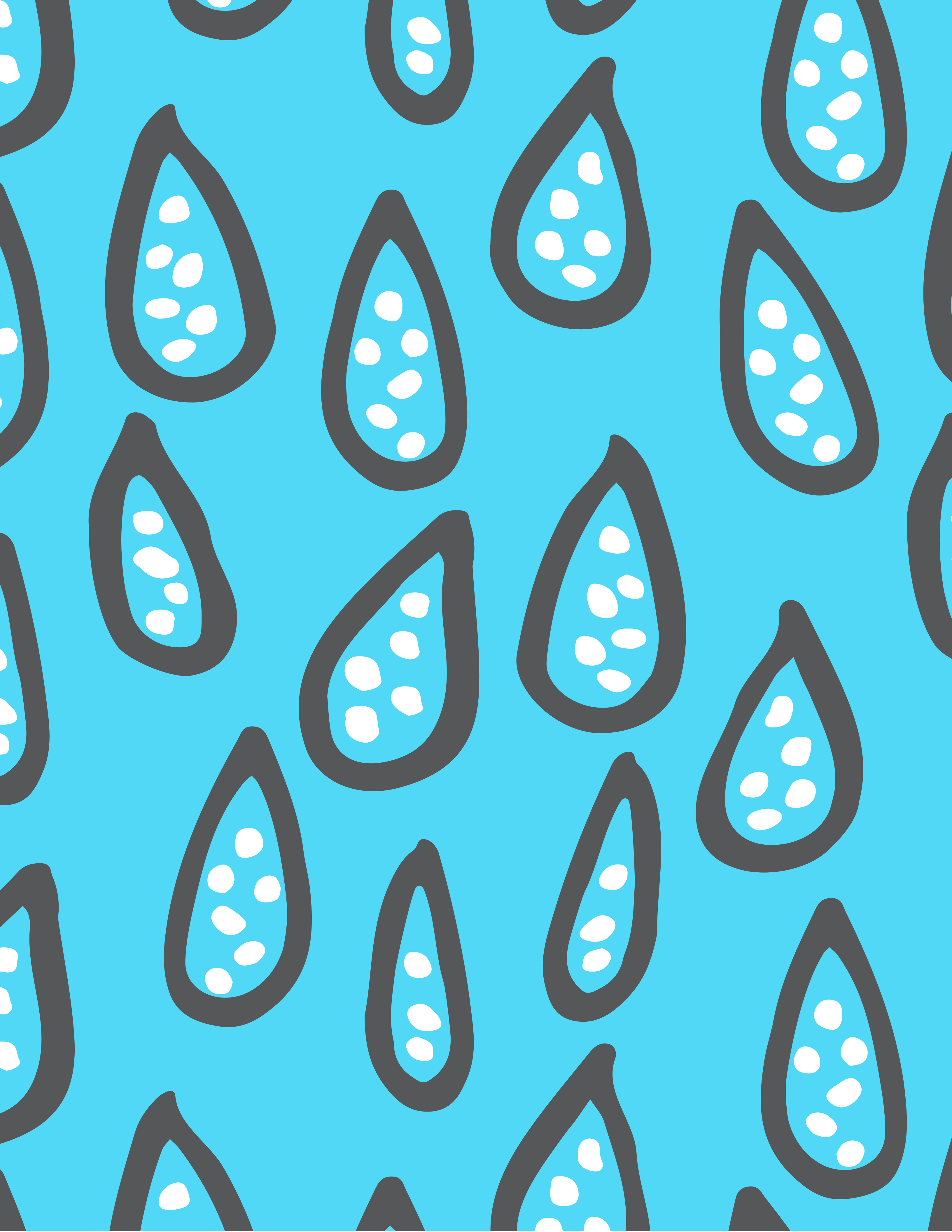


Midterm Exam 1

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) Reliability — how often will it fail
  - (iii) Durability — lifetime → will it meet it
  - (iv) Serviceability — easy to repair
  - (v) Aesthetics — products appearance
  - (vi) Features — what the product does
  - (vii) Perceived Quality — Reputation
  - (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 Factory: The part of the plant the customer doesn't see which deals w/ bad quality

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

will call c2qc ↓ Sometimes called 'critical to quality characteristics'

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: Analyses, interpretation, and presentation of numerical data

Often an upper and lower limit

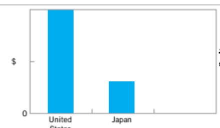
Specifications: Desired measurements for quality characteristics. This is your target value

Nonconformity: Failure to meet spec

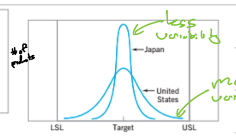
Defect: Nonconformity that limits use

Variability

- Quality is inversely proportional to variability
  - Variability will impact all quality dimensions
- Variability leads to greater cost & waste



■ FIGURE 1.1 Warranty costs for transmissions.



■ FIGURE 1.2 Distributions of critical dimensions for transmissions.

example to right examines 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	Retest
Process control	Failure analysis
Burn-in ← JR's company 20 yr test	Downtime
Training ← Personel	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 results 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.

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

$$= \frac{\$70(100) + \$4(100)(0.25 \cdot 0.60)}{70 + 15} = \boxed{\$22.89}$$

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.

$$\frac{\text{cost}}{\text{good}} = \frac{\$70(100) + \$4(100)(0.05)(0.60)}{95 + 3}$$

$$= \boxed{\$20.53}$$

Quality Engineering

Process: A system of inputs and outputs

Example: Defining 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.

- Controllable input variable: barcoding
- Key output variable: right amount of drug

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

- Centerline: where the characteristic should be if no abnormal sources of variation
- Control Limit: Defined by statistics. Alert the user to unusual variability
  - ↳ Different than spec limits

Designed experiments: discovers key variable that 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 $6\cdot\sigma$ is within the specification limits

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

Why is Six Sigma Selected?

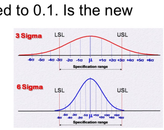
Suppose 100 parts: 3σ : $(.9973)^{100} = .7631$ is probability non-defective → ~24% defective

6σ : $(.9999998)^{100} = .99999998$ non-defective → ~.000002% defective

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?



$$\frac{|USL \text{ or } LSL - \text{target}|}{\sigma\,(\text{standard deviation})} = \frac{|9\,mm - 10\,mm|}{1\,mm} = 1 \quad \text{for old} \quad \text{"}1\sigma\text{ level"}$$

$$= \frac{|9\,mm - 10\,mm|}{0.1\,mm} = 10 \quad \text{for new} \quad \text{"}10\sigma\text{ level"}$$

Lean Manufacturing

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

Process Cycle Efficiency (PCE): (Value added time) / (Process time)

↳ **Value added time** = time spent making the product more valuable

Process Cycle Time (PCT): (Number of items in progress) / (completion rate) ← backlog issues as an example

(number / unit time)

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 personel
  - (iii) More opportunity for damage or loss
  - (iv) Inventory has to be higher
  - (v) More documentation

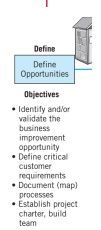
DMAIC Process

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

↳ Define → Measure → Analyze → Improve → Control

Tollgates: where projects are reviewed to ensure on track

5 Stages of DMAIC (will be on HW too)



(i) **Define Stage**

- **Main Obj:** Identify a project opportunity and verify that it represents a legitimate value opportunity or breakthrough
  - o **Breakthrough:** Major improvement to a process or product
  - o **Value Opportunity:** The financial opportunity at stake
- **Tools used**
  - o **SIPOC diagram:** High level map of process
    - → Suppliers: Those who provide information, material or other items that are worked on in the process
    - → Input: The actual information or material provided
    - → Process: The set of steps required to do the work
    - → Output: The product delivered to the customer → Other depts in same company
    - → 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

lean concept

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

Example: Value Stream Mapping (Lawn Mowing)

Old: Gather Tools (30 min) → Fuel Mowers (10 min) → Travel (30-60 min, variable) → Unload (10 min) → Talk to Customer (10-20 min, variable) → Care for Lawn (60-120 min, variable) → Bill Customer (10 min) → Gather Tools/Pack Up (10 min) → Drive to next Customer (30-60 min, variable) → ...

Possible Improvements: Using customer's Lawnmower, storage SOP (leave on truck), prefuel, personal/interns, eliminate/reduce customer interactions, optimize route (graph shortest distance)/logistics

- **Pareto charting:** Plots frequency (# of parts) or cost (time/money) against causes
  - o **Pareto Rule:** 80% of the effects come from 20% of the causes

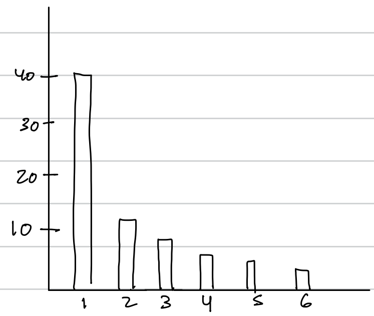
• Graph Pareto by Hand (Example)

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	Misaligned components	21	12,000.00
3	Defective components	15	8,000.00
4	Missing components	10	5,100.00
5	Cold solder joints	7	5,000.00
6	All other causes	5	4,600.00
	Totals	100%	$72,200.00



Using Minitab

1) Input data into Minitab by Type of Defect, and Cost (or other)
2) Go to Stat -> Quality Tools -> Pareto Chart
3) Specify which column the defects or attribute data are in (these are the labels with names)
4) Specify what column the "frequencies" are in (this can be costs, or number of defects – the numerical data)
5) Click ok and check that the chart makes sense

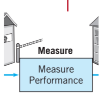
Can also do this with Lean concepts

Project Charter: Short document which describes the product.

Example to right, but varies accross companies

Business Case	Opportunity Statement
• This project supports the business quality goals, namely (a) reduce customer resolution cycle time by x% and (b) improve customer satisfaction by y%.	• 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**
• Reduce the overall response cycle time for returned product from our customers by x% year to year.	• 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**
• Activity — Start — End Define 6/04 6/30 Measure 6/18 7/30 Analyze 7/15 8/30 Improve 8/15 9/30 Control 9/15 10/30 Track Benefits 11/01	• Team Sponsor • Team Leader • Team Members

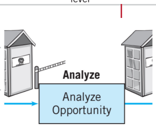
(ii) Measure Stage



Objectives
- Determine what to measure
- Manage measurement data collection
- Develop and validate measurement systems
- Determine Sigma performance level

• Objective: Evaluate & understand the current state of the process
• Baseline Performance: Quantify what the current process is performing at
• Measurement Capability: Quantify if the current measurement systems are adequate
• Tools Used: Charts to understand variability
Histograms, boxplots, run charts, scatter diagrams

(iii) Analyze Stage



Objectives
- Analyze data to understand reasons for variation and identify potential root causes
- Determine process capability, throughput, cycle time
- Formulate, investigate, and verify root cause hypotheses

• Objective: Use data to begin to determine cause and effect relationships and sources of variability
• 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 of variability, errors, or defects in a process or product
↳ For each item, we determine:
(i) Likelihood something will go wrong | (not likely) 1 - 10
(ii) Ability to detect the failure or defect | (likely) 1 - 10
(iii) Severity | (not severe) 1 - 10

Risk Priority #: (i) × (ii) × (iii)

Example

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

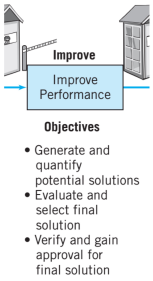
RPN = A X B X C

RPN = Risk Priority Number

FAILURE MODE & EFFECTS ANALYSIS (FMEA) — Date: 1/1/2018 — Revision: 1.3

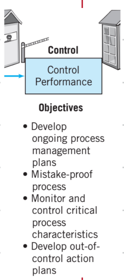
Process Name: Left Front Seat Belt Install — Process Number: SBT 445

Failure Mode	A) Severity Rate 1-10 10=Most Severe	B) Probability of Occurence Rate 1-10 10=Highest Probability	C) Probability of Detection Rate 1-10 10=Lowest Probability	Risk Preference Number (RPN) AxBxC
1) Select Wrong Color Seat Belt	5	4	3	60
2) Seat Belt Bolt Not Fully Tightened	9	2	8	144
3) Trim Cover Clip Misaligned	2	3	4	24



(iv) Improve Stage

- Objective: Use creative thinking & info gathered previously to make specific suggestions that will lead to desired Perfection process. Confirmation is important
- Tools Used
  - Mistake Proofing (example)
  - Optimization
  - Pilot Testing
  - Confirmation Experiments
  - Documentation
    - (i) Solution obtained
    - (ii) Alternatives Considered
    - (iii) Confirmation tests
    - (iv) Risk Analysis



(v) Control Stage

- Objective: Complete all remaining work and hand off to improved process so gains in progress are institutionalized
- Tools Used
  - Control Charts
  - Transition Plans

Chapter 3 - Quantifying Stats in Quality Characteristics

Statistics

Sample Mean: Average Value

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

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

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

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

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

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

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

$$s^2 = \frac{(12-23)^2 + (12-23)^2 + \ldots + (31-23)^2}{17-1} = 37 \qquad s = \sqrt{37} = 6.08$$

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

Illustrate 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 Variability

↳ Often easier to arranged sorted (lowest → highest) data

- **Percentiles:** Tell 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	← Values
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	← Rank (1-indexed)

- **Lower Quartile** (Q_1 or 25th perc.): Value with rank $= (n+1)(\frac{1}{4})$
  here: $\frac{17+1}{4}$ = rank of 4.5 → $\frac{18+21}{2} = \boxed{19.5}$
- **Middle Quartile** (Q_2, or median, or 50th perc.): Value with rank $= (n+1)(\frac{1}{2})$
  here: $\frac{17+1}{2}$ = rank of 9 → $\boxed{25}$
- **Upper Quartile** (Q_3, or 75th perc.): Value with rank $= (n+1)(\frac{3}{4})$
  here: $(17+1)(\frac{3}{4})$ = rank of 13.5 → $\boxed{28.5}$
- **Interquartile Range:** $Q_3 - Q_1$
  here: $Q_3 - Q_1 = 28.5 - 19.5 = 9$
- **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**

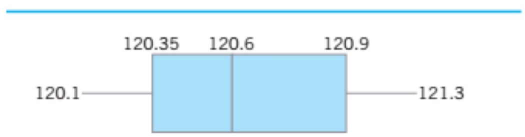
Displays 25th, 50th and 75th quartiles (ends and middle of the box)

- 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

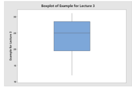
TABLE 3.4
Hole Diameters (in mm) in Wing Leading Edge Ribs

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:

Max= 31
Upper quartile (Q3, 75th percentile) = 28.5
Median is (Q2, 50th percentile): = 25
Lower Quartile (Q1, 25th percentile) = 19.5
min = 12
Whisker Limits:



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

Excel

- Stdev.s is for sample data (use most of time)
- stdev.p is for entire population

- **Histograms:** Plots of frequencies of data occurring vs. the data sorted into bins/intervals
  - ↳ **Bins:** The range that data will be sorted into # bins = ceil($\sqrt{n}$) ← round up bin size = $\frac{\max - \min}{\# \text{ of bins}}$
    - 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 value or category is its own bin

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

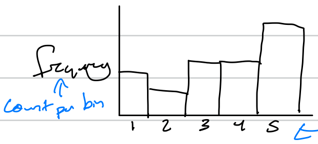
# of bins = $\sqrt{17} = 4.1 \rightarrow 5$

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

Can be fraction any round but data must not overlap

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

Bin 2: bin 1 end + bin size → 15.8+3.8 → [15.8, 19.6)



Usually should be bin limits

Bin 3: [19.6, 25.4)

Bin 4: [23.4, 27.2)

Bin 5: [27.2, 31]

↳ Final bin should be inclusive above

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)

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

102$\bar{0}$0 = 4

How would you write this # with the correct sig. figs?

0.899923 (report 2 sig. figs)
0.90

2300$\bar{0}$.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 same # of places!

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

Probability Distributions Introduction

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

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

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

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

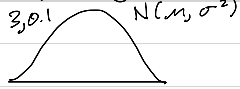
- **Calculating mean and variance**
  - o $$\mu = \begin{cases} \int_{-\infty}^{\infty} x f(x)\,dx, x \text{ continuous} \\ \sum_{i=1}^{\infty} x_i p(x_i), x \text{ discrete} \end{cases}$$
  - o $$\sigma^2 = \begin{cases} \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx, x \text{ continuous} \\ \sum_{i=1}^{\infty}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

- Median: 50th percentile
- Mean: Central Tendency (average)
- Variance: Spread of the distribution

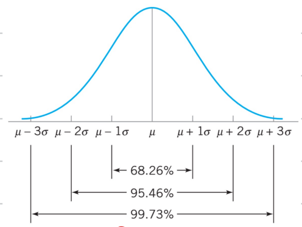
μ, σ^2 (population model params)
$\bar{x}, s^2$ (sample data notation)
Pay attention to this notation

The Normal Distribution

- For a random, normal variable x, the probability distribution of x is: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$
  - ↳ Short hand $N(\mu, \sigma^2)$



  - Continuous, symmetric, "bell-shaped"



Holds for any normally distributed variable, 'x'

General Formula: $P\{a \le x \le b\} = \int_a^b f(x)\,dx$

In a Probability Function:

$P\{(\mu - \sigma) \le x \le (\mu + \sigma)\} = 0.68$

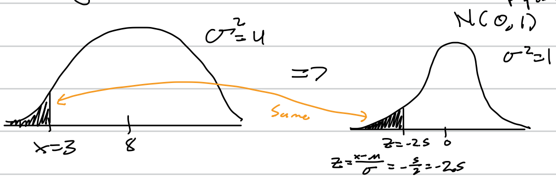
Due to symmetry:

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

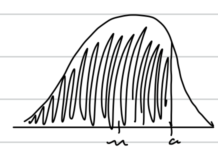
- Getting these percentages: Set $\sigma = 1$, $\mu = 0$ so $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ $\begin{cases} a = \mu - 1\sigma = -1 \\ b = \mu + 1\sigma = 1 \end{cases}$

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

- Standardization: $N(\mu, \sigma^2) \Rightarrow N(0, 1)$
- Standard Normal Distribution: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
- To standardize any random variable x: $z = \frac{x-\mu}{\sigma}$ z is a standard normal variable
- Drawing out normal distributions: Suppose $N(8, 4)$



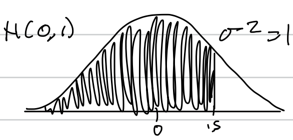
- Cummulative Normal Distribution: The probability that the variable x is less than or equal to some value a



  ↳ Standardization: $P\{x \le a\} = P\{z \le \frac{a-\mu}{\sigma}\}$ ← Go to Appendix II

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

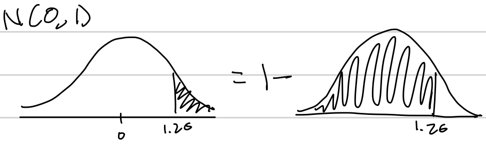
$P\{z \le 1.5\} = 0.93$

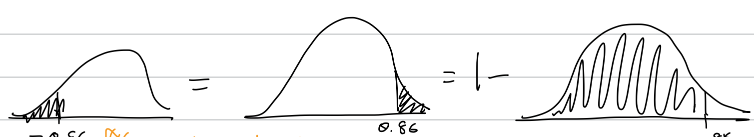


General Probability Demonstration

(i) $P(z > 1.26) = 1 - P(z < 1.26)$ (ii) $P(z < -0.86) = P(z > 0.86) = 1 - P(z < 0.86)$

$N(0,1)$

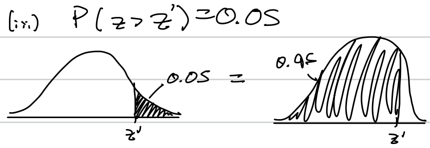




$= 1 - 0.89616 = 0.1039$

$\boxed{0.10}$

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



Chapter 4 - Hypothesis Testing

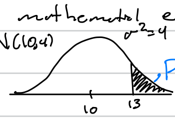
Example: Suppose the current measurements in a strip of wire are assumed to follow a normal distribution with a mean of 10 milliamps and a variance of 4 milliams^2. What is the probability that the measurement will exceed 13 milliamps?

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

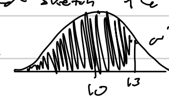
$\mu = 10$

$\sigma^2 = 4$

$P(x \geq 13)$ N(10,4) $\sigma^2=4$



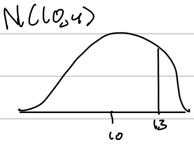
P? $= 1 -$



$\sigma^2 = 4$

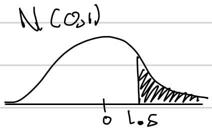
3) Convert sketch and formula to standard normal (if using tables in back of book) and a cumulative normal distribution (so it is compatible with the table in the book and Minitab)
Minitab: Calc > Probability dist. > Normal

N(10,4)



=>

N(0,1)



→ $P(z \geq 1.5)$

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

$$1 - P(z<1.5) = \boxed{0.07}$$

→ 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

12, 12, 15, 18, 21, 22, 22, 23, 25, 25, 26, 27, 28, 29, 30, 30, 31 ← Data
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 ← Rank(j)

How - to (Steps)

1) Ranking data and ordering it ($x_1, x_2 \ldots x_j \ldots x_n$), j= the ranked value, n = total number of values
2) Calculating cumulative frequencies (j-0.5)/n for each data point
   ↳ tells us the % of data that lie below a certain ranked value
   $3 (n=10) = \frac{3-.5}{10} = 0.25$

3) Find a zj value for each rank using the cumulative frequency value is the probability
Find z_j s.t. $P(z<z_j) = \frac{j-0.5}{n}$. Tells us the value of z_j if the data are normally distributed
4) Plot the actual data (x-axis) vs. corresponding zj (y-axis)

z_j data

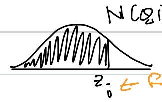
Normal Probability Plots in Minitab – Minitab: Graph > probability plot > in graph variables, choose column with desired data

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. Freq $\left(\frac{j-0.5}{n}\right)$

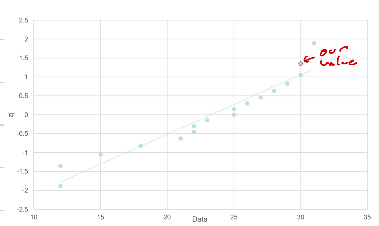
For rank 16: $\frac{16-0.5}{17} = 0.91 \rightarrow z_j$: $P(z<z_j) = 0.91$

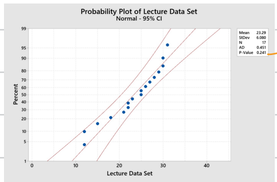
N(0,1)



z_j ← finding this z value

=> $z = 1.35$





P-value < α means not normal

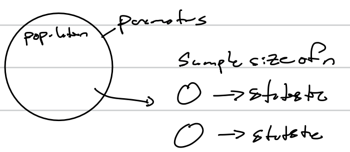
In literature,

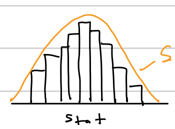
$\alpha = 0.05$

or $\alpha = 0.01$

Statistics Basically

- Statistical Inference: Drawing conclusions about a population based on a sample from that population (of them selected randomly)
- Statistics: Any function of sample data that doesn't contain unknown parameters
- Sampling Distribution: The probability distribution of a statistic





Sampling Distribution

Suppose x is a normally distributed random variable w/ given μ and σ^2

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

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

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

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

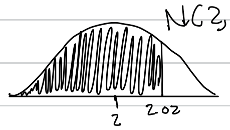
$N\left(\mu, \frac{\sigma^2}{n}\right)$ shorthand

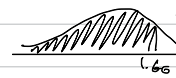
- **Central Limit Theorem:** Implies that the sample mean will be approximately normally distributed for large sample sizes, regardless of the distribution you are sampling from.
  - → 'Large' Sample size: some suggest $n > 20$ (non-normal), $n = 3-4$ (closer to normal)
- **Example** Electrodes are manufactured 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

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

$P(\bar{x} \leq 2.02)$ $N(2, \frac{0.0023}{16} = 0.00014)$ $\Rightarrow$ $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ New! For samples





$Z = 1.66$

$P(Z \leq 1.66) = 0.95$

Hypothesis Testing

- **Parameters:** What probability distributions are 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 the null hypothesis (H_1)

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

$H_0: \mu = 1.500$ in $H_1: \mu \neq 1.500$ two sided

$\begin{cases} \mu < 1.500 \\ \mu > 1.500 \end{cases}$ 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
- **Reference 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
    - ↳ Defined: $\alpha = P\{\text{type I error}\} = P\{\text{reject } H_0 \mid H_0 \text{ is true}\}$ This is specified
  - (ii) **Type II errors:** If the null hypothesis is not rejected when it is false
    - ↳ Defined: $\beta = P\{\text{type II errors}\} = P\{\text{fail to reject } H_0 \mid H_0 \text{ is false}\}$ Hard to control
  - ∘ Type II errors are uncontrolled & unknown, so we try to minimize it / avoid it
- **Conclusions:** Must be worded very clearly
  - ∘ **Rejection of Null:** A strong conclusion because we can control α to be small
  - ∘ **Failing to Reject:** Weak conclusion because it is hard to control β

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

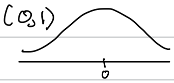
We never say accept the null hypothesis

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 (μ_o).

These are defined

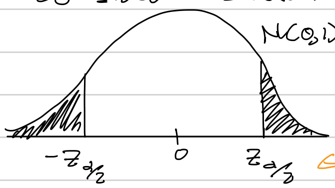
- → **Null Hypothesis:** $\mu = \mu_o$
- → **Alternate Hypothesis:** $\mu \neq \mu_o$
- → **Test Statistic:** $Z_o = \dfrac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}}$
- → **Reference Distribution:** $N(0,1)$



- → **Rejection Criterion (critical region):** Reject H_o if $|Z_o| > Z_{\alpha/2}$

Assume 0.05 unless otherwise stated

◦ Two-sided Sketch



← Critical Value

← 'acceptance' region → critical region

critical region: $Z_o > Z_{\alpha/2}$ or $Z_o < Z_{\alpha/2}$

◦ One-sided Test

$H_o: \mu = \mu_o$

$H_1: \mu > \mu_o$

← 'acceptance' region → critical region

General Procedure (for this class!)

(i) **Parameter of Interest:** From the problem context, identify known/unknown parameter of interest, the test you're going to do and identify type of problem.

(ii) **Null Hypothesis:** State the null hypothesis H_o

(iii) **Alternate Hypothesis:** State the alternative hypothesis H_a

(iv) **Test Statistic:** Determine an appropriate test statistic & reference distributions

(v) **Reject H_o if:** State rejection criterion for the null hypothesis. Should be drawn on plot w/ test statistic

(vi) **Computations:** Compute remaining sample quantities, substitute these into the equation for the test statistic, and compute that value. Compare to rejection criteria

(vii) **Draw Conclusions:** Decide whether or not Ho 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 α = 0.05. A test is conducted executing the command 25 times, and the mean response time is 79.25 ms.

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

↳ Test: One-sample z-test

2) $H_o: \mu = 75$ ms

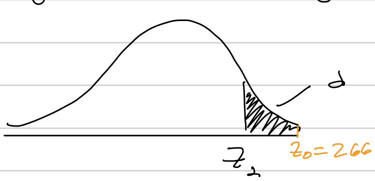
3) $H_1: \mu > 75$ ms (one-sided)

4) Test Statistic: $Z_o = \dfrac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}} = \dfrac{79.25 - 75}{\frac{8}{\sqrt{25}}} = 2.66$

Reference Dist: $N(0,1)$

5) Rejection criteria: Reject if $Z_o > Z_\alpha = Z_{0.05}$

6)



'acceptance' region → critical region

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

$Z_\alpha = 1.65$ critical value

7) Conclusion: $Z_o > Z_\alpha$

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

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

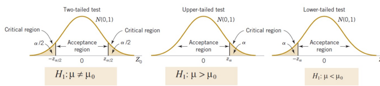
Summary of Z-tests

Testing Hypotheses on the Mean, Variance Known (Z-Tests)

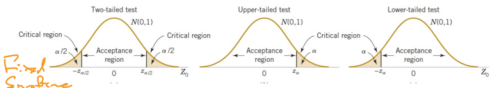
Null hypothesis: $H_0: \mu = \mu_0$

Test statistic: $Z_0 = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu \neq \mu_0$	Probability above $\lvert z_0\rvert$ and probability below $-\lvert z_0\rvert$, $P = 2\left[1-\Phi\left(\lvert z_0\rvert\right)\right]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu > \mu_0$	Probability above z_0, $P = 1-\Phi(z_0)$	$z_0 > z_\alpha$
$H_1: \mu < \mu_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

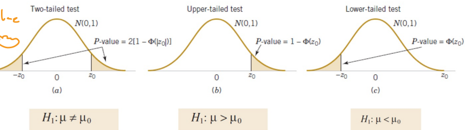


Fixed Significance Level tests



Critical Values Determined by H1 and α

P-value testing



$\Phi(z_0) = P(z < z_0)$

P-values determined by H1 and Test Statistic

Hypothesis Testing Involves

- 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

P-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

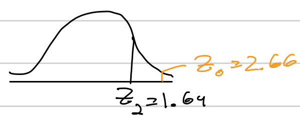
=> Reporting the compatibility of the data with the null hypothesis

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



$$P\text{-value} = 1 - \Phi(z_0)$$
$$= 1 - P\{z < z_0\}$$

Z-test Assumes

- Cts data
- Normally dist. Data
- Random Samples
- Population Std dev is known

Confidence Intervals

The interval between two statistics that includes the true value of that parameter with some probability

→ $100(1-\alpha)\%$ confidence interval

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

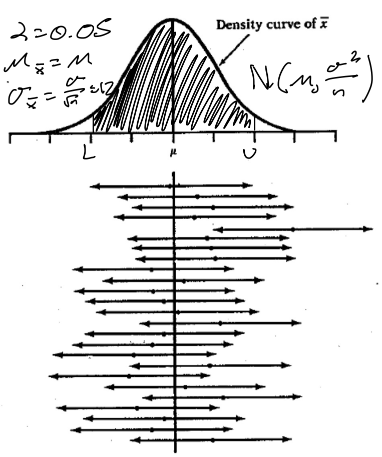
Example

$\mu = 90, \sigma = 36, n = 9$

$\alpha = 0.05$

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

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 12$



Find L and U s.t.

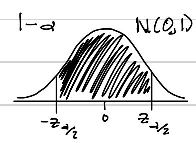
$P(L \leq 90 \leq \mu) = 0.95$

$\rightarrow L = 66.5, U = 113.5$

Assuming L, U are equidistant from the sample mean

→ L, U are 23.5 units from μ

Generally



Recall

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

$$P(L \leq \mu \leq U) = 1 - \alpha$$

$$P\left(-Z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\frac{\alpha}{2}}\right)$$

$$\rightarrow P\left(\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

L   U

CI includes infinity

One sided: $\mu \leq \bar{x} + Z_\alpha \frac{\sigma}{\sqrt{n}}$ or $\bar{x} - Z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu$

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.

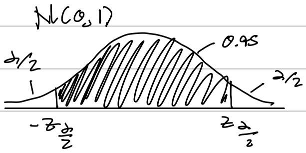
$$P\left\{\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

$n = 20$

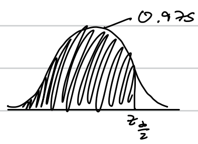
$\sigma^2 = 0.64\ mm^2 \rightarrow \sigma = 0.8$

$\bar{x} = 7.9\ mm,\ \alpha = 0.05$

$N(0,1)$



$\Rightarrow$



$z_{\frac{\alpha}{2}} = 1.96$ from tables

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

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

~~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!

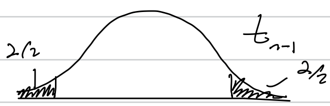
Perhaps a better description would be:

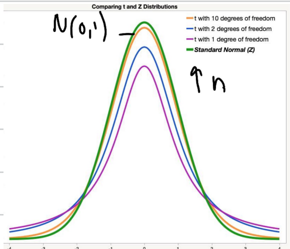
"In repeated sampling, this method will produce intervals that capture the true population mean about 95% of the time."

Use this! (might be on an exam!!)

One Sample t-test

- Inference on the mean distribution, unknown variance.
- Suppose x is a normal random variable with unknown mean μ and unknown variance σ^2
- Null Hypothesis: $\mu = \mu_0$
- Alternative Hypothesis: $\mu \neq \mu_0$
- Test Statistic: $T_0 = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$
- Reference distribution
  - ↳ t-distribution w/ n−1 degrees of freedom
- Rejection Criterion: $|t_0| > t_{\alpha/2, n-1}$ Two-sided





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

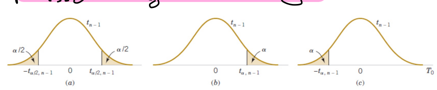
Testing Hypotheses on the Mean of a Normal Distribution, Variance Unknown

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

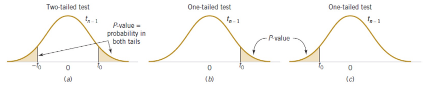
Test statistic: $T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ t dist with n−1 DOF

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 Sig Value Testing



P-Value



t-test Assumes

- Data are cts.
- 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.

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

Use this data

n = 15

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

2) $H_0: \mu = 3200\ cP = \mu_0$ 3) $H_1: \mu \neq 3200\ cP$

4) $T_0 = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$, $\bar{x} = \frac{3193 + 3124 + \ldots + 3170}{15} = 3217$

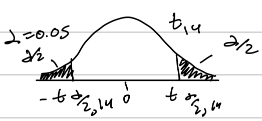
Plug & Chug $= 0.64$ $S = \sqrt{\frac{(3193-3217)^2 + \ldots + (3170-3217)^2}{n-1}} = 106$

5) Reject if $\begin{cases} t_0 > t_{0.025,14} \\ t_0 < -t_{0.025,14} \end{cases}$

7) Conclusion

(Opt 1) The evidence is not sufficient to reject the claim that the mean is equal to 3200 cP

(Opt 2) There is insufficient evidence to support the mean not being equal to 3200 cP.



$t_{\alpha/2} = 2.145$

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:

$$\bar{x} - t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,\, n-1}\frac{s}{\sqrt{n}}$$ **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 $P(t_{n-1} \ge t_{\alpha/2,\, n-1}) = \alpha/2$

$$-\infty \le \mu \le \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$ **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 $P(t_{n-1} \ge t_{\alpha,\, n-1}) = \alpha$

$$\bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} \le \mu \le \infty$$ **Is a 100(1-alpha)% lower confidence bound**

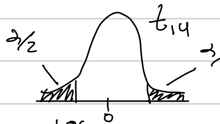
where $t_{\alpha,\, n-1}$ is the percentage point of the t distribution with n-1 degrees of freedom such that $P(t_{n-1} \ge t_{\alpha,\, n-1}) = \alpha$

→ For our previous example

$\bar{x} = 3217$

$s = 106$

$n = 15$



$t_{0.025,14} = 2.145$

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

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

$\Big\}\; 3158 \le \mu \le 3276$

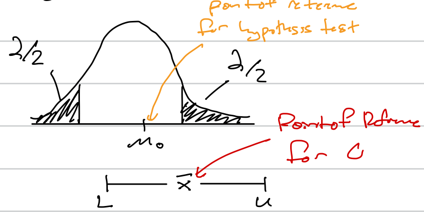
Minitab: Stat>Basic Stat>1-sample t-test, samples in columns (drop down), double click data, check hypothesis test select type and in options select alpha value

CI Relationship w/ Hypothesis Testing

- Both do make interfaces 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

Illustration

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



Inference on the variance of Normal Dist. (1-variance 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, … x_n be a random sample of n observations from this population to test:

Useful for making conclusions about the variability of your process

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

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

Test Statistic: $\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$ s = sample variance with n observations

Reference Distribution: Chi-Squared Distribution

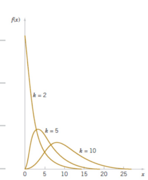
Which Hypothesis should I do?

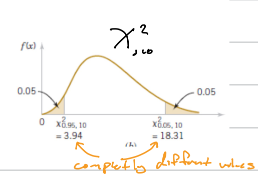
Test	# of Populations	Variance Known?	Parameter of interest?	Reference Dist.
One-sample Z-test	1	Yes	Mean	Standard normal
Two-sample Z-test	2	Yes	Difference between means	Standard normal
One-sample t-test	1	No	Mean	t-distribution
Two-sample (pooled) t-test (equal variance)	2	No	Means of two populations	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 the population are known, it's a probability problem

Chi-squared Distribution

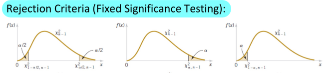
- χ_0^2 has a chi-squared distribution if the null is true
- Key features:
  - (i) Depends on n-1 degrees of freedom
  - (ii) Not symmetric
  - (iii) Not centered @ 0
  - (iv) Total Area under curve is still 1!





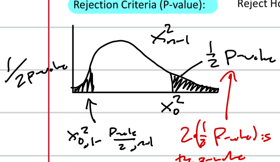
Rejections

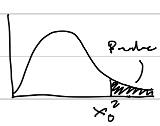
Rejection Criteria (Fixed Significance Testing):



$H_1: \sigma^2 \neq \sigma_0^2$	$H_1: \sigma^2 > \sigma_0^2$	$H_1: \sigma^2 < \sigma_0^2$
Reject Ho if:		
$\chi_0^2 > \chi^2_{\alpha/2,n-1}$ or $\chi_0^2 < \chi^2_{1-\alpha/2,n-1}$	$\chi_0^2 > \chi^2_{\alpha,n-1}$	$\chi_0^2 < \chi^2_{1-\alpha,n-1}$

Rejection Criteria (P-value): Reject Ho if p-value < alpha





Summary

Null hypothesis:	$H_0: \sigma^2 = \sigma_0^2$
Test statistic:	$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$

Ref. Dist: Chi-Square

Alternative Hypothesis	Rejection Criteria
$H_1: \sigma^2 \neq \sigma_0^2$	$\chi_0^2 > \chi^2_{\alpha/2,n-1}$ or $\chi_0^2 < \chi^2_{1-\alpha/2,n-1}$
$H_1: \sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi^2_{\alpha,n-1}$
$H_1: \sigma^2 < \sigma_0^2$	$\chi_0^2 < \chi^2_{1-\alpha,n-1}$

Check Normality! It's 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.

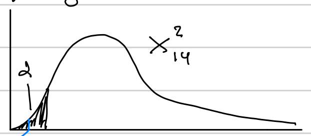
Use $\alpha = 0.05$

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

1) Parameter of interest: σ^2 → 1-variance test
2) $H_0: \sigma_0^2 = 30{,}000\ cP^2$
3) $H_a: \sigma_0^2 < 30{,}000\ cP^2$
4) $\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(15-1)(106)^2}{30{,}000} = 5.2$

Reference Distribution: χ^2_{14}

5) Rejection Criteria: Reject if $\chi_0^2 < \chi^2_{1-\alpha, n-1}$



$\chi_0^2 = 5.2$ $\quad \chi^2_{1-0.05,14} = 6.57$ ← critical value

critical region 'acceptance' region

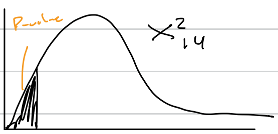
Fixed-Significance Level

6) Computations: $5.2 < 6.57$
7) Conclusion: Reject H_0. We have evidence to reject the claim that the variance is 30,000cP, having evidence that the viscosity that variance is less than 30,000cP.

We accept H_1. ← Can accept H_1 but not accept H_0

p-value

Reject if: p-value < α



5.2 lands btwn 0.990 and 0.975

→ p-value is btwn 1-tails

(0.01 < p < 0.025) → both < α so reject H_0!

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

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

Suppose x is a normal random variable with unknown mean μ and unknown variance σ². From a sample of n observations, the sample mean $\bar{x}$ and sample variance s² are computed. Then:

$(100-\alpha)\%$ two-sided CI

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

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

$(100-\alpha)\%$ upper bound

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

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

$(100-\alpha)\%$ lower bound

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

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

For our example (upper bound)

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

$$\chi^2_{1-\alpha,n-1} = \chi^2_{0.95,14} = 6.57$$

$$\Rightarrow \sigma^2 \le \frac{(14)(106)^2}{6.57} = 239.42\ cP^2$$

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 σ ≤ 173 (from our hypothesis test) and we assume the mean of the population is on target?

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

$$\text{Sigma level} = \frac{|\text{spec limit} - \text{target}|}{\sigma} = \frac{1100}{173}$$

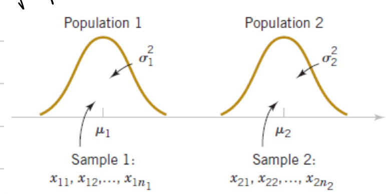
$= 6.35$ → at least 6.35

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 their parameters

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

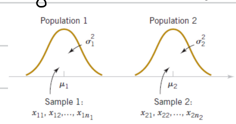
Null: $H_0: \mu_1 - \mu_2 = \cancel{\Delta_0}\ 0$

Alt: $H_1: \mu_1 - \mu_2 \neq 0$ (two-sided)

Test statistic: $Z_0 = \dfrac{\bar{x}_1 - \bar{x}_2 - \cancel{\Delta_0}\ 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Reference Dist: N(0,1)

Rejection Criteria: Same as one-sample z-test



Summary

Ref: $N(0,1)$

Null hypothesis: $H_0\colon \mu_1 - \mu_2 = \Delta_0$

Test statistic: $$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (10\text{-}2)$$

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

Assumptions

(1) $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
(2) $X_{21}, X_{22}, \ldots, X_{2n_2}$ 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

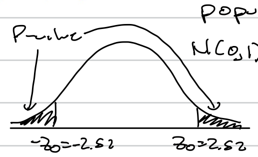
Paint Drying Time A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the 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 $\alpha = 0.05$?

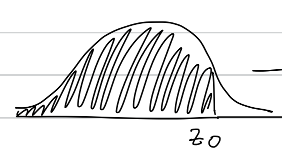
1) Parameter of interest is the difference in the mean drying time ($\mu_1 - \mu_2 = 0$), and we know the standard deviation is 8 min for both populations. → two-sample z-test

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

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

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





→ 0.99413

5) P-value: Reject H_0 if p-value < $\alpha = 0.05$

p-value = 2(1 − 0.99143) = 0.012

6) P-value = 0.012 < $\alpha = 0.05$

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

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 $P(z \geq Z_{\alpha/2}) = \alpha/2$

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

100 (1-α)% one-sided upper confidence bound on $\mu_1 - \mu_2$ (interval includes - infinity)

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

100 (1-α)% one-sided lower confidence bound on $\mu_1 - \mu_2$ (interval includes + infinity)

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

Two-Sample t-test (variances unknown)

Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ← unknown values, 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:

Null: $H_0: \mu_1 - \mu_2 = \cancel{\Delta_0}\ 0 \quad (\mu_1 = \mu_2)$

Alt: $H_a: \mu_1 - \mu_2 \neq \cancel{\Delta_0}\ 0 \quad (\mu_1 \neq \mu_2)$

Test statistic: $$T_0 = \frac{\bar{x}_1 - \bar{x}_2 - \cancel{\Delta_0}\ 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Square root of this

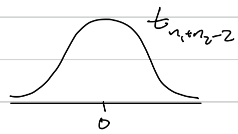
Pooling Sample Variance

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

Weighted average of two sample variances

Reference Distribution:

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

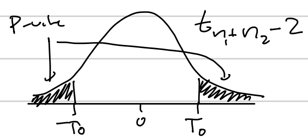


Rejection Criteria: See below exercise

Exercise: Sketch out Fixed rejection criteria & p-value for two-sided H_1

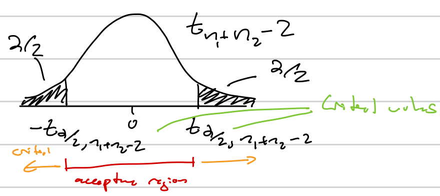
P-value:

Reject H_0 if p-value $< \alpha$



Fixed:

Reject if $T_0 > t_{\alpha/2, n_1+n_2-2}$ 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
	$\bar{x}_1 = 92.255$	$\bar{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

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

1) Interested in difference between 2 means, specifically yield of catalyst 1 and catalyst 2, variance unknown → two-sample t-test

↳ Assume equal variance ← rule of thumb later

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

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

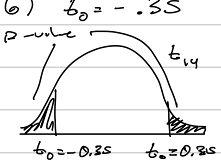
4) $$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{92.255 - 92.733}{2.7\sqrt{\frac{1}{8}+\frac{1}{8}}} = -0.35$$

$$S_p^2 = \frac{(n_1-1)S_1^2 - (n_2-1)S_2^2}{n_1+n_2-2} = \frac{7\cdot 2.39^2 + 7\cdot 2.98^2}{8+8-2} = 7.3$$

$$\rightarrow S_p = \sqrt{S_p^2} = \sqrt{7.3} = 2.7$$

5) Reject if p-value $< \alpha = 0.05$

6) $t_0 = -.35$



7) p-value $> \alpha$

Fail to reject H_0, we don't have enough evidence to say the yield from catalyst 1 is different from catalyst 2.

$0.5 <$ p-value $< .8$

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:

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 Dist: t-dist with v degrees of freedom

where $v = \dfrac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$

If v is not an integer, round down

Example v calculation: $s_1 = 7.63$, $s_2 = 15.3$, $n_1 = n_2 = 10$ → $v = 13.2 \to 13$

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

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$.

Test statistic: $$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (10\text{-}14)$$

Ref. : t-dist with v dof

Alternative Hypotheses	P-Value	Rejection Criterion for Fixed-Level Tests
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert t_0 \rvert$ and probability below $-\lvert t_0 \rvert$	$t_0 > t_{\alpha/2,\, n_1+n_2-2}$ or $t_0 < -t_{\alpha/2,\, n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probability above t_0	$t_0 > t_{\alpha,\, n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probability below t_0	$t_0 < -t_{\alpha,\, n_1+n_2-2}$

Note: replace n1+n2 with v in rejection criteria!

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

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

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

Assumptions

(1) $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
(2) $X_{21}, X_{22}, \ldots, X_{2n_2}$ 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 1

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 on the two populations are observed in pairs where each pair of observations is taken under homogenous conditions

To compare differences, we define Dj:

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

where Dj's are assumed to be normally distributed with mean u_D and variance σ^2_D

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

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

$H_0: \mu_D = 0 = \Delta_0$

$H_a: \mu_D \neq \Delta_0 = 0$

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

where $\bar{D}$ = Sample average of n differences; S_D = Sample standard deviation of n differences

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

Rejection Criterion: Same as one sample t-test

Summary

Null hypothesis: $H_0\colon \mu_D = \Delta_0$ Ref: t-dist w/ $n-1$ dof

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

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

Minitab: stat -> basic stat-> paired t

or input differences data into one-sample t-test

Example

TABLE 4.3 (n=8)
Paired Tensile Strength Data for Example 4.11

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

Get $\bar{D}$ and s_D from these differences

Mean difference: -1.38 $\bar{D}$
Standard deviation of differences: 2.67 s_D
Alpha = 0.05

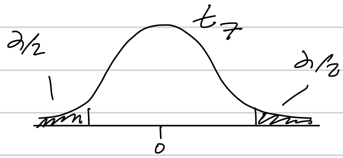
Do these machines yield the same tensile strength values on the same sample?

(i) Interested in the difference in mean tensile strength measured by two devices, same samp → paired data ⇒ paired t-test

(ii) $H_0: \mu_D = 0$ (iii) $\mu_D \neq 0$

(iv) $t_0 = \dfrac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \dfrac{-1.38}{\frac{2.67}{\sqrt{8}}} = -1.46$

(v) Reject if $t_0 > t_{\alpha/2,n-1}$ or $t_0 < -t_{\alpha/2,n-1}$



$\rightarrow t_{\alpha/2,n-1} = 2.365$

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

End of Exam 1 Material

Exam 1 Summary

Which Hypothesis 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 ± 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 ± 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=test or 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 iPad 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

Presenting Conclusions

ECHE 313 Practice Exam 1 Solutions on canvas

Write your name:

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

don't pay attention to this, this exam was take-home

""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 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 formals you want to use first, then write in your subbed in 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 α=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.

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

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 a 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 (or you can say their 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)

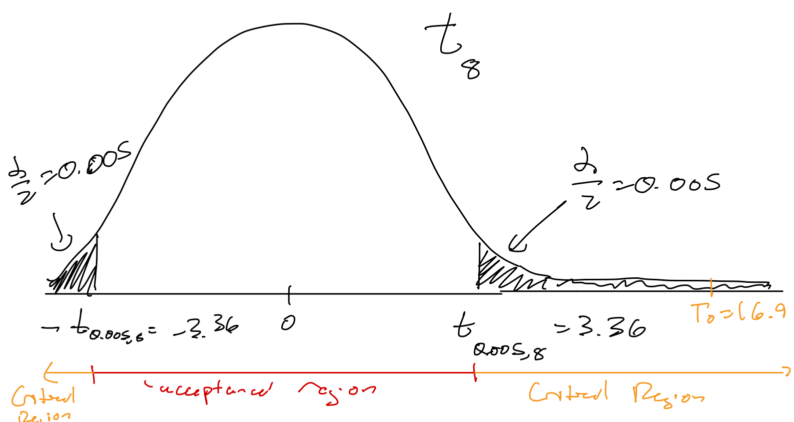
$$T_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{262 - 149}{10.62\sqrt{\frac{1}{5} + \frac{1}{5}}} = 16.9$$

$S_p = 10.62$

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

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

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.

$T_0 < t_{0.005,8}$

Reject Ho, Accept Ha, Context → Moscow, news lower!

See key

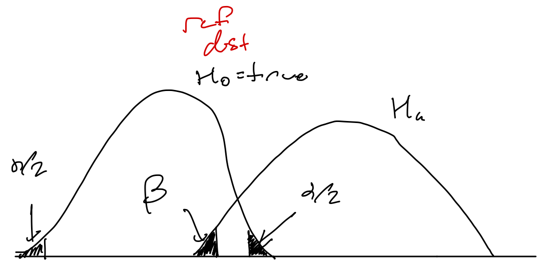
i. (3 points) Explain what plot you would construct to check if the new bioreactor data are normally distributed (what feature(s) of the plot would you check?)

Normal probability plot, see if its linear (looks distinct non-linear shape)

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

α is Type I error =

See Key



Can precisely define
save alpha is

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
Sample Ave.	815.75
Sample Standard Deviation:	6.31

$n=12$

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

$H_o: \mu = 810$ mg/L

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

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

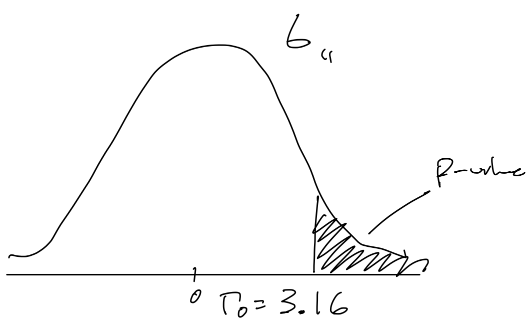
e. (3 points) Write the formula of the test statistic and calculate (report 3 sig figs)

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

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

Reject if p-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.



See Key

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?)

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.

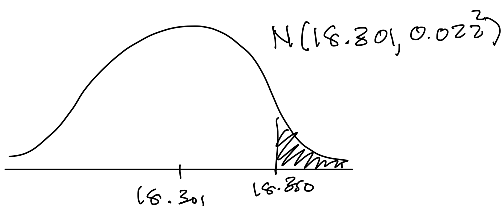
One-sample z-test, and 2-sided due to different

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) A 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.

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



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 and appropriate number of sig figs based on the numbers given in the problem statement.

$$Z = \frac{16.350 - 16.301}{0.022} = 2.23$$

$\rightarrow$ P-value ≈ 0.013

Midterm Exam 2

Chapter 4 Continued - Test 2 Material

Inference on Two populations' variances

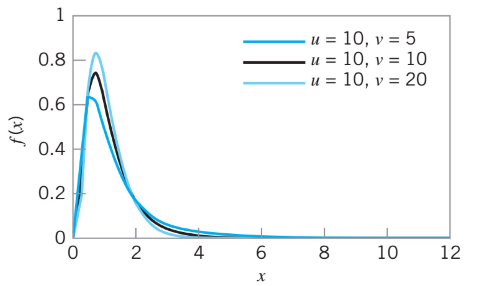
Sensitive to normality – should check!

→ Two independent, normally distributed populations

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

Alt: $H_a: \sigma_1^2 \neq \sigma_2^2$ two sided

Test Stat: $F_0 = \dfrac{S_1^2}{S_2^2}$ ← DoF$_1 = n_1 - 1$ ← DoF$_2 = n_2 - 1$



Reference distribution: F-distribution

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

(ii) Different curve for each set of DoFs

(iii) F statistic ≥ 0

(iv) As DoF increase curve gets more normal

Defined by two DoFs

Note: $F_{0.99, \nu_1, \nu_2} = 1 / F_{0.01, \nu_2, \nu_1}$

Rejection Criteria: (two-sided)

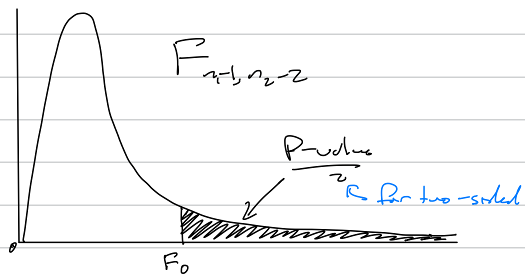
P-value

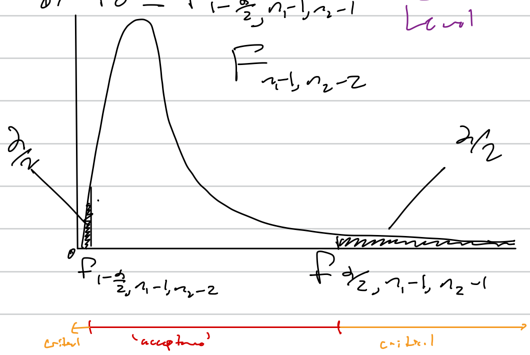
P-value < α

$f_0 \geq f_{\alpha/2, n_1-1, n_2-1}$

or $f_0 \leq f_{1-\frac{\alpha}{2}, n_1-1, n_2-1}$

Fixed Signif. Level





Summary

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

Test statistic: $F_0 = \dfrac{S_1^2}{S_2^2}$ (10-31)

Alternative Hypotheses	Rejection Criterion
$H_1: \sigma_1^2 \neq \sigma_2^2$	$f_0 > f_{\alpha/2, n_1-1, n_2-1}$ or $f_0 < f_{1-\alpha/2, n_1-1, n_2-1}$
$H_1: \sigma_1^2 > \sigma_2^2$	$f_0 > f_{\alpha, n_1-1, n_2-1}$
$H_1: \sigma_1^2 < \sigma_2^2$	$f_0 < f_{1-\alpha, n_1-1, n_2-1}$

Assumptions

Grab from posted notes

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 $\alpha = 0.05$.

Minitab: Stat-> Basic Stat -> Two-sample Variance

1) Parameter of interest: Variance in oxide thickness for two mixtures of gases, unknown variances

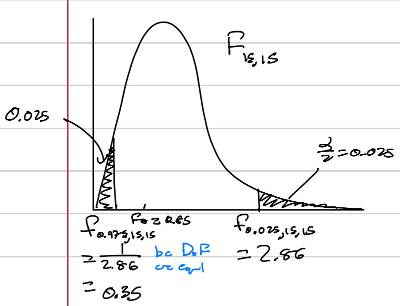
∴ Two-variance test

2) $H_0: \sigma_1^2 = \sigma_2^2$

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

4) $F_0 = \dfrac{S_1^2}{S_2^2} = \dfrac{1.96^2}{2.13^2} = 0.85$

5) Reject if: $F_0 > f_{0.025, 15, 15}$ 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 mixtures 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 gather data such that nuisance variables are eliminated

Factor → Hardwood Concentration (%); levels → 5, 10, 15, 20; Repeats → Observations

Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations 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

# of...
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}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	...	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
.	.	.	...	.	.	.
.	.	.	...	.	.	.
.	.	.	...	.	.	.
a	y_{a1}	y_{a2}	...	y_{an}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

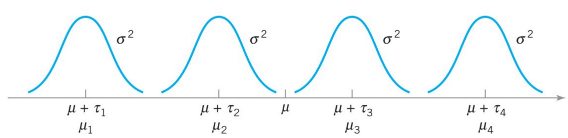
$y_{\cdot\cdot}$ (grand total) $\bar{y}_{\cdot\cdot}$ (grand mean)

a = # of treatments
y_{ij} = j^{th} observation of i^{th} treatment
n = # of observations at each level

Linear Statistical Model

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

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



Hypothesis Testing the Equality of Treatment Effects

Null: $H_0: \tau_1 = \tau_2 = \tau_3 = \ldots = \tau_a = 0$
Alt: $H_a: \tau_i \neq 0$ for at least one i
Test Statistic: $F_0 = \dfrac{SS_{treatments}/a-1}{SS_E/a(n-1)}$ where $\begin{cases} a = \text{\# of treatments} \\ n = \text{\# of replicants @ each treatment level} \end{cases}$
Reference Distribution: F-dist w/ $(a-1)$ [top] and $a(n-1)$ [bottom] DoF
Rejection Criteria: H_0 if $F_0 > F_{\alpha, a-1, a(n-1)}$ 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_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}}$	$\dfrac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n-1)$	MS_E	
Total	SS_T	$an-1$		

$$F_0 = \frac{MS_{trt}}{MS_E} = \frac{SS_{trt}/DoF}{SS_E/DoF}$$

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

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

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

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{i\cdot} = y_{i\cdot}/n \qquad i = 1,2,\ldots,a$$

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

Defining & Development of ANOVA

a=number of treatments

y_{ij} = the j 'th observation of the i 'th treatment

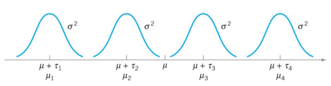
n = number of observations for each level

Typical Data for a Single-Factor Experiment

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

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

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



Hypothesis testing the equality of treatment effects

Null and Alternative Hypothesis:

$H_0: \tau_1 = \tau_2 = \cdots = \tau_a = 0 \qquad H_1: \tau_i \neq 0$ 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

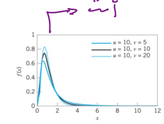
DOF

signal / noise

Rejection Criteria:

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

f_{v_1, v_2}



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	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}}$	$\frac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n-1)$	MS_E	
Total	SS_T	$an-1$		

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

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

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

$y_{i\cdot} = \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{i\cdot} = y_{i\cdot}/n \qquad i = 1,2,\ldots,a$

$y_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/N$ → Total number of observations

General

i = treatment

j = observation

bar = average

dots = total

n = replicates

a = # of treatments

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

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

$y_{\cdot\cdot}$ = total of all observations

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

What does the Calculation look like

Tensile Strength of Paper (psi)

Factor

Treatment Level, $a=4$

Hardwood Concentration (%)	1	2	3	4	5	6	Totals ($y_{i\cdot}$)	Averages ($\bar{y}_{i\cdot}$)
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 ($y_{\cdot\cdot}$)	15.96 ($\bar{y}_{\cdot\cdot}$)

Observations, $n=6$, y_{13}

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

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

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

error

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

all data points

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

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

$\Rightarrow SS_E = 518 - 382.8 = 130.2$

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

1) Parameter of Interest: Multiple pop. means of tensile strength → ANOVA!

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

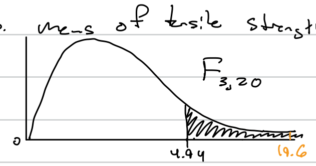
3) $\exists\, i \in [1,4]$ s.t. $\tau_i \neq 0$

4) $F_0 = \dfrac{\frac{SS_{\text{Treatments}}}{(a-1)}}{\frac{SS_E}{a(n-1)}} = \dfrac{\frac{512.96}{3}}{\frac{130.2}{20}} = 19.6$

5) Reject if $F_0 > F_{\alpha, a-1, a(n-1)}$

$F_{0.01, 3, 20} = 4.94$ from charts

6)



Minitab: Stat -> ANOVA -> one-way

7) 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

Fischer's Least Significant Difference (LSD) Test:

Declare Significantly Different if: $|\bar{y}_{i.} - \bar{y}_{j.}| > LSD$

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

Example

Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations 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)

5 vs. 10 : $|10 - 15.67| = 5.67^*$

5 vs. 15 : $|10 - 17| = 7^*$

5 vs. 20 : $= 11.17^*$

10 vs 15 : $= 1.33$

10 vs. 20 : $= 5.5^*$

15 vs. 20 : $= 4.17$

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

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

$= 4.19$

20 : C

15 : C B

10 : B

5 : A

15 & 20 is ok

10 vs 15 is ok

Different treatments

Assumptions

(i) Normal

(ii) No hidden variables

Summary

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 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	17.000	A	B	
10%	6	15.67		B	
5%	6	10.00			C

Fischer's LSD Test

Declare significantly different if: $|\bar{y}_{i.} - \bar{y}_{j.}| > \text{LSD}$

where LSD, the least significant difference, is

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

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

Mini tab 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

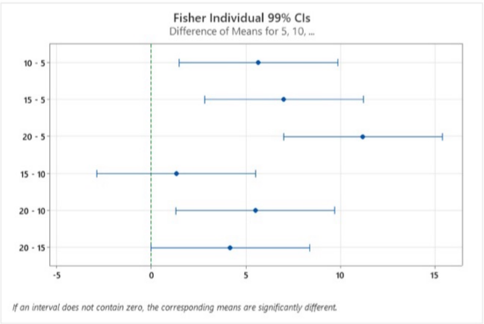
Confidence Intervals

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

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

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

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



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

$y_{ij} = \mu_i + \varepsilon_{ij}$
obs. estimated

$\varepsilon_{ij} = y_{ij} - \bar{y}_{i.}$
residual obs

Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations 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

Residual

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

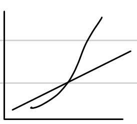
Hardwood Concentration (%)	Residuals					
5	–3.00	–2.00	5.00	1.00	–1.00	0.00
10	–3.67	1.33	–2.67	2.33	3.33	–0.67
15	–3.00	1.00	2.00	0.00	–1.00	1.00
20	–2.17	3.83	0.83	1.83	–3.17	–1.17

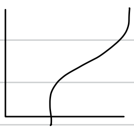
Normal Probability Plot of Residuals

(i) Rank data
(ii) Calc cummulative frequency
(iii) Find z_j's
(iv) Plot residuals vs. z_j's
Looking for straight line

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





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

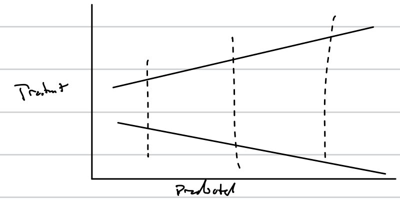
Can also generate in Excel or Minitab: Residuals in one column, predicted value in the other, using scatter plot

Hardwood Concentration (%)	Residuals					
5	–3.00	–2.00	5.00	1.00	–1.00	0.00
10	–3.67	1.33	–2.67	2.33	3.33	–0.67
15	–3.00	1.00	2.00	0.00	–1.00	1.00
20	–2.17	3.83	0.83	1.83	–3.17	–1.17

Tensile Strength of Paper (psi)

Hardwood Concentration (%)	Observations 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

What if there is a pattern?
- Data transformation
- Ratio 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 (or other)

(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

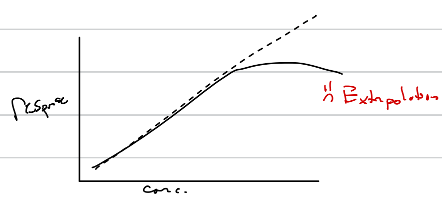
Regression Modeling

- **Regression Modeling:** A mathematical model fit to a set of data relating a response variable (y) to an independent variable ($x_1, x_2, \ldots, x_n$)
- **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

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

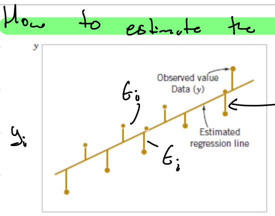
y = expected response
β_0 = intercept
β_1 = slope
regression coef.

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**



$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Method of Least squares

Squares the deviations (ϵ_i)

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \quad \begin{matrix} S_{xy} \\ S_{xx} \end{matrix} \tag{11-8}$$

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

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

Raw data   residual

Example: Fit a simple linear regression line to these data

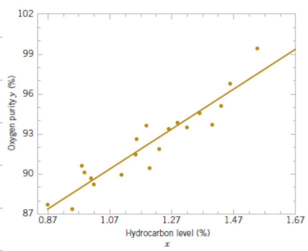
TABLE • 11-1 Oxygen and Hydrocarbon Levels

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

$n = 20$

$\sum x_i = 23.42$ $\quad \sum y_i = 1843.21$ $\quad \sum x_i y_i = 2214$

$\sum x_i^2 = 29.28$ $\quad \sum y_i^2 = 170044$



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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\overbrace{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}^{S_{xy}}}{\underbrace{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}_{S_{xx}}} \tag{11-8}$$

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

$$S_{xx} = 29.28 - \frac{(23.42)^2}{20} = 0.68$$

$$S_{xy} = 2214 - \frac{23.42(1848)}{20} = 10.18$$

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

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

$$= 74.28$$

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

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