Chapter 4 Continued - Test 2 Material

Inference on Two populations' variances

→ Two independent, normally distributed populations

Sensitive to normality – should check!

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

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

Test Stat: $F_0 = \frac{S_1^2}{S_2^2}$ ← $DoF_1 = n_1 - 1$, ← $DoF_2 = n_2 - 1$

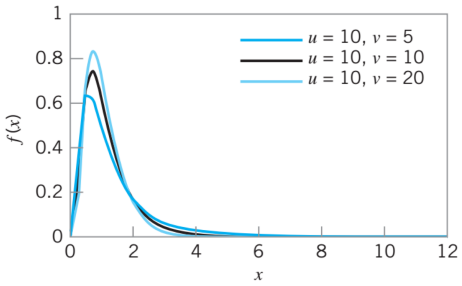
Reference distribution: F - distribution

(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



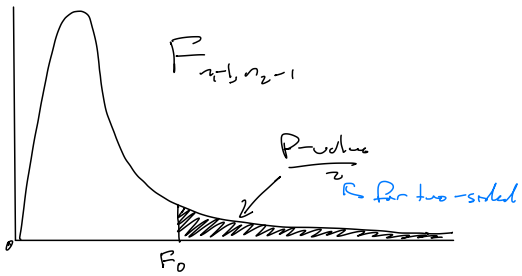
Defined by two DoFs

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

Rejection Criteria: (two-sided)

P-value

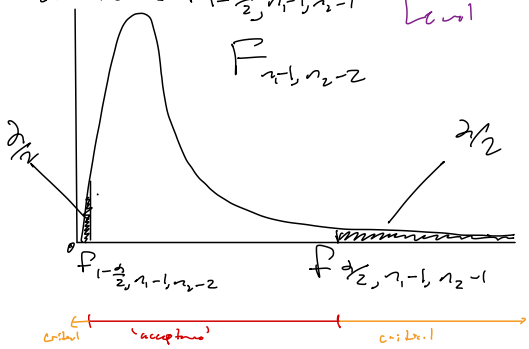
P-value < α



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

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:	$F_0 = \frac{S_1^2}{S_2^2}$ (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

- The data are normally distributed
- Samples are independent

Sensitive to normality!

Example

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

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

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

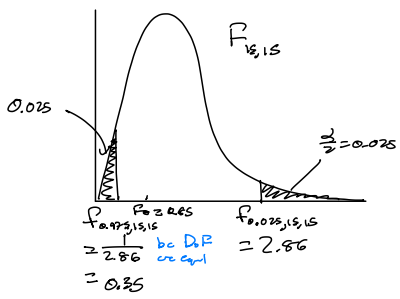
∴ Two-variance test

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

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

4) $F_0 = \frac{s_1^2}{s_2^2} = \frac{1.902}{2.132} = 0.85$

5) Reject if: $F_0 > f_{0.025,15,15}$ or $f_0 < f_{0.975,15,15}$



Fail to reject H_0, medals have enough evidence to support the claim the variance 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 or 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

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

Factor → Hardwood Concentration (%); Levels → 5, 10, 15, 20; Replicates → Observations 1–6

# 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}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

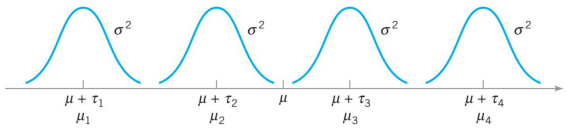
$y_{\cdot\cdot}$ (grand total) $\bar{y}_{\cdot\cdot}$ (grand mean)

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

Linear statistical Model

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

where τ_i = i^{th} treatment effect
ε_{ij} = random error



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: $F_0 = \frac{SS_{treatments}/a-1}{SS_E/a(n-1)}$ where a = # of treatments, n = # of replicants @ each treatment level

Reference Distribution: F-dist ~ $(a-1)$[top] and $a(n-1)$ [bottom] DoF

Rejection Criteria: H_0 if $F_0 > F_{\alpha, a-1, a(n-1)}$ this 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	$SS_{Treatments}$	$a-1$	$MS_{Treatments}$	$\frac{MS_{Treatments}}{MS_E}$
Error	SS_E	$a(n-1)$	MS_E	
Total	SS_T	$an-1$		

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

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

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

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

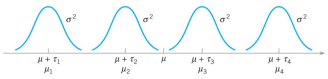
n = number of observations for each level

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

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



Null and Alternative Hypothesis:

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

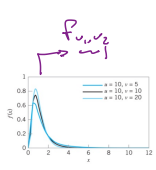
Test Statistic: (reference distribution is the F-distribution)

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

Signal / noise

Rejection Criteria:

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



Hypothesis testing the equality of treatment effects

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	$an-1$		

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

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

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

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

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

Terms

i = treatment
j = observation
bar = average
dots = total
n = replicates
a = # of treatments
y_{ij} = j^{th} observation of i^{th} treatment
$\bar{y}$ = grand mean of all observations
y = total of all observations
$y_{i.}$ = average of observations @ treatment i

What does the Calculation look like

Tensile Strength of Paper (psi)

Factor: Hardwood Concentration (%)	1	2	3	4	5	6	$y_{i.}$ Totals	$\bar{y}_{i.}$ Averages
1	7	8	15	11	9	10	60	10.00
2	10	12	17	13	18	19	15	94	15.67
3	15	14	18	19	17	16	18	102	17.00
4	20	19	25	22	23	18	20	127	21.17
								383	15.96
								$y_{..}$	$\bar{y}_{..}$

Treatment Level: 1, 2, 3, 4 ($a=4$); $n=6$; y_{13} = 15

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

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

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

error

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

all data points

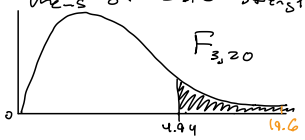
$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2 = (7-15.96)^2 + (8-15.96)^2 + \ldots + (20-15.96)^2 = 513$$

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

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

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) $F_0 = \dfrac{\frac{SS_{\text{Treatments}}}{(a-1)}}{\frac{SS_E}{a(n-1)}} = \dfrac{\frac{382.96}{3}}{\frac{130.2}{20}} = 19.6$
5) Reject H_0 if $F_0 > F_{\alpha, a-1, a(n-1)}$
   $F_{0.01,3,20} = 4.94$ from charts
6)



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!

Minitab: Stat -> ANOVA -> one-way

What do you do when the null is rejected

Multiple Comparisons 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

$$LSD = t_{0.005,20}\sqrt{\frac{2MS_E}{n}} = 2.845\sqrt{2\left(\frac{6.5}{6}\right)} = 4.19$$

20: C
15: C B
10: B
5: A

10 vs 20 is ok
10 vs 15 is ok
Different therapy

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

Summary

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

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

Assumptions

(i) Normal
(ii) No hidden variables

Fischer's LSD Test

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

where LSD, the least significant difference, is

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

Better notes on LSD test:

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

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

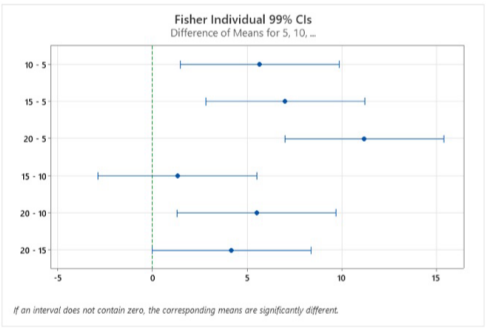
Confidence Intervals

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

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

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

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



ANOVA Assumptions

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

$y_{ij} = \mu_i + \epsilon_{ij}$ (obs. = estimated)

$E_{ij} = y_{ij} - \bar{y}_{i\cdot}$ (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

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

Residual

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

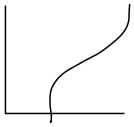
Normal Probability Plot of Residuals

(i) Rank data
(ii) Calc cumulative 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!

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

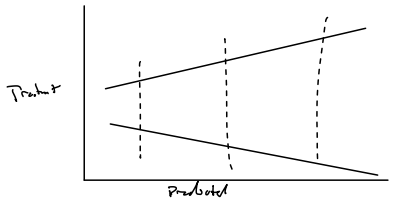
Hardwood Concentration (%)	Residuals					
5	−3.00	−2.00	5.00	1.00	−1.00	0.00
10	−3.67	1.33	−2.67	2.33	3.33	−0.67
15	−3.00	1.00	2.00	0.00	−1.00	1.00
20	−2.17	3.83	0.83	1.83	−3.17	−1.17

Tensile Strength of Paper (psi)

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

What if there is a pattern?

- Data transformation
- Redo 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, plot...)

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

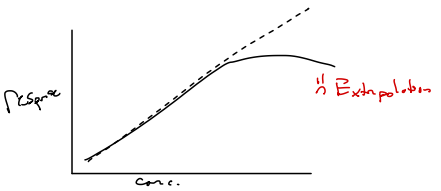
Regression Modeling

- **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

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

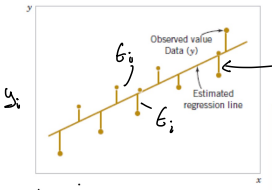
y = expected response
β_0 = intercept
β_1 = slope
} regression coef.

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**



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

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

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

The least squares estimates of the intercept and slope in the simple linear regression model are (Avg of all response values → Avg of all x's) (11-7)

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

S_{xy} / S_{xx}

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

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

Fitted or Estimated Regression Line

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

↑ Predicted Value

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

Each Data ↑ ↑ Residual

← Each Pair Satisfies this

Example: Fit a simple linear regression line to these data

TABLE • 11-1 Oxygen and Hydrocarbon Levels

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

$n = 20$

$\sum x_i = 23.92$ $\quad \sum y_i = 1843.21$ $\quad \sum x_i y_i = 2214$

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

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

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

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

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

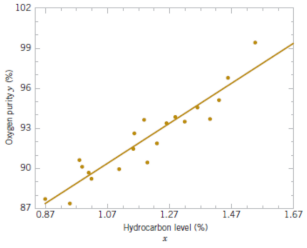
$$S_{xx} = 29.28 - \frac{(23.92)^2}{20} = 0.68$$

$$S_{xy} = 2214 - \frac{23.92(1843)}{20} = 10.18$$

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

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

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



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

ANOVA for Linear Regression

- Null & Alternative Hypotheses:
  $H_0: \beta_1 = 0$
  $H_1: \beta_1 \neq 0$
- Test Statistic:

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

- Reference Distribution: F_{dist} ↳ Units: $F_{1,n-2}$
- Rejection Criterion:
  Reject H_0 if $F_0 > f_{\alpha,1,n-2}$ or p-value $< \alpha$

Calculating the Error

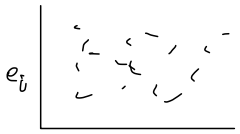
Recall: $SS_T = SS_R + SS_E$

- Residual: $e_i = y_i - \hat{y}_i$ (residual, observed, predicted)
- Error Sum of Squares: $SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- Mean Squared Error: $SS_E / n-2$
- Sum of Squares Total: $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- Regression Sum of Squares: $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Checking Assumptions

1) Errors (and observations) are normally distributed
2) Errors (and observations) are independently distributed
3) Variance is constant

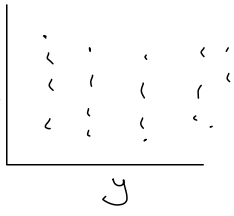
$e_i = y_i - \hat{y}_i \quad i = 1, 2, \ldots, n$ (residual, observed response, estimated value)



Linear? Normality — Plot residuals on Normal Probability Plot

Patterns? Independent Variance — Plot residuals vs. Run order or Business variables

Patterns or Spread? Constant Variance — (1) Plot residuals vs. predicted (common) (2) Plot vs. independent variable (x_i)



Example

TABLE • 11-1 Oxygen and Hydrocarbon Levels

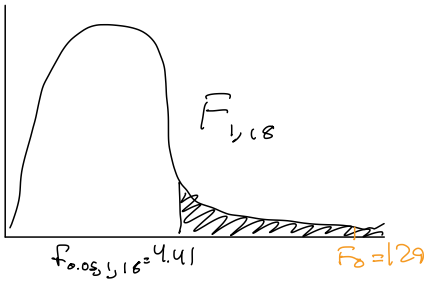
Observation Number	Hydrocarbon Level x(%)	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43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

1) Parameter of interest: Slope of Linear Model; one regressor → simple linear regression

2) $H_0: \beta_1 = 0$ $\quad H_a: \beta_1 \neq 0$

4) $F_0 = \dfrac{SS_R/1}{SS_E/n-2}$

5) Reject if $F_0 > f_{0.05,1,18}$



$\bar{y} = 92.16$

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

$n=1 \Rightarrow \hat{y} = 74.28 + 14.947(0.99) = 88.08$

$\vdots$ all data points

$n=20 \Rightarrow \hat{y} = 74.28 + 14.947(0.95) = 88.48$

$$SS_R = \sum_{i=1}^{n} (\hat{y}-\bar{y})^2 = (88.08-92.16)^2 + \overset{\text{all points}}{\cdots} + (88.48-92.16)^2 = 152.17$$

$$SS_T = \sum_{i=1}^{n} (y_i-\bar{y})^2 = (90.01-92.16)^2 + \overset{\text{all points}}{\cdots} + (87.33-92.16)^2 = 173.38$$

$SS_E = SS_T - SS_R = 21.2$

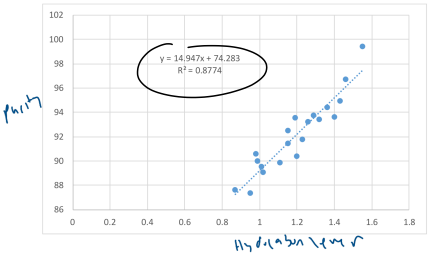
$\Rightarrow F_0 = \dfrac{152.17/1}{21.2/(20-2)} = 129$

$f_{0.05,1,18} = 4.41$ from tables

$129 > 4.41 \rightarrow$ Reject H_0

There is a significant linear relationship between hydrocarbon level and purity

$R^2 = \dfrac{SS_R}{SS_T} = 1 - \dfrac{SS_E}{SS_T} = \dfrac{152}{173} = 0.88$



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

Confidence Intervals around Coefficients

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

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

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

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

How to write and interpret:

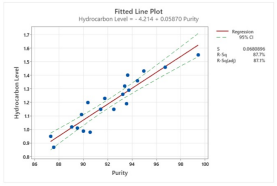
In repeated sampling the intervals calculated capture the true slope or intercept $(1-\alpha)\%$ of the time

Notes on CIs around Response

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

How to Write?

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



How to generate?

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

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

Confidence Intervals around the Response

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

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

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

(Nicely written) ANOVA for Linear Regression

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$
$H_1: \beta_j \neq 0$ for at least one j

Test Statistic:

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

Reference Distribution:

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

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

Regression on Transformed Variables

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

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

Examples

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

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

(ii) Back calculations can be done to report values

Example

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

Observation Number, i	Wind Velocity (mph), xi	DC Output, yi	Observation Number, i	Wind Velocity (mph), xi	DC Output, yi
1	5.00	1.582	14	5.80	1.737
2	6.00	1.822	15	7.40	2.088
3	3.40	1.057	16	3.60	1.137
4	2.70	0.500	17	7.85	2.179
5	10.00	2.236	18	8.80	2.112
6	9.70	2.386	19	7.00	1.800
7	9.55	2.294	20	5.45	1.501
8	3.05	0.558	21	9.10	2.303
9	8.15	2.166	22	10.20	2.310
10	6.20	1.866	23	4.10	1.194
11	2.90	0.653	24	3.95	1.144
12	6.35	1.930	25	2.45	0.123
13	4.60	1.562			

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

Steps

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

We will rely on Minitab a lot more!

1) Scatter Plot doesn't look linear
2-3) Tried Linear → Residuals vs Fit results in obv pattern
↳ Equal variance assumption violated, try a different model
4) Try reciprocal of data
5-6) Results look good now! If they don't, just try a different model
7) y = DC output, x = wind velocity → $y = 2.974 - 6.935\left(\frac{1}{x}\right)$ Always specify variables

Minitab

Transform: Calc > Calculator (input your function)

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

Scatterplot: Graph>Scatterplot> select x and y data

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

Common Transformations

Method	Math Operation	Good for:	Bad for:
Log	$\ln(x)$ $\log_{10}(x)$	Right skewed data $\log_{10}(x)$ is especially good at handling higher order powers of 10 (e.g. 1000, 100000)	Zero values Negative values
Square root	$\sqrt{x}$	Right skewed data	Negative values
Square	x^2	Left skewed data	Negative values
Cube root	$x^{1/3}$	Right skewed data Negative values	Not as effective at normalizing as log transform
Reciprocal	$1/x$	Making small values bigger and big values smaller	Zero values Negative values

Multiple Linear Regression

Multiple Regression Models: A linear regression that contains more than one regressor (K)

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

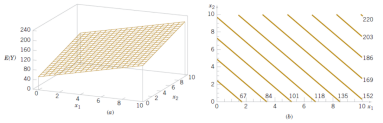
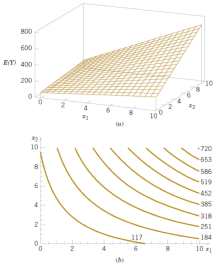


FIGURE 12-1 (a) The regression plane for the model $E(Y) = 50 + 10x_1 + 7x_2$. (b) The contour plot.

Non Linear Surfaces

	Cubic polynomial	Interaction	Second order with interaction
	$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$
Linear Version:	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$



$Y = 50 + 10x_1 + 7x_2 + 5x_1x_2$

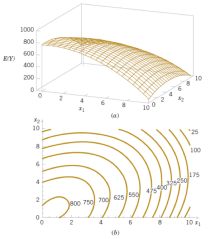


FIGURE 12-3 (a) Three-dimensional plot of the regression model $E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$. (b) The contour plot.

Estimated Coefficients and Variance

Same concepts simple just a lot more terms

Coefficients: Found by method of least squares
↳ Residuals: $e_i = y_i - \hat{y}_i$ Square of these minimized

Variances from different sources

- Variance from Random Error: SS_E from calculation of residuals
  $SS_T = SS_R + SS_E$ → $SS_E = \sum e_i^2$
- Variance from Regression: SS_R = variance from regression
- Variance from individual regression terms: $SS_R(\beta_k)$ = variance coming from specific regressed term
- Partial F-test: Compares variance from specific regressor terms to random error.

Hypothesis Testing

Recall
n = # of data points
k = # of regressor terms

Test for significance of regression:

Ho: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
H1: $\beta_i \neq 0$ for at least one i

Reference Distribution:
F-dist
$F_{k, n-k-1}$

Test Statistic:

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)}$$

Rejection criteria: (Fixed)
Reject Ho if:

$F_0 > F_{\alpha, k, n-k-1}$

or p-value < α

Partial F-test:

Ho: $\beta_i = 0$
H1: $\beta_i \neq 0$

Reference Distribution:
F-dist
$F_{1, n-k-1}$

Test Statistic:

$$F_0 = \frac{SS_R(\beta_i)/1}{SS_E/(n-k-1)}$$

Rejection criteria:
$F_0 > F_{\alpha, 1, n-k-1}$

Nicely Typed Notes

Test for significance of regression:

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

$H_A: \beta_j \neq 0$ for at least one j

F0 = SSR/k / SSE/(n-k-1)

Reference Distribution:
F-distribution: F k, n-k-1

Rejection criteria:
Reject Ho if Fo exceeds F alpha, k, n-k-1
Or if p-value is <alpha

Partial F-test:

Ho: Bj = 0
H1: Bj DNE 0

F0 = SSR (Bj)/1 / SSE/(n-k-1)

Reference Distribution:
F-distribution: F 1, n-k-1

Rejection criteria:
Reject Ho if Fo exceeds F alpha, 1, n-k-1
Or if p-value is <alpha

Adjusted R^2

$SS_T = SS_R + SS_E \Rightarrow R^2 = \frac{SS_R}{SS_T}$ Good for simple linear regression

Use Adj. R^2 for multiple! $R^2 = \frac{SS_R/k}{SS_T/1}$ Accounts overfitting

Normalized by DoF avoids increasing R^2 arbitrarily

Example

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

Brake Horsepower (y) rpm		Road Octane Number	Compression
225	2000	90	100
212	1800	94	95
229	2400	88	110
222	1900	91	96
219	1600	86	100
278	2500	96	110
246	3000	94	98
237	3200	90	100
233	2800	88	105
224	3400	86	97
223	1800	90	100
230	2500	89	104

1) Plot the data
2) Fit a multiple linear regression model, Check the ANOVA results, R2adj, residuals, normality
3) Based on results, consider adding or removing regressors
4) If model is changed – repeat 2-3
5) Generate final model and contour plots

Goal: simplest model which significantly describes the data

Minitab

Plot: Graph > Scatter plot > plot response against all X's

Fit Model: Stat > Regression > Fit Regression Model (add Response, and Predictors by double click) (Determine model by button 'Model" to add or subtract terms)

Contour Plots: Stat- regression > regression > contour or surface plot

Final Result

Regression Equation

Brake Horsepower = -266.0 + 0.01071 rpm + 3.135 Road Octane Number + 1.867 Compression

2 was low → significant

Control Charts

- **Statistical Process Control:** A collection of problem-solving tools useful in achieving process stability and reducing variability

- The Magnificent Seven

1. Histogram or stem-and-leaf plot
2. Check sheet
3. Pareto chart
4. Cause-and-effect diagram
5. Defect concentration diagram
6. Scatter diagram
7. Control chart

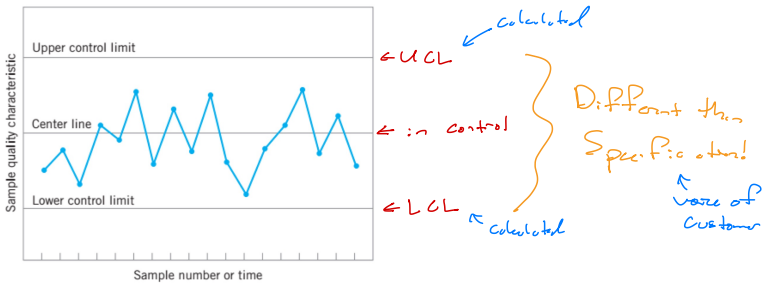
Shewhart's Control Charts & Types of Variability

- **Chance causes of variation:** Inherent variability in a process or background noise - a process operating with only chance causes is in statistical control
- **Assignable Causes of Variation:** Source of variability that are not due to chance - a process operating outside of assignable ranges is an out of control process
  - → **Sources:** Improperly adjusted/controlled machines, Operator errors, Defective raw material ★
  - → **Control Chart:** Online process monitoring technique

Control Chart Uses

- **Monitoring:** Using the control chart to determine if the process is in control, and quickly detect any assignable causes
- **Estimating:** Using the control chart to estimate process parameters (mean or variance)
- **Improving:** Using the control chart to improve process parameters

Features of a Control Chart



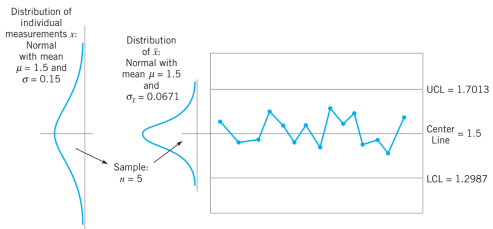
Relationship to Hypothesis Testing

Null Hypothesis: $\mu = \mu_0$ in statistical control

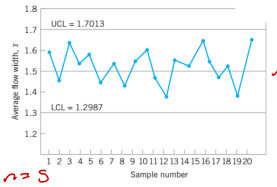
Alternative: $\mu \neq \mu_0$

Type I error: Rejecting H_0 when H_0 is true; concluding the process is out of control when it is not

Type II error: Not rejecting null H_0 when it is false; Assuming a process is in control when it really is out of control



Example



$\bar{x}$ Control chart: uses sample average to monitor a process mean

$100(1-\alpha)\%$ of sample means should fall between control limits

Three Sigma Control limits: Define $Z_{\alpha/2} = 3 \Rightarrow \alpha = 0.0027$

Control limits: (Known variance)

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

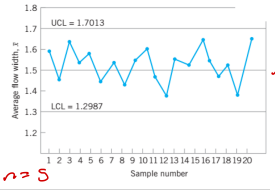
(LCL on the left, UCL on the right)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.15}{\sqrt{5}} = 0.0671$$

$$UCL = 1.5 + 3(.0671) = 1.7013$$

$$LCL = 1.5 - 3(.0671) = 1.2987$$

Example



$\mu_0 = 1.5$
$\sigma = 0.15$
$n = 5$

$\bar{x}$ control chart: uses sample average to monitor a process mean

$100(1-\alpha)\%$ of sample means should fall between control limits

Three Sigma Control limits: Define $Z_{\alpha/2} = 3 \Rightarrow \alpha = 0.0027$

Control limits: (Known μ)

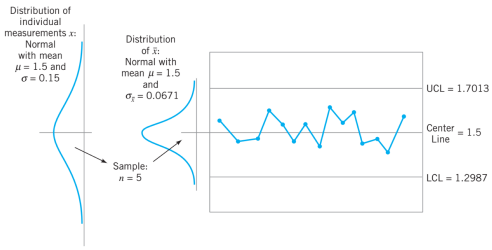
$$\text{LCL} \quad \text{UCL}$$
$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.15}{\sqrt{5}} = 0.0671$$

$UCL = 1.5 + 3(.0671) = 1.7013$

$LCL = 1.5 - 3(.0671) = 1.2987$

General Model



w = quality characteristic of interest

μ_w = mean of w

σ_w = std dev of w

L = distance from centerline expressed in std dev

$UCL = \mu_w + L\sigma_w$

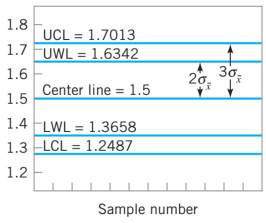
$Center = \mu_w$

$LCL = \mu_w - L\sigma_w$

Design of a Control Chart

- Selecting Control Limits:
  - Set alpha (Europe)
  - Set Sigma Level (US)

3 is pretty standard compromise balancing type I and II error



- Two Limit Control Charts
  - Action limits — outer limits that trigger search for assignable cause
  - Warning Limits — inner limits which raise alarms

Average Run Length (ARL) - the average # of points that must be plotted before a point will indicate an out of control action

$ARL = \frac{1}{p}$; p = probability the point exceeds control limit

Average Run Signal (ARS) - the average amount of time between out of control signals

$ARS = ARL \times$ (time between samples)

Analysis of Patterns

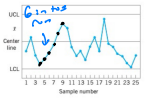
A process can be out of control even if all points are within limits!

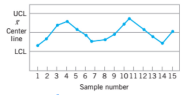
Run: A series of observations of the same type

Run length: How many points are in the series

Pattern: Non-random behavior

Zone Rules: Criteria to help identify paths/runs





Sinusoidal

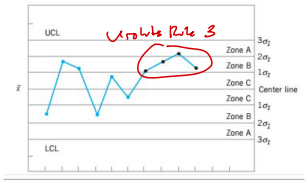
Detecting Patterns and Sensitizing Rules

Standard Action Signal:	1. One or more points outside of the control limits	Western Electric Rules
	2. Two of three consecutive points outside the two-sigma warning limits but still inside the control limits	
	3. Four of five consecutive points beyond the one-sigma limits	
	4. A run of eight consecutive points on one side of the center line	
Sensitizing Rules:	5. Six points in a row steadily increasing or decreasing	
	6. Fifteen points in a row in zone C (both above and below the center line)	
	7. Fourteen points in a row alternating up and down	
	8. Eight points in a row on both sides of the center line with none in zone C	
	9. An unusual or nonrandom pattern in the data	
	10. One or more points near a warning or control limit	

Out of control if yes

If yes, investigate more, get more data
Raises alarm

Sensitizing Rules: Criteria to increase Sensitivity



=> Out of Control example

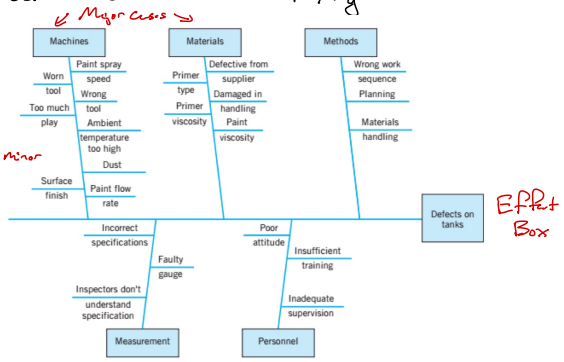
$\bar{x}$ chart: Average of samples
R-Chart: the difference between most extreme readings of a sample

Out of Control Action Plans

OCAP: A flow chart or text-based description of the sequence of actions that take place following an activity event

Checkpoints: Potential assignable causes
Countermeasures: Action taken to Resolve

Cause and Effect Diagram

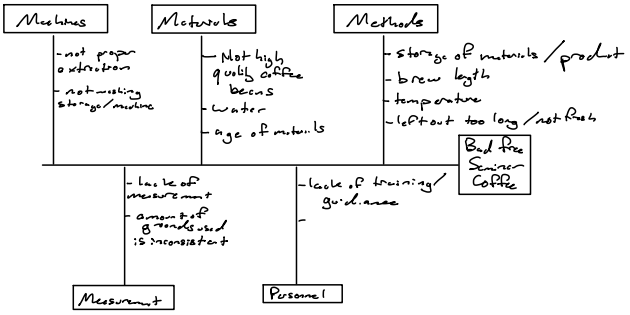


How to construct a C&E diagram

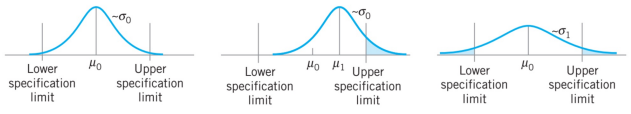
How to Construct a Cause-and-Effect Diagram

1. Define the problem or effect to be analyzed.
2. Form the team to perform the analysis. Often the team will uncover potential causes through brainstorming.
3. Draw the effect box and the center line.
4. Specify the major potential cause categories and join them as boxes connected to the center line.
5. Identify the possible causes and classify them into the categories in step 4. Create new categories, if necessary.
6. Rank order the causes to identify those that seem most likely to impact the problem.
7. Take corrective action.

Example: Free Senior Coffee



Controlling Mean & Variability



Why use both $\bar{x}$ and R charts

Mean Shifts: $\bar{x}$ shows change, R chart won't
Variance Shifts: R chart shows change, $\bar{x}$ won't

Known Population and Variance

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\mu + Z_{\alpha/2}\sigma_{\bar{x}} = \mu + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mu - Z_{\alpha/2}\sigma_{\bar{x}} = \mu - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

LSL USL

Parameters Unknown

Estimated from "in control" samples of at least 20-25 (m) each with at least 4-5 replicates (n)

where m is the total # of samples of size n

$$M = \bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_m}{m}$$

$$\bar{R} = \frac{R_1 + R_2 + \cdots R_m}{m} \rightarrow \bar{R}_m = X_{max} - X_{min}$$

for any sample

Designed to give limits 3 standard deviations away from mean:

Control Limits for the $\bar{x}$ Chart

$$\text{UCL} = \bar{\bar{x}} + A_2\bar{R}$$
$$\text{Center line} = \bar{\bar{x}} \qquad (6.4)$$
$$\text{LCL} = \bar{\bar{x}} - A_2\bar{R}$$

The constant A_2 is tabulated for various sample sizes in Appendix Table VI.

Control Limits for the R Chart

$$\text{UCL} = D_4\bar{R}$$
$$\text{Center line} = \bar{R} \qquad (6.5)$$
$$\text{LCL} = D_3\bar{R}$$

The constants D_3 and D_4 are tabulated for various values of n in Appendix Table VI.

Example

EXAMPLE 6.1 $\bar{x}$ and R Charts for a Manufacturing Process

A hard-bake process (see Section 5.3.1) is used in conjunction with photolithography in semiconductor manufacturing. We wish to establish statistical control of the flow width of the resist in this process using $\bar{x}$ and R charts. Twenty-five samples, each of size five wafers, have been taken when we think the process is in control. The interval of time between samples or subgroups is one hour. The flow width measurement data (in microns) from these samples are shown in Table 6.1.

TABLE 6.1
Flow Width Measurements (microns) for the Hard-Bake Process

Sample Number	Wafers 1	2	3	4	5	$\bar{x}_i$	R_i
1	1.3235	1.4128	1.6744	1.4573	1.6914	1.5119	0.3679
2	1.4314	1.3592	1.6075	1.4666	1.6109	1.4951	0.2517
3	1.4284	1.4871	1.4932	1.4324	1.5674	1.4817	0.1390
4	1.5028	1.6352	1.3841	1.2831	1.5507	1.4712	0.3521
5	1.5604	1.2735	1.5265	1.4363	1.6441	1.4882	0.3706
6	1.5955	1.5451	1.3574	1.3281	1.4198	1.4492	0.2674
7	1.6274	1.5064	1.8366	1.4177	1.5144	1.5805	0.4189
8	1.4190	1.4303	1.6637	1.6067	1.5519	1.5343	0.2447
9	1.3884	1.7277	1.5355	1.5176	1.3688	1.5076	0.3589
10	1.4039	1.6697	1.5089	1.4627	1.5220	1.5134	0.2658
11	1.4158	1.7667	1.4278	1.5928	1.4181	1.5242	0.3509
12	1.5821	1.3355	1.5777	1.3908	1.7559	1.5284	0.4204
13	1.2856	1.4106	1.4447	1.6398	1.1928	1.3947	0.4470
14	1.4951	1.4036	1.5893	1.6458	1.4969	1.5261	0.2422
15	1.3589	1.2863	1.5996	1.2497	1.5471	1.4083	0.3499
16	1.5747	1.5301	1.5171	1.1839	1.8662	1.5344	0.6823
17	1.3680	1.7269	1.3957	1.5014	1.4449	1.4874	0.3589
18	1.4163	1.3864	1.3057	1.6210	1.5573	1.4573	0.3153
19	1.5796	1.4185	1.6541	1.5116	1.7247	1.5777	0.3062
20	1.7106	1.4412	1.2361	1.3820	1.7601	1.5060	0.5240
21	1.4371	1.5051	1.3485	1.5670	1.4880	1.4691	0.2185
22	1.4738	1.5936	1.6583	1.4973	1.4720	1.5390	0.1863
23	1.5917	1.4333	1.5551	1.5295	1.6866	1.5592	0.2533
24	1.6399	1.5243	1.5705	1.5563	1.5530	1.5688	0.1156
25	1.5797	1.3663	1.6240	1.3732	1.6887	1.5264	0.3224
						$\Sigma\bar{x}_i = 37.6400$	$\Sigma R_i = 8.1302$
						$\bar{\bar{x}} = 1.5056$	$\bar{R} = 0.32521$

R-chart

$\bar{R} = \frac{0.3679 + .2517 + \cdots + R_{25}}{25} = 0.32521$

$LCL = \bar{R} D_3 = 0$

$UCL = \bar{R} D_4 = .32521(2.114) = 0.68749$

$\bar{x}$ Chart

$\bar{\bar{x}} = \frac{1.5119 + \cdots R_{25}}{25} = 1.5056$

$UCL = \bar{\bar{x}} + A_2\bar{R} = 1.5056 + .577(.3252) = 1.69325$

$LCL = \bar{\bar{x}} - A_2\bar{R} = 1.5056 - .577(.3252) = 1.31795$

Minitab

Stat -> Control Charts -> Variables Charts for Subgroups -> Select Xbar or R -> dropdown to observations from subgroup are in one row of columns -> select all columns with data (n=number of selections)

Estimates from Control Charts

Estimate process standard deviation:

$\hat{\sigma} = \frac{\bar{R}}{d_2} = \frac{.32521}{2.326} = 0.1398$

Specs ≠ Control Limits

Estimate of process mean

$\hat{\mu} = \mu$

Note: Can chart versus specifications, but be careful

Topics

General:

- how to correctly phrase conclusions in a hypothesis test
- how to correctly phrase what the p-value is
- how to correctly phrase what confidence intervals are representing
- how to sketch hypothesis tests on a reference (sampling) distribution

Inferring the variance of two populations:

- how to conduct the hypothesis test
- how to draw conclusions

ANOVA:

- how to conduct the test (by hand and interpret Minitab) and draw conclusions
- what situations ANOVA is used in (when to pick it as the statistical test)
- what, generally, the test statistic represents
- what the reference distribution is as related to the test statistic and when the test statistic follows the reference distribution
- know how to check assumptions and what assumptions are made
- how and when to perform a posthoc test
- how and when to transform data

Linear Regression:

- know how to perform simple linear regression by hand (how to perform the hypothesis test – won't ask the regression terms by hand)
- how to set up equations and hypothesis for multiple and simple linear regression
- know how to interpret results from Minitab output on simple and multiple linear regression
- explain what R^2 is (+adjusted), what it can and cannot be used for, and how to calculate by hand for simple and multiple linear regression
- know how to check assumptions and what assumptions are made
- how and when to transform data
- describe what the confidence intervals on the response and coefficients are

Control Charts:

- Be able to explain what impact wide or narrow control limits have on type I and type II error, and be able to define type I and type II errors for control charts
- Be able to explain in words and sketches why it is important to use control charts for both process mean and variability
- Calculate control limits and warning limits for xbar and R charts
- Be able to identify and out of control process
- ARL calculation and concept

ECHE 313 Exam 2

Write your name: _____

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

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

_____

The exam layout is such that all problems add up to 100 points. This point total is meant to guide you in time allocation for this test, will be adjusted such that this exam is 25% of your total class grade. You may use your class materials, class notes, homework, book, and Minitab to aid in the completion of the exam, but you are not allowed to use any outside resources. The exam is designed to be completed in 1 hr and 15 min, but you will have between 10 AM 4/9 to 10 AM 4/10 to complete the exam. Note: Once you start the exam, you have 4 hours to complete the test. All of the necessary tables are provided in the back of your book. Please show your work and all steps necessary to arrive at your answer. It is expected you do this entire test by hand, using a calculator and the tables found in the back of your book. So while you are allowed to use Minitab if you wish to check your answers, it should not replace the work you have done by hand using the tables for your answers. Where applicable, it is best to write out the general formals you want to use first, then write in your subbed in values so it is clear. Also, when reporting numbers, keep the same number of decimals as the data you are working with unless there are specific directions on how many sig figs to report. Good luck!

Premise: You work as a process engineer for a company that makes eclipse glasses. What a great time to be in this business! These glasses work by filtering light, reducing light exposure to the eye. Unless otherwise stated, your company uses α=0.05.

Problem 1 (67 points total) Your team wants to change the concentration of materials that are used to block light in your product to save money. Material B is particularly expensive and is typically set at 5.00 wt% in the material. You want to know if the concentration of material B has an impact on the % transmittance of light, so you vary material B and test the amount of light that is transmitted through the material, and get the following data:

	% Transmittance				
	Replicate				
Material B (wt%)	1	2	3	Ave.	Standard Deviation
(typical) 5.00	0.0023	0.0023	0.0024	0.0023	0.000058
4.00	0.0023	0.0025	0.0028	0.0025	0.000252
3.00	0.0027	0.0024	0.0021	0.0024	0.000300
2.00	0.0032	0.0038	0.0035	0.0035	0.000300
1.00	0.0041	0.0039	0.0045	0.0042	0.000306

$\bar{y}_{..} = \frac{0.0023 + 0.0025 + 0.0024 + 0.0035 + 0.0042}{5} = 0.00298$

a. (1 points) List the *name* of the hypothesis test you would use

ANOVA

b. (2 points) Name the parameter of interest in the context of the problem statement

Multiple means, specifically % transmittance at different levels of material B.

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

$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 0$

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

$H_a: \tau_i \neq 0$ for some $i \in (1, 5)$

e. (6 points) Write the formula of the test statistic, and other formulas needed to fully calculate the test statistic from the data above, and calculate its value given that SS_E is 0.00000068

$F_0 = \frac{SS_{Treatments}/(a-1)}{SS_E/(a(n-1))}$ $\begin{matrix} n=5 \\ a=3 \end{matrix}$

$SS_{treatments} = n\sum(\bar{y}_{i.} - \bar{y}_{..})^2$

$= 3\left[(0.0023 - 0.00298)^2 + (0.0025 - 0.00298)^2 + (0.0024 - 0.00298)^2 + (0.0035 - 0.00298)^2 + (0.0042 - 0.00298)^2\right]$

$= 8.364 \times 10^{-6}$ If you use exact values, you get to 0.0000079

$F_0 = \frac{(8.364 \times 10^{-6})/(5-1)}{(0.00000068)/(5(3-1))} = 30.75$

$F_0 = 30.75$ $F_0 = 29$ using exact values

f. (4 points) Describe what the test statistic represents (specifically the numerator and denominator of the test statistic as they related to sources of variation) and what it means if the test statistic is very large versus very small

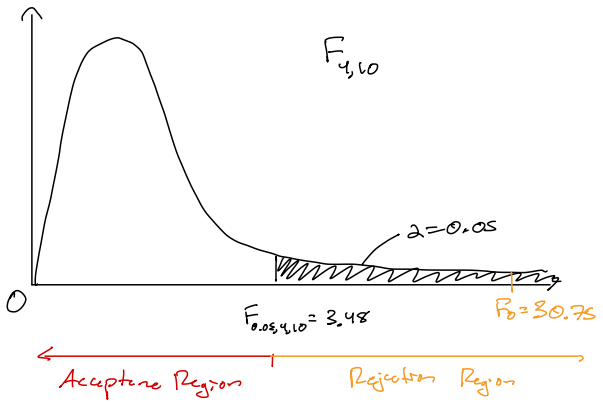
F_0 is effectively $\frac{\text{signal}}{\text{noise}}$. The numerator describes variance from treatments and the denominator describes variance from random error. If F_0 is very large variance is coming from treatments (but hypothesis test determines significance). If it is small, variance is comparable to random error. DoF has an impact on the value.

g. (4 points) Define your general rejection criteria using fixed significance level testing, sub in values and write out the critical value

Reject H_0: $F_0 > F_{\alpha, a-1, a(n-1)}$

$\Rightarrow F_0 > F_{0.05, 4, 10} = 3.48$

h. (10 points) Sketch out the reference distribution for this problem and label: the type of reference distribution used and any defining characteristics (shape, DOF if applicable, and where zero is), the critical value(s), the test statistic, alpha, the acceptance region, and the rejection region.



i. (3 points) Perform any remaining calculations and state your conclusions. Remember to state them in the <u>context of the problem statement</u>, and <u>phrase them appropriately</u>

$F_0 = 30.75 > F_{crit} = 3.48 \rightarrow$ Reject H_0.

Reject the null and accept the alternative. We have significant evidence to indicate that the wt% of material B impacts % transmittance.

j. (7 points) Describe when it is appropriate to do post-hoc testing, and assuming it is appropriate in this case, conduct a Fisher's post hoc test and identify which groups are significantly different than the normal process setting (Material B at 5%).

A post-hoc test is only appropriate once the ANOVA H_0 is rejected!

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

5 vs. 4: $|0.0023 - 0.0028| = 0.0002$

5 vs. 3: $|0.0023 - 0.0024| = 0.0001$

5 vs. 2: $|0.0023 - 0.00385| = 0.0012$ *

5 vs. 1: $|0.0023 - 0.0042| = 0.0019$ *

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

$$= t_{0.025,10}\sqrt{\frac{2\left(\frac{0.00000068}{10}\right)}{3}}$$

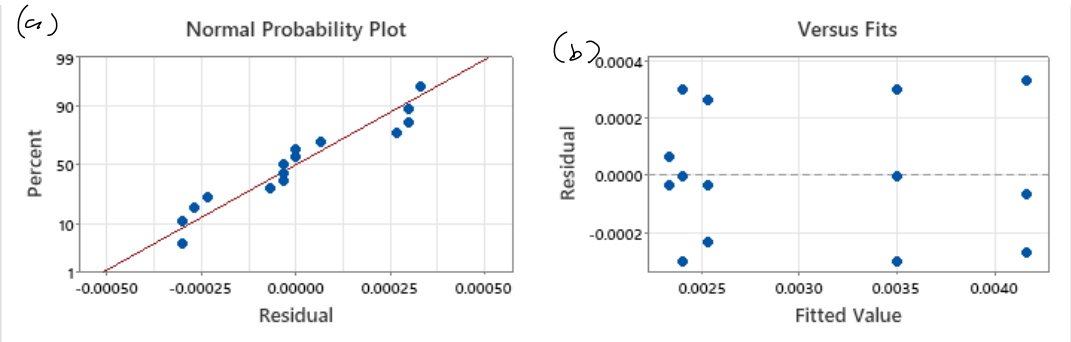
$$= 2.228\,(2.1298\times10^{-4})$$

$$= 4.74\times10^{-4}$$

$$= 0.000474$$

1 and 2 wt% are significantly different than 5 wt% (norm)

k. (4 points) If these are the residual plots:



1. State what assumptions are being checked for each plot

(a) Errors and observations are normally distributed

(b) Constant variance

2. Describe what you should do if the assumptions are violated

Transform the data and perform ANOVA again.

I. You are concerned that the variance of the data for 5 wt% treatment is different than the other treatments.

1. (1 points) What is the name of the hypothesis test I should perform if I want to know if the variance from 5% treatment and 4% treatment are different?

2-variance test!

2. (2 points) Write out the correct null and alternative hypotheses for this test

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

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

where 1 = 5 wt%, and 2 = 4 wt%

3. (5 points) Calculate the test statistic (report 3 significant figures)

Write out the general formula:

$$F_0 = \frac{s_1^2}{s_2^2}$$

Sub in values and calculate:

$$F_0 = \frac{0.000058^2}{0.000252^2} = .052973$$

4. (8 points) 1. Write out the rejection critical using p-value method and 2. determine the p-value from your test statistic (report 2 sig figs in your p-value). 3. Show the p-value graphically. In your sketch, you should have the reference distribution labeled with any defining characteristics (shape, zero, and degrees of freedom if applicable), the test statistic, and the p-value).

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

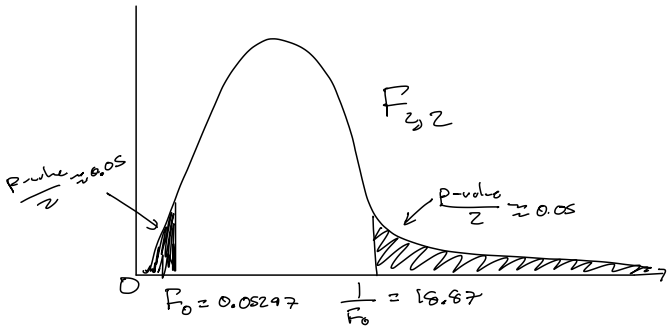
(ii) $F_{2,2}$ (F_{n_1-1, n_2-1})

$19.00 = F_{0.05, 2, 2}$

$\therefore$ p-value $\approx 2(0.05) = 0.10$

$\rightarrow$ p-value ≈ 0.10

(iii)



5. (3 points) State your conclusions. Remember to state them in the context of the problem statement, and phrase them appropriately. Also use your conclusion to answer: is there cause for concern?

p-value = 0.1 > 0.05 ⇒ Fail to reject H_0. We do not have enough evidence to say that the variance at 5 wt% is different than 4 wt%. No cause for concern as this implies constant variance is still valid.

m. (3 points) Provide and answer to your boss who was wondering: "Why is the rejection criteria for ANOVA one sided, whereas sometimes an F-test can be one- or two-sided?"

ANOVA is looking to see if a "signal" is larger than "noise". We therefore only care if the variance of the "signal" is larger than random error, thus it is one sided. If using an F-test we are determining if two variances are different, thus we need a two-sided test.

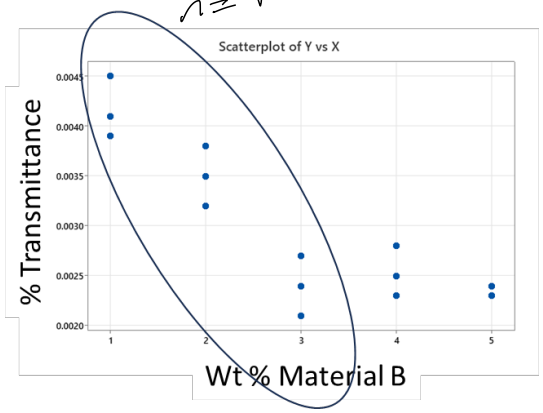
Problem 2. (22 points total) You suspect that the Material B wt% (from 1 to 3 wt% - predictor variable) is linearly correlated to % Transmittance (response variable). You want to perform a hypothesis test to see if the linear correlation is significant.

Performing linear regression on these data

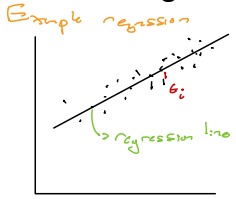
k = # of regression terms

n = # of data points

n = 9



a. (3 points) You perform the liner regression and your estimated coefficient for the slope is -0.00088 and the intercept is 0.0051. What is the name of the method that is used to estimate these coefficients, and describe generally how that method works. (Hint: I am NOT looking for how to calculate S_{xx} or S_{xy} here). You can use a diagram and general formulas for your answer if needed.



The method of least squares is used to estimate the coefficients in linear regression. In this method, the vertical deviations from data and the regression line (ϵ_i's) are minimal. These are the residuals, and the model aims to minimize $\sum_{i=1}^{n} \epsilon_i^2$.

b. This is your Minitab output from the simple linear regression:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression		0.00000468167		51.16	0.000
Error		0.00000064056			
Total					

(1) (2 points) State the degrees of freedom for the $SS_{Regression}$ and the SS_{Error}

$SS_{Regression}$: $k = 1$

SS_{Error}: $n - k - 1 = n - 2 = 7$

(2) (2 points) Write out the formula used and calculate the SS_{Total} using information given in the table above

$SS_{Total} = SS_{12} + SS_E$

$= 0.000004681677 + 0.000000064056$

$= 0.00000053$

(3) (2 points) Write out the formula for and calculate R^2

$\frac{SS_R}{SS_T} = \frac{0.000004681677}{0.00000053} = 0.88$

(4) (3 points) State your conclusions. Remember to state them in the context of the problem statement, and phrase them appropriately. Assume that no issues were noted with the residual plots.

p-value $= 0.2 > \alpha = 0.05 \rightarrow$ Reject H_0 and accept H_a. We have significant evidence to reject the claim that there is no linear relationship between material B from 1-3% and % transmittance. The wt% of material B from 1-3% is linearly correlated with % transmittance.

c. (10 points) You think there might be another variable that could be useful in your model (wt% of another material, Material A). You add in this variable and run multiple linear regression to get the following output:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	0.000005	0.000002	26.21	0.001
Wt% B	1	0.000002	0.000002	19.92	0.004
Wt% A	1	0.000000	0.000000	1.03	0.349
Error	6	0.000001	0.000000		
Total	8	0.000005			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0003018	89.73%	86.30%	76.89%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.005844	0.000760	7.69	0.000	
X	-0.001100	0.000246	-4.46	0.004	4.00
X2	-0.000289	0.000285	-1.02	0.349	4.00

(1) (3 points) The R^2 increased after adding the new variable. Should we keep the variable in the model? Why or why not?

No. R^2 is not a valid tool for determining the significance of a variable in regression. In addition, the partial F-test shows p-value > 0.05 = α for wt% A, so there isn't enough evidence to say it has a significant relationship with % transmittance.

(2) (3 points) R^2 adjusted before adding Material A as a predictor was 86%. Explain why R^2 adjusted did not go up when adding the new variable. Use the formula in your answer.

$R^2_{adj} = \frac{SS_R/k}{SS_T/n-1}$ R^2_{adj} includes degrees of freedom which prevents k^2 from increasing simply by adding more regression terms.

(3) (4 points) State your conclusions appropriately in the context of the problem statement and write out the final model

Fail to reject H_0. We don't have enough evidence to say wt% A has a significant relationship with % transmittance. The final model is: $y = 0.0051 - 0.00088x$

y = % transmittance

x = wt% material B

Problem 3. (11 points total) Your company wants to establish and Xbar control chart on the process for component C wt%. There are 22 preliminary samples, each of size 8, on the internal diameter of the seal. The summary data (in mm) are as follows:

Sample #	Range (R)	Average (Xbar)
1	5	15
2	9	11
3	8	11
4	4	13
5	9	13
6	5	10
7	10	14
8	5	14
9	9	12
10	8	13
11	8	15
12	8	11
13	8	15
14	8	12
15	5	13
16	8	14
17	7	14
18	4	11
19	9	12
20	8	15
21	8	14
22	5	11
Sum	158	283

$\bar{R} = \frac{158}{22} = 7.18$ $\bar{\bar{x}} = \frac{283}{22} = 13.0$

a. (8 Points) Write out the general formulas for and calculate the centerline as well as the three sigma control limits that should be used on the Xbar and R control chart

$\bar{X}$ bar:

$UCL = \bar{\bar{x}} + A_2 \bar{R} = 13.0 + 0.373(7.18) = 15.7$ ≈ 16

$Center = \bar{\bar{x}} = 13.0$ ≈ 13

$LCL = \bar{\bar{x}} - A_2 \bar{R} = 13.0 - 0.373(7.18) = 10.3$ ≈ 10

R:

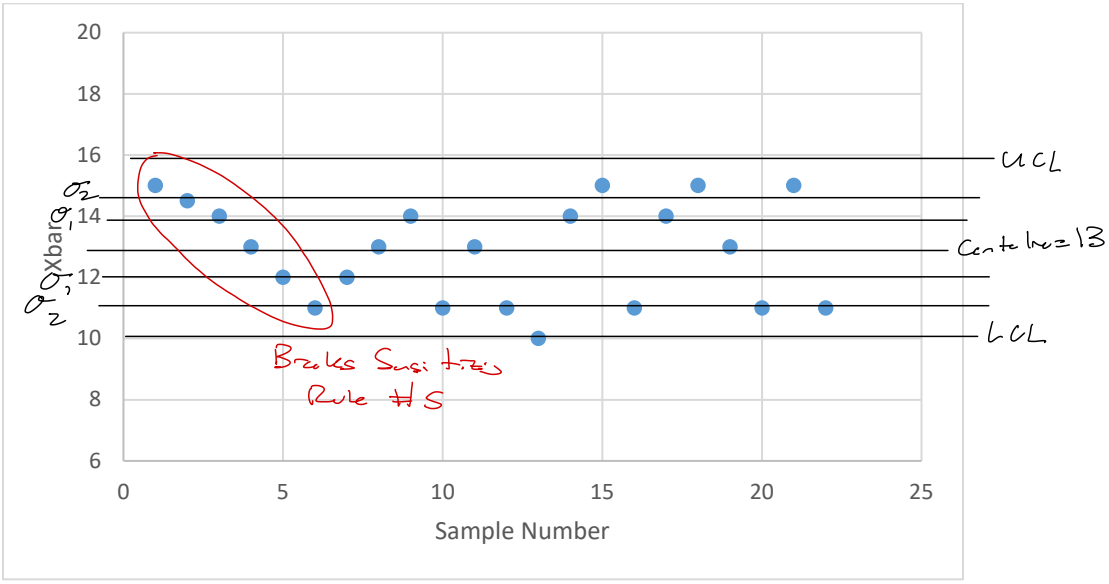
$UCL = D_4 \bar{R} = 1.864(7.18) = 13.4$ ≈ 13

$Center = \bar{R} = 7.18$

$LCL = D_3 \bar{R} = 0.136(7.18) = 0.98$ ≈ 1

b. (3 points) In the next month, you take an additional 20 samples in a similar manner as before. Sketch in your sigma control limits (UCL and LCL) and centerline from part A, as well as the 1 sigma and 2 sigma warning limits on this graph of your data. After analyzing this plot, what is your recommendation?

In control, does not break rules 1-4



Process should be investigated as Sensitizing #5 is broken

Homework 6

Due 3/20/25

Reminder: Reading Chapter 4 of the textbook will help you answer these questions. In lectures 12 and 13 and some of 14 you will have seen relevant background and examples. *Unless otherwise noted, write out by hand how to calculate the desired values (you can abbreviate but sub in some values). Also please pay attention to sig figs, and remember to sketch out the reference distribution when you do the hypothesis tests – each problem should have a reference distribution sketch.*

1) (a) Do problem 4.13 (a) (4.23 in 7th edition) in your book doing the step-by-step hypothesis method by hand (either p-value or critical value is fine). Don't do part b from the book but instead do the following: (b) check your answer in part a in Minitab, print the output and report the p-value you obtained
2) You wish to know if changing the catalyst in a process causes the yield to have more variability. Data are collected on the yield for each catalyst:

Cat. 1	Cat. 2
57.9	66.4
66.2	71.7
65.4	70.3
65.4	69.3
65.2	64.8
62.6	69.6
67.6	68.6
63.7	69.4
67.2	65.3
71.0	68.8

a) Cut and paste this data into Minitab and test the hypothesis that the variances are equal. Assume normality (there is a box to check in this test that assumes normality). Report the output and your conclusions – answer the question of whether you could assume equal variance for these samples if you were to conduct a t-test. Use $\alpha=0.05$
b) Check the normality assumption using Minitab – does it seem appropriate?

3.) Do problem 4.23 from your book (4.39 in the 7th edition) and use Excel to process the data for parts a-c. Provide the excel sheets and write out the equations used to get your answers. For part b, put the calculated residuals you got from Excel into Minitab to produce your normal probability plot. Use α=0.05.

For the final part of this problem (part d) check your answers in Minitab (including the residual plot via the ANOVA function), click assume equal variance box. Report the P-value.

4.) Do problem 4.25 (4.41 in 7th edition) from the book. Use Minitab to conduct the hypothesis testing (assuming equal variances).

a) Conduct the analysis and state your conclusions and sketch out the situation on a reference distribution. Also answer the following questions:

a-1) how many factors are in this experiment?

a-2) how many levels?

a-3) What is the critical value of F if you were to compare the test statistic to it instead of using the P-value?

b) Analyze residuals according to the following directions

b-1) Calculate the residuals in Excel. (Write out the equations needed and report the sheet you used to get your answers).

b-2) Analyze the residuals by plotting them on a normal probability plot in Minitab (cutting and pasting from your table you made in Excel). State if you think normality is an OK assumption.

b-3) Plot the residuals you calculated in Excel, versus the factor levels. (Can use Minitab or Excel) Answer what assumption this graph checks, and comment if the assumption is still valid.

b-4) Plot the residuals vs. the fitted value ($\bar{y_{i.}}$). Can use Minitab or Excel. Do the residuals depend on the response? State what you would do if they did appear to have a pattern.

C) Do this next extra part (not in the book)

c-1) Conduct the Fisher's LSD post-hoc test using Excel given the MSE = 6.625. Compare: 0.37 vs. 0.51, 0.37 vs. 0.71, and 0.51 vs. 0.71 by hand and conduct the full analysis using Minitab

c-2) Construct a plot that represent your data with an appropriate caption by plotting the average ± the standard deviation at each treatment mean, and above the bars, listing the letter assignments from the post hoc testing. For the caption, write something like this

example: “All data are represented by the average ± standard error. ANOVA was performed to confirm statistical significance of (insert what factor you are investigating), followed by the (name here) *post hoc* test to determine differences between treatment means using α=X. Bars that do not share a letter within a graph are statistically different.”

Of course variations of this can be written, but you should include all of the above information.

Remember for sketching reference distributions for a hypothesis test:

P-value method: Sketch out the reference distribution, labeling the test statistic, and the p-value. In addition, fully define the distribution you are sketching on with the correct shape, any defining characteristics and where 0 is. It is also good to draw in alpha – remembering that alpha is not the same as your p-value.

Fixed Significance Level Testing: Sketch out the reference distribution, labeling the acceptance region, the rejection region, α, the critical value(s) and your test statistic. In addition, fully define the distribution you are sketching on with the correct shape, any defining characteristics and where 0 is.

ECHE313 Homework 6 - Due 03/20/25 **Trevor Swan (tcs94)**

4.13 in book

1. $\bar{x}_1 = 9.85$ | $\bar{x}_2 = 8.08$

$s_1^2 = 6.79$ | $s_2^2 = 6.18$

$n_1 = 10$ | $n_2 = 8$

a) Inference on Two chemical processes variance – independent & normality assumed.

↳ Two Sample Variance test

$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \neq \sigma_2^2$

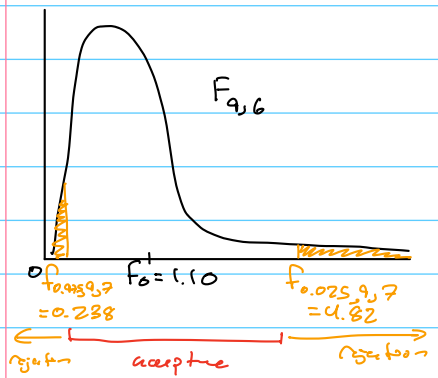
$F_0 = \frac{s_1^2}{s_2^2} = \frac{6.79}{6.18} = 1.10$ ✓

$D.F_1 = 10-1 = 9$

$D.F_2 = 8-1 = 7$

$\alpha = 0.05$

$\frac{\alpha}{2} = 0.025$



Rejection Criteria:

$F_0 > f_{0.025,9,7}$

$F_0 < f_{0.975,9,7} = \frac{1}{f_{0.025,7,9}}$ critical value

$f_{0.025,9,7} = 4.82$ ✓

$f_{0.975,9,7} = \frac{1}{4.20} = 0.238$

$F_0 < f_{0.025,9,7}$ & $F_0 > f_{0.975,9,7}$

Fail to reject H_0. We have insufficient evidence to reject the claim that the variances of the two chemical processes are different. We cannot conclude that the variance of the new purification unit differs from the old purification units variances.

b) Minitab P-value: 0.922 → Fail to reject as p-value = 0.922 > α = 0.05

Test

Null hypothesis	$H_0: \sigma_1^2 / \sigma_2^2 = 1$
Alternative hypothesis	$H_1: \sigma_1^2 / \sigma_2^2 \neq 1$
Significance level	$\alpha = 0.05$

Test

Method	Statistic	DF1	DF2	P-Value
F	1.10	9	7	0.922

b) Residuals

b-1) $E_{ij} = y_{ij} - \bar{y}_{i\cdot}$

$\bar{y}_{1\cdot} = \frac{80+83+83+85}{4} = 82.75$

$\bar{y}_{2\cdot} = \frac{75+75+79+79}{4} = 77$

$\bar{y}_{3\cdot} = \frac{74+73+76+77}{4} = 75$

$\bar{y}_{4\cdot} = \frac{67+72+74+74}{4} = 71.75$

$\bar{y}_{5\cdot} = \frac{62+62+67+69}{4} = 65$

$\bar{y}_{6\cdot} = \frac{60+61+64+66}{4} = 62.75$

$E_{11} = y_{11} - \bar{y}_{1\cdot} = 80 - 82.75 = -2.75$

$E_{12} = y_{12} - \bar{y}_{1\cdot} = 83 - 82.75 = 0.25$

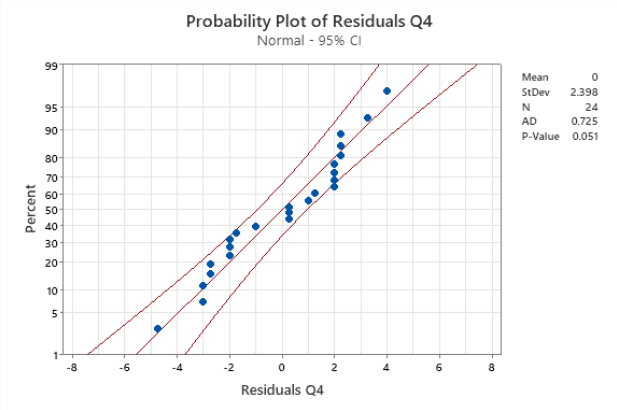
$E_{13} = y_{13} - \bar{y}_{1\cdot} = 83 - 82.75 = 0.25$

$\vdots$

$E_{64} = y_{64} - \bar{y}_{6\cdot} = 66 - 62.75 = 3.25$

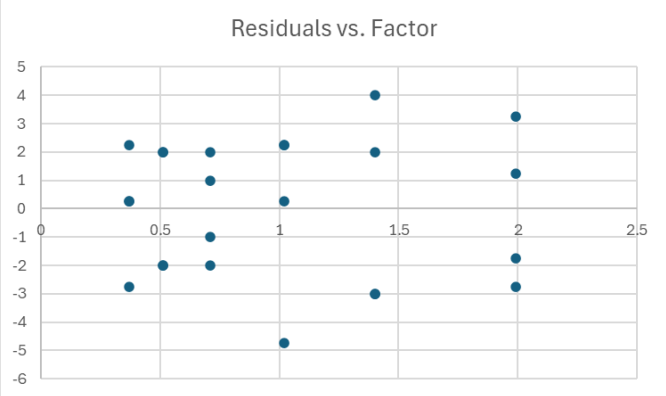
Remaining Residuals on 'Question 4' Sheet in Excel !!

b-2) Normal Probability Plot of Residuals



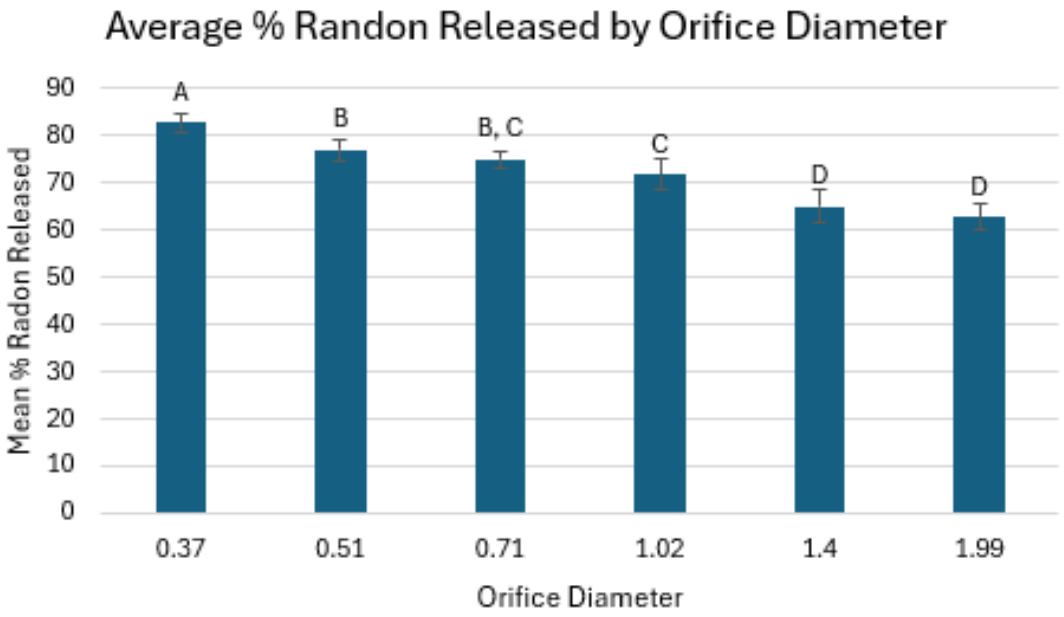
p-value = 0.051 < 0.05 = α, so we fail to Reject the normality assumption. With linearity clearly visible and p-value being so high, normality is an OK assumption. This p-value is very close to α, so I would proceed with caution...

b-3) Residuals vs. Factor Levels



This graph checks equal variance at each factor level. The graph shows the residuals at each factor level being randomly, but evenly, scattered along each factor, so the equal variance assumption is ok here.

c-2)



All data are represented by the average ± standard error. ANOVA was performed to confirm statistical significance of Orifice Diameter differnce on the mean radon percentage released, followed by Fischer's LSD post hoc test to determine differences between treatment means using $\alpha=0.05$. Bars that do not share a letter within a graph are statistically different.

Homework 7

Due 3/27/25 by end of the day

Directions: Reading the supplemental reading posted as Lecture 14 reading and the last part of Chapter 4 will help you answer these questions. In lectures 14, 15 and 16 you will have seen relevant background and examples. Do problem 1-2 by hand using Excel (unless otherwise noted). For the rest of the homework, use Minitab. Try to be aware of the significant figures when reporting model coefficients.

1) Do problem 4.26 (4.43 in 7th edition) in your book, and answer parts a and b by hand (using Excel to aid in the calculations) but use Minitab to check parts a and b and calculate the CI in part c. Keep 3 significant figures and use α=0.05. Also, do extra part (not in book):

   d) calculate the R^2 by hand and explain what this tells you

2) Do problem 4.28 (4.45 in 7th edition) by plotting: (1) residuals on a normal probability plot, (2) residuals versus predicted value, (3) residuals vs. x. To complete these, calculate the residuals in Excel and paste them as part of your answer. You can produce the plots using Excel or Minitab. Remember to comment on if you think the assumptions are valid.
3) Do problem 4.27 (4.44 in the 7th edition) in your book – you can use Minitab for the entire problem (no need to calculate by hand). For part a) also generate a scatter plot of the data and convince yourself a linear model is appropriate and be sure to highlight or report your regression equation. For part b) be sure to state the null and alternative hypothesis and your conclusion. For part c) report the CI and describe what it means. Keep 3 significant figures and use α=0.05. Also add these parts to this problem (not in your book):

   d) use Minitab to generate a plot of the linear regression lines with 95% confidence intervals and explain what the intervals mean

4) An electric utility is interested in developing a model relating peak-hour demand (y in kilowatts) to total monthly energy usage during the month (x, kilowatt hours). Data for 50 residential customers are shown in the table below (last page). You can use Minitab to solve this entire problem.
   a. Plot the scatter diagram of y vs x.
   b. Fit the simple linear regression model (report the equation)
   c. Test for significance of the simple linear regression using alpha=0.05. State the null and alternative hypothesis.
   d. Plot the residuals versus predicted, and comment on the underlying assumptions, specifically, is the equality of variance assumption valid?
   e. Find a simple linear regression model using $\sqrt{y}$ as the response. Does the transformation stabilize the inequality found in part d

5) Do problem 4.30 – 4.33 (4.49-4.52 in the 7th ed) using Minitab, alpha = 0.05. Create a contour plot using Minitab and explain why adding age to the model might be helpful. Using the contour plot, report the range of satisfaction for a 40 year old with a disease severity of 30 and the range for a 40 year old with severity of 60 according to your model.

Data for Problem 4 (cut and paste into Minitab):

	x	y
1	679	0.79
2	292	0.44
3	1012	0.56
4	493	0.79
5	582	2.7
6	1156	3.64
7	997	4.73
8	2189	9.5
9	1097	5.34
10	2078	6.85
11	1818	5.84
12	1700	5.21
13	747	3.25
14	2030	4.43
15	1643	3.16
16	414	0.5
17	354	0.17
18	1276	1.88
19	745	0.77
20	795	3.7
21	540	0.56
22	874	1.56
23	1534	5.28
24	1029	0.64
25	710	4
26	1434	0.31
27	837	4.2
28	1748	4.88
29	1381	3.48
30	1428	7.58
31	1255	2.63
32	1777	4.99
33	370	0.59
34	2316	8.19
35	1130	4.79
36	436	0.51
37	770	1.74
38	724	4.1
39	808	3.94
40	790	0.96
41	783	3.29
42	406	0.44
43	1242	3.24
44	658	2.14
45	1746	5.71
46	895	4.12
47	1114	1.9
48	413	0.51
49	1787	8.33
50	3560	14.94

ECHE313 Homework 7 - Due 03/27/25

Trevor Swan (tcs94)

4.26 in book

1. **Table 4E.6** Tensile Strength Data for Exercise 4.26

Strength	Percentage Hardwood	Strength	Percentage Hardwood
160	10	181	20
171	15	188	25
175	15	193	25
182	20	195	28
184	20	200	30

a) $n=10$

$\Sigma x_i = 160+171+\dots+200 = 1829$

$\Sigma x_i^2 = 160^2+171^2+\dots+200^2 = 335825$

$\Sigma y_i = 10+15+\dots+30 = 208$

$\Sigma y_i^2 = 10^2+15^2+\dots+30^2 = 4684$

$\Sigma x_i y_i = 160(10)+171(15)+\dots+200(30) = 38715$

Should've been flipped → x = strength; y = % Hardwood

$$S_{xy} = \Sigma x_i y_i - \frac{(\Sigma y_i)(\Sigma x_i)}{n} = 38715 - \frac{(208)(1829)}{10} = 671.8$$

$$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = 335825 - \frac{(1829)^2}{10} = 1300.9$$

$$\tilde{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{671.8}{1300.9} = 0.5164 \quad ; \quad \bar{x} = \frac{1}{10}(\Sigma x_i) = \frac{1}{10}(1829) = 182.9$$

$$\bar{y} = \frac{1}{10}(\Sigma y_i) = \frac{1}{10}(208) = 20.8$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \cdot \bar{x} = 20.8 - 0.5164(182.9) = -73.65$$

~~$\hat{y} = 0.516x - 73.7$~~ where x = strength, y = Percentage Hardwood

$\hat{y} = 144 + 1.88x$

correct procedure

b) ANOVA for linear regression

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

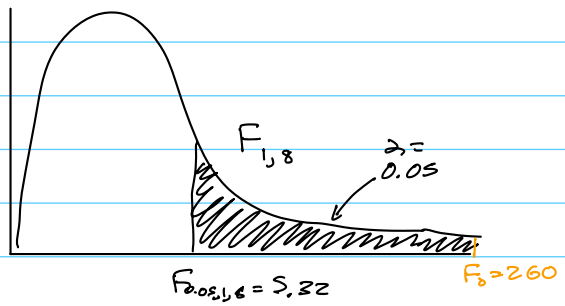
Test Statistic:

$$SS_R = \Sigma(\hat{y}_i - \bar{y})^2 = (8.97-20.8)^2+(14.65-20.8)^2+\dots+(29.63-20.8)^2 = 346.93$$

$$SS_E = \Sigma(y_i - \hat{y}_i)^2 = (10-8.97)^2+\dots+(30-29.63)^2 = 10.67$$

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{346.93}{10.67/(8)} = 260$$

Reject if: $F_0 > F_{0.05,1,8}$



$260 > 5.32 \rightarrow$ Reject H_0

We have sufficient evidence to support the claim that there is a significant linear relationship between percentage hardwood (y) and strength (x).

c)

Regression Equation

Percentage Hardwood = -73.65 + 0.5164 Strength

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-73.65	5.87	(-87.19, -60.12)	-12.55	0.000	
Strength	0.5164	0.0320	(0.4426, 0.5903)	16.12	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
1.15513	97.01%	96.64%	17.4802	95.11%	39.03	35.94

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	346.93	97.01%	346.93	346.925	260.00	0.000
Strength	1	346.93	97.01%	346.93	346.925	260.00	0.000
Error	8	10.67	2.99%	10.67	1.334		
Total	9	357.60	100.00%				

95% CI on β_1 (slope) — 1 Propagated errors

(0.4426, 0.5903) [1.6, 2.15]

In repeated sampling, the interval calculated captures the slope β_1 95% of the time.

Also calculations from previous page (error sum of squares) match the minitab output

d) $R^2 = \frac{SS_R}{SS_T} = \frac{SS_R}{SS_R + SS_E} = \frac{346.93}{346.93 + 10.67} = 97.0\%$

97% of the variability in the target variable is accounted for by the regression model.

4.28 in book

2.

Data from problem ①

Table 4E.6 Tensile Strength Data for Exercise 4.26

Strength	Percentage Hardwood	Strength	Percentage Hardwood
160	10	181	20
171	15	188	25
175	15	193	25
182	20	195	28
184	20	200	30

x = strength; y = % Hardwood

Residuals

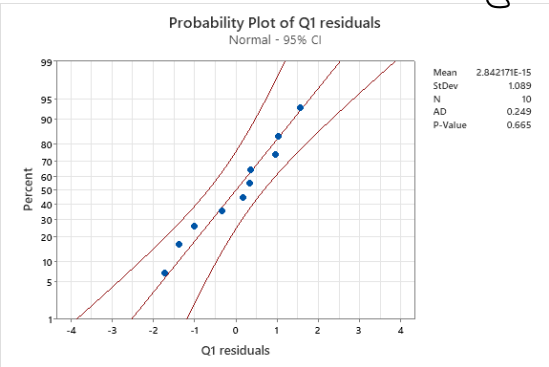
$e_1 = 10 - 8.97 = 1.03$

$e_2 = 15 - 14.65 = .345$

$\vdots$

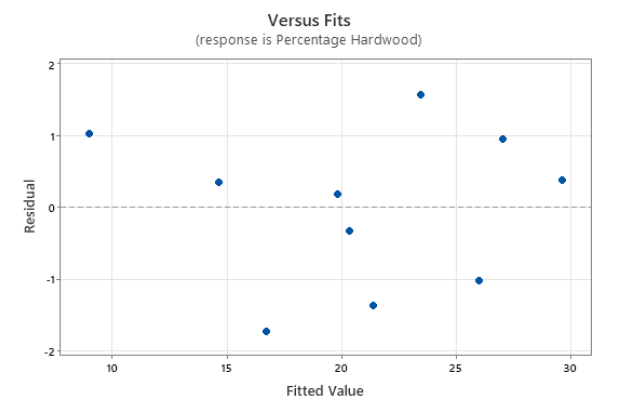
$e_{10} = 30 - 29.63 = 0.369$

a) Residual Normal Probability Plot



Insufficient evidence to reject the claim that the data is normal. As p-value = 0.665 > 0.05 = α, we can support the normality assumption of the model.

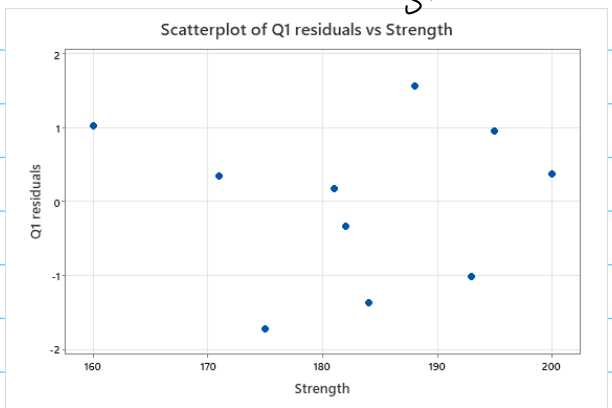
b) Residuals vs. Fitted Value



The residuals are evenly distributed about $y = 0$. There is no clear pattern, and thus the equal variance assumption is valid here.

If there was a pattern, I would apply a transformation

c) Residuals vs. Strength values (x)



The residuals show no patterns and are evenly distributed about 0. the equal variance assumption is again valid.

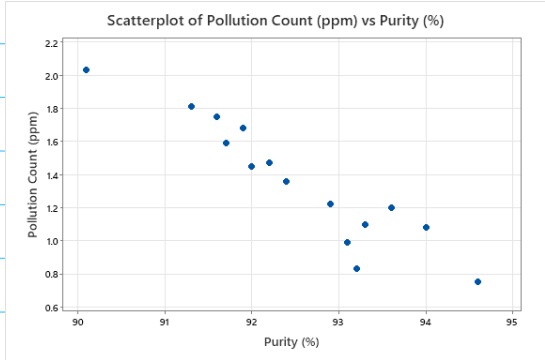
4.27 in book

3.

Purity (%)		93.3	92.0	92.4	91.7	94.0	94.6	93.6
Pollution count (ppm)		1.10	1.45	1.36	1.59	1.08	0.75	1.20
Purity (%)	93.1	93.2	92.9	92.2	91.3	90.1	91.6	91.9
Pollution count (ppm)	0.99	0.83	1.22	1.47	1.81	2.03	1.75	1.68

Missing 1 obs, got both.

a)



Without a fit, there is a clear negative linear relationship between Pollution Count(y) and Purity Percent (x). I'm convinced!

Regression:

$\hat{y} = 29.23 - .303x$

$\hat{y}$ = Pollution Count(ppm)
x = Purity Percent

b) ANOVA test for linear regression

$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$

Minitab p-value $\approx 0 < \alpha = 0.05$

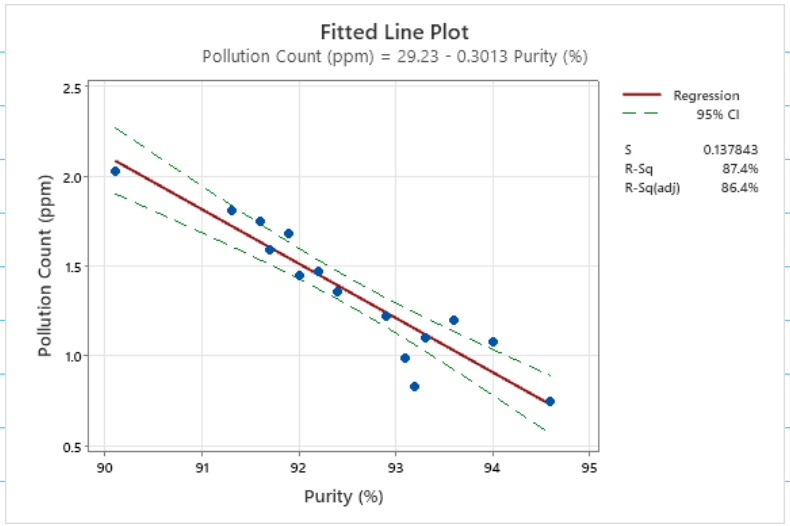
p-value $< \alpha$, so we have significant evidence to reject the claim that there is no linear relationship. We can accept there is a significant linear relationship between Pollution Count(y) and Purity Percent (x).

c) 95% CI for: Purity % = (-0.3698, -.2327) ← slope

Pollution Count(Constant) = (22.89, 35.57) ← intercept

In repeated sampling, the intervals calculated capture the true slope or intercept 95% of the time.

d)

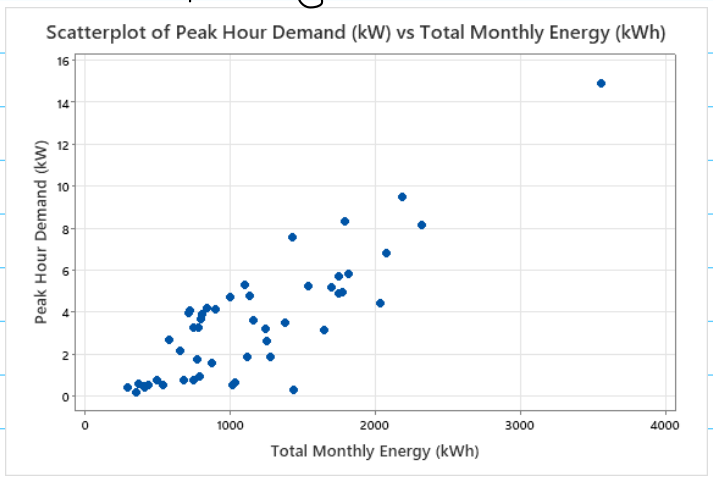


In repeated experiments, there is a 95% chance that the CI on the responses calculated contain the true responses.

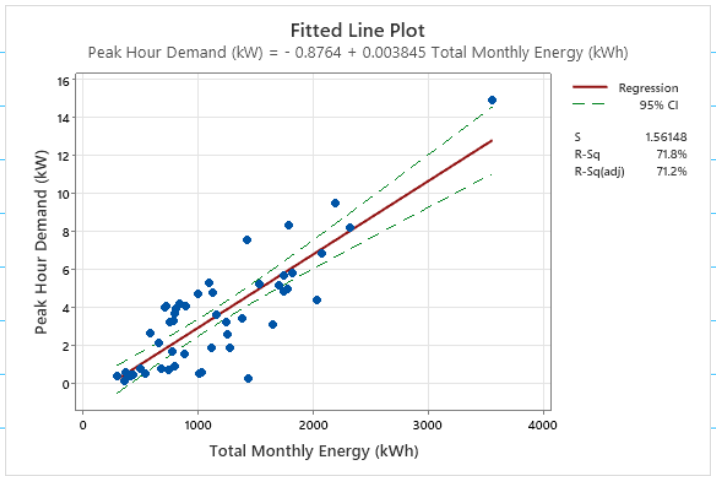
4. use $\begin{cases} x = \text{Total Monthly Energy Used (kWh)} \\ y = \text{Peak Hour Demand (KW)} \end{cases}$

a) Scatter plot of y vs x

Missing data, got better



b) Linear Fit



$\hat{y} = -0.8764 + .003845x$

$\begin{cases} x = \text{Total Monthly Energy Used (kWh)} \\ y = \text{Peak Hour Demand (KW)} \end{cases}$

c) $H_0: \beta_1 = 0$ $\quad H_a: \beta_1 \neq 0$

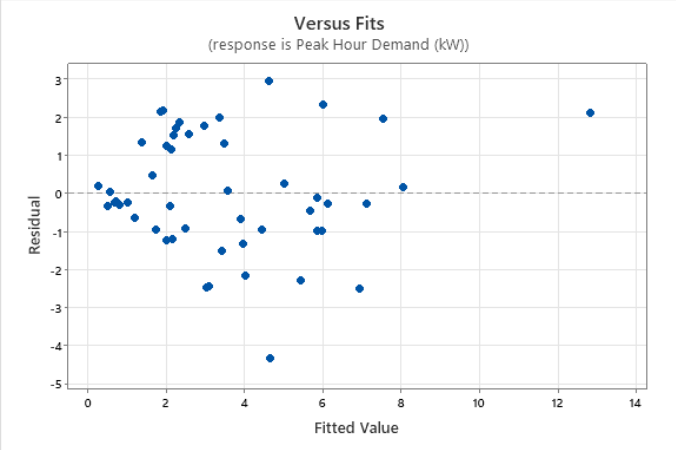
Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	297.7	71.78%	297.7	297.728	122.11	0.000
Total Monthly Energy (kWh)	1	297.7	71.78%	297.7	297.728	122.11	0.000
Error	48	117.0	28.22%	117.0	2.438		
Total	49	414.8	100.00%				

p-value $\approx 0 < \alpha = 0.05$

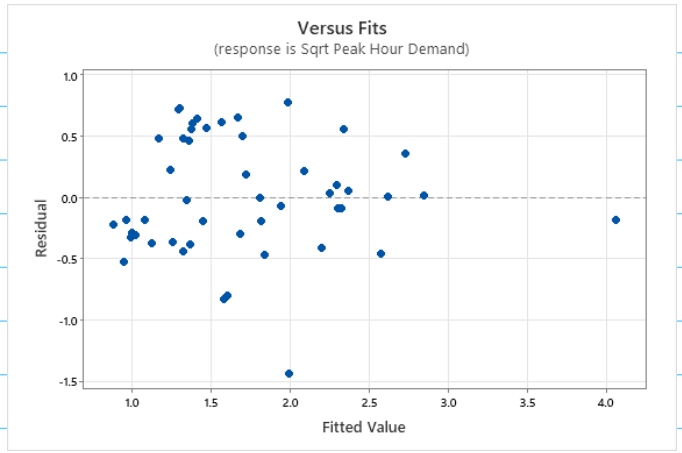
p-value $< \alpha$, so we have significant evidence to reject the claim that there is no linear relationship. We can accept there is a significant linear relationship between the Total Monthly Energy Used (x), and the Peak Hour Demand (y).

d)



There seems to be a factor like pattern, despite being roughly displaced about 0. At lower fitted values, residuals are closer together. Thus, the equality of variance may not be a good assumption and we should apply a data transformation to remedy this

e)



Yes, the data is more evenly distributed about zero and there is not a pattern corresponding to the fitted values. With this transformation, the equality of variance assumption appears to be a valid assumption.

5. a) Problem 4.30

(i) Satisfaction (y) = 131.10 − 1.240 · Age (x_1)

y = Satisfaction
x_1 = Age

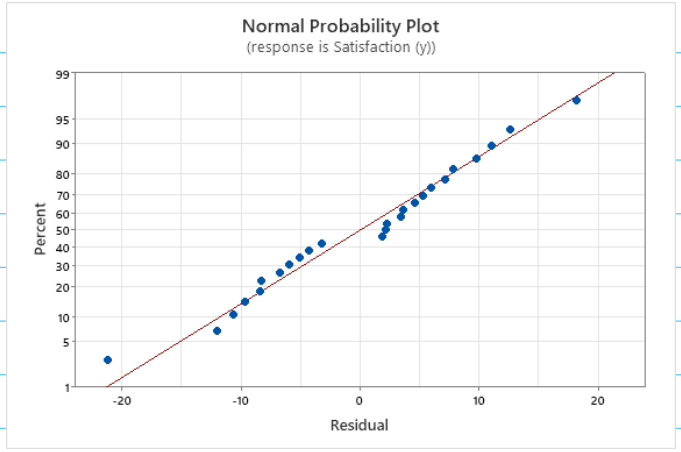
(ii) $H_0: \beta_1 = 0$ $\quad H_a: \beta_1 \neq 0$

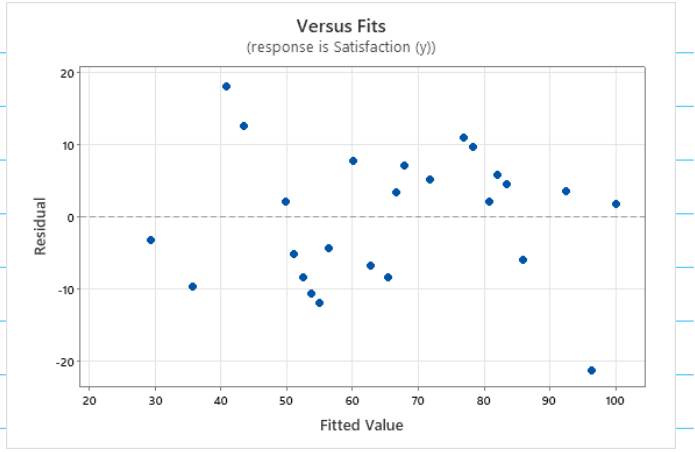
Minitab → p-value = 0 < α = 0.05

we have significant evidence to reject the claim that there is no linear relationship. We can accept that there is a significant linear relationship between the Age of the patient (x_1), and their Satisfaction (y).

(iii) 81.24% of the variability is accounted for by the regressor variable age.

b) Problem 4.31





The normal probability plot of the residuals is linear with no obvious variability, so the normality assumption is valid. The residuals vs. Fitted value is evenly distributed about 0, and no clear pattern is seen. Thus the equality of variance assumption is valid. The model seems to be a good fit.

c) Problem 4.32

(i) Satisfaction (y) = 143.47 − 1.031 · Age (x_1) − 0.556 · Severity (x_2)

y = Satisfaction x_1 = Age x_2 = Severity

H_0: $B_1 = B_2 = 0$ H_a: $B_i \neq 0$ for at least one i

Minitab → p-value = $0 < \alpha = 0.05$

We have significant evidence to reject the claim that there is no linear relationship. We can accept there is a significant linear relationship between the Age of the patient (x_1), Severity of the illness (x_2), and their Satisfaction (y).

(ii) $R^2(x_1) = 81.24\%$ ← Age

$R^2(x_2) = 8.42\%$ ← Severity

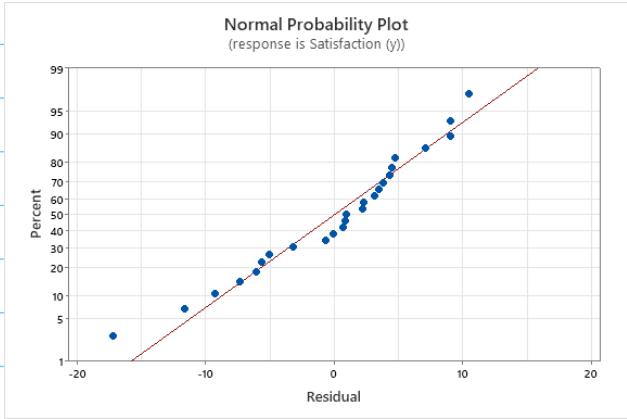
The portion of variability is significant ($> \alpha = .05$) in both cases. Both regressors are needed.

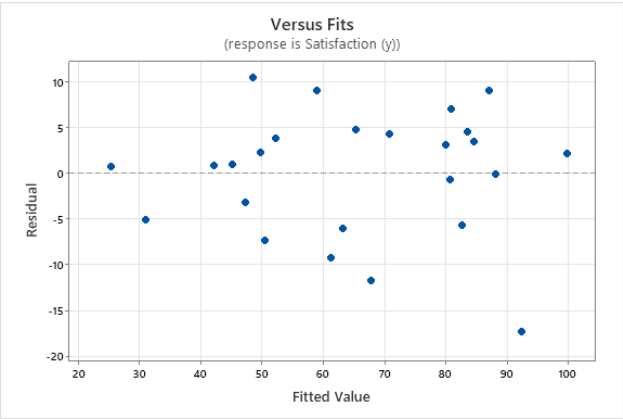
(iii) $R^2_{mult.} = 88.72\%$ (adj)

$R^2_{single} = 80.43\%$ (adj)

The adjusted R^2 value which accounts for overfitting effects improved with the full fit expression. Thus we can conclude adding severity improved the model's quality.

d) Problem 4.33





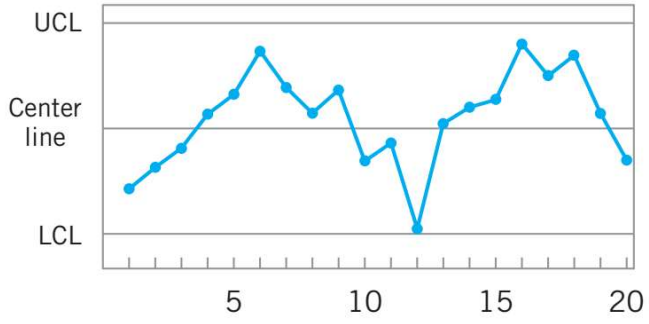
The normal probability plot of the residuals is linear with no outstanding variability, so the normality assumption is valid. The residuals vs. Fitted value is evenly distributed about 0, and no clear pattern is seen. Thus the equality of variance assumption is valid. The model seems to be a good fit.

Homework 8

Due 4/3/25 by end of day

Directions: Reading the chapter 5 and 6 will help solve these problems along with content seen in class.

1) Most control charts choose 3 sigma levels for their control limits because it seems to be a good balance between type II and type II errors. Discuss the following:
    a. If narrower limits are chosen, what happens to the magnitude of type I and II error?
    b. What effect does the sigma level (high or low) have on alpha?
2) Laboratory glassware shipped from the manufacturer to Dr. Renner's Lab via an overnight package service has arrived damaged. Develop a cause-and-effect diagram that identifies and outlines the possible causes of this event. You won't necessarily be graded on the details, but include the major components of the "fishbone" diagram.
3) Sketch out diagrams and explain why it is important to control both process mean and variability (see Figure 6.1 in the book)
4) Problem 5.16, 5.17, 5.18 (5.19, 5.22, and 5.23in the 7th edition) – use the below control chart (you can cut and paste into as separate file and print it off as part of your answer). For 5.17 use Sensitizing Rules 5-10, and for 5.18 use the Western Electric Rules (1-4).



5) Problem 6.1
6) A hospital emergency department is monitoring the time require to admit a patient using $\bar{x}$ and R charts. The table below presents summary data for 20 subgroups of two patients each (time is in minutes)

Subgroup	Xbar	R
1	8.3	2
2	8.1	3
3	7.9	1
4	6.3	5
5	8.5	3
6	7.5	4
7	8	3
8	7.4	2
9	6.4	2
10	7.5	4
11	8.8	3
12	9.1	5
13	5.9	3
14	9	6
15	6.4	3
16	7.3	3
17	5.3	2
18	7.6	4
19	8.1	3
20	8	2

a) Use these data to determine the control limits for the $\bar{x}$ and R control charts for this patient admitting process.
b) Plot the preliminary data form the first 20 samples on the control charts that you set up in part (a). Is the process in statistical control?

7) A high-voltage power supply should have a nominal output voltage of 350V. A sample of four units is selected each day and tested for process-control purposes. The data shown in the table below give the difference between the observed reading on each unit and the nominal voltage times 10; that is xi = (observed voltage unit on i – 350)*10. Use Minitab to set up the $\bar{x}$ and R charts on this process. Is the process in statistical control?

Sample #	X1	X2	X3	X4
1	6	9	10	15
2	10	4	6	11
3	7	8	10	5
4	8	9	6	13
5	9	10	7	13
6	12	11	10	10

7	16	10	8	9
8	7	5	10	4
9	9	7	8	12
10	15	16	10	13
11	8	12	14	16
12	6	13	9	11
13	16	9	13	15
14	7	13	10	12
15	11	7	10	16
16	15	10	11	14
17	9	8	12	10
18	15	7	10	11
19	8	6	9	12
20	13	14	11	15

ECHE313 Homework 8 - Due 4/3/25 **Trevor Swan (tcs94)**

1. Control Charts

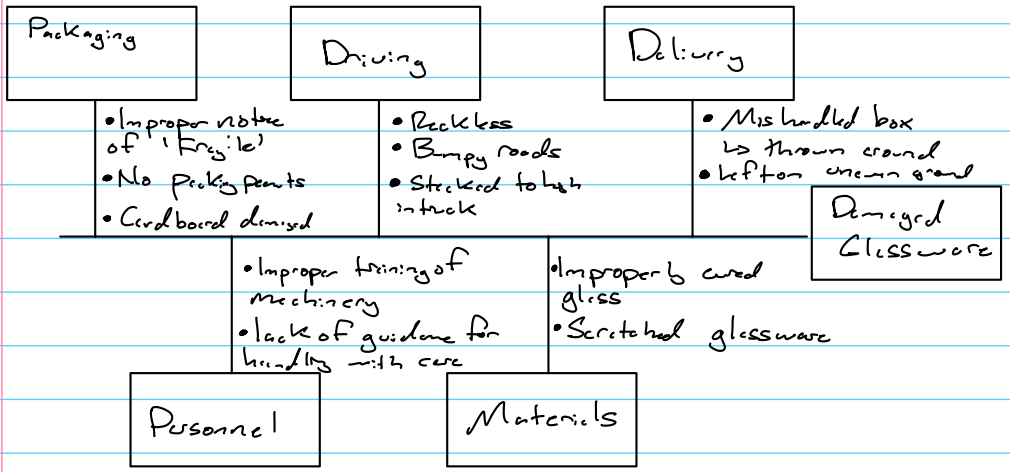
a) Effect of Narrower limits on mag. on type I & II Errors?

If we choose narrower limits, Type I error magnitude (false alarm) increases. Narrower limits → more points fall outside acceptable range → increased chance of false alarms. Type II error decreases bc, as narrower limits make it more likely to detect small shifts, reducing β.

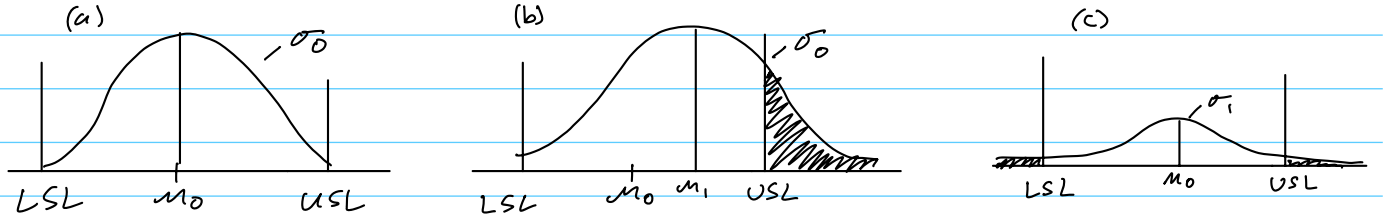
b) Effect of Sigma level on Alpha

A high sigma level increases the control limit range → harder for points to fall outside control limit range → reduces α, but increases β. Lower Sigma level results in tighter limits → more points fall outside → increases α but reduces β.

2. Cause & Effect Diagram for Glassware Shipping Error



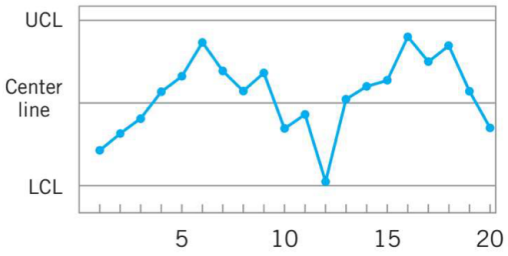
3. Sketches to explain why controlling mean & variability is needed



Normal Operation — Process mean $\mu_1 > \mu_0$ — Process std dev $\sigma_1 > \sigma_0$

Controlling mean ensures process lies within specs as in (a), while preventing systematic errors which cause consistent deviation like in (b). Controlling variability reduces out-of-spec variability as in (a), prevents situations like in (c). Controlling both ensures high quality, consistent output, within specs.

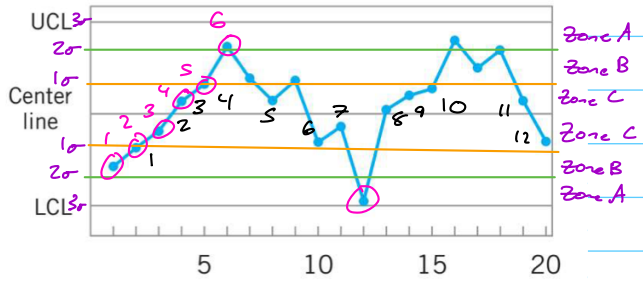
4.



a) Problem S.16

No, it does not appear random. The sharp dip in between two roughly random regions suggests assignable cause. The plot could also be viewed as sinusoidal, suggests a lack of randomness.

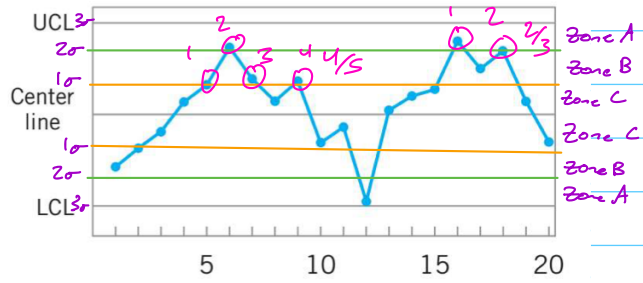
b) Problem S.17 – Sensitizing Rules (5-10)



5. Six points in a row steadily increasing or decreasing	✓
6. Fifteen points in a row in zone C (both above and below the center line)	✗
7. Fourteen points in a row alternating up and down	✗
8. Eight points in a row on both sides of the center line with none in zone C	✗
9. An unusual or nonrandom pattern in the data	✓
10. One or more points near a warning or control limit	✓

Rules 5, 9, and 10 are violated on this control chart, as noted directly on the chart. This indicates potentially out of control conditions which should be investigated.

c) Problem S.18 – Western Electric Rules (1-4)



1. One or more points outside of the control limits	✗
2. Two of three consecutive points outside the two-sigma warning limits but still inside the control limits	✓
3. Four of five consecutive points beyond the one-sigma limits	✓
4. A run of eight consecutive points on one side of the center line	✗

Rules 2 and 3 are violated on this chart, noted directly on the chart. This indicates a potential out of control process or conditions which should be investigated.

5. Problem 6.1

A manufacturer of components for automobile transmissions wants to use control charts to monitor a process producing a shaft. The resulting data from 20 samples of 4 shaft diameters that have been measured are:

$$\sum_{i=1}^{20} \bar{x}_i = 10.275, \quad \sum_{i=1}^{20} R_i = 1.012$$

a. Find the control limits that should be used on the $\bar{x}$ and R control charts.

b. Assume that the 20 preliminary samples plot in control on both charts. Estimate the process mean and standard deviation.

a) $\bar{\bar{x}}$ Chart

$UCL = \bar{\bar{x}} + A_2 \bar{R}$

$LCL = \bar{\bar{x}} - A_2 \bar{R}$

R Chart

$UCL = D_4 \bar{R}$

$LCL = D_3 \bar{R}$

Appendix ← n = 20

$A_2 = 0.180$

$D_3 = 0.415$

$D_4 = 1.585$

$\bar{\bar{x}} = \frac{\bar{x}}{n} = \frac{10.275}{20} = 0.51375$

$\bar{R} = \frac{R}{n} = \frac{1.012}{20} = 0.0506$

$\bar{x}$: $UCL = 0.51375 + (0.180)(0.0506) = 0.523$

$LCL = 0.51375 - (0.180)(0.0506) = 0.505$

R: $UCL = (1.585)\,0.0506 = 0.0802$

$LCL = (0.415)\,0.0506 = 0.0210$

$\bar{x}$: UCL = 0.523, LCL = 0.505

R: UCL = 0.0802, LCL = 0.0210

b) $\mu \approx \bar{\bar{x}} = 0.51375$

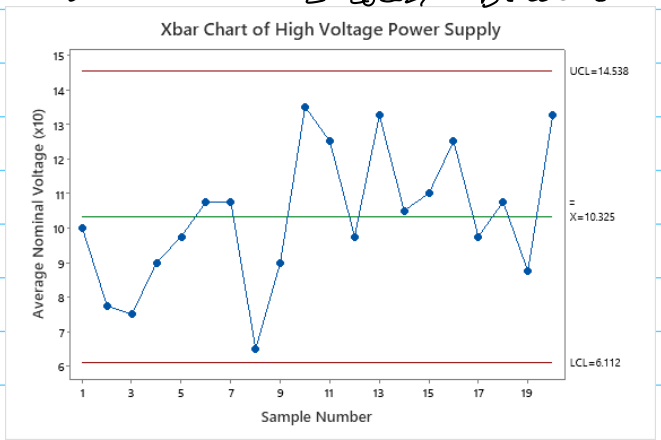
Appendix n = 20

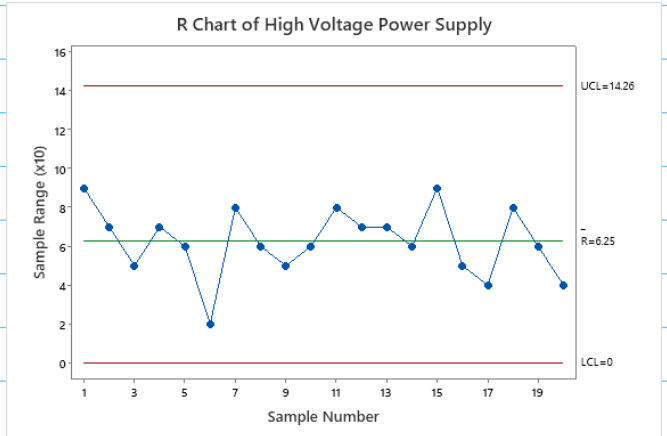
$d_2 = 3.735$

$\hat{\sigma} \approx \frac{\bar{R}}{d_2} = \frac{0.0506}{3.735} = 0.0135$

Estimates: $\mu = 0.514$, $\sigma = 0.014$

7. All points on the $\bar{x}$ and R charts are within the LCL and UCL shown below.





6.

Subgroup	Xbar	R
1	8.3	2
2	8.1	3
3	7.9	1
4	6.3	5
5	8.5	3
6	7.5	4
7	8	3
8	7.4	2
9	6.4	2
10	7.5	4
11	8.8	3
12	9.1	5
13	5.9	3
14	9	6
15	6.4	3
16	7.3	3
17	5.3	2
18	7.6	4
19	8.1	3
20	8	2

a) $n = 20$

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{20} \bar{x}_i}{20} = \frac{8.3 + 8.1 + 7.9 + \cdots + 8.1 + 8}{20} = 7.57$$

$$\bar{R} = \frac{\sum_{i=1}^{20} R}{20} = \frac{2 + 3 + 1 + \cdots + 3 + 2}{20} = 3.15$$

Apendix: $A_2 : 0.180$

$D_3 : 0.415$

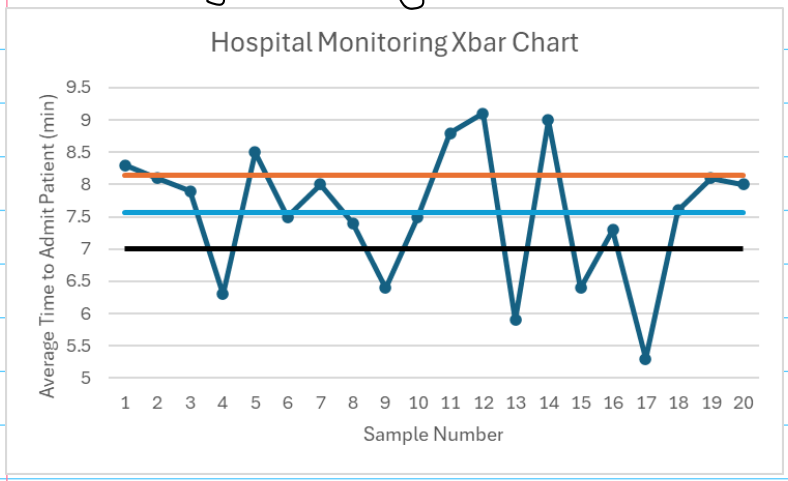
$D_4 : 1.585$

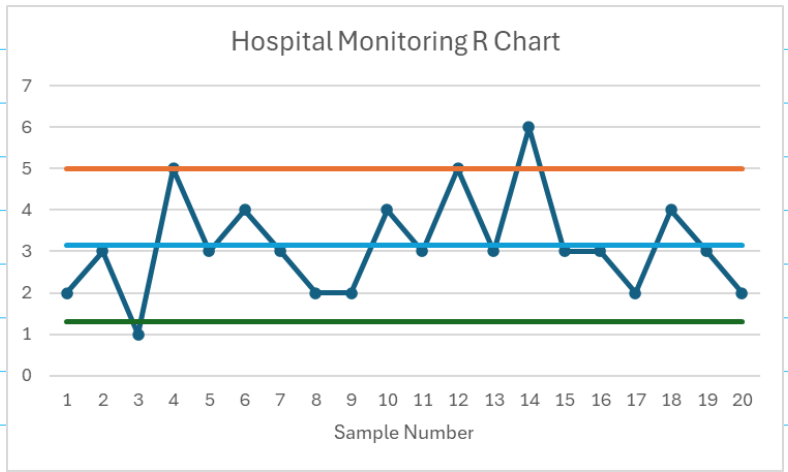
$$\bar{x}\begin{cases} UCL = 7.57 + 0.180(3.15) = 8.14 \\ LCL = 7.57 - 0.180(3.15) = 7 \end{cases}$$

$$R\begin{cases} UCL = 1.585(3.15) = 4.99 \\ LCL = 0.415(3.15) = 1.31 \end{cases}$$

$$\bar{x}\begin{cases} UCL = 8.14 \\ LCL = 7 \end{cases} \qquad R\begin{cases} UCL = 4.99 \\ LCL = 1.31 \end{cases}$$

b) Plots generated using excel





No, both the Xbar and R chart have points that lay outside the upper and lower control limits indicating that the process is not in Statistical control.