04b Sample Examination Problems Chapter 9 SOLUTIONS

1. (a) Explain and discuss the difference between one-way and two-way analysis of variance.

Whereas one-way analysis of variance (ANOVA) tests measure significant effects of one factor only, two-way analysis of variance (ANOVA) tests (also called two-factor analysis of variance) measure the effects of two factors simultaneously. For example, an experiment might be defined by two parameters, such as treatment and time point. One-way ANOVA tests would be able to assess only the treatment effect or the time effect. Two-way ANOVA on the other hand would not only be able to assess both time and treatment in the same test, but also whether there is an interaction between the parameters.

One way ANOVA
Investigates how much of variations in grouped data comes from differences between the groups and how much is just random observational error.
performs a hypothesis test:

H_0 : $\mu_1 = \mu_2 = ... = \mu_k$

H_1 : At least one is different i.e. $\mu_i \neq \mu_j$ **for i≠j**

Two way ANOVA : Allows us to analyze both row effect and column effect,i.e. differences between rows and differences between columns.
splits the variations among the observations into row effect , column effect and random error.
Performs a hypothesis test :

H_0 : $\alpha_i = 0$ **for all i (testing no row effect at all)**

H_1 : At least one is none zero

H_0 : $\beta_j = 0$ **for all j (testing no column effect at all)**

H_1 : At least one is none zero

(b) Explain qualitatively why in a one-way analysis of variance one rejects the null hypothesis of no differences between group means if the mean sum of squares between groups is large compared to the mean sum of squares within groups.

In one-way ANOVA ,We reject the null hypothesis :

H_0 : $\mu_1 = \mu_2 = \ldots\ldots = \mu_k$

When we obtain large value of the test statistic:

$\frac{S_B^2}{S^2}$ **~ $F_{n-1,n-k}$**

Where S_B^2 is the between group sum squared which is responsible for between group variations and S^2 is the within group sum squared which is responsible for random observational error.

Large value of the test statistic occurs when

$S_B^2 >> S^2$

If the null is not true , then there will be variations between the groups and hence S_B^2 will be large.

(c) The table below shows measurements of sections taken from five European larch trees of the same age. Each section gives rise to 4 measurements of the trachoid length from each of the four aspects North, South, East and West.

	Aspect			
Tree	East	South	West	North
1	3.4	3.5	3.1	3.5
2	2.8	3.1	3.0	3.0
3	3.0	3.2	3.3	3.3
4	3.0	3.0	2.5	2.8
5	3.3	3.5	3.7	3.6

i. Give the analysis of variance table for a two-way analysis of variance for these data, using the classification by aspects and by tree number.

ii. Test the hypothesis that there is no difference between the trachoid lengths from different aspects.

i. Assume that the trachoid length measurements are Normally distributed with constant population variance, σ^2

Let the population mean μ_{ij} for tree i in aspect j be of the

form $\mu_{ij} = \mu + \alpha_i + \beta_j$

<u>Check :</u> row data has no <u>outliers</u>

<u>Row means are</u> : $\bar{x}_{1.}$ **= 3.375**

$\bar{x}_{2.}$ **= 2.975**

$\bar{x}_{3.}$ **= 3.2**

$\bar{x}_{4.}$ **= 2.825**

$\bar{x}_{5.}$ **= 3.525**

Overall mean $\bar{x}_{..}$ **= 3.18**

The estimated row effects $\hat{\alpha}_i$ **are the differences between The row means and the overall mean**

$\hat{\alpha}_1$ **= 3.375 – 3.18 = 0.195**

$\hat{\alpha}_2$ **= 2.975 – 3.18 = -0.205**

$\hat{\alpha}_3$ **= 3.2 – 3.18 = 0.02**

$\hat{\alpha}_4$ **= 2.825 – 3.18 = -0.355**

$\hat{\alpha}_5$ **= 3.525 – 3.18 = 0.345**

Sample variance of row means : S_R^2 **= 0.081375**

Mean sum squares between rows = c S_R^2 **(c = No. of columns) = 4(0.081375) = 0.3255**

sum squares between rows = (r-1) c S_R^2 **(r= no. of rows)**

= 4(0.3255) = 1.302

<u>Columns means are</u> : $\bar{x}_{.1}$ **= 3.1**

$\bar{x}_{.2}$ **= 3.26**

$\bar{x}_{.3}$ **= 3.12**

$\bar{x}_{.4}$ **= 3.24**

Overall mean $\bar{x}_{..}$ **= 3.18**

The estimated column effects $\hat{\beta}_i$ **are the differences between the column means and the overall mean**

$\hat{\beta}_1$ **= 3.1 – 3.18 = 0.195**

$\hat{\beta}_2$ **= 3.26 – 3.18 = 0.08**

$\hat{\beta}_3$ **= 3.12 – 3.18 = -0.06**

$\hat{\beta}_4$ **= 2.24 – 3.18 = 0.06**

Sample variance of column means : S_C^2 **= 0.006667**

Mean sum squares between columns = $r S_C^2$
= 5(0.006667) = 0.0333

sum squares between columns = (c-1) r S_C^2
= 3(0.03333) = 0.1

Sample variance of all observations : $S_T^2 =$ **0.093263**

Total sum squares : (rc - 1) $S_T^2 =$ **19(0.093263)= 1.772**
Residual sum squares
= Total SS – SS between rows – SS between columns
= 1.772 - 1.302 - 0.1 = 0.37

Mean Residual SS = $S^2 = \frac{ResidualSS}{(r-1)(c-1)} = \frac{0.37}{4 \times 3} = 0.030833$

ANOVA TABLE $F_R = c S_R^2 / S^2 = 4(0.081375)/0.030833 = 10.56$

$F_C = r S_C^2 / S^2 = 5(0.006667)/\ 0.030833 = 1.08$

Source	D.F.	Sum squresSS	MSE	F-value
Trees(row)	r - 1	1.30	0.33	$F_R = c S_R^2 / S^2$
Aspects(column)	c - 1	0.1	0.03	$F_C = r S_C^2 / S^2$
Error	(r-1)(c-1)	0.37	0.03	
Total	rc – 1	1.77		

Source	D.F.	Sum squares SS	MSE	F-value
Trees(row)	4	1.30	0.33	10.56
Aspects(column)	3	0.1	0.03	1.08
Error	12	0.37	0.03	
Total	19	1.77		

ii. H_0 **: there is no difference between the trachoid lengths from different aspects**

H_0 : $\beta_j = 0 \quad \forall j$

H_1 **: non zero column effect**

Under H_0 : $F_C = \frac{rS_C^2}{S^2} \sim F_{c-1,(r-1)(c-1)}$

The criterion : Reject H_0 **if** $F_C \geq F_{\alpha, c-1,(r-1)(c-1)} = F_{0.05,3,12} = 3.49$

$r S_C^2 / s^2 = 1.08 < 3.49$ **, we do not reject** H_0 **and therefore there is no evidence to support a difference in trachoid length from different aspects.**

If you were asked to test rows effects(trees): $F_R = c S_R^2 / S^2 = 10.56$

Reject H_0 if $F_R \geq F_{\alpha, r-1, (r-1)(c-1)} = F_{0.05, 4, 12} = 3.26$

Since 10.56 > 3.26 , we reject H_0

---

2. (a) Sometimes it is suggested that one carries out an analysis of variance on the logarithms of the original data. Why might this be a sensible transformation?

   (b) The table below shows the percentage vote for the Democratic Party in US presidential elections of several different campaigns for different counties of Connecticut.

	Lich	Fairf	Middx	Toll
1920	32.5	30.9	33.1	31.0
1924	30.0	24.5	29.9	30.3
1928	36.0	43.7	39.7	39.6
1932	41.9	47.1	46.3	46.0

   i. Give the analysis of variance table for a two-way analysis of variance for these data, using the classification by counties and by years.

   ii. Are some year effects significantly different from 0?

   iii. Are these data suitable for this form of analysis?

(a) Logarithms are used to convert multiplication into addition. In two – way ANOVA we have the assumption that all the population means have addition structure :

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

If the initial variables consist of multiplication behavior , then a logarithm transformation is sensible.

(b) i. Assume that the votes percentages are Normally distributed with constant population variance, σ^2

Let the population mean μ_{ij} **for vote i in county j be of the form** $\mu_{ij} = \mu + \alpha_i + \beta_j$

Check : row data has no outliers

Row means are : $\bar{x}_{1.} = 31.875$

$\bar{x}_{2.} = 28.675$

$\bar{x}_{3.} = 39.75$

$\bar{x}_{4.} = 45.325$

Overall mean $\bar{x}_{..} = 36.40625$

Sample variance of row means : S_R^2 = 57.01

Mean sum squares between rows = c S_R^2 (c = No. of columns) = 228.04

sum squares between rows = (r-1) c S_R^2 (r= no. of rows)
= 684.12

<u>Columns means are</u> : $\bar{x}_{.1}$ = 35.1

$\bar{x}_{.2}$ = 36.55

$\bar{x}_{.3}$ = 37.25

$\bar{x}_{.4}$ = 36.725

Overall mean $\bar{x}_{..}$ = 36.40625

Sample variance of column means : S_C^2 = 0.846823

Mean sum squares between columns = r S_C^2 = 3.39

sum squares between columns = (c-1) r S_C^2 = 10.16

Sample variance of all observations : $S_T^2 = 50.47$

Total sum squares : (rc - 1) $S_T^2 =$ 757.03

Residual sum squares
= Total SS – SS between rows – SS between columns
= 757.03 - 684.12 - 10.16 = 62.75

Mean Residual SS = S^2 = $\dfrac{\text{Re}\,sidualSS}{(r-1)(c-1)} = \dfrac{62.75}{3\times 3} = 6.97$

ANOVA TABLE

Source	D.F.	Sum squresSS	MSE	F-value
Year(row)	r - 1	684.12	228.04	F_R= c S_R^2/S^2
county(column)	c - 1	10.16	3.39	F_C= r S_C^2/S^2
Error	(r-1)(c-1)	62.75	6.97	
Total	rc – 1	757.03		

Source	D.F.	Sum squares SS	MSE	F-value
Year(row)	3	684.12	228.04	32.71
County(column)	3	10.16	3.39	0.49
Error	9	62.75	6.97	
Total	15	757.03		

ii. H_0 : there is no difference between the vote percentage for the democratic party in different years.

H_0 : $\alpha_i = 0 \quad \forall i$

H_1 : non zero row effect

Under H_0 : $F_R = c S_R^2 / S^2$= 32.71 ~ $F_{r-1,(r-1)(c-1)}$

The criterion : Reject H_0 if Reject H_0 if

$$F_R \geq F_{\alpha, r-1,(r-1)(c-1)} = F_{0.05,3,9} = 3.86$$

Since 32.71 > 3.86 , we reject H_0

iii. Challenge assumptions :

Normally distributed

Constant population variance , σ^2

cell population means μ_{ij} have

additive structure: $\mu_{ij} = \mu + \alpha_i + \beta_j$

---

3. (a) Give a model for the two-way analysis of variance, specifying the distribution of any random variables included in your model.

2 – way ANOVA :

- We need a continuous r.v.
- 2 characteristics described as r rows , c columns such that the r.v. X_{ij} appears in the cell corresponding to the ith row and jth column i = 1,…,r and j= 1,….,c

-Define $E[X_{ij}] = \mu_{ij}$

-Key Model assumptions : X_{ij} are normal r.v.'s with mean μ_{ij} and constant variance σ^2 : $X_{ij} \sim N(\mu_{ij}, \sigma^2)$

-Additional assumption: cell population means μ_{ij} have additive structure: $\mu_{ij} = \mu + \alpha_i + \beta_j$

μ : overall mean , α_i : row effect , β_j : column effect

(b) Explain what is meant by interaction in a two-way analysis of variance.

Additive structure $\mu_{ij} = \mu + \alpha_i + \beta_j$ allows us to talk clearly about differences in row/column differences
2-way ANOVA allows us to analyze the interaction between the two variables (row variables & column variables)so we can study how combinations of these variables influence behavior.
Interaction describes how the effect of one independent variable is influenced/effected by the value of the other independent variable.
Equivalently , the effect of one independent variable varies with the value of the other independent variable.

(c) The table below shows the values of price index numbers for glasshouse fruit and vegetables (with base January 1969 at 100).

i. Give the analysis of variance table for a two-way analysis of variance for these data, using the classification by years and by months.

Usual 2-way ANOVA assumptions
Assume that the entries are Normally
distributed with constant population variance, σ^2
Let the population mean μ_{ij} for vote i in county j be of the form $\mu_{ij} = \mu + \alpha_i + \beta_j$
<u>Check :</u> row data has no <u>outliers</u>

<u>Row means are</u> : $\bar{x}_{1.}$ **= 207.04**
$\bar{x}_{2.}$ **= 201.0**
$\bar{x}_{3.}$ **= 216.0**
$\bar{x}_{4.}$ **= 192.0**
$\bar{x}_{5.}$ **= 246.2**

Overall mean $\bar{x}_{..}$ **= 212.52**
Sample variance of row means : S_R^2 **= 431.612**
Mean sum squares between rows = c S_R^2 **= 2158.06**
sum squares between rows = (r-1) c S_R^2 **=4(2158.06)= 8632.24**

<u>Columns means are</u> : $\bar{x}_{.1}$ = 261.6

$\bar{x}_{.2}$ = 261.8

$\bar{x}_{.3}$ = 210.4

$\bar{x}_{.4}$ = 193.0

$\bar{x}_{.4}$ = 135.8

Overall mean $\bar{x}_{..}$ = 212.52

Sample variance of column means : S_C^2 = 2777.212

Mean sum squares between columns = rS_C^2 = 13886.06

sum squares between columns = (c-1) r S_C^2 = 55544.24

Sample variance of all observations : $S_T^2 =$ 3601.76

Total sum squares : (rc - 1) $S_T^2 =$ 86442.24

Residual sum squares
= Total SS – SS between rows – SS between columns
= 86442.24 – 8632.24 – 55544.24 = 22265.76

Mean Residual SS = S^2 = $\dfrac{ResidualSS}{(r-1)(c-1)} = \dfrac{22265.76}{16} = 1391.61$

ANOVA TABLE

Source	D.F.	Sum squresSS	MSE	F-value
Year(row)	r - 1	8632.24	2158.06	$F_R = c S_R^2 / S^2$
month(column)	c - 1	55544.24	13886.06	$F_C = r S_C^2 / S^2$
Error	(r-1)(c-1)	22265.76	1391.06	
Total	rc – 1	86442.24		

Source	D.F.	Sum squares SS	MSE	F-value
Year(row)	4	8632.24	2158.06	1.55
County(column)	4	55544.24	13886.06	9.98
Error	16	22265.76	1391.06	
Total	24	86442.24		

If you were to test H_0 (rows effect) : H_0 : $\alpha_i = 0 \quad \forall i$

$F_{0.05,4,16}$ = 6.39 , since 1.55 < 6.39 , do not reject H_0

If you were to test H_0 (rows effect) : H_0 : $\beta_j = 0 \quad \forall j$

Reject H_0 .

ii. Give a set of 90% simultaneous confidence intervals for the differences between the first three years.

	Jan	Feb	March	April	May
1970	261	276	193	160	147
1971	214	239	193	2210	138
1972	332	248	208	164	128
1973	173	232	199	211	145
1974	328	314	259	209	121

Simultaneous C.I. (SCI) uses the Scheffe's method and the F distribution , used for all possible pairs of α_i's :

A(1-α)% set of SCI for every linear combination $\sum_{i=1}^{k} d_i \alpha_i$

Where $\sum_{i=1}^{k} d_i = 0$ is given by :

$$\sum_{i=1}^{k} d_i \overline{X}_i \pm S\sqrt{(r-1)F_{\alpha,r-1,(r-1)(c-1)} \Sigma d_i^2 / c}$$

The confidence interval in case of difference of population means α_i - α_j :

$$\overline{X}_i - \overline{X}_j \pm S\sqrt{(r-1)F_{\alpha,r-1,(r-1)(c-1)}(2/c)}$$

Here we seek 3 SCI :differences between the first 3 years 1970-1971 , 1970-1972 , 1971-1971

α_1 - α_2 , α_1 - α_3 , α_2 - α_3:

$\overline{X}_i - \overline{X}_j \pm S\sqrt{(r-1)F_{0.1,4,16}(2/c)}$ **, $F_{0.1,4,16}$ = 2.33**

For α_1 - α_2 , (1970 , 1971) :

207.4 – 201.0 $\pm$ $\sqrt{1391.61}\sqrt{4\times 2.33\times 2/5}$

= 6.4 $\pm$ 72.07 = (-65.67 , 78.47)

For α_1 - α_3 , (1970,1972)

207.4 – 216.0 $\pm$ 72.07 = (-80.67,63.47)

For α_2 - α_3 , (1971 , 1972) :

201.0 – 216.0 $\pm$ 72.07 = (-87.07,57.07)

Remark: Note that all SCI's contain 0

Compare F_R = 1.55 , H_0 : $\alpha_i = 0 \quad \forall i$, here all possible pairs of SCI would include 0