Handout 1

Two-factor studies (balanced)

Consider a two-factor case where factor A has a levels, factor B has b levels, and there are n observations for each of the ab combinations of the factors. For the Hay Fever Relief example, 9 compounds for hay fever relief are made by varying levels of the two basic ingredients. Ingredient 1 (factor A) can be at $a = 3$ levels: low ($i = 1$), medium ($i = 2$) and high ($i = 3$). Similarly, ingredient 2 (factor B) has $b = 3$ levels: low ($j = 1$), medium ($j = 2$) and high ($j = 3$). A total of 36 subjects (suffering from hay fever) are selected and each of the 9 compounds are given to randomly selected $n = 4$ individuals. Let Y_{ijk} be "hours of relief" for the k^{th} individual, $k = 1, \ldots, n = 4$, who is given the compound with ingredient 1 is at level i and ingredient 2 is at level j. Let μ_{ij} be the population mean when factor A (ingredient 1) is at level i and factor B (ingredient 2) is at level j. The cell means model for the data is

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, k = 1, ..., n, j = 1, ..., b, i = 1, ..., a,$$

where ε_{ijk}'s are independent $N(0, \sigma^2)$. We can interpret this model as follows. There are ab normal populations with means μ_{11}, μ_{12},...,μ_{ab}, and these populations have the same variance σ^2. As in the one-factor case, we can check these assumptions of equal variance, normality, independence etc. using graphical as well as formal methods. For the model above, the total number of observations is equal to $n_T = nab$.

Decomposition of μ_{ij}

Let

$$\mu_{\cdot\cdot} = \sum\sum \mu_{ij}/(ab), \quad \mu_{i\cdot} = \sum \mu_{ij}/b, \quad \mu_{\cdot j} = \sum \mu_{ij}/a.$$

Define

$$\alpha_i = \mu_{i\cdot} - \mu_{\cdot\cdot}, \quad \beta_j = \mu_{\cdot j} - \mu_{\cdot\cdot}, \quad (\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot} = \mu_{ij} - (\mu_{\cdot\cdot} + \alpha_i + \beta_j).$$

Here $\mu_{\cdot\cdot}$ is the overall mean, α_i's are the main effects of factor A, β_j's are the main effects of factor B, and $(\alpha\beta)_{ij}$'s are the interactions. Note that we can reexpress μ_{ij} as

$$\begin{aligned} \mu_{ij} &= \mu_{\cdot\cdot} + (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot}) + (\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}), \quad \textit{or} \\ \mu_{ij} &= \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij}. \end{aligned}$$

In words, we can explain this as follows. The mean μ_{ij} is the sum of an overall mean $\mu_{\cdot\cdot}$, the main effect α_i of factor A, the main effect β_j of factor B and the interaction effect $(\alpha\beta)_{ij}$
When the interaction effects are absent, i.e., $(\alpha\beta)_{ij} = 0$ for all i and j, we have

$$\mu_{ij} = \mu_{\cdot\cdot} + \alpha_i + \beta_j.$$

This is called an **additive model** for μ_{ij} , i.e., it is additive in the factor effects.

The main effects and the interactions satisfy some constraints

$$\sum_{i=1}^{a} \alpha_i = 0, \sum_{j=1}^{b} \beta_j = 0,$$

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 \text{ for all } j, \; \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0 \text{ for all } i.$$

Example 1. Consider the following artificial data for two states (California and Colorado) where the mean recovery times from knee surgery are recorded from the data available from the hospitals. For any state, let μ_{ij} denote the mean recovery time when the patient is at physical fitness status i and age group j, where

$i = 1$: below average, $i = 2$:, average, and $i = 3$: above average

$j = 1$: young, $j = 2$: old.

Thus there are two factors: fitness (factor A) at $a = 3$ levels and age (factor B) at $b = 2$ levels.

	Table of μ_{ij}'s						
	California				Colorado		
	$j = 1$	$j = 2$	$\mu_{i\cdot}$		$j = 1$	$j = 2$	$\mu_{i\cdot}$
$i = 1$	21	27	24		23	25	24
$i = 2$	18	24	21		19	23	22
$i = 3$	12	18	15		9	21	15
$\mu_{\cdot j}$	17	23	$\mu_{\cdot\cdot} = 20$		17	23	$\mu_{\cdot\cdot} = 20$

For California, $\alpha_1 = 4$, $\alpha_2 = 1$, $\alpha_3 = -5$, and, $\beta_1 = -3$ and $\beta_2 = 3$.

For Colorado, $\alpha_1 = 4$, $\alpha_2 = 1$, $\alpha_3 = -5$, and, $\beta_1 = -3$ and $\beta_2 = 3$.

The following is the table of interactions.

	Table of $(\alpha\beta)_{ij}$'s						
	California				Colorado		
	$j = 1$	$j = 2$	row sum		$j = 1$	$j = 2$	row sum
$i = 1$	0	0	0		2	-2	0
$i = 2$	0	0	0		1	-1	0
$i = 3$	0	0	0		-3	3	0
col sum	0	0	0		0	0	0

Thus we have

$$\mu_{ij} = \mu_{\cdot\cdot} + \alpha_i + \beta_j, \text{ California},$$

$$\mu_{ij} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \text{ Colorado}.$$

Clearly, the means $\{\mu_{ij}\}$ for California can be described by an additive model.

For California, changes in mean recovery time from $j = 1$ to $j = 2$ are the same for the three fitness groups and that change is equal to 6, i.e., the change in mean recovery time from $j = 1$ to $j = 2$ does not depend on i. Therefore the interaction effects are not present and this results in parallel mean plots.

For Colorado, changes in mean recovery time from $j = 1$ to $j = 2$ are 2 (at $i = 1$), 4 (at $i = 4$) and 12 (at $i = 3$), respectively, i.e., the change in mean recovery time from $j = 1$ to $j = 2$ depends on the fitness status i. Thus the interaction effects are present and the mean plot lacks parallelism.

Model for the data $\{Y_{ijk}\}$

Let us recall that for a two-factor study we have considered two models. A cell means model

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk},$$

and the factor effects model

$$Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}.$$

Note that the factor effects model and the cell means model are the same model. One is simply a reformulation of the other. However, the additive model for the data is different since it is applicable only if we can ascertain that the interaction are absent. The data model for additive effects is

$$Y_{ijk} = \mu_{\cdot\cdot} + \alpha_i + \beta_j + \varepsilon_{ijk}.$$

Analysis of additive models may be a bit different from the full model and we come back to this issue later.

Why additive model?

It is always preferable to work with an additive model (instead of the full model with interactions) if such a model provides an adequate description of the data. Here are two principal reasons.

(a) An additive model has fewer parameters in comparison to a full model. In general, we should look for a model with as few parameters as possible. Estimation of a model with fewer parameters has more accuracy than a model with many parameters.

(b) An additive model has the nice property that in order to understand the influence of the factors, we need to examine only the main effects.

If μ_{ij} is not additive in the factors, sometimes a transformation (Box-Cox type) can make it additive. Here is an example from the text.

Example 2: Suppose the treatment means $\{\mu_{ij}\}$ are given in the following table

	Factor B	
Factor A	$j = 1$	$j = 2$
$i = 1$	16	64
$i = 2$	49	121
$i = 3$	64	144

Note that the values of $\mu_{i2} - \mu_{i1}$ are different for $i = 1, 2, 3$. This tells us interaction effects are present.

Now consider the table of square root $\{\sqrt{\mu_{ij}}\}$ of the means.

	Factor B	
Factor A	$j = 1$	$j = 2$
$i = 1$	4	8
$i = 2$	7	11
$i = 3$	8	12

Note that the interaction effects are zero for the square roots of the means. So $\sqrt{\mu_{ij}}$ is additive in the factor effects.

Estimation

We now obtain the parameter estimates for a two-factor balanced study where the observations are $\{Y_{ijk} : k = 1, \ldots, n, j = 1, \ldots, b, i = 1, \ldots, a\}$. Define:

$$\bar{Y}_{...} = \sum\sum\sum Y_{ijk}/(nab), \quad \bar{Y}_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}/(nb),$$
$$\bar{Y}_{.j.} = \sum_{i=1}^{a}\sum_{k=1}^{n} Y_{ijk}/(nb), \quad \bar{Y}_{ij.} = \sum_{k=1}^{n} Y_{ijk}/n.$$

Estimates of the parameters are given in the following table.

Parameter	μ_{ij}	$\mu_{..}$	$\mu_{i.}$	$\mu_{.j}$	α_i	β_j	$(\alpha\beta)_{ij}$
Estimate	$\bar{Y}_{ij.}$	$\bar{Y}_{...}$	$\bar{Y}_{i..}$	$\bar{Y}_{.j.}$	$\bar{Y}_{i..} - \bar{Y}_{...}$	$\bar{Y}_{.j.} - \bar{Y}_{...}$	$\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...} = \overline{Y}_{ij.} - (\overline{Y}_{...} + \hat{\alpha}_i + \hat{\beta}_j)$

Estimates of the factor effects satisfy the constraints

$$\sum_{1 \le i \le a} \hat{\alpha}_i = 0, \sum_{1 \le j \le b} \hat{\beta}_j = 0,$$
$$\sum_{1 \le i \le a} \widehat{(\alpha\beta)}_{ij} = 0 \text{ for all } j, \sum_{1 \le j \le b} \widehat{(\alpha\beta)}_{ij} = 0 \text{ for all } i.$$

Fitted values and residuals

The fitted values and the residuals are $\hat{Y}_{ijk} = \bar{Y}_{ij.}$ and $e_{ijk} = Y_{ijk} - \bar{Y}_{ij.}$. Note that the residual e_{ijk} is an estimate of ε_{ijk}.

Residuals sum of squares is

$$SSE = \sum\sum\sum e_{ijk}^2 = \sum\sum\sum (Y_{ijk} - \overline{Y}_{ij.})^2,$$
$$df(SSE) = n_T - ab = nab - ab = (n-1)ab$$

The mean square error is defined as $MSE = \frac{SSE}{df(SSE)} = \frac{SSE}{(n-1)ab}$. The common variance of the populations σ^2 is estimated by MSE and it is a consequence of the fact that $E(MSE) = \sigma^2$.

Decomposition of the total sum of squares

The total sum of squares is

$$SSTO = \sum\sum\sum(Y_{ijk} - \bar{Y}_{\cdots}),\ df(SSTO) = n_T - 1 = nab - 1.$$

The residual sum of squares is

$$SSE = \sum\sum\sum e_{ijk}^2 = \sum\sum\sum(Y_{ijk} - \overline{Y}_{ij\cdot})^2, df = (n-1)ab.$$

Sum of squares due the main effects of factor A is

$$SSA = \sum\sum\sum \hat{\alpha}_i^2 = nb\sum_{i=1}^{a} \hat{\alpha}_i^2,\ df(SSA) = a - 1.$$

Sum of squares due to the main effects of factor B is

$$SSB = \sum\sum\sum \hat{\beta}_j^2 = na\sum_{j=1}^{b} \hat{\beta}_j^2,\ df(SSB) = b - 1.$$

Sum of squares due to interaction effects is

$$SSAB = \sum\sum\sum \widehat{(\alpha\beta)}_{ij}^2 = n\sum_{i=1}^{a}\sum_{j=1}^{b} \widehat{(\alpha\beta)}_{ij}^2,\ df(SSAB) = (a-1)(b-1).$$

The following decompositions of the total sum squares and its df hold

$$SSTO = SSA + SSB + SSAB + SSE,$$
$$df(SSTO) = df(SSA) + df(SSB) + df(SSAB) + df(SSE).$$

Mean squares

As usual the mean squares are defined as the sums of squares divided by their degrees of freedom. So we have

$$MSA = \frac{SSA}{a-1},\ MSB = \frac{SSB}{b-1},\ MSAB = \frac{SSAB}{(a-1)(b-1)},\ MSE = \frac{SSE}{(n-1)ab}.$$

The following fact is important as it provides intuitive justifications for various tests.

Fact 1: (i) $E(MSE) = \sigma^2$, (ii) $E(MSA) = \sigma^2 + nb\sum\alpha_i^2/(a-1)$, (iii) $E(MSB) = \sigma^2 + na\sum\beta_j^2/(b-1)$, (iv) $E(MSAB) = \sigma^2 + n\sum\sum(\alpha\beta)_{ij}^2/((a-1)(b-1))$.

F-tests:

For each of the F-tests below, the level of significance is assumed to be equal to α.

Test for interactions: $H_0 : (\alpha\beta)_{ij} = 0$ for i and j, H_1 : not all $(\alpha\beta)_{ij}$ are zero.

Let $F^* = MSAB/MSE,\ df = ((a-1)(b-1), (n-1)ab)$.

Reject H_0 if $F^* > F(1-\alpha; (a-1)(b-1), (n-1)ab)$.

Test for the main effect of factor A: $H_0 : \alpha_i = 0$ for all i, H_1 : not all α_i are zero.

Let $F^* = MSA/MSE$, $df = (a-1, (n-1)ab)$.

Reject H_0 if $F^* > F(1-\alpha; a-1, (n-1)ab)$.

Test for the main effect of factor B: $H_0 : \beta_j = 0$ for all j, H_1 : not all β_j are zero.

Let $F^* = MSB/MSE$, $df = (b-1, (n-1)ab)$.

Reject H_0 if $F^* > F(1-\alpha; b-1, (n-1)ab)$.

Justification of the F-tests.

Before we state the main results, let us see intuitively why the F-tests given above are justified. Consider the test for the main effects for factor A: $H_0 : \alpha_i = 0$ for all i, H_1 : not all α_i are zero. Since

$$E(MSA) = \sigma^2 + nb\sum \alpha_i^2/(a-1),$$

from part (ii) of Fact 1, we can intuitively say that MSA fluctuates about $\sigma^2 + nb\sum \alpha_i^2/(a-1)$. Similarly we can say that MSE fluctuates about σ^2. Thus, we expect the F-statistic $F = MSA/MSE$ to fluctuate about

$$\left[\sigma^2 + nb\sum \alpha_i^2/(a-1)\right]/\sigma^2 = 1 + nb\sum \alpha_i^2/[\sigma^2(a-1)].$$

Hence $F = MSA/MSE$ fluctuates about 1 when and only when H_0 is true, otherwise it fluctuates about a quantity larger than 1. So if the value of $F = MSA/MSE$ obtained from the data is quite a bit larger than 1, it indicates that H_0 may be false. How large is large? F-distribution comes to the rescue and it provides a cutoff point (critical value) to decide of the observed F value is indeed much larger than 1.

These same intuitive arguments are also valid for testing for the main effects of B and for testing for interaction effects.

Fact 2. The following are true.

(a) Under $H_0 : \alpha_i = 0$ for all i, $F = MSA/MSE \sim F_{a-1,(n-a)ab},.$

(b) Under $H_0 : \beta_j = 0$ for all j, $F = MSB/MSE \sim F_{b-1,(n-a)ab}$,

(c) Under $H_0 : (\alpha\beta)_{ij} = 0$ for i and j, $MSAB/MSE \sim F_{(a-1)(b-1),(n-1)ab}$.

Remark 1 *What is the distribution of $F = MSA/MSE$ when "$H_0 : \alpha_i = 0$ for all i" is false? In general, it has a non-central F-distribution with degrees of freedom $(a-1, (n-1)ab)$ and non-centrality parameter $nb\sum \alpha_i^2/[\sigma^2(a-1)]$. Under H_0, the non-centrality parameter equals 0 and the non-central F distribution is same as the usual (central) F distribution with df $(a-1, (n-1)ab)$. Non-central F distribution has a very complicated form, but is useful in calculating the power of an F-test.*

The other F-statistics $F = MSB/MSE$ and $F = MSAB/MSE$ also have non-central F-distributions with corresponding degrees of freedom and non-centrality parameters.

Hay fever relief

For the Hay Fever Relief data, $a = 3, b = 3, n = 4$ and the total number of observations is $n_T = nab = 36$. Here Y_{ijk} is the hours of relief for the k^{th} patient when the compound has ingredient 1 at level i and ingredient 2 is at level j, $[i = 1, \ldots a, j = 1, \ldots b.]$

The table of estimated means and the main effects is given below.

	$\bar{Y}_{ij\cdot}$				
	$j = 1$	$j = 2$	$j = 3$	$\bar{Y}_{i\cdot\cdot}$	$\hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot}$
$i = 1$	2.475	4.600	4.575	3.883	-3.3
$i = 2$	5.450	8.925	9.125	7.833	0.65
$i = 3$	5.975	10.275	13.250	9.833	2.65
$\bar{Y}_{\cdot j\cdot}$	4.633	7.933	8.983	$\bar{Y}_{\cdot\cdot\cdot} = 7.183$	
$\hat{\beta}_j = \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdot\cdot\cdot}$	-2.55	0.75	1.80		

The table of estimated interaction effects is given below.

	$\widehat{(\alpha\beta)}_{ij}$			
	$j = 1$	$j = 2$	$j = 3$	Row sum
$i = 1$	1.142	-0.033	-1.108	-0.001
$i = 2$	0.167	0.342	-0.508	0.001
$i = 3$	-1.308	-0.308	1.617	0.001
Col. sum	0.001	0.001	0.001	

Note that the row sums and the column sums should be equal to zero. But they are not due to round off errors.

The Analysis of variance table for this data set is given below.

Source	df	SS	MS	F	p-value
Factor A	$a - 1 = 2$	220.020	110.010	1827.86	0.000
Factor B	$b - 1 = 2$	123.660	61.830	1027.33	0.000
Interaction	$(a - 1)(b - 1) = 4$	29.425	7.356	122.23	0.000
Error	$(n - 1)ab = 27$	1.625	0.060		
Total	$nab - 1 = 35$	374.730			

Here the interaction effects and main effects of factors A and B are statistically significant, i.e., we can reject all the three hypotheses $H_0 : (\alpha\beta)_{ij} = 0$ for all i and j, $H_0 : \alpha_i = 0$ for all i and $H_0 : \beta_j = 0$ for all j.

Diagnostics:

The methods for diagnostics are the same as in the one-factor case. You will find below the plot of the residuals $\{e_{ijk}\}$ against the fitted values $\{\hat{Y}_{ijk} = \bar{Y}_{ij\cdot}\}$ and the normal probability plot of the residuals. Plot of residuals against fitted values indicate that the assumption of equal variance is reasonable. The normal probability plot indicate the assumption of normality is also reasonable.

Appendix: Justification of Technical Results.

It is important to note that, for any linear model including ANOVA models, the fitted Y-values are independent of the residuals when the error terms are assumed to be iid $N(0, \sigma^2)$. Thus, in the two factor ANOVA case as discussed here, $\{\hat{Y}_{ijk} = \bar{Y}_{ij\cdot}\}$ are independent of $\{e_{ijk} = Y_{ijk} - \bar{Y}_{ij\cdot}\}$. This result is the backbone of all the inferential results such as various F-tests and t-tests, and construction of confidence intervals.

There is an elementary result from STA 131A that is very useful.

Fact 3: (a) Let $X_1, \ldots, X_n$ be iid with mean μ and variance σ^2, then $E\left[\sum_{k=1}^n (X_i - \bar{X})^2\right] = (n-1)\sigma^2$.
(b) If $\{X_i\}$ are assumed to be iid $N(\mu, \sigma^2)$, then $\sum(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$.

Proof of Fact 1:

(i) Check that

$$e_{ijk} = Y_{ijk} - \bar{Y}_{ij\cdot} = \varepsilon_{ijk} - \bar{\varepsilon}_{ij\cdot}.$$

Fix i and j, and then by Fact 3 (a),

$$E\left[\sum_{k=1}^n e_{ijk}^2\right] = E\left[\sum_{k=1}^n (\varepsilon_{ijk} - \bar{\varepsilon}_{ij\cdot})^2\right] = (n-1)\sigma^2.$$

Consequently,

$$E(SSE) = E\left[\sum_i \sum_j \sum_k e_{ijk}^2\right] = \sum_i \sum_j E\left[\sum_{k=1}^n e_{ijk}^2\right]$$
$$= \sum_i \sum_j (n-1)\sigma^2 = (n-1)ab\sigma^2.$$

.This shows that $E(MSE) = \sigma^2$.

(ii) Check that

$$\hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot} = \alpha_i + \bar{\varepsilon}_{i\cdot\cdot} - \bar{\varepsilon}_{\cdot\cdot\cdot}.$$

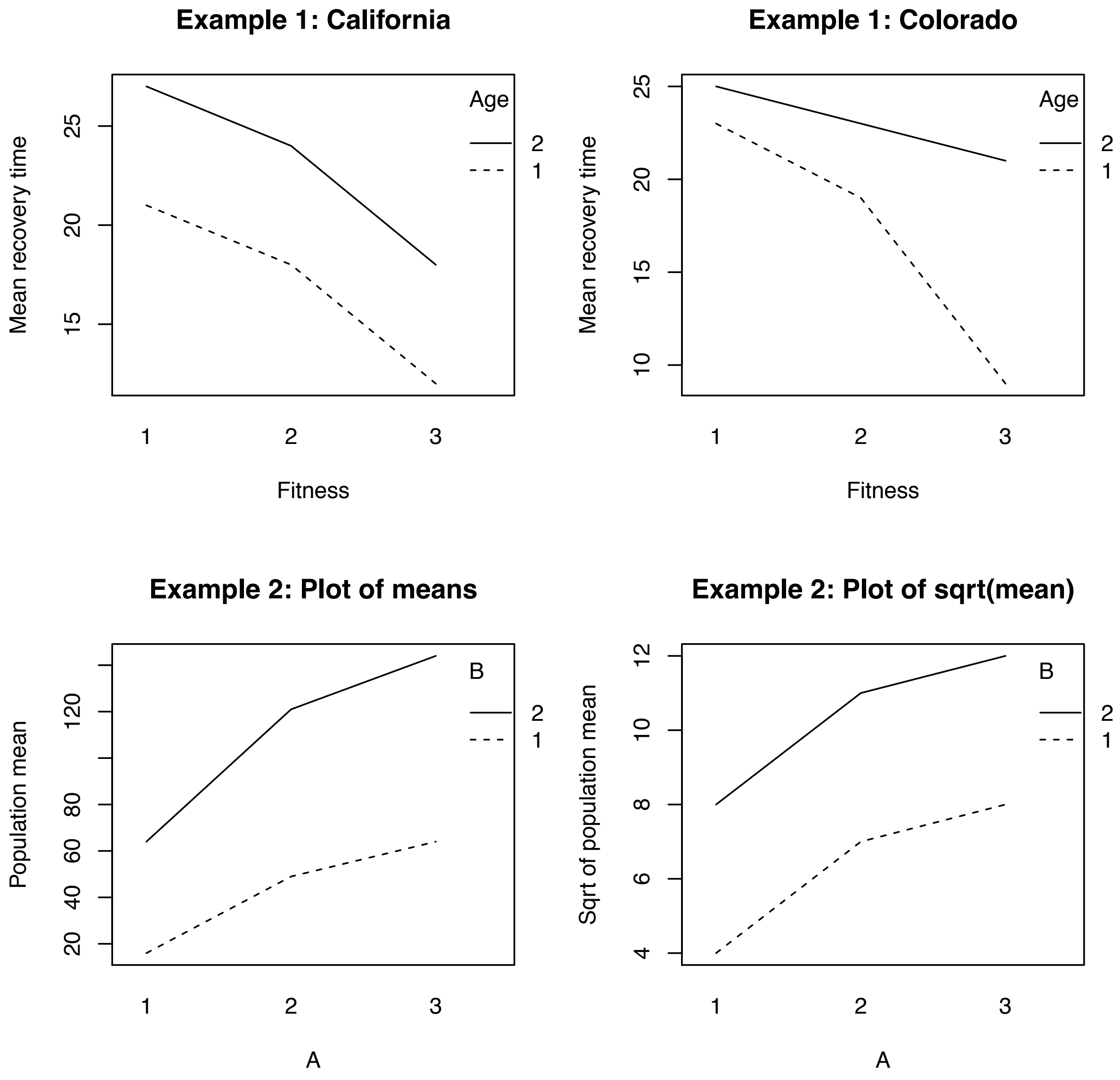
Hence

$$\sum \hat{\alpha}_i^2 = \sum (\bar{\varepsilon}_{i\cdot\cdot} - \bar{\varepsilon}_{\cdot\cdot\cdot})^2 + 2\sum (\bar{\varepsilon}_{i\cdot\cdot} - \bar{\varepsilon}_{\cdot\cdot\cdot})\alpha_i + \sum \alpha_i^2.$$

Since $\{\bar{\varepsilon}_{i\cdot\cdot} : i = 1, \ldots, a\}$ are iid with mean 0 and variance $\sigma^2/(nb)$, $\bar{\varepsilon}_{\cdot\cdot\cdot} = (1/a)\sum \bar{\varepsilon}_{i\cdot\cdot}$, we have using Fact 3(a)

$$E\left[\sum (\bar{\varepsilon}_{i\cdot\cdot} - \bar{\varepsilon}_{\cdot\cdot\cdot})^2\right] = (a-1)[\sigma^2/(nb)].$$

Figure 1: Examples 1 and 2



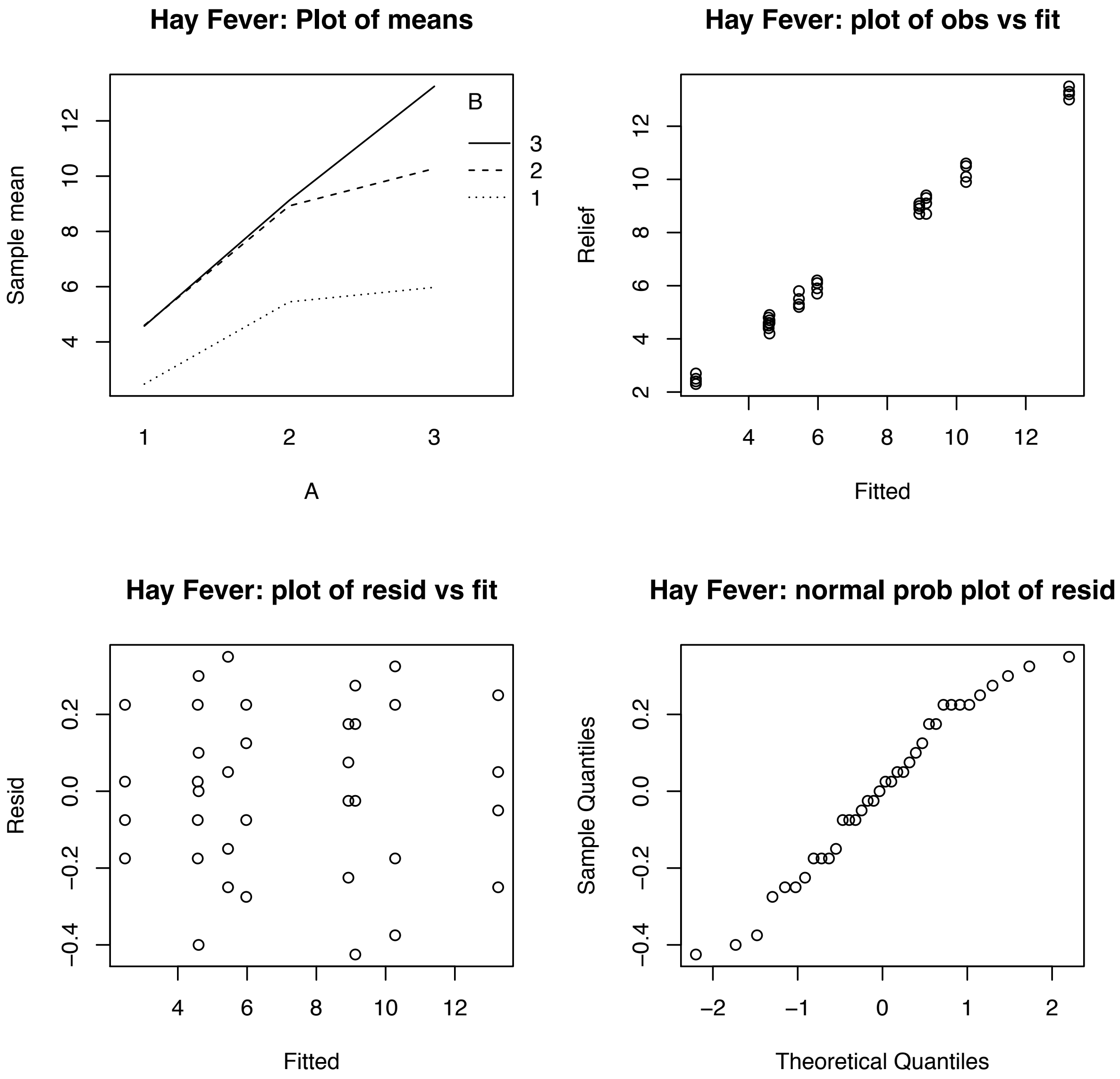
Example 1: California

Example 1: Colorado

Example 2: Plot of means

Example 2: Plot of sqrt(mean)

Figure 2: Hay Fever Data



Consequently,

$$\begin{aligned} E\left[\sum \hat{\alpha}_i^2\right] &= (a-1)[\sigma^2/(nb)] + \sum \alpha_i^2, \text{ and} \\ E[SSA] &= nb\left\{(a-1)[\sigma^2/(nb)] + \sum \alpha_i^2\right\} \\ &= (a-1)\sigma^2 + nb\sum \alpha_i^2. \end{aligned}$$

This validates the expression for $E[MSA]$.
(iii) and (iv) The justifications for these parts are the same as those for part (ii).

Proof of Fact 2.

At the beginning of the Appendix we have mentioned a basic result. When the errors $\{\varepsilon_{ijk}\}$ are iid $N(0, \sigma^2)$, the fitted Y-values $\{\hat{Y}_{ijk} = \bar{Y}_{ij\cdot}\}$ and the residuals $\{e_{ijk} = Y_{ijk} - \bar{Y}_{ij\cdot}\}$ are independent. Note that MSE is a function of the residuals, and MSA, MSB and $MSAB$ are functions of the fitted Y-values. Thus MSE is independent of MSA, MSB and $MSAB$.

Next note that. for any given levels i and j of factors A and B, we have n residuals $\{e_{ijk} : k = 1, \ldots, n\}$ and, by part (b) of Fact 3, $\sum_{k=1}^n e_{ijk}^2/\sigma^2 = \sum_{k=1}^n (\varepsilon_{ijk} - \bar{\varepsilon}_{ij\cdot})^2/\sigma^2 \sim \chi^2_{n-1}$. Since factor A has a levels and factor B has b levels, there are a total of ab such collection of residuals and they are all independent. Since the sum of independent chi-square random variables also follow a chi-square distribution whose degrees of freedom is the sum of the df's of the component ch-squares, we can conclude that (denoting SSE/σ^2 by R_0^2)

$$R_0^2 = SSE/\sigma^2 = \sum_{i=1}^a \sum_{j=1}^b \left[\sum_{k=1}^n e_{ijk}^2/\sigma^2\right] \sim \chi^2_{(n-1)ab}.$$

(a) When $\alpha_i = 0$ for all i, we have $\hat{\alpha}_i = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot} = \bar{\varepsilon}_{i\cdot\cdot} - \bar{\varepsilon}_{\cdot\cdot\cdot}$. Now $\{\bar{\varepsilon}_{i\cdot\cdot} : i = 1, \ldots, a\}$ are iid $N(0, \sigma^2/(nb))$ and thus by Fact 3(b) we have

$$\sum \hat{\alpha}_i^2/[\sigma^2/(nb)] \sim \chi^2_{a-1}.$$

Since $SSA = nb\sum \hat{\alpha}_i^2$, we can therefore say that (denoting SSA/σ^2 by R_1^2)

$$R_1^2 = SSA/\sigma^2 = \sum \hat{\alpha}_i^2/[\sigma^2/(nb)] \sim \chi^2_{a-1}.$$

Now

$$F = MSA/MSB = \frac{MSA/\sigma^2}{MSE/\sigma^2} = \frac{R_1^2/(a-1)}{R_0^2/[(n-1)ab]}.$$

Since MSA and MSB are independent, R_0^2 and R_1^2 are independent chi-square random variables. Thus the F-ratio MSA/MSE has an F-distribution with df $(a-1, (n-1)ab)$ when $\alpha_i = 0$ for all i.
(b) and (c). Proofs of these parts are similar to that of (a).