

One/Two Way ANOVA

Due: Monday 11/15 (11.00PM)

Submission: On Gradescope

Part I: Practice Questions

1. A recent study claims that using music in a class enhances the concentration and consequently helps students absorb more information. To test this claim, three different groups of ten randomly selected students (all of the same age) from three different classrooms were selected. Each classroom was provided with a different environment for students to study. *Classroom A* had constant music being played in the background, *Classroom B* had variable music being played and *Classroom C* was a regular class with no music playing. After one month, the students were given a test and their scores were collected (out of 10). The test scores were:

		Test Scores j										Mean
Classroom i	A: constant sound	7	9	5	8	6	8	6	10	7	4	7
	B: variable sound	4	3	6	2	7	5	5	4	1	3	4
	C: no sound	6	1	3	5	3	4	6	5	7	3	4.3

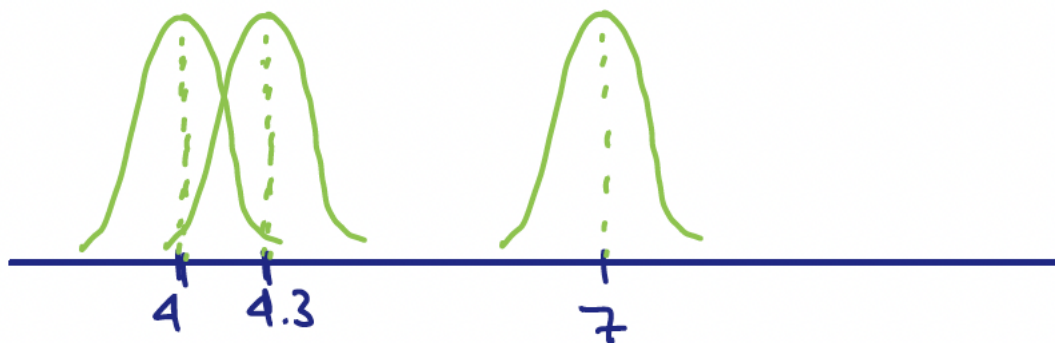
Using the information above, answer the following questions:

- (a) Estimate the mean scores of the students in each type of classroom.

In this question we are looking for the point estimates $\hat{\mu}$. Specifically,

$$\hat{\mu}_A = \bar{Y}_{1.} = 7, \quad \hat{\mu}_B = \bar{Y}_{2.} = 4, \quad \hat{\mu}_C = \bar{Y}_{3.} = 4.3$$

- (b) Draw an illustration of the ANOVA model.



- (c) Write down the ANOVA model for this problem, including the assumptions. Explain your notation.

All the following models are equivalent:

- Cell Means Model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Y_{ij} : test score of student j in classroom i

μ_i : mean test score of students in classroom i

Note that μ_i is a parameter, not a statistic, which means that it is an unknown quantity that we want to estimate.

The error terms satisfy the usual assumption $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

- Factor Effects Model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Y_{ij} : test score of student j in classroom i

μ : mean test score of all students (grand mean)

α_i : effect of classroom i on test score

The error terms satisfy the usual assumption $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

In order for the factor effects model to be equivalent with the Cell Means Model, we need to impose constraints on the model parameters. Different constraints will lead to different parametrizations. All are equivalent and will lead to the same conclusion, however the estimated values and the interpretation will be different.

My favorite constraint is the **sum** constraint, according to which we impose $\sum_i \alpha_i = 0$.

- Regression Model

In order to write the equivalent regression model, we need to define dummy variables. Our factor has 3 levels, so we will introduce 3-1=2 dummy variables. Those are defined as follows:

$$Z_1 = \begin{cases} 1, & \text{if in Classroom A} \\ 0, & \text{otherwise} \end{cases}, Z_2 = \begin{cases} 1, & \text{if in Classroom B} \\ 0, & \text{if otherwise} \end{cases}$$

So, the model is

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon$$

Note that this is the model that you see when you run the "summary" function of the lm fitted model.

Depending on the model (as you can see) the parameters are different. However, when we test for the significance of the corresponding factors, then the results we will obtain will be equivalent.

Below, we choose to work with the Cell Means Model notation:

- (d) Fill in the degrees of freedom, MS and F -value in the ANOVA Table:

<i>Source of Variation</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>p-value</i>
Between Groups	3-1=2	54.6	27.3	8.18	0.0017
Within Groups	29-2=27	90.1	3.34		
Total	30-1=29	144.756			

- (e) Test whether the classroom environment affects students concentration. Use $\alpha = 5\%$. State the null/alternative hypotheses, decision rule and conclusion in the context of the problem.

$$\begin{cases} H_0 : \mu_A = \mu_B = \mu_C \\ H_\alpha : \text{not all } \mu' \text{'s are equal} \end{cases}$$

We look at the p -value that is equal to 0.0017; 0.05, so we reject the null and we conclude that there is at least on mean that is different than the rest.

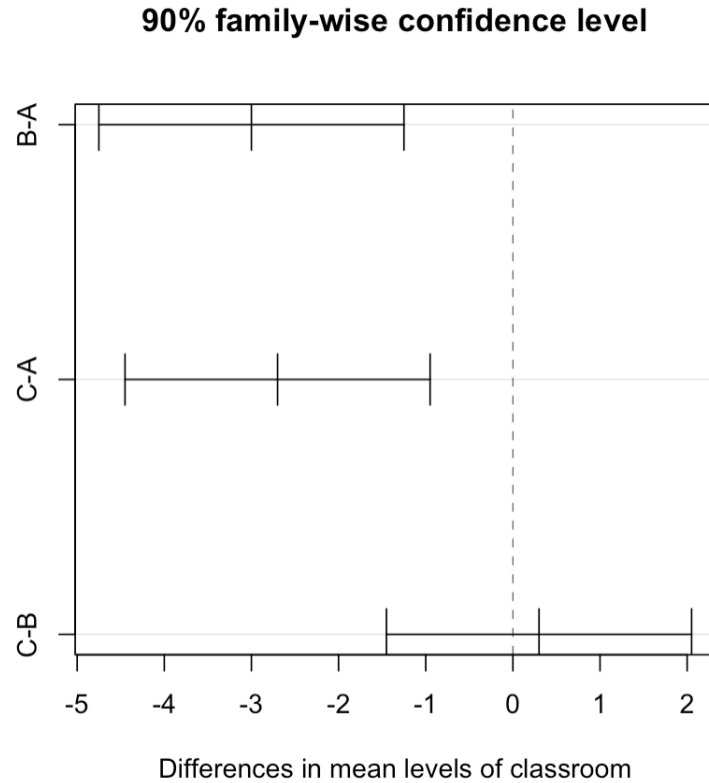
- (f) Construct 90% family confidence intervals for all pairwise comparisons of classroom environments. State your conclusions.

We are going to use R to do this part, and we will focus on interpreting the results:

```
> TukeyHSD(aov(score ~ classroom, data=class.data), "classroom", conf.level=0.9)
Tukey multiple comparisons of means
 90% family-wise confidence level

Fit: aov(formula = score ~ classroom, data = class.data)

$classroom
      diff      lwr      upr    p adj
B-A -3.0 -4.750606 -1.2493942 0.0029108
C-A -2.7 -4.450606 -0.9493942 0.0073313
C-B  0.3 -1.450606  2.0506058 0.9285599
```



Based on the plot and the output, we can conclude that:

- Scores of students in Classrooms B and C do not differ.
- Scores of students in Classrooms C and A differ significantly with the scores of students in classroom A being higher than those in classroom C.
- Scores of students in Classrooms B and A differ significantly with the scores of students in classroom A being higher than those in classroom B.

So, in the context of the problem, we can say that students in a classroom with *constant sound* performed better in the test when compared to students in classrooms B and C, while students in classrooms with variable or no sound performed similarly.

(g) Estimate the following contrast with a 95% confidence interval:

$$L = 2\mu_1 - \mu_2 - \mu_3$$

The ingredients we need to estimate L are computed as follows:

$$\hat{L} = 2\bar{Y}_{1.} - \bar{Y}_{2.} - \bar{Y}_{3.} = 2(7) - 4 - 4.3$$

$$s_{\hat{L}}^2 = \frac{MSE}{n}(2^2 + (-1)^2 + (-1)^2)$$

where $MSE = 3.34$ and $n = 10$.

So, the interval will be

$$L \in \left(\hat{L} \pm t_{27}(0.05/2)s_{\hat{L}} \right)$$

2. Consider the ANOVA model and the difference estimator $D = \mu_i - \mu_{i'}$. Show that $E(\hat{D}) = \mu_i - \mu_{i'}$ and that its estimated variance is

$$s_D^2 = MSE \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)$$

Construct an $(1 - \alpha)100\%$ confidence interval for D .

$$\begin{aligned} E(\hat{D}) &= E(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}) \\ &= E(\bar{Y}_{i\cdot}) - E(\bar{Y}_{i'\cdot}) = \mu_i - \mu_{i'} \end{aligned}$$

$$\begin{aligned} Var(\hat{D}) &= Var(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}) \\ &= Var(\bar{Y}_{i\cdot}) + Var(\bar{Y}_{i'\cdot}) = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_{i'}} \end{aligned}$$

Since σ^2 is unknown, we estimate it with MSE , so we obtain that

$$s_D^2 = MSE \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)$$

An $(1 - \alpha)100\%$ confidence interval for D is given by

$$D \in \left(\hat{D} \pm t_{n_T-r} s_D \right)$$

3. An experiment was conducted to determine the effects of four different pesticides on the yield of fruit from three different varieties of a citrus tree. Eight trees of each variety were randomly selected from an orchard. The four pesticides were randomly assigned to two trees of each variety and applications were made according to recommended levels. Yields of fruit (in bushels) were obtained after the test period.

The data are shown below:

rep	Pest	Var	Yield
1	A	1	49
2	A	1	39
1	B	1	50
2	B	1	55
1	C	1	43
2	C	1	38
1	D	1	53
2	D	1	48
1	A	2	55
2	A	2	41
1	B	2	67
2	B	2	58
1	C	2	53
2	C	2	42
1	D	2	85
2	D	2	73
1	A	3	66
2	A	3	68
1	B	3	85
2	B	3	92
1	C	3	69
2	C	3	62
1	D	3	85
2	D	3	99

Observe that for every factor level combination we have $n = 2$ data points (this is the "rep" column). Factor A (pesticide) has 4 levels (i.e. $a=4$), factor B (variety) has 3 levels (i.e. $b=3$).

- (a) Write down the factor effects model that corresponds to this experiment.

The factor effects model is as follows:

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

Y_{ijk} : fruit yield for pesticide i , variety j , tree j

μ : mean fruit yield of all trees (grand mean)

α_i : effect of Pesticide i on fruit yield

β_j : effect of Variety j on fruit yield

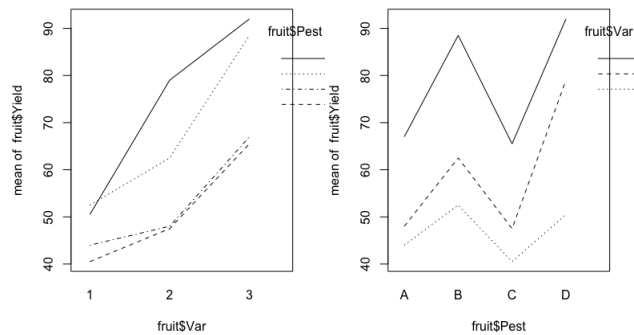
$(\alpha\beta)_{ij}$: interaction term

The error terms satisfy the usual assumption $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

Sum Constraints: $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$.

- (b) Prepare an estimated interaction plot. What is your conclusion about the presence of interactions.

We do this in R and we obtain:



Some of the lines in the plots intersect, so interactions (probably mild) are present.

- (c) Fit an ANOVA model with *yield* as the response, and test whether the interaction term is statistically significant. Use $\alpha = 0.05$. State the alternatives, decision rule and conclusion.

```
> fruit.modell1 <- lm(Yield ~ Var*Pest, data=fruit)
> anova(fruit.modell1)
Analysis of Variance Table

Response: Yield
      Df Sum Sq Mean Sq F value    Pr(>F)    
Var       2 3996.1  1998.04  47.2443 2.048e-06 ***
Pest      3  2227.5   742.49  17.5563 0.0001098 ***
Var:Pest  6   456.9    76.15   1.8007 0.1816844
Residuals 12   507.5    42.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

The hypothesis we test is

$$\begin{cases} H_0 : (\alpha\beta)_{ij} = 0, \forall i, j \\ H_a : \text{not all } (\alpha\beta)_{ij} = 0 \end{cases}$$

Based on the p -value above, we conclude that the interaction term is not statistically significant. So, we can remove it from the model.

Note that we can also perform a partial F test with Full/Reduced model. The result will be the same.

- (d) Test whether the main effects are statistically significant. Use $\alpha = 0.05$. State the alternatives, decision rule and conclusion.

We will formulate this one as a partial F tests

- Factor A

$$\begin{cases} H_0 : Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk} \\ H_\alpha : Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \end{cases}$$

```
> fruit.model2 <- lm(Yield ~ Var, data=fruit)
> fruit.model3 <- lm(Yield ~ Var + Pest, data=fruit)
> anova(fruit.model2, fruit.model3)
Analysis of Variance Table

Model 1: Yield ~ Var
Model 2: Yield ~ Var + Pest
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     21 3191.9
2     18  964.4  3    2227.5 13.858 6.31e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 0.05, and therefore we conclude that the reduced model is not adequate and factor A is statistically significant.

- Factor B

$$\begin{cases} H_0 : Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \\ H_\alpha : Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \end{cases}$$

```
> fruit.model4 <- lm(Yield ~ Pest, data=fruit)
> fruit.model3 <- lm(Yield ~ Var + Pest, data=fruit)
> anova(fruit.model4, fruit.model3)
Analysis of Variance Table

Model 1: Yield ~ Pest
Model 2: Yield ~ Var + Pest
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     20 4960.5
2     18  964.4  2    3996.1 37.292 3.969e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 0.05, and therefore we conclude that the reduced model is not adequate and factor B is statistically significant.

So, we conclude that the additive model is the best one here.

```
> anova(additive.model)
Analysis of Variance Table

Response: Yield
      Df Sum Sq Mean Sq F value    Pr(>F)
Var      2 3996.1  1998.04  37.292 3.969e-07 ***
Pest      3 2227.5   742.49  13.858 6.310e-05 ***
Residuals 18  964.4    53.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- (e) Estimate the mean difference in yields of fruit between Variety 1 and Variety 2 with a 95% confidence interval.

What we are looking for here is a 95% CI for

$$D = \mu_{.1} - \mu_{.2}$$

(the order in the difference does not matter - if we flip the 1,2 the sign is the only thing that will change).

- Estimator $\hat{D} = \bar{Y}_{.1.} - \bar{Y}_{.2.}$.
- Variance: $s_D^2 = \frac{2MSE}{na}$, where $MSE = 53.58$, $n = 2$ and $a = 4$.

So, the interval is

$$D \in (\hat{D} \pm t_{18}(0.05/2)s_{\hat{D}})$$

(f) Estimate the following contrast with a 95% confidence interval

$$L = \frac{\mu_{A.} + \mu_{B.}}{2} - \frac{\mu_{C.} + \mu_{D.}}{2}$$

- Estimator $\hat{L} = \frac{1}{2}\bar{Y}_{1..} + \frac{1}{2}\bar{Y}_{2..} - \frac{1}{2}\bar{Y}_{3..} - \frac{1}{2}\bar{Y}_{4..}$.
- Variance: $s_L^2 = \frac{MSE}{nb}((1/2)^2 + (1/2)^2 + (1/2)^2 + (1/2)^2)$, where $MSE = 53.58$, $n = 2$ and $b = 3$.

So, the interval is

$$L \in (\hat{L} \pm t_{18}(0.05/2)s_{\hat{L}})$$

Part II: Homework Questions – to be submitted

1. A manufacturer of television sets is interested in the effect of tube conductivity of **four** different types of coating color picture tubes. An experiment is conducted and the following conductivity data are obtained:

Coating Type	Conductivity			
1	143	141	150	146
2	152	149	137	143
3	134	136	132	127
4	129	127	132	129

- Is there a difference in conductivity due to coating type? Use $\alpha = 0.05$.
 - Compute a 95% confidence interval for the mean of coating type 4.
 - Compute a 99% confidence interval for the for the mean difference between coating types 1 and 4.
 - Test all pairwise differences in means with a family confidence coefficient 90% (you can choose Bonferroni, Scheffe or Tukey). Based on the results, which coating type produces the highest conductivity?
 - Assuming that coating type 4 is currently in use, what are your recommendations to the manufacturer if they wish to minimize conductivity?
2. Consider the *butterfat* data set in the *Faraway* library. This data set contains information about the percent of butter fat (more is better) in the milk taken from 100 cows. In the study, there are 5 different breeds of cows and 2 different ages. We are interested in assessing if *Age* and *Breed* affect the butterfat content.
- Write down the factor effects model that corresponds to this problem.
 - Prepare an estimated interaction plot. What is your conclusion about the presence of interactions.
 - Fit an ANOVA model with $\log(\text{butterfat})$ are the response, and test whether the interaction term is statistically significant. Use $\alpha = 0.05$. State the alternatives, decision rule and conclusion.
 - Test whether the main effects are statistically significant. Use $\alpha = 0.05$. State the alternatives, decision rule and conclusion.
 - Estimate the mean difference in butterfat content between Mature and 2year cows with a 95% confidence interval.
 - Estimate the following contrast with a 95% confidence interval

$$L = \frac{\mu_{Ayrshire,\cdot} + \mu_{Canadian,\cdot}}{2} - \frac{\mu_{Guernsey,\cdot} + \mu_{Jersey,\cdot}}{2}$$