Solutions

Quiz 7

```
age <- c(288,228,240,384,264,247,253,324,240,288,262,257,260,361,252)
row <- c(rep("1",3), rep("2",3), rep("3",3), rep("4",3), rep("5",3))
as.factor(row)

##  [1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5
## Levels: 1 2 3 4 5

class.data <- rbind(age,row)
lm.fit <- lm(age~row)
```

Above is the age row data we collected in class on Wednesday. Answer the following questions.

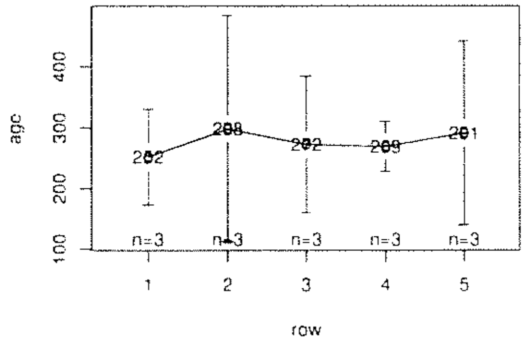
1. Is this a balanced design? Yes – 3 students in each row
2. Refer to the plot below.
    (a) What was the average age of students in row 3? List the average age of students for each row.

    row 1 - 252
    2) 298
    3) 272
    4) 269
    5) 291

    (b) Which row had the largest spread of ages?

    row 2

```
require(gplots)
plotmeans(age~row, mean.labels = T, digits = 2)
```



3. I fit a linear model and below is a summary of the model.

```
summary(lm.fit)

##
## Call:
## lm(formula = age ~ row)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -51.3  -31.7  -12.0   27.5   85.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    252.0       29.0    8.70  5.6e-06 ***
## row2            46.3       41.0    1.13     0.28
## row3            20.3       41.0    0.50     0.63
## row4            17.0       41.0    0.41     0.69
## row5            39.0       41.0    0.95     0.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.2 on 10 degrees of freedom
## Multiple R-squared:  0.139,Adjusted R-squared:  -0.205
## F-statistic: 0.405 on 4 and 10 DF,  p-value: 0.801
```

4. Refer to 2(a) where you listed the average age of students for each row. Can you find these numbers in the output above? Hint: you may have to do some arithmetic. Show, specifically, how to use the output to get to each of the row averages.

row1 = 252

row2 = 252 + 46.3 = 298

row3 = 252 + 20.3 = 272

row4 = 252 + 17 = 269

row5 = 252 + 39 = 291

5. State the hypotheses for conducting a one-way ANOVA for these data.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_A: ~~at least~~ not all u_j are equal

6. The anova is below. What is the F-statistic and the p-value?

F-stat: 0.41 p-value: 0.8

```
anova(lm.fit)

## Analysis of Variance Table
##
## Response: age
##           Df Sum Sq Mean Sq F value Pr(>F)
## row        4   4082    1021    0.41    0.8
## Residuals 10  25193    2519
```

7. What is your decision at a significance level of 0.05?

   Fail to reject H_0

8. What is your conclusion?

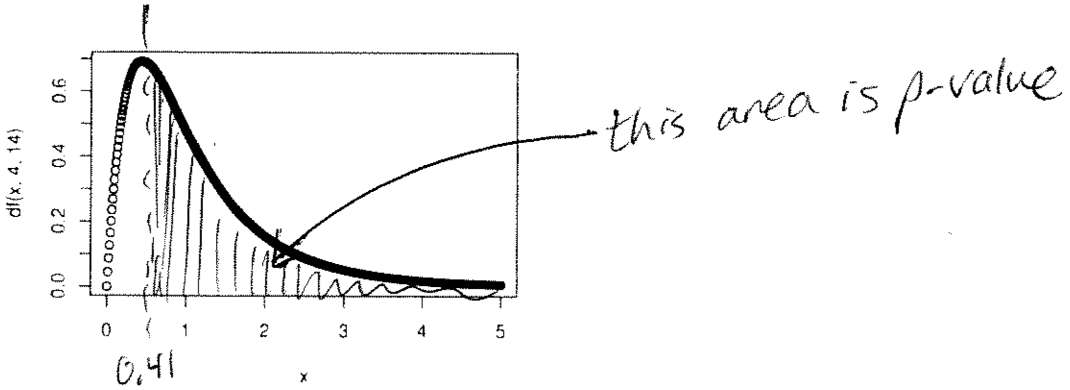
   There is no evidence of at least one difference in the mean age of students across row (p-value = 0.8 from F-stat = 0.40 on 4 and 10 df).

9. It turns out that the F-statistic follows an F-distribution with 4 and 10 degrees of freedom under the null hypotheses. The F-distribution is similar to the t-distribution in that if I ask:

   What is the distribution of the F-statistic under the null hypothesis?

   It is not enough to say "the F-distribution". You should say "the F-distribution with 4 and 10 degrees of freedom". This is because the F-distribution changes when the degrees of freedom changes. Note the degrees of freedom are found in the df column in the ANOVA table above.

   The plot below shows the F distribution with 4 and 10 degrees of freedom. Draw a long vertical line at your F-statistic. Shade the area that is your p-value. Note that F-tests are always upper tailed.



10. Does the shaded area appear to be consistent with the p-value given in the ANOVA?

    Yes the shaded area is about 80% of the plot

Let's look at a new dataset. Cuckoos are known to lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds. These data were originally collected by O. M. Latter in 1902 to see how the size of a cuckoo egg is related to the species of the host bird.

11. What parameterization of the ANOVA model does the following code create?

```
fit.bird <- lm(length~species, data = cuckoo)
```

(a) Cell Means

[✓ (b) Refrence Coded]

12. Use the following output to write the estimated mean egg size for the hedge sparrow and the tree pipet.

```
summary(fit.bird)
```

```
##
## Call:
## lm(formula = length ~ species, data = cuckoo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6489 -0.4489 -0.0489  0.5511  2.1511
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           23.1214     0.2430   95.15  < 2e-16
## speciesmeadow pipet  -0.8225     0.2783   -2.96   0.0038
## speciespied wagtail  -0.2181     0.3379   -0.65   0.5199
## speciesrobin         -0.5464     0.3328   -1.64   0.1033
## speciestree pipet    -0.0314     0.3379   -0.09   0.9261
## specieswren          -1.9914     0.3379   -5.89  3.9e-08
##
## Residual standard error: 0.909 on 114 degrees of freedom
## Multiple R-squared:  0.313,  Adjusted R-squared:  0.283
## F-statistic: 10.4 on 5 and 114 DF,  p-value: 3.15e-08
```

- Hedge Sparrow:

  23.12

- Tree Pipet:

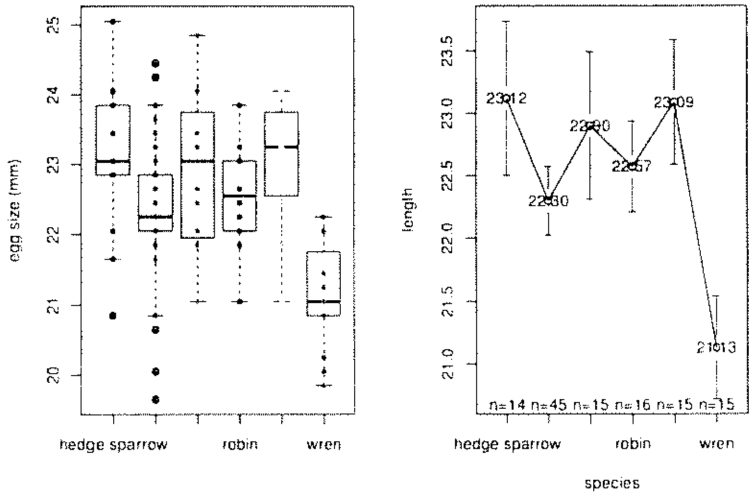
  23.12 -0.03 = 23.09

13. Below are boxplots for the egg size (mm) for all species of host birds, an effects plot and summary statistics for each host species.

```
names(cuckoo)
```

```
## [1] "length"  "species"
```

```
par(mfrow = c(1,2))
boxplot(length~species, data = cuckoo, ylab = "egg size (mm)")
points(length~species, data = cuckoo, col = as.numeric(species), pch = 20)
#install.packages("gplots")
require(gplots)
plotmeans(length ~species, data = cuckoo, mean.labels = T, digits = 2)
```



```
favstats(length~species, data = cuckoo)
```

```
##           .group   min    Q1 median    Q3   max  mean     sd  n missing
## 1 hedge sparrow 20.85 22.90  23.05 23.85 25.05 23.12 1.0687 14       0
## 2  meadow pipet 19.65 22.05  22.25 22.85 24.45 22.30 0.9206 45       0
## 3  pied wagtail 21.05 21.95  23.05 23.75 24.85 22.90 1.0676 15       0
## 4         robin 21.05 22.05  22.55 23.05 23.85 22.57 0.6846 16       0
## 5    tree pipet 21.05 22.55  23.25 23.75 24.05 23.09 0.9014 15       0
## 6          wren 19.85 20.85  21.05 21.75 22.25 21.13 0.7437 15       0
```

14. TRUE or [✓ FALSE] This a balanced design (circle one).

15. State the hypotheses for conducting a one-way ANOVA for these data

H_0 : $\mu_{HS} = \mu_{MP} = \mu_{PW} = \mu_R = \mu_{TP} = \mu_W$

H_a : Not all μ_j are equal

OR

$H_0 : \tau_{HS} = \tau_{MP} = \tau_{PW} = \tau_R = \tau_{TP} = \tau$
$= 0$

H_A : Not all $\tau_j = 0$

16. The ANOVA is below.

```
## Analysis of Variance Table
##
## Response: length
##            Df Sum Sq Mean Sq F value  Pr(>F)
## species     5   42.9    8.59    10.4 3.2e-08 ***
## Residuals 114   94.2    0.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

17. What is the distribution of the F statistic under H_0?

F distribution with 5 and 114 df

18. What is the F statistic and corresponding p-value?

F-stat: 10.4 p-value: 3.2×10^{-8}

19. Which of the following is true about the F-statistic?

A. The F-statistic is small because the variation in average egg size across bird type is large compared to the variation in the egg sizes within a bird type.

[✓ B.] The F-statistic is large because the variation in average egg size across bird types is large compared to the variation in the egg sizes within a bird type.

C. The F-statistic is large because the variation in average egg size across bird types is small compared to the variation in the egg sizes within a bird type.

20. What is your conclusion?

There is strong evidence of at least one difference in mean egg size across host bird type (p-value < 0.0001 from F-stat = 10.4 on 5 and 114 df).