ST 518 Homework 6

Eric Warren

October 28, 2023

Contents

1 Problem 1

A person's blood-clotting ability is typically expressed in terms of a "prothrombin time," which is defined to be the interval between the initiation of the prothrombin-thrombin (two proteins) reaction and the formation of the final clot. Does *aspirin* affect this function? Measurements made before administration of two tablets and three hours after.

First we are going to read in the data to do our future analysis.

```
library(tidyverse)
drug <- read_table("prothrombin.dat")
(
  drug <- drug %>%
    mutate(subject = row_number(),
```

```
           difference = before - after) %>%
    dplyr::select(subject, everything())
)
```

```
## # A tibble: 12 x 4
##    subject before after difference
##      <int>  <dbl> <dbl>      <dbl>
##  1       1   12.3  12         0.300
##  2       2   12    12.3      -0.300
##  3       3   12    12.5      -0.5
##  4       4   13    12         1
##  5       5   13    13         0
##  6       6   12.5  12.5       0
##  7       7   11.3  10.3       1
##  8       8   11.8  11.3       0.5
##  9       9   11.5  11.5       0
## 10      10   11    11.5      -0.5
## 11      11   11    11         0
## 12      12   11.3  11.5      -0.200
```

1.1 Part A

Carry out a paired t-test of the hypothesis that prothrombin time is unaffected by aspirin.

To do this in R, we are going to use the `t.test()` function that will be shown below. We are looking at the difference between the before and after effects of taking aspirin on the prothrombin time. In this case, we can say our null hypothesis is $H_0 : \mu_{before} - \mu_{after} = 0$ with our alternative hypothesis saying that $H_A : \mu_{before} - \mu_{after} \neq 0$. We are going to do this test below.

```
# Transform the data to make it applicable for R t.test
drug_transform <- drug %>%
  pivot_longer(cols = c("before", "after"),
               names_to = "time",
               values_to = "response") %>%
  mutate(subject = as.factor(subject)) %>%
  select(- difference)

# Do t.test
t.test(formula = drug_transform$response ~ drug_transform$time,
       alternative = "two.sided",
       mu = 0,
       paired = TRUE,
       var.equal = TRUE,
       conf.level = 0.95) -> t_test_result

t_test_result
```

```
##
##  Paired t-test
##
## data:  drug_transform$response by drug_transform$time
## t = -0.73998, df = 11, p-value = 0.4748
```

```
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.4305585  0.2138918
## sample estimates:
## mean difference
##      -0.1083333
```

As we can see from our statistical test, we get a test statistic value (t) of -0.7399797, a degrees of freedom of 11, and a resulting p-value is 0.4748097. This p-value is higher than most significance levels we would use for hypothesis testing. Because of this we are failing to reject the null hypothesis that prothrombin time is unaffected by aspirin. As a result we can say that the difference in prothrombin time does not differ significantly after taking aspirin.

1.2 Part B

Carry out an F-test of the same hypothesis treating subjects as blocks in an analysis for a RCBD.

Here we are going to make a model by saying $Y_{ij} = \mu + \alpha_i + B_j + E_{ij}$ where $i = 1, 2$ (represents the before and after taking aspirin) and $j = 1, 2, ..., 12$ (the number of subjects). We are going to fit this model below also using our transformed data.

```
problem1_blocking_model <- lm(response ~ subject + time, drug_transform)
```

Now we are going to do our test which is looking at the difference between the before and after effects of taking aspirin on the prothrombin time. In this case, we can say our null hypothesis is $H_0 : \alpha_1 - \alpha_2 = 0$ with our alternative hypothesis saying that $H_A : \alpha_1 - \alpha_2 \neq 0$. We are going to do an F-test by using the `anova()` function on our fitted model and see what the `time` variable gives us. We are going to do this test below.

```
problem1_anova <- anova(problem1_blocking_model)

problem1_anova
```

```
## Analysis of Variance Table
##
## Response: response
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## subject   11 10.2113 0.92830  7.2186 0.001377 **
## time       1  0.0704 0.07042  0.5476 0.474810
## Residuals 11  1.4146 0.12860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see from our statistical test, we get a test statistic value (F) of 0.54757, a degrees of freedom of numerator 1 and denominator 11, and a resulting p-value is 0.4748097. As this is the same p-value as in **Part A**, this p-value is higher than most significance levels we would use for hypothesis testing. Because of this we are failing to reject the null hypothesis that prothrombin time is unaffected by aspirin. As a result we can say that the difference in prothrombin time does not differ significantly after taking aspirin.

1.3 Part C

Show that, in general, the paired t-test is equivalent to the F-test for the RCBD with block size equal to 2.

We know from our class (I think covered in Module 2 or 3) but also in the practice of statistics that there is a relationship between the F-distribution and t-distribution. It follows that the F-value of something with numerator degrees of freedom 1 and denominator degrees of freedom m is equal to the squared value t-value with m degrees of freedom; moreover, $F(1,m) = [t(m)]^2$. Well whenever we have a size of two for our blocks when doing a F-test, the numerator degrees of freedom will always be $2-1=1$. The denominator degrees of freedom will always be the observation size in the data $2n$, where n is the number of subjects, subtracting the number of degrees of freedom for subject + blocks + 1. Because of 2 blocks, degrees of freedom for blocks is $2-1=1$ and degrees of freedom for subject is $n-1$ where n is the number of subjects. Therefore, the denominator degrees of freedom, $m = 2n-(1+n-1+1) = 2n-(n+1) = n-1$. Now if we can prove that in a paired t-test that the degrees of freedom, $m = n-1$ then we will have shown that when the number of blocks is true, then we have $F(1,m) = [t(m)]^2$. For a two sample t-test we know the degrees of freedom is $m = n-1$. Therefore, the degrees of freedom for a t-test is the same as the denominator degrees of freedom for a F-test when having 2 blocks. Therefore, with 2 blocks $F(1,m) = [t(m)]^2$ holds and why that in general the paired t-test is equivalent to the F-test for the RCBD with block size equal to 2.

1.4 Part D

Consider the linear mixed effects model (or just "mixed model"), $Y_{ij} = \mu + \alpha_i + B_j + E_{ij}$ where $B_j \sim^{iid} N(0,\sigma_B^2), E_{ij} \sim^{iid} N(0,\sigma^2)$ with $B \perp E$ for $i = 1,...,a = 2, j = 1,...,b = 12$.

1.4.1 Part i

Show that $E[MS(block)] = \sigma^2 + a\sigma_B^2$.

We from our lecture that $\hat{\sigma_B^2} = \frac{1}{a}(MS(Block) - MS(E))$. We are going to manipulate this formula to get $MS(Block)$ by itself. So,

$$\hat{\sigma_B^2} = \frac{1}{a}(MS(Block) - MS(E))$$
$$a\hat{\sigma_B^2} = MS(Block) - MS(E)$$
$$a\hat{\sigma_B^2} + MS(E) = MS(Block)$$

Now we want to find $E[MS(Block)]$ so using $MS(Block) = a\hat{\sigma_B^2} + MS(E)$, we get that $E[MS(Block)] = E[a\hat{\sigma_B^2} + MS(E)] = E[a\hat{\sigma_B^2}] + E[MS(E)] = a\sigma_B^2 + \sigma^2$ since we know from past lectures that $E[MS(E)] = \sigma^2$. So $E[MS(Block)] = a\sigma_B^2 + \sigma^2$.

1.4.2 Part ii

Use this result to estimate the variance component for subject effects in a mixed model for the prothrombin data.

We know that $E[MS(Block)] = a\sigma_B^2 + \sigma^2$ which means that $MS(Block) = a\hat{\sigma_B^2} + \hat{\sigma^2}$. We want to find $\hat{\sigma_B^2}$ so we can change our formula $MS(Block) = a\hat{\sigma_B^2} + \hat{\sigma^2}$ to get this by saying $MS(Block) = a\hat{\sigma_B^2} + \hat{\sigma^2}$ <===> $MS(Block) - \hat{\sigma^2} = a\hat{\sigma_B^2}$ <===> $\frac{MS(Block)-\hat{\sigma^2}}{a} = \hat{\sigma_B^2}$. So $\hat{\sigma_B^2} = \frac{MS(Block)-\hat{\sigma^2}}{a}$ and we can use our anova table to help get these values. A reminder of our anova table is below.

```
problem1_anova
```

```
## Analysis of Variance Table
##
## Response: response
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## subject   11 10.2113 0.92830  7.2186 0.001377 **
## time       1  0.0704 0.07042  0.5476 0.474810
## Residuals 11  1.4146 0.12860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see from this table, we can find that $MS(Block) = MS(Subject) = 0.9282955$. We also know that $\hat{\sigma^2} = MS(E) = 0.1285985$. Lastly, a is the number of levels in our other variable (in this case `time`) which is 2. So $a = 2$. Therefore, $\hat{\sigma_B^2} = \frac{MS(Block) - \hat{\sigma^2}}{a} = (1/2)*(0.9282955 - 0.1285985) = 0.3998485$. Therefore, $\hat{\sigma_B^2} = 0.3998485$.

1.4.3 Part iii (Told in Discussion Board we do not have to do)

Report an estimate of the intra-subject correlation. Is the scatterplot above consistent with this estimate?

Based on the discussion board post, we are told to solve this by using the formula $\frac{\sigma_S^2}{\sigma_S^2 + \sigma^2}$. We know that $\sigma_S^2 = \sigma_B^2 = 0.3998485$ and $\sigma^2 = 0.1285985$. Therefore, the correlation is 0.3998485 / (0.3998485 + 0.1285985) = 0.7566483, which is a fairly strong correlation. When comparing to the scatterplot, there seems to be weak correlation on this so the results from the estimate and the scatterplot do not seem to match up.

2 Problem 2

Fuel efficiency of four blends of gasoline is measured in MPG. There is considerable variability due to driver. Another source of variability is model of car. An experiment randomizes four models of car and gasoline blends (A,B,C,D) to drivers according to the design below:

Driver	Model 1	Model 2	Model 3	Model 4
1	15.5(A)	33.8(B)	13.7(C)	29.2(D)
2	16.3(B)	26.4(C)	19.1(D)	22.5(A)
3	10.5(C)	31.5(D)	17.5(A)	30.1(B)
4	14.0(D)	34.5(A)	19.7(B)	21.6(C)

2.1 Part A

Assuming normally distributed data, propose a model in which the effects of model, driver and blend are additive on the mean.

We can make a model that is found from what we have learned for Latin Squares (as this experiment is designed in this way) as $Y_{ij} = \mu + \rho_i + \kappa_j + \tau_k + E_{ij}$ where $i = 1, 2, 3, 4; j = 1, 2, 3, 4; E_{ij} \sim^{iid} N(0, \sigma^2)$ and k is determined by the design.

Now we are going to put this model into R. First we are going to store the data and then we are going to make the appropriate model.

```r
# Make vector of rows
model <- rep(1:4, 4)
driver <- c(rep(1, 4), rep(2, 4), rep(3, 4), rep(4, 4))
gas_blend <- c(LETTERS[1:4], LETTERS[c(2:4, 1)], LETTERS[c(3:4, 1:2)], LETTERS[c(4, 1:3)])
mpg <- c(15.5, 33.8, 13.7, 29.2,
         16.3, 26.4, 19.1, 22.5,
         10.5, 31.5, 17.5, 30.1,
         14.0, 34.5, 19.7, 21.6)

# Make data frame
mpg_df <- data.frame(driver, model, gas_blend, mpg) %>%
  mutate(driver = as.factor(driver),
         model = as.factor(model),
         gas_blend = as.factor(gas_blend))

# Make model
mpg_model <- lm(mpg ~ driver + model + gas_blend, mpg_df)
```

2.2 Part B

Obtain an ANOVA table for this model.

Here we are going to show the ANOVA table below for our model

```r
(problem2_anova <- anova(mpg_model))
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value     Pr(>F)
## driver     3   8.33   2.777  0.6443    0.61428
## model      3 755.37 251.791 58.4116 0.00007838 ***
## gas_blend  3 106.27  35.424  8.2178    0.01515 *
## Residuals  6  25.86   4.311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Something to note is that it seems to be that the driver is said to not be significantly different so it is plausible to think that driver could potentially not make a difference.

2.3 Part C

Report the average fuel efficiency for each blend and the lowest level of significance α at which these averages can be said to differ significantly.

First we are going to find the average fuel efficiency for each blend.

```r
(
  mpg_blend_avgs <- mpg_df %>%
    group_by(gas_blend) %>%
    summarize(avg_mpg = mean(mpg))
)
```

```
## # A tibble: 4 x 2
##   gas_blend avg_mpg
##   <fct>       <dbl>
## 1 A            22.5
## 2 B            25.0
## 3 C            18.0
## 4 D            23.4
```

We can see the average mpg for each gas blend with gas blend **B** having the highest and gas blend **C** having the lowest. From our anova table we can see that the `gas_blend` variable p-value is 0.0151489 which by definition is also the lowest level of significance α at which these averages can be said to differ significantly. Therefore, the lowest level of significance α at which these averages can be said to differ significantly is 0.0151489.

3 Problem 3

For each of several species, an Ecology researcher ran an exposure assay in which groups of $n = 50$ ants are measured for mortality after exposure to one of three different bacteria. For each species, ants are randomized to the three bacteria treatments within each of 15 colonies. That is, a randomized complete block design is used for each species, with colonies serving as complete blocks. Find the results for the DB species on moodle as **"DBdat.mtx"**.

3.1 Part A

Watch the video on the `hiddenf` package. What is the name of the function in R that will create a directory system containing all of the help files that a developer can complete to create documentation for a package? `______skeleton`.