

Homework 4 - Key

Math 530/630

1. One of the goals of the Edinburgh Artery Study was to investigate the risk factors for peripheral arterial disease among persons 55 to 75 years of age. You wish to compare mean LDL cholesterol levels, measured in mmol/liter, among four different populations of subjects: patients with intermittent claudication or interruptions in movement, those with major asymptomatic disease, those with minor asymptomatic disease, and those with no evidence of disease at all. Samples are selected from each population; summary statistics are shown below:
 - a. Use a one-way ANOVA to test the null hypothesis that the mean LDL cholesterol levels are the same for each of the four populations. Calculate this by-hand, and include your ANOVA table.
 - b. If $\alpha = .05$, what do you conclude?
 - c. What assumptions about the data must be true for you to use one-way ANOVA?
 - d. Is it necessary to take any additional steps in this analysis? If so, what?

	n	\bar{x}	s
Intermittent Claudication	73	6.22	1.62
Major Asymptomatic Disease	105	5.81	1.43
Minor Asymptomatic Disease	240	5.77	1.24
No Disease	1080	5.47	1.31

Answers

a.

```
ns = c(73, 105, 240, 1080)
means = c(6.22, 5.81, 5.77, 5.47)
sd = c(1.62, 1.43, 1.24, 1.31)
ttl_n = sum(ns)
df_between_groups = length(ns) - 1
df_within_groups = ttl_n - 1
MSE_W = sum((ns-1)*sd^2)/df_within_groups
grand_mean = sum((means*ns))/ttl_n
MSE_B = sum(ns*(means-grand_mean)^2)/df_between_groups
(F_stat = MSE_B/MSE_W)
```

```
[1] 10.88783
```

```
(F_crit = qf(.95,3,1494))
```

```
[1] 2.610859
```

b.

- $F(3,1494) = 10.87$
- $F_{\text{crit}} = 2.61$
- $F(3,1494) = 10.87 > F_{\text{crit}} = 2.61$
- We can reject the null hypothesis, that LDL Cholesterol is the same for all the four groups.

c. Assumptions for one way ANOVA:

- The k samples are randomly selected from the k populations of interest
- Each of the k populations have a normal distribution

- All k populations have the same variance
- d. The ANOVA F-test answers the question whether there are significant differences in the 4 population means. However, it does not provide us with any information about how they differ. Therefore when we reject H_0 in ANOVA, additional analyses are required to determine what is driving the difference in means.
2. In the early 1900s, Latter (1902) investigated the behavior of female cuckoos, that lay their eggs on the ground and then move them to the nests of other birds. In particular, Latter gathered data on the lengths of the cuckoo eggs found in these foster-nests. Data based on this work is used in (Tippett, 1952) and is located in the file `cuckoos`. The data contains the lengths, in millimeters, of the lengths of cuckoo eggs and the species of the nests where the eggs were placed. Get the data by installing and loading the `resampled` R package, and use the `Cuckoos` dataset.

```
library(resampled)
cuckoos <- Cuckoos
head(cuckoos)
```

	ID	Eggs	Bird
1	1	19.65	MeadowPipit
2	2	20.05	MeadowPipit
3	3	20.65	MeadowPipit
4	4	20.85	MeadowPipit
5	5	21.65	MeadowPipit
6	6	21.65	MeadowPipit

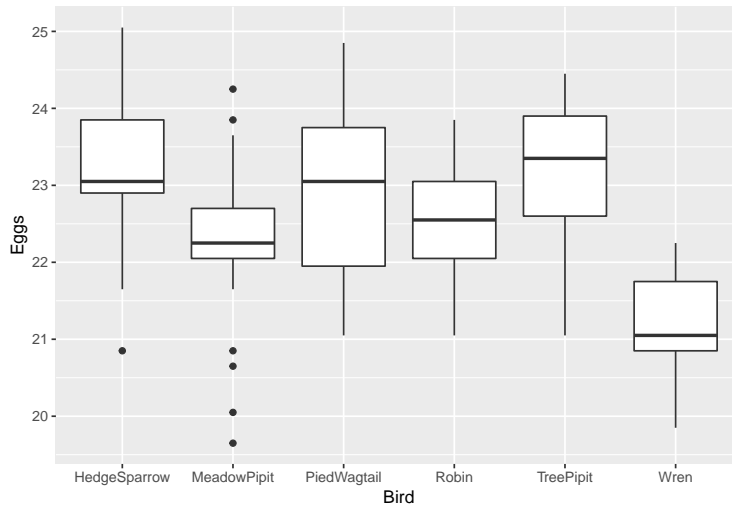
- Create side-by-side boxplots (in R) to compare the distribution of lengths across the different foster nests.
- Conduct an ANOVA test (also in R) to see if the mean lengths of the cuckoo eggs are the same across the different foster nests.
- Perform the Tukey Honestly Significant Difference test (without p-value adjustment) to compare all pairwise means. What can you conclude from this analysis?
- Do the Tukey HSD test using the p-value adjustment method of your choice. Do your conclusions from “2c” change? Given the number of pairwise contrasts, without p-value adjustment, what would be your family-wise error rate if you were to conduct each pairwise contrast at $\alpha = .05$?

Answers

a.

```
library(ggplot2)
library(dplyr)

ggplot(aes(y = Eggs, x = Bird), data = cuckoos) + geom_boxplot()
```



```
group_by(cuckoos, Bird) %>%
  summarise(
    count = n(),
    mean = mean(Eggs, na.rm = TRUE),
    sd = sd(Eggs, na.rm = TRUE)
  )
```

```
# A tibble: 6 x 4
  Bird      count mean  sd
<fct>    <int> <dbl> <dbl>
1 HedgeSparrow    14  23.1 1.07
2 MeadowPipit     44  22.2 0.870
3 PiedWagtail     15  22.9 1.07
4 Robin           16  22.6 0.685
5 TreePipit       16  23.2 0.935
6 Wren            15  21.1 0.744
```

b.

```
aov_model <- aov(Eggs ~ Bird, data = cuckoos)
summary(aov_model)
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)
Bird         5  45.94    9.188    11.48 5.49e-09 ***
Residuals   114  91.25     0.800
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#Let's look at the linear model and confidence intervals for the coefficients.

```
cuckmod <- lm(Eggs ~ Bird, data = cuckoos)
summary(cuckmod)
```

Call:

```
lm(formula = Eggs ~ Bird, data = cuckoos)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-2.6    -0.4     0.0     0.6     2.0
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	23.12143	0.23911	96.697	< 2e-16 ***
BirdMeadowPipit	-0.87143	0.27453	-3.174	0.00193 **
BirdPiedWagtail	-0.21810	0.33247	-0.656	0.51316
BirdRobin	-0.54643	0.32742	-1.669	0.09788 .
BirdTreePipit	0.05357	0.32742	0.164	0.87032
BirdWren	-1.99143	0.33247	-5.990	2.5e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8947 on 114 degrees of freedom

Multiple R-squared: 0.3349, Adjusted R-squared: 0.3057

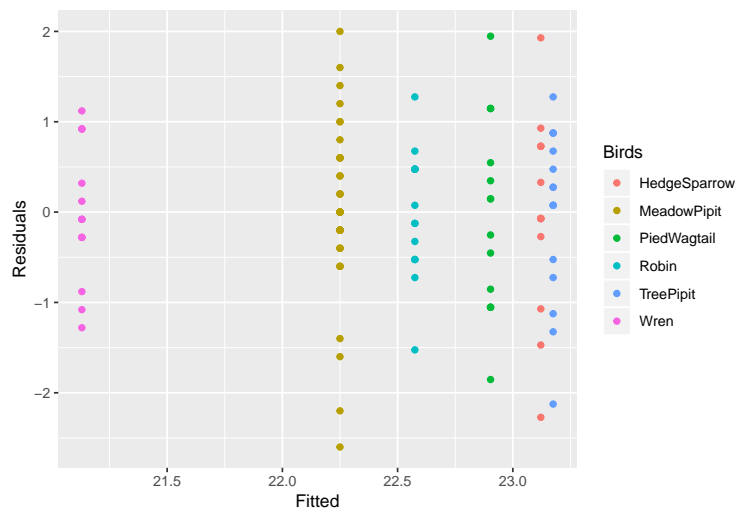
F-statistic: 11.48 on 5 and 114 DF, p-value: 5.494e-09

```
confint(cuckmod)
```

	2.5 %	97.5 %
(Intercept)	22.6477513	23.5951059
BirdMeadowPipit	-1.4152674	-0.3275897
BirdPiedWagtail	-0.8767168	0.4405263
BirdRobin	-1.1950379	0.1021808
BirdTreePipit	-0.5950379	0.7021808
BirdWren	-2.6500501	-1.3328070

#While we're being diligent, let's check the residuals

```
cuckmod1 = data.frame(Fitted = fitted(cuckmod),
  Residuals = resid(cuckmod), Birds = cuckoos$Bird)
ggplot(cuckmod1, aes(Fitted, Residuals, colour = Birds)) + geom_point()
```



Looking at the “Analysis of Variance Table”, we see that the means lengths of cuckoo eggs differ across foster nests.

c.

```
library(multcomp)
mcp <- glht(aov_model, lincfct=mcp(Bird="Tukey"))
summary(mcp, test=univariate()) #the test=univariate() => unadjusted p-value
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

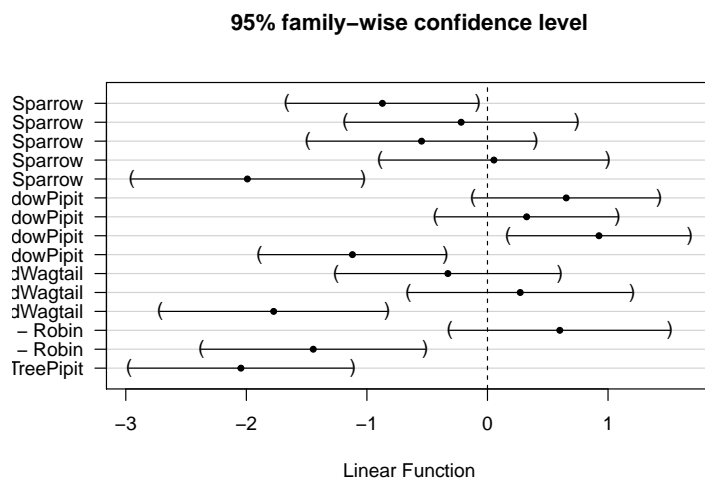
Fit: `aov(formula = Eggs ~ Bird, data = cuckoos)`

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)	
MeadowPipit - HedgeSparrow == 0	-0.87143	0.27453	-3.174	0.001932	**
PiedWagtail - HedgeSparrow == 0	-0.21810	0.33247	-0.656	0.513156	
Robin - HedgeSparrow == 0	-0.54643	0.32742	-1.669	0.097879	.
TreePipit - HedgeSparrow == 0	0.05357	0.32742	0.164	0.870321	
Wren - HedgeSparrow == 0	-1.99143	0.33247	-5.990	2.50e-08	***
PiedWagtail - MeadowPipit == 0	0.65333	0.26750	2.442	0.016125	*
Robin - MeadowPipit == 0	0.32500	0.26119	1.244	0.215936	
TreePipit - MeadowPipit == 0	0.92500	0.26119	3.542	0.000577	***
Wren - MeadowPipit == 0	-1.12000	0.26750	-4.187	5.59e-05	***
Robin - PiedWagtail == 0	-0.32833	0.32154	-1.021	0.309361	
TreePipit - PiedWagtail == 0	0.27167	0.32154	0.845	0.399945	
Wren - PiedWagtail == 0	-1.77333	0.32669	-5.428	3.25e-07	***
TreePipit - Robin == 0	0.60000	0.31631	1.897	0.060379	.
Wren - Robin == 0	-1.44500	0.32154	-4.494	1.69e-05	***
Wren - TreePipit == 0	-2.04500	0.32154	-6.360	4.33e-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Univariate p values reported)

#summary(mcp) #is the same as ...#TukeyHSD(x=aov_model, conf.level=0.95) #and includes adjusted p-value
`plot(mcp)`



Post-hoc analysis provides insight into any differences between the groups that may drive the significant omnibus F-statistic found in b.

Looking at the pairwise differences above, we see that 8 pairs (all of those including Wrens, and all but one including MeadowPipits) are significantly different using an un-adjusted p-value of 0.05.

d.

```
summary(mcp, test=adjusted(type="BH"))
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = Eggs ~ Bird, data = cuckoos)

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)	
MeadowPipit - HedgeSparrow == 0	-0.87143	0.27453	-3.174	0.004139	**
PiedWagtail - HedgeSparrow == 0	-0.21810	0.33247	-0.656	0.549810	
Robin - HedgeSparrow == 0	-0.54643	0.32742	-1.669	0.146819	
TreePipit - HedgeSparrow == 0	0.05357	0.32742	0.164	0.870321	
Wren - HedgeSparrow == 0	-1.99143	0.33247	-5.990	1.88e-07	***
PiedWagtail - MeadowPipit == 0	0.65333	0.26750	2.442	0.030235	*
Robin - MeadowPipit == 0	0.32500	0.26119	1.244	0.294458	
TreePipit - MeadowPipit == 0	0.92500	0.26119	3.542	0.001443	**
Wren - MeadowPipit == 0	-1.12000	0.26750	-4.187	0.000168	***
Robin - PiedWagtail == 0	-0.32833	0.32154	-1.021	0.386701	
TreePipit - PiedWagtail == 0	0.27167	0.32154	0.845	0.461475	
Wren - PiedWagtail == 0	-1.77333	0.32669	-5.428	1.62e-06	***
TreePipit - Robin == 0	0.60000	0.31631	1.897	0.100631	
Wren - Robin == 0	-1.44500	0.32154	-4.494	6.33e-05	***
Wren - TreePipit == 0	-2.04500	0.32154	-6.360	6.49e-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Adjusted p values reported -- BH method)

```
summary(mcp, test=adjusted(type="bonferroni"))
```

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = Eggs ~ Bird, data = cuckoos)

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)	
MeadowPipit - HedgeSparrow == 0	-0.87143	0.27453	-3.174	0.028975	*
PiedWagtail - HedgeSparrow == 0	-0.21810	0.33247	-0.656	1.000000	
Robin - HedgeSparrow == 0	-0.54643	0.32742	-1.669	1.000000	
TreePipit - HedgeSparrow == 0	0.05357	0.32742	0.164	1.000000	
Wren - HedgeSparrow == 0	-1.99143	0.33247	-5.990	3.75e-07	***
PiedWagtail - MeadowPipit == 0	0.65333	0.26750	2.442	0.241881	
Robin - MeadowPipit == 0	0.32500	0.26119	1.244	1.000000	
TreePipit - MeadowPipit == 0	0.92500	0.26119	3.542	0.008661	**
Wren - MeadowPipit == 0	-1.12000	0.26750	-4.187	0.000839	***
Robin - PiedWagtail == 0	-0.32833	0.32154	-1.021	1.000000	
TreePipit - PiedWagtail == 0	0.27167	0.32154	0.845	1.000000	

Wren - PiedWagtail == 0	-1.77333	0.32669	-5.428	4.87e-06	***
TreePipit - Robin == 0	0.60000	0.31631	1.897	0.905678	
Wren - Robin == 0	-1.44500	0.32154	-4.494	0.000253	***
Wren - TreePipit == 0	-2.04500	0.32154	-6.360	6.49e-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Adjusted p values reported -- bonferroni method)

Using a Benjamini-Hochberg p-value adjustment, we replicate the results for the un-adjusted p-values. Using a Bonferroni adjusted p-value, we now find that the difference between lengths for the PiedWagtail-MeadowPipit foster nests is no longer significant.

Calculation of Familywise Error Rate:

- $\alpha = 0.05$
- $c = \text{number of comparisons}$
- $\text{FamilyWiseError} = 1 - (1 - \alpha)^c$
- $\text{FamilyWiseError} = 1 - (1 - 0.05)^{15}$
- $\text{FamilyWiseError} = 0.5367$

Probability of Type I error is over 53% given 15 comparisons.