

# 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)
grand_mean = mean(means)
sigmas = 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
SSW_var = sum((ns-1)*sigmas^2)/df_within_groups
x_bar = sum(ns*means)/ttl_n
SSB_var = sum(ns*(means-x_bar)^2)/df_between_groups
(F_stat = SSB_var/SSW_var)
```

```
## [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. The function `pairwise.t.test` computes the pair-wise comparisons between group means with corrections for multiple testing.

`pairwise.t.test(reponse, factor, p.adjust = method, alternative = c("two.sided", "less", "greater"))` Here response is a vector of observations (the response variable), factor a list of factors and p.adjust is the correction method (e.g., "Bonferroni").

`*pairwise.t.test(cholesterol level, disease, p.adjust = "bonferroni")`

Another multiple comparisons procedure is Tukey's method `TukeyHSD(x, conf.level = 0.95)` where x is a fitted model object.

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
```

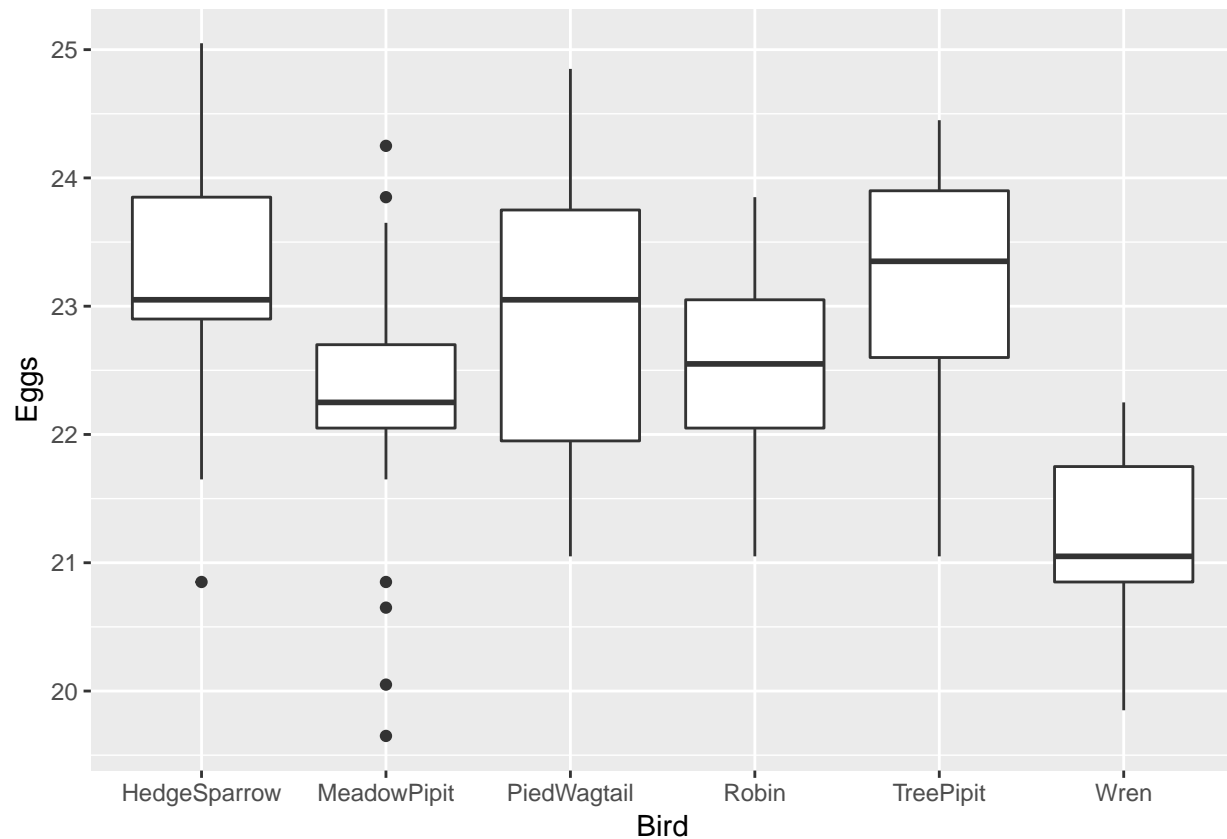
- Create side-by-side boxplots (in R) to compare the distribution of lengths across the different foster nests.
- Conduct an ANOVA test (also in R) to see if the mean lengths of the cuckoo eggs are the same across the different foster nests.
- Perform the Tukey Honestly Significant Difference test (without 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.

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

```
## Analysis of Variance Table
##
## Response: Eggs
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Bird         5  45.938   9.1876  11.478 5.494e-09 ***
## Residuals  114  91.250   0.8004
```

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

```
summary(cuckmod)
```

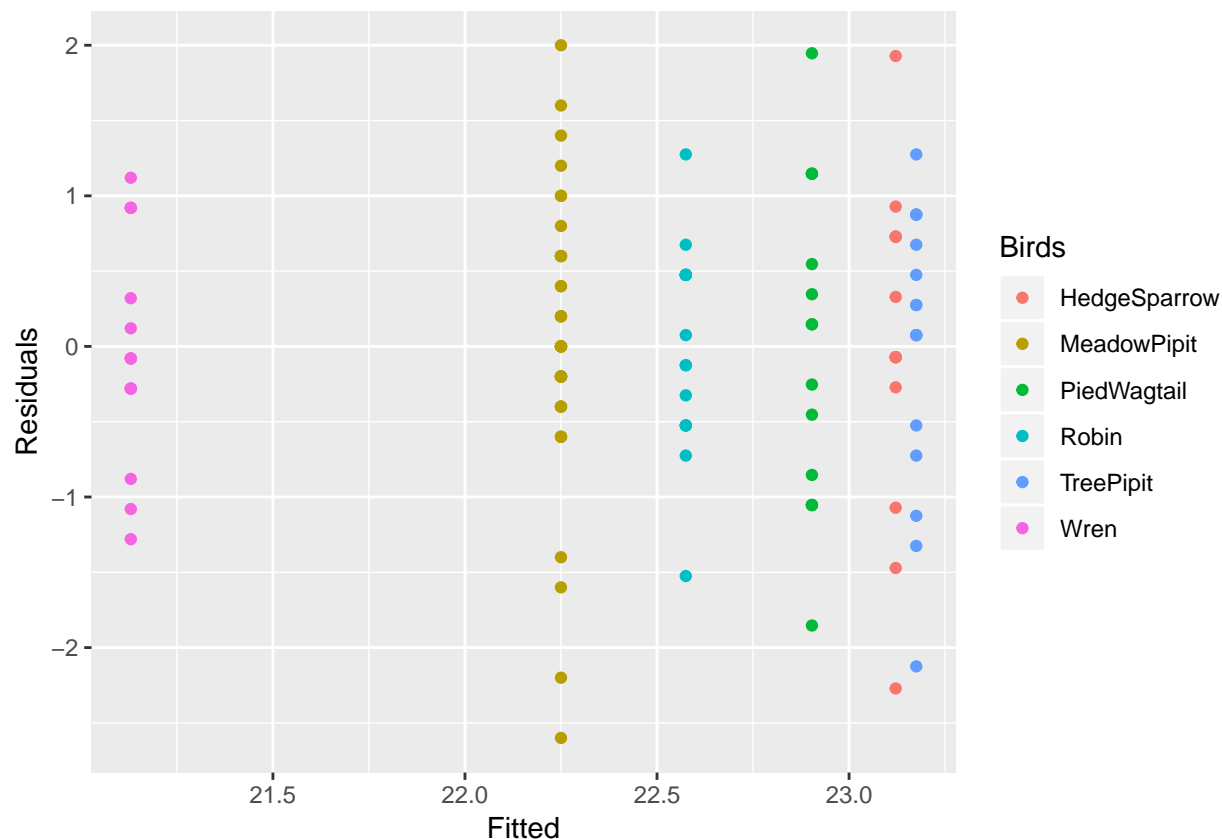
```
##
## Call:
## lm(formula = Eggs ~ Bird, data = cuckoos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##     -2.6     -0.4       0.0       0.6       2.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23.12143     0.23911  96.697 < 2e-16 ***
## 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
```

```
cuckmod1 = data.frame(Fitted = fitted(cuckmod),
  Residuals = resid(cuckmod), Birds = cuckoos$Bird)
```

```
ggplot(cuckmod1, aes(Fitted, Residuals, colour = Birds)) + geom_point()
```



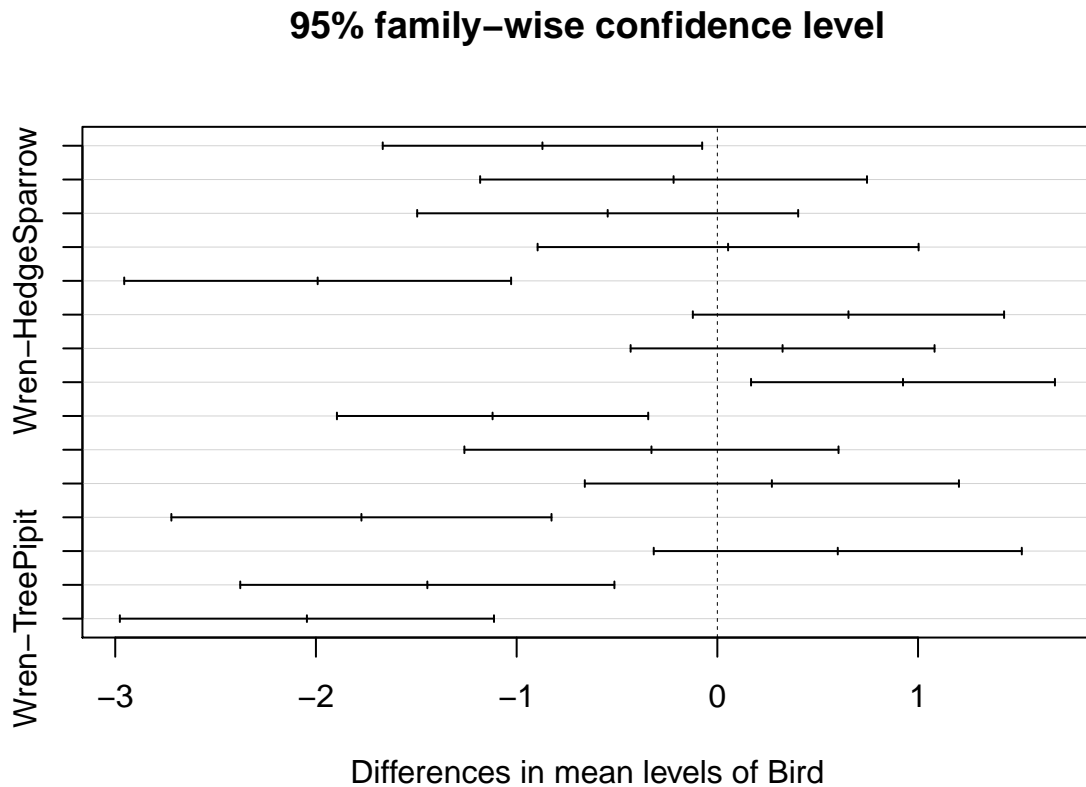
c.

```
aov.fit1 <- aov(Eggs ~ Bird, data=cuckoos)
posthoc <- TukeyHSD(x=aov.fit1, conf.level=0.95)
posthoc
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Eggs ~ Bird, data = cuckoos)
##
## $Bird
##
```

	diff	lwr	upr	p adj
MeadowPipit-HedgeSparrow	-0.87142857	-1.6672254	-0.07563177	0.0231538
PiedWagtail-HedgeSparrow	-0.21809524	-1.1818530	0.74566255	0.9862488
Robin-HedgeSparrow	-0.54642857	-1.4955355	0.40267841	0.5550394
TreePipit-HedgeSparrow	0.05357143	-0.8955355	1.00267841	0.9999833
Wren-HedgeSparrow	-1.99142857	-2.9551864	-1.02767078	0.0000004
PiedWagtail-MeadowPipit	0.65333333	-0.1220788	1.42874546	0.1506847
Robin-MeadowPipit	0.32500000	-0.4321254	1.08212545	0.8140113
TreePipit-MeadowPipit	0.92500000	0.1678746	1.68212545	0.0074440
Wren-MeadowPipit	-1.12000000	-1.8954121	-0.34458788	0.0007787
Robin-PiedWagtail	-0.32833333	-1.2604146	0.60374792	0.9099884
TreePipit-PiedWagtail	0.27166667	-0.6604146	1.20374792	0.9583313
Wren-PiedWagtail	-1.77333333	-2.7203288	-0.82633783	0.0000048
TreePipit-Robin	0.60000000	-0.3169245	1.51692445	0.4094264
Wren-Robin	-1.44500000	-2.3770813	-0.51291874	0.0002402

```
## Wren-TreePipit -2.04500000 -2.9770813 -1.11291874 0.0000001
plot(posthoc)
```



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, we see that the below pairs are significantly different using an un-adjusted p-value of 0.05.

MeadowPipit-HedgeSparrow Wren-HedgeSparrow TreePipit-MeadowPipit Wren-MeadowPipit Wren-PiedWagtail Wren-Robin Wren-TreePipit

d.

```
posthoc <- TukeyHSD(x=aov.fit1, conf.level=0.95)
old <- round(posthoc$Bird[,4],5)
(adjusted <- round(p.adjust(posthoc$Bird[,4]),5))
```

## MeadowPipit-HedgeSparrow	PiedWagtail-HedgeSparrow	Robin-HedgeSparrow
## 0.20838	1.00000	1.00000
## TreePipit-HedgeSparrow	Wren-HedgeSparrow	PiedWagtail-MeadowPipit
## 1.00000	0.00001	1.00000
## Robin-MeadowPipit	TreePipit-MeadowPipit	Wren-MeadowPipit
## 1.00000	0.07444	0.00857
## Robin-PiedWagtail	TreePipit-PiedWagtail	Wren-PiedWagtail
## 1.00000	1.00000	0.00006
## TreePipit-Robin	Wren-Robin	Wren-TreePipit
## 1.00000	0.00288	0.00000

Using a Bonferroni adjusted p-value, we now find only 5 pairs that are significantly different - those involving Wrens.

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.