STAT461 Undergraduate Song Knowledge

Neil Hatfield and Your Name Here

Date

Introduction and Background

Bar or pub trivia are a growing event in American bars/pubs that came from the 1970s British culture. While increasing in popularity, pub trivia is emblematic of a common phenomenon of trivia and game nights. A common round of pub trivia is a music round where short snippets of songs are played over the PA system. Trivia contestants then try to identify the artist and title of each song.

Dr. Hatfield is a big fan of pub trivia and has noticed that the ages of contestants seemed to play a role in how people did at identifying songs. To make exploring his noticing accessible to his students, he designed an sequence of activities to allow STAT 461 students to not only design a study to investigate, but also collect and analyze data.

To this end, we want to explore whether a STAT461 undergraduate student's year in school impacts how well they can identify song title and artist (i.e., their song knowledge)?

Study Design and Methods

In order to investigate our research question, we've designed a quasi-experimental study. Dr. Hatfield constructed a set of 20-second snippets of 10 different songs. Each member of STAT461 present on the quizzing day, took part in the activity. They recorded their year in school (Sophomore, Junior, Senior) and then wrote down the titles and main artist for each of the 10 songs. Baring spelling issues, students received one point for each correct answer.

Dr. Hatfield collected the un-scored answer sheets from all present students and sorted them into three strata based upon year in school. From each stratum, he used R to randomly sample, without replacement, five answer sheets to score and make up our data set. The data is publicly available[1].

Our primary response is a student's level of song knowledge. We've operationalized this through the score based upon their correct identification of song title and primary artist. Our only factor of interest is their categorical year in school. Taken together and with our research question, ANOVA methods appear to be appropriate. Figure 1 shows the Hasse diagram for this study. We can see that an additive model will work and that we have sufficient degrees of freedom to estimate effects and residuals/errors.

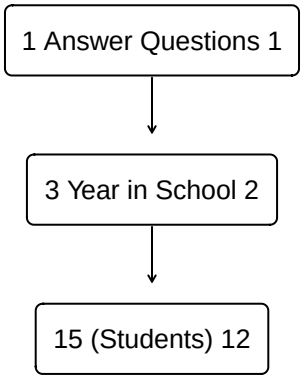


Figure 1: Hasse Diagram for the Song Knowledge Study

[1]Data available at https://raw.github.com/neilhatfield/STAT461/master/dataFiles/songKnowledge_Spring2022.csv

Thus, we will adopt the following null hypothesis: there is no statistically significant impact of year in school on song knowledge score. Our alternative hypothesis is then: there is a statistically significant impact of year in school on song knowledge score. We will control our overall Type I risk at 6%, and we'll use a personal unusualness threshold of 4%[2]. For multiple comparisons, we'll use Tukey's HSD.

Exploration of the Data

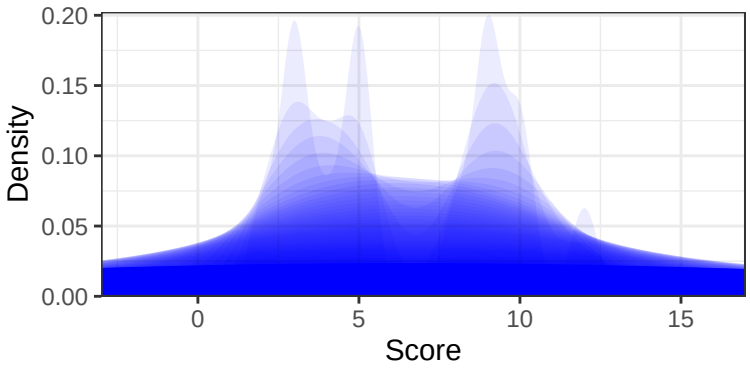


Figure 2: Shadowgram of Song Knowledge Scores

Figure 2 provides the shadowgram for our 15 song knowledge scores. In examining the shadowgram, we can see two modal clumps: the first covering scores of one to six points and the second stretching from eight to eleven points. While we know that we have three groups based upon year in school, Figure 2 suggests that we might only have two.

Table 1: Summary Statistics for Song Knowledge Scores

	n	Min	Q1	Median	Q3	Max	MAD	SAM	SASD	Sample Skew	Sample Ex. Kurtosis
sophomore	5	2	5	8	9	10	2.965	6.8	3.271	-0.407	-1.827
junior	5	5	9	9	10	12	1.483	9.0	2.550	-0.434	-1.400
senior	5	3	3	3	4	5	0.000	3.6	0.894	0.604	-1.670

Table 1 shows the values of various descriptive statistics broken out by year in school. Visually, we can see that the seniors tend to have lower scores than juniors; the max score of the seniors is the same as the minimum score of the juniors. Sophomores cover both. Visually, we can see this in the box plots of Figure 3. There does appear to be differences in the performance of each year when we look at values of the *Sample Arithmetic Mean* (*SAM*). There also appears to be different amounts of variation with each year as evidenced by the values of the *Sample Arithmetic Standard Deviation* (*SASD*).

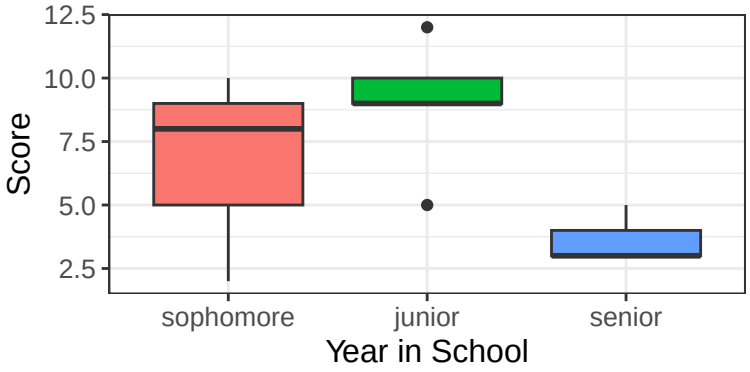


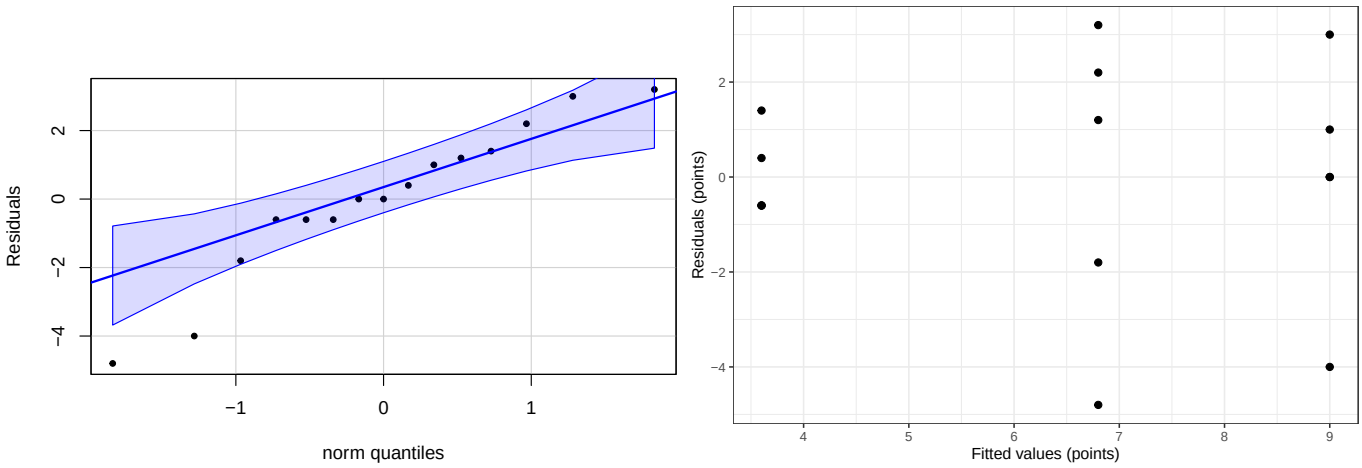
Figure 3: Side-by-side Box Plots of Score by Year

[2]We encourage our readers to select their own unusualness thresholds when examining our results.

Results

To answer our research question, we will seek to use the parametric shortcut known as the ANOVA F test. There are three assumptions that our data must satisfy to use this approach: residuals follow a Gaussian distribution, homoscedasticity, and independence of observations.

Assumptions



(a) QQ Plot (b) Srip Chart

Figure 4: Assessing Assumptions for Song Knowledge Study

Let us first turn towards the Gaussian assumption. Figure 4 shows the QQ plot for our residuals with a 90% confidence envelope. Only two observations (~13%) fall outside of this envelope. Additionally from Table 1, we can see that two of the groups (sophomores and juniors) have negative skewness while the seniors have positive skewness. All groups have negative excess kurtosis. Taken together, we have questions about the whether we've fully satisfied the Gaussian assumption.

Further, in Figure 4 we can also see the strip chart for assessing the homoscedasticity assumption. The groups to the middle and right appear to have more than twice the variation of the leftmost group. This raises a question about whether we satisfy the homoscedasticity assumption.

For the issue of independence of observations, we know that each student completed the quizzes individually. Further, Dr. Hatfield used a computerized system to carry out the selection of individuals from the broader pools. Thus, we can be relatively assured that we have the independence of observations.

Given that we have a balanced design, we will cautiously proceed with the parametric ANOVA F test and our planned post hoc analysis.

Omnibus

Table 2: ANOVA Table for Song Knowledge Study

Source	SS	df	MS	F	p-value	Eta Sq.	Omega Sq.	Epsilon Sq.
year	73.7333	2	36.8667	6.1444	0.0145	0.5059	0.4069	0.4236
Residuals	72.0000	12	6.0000					

As we can see from Table 2, a STAT461 undergradute's year in school accounts for 6.14 times as much variation as the residuals. Since our p-value is less than our unusualness threshold ($0.0145 < 0.04$), we will reject the null hypothesis and decide to act as if a STAT461 undergraduate student's year in college does impact their song knowledge score.

In particular, their year in school accounts for around 41% of the variation in their score ($\omega^2 = 0.41$, $\epsilon^2 = 0.42$, $\eta^2 = 0.51$).

Table 3: Point Estimates from the Song Knowledge Study

	Estimate
Grand Mean	6.47
Sophomores	0.33
Juniors	2.53
Seniors	-2.87

We can also see the factor level (treatment) effects ($\widehat{\alpha_i}$) estimates. For Juniors, they accumulated an additional 2.53 points per student where as the Sophomores only accumulated 0.33 points per student and the Seniors accumulated -2.87 points per student. This suggests that Seniors performing worse that baseline ($GSAM$).

Post Hoc

For our post hoc analyses, we are interested in all of the pairwise comparisons between the three years in school (sophomore, junior, and senior).

Table 4: Post Hoc Tukey HSD Comparisons

	Difference	Lower Bound	Upper Bound	Adj. p-Value
junior-sophomore	2.2	-1.772	6.172	0.362
senior-sophomore	-3.2	-7.172	0.772	0.139
senior-junior	-5.4	-9.372	-1.428	0.012

From Table 4, we can see that the differences in performance between juniors and sophomores and seniors and sophomores are at last as large in magnitude as what we observed 36.2% and 13.9% of time under the null hypothesis. We will take these as usual or typical events when there is no distinction between juniors and sophomores and seniors and sophomores. However, there does appear to a statistical difference between juniors and seniors. We would only anticipate seeing a difference at least as large as 5.4 (in magnitude) about 1% of the time if there was truly no difference between these groups. Approximately 98% of the time we randomly select a junior from STAT461 and a senior from STAT461, the junior will have the higher song knowledge score (see Table 5).

Table 5: Post Hoc Comparison Effect Sizes

Pairwise Comparison	Cohen's d	Hedge's g	Prob. Superiority
sophomore vs. junior	-0.750	-0.678	0.298
sophomore vs. senior	1.334	1.205	0.827
junior vs. senior	2.826	2.553	0.977

Discussion and Limitations

We explored our research question of whether a STAT461 undergraduate student's year in school impacted their song knowledge. From our data, we found that year in school does appear to influence the score they got on the trivia quiz. However, this effect appears to be limited to just juniors and seniors.

Our study has several limitations. First, we are looking at a niche population of just students taking STAT461 in the Spring 2022 semester. In future work, we may want to broaden this to a larger population. Second, our total sample size was only 15 students. We may want to increase this sample size along with the broadening of the population.

Additionally, we may want to incorporate additional attributes that will allow us to more accurately investigate what might be going on. For example, we might collect information on what a student's most commonly listened to genre of music might be.

Author Contributions

The authors of this report would like to acknowledge their individual contributions to the report.

- Dr. Hatfield contributed to the design of the study, collection of data, analysis of data, and writing of the report.
- [Your Name] contributed to the design of the study, participated in the study, analyzed the data, and the writing of the report.

Code Appendix

```
# Setting Document Options ----
knitr::opts_chunk$set(
  echo = FALSE,
  warning = FALSE,
  message = FALSE,
  fig.align = "center",
  dpi = 300
)

# Add additional packages by name to the following list
packages <- c(
  "tidyverse", "knitr", "kableExtra", "hasseDiagram",
  "psych", "car", "parameters"
  )
lapply(X = packages, FUN = library, character.only = TRUE)

# Loading Helper Files and Setting Global Options
options(knitr.kable.NA = "")
options("contrasts" = c("contr.sum", "contr.poly"))
source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/ANOVATools.R")

source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/shadowgram.R")

songData <- read.table(
  file = "https://raw.github.com/neilhatfield/STAT461/master/dataFiles/songKnowledge_Spring2022.csv",
  header = TRUE,
  sep = ","
)

songData$year <- factor(
  x = songData$year,
  levels = c("sophomore", "junior", "senior")
)
# Create a Hasse diagram for the study
# Feel free to use the Hasse diagram wizard app to generate the code for you:
## https://psu-eberly.shinyapps.io/Hasse_Diagrams/

modelLabels <- c("1 Answer Questions 1", "3 Year in School 2", "15 (Students) 12")
modelMatrix <- matrix(
  data = c(FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, TRUE, TRUE, FALSE),
  nrow = 3,
  ncol = 3,
  byrow = FALSE
)
hasseDiagram::hasse(
 data = modelMatrix,
 labels = modelLabels
)

# Note: you do not have to use shadowgrams.
# You can use a histogram or any other kind of data visualization.
shadowgram(
  dataVec = songData$score,
  label = "Score",
  layers = 50,
```

```
  color = "blue",
  aStep = 4
)

scoreStats <- psych::describeBy(
  x = songData$score,
  group = songData$year,
  na.rm = TRUE,
  skew = TRUE,
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = FALSE,
  mat = TRUE,
  digits = 4
)

scoreStats %>%
  tibble::remove_rownames() %>%
  tibble::column_to_rownames(
    var = "group1"
  ) %>%
  dplyr::select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd, skew, kurtosis
  ) %>%
  knitr::kable(
    caption = "Summary Statistics for Song Knowledge Scores",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "SAM", "SASD",
                  "Sample Skew", "Sample Ex. Kurtosis"),
    booktabs = TRUE
  )  %>%
  kableExtra::kable_styling(
    font_size = 12,
    latex_options = c("scale_down", "HOLD_position")
  )

ggplot(
  data = songData,
  mapping = aes(x = year, y = score, fill = year)
) +
  geom_boxplot() +
  theme_bw() +
  xlab("Year in School") +
  ylab("Score") +
  theme(
    legend.position = "none",
    text = element_text(size = 12)
  )

songModel <- aov(
  formula = score ~ year,
  data = songData,
  na.action = "na.omit"
)
```

```r
# Gaussian Residuals Assumption ----
car::qqPlot(
  x = songModel$residuals,
  distribution = "norm",
  envelope = 0.90,
  id = FALSE,
  pch = 20,
  ylab = "Residuals"
)

# Strip Chart for Homoscedasticity ----
ggplot(
  data = data.frame(
    residuals = songModel$residuals,
    fitted = songModel$fitted.values
  ),
  mapping = aes(x = fitted, y = residuals)
) +
  geom_point(size = 2) +
  theme_bw() +
  xlab("Fitted values (points)") +
  ylab("Residuals (points)")

# Modern ANOVA Table ----
parameters::model_parameters(
  model = songModel,
  effectsize_type = c("eta", "omega", "epsilon")
) %>%
  knitr::kable(
  digits = 4,
  col.names = c(
    "Source", "SS", "df", "MS", "F", "p-value",
    "Eta Sq.", "Omega Sq.", "Epsilon Sq."),
  caption = "ANOVA Table for Song Knowledge Study",
  booktabs = TRUE,
  align = c("l", rep("c", 8))
  ) %>%
  kableExtra::kable_styling(
    font_size = 10,
    latex_options = c("HOLD_position")
  )

# Point Estimates for Parametric Shortcut ----
pointEst <- dummy.coef(songModel)
pointEst <- unlist(pointEst)
names(pointEst) <- c("Grand Mean", "Sophomores", "Juniors",
                     "Seniors")

data.frame("Estimate" = pointEst) %>%
  knitr::kable(
  digits = 2,
  caption = "Point Estimates from the Song Knowledge Study",
  format = "latex",
  booktabs = TRUE,
  align = "c"
  ) %>%
```

```
  kableExtra::kable_styling(
    font_size = 12,
    latex_options = c("HOLD_position")
  )

# Nonparametric test ----
songKW <- kruskal.test(
  formula = score ~ year,
  data = songData,
  na.action = "na.omit"
)

songEffectSize <- rcompanion::epsilonSquared(
  x = songData$score,
  g = songData$year,
  digits = 4
)

# Post Hoc via Tukey HSD ----
hsdSong <- TukeyHSD(
  x = songModel, # Your aov/lm object
  conf.level = 0.94 # 1 -- Your overall Type I Error level
)

## Kable Code for Tukey HSD
knitr::kable(
  x = hsdSong$year, # Notice the factor's name
  digits = 3,
  caption = "Post Hoc Tukey HSD Comparisons",
  col.names = c("Difference", "Lower Bound",
                "Upper Bound", "Adj. p-Value"),
  align = 'cccc',
  booktabs = TRUE,
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("condensed", "boardered"),
    font_size = 12,
    latex_options = "HOLD_position"
  )

# Post Hoc Effect Sizes ----
anova.PostHoc(songModel) %>%
  knitr::kable(
    digits = 3,
    caption = "Post Hoc Comparison Effect Sizes",
    col.names = c("Pairwise Comparison","Cohen's d", "Hedge's g",
                  "Prob. Superiority"),
    align = 'lccc',
    booktabs = TRUE
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("condensed", "boardered"),
    font_size = 12,
    latex_options = "HOLD_position"
  )
```

```r
# Nonparametric Post Hoc via DSCF ----
dscfSong <- dscfTest(
  response = songData$score,
  factor = songData$year
)

# Kable Code for DSCF
knitr::kable(
  x = dscfSong,
  digits = 3,
  col.names = c("Comparison", "Observed W", "Adj. p-value"),
  caption = paste("Post Hoc-Dwass-Steel-Critchlow-Fligner Tests"),
  align = 'lcc',
  booktabs = TRUE
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("condensed"),
    font_size = 12,
    latex_options = "HOLD_position"
  )

# Nonparamertic Post Hoc Effect Sizes
kw.PostHoc(
    response = songData$score,
    treatments = songData$year
  ) %>%
knitr::kable(
  digits = 3,
  caption = "Post Hoc Comparison Effect Sizes",
  col.names = c("Pairwise Comparison","Hodges Lehmann Estimate",
                "Prob. Superiority"),
  align = 'lcc',
  booktabs = TRUE
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("condensed", "boardered"),
    font_size = 12,
    latex_options = "HOLD_position"
  )
```