Quarto Set Up, Descriptive Statistics and Hypotheses Testing

Yimeng (Amy) CHENG

Department of Psychology, University of Chicago

Author Note

Yimeng (Amy) CHENG ⓘ https://orcid.org/0009-0007-1238-3749

Correspondence concerning this article should be addressed to Yimeng (Amy) CHENG,

Department of Psychology, University of Chicago, 5801 S Ellis Ave, Chicago, IL 60615, USA,

Email: yimengcheng@uchicago.edu

Abstract

This assessment uses the math anxiety dataset to perform various statistical analyses and hypotheses testing.

Keywords: Math Anxiety, Cognitive Concerns, Perfectionism, Statistical Anxiety

Quarto Set Up, Descriptive Statistics and Hypotheses Testing

Objective-modified

This assessment is based on '04_data-analysis/01_descriptives', but it has been greatly modified to satisfy assessed and unassessed goals in the following areas:

1. Basic R set up.

2. Quarto set up.

3. APA formatting and reference.

4. Creating figures using ggplot2.

5. Perform statistical analyses, including descriptive and hypotheses testing, and present the results in a proper way.

6. Render to PDF.

The dataset being used in the assessment is 'math-anxiety.csv' from Natalie's '04_data-analysis' folder. You may also notice that a lot of instructions are also from the original '04_data-analysis/01_descriptive/descriptive-stats.qmd' file. That is because I was originally doing this task but then thought I could play around with this task and meet extra unique goals. However, the end product did turnout to be quite messy and I apologize for that :(I promise that I will do better in my final research project.

Note: To meet APA formatting requirements, in the **PDF** rendered, **only the proper APA tables and figures will be presented**, and they are at the very end of the document. However, the codes for the walkthrough process and intermediate is figures still presented.

Objective-original

The objective of this assignment is to practice using R to calculate and interpret descriptive statistics. Unlike in previous assignments, you will not use a built-in dataset. Instead, you will use data from a published, open-access dataset capturing relationships between math anxiety and self-perception:

Mackinnon, S. P., McCaughey, N., & Hill, T. G. (2024, July 25). The Association of Self-Efficacy, Anxiety Sensitivity, and Self-Critical Perfectionism with Statistics and Math Anxiety. https://doi.org/10.17605/OSF.IO/NZHQ6

You can access the full dataset and codebook here. The data you will need is included in the .csv file within this project's directory.

Before you begin diving into the data summary, you should take some time to familiarize yourself with the dataset, including the different categories of variables and how they can be interpreted.

Background information

It is widely recognized that anxiety is a common problem among university students(LeViness et al., 2017). However, fewer people know that besides general anxiety, domain-specific anxieties are also prevalent. These includes math anxiety , statistical anxiety and so forth (Onwuegbuzie & Wilson, 2003).

The current project investigates the relationships among domain-specific anxieties and its related constructs, as well as some demographic variables among university students. The dataset is from McCaughey et al. (2022).

Set Up

Read in the data

Run this chunk to read in the dataset and view the first few rows:

```
math.anxiety <- read_csv("math-anxiety.csv")
```

```
# View first 6 rows
head(math.anxiety)
```

Examine the structure of the df. Take note of data types and review factor levels.

```
# Examine the structure of the df

str(math.anxiety)


# Seems like there are only two types of data: numbers and characters.


#change character data to factor so that I can see different levels

math.anxiety <- math.anxiety %>%

  mutate(across(where(is.character), as.factor))


math.anxiety %>%

  select(where(is.factor)) %>%

  map(levels)
```

Data Preparation

Not all variables are currently the data type we'd like them to be. Based on what you see
when you look at the structure of the df and using the data's codebook, which columns are not
currently in the correct data type? What changes do you need to make?

Answer: Program type should not be numeric but should be factor. Id should also not be
numeric but should be character.

In the chunk below, perform simple data cleaning to retype the columns as needed.

```
# Change the data type of columns as needed

math.anxiety <- math.anxiety %>%

   mutate(program.type = factor(program.type, levels = c(0, 1), labels = c("Undergraduat


# Change program.type to factor, with two levels: 1 prepersents yes and 0 represents no.

#Change id to character so that it would not be calculated in summary.
```

In an effort to simplify things at least a little, let's drop the columns that are item

responses, keeping participant info and composite survey scores. Use the codebook to identify
which columns are item responses, then modify the dataframe to drop those columns. We're
assigning this to a new object, `math.anx`, to keep the original data intact.

```
# There are (at least) two ways to approach this, either:
# 1. Notice that all the item responses follow a naming pattern, then use select helpers
# 2. Use the codebook to identify the columns that are *not* item responses and keep the

# Assign your new long df to math.anx
math.anx <- math.anxiety %>%
  select(id, age, gender.category, ethnicity.r, program.type, major, dal.course.code, fa
```

Descriptive Statistics

Center and Spread

We can see basic summary statistics from the wide data (`math.anx`):

```
# Use the appropriate function to see basic summary statistics for the wide data

summary(math.anx)
```

But that's kind of a mess, and it doesn't let us use the summary statistics for any purpose
beyond just looking at them. We can use the long version to calculate summary statistics for each
survey score.

The dataset is not currently in a tidy format, at least not for our purposes. The survey
scores are spread across multiple columns. If we want to use a survey score as a grouping variable
or calculate summary statistics, we need to reshape the data so that each survey score is in a single
column.

Now pivot the data so that the survey scores are in a single column. You'll want to retain
the participant information in the long format as well.

```r
# Create a list of the survey score variables



# Use pivot_longer to reshape the data so that the survey scores are in a single column



math.anx <- math.anx %>%

  pivot_longer(

    cols = c(asi.phys, asi.cog, asi.social, asi, perf.cog, perf.sp,

             frost.ps, frost.com, frost.da, stat.anx.tc, stat.anx.i,

             stat.anx.ah, stat.anx.ws, stat.anx.fst, stat.anx.sc, math.anx),

    names_to = "Survey_Type",   # Create a new column named "survey_type"

    values_to = "Score"         # Put survey values to "score"

  )


head(math.anx)
```

With the data in a tidy format, we can both calculate summary statistics and visualize the data more easily. Let's start with summarizing mean, median, sd, min/max, and range:

```r
# Use the long data to calculate summary statistics for each survey score


summary(math.anx) #this is the most dirct way


#I can also try a more sophisticated way:


# Calculate summary statistics for each survey score

survey_summary <- math.anx %>%
```

```
  group_by(Survey_Type) %>%  # Group by survey type

  summarise(

    Mean = mean(Score, na.rm = TRUE),  # Mean value

    Median = median(Score, na.rm = TRUE),  # Median value

    SD = sd(Score, na.rm = TRUE),  # Standard deviation

    Min = min(Score, na.rm = TRUE),  # Minimum value

    Max = max(Score, na.rm = TRUE),  # Maximum value

    Range = Max - Min  # Calculate range

  )
```

There's a problem with the summary above. Missing values for some scores for some participants mean that the summary calculations are also missing. In some cases that may be what we want, but in this case we have enough data that we can just ignore missing values.

For some reason I just could not load flextable. I tried search on Google and ask AI as well as friends but nothing worked so far. For now, I will use kableExtra instead to create APA formatted table.

```
# Rename survey types for plain English labels

math.anx <- math.anx %>%

  mutate(Survey_Type = recode(Survey_Type,

    "math.anx" = "Math Anxiety",

    "asi" = "Global ASI",

    "asi.phys" = "Physical concerns",

    "asi.cog" = "Cognitive concerns",

    "asi.social" = "Social concerns",

    "perf.cog" = "Perfectionism Cognitions",

    "perf.sp" = "Perfectionism Self-Presentation",

    "frost.ps" = "Perfectionism Personal Standard",
```

```r
    "frost.com" = "Concern Over Mistakes",

    "frost.da" = "Doubt about Actions",

    "stat.anx.tc" = "Test and Class Anxiety",

    "stat.anx.i" = "Interpretation",

    "stat.anx.ah" = "Asking for help",

    "stat.anx.ws" = "Worth of Statistics",

    "stat.anx.fst" = "Fear of Statistics",

    "stat.anx.sc" = "Computational Self-concept"

  ))


# Summarize survey scores


survey_summary <- math.anx %>%

  group_by(Survey_Type) %>%

  summarise(

    Mean = round(mean(Score, na.rm = TRUE), 2),

    Median = round(median(Score, na.rm = TRUE), 2),

    SD = round(sd(Score, na.rm = TRUE), 2),

    Min = round(min(Score, na.rm = TRUE),0),

    Max = round(max(Score, na.rm = TRUE),0),

    Range = Max - Min

  ) %>%

  rename("Survey Type" = Survey_Type

  )
```

Because I will need to create APA table multiple times, I will just define a function to simplify the process. The function includes a few essential lines to produce an APA table, but additional lines are needed to talior the table to specific needs:

```r
kable_apa <- function(summary_data) {

  summary_data %>%

    kable("latex", booktabs = TRUE) %>%

    kable_styling(latex_options = c("striped", "scale_down"), font_size = 10, , position
}
```

```r
# Convert to APA formatted table


kable_apa(survey_summary) %>%

  add_header_above(c(" " = 1, "Descriptive Statistics" = 6)) %>%

  column_spec(1, bold = TRUE)
```

As can be observed from Table 1, for most variables, the median and mean are close to each other, indicating a relatively symmetric distribution.

We can also visualize the data to get a better sense of the distribution of scores. Let's start with a histogram of the math anxiety scores.

```r
math_anxiety_data <- math.anx %>%

  filter(Survey_Type == "Math Anxiety")


ggplot(math_anxiety_data, aes(x = Score)) +

  geom_histogram(binwidth = 0.5, fill = "steelblue", color = "black", size = 1.2, alpha

  labs(

    title = "Distribution of Math Anxiety Scores",

    x = "Math Anxiety Score",

    y = "Frequency"

  ) +

  theme_minimal() +
```

```
  theme(

    plot.title = element_text(hjust = 0.5, face = "bold", size = 14)

  )
```

Extra step: Math Anxiety vs. Statistics Anxiety Scores

```
ggplot(math.anx %>% filter(Survey_Type %in% c("Math Anxiety", "Test and Class Anxiety"))

  geom_histogram(position = "identity", alpha = 0.5, binwidth = 0.5, color = "black") +

  labs(

    title = "Comparison of Math and Statistics Anxiety Scores",

    x = "Score",

    y = "Frequency",

    fill = "Survey Type"

  ) +

  theme_minimal() +

  theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 14))
```

We can observe from Figure 1 and Figure 2 that math anxiety scores are relatively normally distributed with a slight positive skew, and test and class anxiety scores are negatively skewed.

I can also apply facet_wrap() and facet_grid() functions to compare the math anxiety and Test and Class Anxiety scores by program type:

Figure 3 shows that for both graduates and undergraduates, the math anxiety score is slightly positively skewed, where as text and class anxiety is negatively skewed.

Now generate a violin plot of all the survey scores to see differences in distributions across surveys.

Creating publication quality plots using theme and labs layers

```
# Create violin plot

ggplot(math.anx, aes(x = Survey_Type, y = Score, fill = Survey_Type)) +

  geom_violin(trim = FALSE, alpha = 0.6, color = "black", size = 1.2) +  # Static aesthe

  geom_boxplot(width = 0.1, outlier.shape = NA, color = "black", alpha = 0.7) +  # Add b

  labs(

    title = "Distribution of Anxiety Scores Across Surveys",

    subtitle = "Comparing different types of anxiety and performance-related scores",

    x = "Survey Type",

    y = "Score",

    fill = "Survey Type"

  ) +

  theme_classic() +  # Professional minimalist theme

  theme(

    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),  # Centered bold t

    plot.subtitle = element_text(size = 12, hjust = 0.2, color = "gray40"),  # Subtle su

    axis.title.x = element_text(size = 14, face = "bold"),  # X-axis label bold

    axis.title.y = element_text(size = 14, face = "bold"),  # Y-axis label bold

    axis.text.x = element_text(size = 12, angle = 45, hjust = 1),  # Rotate x-axis label

    axis.text.y = element_text(size = 12),  # Readable y-axis labels

    legend.position = "none",  # Remove legend since labels are already clear

     plot.margin = margin(10, 10, 10, 50) #Increase margin so that all labels can be see

  )
```

Figure 4 tells us that Perfectionism Self−Presentation is quite negatively skewed,
potentially suggesting the presence of outliers or extreme values.

Next, I will analyze all survey scores by program type.

```r
summary_stats <- math.anx %>%

  group_by(program.type, Survey_Type) %>%

  summarise(

    Mean = round(mean(Score, na.rm = TRUE), 2),

    Median = round(median(Score, na.rm = TRUE), 2),

    SD = round(sd(Score, na.rm = TRUE), 2),

    Min = round(min(Score, na.rm = TRUE),2),

    Max = round(max(Score, na.rm = TRUE),2),

    Range = Max - Min,

    IQR = round(IQR(Score, na.rm = TRUE),2)

  ) %>%

  ungroup()


# Change the variable names

program_type_summary <- summary_stats %>%

  rename(

    "Program Type" = program.type,

    "Survey Type" = Survey_Type

  )
```

I will use pivot_wider() to make a more accessible table.

```r
summary_stats_wide <- summary_stats %>%

  pivot_wider(names_from = program.type, values_from = c(Mean, Median, SD, Min, Max, Ran

  rename_with(~ gsub("_", " ", .x))
```

From Table 2 we can see that for undergraduates, the mean value for Perfectionism Self-Presentation is 5.28, whereas for graduates it's 4.77.

I want to test whether this difference is significant using one-tailed t-test.

My hypotheses are:

Null Hypothesis (H_0): Undergraduate's Perfectionism Self-Presentation score does not significantly differ from graduates' score.

Alternative Hypothesis (H_A): Undergraduate's Perfectionism Self-Presentation score is significantly higher than graduates' score.

```r
# Extract scores for each group

undergrad_scores <- math.anxiety %>%

  filter(program.type == "Undergraduate") %>%

  pull(perf.sp)


grad_scores <- math.anxiety %>%

  filter(program.type == "Graduate") %>%

  pull(perf.sp)


# Perform One-Tailed Independent t-Test

t_test_result <- t.test(undergrad_scores, grad_scores,

                        alternative = "greater",

                        var.equal = TRUE)  # Assume equal variance
```

As shown in Table 3, the p-value is <0.001, which is significant. Therefore we can reject the null hypothesis and say that undergraduates' Perfectionism Self-Presentation score is significantly higher than graduates' score.

Let's calculate Cohen's d to see the effect size.

```r
# Define a function to compute Cohen's d

calculate_cohens_d <- function(x, y) {

  mean_diff <- mean(x, na.rm = TRUE) - mean(y, na.rm = TRUE)

  pooled_sd <- sqrt((sd(x, na.rm = TRUE)^2 + sd(y, na.rm = TRUE)^2) / 2)
```

```
  cohens_d <- mean_diff / pooled_sd

  return(cohens_d)

}


cohens_d_value <- calculate_cohens_d(undergrad_scores, grad_scores)


print(paste("Cohen's d:", cohens_d_value))
```

```
[1] "Cohen's d: 0.356169195178495"
```

Cohen's d = 0.356 suggests that the difference in means between undergraduates and graduates is small to moderate in standardized terms.

Now I want to further test whether students' Perfectionism Self-Presentation scores differ by ethnicity.

My hypothesis is that there will be no differences: Null Hypothesis (H_0): Students' Perfectionism Self-Presentation scores do not significantly differ by ethnicity.

First, I will set white people as the reference group. Then, I will keep the top three frequent ethinicity and combine the rest as "Other".

```
# First, I will set white people as the reference group
math.anxiety.ethnicity <- math.anxiety %>%
  mutate(ethnicity.r = fct_relevel(ethnicity.r, "White"))


# Keep the top three frequent ethinicity and combine the rest as "Other"

math.anxiety.ethnicity <- math.anxiety.ethnicity %>%
  mutate(ethnicity.r = fct_lump_n(ethnicity.r, n = 3, other_level = "Other"))
```

Next, I will change the numeric scores of Perfectionism Self-Presentation to three levels: low (score < 3), medium $3 \leq$ score < 5) and high ($5 \leq$ score < 7).

```r
math.anxiety.ethnicity <- math.anxiety.ethnicity %>%

  mutate(

    Anxiety_Level = case_when(

      perf.sp < 3 ~ "Low",

      perf.sp >= 3 & perf.sp < 5 ~ "Medium",

      perf.sp >= 5 ~ "High"

    ),

    High_Anxiety = if_else(perf.sp >= 4, "Yes", "No")

  )
```

Note that I have also created two levels, Perfectionism Self-Presentation ≥4 to be high anxiety and <4 to be low anxiety. This is independent from the three "low" "medium" "high" anxiety levels. I have to admit that for this mini-project this is purely for the sake of meeting the object of using both "case_when" and "if_else", but in reality one can also perform Chi-square Test using these two levels.

Now I will perform ANOVA and produce APA table.

```r
anova_ethnicity <- aov(perf.sp ~ ethnicity.r, data = math.anxiety.ethnicity)


anova_table <- broom::tidy(anova_ethnicity) %>%

  rename(

    `Source` = term,

    `Df` = df,

    `Sum Sq` = sumsq,

    `Mean Sq` = meansq,

    `F value` = statistic,
```

```
    `p-value` = p.value

) %>%

mutate(

  `p-value` = ifelse(`p-value` < .001, "< .001", round(`p-value`, 3))

)


kable_apa(anova_table) %>%

column_spec(1, bold = TRUE)
```

As shown in Table 4, there was no significant effect of ethnicity, F(1.10, 3, 449) = 1.10, p = 0.348.

This means the null hypothesis was not rejected and Perfectionism Self-Presentation does not vary across different ethnicity.

References

LeViness, P., Bershad, C., Gorman, K., Braun, L., & Murray, T. (2017). The association for university and college counseling center directors annual survey. *Retrieved Fro m Https://Www. Aucccd. Org/Assets/Documents/Governance/2017% 20aucccd% 20surveypublicapr26. Pdf.*

McCaughey, N. J., Hill, T. G., & Mackinnon, S. P. (2022). The association of self-efficacy, anxiety sensitivity, and perfectionism with statistics and math anxiety. *Personality Science*, *3*(1), e7091.

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments–a comprehensive review of the literature. *Teaching in Higher Education*, *8*(2), 195–209.

Table 1

Summary Statistics for Survey Scores

Survey Type	Descriptive Statistics					
	Mean	Median	SD	Min	Max	Range
Asking for help	3.18	3.00	1.25	1	5	4
Cognitive concerns	2.28	2.17	1.02	1	5	4
Computational Self-concept	2.05	2.00	1.14	1	5	4
Concern Over Mistakes	2.85	2.80	0.98	1	5	4
Doubt about Actions	3.15	3.25	1.01	1	5	4
Fear of Statistics	2.09	2.00	0.89	1	5	4
Global ASI	2.45	2.33	0.79	1	5	4
Interpretation	2.74	2.67	1.00	1	5	4
Math Anxiety	2.61	2.56	0.87	1	5	4
Perfectionism Cognitions	2.83	3.00	0.93	1	5	4
Perfectionism Personal Standard	3.66	3.75	0.82	1	5	4
Perfectionism Self-Presentation	5.12	5.33	1.43	1	7	6
Physical concerns	2.15	1.83	0.95	1	5	4
Social concerns	2.94	2.83	0.92	1	5	4
Test and Class Anxiety	3.67	3.67	0.97	1	5	4
Worth of Statistics	1.79	1.67	0.81	1	5	4

Table 2

Survey Scores by Program Type Wide Table

Survey Type						Descriptive Statistics							
	Mean Undergraduate	Mean Graduate	Median Undergraduate	Median Graduate	SD Undergraduate	SD Graduate	Min Undergraduate	Min Graduate	Max Undergraduate	Max Graduate	Range Undergraduate	Range Graduate	IQR Undergraduate	IQR Graduate
Asking for help	3.24	3.03	3.00	3.00	1.28	1.18	1.00	1.00	5.00	5.00	4.00	4.00	2.50	2.00
Cognitive concerns	2.24	2.37	2.17	2.17	1.02	1.03	1.00	1.00	5.00	4.83	4.00	3.83	1.67	1.50
Computational Self-concept	2.11	1.91	2.00	1.50	1.17	1.07	1.00	1.00	5.00	5.00	4.00	4.00	2.00	1.50
Concern Over Mistakes	2.93	2.67	3.00	2.60	0.99	0.95	1.00	1.00	5.00	4.80	4.00	3.80	1.40	1.40
Doubt about Actions	3.13	3.18	3.25	3.25	1.04	0.93	1.00	1.25	5.00	5.00	4.00	3.75	1.50	1.38
Fear of Statistics	2.10	2.08	2.00	2.00	0.88	0.93	1.00	1.00	5.00	4.33	4.00	3.33	1.33	1.67
Global ASI	2.42	2.53	2.33	2.33	0.79	0.80	1.06	1.28	4.56	4.88	3.50	3.60	1.17	1.14
Interpretation	2.88	2.41	3.00	2.33	1.01	0.90	1.00	1.00	5.00	4.33	4.00	3.33	1.67	1.33
Math Anxiety	2.74	2.33	2.67	2.33	0.89	0.77	1.00	1.00	5.00	4.44	4.00	3.44	1.33	1.11
Perfectionism Cognitions	2.67	3.20	2.67	3.33	0.84	1.00	1.00	1.00	4.00	5.00	3.00	4.00	1.33	1.50
Perfectionism Personal Standard	3.71	3.54	3.75	3.50	0.81	0.81	1.00	1.00	5.00	5.00	4.00	4.00	1.00	1.00
Perfectionism Self-Presentation	5.28	4.77	5.50	4.67	1.40	1.42	1.00	1.00	7.00	7.00	6.00	6.00	1.67	1.83
Physical concerns	2.13	2.17	1.83	2.00	0.96	0.94	1.00	1.00	5.00	5.00	4.00	4.00	1.50	1.17
Social concerns	2.89	3.03	2.83	3.00	0.92	0.93	1.17	1.17	5.00	5.00	3.83	3.83	1.33	1.48
Test and Class Anxiety	3.80	3.37	4.00	3.67	0.93	0.98	1.00	1.00	5.00	5.00	4.00	4.00	1.33	1.33
Worth of Statistics	1.81	1.75	1.67	1.67	0.80	0.82	1.00	1.00	4.67	5.00	3.67	4.00	1.00	1.00

Table 3

One-Tailed t-Test Results for Perfectionism Self-Presentation

Statistic	Value
t-value	3.504
df	451
p-value	<0.001
Mean (Undergraduates)	5.276
Mean (Graduates)	4.773
95% CI Lower	0.266
95% CI Upper	Inf

Table 4

Source	Df	Sum Sq	Mean Sq	F value	p-value
ethnicity.r	3	6.704285	2.234762	1.100839	0.348
Residuals	449	911.494145	2.030054	NA	NA

Figure 1

Distribution of Math Anxiety Scores
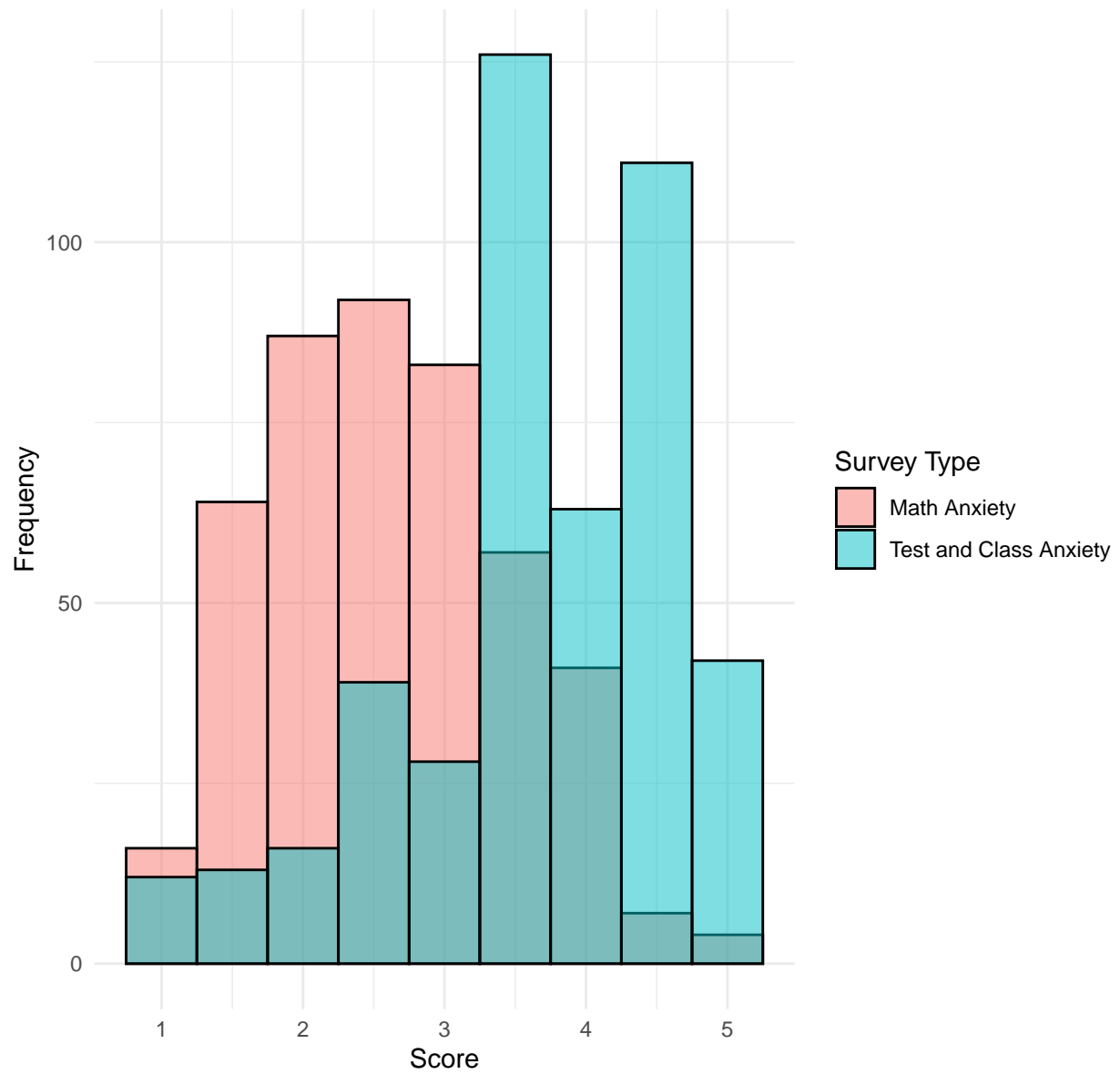

Distribution of Math Anxiety Scores
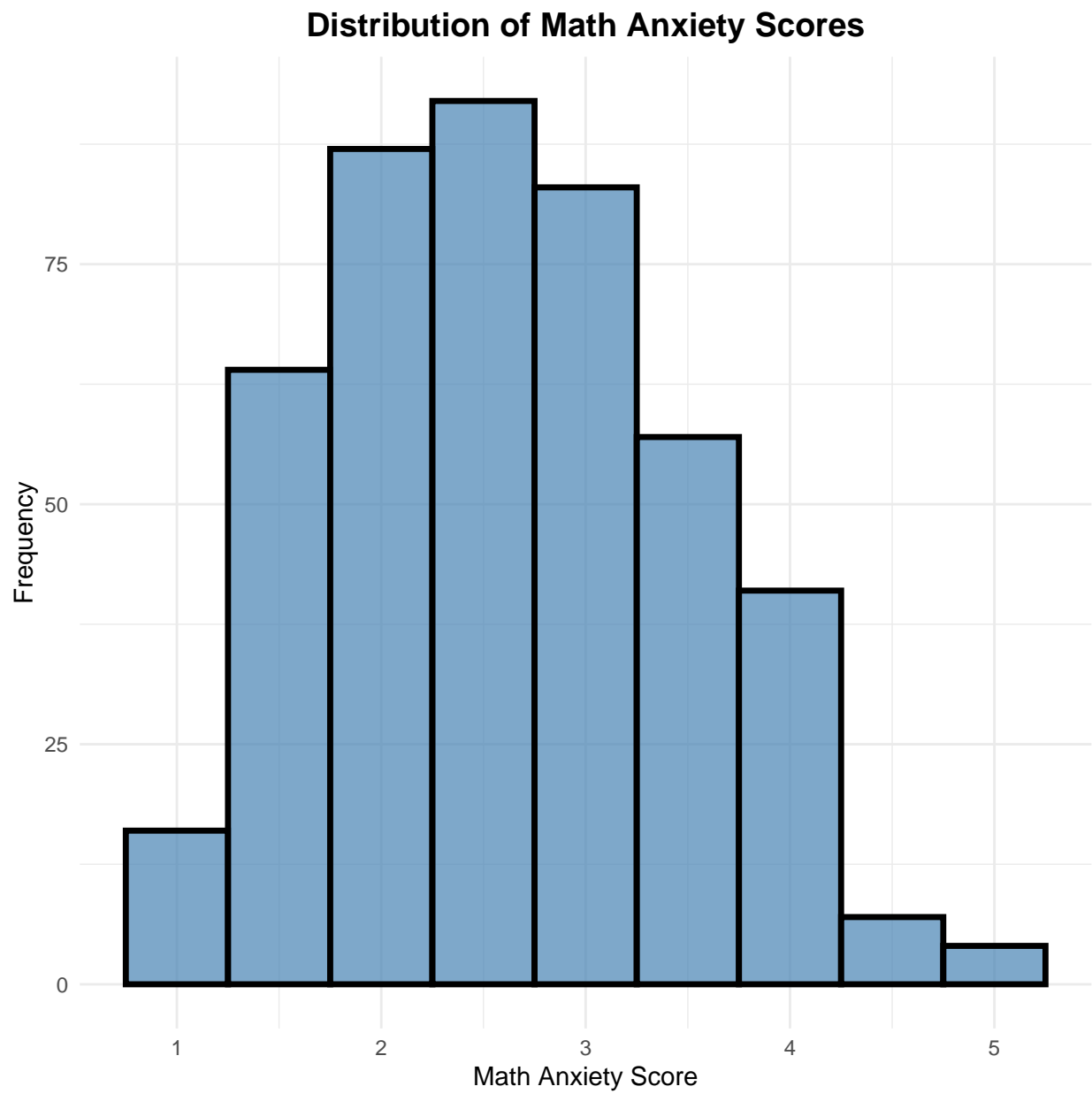
Figure 2

Comparison of Math and Statistics Anxiety Scores

Figure 3

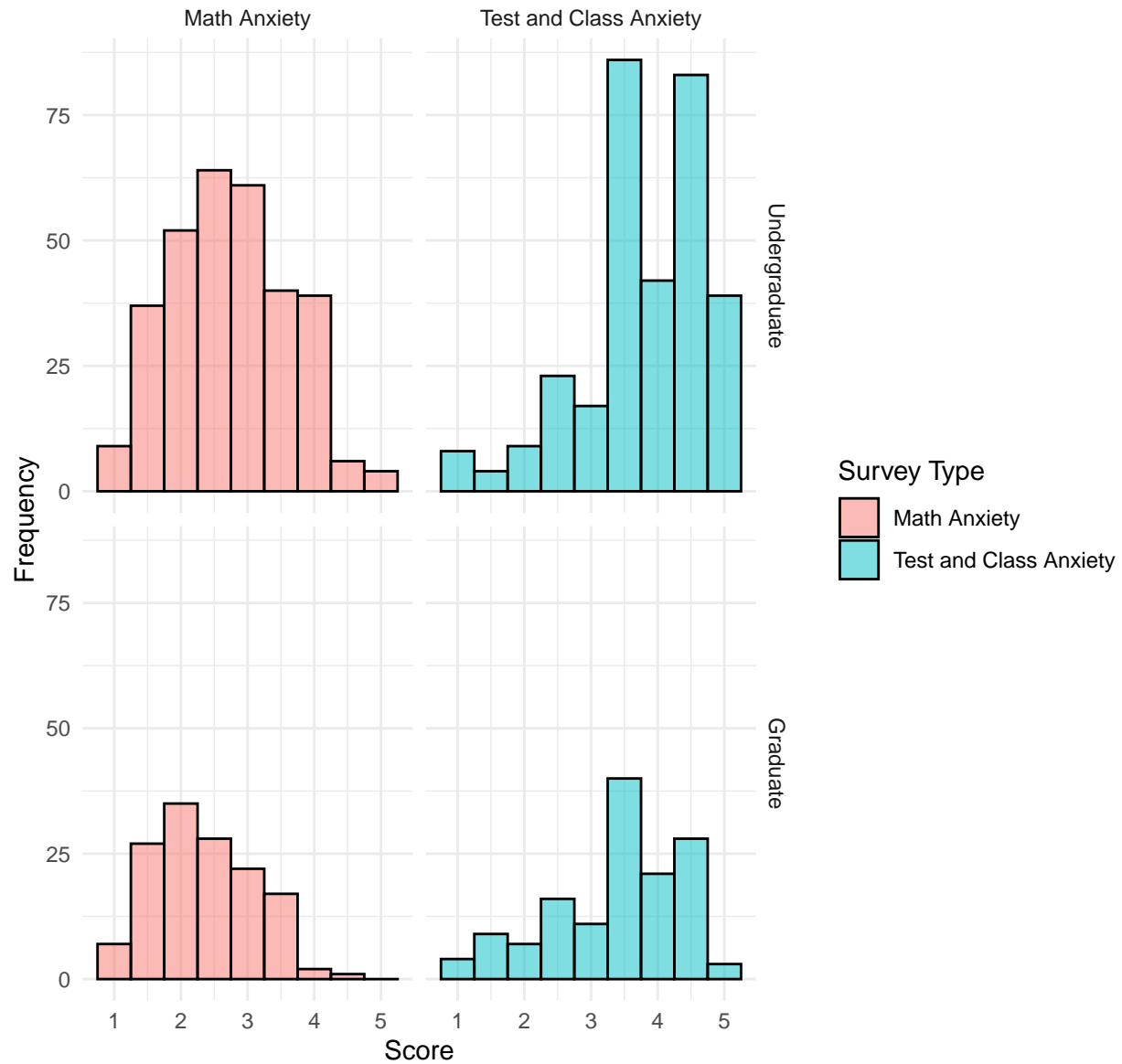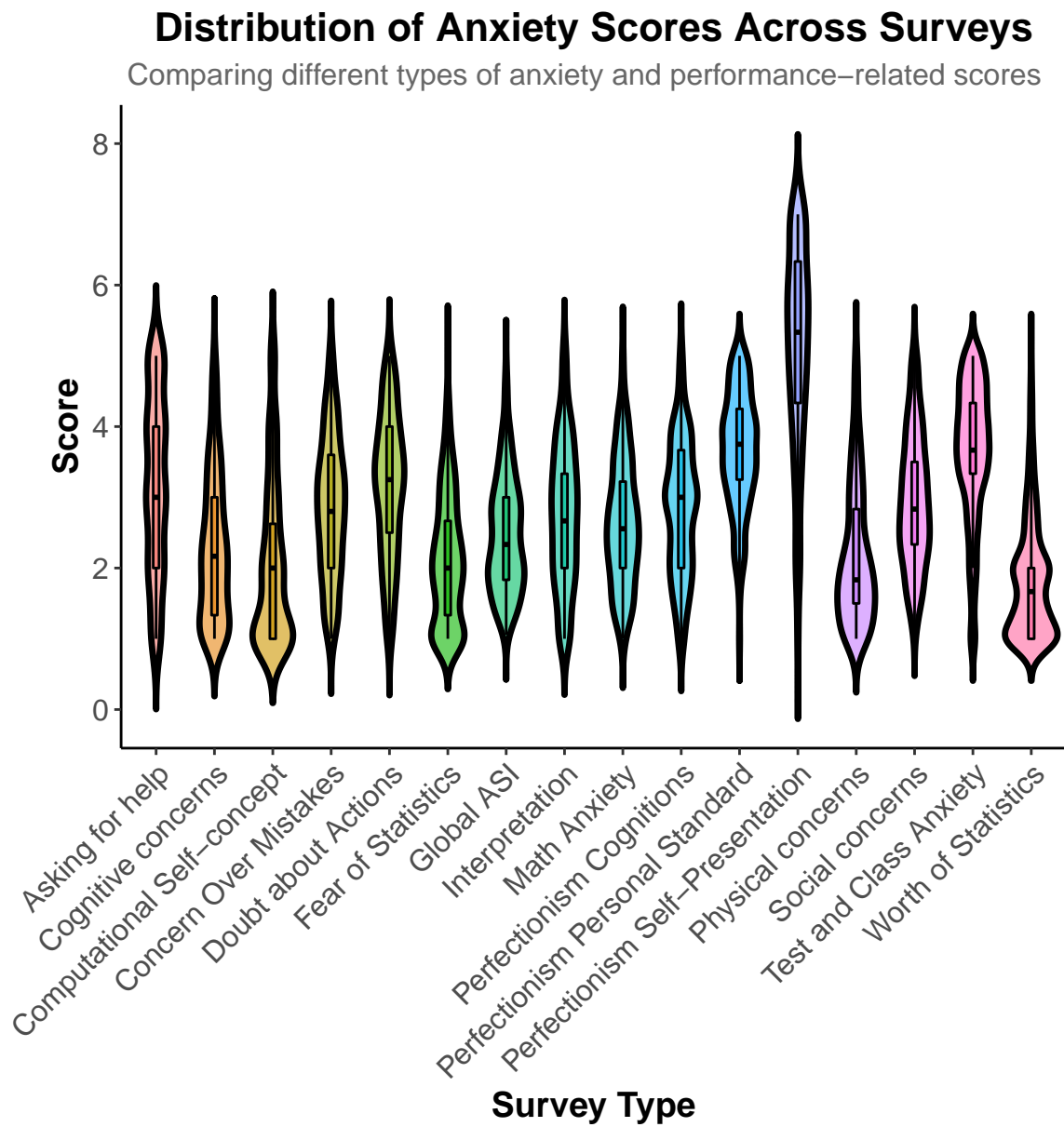**ison of Math and Statistics Anxiety Scores by Program Type**

Figure 4

Distributions of Survey Scores (Publication Quality)

Appendix

Title for Appendix