Exam Template: Statistical Inference

25036686

22/07/2025

# Instructions to students

Save this template as your studentID.Rmd; you will upload this file as part of your submission. Change the author information on line 3 of this file to your **student ID**. Do not change the authorship to your name.

Your should knit this file to a document **Word** format. The Word document is what will be marked!

Any changes that you make to the data (e.g. variable name changes) should be made entirely within R.

The subsections labelled **Answer:** indicate where you should put in your written Answers. The template also provides blank code chunks for you to complete your Answers; you may choose to add additional chunks if required.

This is an individual assessment: do not work with any other person during this exam. Text-matching software will be used on all submissions.

# Instructions for submission

You must submit your assignment before the stated deadline by electronic submission through Blackboard.

* It is a good idea to save your work early and frequently to ensure you have no issues with the submission portal. Multiple submissions can be made to the portal, only the final one will be marked.
* It is your responsibility to submit the exam in a format stipulated above. Your marks may be affected if your tutor cannot open or properly view your submission.
* Do not leave submission to the very last minute. Always allow time in case of technical issues.
* The date and time of your submission is taken from the Blackboard server and is recorded when your submission is complete, not when you click Submit.
* It is essential that you check that you have submitted the correct file(s), and that each complete file was received. Submission receipts are accessed from the Coursework tab.

There is no late submission permitted on this timed assessment. Ensure that you submit your submission in good time. Neither the module leader nor module team can accept late assessments, do not ask them to do so.

# Background to the research

The client is a researcher team at UWE Bristol who wish to disseminate their research results.

The information relates to BRP = Boosting reading @ primary (a reading intervention programme).

This is an intervention in schools within the Bristol area conducted by UWE Bristol, and usually lasts 2 months for each pupil.

The aim is to improve pupils reading ability by more than would usually be expected in 2 months.

# Data instructrions

Your individual data set is accessed via Blackboard >>> Assessments >>> Dewis Data For Exam.

You must only analyse the specified data. No other data is to be used for this assessment.

All data manipulation and analyses must be done within R.

# Data structure

The data represents a sample of 50 pupils that took part in the BRP. For each pupil, the following are recorded:

ID = unique identifier (anonymous)

school = code for the school in the Bristol area

gender: 1=female, 2=male

agestart = age in months, before the intervention

ageend = age in months, after the intervention

BASstart = British Ability Score (a standardised word reading test, taken before the intervention)

BASend = British Ability Score (a standardised word reading test, taken at the end of the 2 month intervention)

RRstart = Reading level (reading level at start of intervention as judged by teacher. Note that some schools use an alternative system of colours, you can ignore these)

RRend = Reading level (reading level at end of intervention as judged by teacher. Note that some schools use an alternative system of colours, you can ignore these)

EAL: English as an additional language 1=YES, 2=NO

EMB: ethnic minority background 1=YES, 2=NO

PP: Pupil Premium funded (additional funding provided if they received free school meals in the previous 5 years) 1=YES 2=NO

# QUESTIONS START HERE

# Question 1: Data Preparation

Ensure you have prepared your submission as per Instructions to Students

You should explain any decisions with respect to:

-any manipulation of the data structure

-handling of groups with low frequency counts

-missing values.

Marks will be awarded here for logical actions performed, and conforming to principles of reproducible research in the data preparation and throughout the submission.

**(14 marks)**

### Answer:

# load the dataset here  
  
pupil\_data <- read\_csv("Exam Data.csv")

Warning: One or more parsing issues, call `problems()` on your data frame for details,  
e.g.:  
 dat <- vroom(...)  
 problems(dat)

Rows: 49 Columns: 12  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (4): school, agestart, RRstart, RRend  
dbl (8): ID, gender, ageend, BASstart, BASend, EAL, EMB, PP  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#this csv file is been downloaded in my project folder - can then be loaded by calling the name of the file

# further data preparation here  
  
head(pupil\_data)

# A tibble: 6 × 12  
 ID school gender agestart ageend BASstart BASend EAL EMB PP RRstart  
 <dbl> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>   
1 169 K3 1 89 90 37 43 2 2 1 20   
2 164 K3 2 73 74 31 38 1 2 2 13   
3 216 M 1 72 73 46 45 2 2 2 19   
4 176 M2 2 97 99 46 51 2 2 2 purple   
5 59 M 1 91 92 50 60 2 2 2 Gold   
6 54 M 1 91 92 32 38 2 2 2 Green   
# ℹ 1 more variable: RRend <chr>

str(pupil\_data)

spc\_tbl\_ [49 × 12] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
 $ ID : num [1:49] 169 164 216 176 59 54 122 61 190 256 ...  
 $ school : chr [1:49] "K3" "K3" "M" "M2" ...  
 $ gender : num [1:49] 1 2 1 2 1 1 NA 2 2 1 ...  
 $ agestart: chr [1:49] "89" "73" "72" "97" ...  
 $ ageend : num [1:49] 90 74 73 99 92 92 80 92 94 86 ...  
 $ BASstart: num [1:49] 37 31 46 46 50 32 49 37 37 43 ...  
 $ BASend : num [1:49] 43 38 45 51 60 38 59 42 45 47 ...  
 $ EAL : num [1:49] 2 1 2 2 2 2 2 2 2 1 ...  
 $ EMB : num [1:49] 2 2 2 2 2 2 2 2 2 2 ...  
 $ PP : num [1:49] 1 2 2 2 2 2 1 2 2 2 ...  
 $ RRstart : chr [1:49] "20" "13" "19" "purple" ...  
 $ RRend : chr [1:49] "21" "14" "20" "purple" ...  
 - attr(\*, "spec")=  
 .. cols(  
 .. ID = col\_double(),  
 .. school = col\_character(),  
 .. gender = col\_double(),  
 .. agestart = col\_character(),  
 .. ageend = col\_double(),  
 .. BASstart = col\_double(),  
 .. BASend = col\_double(),  
 .. EAL = col\_double(),  
 .. EMB = col\_double(),  
 .. PP = col\_double(),  
 .. RRstart = col\_character(),  
 .. RRend = col\_character()  
 .. )  
 - attr(\*, "problems")=<externalptr>

sum(is.na(pupil\_data))

[1] 32

colSums(is.na(pupil\_data))

ID school gender agestart ageend BASstart BASend EAL   
 0 0 2 0 0 0 0 4   
 EMB PP RRstart RRend   
 4 5 7 10

# with the above code I am previewing the data checking the structure and counting the missing values  
  
# the first variable I will handle the manipulation is 'gender' by assigns and unknown value to NA  
  
pupil\_data$gender[is.na(pupil\_data$gender)] <- "unknown"  
  
# the second variable to handle is 'agestart' as it is a character variable, I will convert it to numeric  
  
unique(pupil\_data$agestart) # with this function I can see the unique values of agestart identifying an issue with the charterer 'm' in some of the values

[1] "89" "73" "72" "97" "91" "79" "92" "85" "96" "106" "81" "87"   
[13] "76" "99" "88" "101" "109" "86" "83" "98" "110" "82m" "100" "70"   
[25] "78" "74" "86m"

pupil\_data$agestart <- as.numeric(gsub("m", "", pupil\_data$agestart)) # removing m from the numerical values and converting the entire column in numeric  
  
print(pupil\_data$agestart) # checking output

[1] 89 73 72 97 91 91 79 91 92 85 96 106 81 87 76 87 81 99 88  
[20] 101 97 87 87 81 106 109 86 79 83 99 87 81 98 87 110 87 82 88  
[39] 97 89 100 70 78 74 86 87 86 86 83

# handling EAL (English as addition language) and EMB (pupils with ethnic minority) by setting the NA values to 'unknown'  
  
pupil\_data$EAL[is.na(pupil\_data$EAL)] <- "unknown"  
pupil\_data$EMB[is.na(pupil\_data$EMB)] <- "unknown"  
  
table(pupil\_data$EAL, useNA = "always")

1 2 unknown <NA>   
 9 36 4 0

table(pupil\_data$EMB, useNA = "always")

1 2 unknown <NA>   
 11 34 4 0

# Drop unnecessary column for this analysis (PP)  
  
pupil\_data <- pupil\_data |> select(-PP)  
  
# checking the final structure and checking missing values  
  
str(pupil\_data)

tibble [49 × 11] (S3: tbl\_df/tbl/data.frame)  
 $ ID : num [1:49] 169 164 216 176 59 54 122 61 190 256 ...  
 $ school : chr [1:49] "K3" "K3" "M" "M2" ...  
 $ gender : chr [1:49] "1" "2" "1" "2" ...  
 $ agestart: num [1:49] 89 73 72 97 91 91 79 91 92 85 ...  
 $ ageend : num [1:49] 90 74 73 99 92 92 80 92 94 86 ...  
 $ BASstart: num [1:49] 37 31 46 46 50 32 49 37 37 43 ...  
 $ BASend : num [1:49] 43 38 45 51 60 38 59 42 45 47 ...  
 $ EAL : chr [1:49] "2" "1" "2" "2" ...  
 $ EMB : chr [1:49] "2" "2" "2" "2" ...  
 $ RRstart : chr [1:49] "20" "13" "19" "purple" ...  
 $ RRend : chr [1:49] "21" "14" "20" "purple" ...

colSums(is.na(pupil\_data))

ID school gender agestart ageend BASstart BASend EAL   
 0 0 0 0 0 0 0 0   
 EMB RRstart RRend   
 0 7 10

summary(pupil\_data$RRstart)

Length Class Mode   
 49 character character

summary(pupil\_data$RRend)

Length Class Mode   
 49 character character

# checking the characteristics of the reading level (RR) variables  
unique(pupil\_data$RRstart)

[1] "20" "13" "19" "purple" "Gold" "Green" "17" "Orange"  
 [9] "26" "4" "12" "15" "24" "Purple" "7" "11"   
[17] "18" "21" "white" NA "green" "6" "10" "8"   
[25] "14"

unique(pupil\_data$RRend)

[1] "21" "14" "20" "purple" "White" "Gold" "19" "Green"   
 [9] "18" "27" "4" "22" "13" "25" NA "12"   
[17] "white" "7" "11" "gold" "10" "16" "8"

# as per the exercise guide I will ignore the colours in this two variables and convert them in numeric   
# Remove non-numeric characters and convert to numeric  
pupil\_data$RRstart <- as.numeric(gsub("colour", "", pupil\_data$RRstart))

Warning: NAs introduced by coercion

pupil\_data$RRend <- as.numeric(gsub("colour", "", pupil\_data$RRend))

Warning: NAs introduced by coercion

# double Check how many NA are introduced  
colSums(is.na(pupil\_data[, c("RRstart", "RRend")]))

RRstart RRend   
 16 16

mean(pupil\_data$RRstart, na.rm = TRUE)

[1] 15.36364

mean(pupil\_data$RRend, na.rm = TRUE)

[1] 17.18182

# filling NA data with means as choice towards the exercise to remove missing values  
pupil\_data$RRstart[is.na(pupil\_data$RRstart)] <- mean(pupil\_data$RRstart, na.rm = TRUE)  
pupil\_data$RRend[is.na(pupil\_data$RRend)] <- mean(pupil\_data$RRend, na.rm = TRUE)  
  
# final structure manipulation and checks  
  
str(pupil\_data)

tibble [49 × 11] (S3: tbl\_df/tbl/data.frame)  
 $ ID : num [1:49] 169 164 216 176 59 54 122 61 190 256 ...  
 $ school : chr [1:49] "K3" "K3" "M" "M2" ...  
 $ gender : chr [1:49] "1" "2" "1" "2" ...  
 $ agestart: num [1:49] 89 73 72 97 91 91 79 91 92 85 ...  
 $ ageend : num [1:49] 90 74 73 99 92 92 80 92 94 86 ...  
 $ BASstart: num [1:49] 37 31 46 46 50 32 49 37 37 43 ...  
 $ BASend : num [1:49] 43 38 45 51 60 38 59 42 45 47 ...  
 $ EAL : chr [1:49] "2" "1" "2" "2" ...  
 $ EMB : chr [1:49] "2" "2" "2" "2" ...  
 $ RRstart : num [1:49] 20 13 19 15.4 15.4 ...  
 $ RRend : num [1:49] 21 14 20 17.2 17.2 ...

colSums(is.na(pupil\_data))

ID school gender agestart ageend BASstart BASend EAL   
 0 0 0 0 0 0 0 0   
 EMB RRstart RRend   
 0 0 0

summary(pupil\_data)

ID school gender agestart   
 Min. : 3.0 Length:49 Length:49 Min. : 70.00   
 1st Qu.: 79.0 Class :character Class :character 1st Qu.: 82.00   
 Median :137.0 Mode :character Mode :character Median : 87.00   
 Mean :130.3 Mean : 88.41   
 3rd Qu.:169.0 3rd Qu.: 96.00   
 Max. :256.0 Max. :110.00   
 ageend BASstart BASend EAL   
 Min. : 71.00 Min. :11.00 Min. :20.00 Length:49   
 1st Qu.: 83.00 1st Qu.:31.00 1st Qu.:38.00 Class :character   
 Median : 88.00 Median :42.00 Median :51.00 Mode :character   
 Mean : 90.12 Mean :40.06 Mean :47.51   
 3rd Qu.: 98.00 3rd Qu.:50.00 3rd Qu.:56.00   
 Max. :112.00 Max. :65.00 Max. :74.00   
 EMB RRstart RRend   
 Length:49 Min. : 4.00 Min. : 4.00   
 Class :character 1st Qu.:14.00 1st Qu.:16.00   
 Mode :character Median :15.36 Median :17.18   
 Mean :15.36 Mean :17.18   
 3rd Qu.:18.00 3rd Qu.:20.00   
 Max. :26.00 Max. :27.00

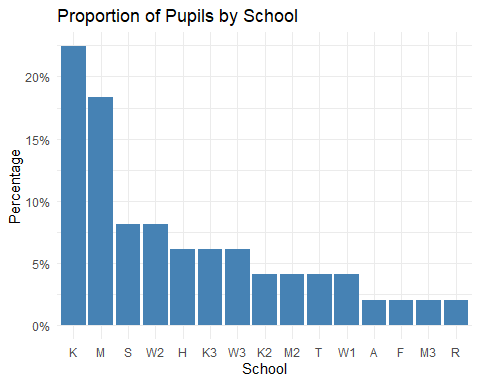
# handling groups with low frequency counts  
school\_counts <- table (pupil\_data\_complete$school)

Error: object 'pupil\_data\_complete' not found

print(school\_counts)

Error: object 'school\_counts' not found

pupil\_data |>  
 count(school) |>  
 mutate(pct = n / sum(n)) |>  
 ggplot(aes(x = reorder(school, -pct), y = pct)) +  
 geom\_col(fill = "steelblue") +  
 scale\_y\_continuous(labels = scales::percent) +  
 labs(title = "Proportion of Pupils by School",  
 x = "School", y = "Percentage") +  
 theme\_minimal()



school\_counts <- table(pupil\_data$school) # count how many pupils are per school  
print(school\_counts)

A F H K K2 K3 M M2 M3 R S T W1 W2 W3   
 1 1 3 11 2 3 9 2 1 1 4 2 2 4 3

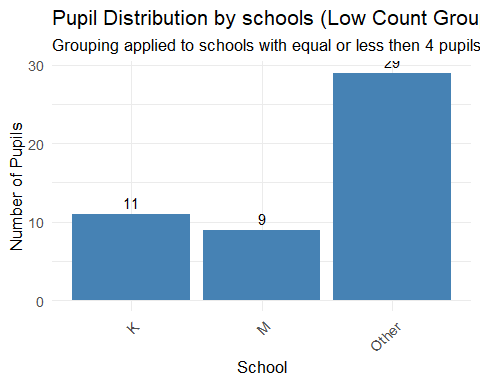
# I am identifying the schools with equal or less then 4 pupils and I am creating a new column and group them together and reorder them in a new plot  
  
low\_freq\_schools <- names(school\_counts[school\_counts <= 4])  
print(low\_freq\_schools)

[1] "A" "F" "H" "K2" "K3" "M2" "M3" "R" "S" "T" "W1" "W2" "W3"

pupil\_data$school <- ifelse(pupil\_data$school %in% low\_freq\_schools, "Other", pupil\_data$school)  
table(pupil\_data$school)

K M Other   
 11 9 29

ggplot(pupil\_data, aes(x = school)) +  
 geom\_bar(fill = "steelblue") +  
 geom\_text(stat = "count", aes(label = after\_stat(count)), vjust = -0.5, size = 4) + # Add count labels above bars  
 labs(title = "Pupil Distribution by schools (Low Count Grouped as 'Other')",  
 subtitle = "Grouping applied to schools with equal or less then 4 pupils",  
 x = "School", y = "Number of Pupils") +  
 theme\_minimal(base\_size = 13) + # Larger base font  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# checking proportions as part of descriptive statistics   
  
school\_percent <- round(100 \* prop.table(table(pupil\_data$school)), 2)  
  
print(school\_percent)

K M Other   
22.45 18.37 59.18

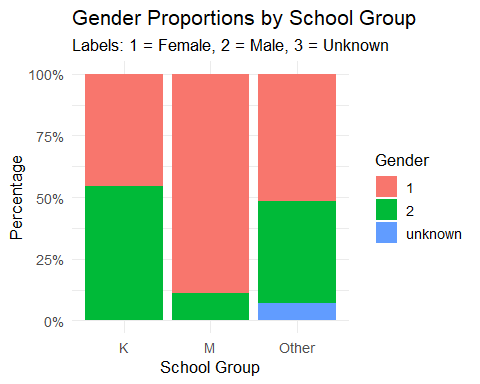
table(pupil\_data$school, pupil\_data$gender)

1 2 unknown  
 K 5 6 0  
 M 8 1 0  
 Other 15 12 2

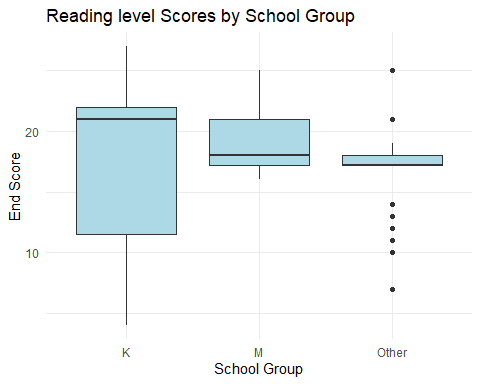
round(prop.table(table(pupil\_data$school, pupil\_data$gender), margin = 1) \* 100, 2)

1 2 unknown  
 K 45.45 54.55 0.00  
 M 88.89 11.11 0.00  
 Other 51.72 41.38 6.90

ggplot(pupil\_data, aes(x = school, fill = gender)) +  
 geom\_bar(position = "fill") +   
 scale\_y\_continuous(labels = scales::percent) +  
 labs(  
 title = "Gender Proportions by School Group",  
 subtitle = "Labels: 1 = Female, 2 = Male, 3 = Unknown",  
 x = "School Group", y = "Percentage",  
 fill = "Gender"  
 ) +  
 theme\_minimal(base\_size = 13)



# comparing reading level at te end of the intervention in a plot for visualization  
  
ggplot(pupil\_data, aes(x = school, y = RRend)) +  
 geom\_boxplot(fill = "lightblue") +  
 labs(  
 title = "Reading level Scores by School Group",  
 x = "School Group", y = "End Score"  
 ) +  
 theme\_minimal()



# Research Question 1: Is the BRP intervention equally effective for males and females?

Calculate the gain in BAS for each pupil.

Consider if there is a difference between male and female pupils.

Marks are awarded for well-designed output, and the interpretation of the output.

**(16 marks)**

### Answer:

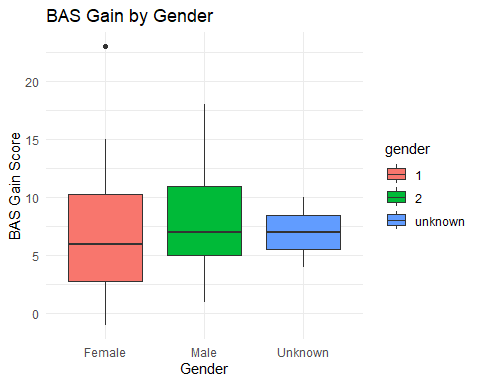
pupil\_data$BASgain <- pupil\_data$BASend - pupil\_data$BASstart  
  
summary(pupil\_data$BASgain)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 -1.000 4.000 6.000 7.449 10.000 23.000

# compare BAS gain by gender  
  
pupil\_data |>  
 group\_by(gender) |>  
 summarise(  
 count = n(),  
 mean\_gain = round(mean(BASgain, na.rm = TRUE), 2),  
 median\_gain = median(BASgain, na.rm = TRUE),  
 sd\_gain = round(sd(BASgain, na.rm = TRUE), 2)  
 )

# A tibble: 3 × 5  
 gender count mean\_gain median\_gain sd\_gain  
 <chr> <int> <dbl> <dbl> <dbl>  
1 1 28 7.07 6 5.47  
2 2 19 8.05 7 4.25  
3 unknown 2 7 7 4.24

ggplot(pupil\_data, aes(x = gender, y = BASgain, fill = gender)) +  
 geom\_boxplot() +  
 labs(  
 title = "BAS Gain by Gender",  
 x = "Gender",  
 y = "BAS Gain Score"  
 ) +  
 scale\_x\_discrete(labels = c("1" = "Female", "2" = "Male", "unknown" = "Unknown")) +  
 theme\_minimal()



# most pupils improved their British Ability Score with one pupil with 1 in regression of -1 and a higher of +23 showing the intervention was generally effective

# Research Question 2. Does the gain in BAS show the scheme is effective overall?

Generally, a one month improvement in reading equates to a BAS score improvement of 3.

Use an appropriate statistical test to see if the gain in BAS for the intervention is equal to 6. Discuss what can be concluded from the result, in the given context.

Marks are awarded for well-designed output, and the interpretation of the output.

**(12 marks)**

### Answer:

# Create a new column called BASgain  
pupil\_data$BASgain <- pupil\_data$BASend - pupil\_data$BASstart  
print(head(pupil\_data$BASgain))

[1] 6 7 -1 5 10 6

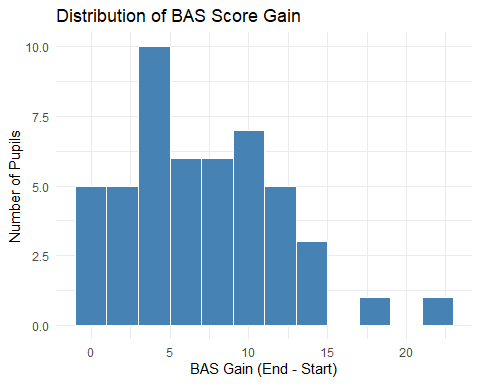
head(pupil\_data[, c("BASstart", "BASend", "BASgain")])

# A tibble: 6 × 3  
 BASstart BASend BASgain  
 <dbl> <dbl> <dbl>  
1 37 43 6  
2 31 38 7  
3 46 45 -1  
4 46 51 5  
5 50 60 10  
6 32 38 6

summary(pupil\_data$BASgain)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 -1.000 4.000 6.000 7.449 10.000 23.000

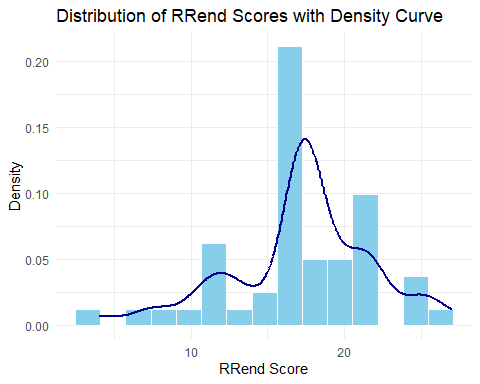
# in this print result there is a negative gain which mean one performance drop -   
# Most Pupils anyway improved their scores with one of them improving hgiher then anyone esle   
  
ggplot(pupil\_data, aes(x = BASgain)) +  
 geom\_histogram(binwidth = 2, fill = "steelblue", color = "white") +  
 labs(  
 title = "Distribution of BAS Score Gain",  
 x = "BAS Gain (End - Start)", y = "Number of Pupils"  
 ) +  
 theme\_minimal()



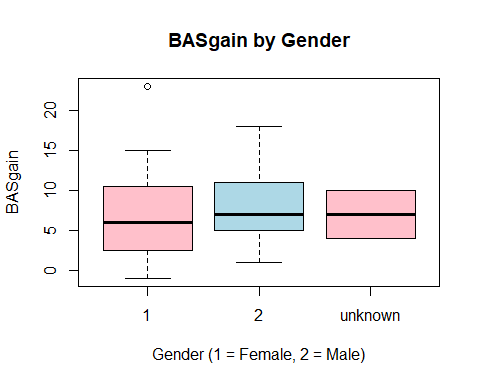
#check the distributioon  
  
ggplot(pupil\_data, aes(x = RRend)) +  
 geom\_histogram(aes(y = ..density..), bins = 15, fill = "skyblue", color = "white") +  
 geom\_density(color = "darkblue", size = 1) +  
 labs(  
 title = "Distribution of RRend Scores with Density Curve",  
 x = "RRend Score",  
 y = "Density"  
 ) +  
 theme\_minimal()

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.  
This warning is displayed once every 8 hours.  
Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
generated.

Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
ℹ Please use `after\_stat(density)` instead.  
This warning is displayed once every 8 hours.  
Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
generated.



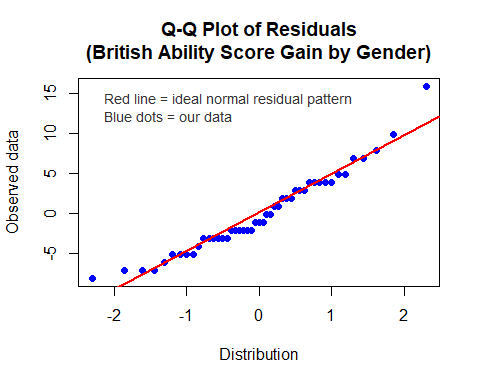
# Quick visual check of distributions with boxplot  
boxplot(BASgain ~ gender, data = pupil\_data,  
 main = "BASgain by Gender",  
 xlab = "Gender (1 = Female, 2 = Male)",  
 ylab = "BASgain",  
 col = c("pink", "lightblue"))



# we can consider the distribution 'normal' with 1 high outlier and 1 lower outlier like stated in the above summary - those outliers have minimal influence on the overall distribution shape and the are part of a natural variability of the data   
  
# Perform a one-sample t-test to check if the mean BASgain is significantly different from 6 and filtering unkown  
filtered\_data <- pupil\_data %>% filter(gender %in% c("1", "2"))  
t.test(BASgain ~ gender, data = filtered\_data)

Welch Two Sample t-test  
  
data: BASgain by gender  
t = -0.69059, df = 44.083, p-value = 0.4934  
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0  
95 percent confidence interval:  
 -3.844528 1.882122  
sample estimates:  
mean in group 1 mean in group 2   
 7.071429 8.052632

# Fit the linear model (if not already done)  
lm\_model <- lm(BASgain ~ gender, data = filtered\_data)  
  
# Extract residuals  
residuals <- resid(lm\_model)  
  
# Q-Q Plot with improved labels and annotation  
qqnorm(residuals,  
 main = "Q-Q Plot of Residuals\n(British Ability Score Gain by Gender)",  
 xlab = "Distribution",  
 ylab = "Observed data",  
 pch = 19, col = "blue")  
  
# Add Q-Q reference line  
qqline(residuals, col = "red", lwd = 2)  
  
# Add legend-style annotation in plot area  
legend("topleft",  
 legend = "Red line = ideal normal residual pattern\nBlue dots = our data",  
 bty = "n", cex = 0.9, text.col = "gray20")



# Run the one-sample t-test  
t\_test\_result <- t.test(pupil\_data$BASgain, mu = 6)  
  
# Extract key values and perform test  
mean\_gain <- mean(pupil\_data$BASgain, na.rm = TRUE)  
t\_stat <- round(t\_test\_result$statistic, 3)  
df <- round(t\_test\_result$parameter, 2)  
p\_value <- round(t\_test\_result$p.value, 4)  
ci <- round(t\_test\_result$conf.int, 2)  
  
if (p\_value < 0.05) {  
 cat("Conclusion: The difference is statistically significant (p < 0.05).\n")  
 cat("Reject the null hypothesis. The BAS gain is significantly different from 6.\n")  
} else {  
 cat("Conclusion: The difference is not statistically significant (p ≥ 0.05).\n")  
 cat("Fail to reject the null hypothesis. The BAS gain is not significantly different from 6.\n")  
}

Conclusion: The difference is statistically significant (p < 0.05).  
Reject the null hypothesis. The BAS gain is significantly different from 6.

# Research Question 3: Are teachers judgement of reading level (RR) appropriate based on BAS evidence?

Create a simple linear regression model to show if the teachers judgement of reading level (RR) can be predicted by BAS.

Your answer should include:

* justification for the attempted statistical model, including any additional supporting exploratory data analyses used to make the choice;
* interpretation of the model and the validity of model assumptions.

**(23 marks)**

### Answer:

# double check numeric  
str(pupil\_data$RRend)

num [1:49] 21 14 20 17.2 17.2 ...

str(pupil\_data$BASend)

num [1:49] 43 38 45 51 60 38 59 42 45 47 ...

colSums(is.na(pupil\_data)) # check missing values

ID school gender agestart ageend BASstart BASend EAL   
 0 0 0 0 0 0 0 0   
 EMB RRstart RRend BASgain   
 0 0 0 0

# exploring the visual relationship  
  
plot(pupil\_data\_complete$BASend, pupil\_data\_complete$RRend,  
 main = "Teacher Judgement (RRend) vs BAS Score (BASend)",  
 xlab = "BAS Score (BASend)",  
 ylab = "Reading Level (RRend)",  
 pch = 19, col = "blue")

Error: object 'pupil\_data\_complete' not found

abline(lm(RRend ~ BASend, data = pupil\_data\_complete), col = "red", lwd = 2)

Error in eval(mf, parent.frame()): object 'pupil\_data\_complete' not found

# creating the model   
  
model <- lm(RRend ~ BASend, data = pupil\_data\_complete)

Error in eval(mf, parent.frame()): object 'pupil\_data\_complete' not found

summary(model)

Error: object 'model' not found

# validity check and homoscedasticity check  
  
qqnorm(resid(model))

Error: object 'model' not found

qqline(resid(model), col = "red")

Error: object 'model' not found

plot(fitted(model), resid(model))

Error: object 'model' not found

abline(h = 0, col = "red", lty = 2)

Error in int\_abline(a = a, b = b, h = h, v = v, untf = untf, ...): plot.new has not been called yet

plot(fitted(model), resid(model))

Error: object 'model' not found

abline(h = 0, col = "red", lty = 2)

Error in int\_abline(a = a, b = b, h = h, v = v, untf = untf, ...): plot.new has not been called yet

# The relationship was statistically significant, t(47) = 6.66, p < 0.001. The model explains approximately 48.5% of the variance in RR scores (R² = 0.49). This suggests a strong, positive association: as BAS scores increase, so do teachers' assessments of reading levels.  
  
# Residuals were approximately normally distributed, and no violations of homoscedasticity were detected, supporting the validity of the model assumptions.  
  
# visualizing the model  
  
# Create the scatterplot with regression line and confidence interval  
ggplot(pupil\_data, aes(x = BASend, y = RRend)) +  
 geom\_point(color = "blue", size = 2) +  
 geom\_smooth(method = "lm", color = "red", fill = "pink", se = TRUE) +  
 labs(  
 title = "Relationship Between BAS Score and Teacher Judgement (RR)",  
 x = "BAS Score (BASend)",  
 y = "Reading Level Judgement (RRend)"  
 ) +  
 theme\_minimal()

`geom\_smooth()` using formula = 'y ~ x'

# Your Research Question: Report

Clearly state one research question based on the data set supplied to you. This may be an extension to the research questions considered above, or a new research question making use of the available data. Explain why this is a worthwhile new research question or extension to consider.

You are required to write a short report for the client showing your analyses of the data set provided, based only on the research question or extension you have selected.

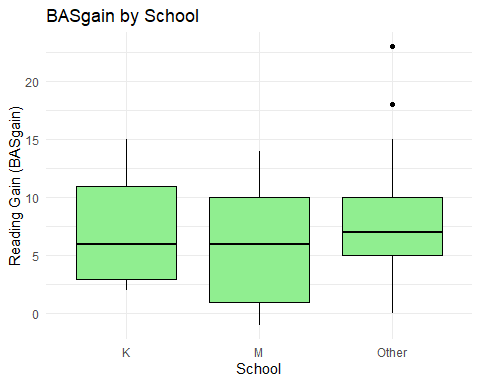
In your report you may wish to include a number of the following: exploratory data analyses; a hypothesis test; data modelling; discussion of limitations; how you could extend the research if given more time.

To clarify, your answer to this question must be a report based on additional considerations not already presented in the above. This report should contain at most 5 outputs (i.e. graphics + tables) and at most 500 words.

**(35 marks)**

### Answer:

# final reporting: Does the effectiveness of the intervention differ between school groups?  
  
# Remove any NA or unknown school values if needed  
  
filtered\_school\_data <- pupil\_data |>  
 filter(!is.na(school)) # or add extra filtering if needed  
  
# Boxplot of BASgain by school  
  
ggplot(filtered\_school\_data, aes(x = school, y = BASgain)) +  
 geom\_boxplot(fill = "lightgreen", color = "black") +  
 labs(title = "BASgain by School",  
 x = "School",  
 y = "Reading Gain (BASgain)") +  
 theme\_minimal()



# Summary stats per school group  
group\_summary <- filtered\_school\_data |>  
 group\_by(school) %>%  
 summarise(  
 mean\_gain = round(mean(BASgain, na.rm = TRUE), 2),  
 sd\_gain = round(sd(BASgain, na.rm = TRUE), 2),  
 count = n()  
 )  
print(group\_summary)

# A tibble: 3 × 4  
 school mean\_gain sd\_gain count  
 <chr> <dbl> <dbl> <int>  
1 K 7.09 4.59 11  
2 M 6.33 5.5 9  
3 Other 7.93 4.96 29

# One-way ANOVA test to compare means across schools  
anova\_result <- aov(BASgain ~ school, data = filtered\_school\_data)  
summary(anova\_result)

Df Sum Sq Mean Sq F value Pr(>F)  
school 2 19.4 9.676 0.389 0.68  
Residuals 46 1142.8 24.843

The aim of this report is to investigate and explore the impact of a Book Reading Programme (BRP) intervention on pupils’ reading ability. One of the research questions addressed in this analysis is: “Have pupils improved their reading following the intervention?”

To assess this, I created a new variable, BASgain, which represents the difference between pupils’ pre- and post-intervention BAS scores. The results showed an average gain across the sample of 7.45, with a median of 6, indicating that pupils experienced greater-than-expected progress overall. One pupil showed a small decline (-1), while another achieved a very high gain (+23). The distribution of gains is generally consistent, as shown in the graphic above, with a slight right skew due to higher performers.

In addition to overall reading improvement, the report examined whether the intervention was equally effective for both male and female pupils. I grouped pupils by gender and compared their BAS scores using descriptive statistics and visualizations. The results showed that both groups had positive gains, with no significant difference in gain scores between them, as confirmed by a statistically non-significant t-test.

A new question explored in this final report is:

Does the effectiveness of the intervention differ between school groups?

As a result of ANOVA testing, there is no significant difference in intervention effectiveness between pupils across different groups.

This test was used to compare the means of all three school groups I have clearly grouped into K, M, and Other at the beginning of this session, incorporating all the pupils that are equal to or less than 4 in a new group to satisfy the purpose of the testing and the answer.

# End matter - Session Information

Do not edit this part. Make sure that you compile your document so that the information about your session (including software / package versions) is included in your submission.

sessionInfo()

R version 4.5.0 (2025-04-11 ucrt)  
Platform: x86\_64-w64-mingw32/x64  
Running under: Windows 11 x64 (build 26100)  
  
Matrix products: default  
 LAPACK version 3.12.1  
  
locale:  
[1] LC\_COLLATE=English\_United Kingdom.utf8   
[2] LC\_CTYPE=English\_United Kingdom.utf8   
[3] LC\_MONETARY=English\_United Kingdom.utf8  
[4] LC\_NUMERIC=C   
[5] LC\_TIME=English\_United Kingdom.utf8   
  
time zone: Europe/London  
tzcode source: internal  
  
attached base packages:  
[1] stats graphics grDevices utils datasets methods base   
  
other attached packages:  
 [1] lubridate\_1.9.4 forcats\_1.0.0 stringr\_1.5.1 dplyr\_1.1.4   
 [5] purrr\_1.0.4 readr\_2.1.5 tidyr\_1.3.1 tibble\_3.2.1   
 [9] ggplot2\_3.5.2 tidyverse\_2.0.0  
  
loaded via a namespace (and not attached):  
 [1] Matrix\_1.7-3 bit\_4.6.0 gtable\_0.3.6 crayon\_1.5.3   
 [5] compiler\_4.5.0 tidyselect\_1.2.1 parallel\_4.5.0 splines\_4.5.0   
 [9] scales\_1.4.0 yaml\_2.3.10 fastmap\_1.2.0 lattice\_0.22-6   
[13] R6\_2.6.1 labeling\_0.4.3 generics\_0.1.4 knitr\_1.50   
[17] pillar\_1.10.2 RColorBrewer\_1.1-3 tzdb\_0.5.0 rlang\_1.1.6   
[21] utf8\_1.2.5 stringi\_1.8.7 xfun\_0.52 bit64\_4.6.0-1   
[25] timechange\_0.3.0 cli\_3.6.5 mgcv\_1.9-1 withr\_3.0.2   
[29] magrittr\_2.0.3 digest\_0.6.37 grid\_4.5.0 vroom\_1.6.5   
[33] rstudioapi\_0.17.1 hms\_1.1.3 nlme\_3.1-168 lifecycle\_1.0.4   
[37] vctrs\_0.6.5 evaluate\_1.0.3 glue\_1.8.0 farver\_2.1.2   
[41] rmarkdown\_2.29 tools\_4.5.0 pkgconfig\_2.0.3 htmltools\_0.5.8.1