Thesis Data Analysis: The Psychology of Scientific Fraud

Benjamin Zubaly

2024-03-04

Table of contents

Introduction **2**
- Overview of Sections . . . 2
- Directory Set-Up . . . 2
- Variable Definitions . . . 2
- Initial Package Installations . . . 3
- Initial Import of Data . . . 5

Data Cleaning **6**

Data Exploration **7**
- Missing Data . . . 7
- Descriptive Statistics . . . 13
- Data Visualization . . . 29
- Bivariate Correlations . . . 43

Testing Hypotheses **46**
- **Hypothesis 1: Linguistic Obfuscation Hypothesis** . . . 46
- **Hypothesis 2: References Hypothesis** . . . 55
- **Hypothesis 3: Certainty** . . . 61
- **Hypothesis 4:** . . . 64

Saving Data **66**

Introduction

This document is intended to track the data analysis for my undergraduate thesis project on the psychology of scientific fraud. I have already planned out (and preregistered) my data analysis plan (see here), and I will note throughout the document if I deviate from this plan and why. This project is also being tracked in a private GitHub repository in case I need to revert back to a previous version of the project due to a fatal error.

Overview of Sections

- **Data Cleaning:** The data cleaning section will take the finalized dataset that I have included in the project directory and use it to create the remaining variables that we need for our analysis. We will also save this dataset.
- **Data Exploration:** In the data exploration section, we will check for and deal with missing data, run descriptive statistics of our outcome variables, visualize our data, and run bivariate correlations.
- **Testing Hypotheses:** In the hypothesis testing section, we will test each of our hypotheses one-by-one, according to our analysis plan.

Directory Set-Up

In order to ensure that this analysis is computationally reproducible, I have included everything that is needed to complete this analysis (just the finalized data file, "study_dataset.csv") in the current working directory. The code will be displayed before the output of each analysis.

Variable Definitions

Although the following variables may not exist until after the data cleaning section, here is what each variable name refers to, so that you can refer to them while examining the code.

`DOI`: Unique identifier for each paper (i.e., the paper DOI).

`PaperType`: Categorical variable indicating SAFP, SAGP, MAFP, or MAGP.

`LingObf`: Continuous variable for linguistic obfuscation.

`CertSent`: Continuous variable for certainty sentiment.

`Refs`: Count variable for references.

`FraudCorrAuth`: Dichotomous variable indicating if the fraudulent author is the corresponding author (1) or is not (0). Unknown cases will be marked in a seperate variable with this variable left blank.

`NumAuth`: Count variable indicating the number of authors for each paper.

`abstraction`: Abstraction index composed of the sum of standardized scores for `article`, `prep`, and `quantity`.

`article`: Articles from LIWC.

`prep`: Prepositions from LIWC.

`quantity`: Quantities from LIWC.

`cause`: Causation terms from LIWC.

`jargon`: The percent of words not captured by LIWC (100-`Dic`).

`Dic`: The percentage of words captured by all LIWC dictionaries.

`emo_pos`: Positive emotion terms from LIWC.

`flesch_re`: Flesch Reading Ease from ARTE.

Initial Package Installations

If you would like to reproduce this analysis, here are the package I will be using, so that they can be cued before starting.

```
install.packages("readr")       # For reading data
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)

The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//RtmpFV2ZlM/downloaded_packages
```

```
install.packages("dplyr")       # For data manipulation and handling of missing data
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)

The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//RtmpFV2ZlM/downloaded_packages
```

```r
install.packages("psych")       # For descriptive statistics, correlations
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//RtmpFV2ZlM/downloaded_packages
```

```r
install.packages("ggplot2")     # For data visualization
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//RtmpFV2ZlM/downloaded_packages
```

```r
install.packages("car")         # For diagnostic tests such as Levene's test
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//RtmpFV2ZlM/downloaded_packages
```

```r
install.packages("effsize")     # For calculating effect sizes
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)
```

```
The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//RtmpFV2ZlM/downloaded_packages
```

```r
install.packages("rcompanion")  # For Games-Howell post-hoc test (if applicable)
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)

The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//RtmpFV2ZlM/downloaded_packages
```

```r
install.packages("dunn.test")   # For Dunn post-hoc test (if applicable)
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)

The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//RtmpFV2ZlM/downloaded_packages
```

```r
install.packages("boot")        # For bootstrapped hypothesis tests (if applicable)
```

```
Installing package into '/Users/benjaminzubaly/Library/R/x86_64/4.3/library'
(as 'lib' is unspecified)

The downloaded binary packages are in
    /var/folders/38/1ybnplc53zdb089bn6drqfn00000gn/T//RtmpFV2ZlM/downloaded_packages
```

Initial Import of Data

We will not load in the dataset "study_dataset.csv" from the working directory.

```r
library(readr) # Loading the readr package

data <- read_csv("study_dataset.csv") # Loading in study dataset as "data"
```

```
Rows: 88 Columns: 207
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (26): DOI, Title, Subject, Institution, Journal, Publisher, Country, Au...
dbl (181): Record ID, RetractionPubMedID, OriginalPaperDate, year, OriginalP...

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.
```

I have viewed the data frame, and the data seems to have loaded correctly.

Data Cleaning

To conduct the analysis, we will first need to calculate the `LingObf` variable by calculating the `abstraction` index and `jargon` words; creating standardized scores for `abstraction`, `cause`, `jargon`, `emo_pos`, and `flesch_re`; and calculating the `LingObf` composite variable from these standardized scores.

1. First, we will calculate the `abstraction` index by creating standardized scores for `article`, `prep`, and `quantity` and summing them.

```
# Calculate standardized scores for article, prep, and quantity and add them to the datase
data$articles_standardized <- scale(data$article)
data$prep_standardized <- scale(data$prep)
data$quantity_standardized <- scale(data$quantity)

# Create the new variable 'abstraction' as the sum of the three standardized variables
data$abstraction <- (data$articles_standardized + data$prep_standardized + data$quantity_s
```

- After viewing the data, the transformations and variable calculation seem to have occurred appropriately.

2. Next, we will calculate the `jargon` words by subtracting `Dic` from 100.

```
# Calculate the new variable 'jargon' by subtracting 'Dic' from 100
data$jargon <- (100 - data$Dic)
```

- After viewing the data, the variable calculation seem to have occurred appropriately.

3. Next, we will create standardized scores for each subcomponent of the `LingObf`.

```
# Standardize the new set of variables and add them to the dataset
data$abstraction_standardized <- scale(data$abstraction)
data$cause_standardized <- scale(data$cause)
data$jargon_standardized <- scale(data$jargon)
data$emo_pos_standardized <- scale(data$emo_pos)
data$flesch_re_standardized <- scale(data$flesch_re)
```

- After viewing the data, the variable transformations seem to have occurred appropriately.

4. Now we will calculate the `LingObf` variable using the following formula: [cause_standardized + abstraction_standardized + jargon_standardized] – [emo_pos_standardized + flesch_re_standardized].

```
# Calculate 'LingObf'
data$LingObf <- (data$cause_standardized + data$abstraction_standardized + data$jargon_sta
```

- After viewing the data, the variable calculation seem to have occurred appropriately.

5. Lastly, our variable that indicates certainty sentiment is currently `certainty_avg`, but to make things easier I am going to copy this data into a new variables called `CertSent`.

```
data$CertSent <- data$certainty_avg
```

To ensure that our clean data is saved, we will write the dataset to the current working directory.

```
# Writing our data as a csv file in the current working directory
write.csv(data, "clean_study_data.csv")
```

- I have opened the saved data file outside of Rstudio, and it seems to have been written correctly.

Data Exploration

Missing Data

1. Data will be first inspected for missing scores.

```
library(dplyr) # Loading the dplyr package for data manipulation and handling missing valu
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
# To summarize the number of missing values in each column
missing_data_summary <- sapply(data, function(x) sum(is.na(x)))

print(missing_data_summary) # To see summary of missing values for all columns
```

```
                    DOI                 Record ID
                      0                        44
                  Title                   Subject
                      0                        44
            Institution                   Journal
                      0                         0
              Publisher                   Country
                      0                         0
                 Author                      URLS
                      0                        68
            ArticleType            RetractionDate
                      0                        44
         RetractionDOI         RetractionPubMedID
                     48                        51
      OriginalPaperDate                      year
                     44                         0
       OriginalPaperDOI       OriginalPaperPubMedID
                      0                        51
       RetractionNature                    Reason
                     44                        44
             Paywalled                     Notes
                     44                        70
         simple_reason                   NumAuth
                     44                         0
             inst_pres                 PaperType
                      0                         0
                gender       matched_DOI_SAFP_MAFP
                      0                        44
 matched_DOI_SAFP_SAGP       matched_DOI_MAFP_MAGP
                     44                        44
  matching_concessions               FraudCorrAuth
                     42                        80
           FCA_unknown                  FCA_notes
                     66                        66
    different_country             year_difference
                     22                        22
     different_gender           different_inst_pres
```

22	22
different_journal	Refs
66	0
Abstract	flesch_re
0	0
wordcount	valence_min
0	0
valence_max	valence_avg
0	0
valence_std	extremity_min
0	0
extremity_max	extremity_avg
0	0
extremity_std	extremity_min_pos
0	0
extremity_max_pos	extremity_avg_pos
0	0
extremity_std_pos	extremity_min_neg
0	1
extremity_max_neg	extremity_avg_neg
1	1
extremity_std_neg	extremity_PosMinNeg
5	0
emotionality_min	emotionality_max
0	0
emotionality_avg	emotionality_std
0	0
emotionality_min_pos	emotionality_max_pos
0	0
emotionality_avg_pos	emotionality_std_pos
0	0
emotionality_min_neg	emotionality_max_neg
1	1
emotionality_avg_neg	emotionality_std_neg
1	5
emotionality_PosMinNeg	count_unique_evaluative_pos
0	0
count_total_evaluative_pos	count_unique_evaluative_neg
0	0
count_total_evaluative_neg	count_unique_evaluative
0	0
count_total_evaluative	count_evaluative_PosMinNeg
0	0

ambivalent	pos_dichotomous
0	87
certainty_min	certainty_max
0	0
certainty_avg	certainty_std
0	0
count_unique_certainty	count_total_certainty
0	0
Segment	WC
0	0
Analytic	Clout
0	0
Authentic	Tone
0	0
WPS	BigWords
0	0
Dic	Linguistic
0	0
function	pronoun
0	0
ppron	i
0	0
we	you
0	0
shehe	they
0	0
ipron	det
0	0
article	number
0	0
prep	auxverb
0	0
adverb	conj
0	0
negate	verb
0	0
adj	quantity
0	0
Drives	affiliation
0	0
achieve	power
0	0
Cognition	allnone

0	0
cogproc	insight
0	0
cause	discrep
0	0
tentat	certitude
0	0
differ	memory
0	0
Affect	tone_pos
0	0
tone_neg	emotion
0	0
emo_pos	emo_neg
0	0
emo_anx	emo_anger
0	0
emo_sad	swear
0	0
Social	socbehav
0	0
prosocial	polite
0	0
conflict	moral
0	0
comm	socrefs
0	0
family	friend
0	0
female	male
0	0
Culture	politic
0	0
ethnicity	tech
0	0
Lifestyle	leisure
0	0
home	work
0	0
money	relig
0	0
Physical	health
0	0

illness	wellness
0	0
mental	substances
0	0
sexual	food
0	0
death	need
0	0
want	acquire
0	0
lack	fulfill
0	0
fatigue	reward
0	0
risk	curiosity
0	0
allure	Perception
0	0
attention	motion
0	0
space	visual
0	0
auditory	feeling
0	0
time	focuspast
0	0
focuspresent	focusfuture
0	0
Conversation	netspeak
0	0
assent	nonflu
0	0
filler	AllPunc
0	0
Period	Comma
0	0
QMark	Exclam
0	0
Apostro	OtherP
0	0
Emoji	articles_standardized
0	0
prep_standardized	quantity_standardized

```
                         0                          0
               abstraction                    jargon
                         0                          0
 abstraction_standardized          cause_standardized
                         0                          0
      jargon_standardized        emo_pos_standardized
                         0                          0
   flesch_re_standardized                     LingObf
                         0                          0
                 CertSent
                         0
```

- Taking a look at our outcome variables, there are no missing scores, so we do not need to impute any values.

Descriptive Statistics

1. I will now generate descriptive statistics for each relevant variable in the dataset. I originally (in the preregistered plan) was simply going to deploy the `describe()` function on the entire dataset, but because I retained all of the columns from the Retraction Watch Database (*Retraction Watch Database*, 2023) and all of the output from the text analysis packages (Aggarwal, 2022; Boyd et al., 2022; Rocklage et al., 2023) there are currently 219 variables. Because this would be unmanageable, I am going to only calculate descriptive statistics for a selection of variables of interest. I will first create a dataframe with only the continuous variables that I am interested in generating descriptive statistics for, and I will use the `psych` package to produce the descriptive statistics for these variables.

```r
library(psych) # Loading the psych package

# Selecting the continuous variables I am interested in getting descriptive statistics for
continuous_data_for_descriptives <- data[c("year", "Refs", "flesch_re", "WC", "abstraction

# Generating descriptive statistics for variables of interest
continuous_descriptive_stats_all <- describe(continuous_data_for_descriptives)

# Displaying the results of the descriptive stats for the variables listed above
continuous_descriptive_stats_all
```

```
           vars  n    mean      sd  median trimmed     mad     min      max
year          1 88 2009.70   10.47 2013.00 2011.14    7.41 1980.00  2022.00
Refs          2 88   45.82   23.47   39.00   43.21   20.02   12.00   146.00
```

```
flesch_re      3 88   34.90   10.94   36.45   35.46   10.28  -10.89    56.59
WC             4 88 4436.25 2248.24 3951.50 4238.17 1956.29 1124.00 13901.00
abstraction    5 88    0.00    1.83    0.09    0.04    1.86   -4.46     3.55
jargon         6 88   33.33    7.47   34.76   33.50    7.84   17.91    46.67
CertSent       7 88    6.17    0.25    6.18    6.18    0.22    5.21     6.76
LingObf        8 88    0.00    1.67   -0.14   -0.01    1.54   -5.84     4.28
cause          9 88    2.80    1.22    2.62    2.69    1.02    0.89     6.01
emo_pos       10 88    0.09    0.22    0.04    0.05    0.06    0.00     1.67
article       11 88    8.38    2.20    8.31    8.33    2.13    3.53    14.47
prep          12 88   15.43    1.68   15.66   15.51    1.34    9.76    20.67
quantity      13 88    4.23    1.32    4.08    4.14    1.33    1.69     8.66
PaperType*    14 88    2.50    1.12    2.50    2.50    1.48    1.00     4.00
               range  skew kurtosis     se
year           42.00 -1.25     0.74   1.12
Refs          134.00  1.41     2.87   2.50
flesch_re      67.48 -0.85     2.16   1.17
WC          12777.00  1.21     2.18 239.66
abstraction     8.01 -0.20    -0.42   0.20
jargon         28.76 -0.27    -0.94   0.80
CertSent        1.55 -0.69     1.42   0.03
LingObf        10.12 -0.21     1.40   0.18
cause           5.12  0.80     0.14   0.13
emo_pos         1.67  5.66    34.22   0.02
article        10.94  0.33    -0.19   0.23
prep           10.91 -0.44     1.22   0.18
quantity        6.97  0.67     0.42   0.14
PaperType*      3.00  0.00    -1.40   0.12
```

- I won't make too many comments here, because for most of these measures there are no formal or informal norms against which to judge them. That being said:
  - **Year:** The mean year is mid-2009, with a standard deviation of 10 years (max 2022 and min 1980), indicating that the sample is recent enough to be relevant but also spans quite a number of years. This I think is good insofar as the recent history of academic publishing is represented more fully (some recent investigations limited their search to the three years prior to publication). There is some negative skew, which I suspect is due to the 1980 paper being quite a bit older than most papers.
  - **Refs:** The mean number of references is 45.82 with a standard deviation of 23.47. At least intuitively, this seems like a pretty standard distribution of references if we were to randomly select papers from the literature. However, the range is huge, with one paper showing 146 references—perhaps why the skew is positive. This may be something to consider when making group comparisons, as this point may have significant leverage.

- **WC:** The average word count is 4436.25, and the standard deviation is 2248.24. It is also positively skewed, and although this should not affect our linguistic dependent variables it may be a confounding factor for references. Indeed, taking a look at the dataset I can see that the second highest value for word count (9670) also has the maximum value for references (the outlier noted above at 146 references). This will be something to watch out for.

- Now we will create descriptive statistics for each of the continuous variables above within the `PaperType` groups.

```
# Making PaperType a factor variable in the continuous variable dataframe to allow for gro
continuous_data_for_descriptives$PaperType <- factor(continuous_data_for_descriptives$Pape

# Generating descriptive statistics within PaperType groups
descriptive_stats_by_PaperType <- describeBy(continuous_data_for_descriptives, group = con

descriptive_stats_by_PaperType
```

```
 Descriptive statistics by group
group: Single-Authored Fraudulent Papers
            vars  n    mean      sd  median trimmed     mad     min     max
year           1 22 2008.95   11.74 2013.00 2010.39    6.67 1983.00 2022.00
Refs           2 22   44.18   19.93   36.00   42.89   17.79   18.00   87.00
flesch_re      3 22   36.09    9.17   37.06   36.29   10.53   18.49   54.36
WC             4 22 3998.91 1741.31 3796.50 3911.39 1829.53 1273.00 7330.00
abstraction    5 22    0.21    1.54    0.04    0.27    1.25   -3.27    2.95
jargon         6 22   34.23    8.33   36.94   34.90    8.49   17.91   44.35
CertSent       7 22    6.21    0.21    6.18    6.20    0.18    5.87    6.66
LingObf        8 22   -0.24    1.79   -0.04   -0.05    1.97   -4.23    2.25
cause          9 22    2.71    1.48    2.50    2.56    1.07    0.89    6.01
emo_pos       10 22    0.15    0.35    0.06    0.08    0.06    0.00    1.67
article       11 22    8.35    2.60    8.48    8.41    2.90    3.53   12.25
prep          12 22   15.75    1.60   15.75   15.79    1.43   11.79   18.42
quantity      13 22    4.27    0.97    4.11    4.21    1.13    2.59    6.59
PaperType     14 22    1.00    0.00    1.00    1.00    0.00    1.00    1.00
              range  skew kurtosis     se
year          39.00 -1.02    -0.29   2.50
Refs          69.00  0.56    -1.06   4.25
flesch_re     35.87 -0.01    -0.63   1.95
WC          6057.00  0.54    -0.67 371.25
abstraction    6.21 -0.23    -0.33   0.33
jargon        26.44 -0.56    -1.04   1.78
```

```
CertSent       0.78  0.44    -0.57   0.04
LingObf        6.48 -0.70    -0.32   0.38
cause          5.12  0.93    -0.05   0.32
emo_pos        1.67  3.81    13.82   0.07
article        8.72 -0.16    -1.18   0.55
prep           6.63 -0.34    -0.15   0.34
quantity       4.00  0.49    -0.41   0.21
PaperType      0.00   NaN      NaN   0.00
------------------------------------------------------------
group: Multi-Authored Fraudulent Papers
            vars  n    mean      sd  median trimmed     mad     min     max
year           1 22 2010.14    9.59 2012.50 2011.67    8.15 1982.00 2022.00
Refs           2 22   44.00   19.66   39.50   42.06   17.05   18.00   94.00
flesch_re      3 22   34.18    9.64   31.41   33.58    9.38   20.08   56.59
WC             4 22 4705.91 2047.63 4274.00 4559.39 1931.83 1823.00 9004.00
abstraction    5 22   -0.45    1.83   -0.53   -0.44    1.95   -4.10    3.55
jargon         6 22   35.97    6.97   37.36   36.23    6.28   23.85   46.48
CertSent       7 22    6.22    0.18    6.20    6.23    0.24    5.82    6.47
LingObf        8 22    0.47    1.43    0.18    0.36    1.57   -1.66    4.13
cause          9 22    2.84    1.08    2.74    2.75    1.03    1.14    5.66
emo_pos       10 22    0.03    0.04    0.01    0.02    0.02    0.00    0.15
article       11 22    8.19    1.94    8.51    8.12    1.53    4.90   12.66
prep          12 22   14.84    1.64   15.19   14.84    1.75   11.80   17.58
quantity      13 22    4.22    1.39    4.10    4.15    1.54    2.27    6.95
PaperType     14 22    2.00    0.00    2.00    2.00    0.00    2.00    2.00
              range  skew kurtosis     se
year          40.00 -1.37     1.52   2.05
Refs          76.00  0.86    -0.01   4.19
flesch_re     36.51  0.59    -0.62   2.05
WC          7181.00  0.56    -0.81 436.56
abstraction    7.65  0.02    -0.64   0.39
jargon        22.63 -0.38    -1.16   1.49
CertSent       0.64 -0.27    -1.00   0.04
LingObf        5.79  0.68    -0.15   0.30
cause          4.52  0.78     0.30   0.23
emo_pos        0.15  1.55     1.39   0.01
article        7.76  0.16    -0.34   0.41
prep           5.78 -0.11    -1.18   0.35
quantity       4.68  0.46    -1.11   0.30
PaperType      0.00   NaN      NaN   0.00
------------------------------------------------------------
group: Single-Authored Genuine Papers
            vars  n    mean      sd  median trimmed     mad     min     max
```

```
year           1 22 2008.95   12.24 2013.50 2010.61    5.93 1980.00 2022.00
Refs           2 22   51.05   32.53   40.50   46.28   28.91   12.00  146.00
flesch_re      3 22   34.46   15.72   38.40   36.15   10.34  -10.89   54.97
WC             4 22 4632.36 2424.81 4148.00 4491.00 2364.01 1141.00 9670.00
abstraction    5 22    0.80    1.89    1.00    0.91    2.13   -3.39    3.43
jargon         6 22   29.50    6.19   31.79   29.61    5.77   19.26   38.66
CertSent       7 22    6.07    0.28    6.13    6.07    0.31    5.55    6.49
LingObf        8 22   -0.25    1.77   -0.30   -0.15    1.56   -5.84    2.97
cause          9 22    2.73    1.04    2.55    2.67    1.05    1.18    4.91
emo_pos       10 22    0.12    0.26    0.06    0.06    0.09    0.00    1.23
article       11 22    9.13    2.45    8.61    8.95    2.64    5.91   14.47
prep          12 22   15.66    1.15   15.89   15.83    0.76   12.82   17.22
quantity      13 22    4.66    1.64    4.46    4.55    1.43    2.05    8.66
PaperType     14 22    3.00    0.00    3.00    3.00    0.00    3.00    3.00
               range  skew kurtosis     se
year           42.00 -1.11    -0.07   2.61
Refs          134.00  1.32     1.43   6.94
flesch_re      65.86 -1.16     0.90   3.35
WC           8529.00  0.42    -0.93 516.97
abstraction     6.82 -0.42    -0.84   0.40
jargon         19.40 -0.30    -1.34   1.32
CertSent        0.94 -0.27    -1.13   0.06
LingObf         8.81 -1.03     2.29   0.38
cause           3.73  0.42    -0.61   0.22
emo_pos         1.23  3.38    11.22   0.06
article         8.56  0.48    -0.91   0.52
prep            4.40 -1.22     0.76   0.25
quantity        6.61  0.62    -0.27   0.35
PaperType       0.00   NaN      NaN   0.00
------------------------------------------------------------
group: Multi-Authored Genuine Papers
            vars  n    mean      sd  median trimmed     mad     min      max
year           1 22 2010.77    8.50 2012.50 2011.89    8.15 1990.00  2022.00
Refs           2 22   44.05   20.05   36.50   41.94   15.57   19.00    91.00
flesch_re      3 22   34.86    8.34   36.17   35.12    6.35   15.95    49.45
WC             4 22 4407.82 2741.91 3624.00 4009.67 1307.65 1124.00 13901.00
abstraction    5 22   -0.56    1.83   -0.42   -0.48    1.65   -4.46     3.07
jargon         6 22   33.61    7.19   35.03   33.51    6.42   20.86    46.67
CertSent       7 22    6.19    0.31    6.22    6.20    0.19    5.21     6.76
LingObf        8 22    0.02    1.69   -0.45   -0.16    1.05   -2.50     4.28
cause          9 22    2.90    1.30    2.44    2.83    0.79    1.04     5.56
emo_pos       10 22    0.04    0.05    0.04    0.04    0.05    0.00     0.17
article       11 22    7.88    1.62    7.71    7.78    1.06    4.22    12.06
```

```
prep          12 22   15.48    2.15   15.66   15.58    1.41    9.76    20.67
quantity      13 22    3.75    1.09    3.74    3.76    1.21    1.69     5.53
PaperType     14 22    4.00    0.00    4.00    4.00    0.00    4.00     4.00
               range  skew kurtosis     se
year           32.00 -0.95     0.15   1.81
Refs           72.00  0.77    -0.61   4.27
flesch_re      33.50 -0.41    -0.37   1.78
WC          12777.00  1.90     3.92 584.58
abstraction     7.53 -0.26    -0.37   0.39
jargon         25.81 -0.16    -1.01   1.53
CertSent        1.55 -1.02     2.09   0.07
LingObf         6.78  1.08     0.44   0.36
cause           4.52  0.65    -0.77   0.28
emo_pos         0.17  0.88     0.00   0.01
article         7.84  0.52     0.94   0.35
prep           10.91 -0.37     1.33   0.46
quantity        3.84 -0.05    -1.19   0.23
PaperType       0.00   NaN      NaN   0.00
```

- Some quick notes:
  - **Year:** The year seems to be relatively similar between groups, with the single-author groups (SAGP and SAFP) both having mid-2008 as the mean and the multi-author groups both having mid-2010 as the mean. This may be partially due to there being single outliers (matched papers that are both outliers) at the lower end of the distribution.
  - **Refs:** As I suspected, the outlier of 146 seems to be pulling the mean for SAGP up to 51.05 (the means for all other groups are around 44). This is corroborated by the skew statistics which show that only the SAGP group has a skew value above 1. I may run the analyses where references is the outcome variable both with and without this case in the model, or I may use word count as a covariate in the model to try to account for it.
  - **WC:** For some reason, there seems to be a notable difference in mean word count between the single-authored fraudulent group and the rest of the groups. I could speculate as to why this is the case, but I may do an unplanned follow-up analysis of this.

2. Frequencies will now be produced for categorical variables, both for the data in general and within `PaperType` groups.

- First, we will make the variables `inst_pres`, `gender`, `simple_reason`, `Country`, and `PaperType` factor variables.

```r
# Changing inst_pres (institutional prestige), gender, simple_reason, Country, and PaperTy
data$inst_pres <- factor(data$inst_pres, levels = c(0, 1), labels = c("Not Major Research
data$gender <- factor(data$gender, levels = c("FEMALE", "MALE"), labels = c("Female", "Mal
data$simple_reason <- factor(data$simple_reason, levels = c("f_data", "f_image", "m_image"
data$Country <- factor(data$Country)
data$PaperType <- factor(data$PaperType)
```

- Next, we will produce frequency tables for each categorical variable of interest for the whole dataset.

```r
# Creating the frequency tables for each categorical variable
freq_tab_inst_pres <- table(data$inst_pres)
freq_tab_gender <- table(data$gender)
freq_tab_simple_reason <- table(data$simple_reason)
freq_tab_Country <- table(data$Country)
freq_tab_PaperType <- table(data$PaperType)

# Displaying the frequency tables
freq_tab_inst_pres
```

```
Not Major Research Institution     Major Research Institution
                             70                             18
```

```r
freq_tab_gender
```

```
Female   Male
    15     73
```

```r
freq_tab_simple_reason
```

```
         Fabricated/Falsified Data          Fabricated/Falsified Image
                                27                                   3
                Manipulated Image Fabricated/Falsified Data and Image
                                12                                   2
```

```r
freq_tab_Country
```

```
     Australia        Belgium          China          Egypt       Ethiopia
             1              1             10              3              2
         India           Iran         Israel          Italy          Japan
            10              1              1              1              3
        Latvia       Malaysia    Netherlands       Pakistan         Poland
             1              2              6              3              2
      Portugal   South Africa    South Korea         Taiwan         Turkey
             2              1              2              1              2
United Kingdom  United States
             4             29
```

```
freq_tab_PaperType
```

```
MAFP MAGP SAFP SAGP
  22   22   22   22
```

- Let's take a look at the way these are split:
  - First, there seems to ba a preponderance of lower status (not major) research institutions in the whole dataset (70:18).
  - Next, most of the papers were retracted due to fabricated/falsified data ($n = 27$), followed by those that manipulated images ($n = 12$), fabricated/falsified images ($n = 3$), or both fabricated/falsified data and images ($n = 2$).
  - For country, the most common category of papers seem to be from the United States ($n = 29$), followed by China and India (both $n = 10$) and the Netherlands ($n = 6$), with no other countries totaling more than $n = 4$.
  - Finally, we can see that the papers are evenly split between each of our groups ($n = 22$; $N = 88$).
- Next, we will produce the frequencies within `PaperType` groups. Because we were not able to perfectly match across our matching characteristics, we will take this opportunity to look at how far off we were. A proportion table will be produced to more easily compare frequencies across `PaperType` groups. First, however, we need to make the matching dummy variables into factor variables with labels to make them more interpretable.

```
# Creating factor variables out of the dummy variables that track matching across groups
data$different_country <- factor(data$different_country, levels = c(0, 1), labels = c("Sam
data$different_gender <- factor(data$different_gender, levels = c(0, 1), labels = c("Same
data$different_inst_pres <- factor(data$different_inst_pres, levels = c(0, 1), labels = c(
data$different_journal <- factor(data$different_journal, levels = c(0, 1), labels = c("Sam
```

- I have checked the data set, and the labels seem to have been assigned correctly. Now we will produce the frequency tables and proportion tables.
  - Because we are taking a look at matching characteristics here, I will also take a look at the `year_difference` (the difference between the years of publication for the matched papers) variable within each group by using the `describe()` function from the `psych` package (like for the continuous variables above). In addition, I will take a look at the `NumAuth` (number of authors) for each of the groups.

```
# Frequency tables and proportion tables
  # institutional prestige
frequency_table_by_PaperType_inst_pres <- table(data$PaperType, data$inst_pres)
frequency_table_by_PaperType_inst_pres
```

```
     Not Major Research Institution Major Research Institution
MAFP                             17                           5
MAGP                             18                           4
SAFP                             18                           4
SAGP                             17                           5
```

```
proportion_table_by_PaperType_inst_pres <- prop.table(frequency_table_by_PaperType_inst_pr
proportion_table_by_PaperType_inst_pres
```

```
     Not Major Research Institution Major Research Institution
MAFP                       0.7727273                   0.2272727
MAGP                       0.8181818                   0.1818182
SAFP                       0.8181818                   0.1818182
SAGP                       0.7727273                   0.2272727
```

```
  # gender
frequency_table_by_PaperType_gender <- table(data$PaperType, data$gender)
frequency_table_by_PaperType_gender
```

```
     Female Male
MAFP      4   18
MAGP      4   18
SAFP      3   19
SAGP      4   18
```

```r
proportion_table_by_PaperType_gender <- prop.table(frequency_table_by_PaperType_gender, ma
proportion_table_by_PaperType_gender
```

```
        Female      Male
MAFP 0.1818182 0.8181818
MAGP 0.1818182 0.8181818
SAFP 0.1363636 0.8636364
SAGP 0.1818182 0.8181818
```

```r
  # reason
frequency_table_by_PaperType_reason <- table(data$PaperType, data$simple_reason)
frequency_table_by_PaperType_reason
```

```
     Fabricated/Falsified Data Fabricated/Falsified Image Manipulated Image
MAFP                        12                          1                 7
MAGP                         0                          0                 0
SAFP                        15                          2                 5
SAGP                         0                          0                 0

     Fabricated/Falsified Data and Image
MAFP                                  2
MAGP                                  0
SAFP                                  0
SAGP                                  0
```

```r
proportion_table_by_PaperType_reason <- prop.table(frequency_table_by_PaperType_reason, ma
proportion_table_by_PaperType_reason
```

```
     Fabricated/Falsified Data Fabricated/Falsified Image Manipulated Image
MAFP                0.54545455                 0.04545455        0.31818182
MAGP
SAFP                0.68181818                 0.09090909        0.22727273
SAGP

     Fabricated/Falsified Data and Image
MAFP                           0.09090909
MAGP
SAFP                           0.00000000
SAGP
```

```r
  # country
frequency_table_by_PaperType_country <- table(data$PaperType, data$Country)
frequency_table_by_PaperType_country
```

```
     Australia Belgium China Egypt Ethiopia India Iran Israel Italy Japan
MAFP         0       0     4     0        0     3    0      0     0     1
MAGP         0       0     4     0        0     3    0      0     0     1
SAFP         0       1     2     2        1     4    0      0     0     1
SAGP         1       0     0     1        1     0    1      1     1     0

     Latvia Malaysia Netherlands Pakistan Poland Portugal South Africa
MAFP      0        1           2        1      1        1            0
MAGP      0        1           1        1      1        1            0
SAFP      0        0           1        0      0        0            0
SAGP      1        0           2        1      0        0            1

     South Korea Taiwan Turkey United Kingdom United States
MAFP           0      0      0              0             8
MAGP           0      0      0              0             9
SAFP           1      0      1              1             7
SAGP           1      1      1              3             5
```

```r
proportion_table_by_PaperType_country <- prop.table(frequency_table_by_PaperType_country,
proportion_table_by_PaperType_country
```

```
      Australia    Belgium      China      Egypt   Ethiopia      India
MAFP 0.00000000 0.00000000 0.18181818 0.00000000 0.00000000 0.13636364
MAGP 0.00000000 0.00000000 0.18181818 0.00000000 0.00000000 0.13636364
SAFP 0.00000000 0.04545455 0.09090909 0.09090909 0.04545455 0.18181818
SAGP 0.04545455 0.00000000 0.00000000 0.04545455 0.04545455 0.00000000

           Iran     Israel      Italy      Japan     Latvia   Malaysia
MAFP 0.00000000 0.00000000 0.00000000 0.04545455 0.00000000 0.04545455
MAGP 0.00000000 0.00000000 0.00000000 0.04545455 0.00000000 0.04545455
SAFP 0.00000000 0.00000000 0.00000000 0.04545455 0.00000000 0.00000000
SAGP 0.04545455 0.04545455 0.04545455 0.00000000 0.04545455 0.00000000

     Netherlands   Pakistan     Poland   Portugal South Africa South Korea
MAFP  0.09090909 0.04545455 0.04545455 0.04545455   0.00000000  0.00000000
MAGP  0.04545455 0.04545455 0.04545455 0.04545455   0.00000000  0.00000000
```

```
SAFP  0.04545455 0.00000000 0.00000000 0.00000000   0.00000000  0.04545455
SAGP  0.09090909 0.04545455 0.00000000 0.00000000   0.04545455  0.04545455

         Taiwan     Turkey United Kingdom United States
MAFP 0.00000000 0.00000000     0.00000000    0.36363636
MAGP 0.00000000 0.00000000     0.00000000    0.40909091
SAFP 0.00000000 0.04545455     0.04545455    0.31818182
SAGP 0.04545455 0.04545455     0.13636364    0.22727273
```

```
# And for the matching characteristic dummy variables
  # Country different
frequency_table_by_PaperType_different_country <- table(data$PaperType, data$different_cou
frequency_table_by_PaperType_different_country
```

```
     Same Country Different Country
MAFP           14                 8
MAGP           21                 1
SAFP            0                 0
SAGP           10                12
```

```
proportion_table_by_PaperType_different_country <- prop.table(frequency_table_by_PaperType
proportion_table_by_PaperType_different_country
```

```
     Same Country Different Country
MAFP   0.63636364        0.36363636
MAGP   0.95454545        0.04545455
SAFP
SAGP   0.45454545        0.54545455
```

```
  # Gender different
frequency_table_by_PaperType_different_gender <- table(data$PaperType, data$different_gend
frequency_table_by_PaperType_different_gender
```

```
     Same Gender Different Gender
MAFP          18                4
MAGP          20                2
SAFP           0                0
SAGP          19                3
```

```r
proportion_table_by_PaperType_different_gender <- prop.table(frequency_table_by_PaperType_
proportion_table_by_PaperType_different_gender
```

```
     Same Gender Different Gender
MAFP  0.81818182       0.18181818
MAGP  0.90909091       0.09090909
SAFP
SAGP  0.86363636       0.13636364
```

```r
  # Institutional prestige different
frequency_table_by_PaperType_different_inst_pres <- table(data$PaperType, data$different_i
frequency_table_by_PaperType_different_inst_pres
```

```
     Same Institutional Prestige Different Institutional Prestige
MAFP                          17                                5
MAGP                          21                                1
SAFP                           0                                0
SAGP                          21                                1
```

```r
proportion_table_by_PaperType_different_inst_pres <- prop.table(frequency_table_by_PaperTy
proportion_table_by_PaperType_different_inst_pres
```

```
     Same Institutional Prestige Different Institutional Prestige
MAFP                  0.77272727                        0.22727273
MAGP                  0.95454545                        0.04545455
SAFP
SAGP                  0.95454545                        0.04545455
```

```r
  # Journal different
frequency_table_by_PaperType_different_journal <- table(data$PaperType, data$different_jou
frequency_table_by_PaperType_different_journal
```

```
     Same Journal Different Journal
MAFP           11                11
MAGP            0                 0
SAFP            0                 0
SAGP            0                 0
```

```r
proportion_table_by_PaperType_different_journal <- prop.table(frequency_table_by_PaperType
proportion_table_by_PaperType_different_journal
```

```
     Same Journal Different Journal
MAFP          0.5               0.5
MAGP
SAFP
SAGP
```

```r
# Descriptive statistics for year off within PaperType groups
year_difference_descriptives <- describeBy(data$year_difference, group = data$PaperType, n
```

```
Warning in min(x, na.rm = na.rm): no non-missing arguments to min; returning
Inf
```

```
Warning in max(x, na.rm = na.rm): no non-missing arguments to max; returning
-Inf
```

```r
year_difference_descriptives
```

```
 Descriptive statistics by group
group: MAFP
   vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 22 -1.36 6.15      0   -0.22 1.48 -21   8    29 -1.9     3.45 1.31
------------------------------------------------------------
group: MAGP
   vars  n  mean   sd median trimmed mad min max range  skew kurtosis   se
X1    1 22 -0.64 2.15      0   -0.22   0 -10   1    11 -3.72    13.37 0.46
------------------------------------------------------------
group: SAFP
   vars n mean sd median trimmed mad min  max range skew kurtosis se
X1    1 0  NaN NA     NA     NaN  NA Inf -Inf  -Inf   NA       NA NA
------------------------------------------------------------
group: SAGP
   vars  n mean   sd median trimmed mad min max range skew kurtosis   se
X1    1 22    0 1.35      0   -0.06   0  -4   4     8 0.11     4.36 0.29
```

```r
num_auth_descriptives <- describeBy(data$NumAuth, group = data$PaperType, na.rm = TRUE)
num_auth_descriptives
```

```
 Descriptive statistics by group
group: MAFP
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 22 3.73 1.32      4    3.61 1.48   2   7     5  0.6    -0.17 0.28
------------------------------------------------------------
group: MAGP
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 22 4.05 1.91      4    3.78 1.48   2   9     7 0.99     0.29 0.41
------------------------------------------------------------
group: SAFP
   vars  n mean sd median trimmed mad min max range skew kurtosis se
X1    1 22    1  0      1       1   0   1   1     0  NaN      NaN  0
------------------------------------------------------------
group: SAGP
   vars  n mean sd median trimmed mad min max range skew kurtosis se
X1    1 22    1  0      1       1   0   1   1     0  NaN      NaN  0
```

- **Overview of Categorical Variables:**
  - **Institutional Prestige:** The split for frequencies/proportion of is very close to even across groups. The MAFP and SAGP groups have 17 papers from major research institutions and 5 from lower status research institutions (.77/.22), and the MAGP and SAFP groups have 18 papers from major research institutions and 4 from lower status research institutions (.82/.18).
  - **Gender:** The split between groups is also very similar for gender, with MAFP, MAGP, and SAGP groups each having 4 female first authors and 18 male first authors (.18/.82). SAFP had 3 female first authors and 19 male first authors (.14/.86).
  - **Reason (Not Matched Between Groups During Data Collection):** The reason for retraction is only relevant for the fraudulent paper groups, and there is a similar split between them. The SAFP had 15 papers in the group for Fabricated/Falsified Data, 2 paper for Fabricated/Falsified Images, 5 for Manipulated Images, and 0 that Fabricated/Falsified Data and Images (.68/.09/.22/.00). The MAFP had 12 papers in the group for Fabricated/Falsified Data, 1 paper for Fabricated/Falsified Images, 7 for Manipulated Images, and 2 for Fabricated/Falsified Data and Images (.55/.05/.32/.09).
  - **Country:** Because there are many countries, I will not delineate them all here but rather refer you to the data printed above if you are interested in diving deeper.

To highlight one thing, we can see that papers from the United States (our most common country in the whole data set) tend to be overrepresented in the frequencies for each group (MAFP, 8/.36; MAGP, 9/.40; SAFP, 7/.32; SAGP, 5/.23). There are no other countries with proportions that exceed the lowest proportion of United States papers in any group (all<.23).

- **Overview of Matching Characteristic Dummy Variables and Average Differences:**
  - Although the frequencies and proportions of the matched variables within groups above indicates to some degree how similar each group is, the more important metric of how closely matched the groups are is between the groups for which comparisons are to be made. For this reason, I dummy coded whether each paper matched it's pair on each characteristic or not as a binary variable (which we labeled above) before loading the data. Because the SAFP papers were not gathered by matching to any group, there are no data for the SAFP group for any of the dummy variables, and journal only didn't match across some of the SAFP and MAFP pairs. Each of the other groups is in reference to the group against which it will be compared (i.e., data for MAFP represent how well they match SAFP papers, data for MAGP represent how well they match with MAFP, data for SAGP represent how well they match with SAFP). In addition, I coded how different the year was between papers, such that positive values indicate the number of years the original paper was published after the matched paper (negative values indicate the matched paper was published after the paper it was being matched to).
  - **MAFP compared with SAFP:**
    - Country was the same for 14 (.64) pairs and different for 8 (.36) pairs.
    - Gender was the same for 18 (.82) pairs and different for 4 (.18) pairs.
    - Institutional prestige was the same for 17 (.77) pairs and different for 5 (.23) pairs.
    - Journal was the same for 11 (.50) pairs and different for 11 (.50) pairs.
    - The average year difference was -1.36, indicating that the MAFP tended to be slightly more recent than the SAFP.
    - Finally, the mean number of authors for MAFP is 3.73.
  - **MAGP compared with MAFP:**
    - Country was the same for 21 (.95) pairs and different for 1 (.05) pairs.
    - Gender was the same for 20 (.91) pairs and different for 2 (.09) pairs.

    - Institutional prestige was the same for 21 pairs (.95) and different for 1 (.05) pair.
    - Journal was the same for every paper.
    - The average year difference was -.64, indicating that the MAGP tended to be slightly more recent than the MAFP.
    - Finally, the mean number of authors for MAFP is 4.05.
  - **SAGP compared with SAFP:**
    - Country was the same for 10 (.45) pairs and different for 12 (.55) pairs.
    - Gender was the same for 19 (.86) pairs and different for 3 (.14) pairs.
    - Institutional prestige was the same for 21 pairs (.95) and different for 1 (.05) pair.
    - Journal was the same for every paper.
    - The average year difference was 0, indicating that the SAGP and SAFP are exactly matched for the average year of publication.

- **Summary of matching characteristics:**
  - Looking at the data, the most difficult characteristic to match was country across all categories of papers, except for MAFP compared with SAFP, where journal was the most difficult to match. Nevertheless, for many of the situations where we were unable to match the specific paper, this was balanced out by not being able to match another paper in the opposite direction, as the frequencies within paper types demonstrates.

Data Visualization

1. Using the `ggplot2` package, frequency distributions and box plots will be generated for `LingObf`, `CertSent`, and `Refs`. A bar plot will be produced for `FraudCorrAuth`.

```
library(ggplot2) # To load the ggplot2 package
```

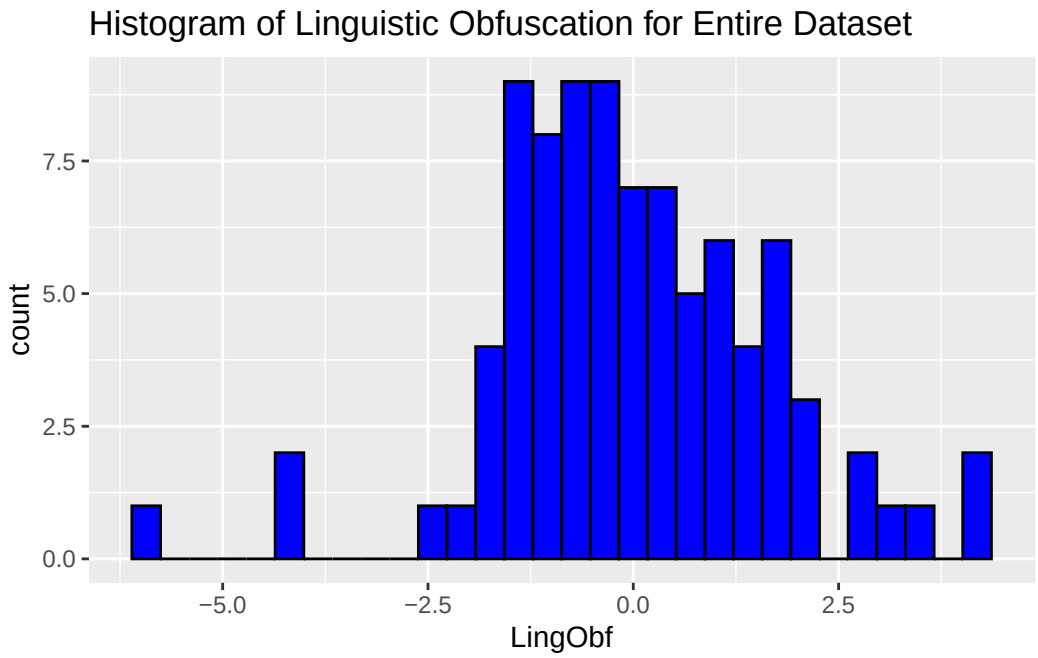
```
Attaching package: 'ggplot2'
```

```
The following objects are masked from 'package:psych':
```

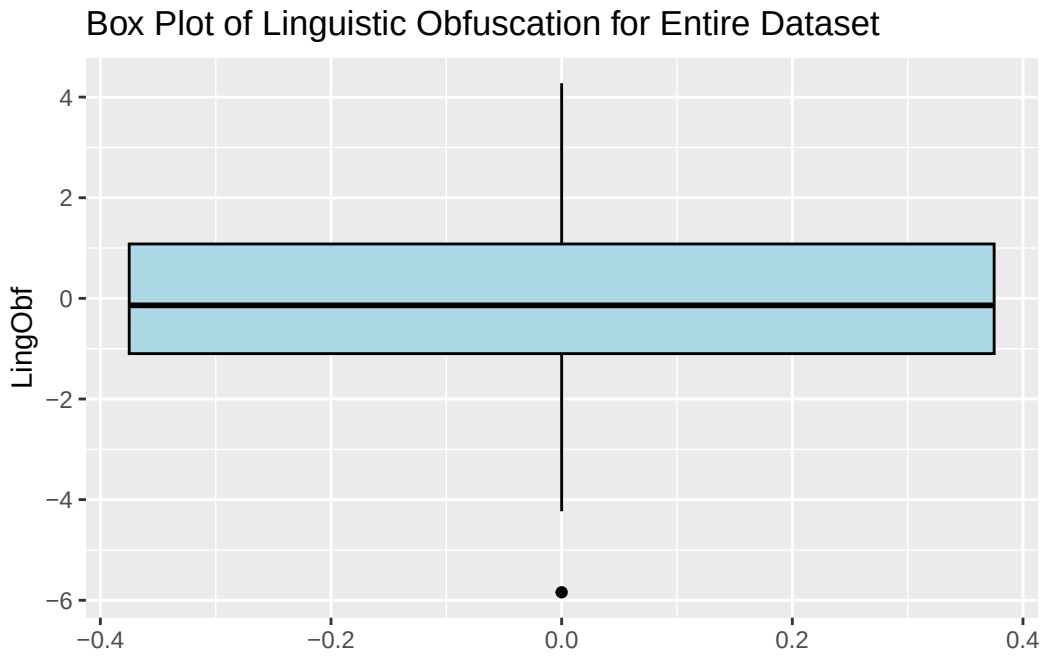
```
    %+%, alpha
```

```r
# To produce histogram and box plot for LingObf
ggplot(data, aes(x = LingObf)) +
  geom_histogram(fill = "blue", color = "black") +
  labs(title = "Histogram of Linguistic Obfuscation for Entire Dataset")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

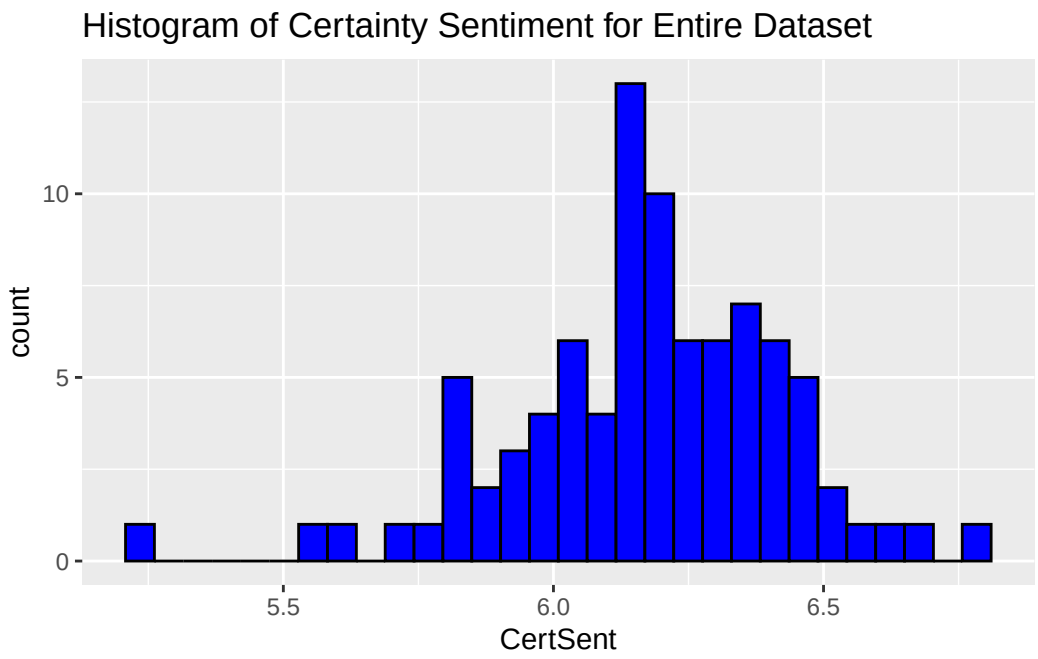



```r
ggplot(data, aes(y = LingObf)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Box Plot of Linguistic Obfuscation for Entire Dataset")
```

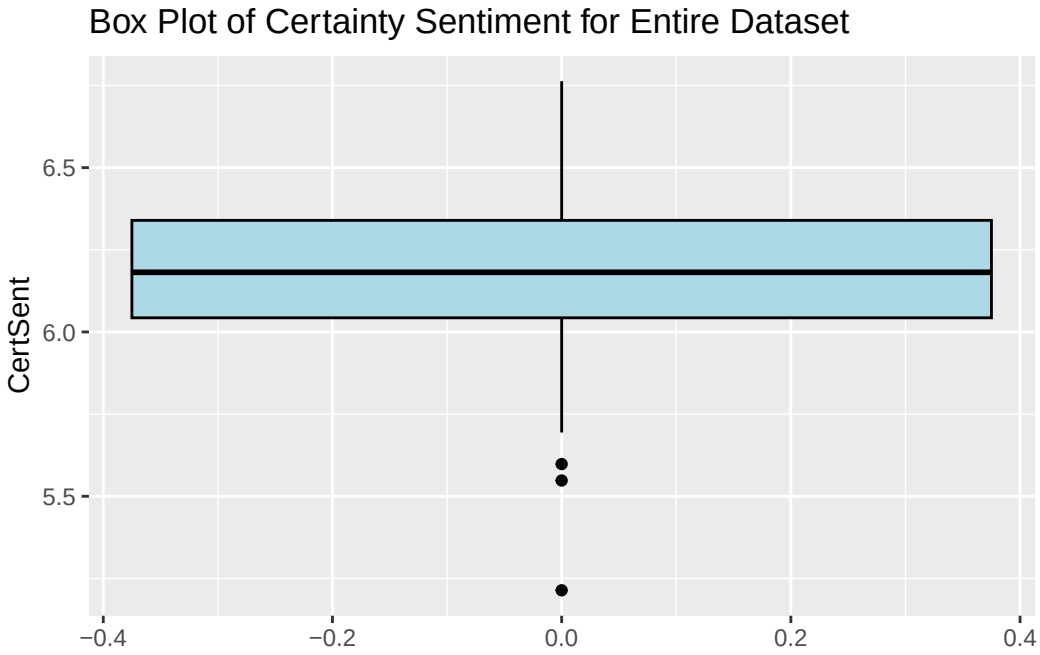



```
# To produce histogram and box plot for CertSent
ggplot(data, aes(x = CertSent)) +
  geom_histogram(fill = "blue", color = "black") +
  labs(title = "Histogram of Certainty Sentiment for Entire Dataset")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

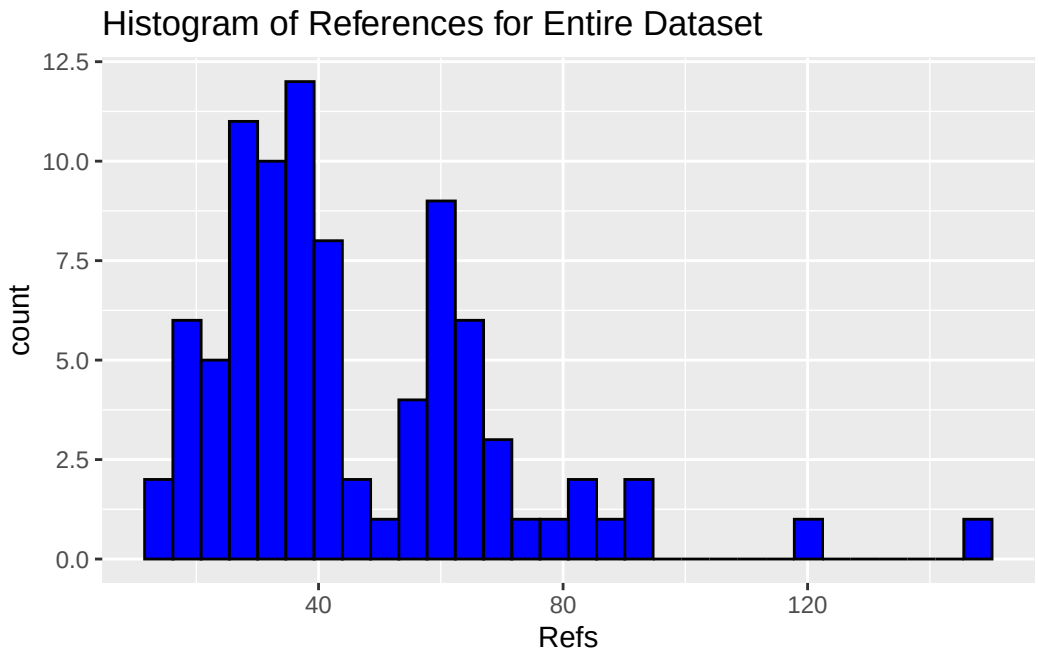



```
ggplot(data, aes(y = CertSent)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Box Plot of Certainty Sentiment for Entire Dataset")
```

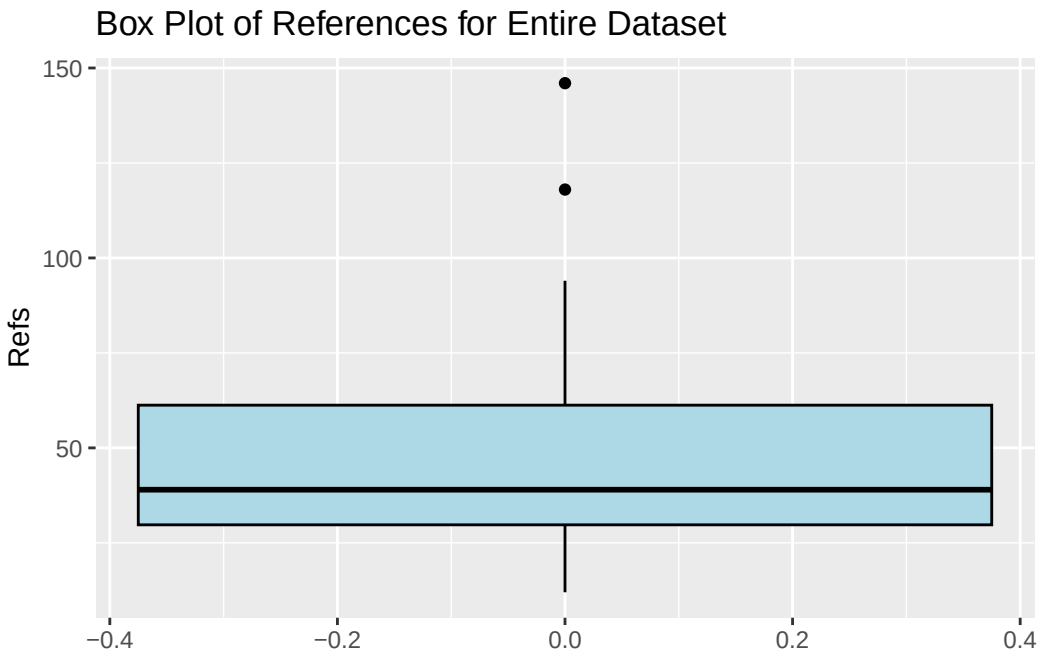



```
# To produce histogram and box plot for Refs
ggplot(data, aes(x = Refs)) +
  geom_histogram(fill = "blue", color = "black") +
  labs(title = "Histogram of References for Entire Dataset")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

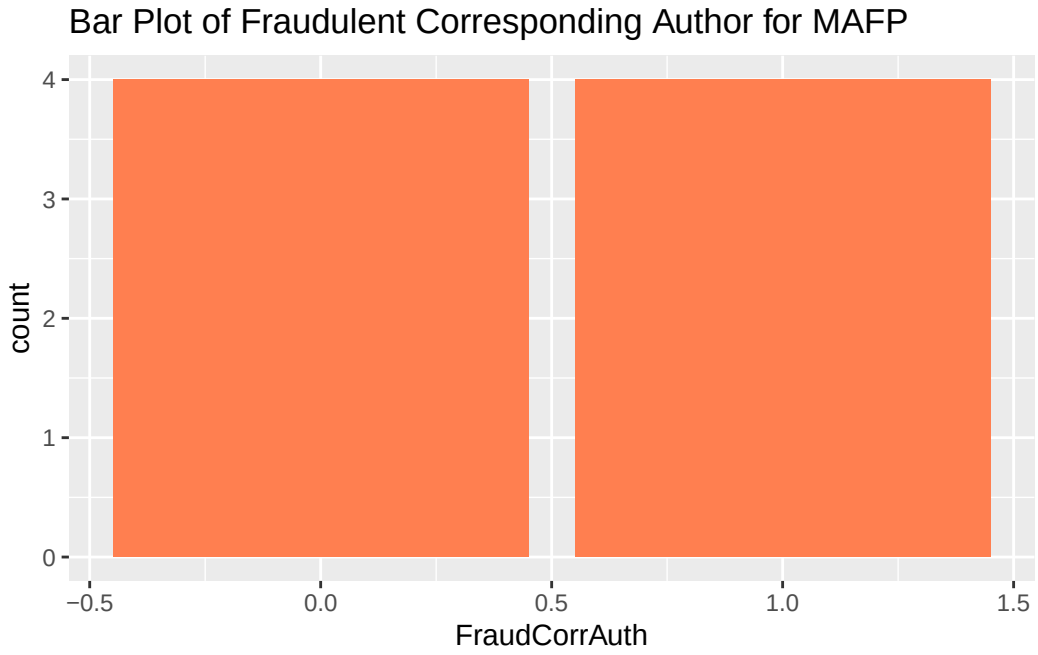



```r
ggplot(data, aes(y = Refs)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Box Plot of References for Entire Dataset")
```




```
# To produce bar plot for FraudCorrAuth
ggplot(data, aes(x = FraudCorrAuth)) +
  geom_bar(fill = "coral") +
  labs(title = "Bar Plot of Fraudulent Corresponding Author for MAFP")
```

```
Warning: Removed 80 rows containing non-finite outside the scale range
(`stat_count()`).
```



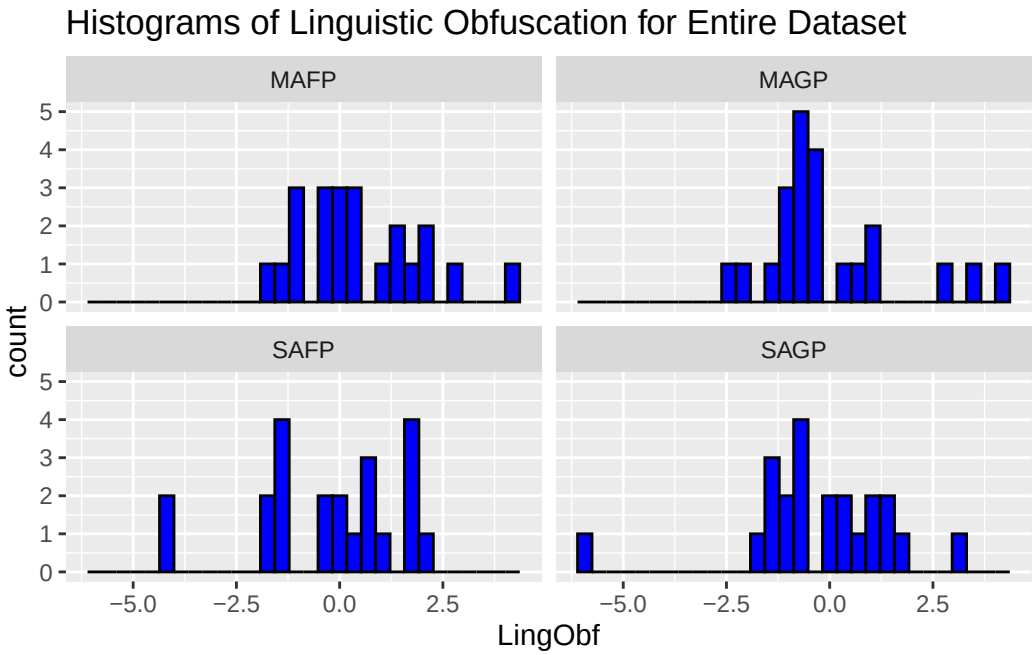
- **LingObf:** Linguistic obfuscation seems to have a bit of a wonky distribution with a few outliers at the low end of the distribution (M = 0, SD = 1.67; from descriptives above). This can be observed in the box plot as well, which shows one of the particularly low values.
- **CertSent:** Certainty sentiment does not seem to be skewed despite having an outlier at the low end of the distribution as well. The box plot highlights three values at the low end of the scale as being outliers.
- **Refs:** References has a strange distribution, one that looks like it is actually composed of two latent classes. It also has two outlier values way at the high end of the distribution, as shown by the histogram and the box plot.
- **FraudCorrAuth:** The fraudulent corresponding author variable shows an even split between papers for which the fraudulent author was and was not the corresponding author.

2. Using the `ggplot2` package, frequency distributions and box plots will be generated for `LingObf`, `CertSent`, and `Refs` within `PaperType` groups.

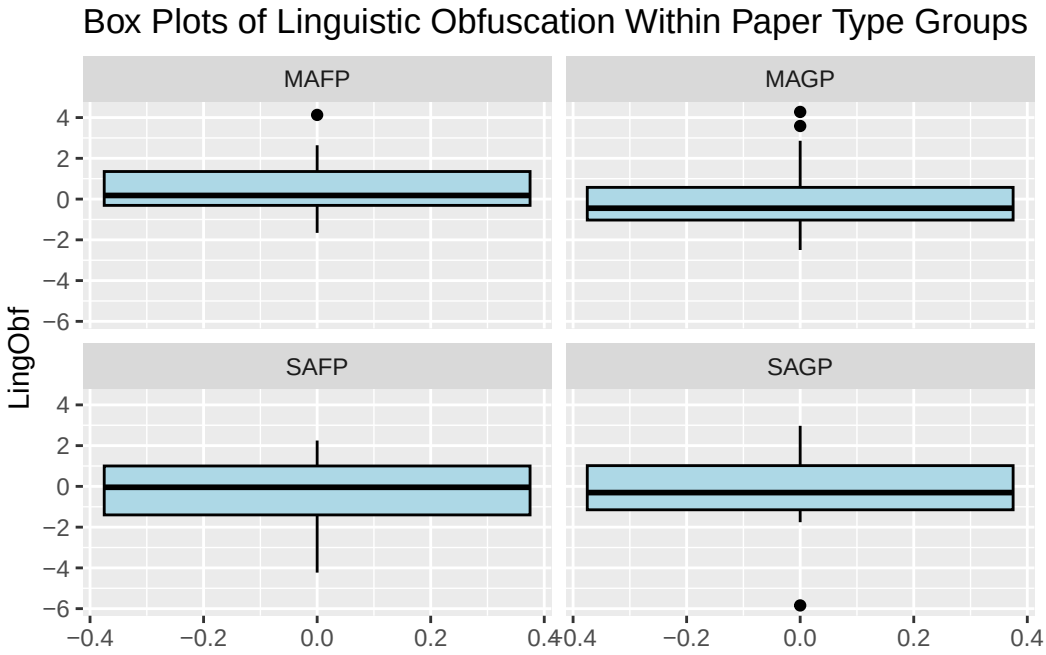
```
# To produce histogram and box plot for LingObf
ggplot(data, aes(x = LingObf)) +
  geom_histogram(fill = "blue", color = "black") +
  facet_wrap(~ PaperType) +
```

```
    labs(title = "Histograms of Linguistic Obfuscation for Entire Dataset")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

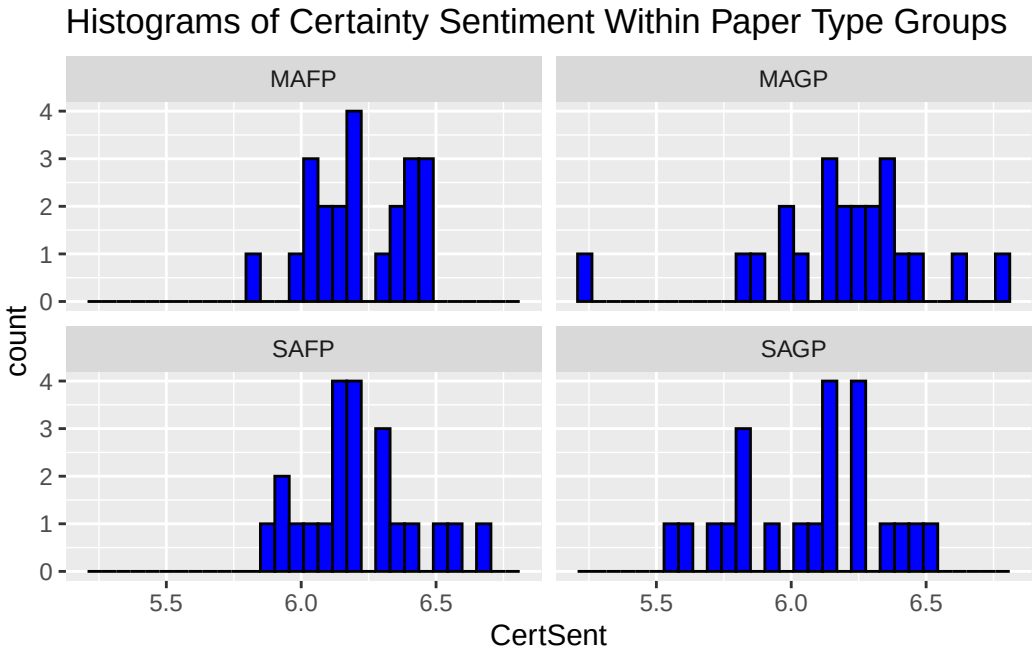



```
ggplot(data, aes(y = LingObf)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  facet_wrap(~ PaperType) +
  labs(title = "Box Plots of Linguistic Obfuscation Within Paper Type Groups")
```

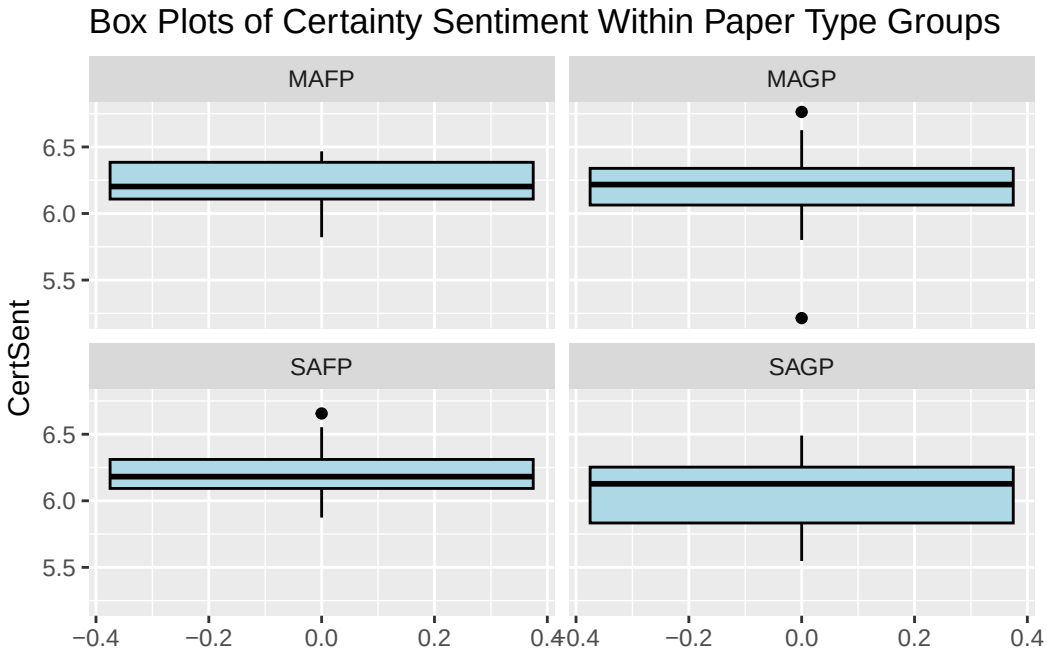



```r
# To produce histogram and box plot for CertSent
ggplot(data, aes(x = CertSent)) +
  geom_histogram(fill = "blue", color = "black") +
  facet_wrap(~ PaperType) +
  labs(title = "Histograms of Certainty Sentiment Within Paper Type Groups")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

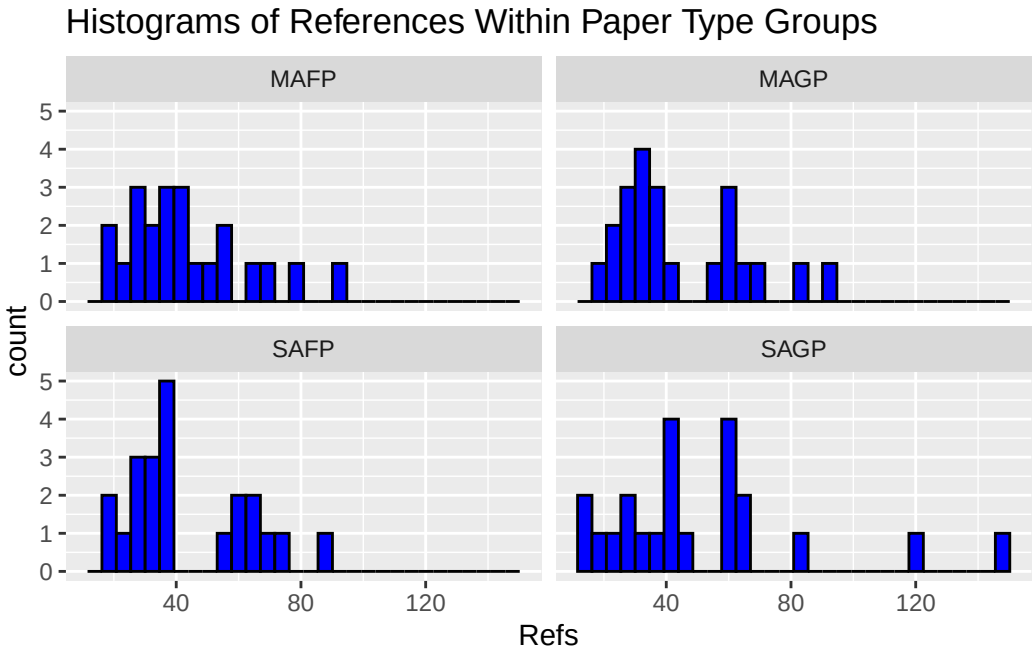



```
ggplot(data, aes(y = CertSent)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  facet_wrap(~ PaperType) +
  labs(title = "Box Plots of Certainty Sentiment Within Paper Type Groups")
```

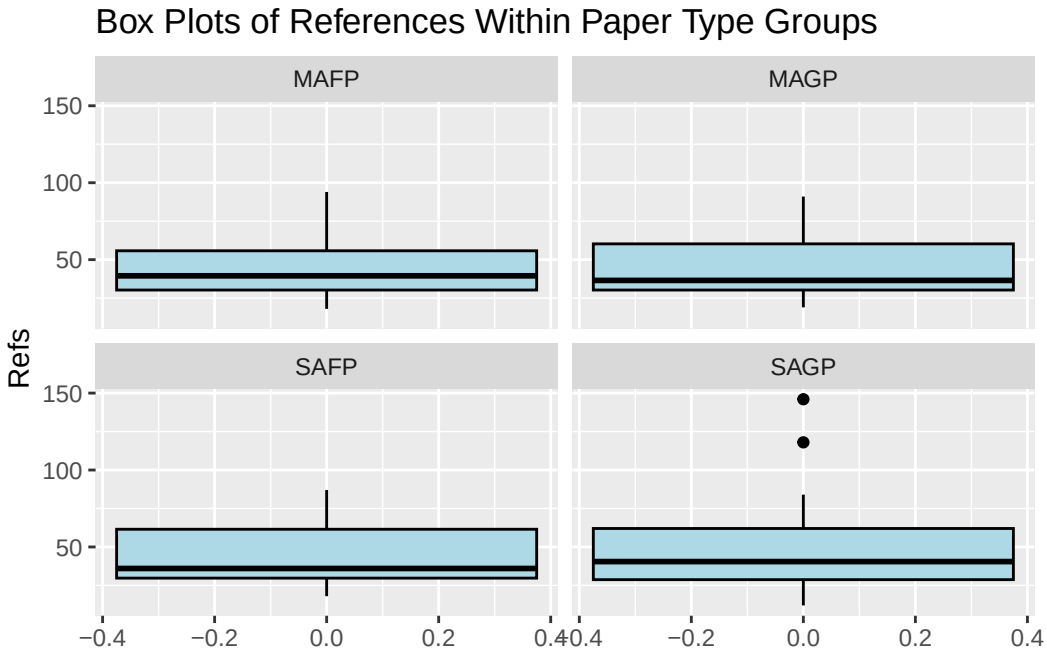



```r
# To produce histogram and box plot for Refs
ggplot(data, aes(x = Refs)) +
  geom_histogram(fill = "blue", color = "black") +
  facet_wrap(~ PaperType) +
  labs(title = "Histograms of References Within Paper Type Groups")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data, aes(y = Refs)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  facet_wrap(~ PaperType) +
  labs(title = "Box Plots of References Within Paper Type Groups")
```



- A few quick notes:
  - **LingObf:**
    - Positive outliers for linguistic obfuscation seem to be represented in the multi-author groups, and negative outliers seem to be in the single-author groups (although the lowest values do not show up on the SAFP box plot, perhaps because there are two of them that can be seen in the histogram which suggests that they may skew the spread statistics enough to envelop them within the whiskers).
  - **CertSent:**
    - The otulier at the low end of the certainty sentiment distribution seems to belong to the MAGP group.
  - **Refs:**
    - The positive outliers we have noted a couple times now for references seem to be in the SAGP group, corroborating our speculation above (see Descriptive Statistics).
    - Strangely, we noted that the references seemed to have two latent classes, but, at least visually, this does not seem to represent any particular paper group, as the latent classes show up in the SAFP and MAGP groups (as well as perhaps the SAGP group). I am not sure why this is.

- Overall, the strange values are nto all within one group, which suggests that they are due to random variation rather than one paper that was not read appropriately by the text analysis programs (or some other form of systematic problem with one or two observations in the data set).

Bivariate Correlations

In order to further explore data characteristics, bivariate correlations between outcome variables will be produced. As these are only for exploring data and not testing predictions, assumptions will not be tested, and the relationships will not be probed for statistical significance. Then, correlations for subcomponents of composite variables will be produced.

1. First, Pearson bivariate correlations will be produced for `LingObf`, `CertSent`, and `Refs`.

```
# To produce correlation matrix for LingObf, CertSent, and Refs variables within whole dat
  # To calculate the Pearson correlation coefficients between the numerical outcome variab
correlation_matrix_num_outcomes <- cor(data[c("LingObf", "CertSent", "Refs")], use = "comp
  # To display the correlation matrix for numerical outcome variables
correlation_matrix_num_outcomes
```

```
             LingObf    CertSent        Refs
LingObf   1.00000000  0.01747474 -0.08348489
CertSent  0.01747474  1.00000000 -0.13041183
Refs     -0.08348489 -0.13041183  1.00000000
```

- All of the variables here show negligible bivariate relationships.

2. Next, here are the point-biserial correlations between `FraudCorrAuth` and `LingObf`, `CertSent`, and `Refs`.

```
# Calculating Point-biserial correlations between FraudCorrAuth and numerical outcome vari
pbcorr_FraudCorrAuth_LingObf <- cor(data$FraudCorrAuth, data$LingObf, use = "complete.obs"
pbcorr_FraudCorrAuth_CertSent <- cor(data$FraudCorrAuth, data$CertSent, use = "complete.ob
pbcorr_FraudCorrAuth_Refs <- cor(data$FraudCorrAuth, data$Refs, use = "complete.obs", meth

# Displaying the correlation coefficients
pbcorr_FraudCorrAuth_LingObf
```

```
          [,1]
[1,] 0.5537262
```

```
pbcorr_FraudCorrAuth_CertSent
```

```
[1] 0.8266114
```

```
pbcorr_FraudCorrAuth_Refs
```

```
[1] -0.2157277
```

- While the relationship between FraudCorrAuth and LingObf is large (r = .55) and the relationship between FraudCorrAuth and CertSent is very large (r = .83), I will refrain from interpreting these because the sample is severely underpowered with a mere $n = 8$ observations.

3. Now, we will produce bivariate correlations for the subcomponents of the `abstraction` index (see Data Cleaning for more details.

```
# Calculate pairwise correlations among the subcomponents of the abstraction index
correlation_matrix_abstraction_subcomponents <- cor(data[,c("article", "prep", "quantity")

# Display the correlation matrix
correlation_matrix_abstraction_subcomponents
```

```
            article       prep   quantity
article  1.00000000 0.09553478 0.05573919
prep     0.09553478 1.00000000 0.02835620
quantity 0.05573919 0.02835620 1.00000000
```

- The correlations here (.03 < r < .09) are quite a bit smaller (i.e., nonexistent) than Markowitz & Hancock (2016) found for their sample when they constructed the abstraction index (.176 < r < .263). This calls into question the reliability of the abstraction index as a subcomponent of the linguistic obfuscation index in our sample, and it calls into question the validity of each of the indicators of the abstraction index (articles, prepositions, and quantity words) in our sample.

4. Now we will produce Pearson bivariate correlations for the subcomponents of the `LingObf` index.

```
# Calculate pairwise correlations among the subcomponents of LingObf
correlation_matrix_LingObf_subcomponents <- cor(data[,c("cause", "abstraction", "jargon",
```

```
# Display the correlation matrix
correlation_matrix_LingObf_subcomponents
```

```
                cause abstraction     jargon     emo_pos  flesch_re
cause        1.0000000 -0.19392253 -0.2100682  0.25003236 -0.3237214
abstraction -0.1939225  1.00000000 -0.2967321 -0.05026066  0.3604363
jargon      -0.2100682 -0.29673215  1.0000000 -0.21731289  0.1764547
emo_pos      0.2500324 -0.05026066 -0.2173129  1.00000000 -0.2037766
flesch_re   -0.3237214  0.36043631  0.1764547 -0.20377661  1.0000000
```

- If we let the Linguistic Obfuscation Index be *LFI*, causal terms be C, the abstraction index be A, jargon be J, positive emotion terms be P, and Flesch reading ease be F, the mathematical model for the construction of the Linguistic Obfuscation index for an observation i can be stated as follows:

$$LFI_i = (C_i + A_i + J_i) - (P_i + F_i)$$

- If this model is internally consistent, we should expect that each of the variables within the aggregated variables (i.e., that are within parentheses with one another) should be positively correlate. We should also expect that variables will be negatively correlated across the aggregates (i.e., each of the variables that are in opposite aggregates).
  - Expected Positive Correlations:
    - We should expect the relationships between causal terms, the abstraction index, and jargon words to be positive. However, the relationships all seem to be (small) negative relationships.
    - We also should expect the relationship between positive emotion words and Flesch reading ease to be positive, but it is also a small, negative correlation.
  - Expected Negative Correlations:
    - As we should expect, there is a moderate negative relationship between causal terms and Flesch reading ease.
    - However, we should expect the relationship between causal terms and positive emotion terms to be negative, but it is a small, positive correlation.
    - We should expect a negative relationship between the abstraction index and positive emotion terms, but there is no relationship to speak of.
    - We also should expect a negative relationship between the abstraction index and Flesch reading ease, but there is a moderate positive correlation.

    * As should expect, there is a small negative relationship between Jargon and positive emotion terms.
    * Lastly, we should expect a negative correlation between jargon and Flesch reading ease, but there is a small, positive correlation between them.
- The observed relationships here may be expected *a priori* in some cases (e.g., the relationship between jargon and Flesch reading ease may be expected to be a positive relationship because the longer sentences someone writes the more likely they may be to use nonstandard words, simply by volume). However, from an internal consistency perspective, most of the relationships (8/10) we have observed here are either in the wrong direction or are too small to be considered.
    * This warrants serious consideration for whether our findings for hypothesis A can be interpreted appropriately.

Testing Hypotheses

Hypothesis 1: Linguistic Obfuscation Hypothesis

Hypothesis 1 consists of two variants. The first, **Hypothesis 1a**, is the general linguistic obfuscation hypothesis tested by Markowitz & Hancock (2016) which we will test in order to replicate their previous findings. The second, **Hypothesis 1b**, consists of our specific variant of the linguistic obfuscation hypothesis that states fraudulent scientists obfuscate their writing to make their work less accessible to their research group, rather than those outside their research group. This leads to the prediction that SAFP—which presumably are written without the presence of a research group—will use less linguistic obfuscation. Specifically, these hypotheses are stated follows:

Hypothesis 1a: Fraudulent research will be written with more linguistic obfuscation than non-fraudulent research. [Replication]

Hypothesis 1b: Single-author fraudulent research will be written with more linguistic obfuscation than non-fraudulent research but less linguistic obfuscation than multi-author fraudulent research. [Novel]

Testing Hypothesis 1a

To test **Hypothesis 1a**, we will conduct an independent samples t-test comparing the mean levels of `LingObf` of fraudulent papers (SAFP and MAFP combined) with genuine papers (SAGP and MAGP combined). That is, we will see whether the mean linguistic obfuscation differs between fraudulent papers and genuine papers.

1. To compare fraudulent publications with genuine publications, we will first add a new dichotomous grouping variable for genuine or fraudulent papers (`Genuine_or_Fraudulent`). That is, the new variable will combine SAGP and MAGP into one category (`GPaper`) and SAFP and MAFP into another category (`FPaper`). This will be done using the `dplyr` package. Then we will make the variable a factor variable.

```
data$Genuine_or_Fraudulent <- ifelse(data$PaperType %in% c("SAFP", "MAFP"), "FPaper", "GPa
```

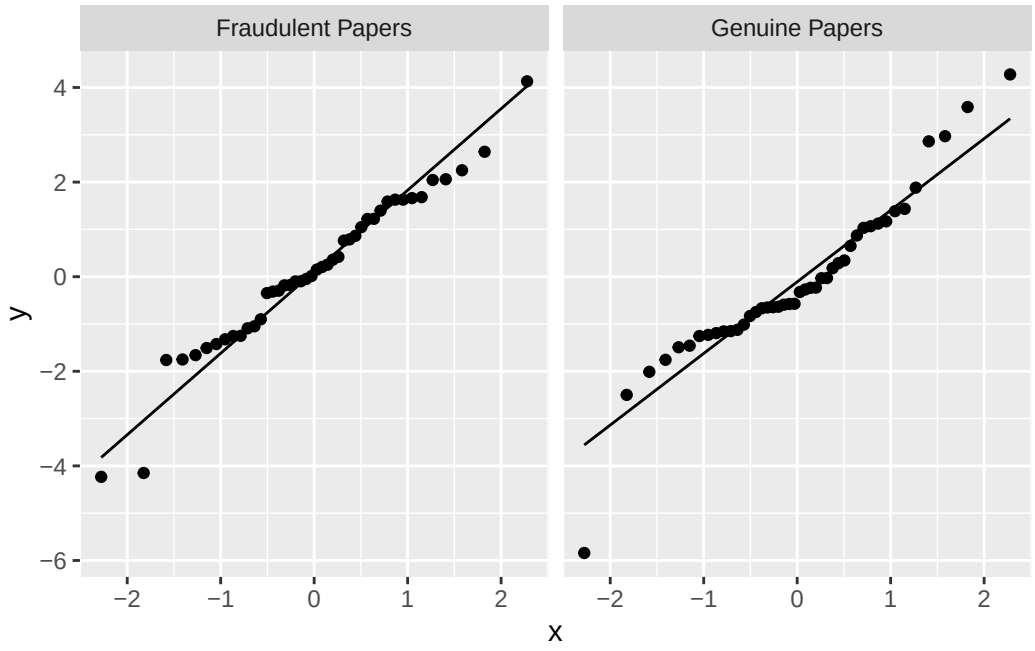
- I have checked the data frame, and this transformation seems to have been done correctly.
- Now to make it a factor variable.

```
data$Genuine_or_Fraudulent <- factor(data$Genuine_or_Fraudulent, levels = c("FPaper", "GPa
```

- This transformation also seems to have been done correctly.

2. To check the assumption of normality we will use the `ggplot2` package to create Q-Q plots for `LingObf` within the `FPaper`and `GPaper` groups. The Q-Q plots will be investigated visually. Because each group will has $n = 44$ cases, our t-test will not necessarily be rhobust to violations of this assumption. If the assumption of normality is violated and we cannot fix it, then we will use a bootstrapped t-test to fit a more robust model.

```
# To produce Q-Q plots for LingObf within the FPaper and GPaper groups
ggplot(data, aes(sample = LingObf)) +
  stat_qq() +
  facet_wrap(~ Genuine_or_Fraudulent) +
  geom_qq_line()
```



- The QQ plots are relatively close to the theoretical values, but there is some deviation at the ends of the distributions (particularly in the genuine papers group with a huge outlier). Because of this, I will run a Shapiro-Wilk test on each of the groups.

```
# Performing the Shapiro-Wilks test for normality within each group
sw_results_LingObf_GorP <- data %>%
  group_by(Genuine_or_Fraudulent) %>%
  summarise(SW_test_result = list(shapiro.test(LingObf)))

# Displaying the results
sw_results_LingObf_GorP$SW_test_result[[1]]
```

```
    Shapiro-Wilk normality test

data:  LingObf
W = 0.9676, p-value = 0.2484
```

```
sw_results_LingObf_GorP$SW_test_result[[2]]
```

```
    Shapiro-Wilk normality test
```

```
data:  LingObf
W = 0.92897, p-value = 0.009613
```

- The Shapiro-Wilk tests indicate that for the Fraudulent Paper group we should accept the null hypothesis that the data are normally distributed, and for the Genuine Paper group we should reject the null hypothesis that the data are normally distributed. I am now going to try to remove the outlier (lowest value for `LingObf`) and rerun these tests to see if that takes care of the issue. I will create a new data frame (`data_no_gpaper_obf_outlier`) without the outlier so that I can preserve the original data for later analyses.

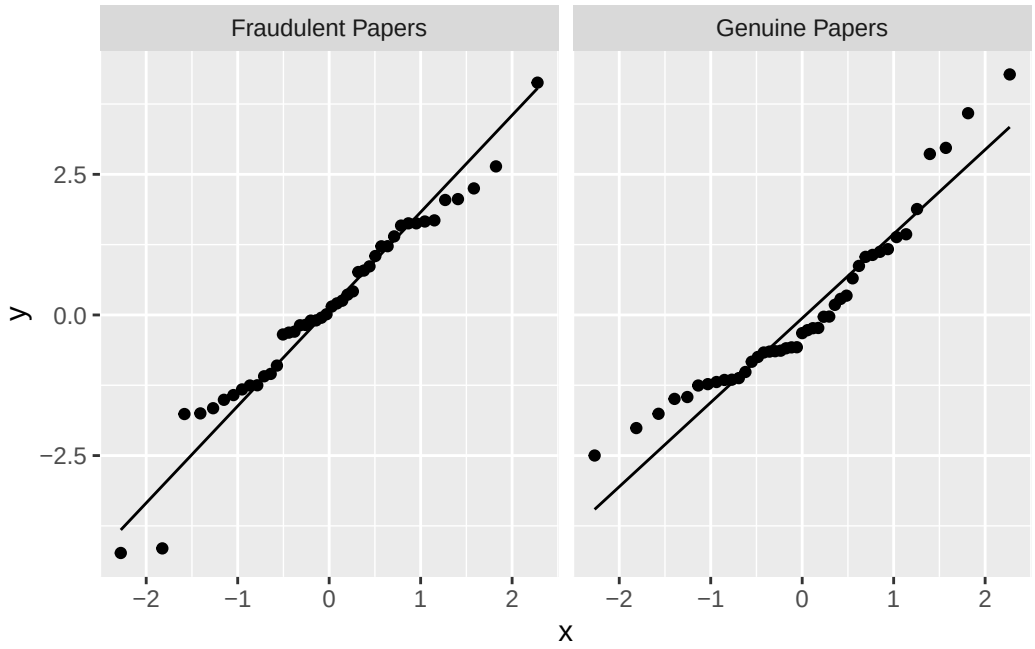
```
min_value_obf <- min(data$LingObf, na.rm = TRUE)

# Remove the row with the minimum LingObf value from the entire dataset
data_no_gpaper_obf_outlier <- data %>%
  filter(LingObf != min_value_obf)
```

```
Warning: Using one column matrices in `filter()` was deprecated in dplyr 1.1.0.
i Please use one dimensional logical vectors instead.
```

- Now that we have removed the lowest value for GPaper, we will produce the QQ-plots and the Shapiro-Wilk normality tests again.

```
# To produce Q-Q plot for LingObf within the FPaper and GPaper groups in the new data fram
ggplot(data_no_gpaper_obf_outlier, aes(sample = LingObf)) +
  stat_qq() +
  facet_wrap(~ Genuine_or_Fraudulent) +
  geom_qq_line()
```




```r
# Performing the Shapiro-Wilks test for normality within each group in the new data frame
sw_results_LingObf_GorP_2 <- data_no_gpaper_obf_outlier %>%
  group_by(Genuine_or_Fraudulent) %>%
  summarise(SW_test_result2 = list(shapiro.test(LingObf)))

# Displaying the results
sw_results_LingObf_GorP_2$SW_test_result2[[1]]
```

```
    Shapiro-Wilk normality test

data:  LingObf
W = 0.9676, p-value = 0.2484
```

```r
sw_results_LingObf_GorP_2$SW_test_result2[[2]]
```

```
    Shapiro-Wilk normality test

data:  LingObf
W = 0.92355, p-value = 0.007035
```

- Unfortunately, the result of the Shapiro-Wilk normality test indicates that removal of the outlier did not improve the situation. Instead, I will run the t-test with bootstrapped confidence intervals to attempt to make it robust to violations of the assumption of normality. Because bootstrapped t-tests are also robust to violations of the assumption of homogeneity of variance, we will skip this step.

3. Now we will run the independent samples t-test with bootstrapped confidence intervals.

```
# Load necessary libraries
library(boot)
```

```
Attaching package: 'boot'
```

```
The following object is masked from 'package:psych':

    logit
```

```
library(effsize)
```

```
Attaching package: 'effsize'
```

```
The following object is masked from 'package:psych':

    cohen.d
```

```
# Setting the seed
set.seed(616913)

# Defining the statistic function for bootstrapping
t_stat_boot_obf <- function(data, index) {
  resampled_data_obf <- data[index, ] # Resample the data with replacement
  t_result <- t.test(LingObf ~ Genuine_or_Fraudulent, data=resampled_data_obf)
  return(t_result$statistic)
}

# Performing the bootstrapping
boot_results_obf <- boot(data = data, statistic = t_stat_boot_obf, R = 1000)

# Performing the regular t-test
```

```r
stud_t_test_obf <- t.test(LingObf ~ Genuine_or_Fraudulent, data = data)

# Calculating Cohen's d using the effsize package
d_obf <- cohen.d(LingObf ~ Genuine_or_Fraudulent, data = data)

# Calculate the 95% confidence interval from the bootstrapped results
boot_ci_obf <- boot.ci(boot_results_obf, type = "bca")

# Report the results
cat("Mean Difference: ", stud_t_test_obf$estimate, "\n",
    "95% CI: ", boot_ci_obf$bca[4], boot_ci_obf$bca[5], "\n",
    "t-value: ", stud_t_test_obf$statistic, "\n",
    "p-value: ", stud_t_test_obf$p.value, "\n",
    "Cohen's d: ", d_obf$estimate, "\n")
```

```
Mean Difference:  0.1159559 -0.1159559
 95% CI:  -1.611907 2.615176
 t-value:  0.6479212
 p-value:  0.5187658
 Cohen's d:  0.1381373
```

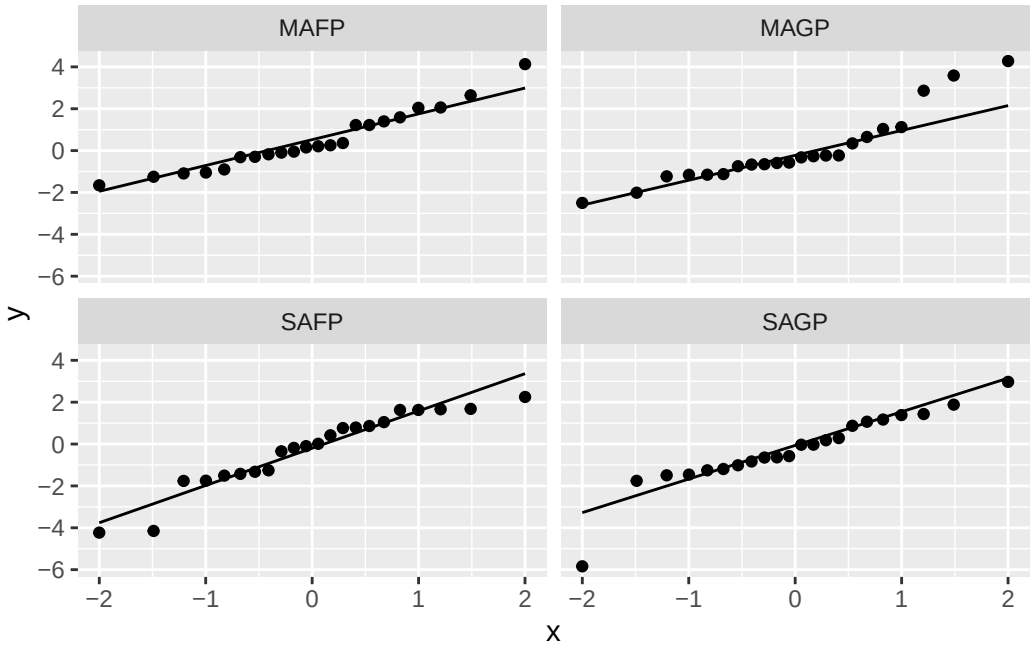
- The results of this bootstrapped t-test with 1,000 resamples does not support **Hypothesis 1a**. We should accept the null hypothesis that the means of the two groups are the same.

Testing Hypothesis 1b

To test **Hypothesis 1b** we will conduct a one-way analysis of variance (ANOVA) to compare group means of `LingObf` for SAFP, MAFP, SAGP, and MAGP. That is, we will determine whether SAFP contains more linguistic obfuscation than the genuine paper groups but less linguistic obfuscation than the MAFP.

1. To check the assumption of normality we will use the `ggplot2` package to create Q-Q plots for `LingObf` using our original study dataset. The Q-Q plots will be investigated visually. If the assumption of normality is violated we will use a bootstrapped t-test to fit a more robust model.

```r
# To produce Q-Q plots for LingObf within the PaperType groups
ggplot(data, aes(sample = LingObf)) +
  stat_qq() +
  facet_wrap(~ PaperType) +
  geom_qq_line()
```



- I am not satisfied with the plots within PaperType groups for SAFP, MAGP, or SAGP. Therefore, we will test for non-normality using a Shapiro-Wilk test within each group.

```
# Performing the Shapiro-Wilks test for normality within each group
sw_results_LingObf_PT <- data %>%
  group_by(PaperType) %>%
  summarise(SW_test_result2 = list(shapiro.test(LingObf)))

# Displaying the results
sw_results_LingObf_PT$SW_test_result2[[1]]
```

```
    Shapiro-Wilk normality test

data:  LingObf
W = 0.94773, p-value = 0.2845
```

```
sw_results_LingObf_PT$SW_test_result2[[2]]
```

```
    Shapiro-Wilk normality test

data:  LingObf
W = 0.87454, p-value = 0.009519
```

```
sw_results_LingObf_PT$SW_test_result2[[3]]
```

```
    Shapiro-Wilk normality test

data:  LingObf
W = 0.91969, p-value = 0.0749
```

```
sw_results_LingObf_PT$SW_test_result2[[4]]
```

```
    Shapiro-Wilk normality test

data:  LingObf
W = 0.9018, p-value = 0.0323
```

- As I suspected, the results for the Shapiro-Wilk test for normality indicate that for the SAFP, MAGP, or SAGP we should reject the null hypothesis that the data are normally distributed. Therefore, we will run a bootstrapped ANOVA to test for differences between groups, and we will skip the test for assuming homogeneity of variances again as well.

2. We will now run the one-way ANOVA with bootstrapped confidence intervals.

```
# Setting the seed
set.seed(616913)

# Define the statistic function for bootstrapping ANOVA
f_stat_boot_obf <- function(data, index) {
  resampled_data <- data[index, ] # Resample the data with replacement
  aov_result <- aov(LingObf ~ PaperType, data=resampled_data)
  return(summary(aov_result)[[1]]$F[1]) # Extract the F-statistic
}

# Perform the bootstrapping
boot_results_f_obf <- boot(data = data, statistic = f_stat_boot_obf, R = 1000)

# Perform and print the original ANOVA to get the original F-statistic and p-value
original_aov_obf <- aov(LingObf ~ PaperType, data=data)
summary(original_aov_obf)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
PaperType    3   7.57   2.523   0.898  0.446
Residuals   84 236.01   2.810
```

```r
# Calculate and report the bootstrapped confidence interval for the F-statistic
boot_f_ci_obf <- boot.ci(boot_results_f_obf, type="bca")
```

```
Warning in norm.inter(t, adj.alpha): extreme order statistics used as endpoints
```

```r
cat("Bootstrapped 95% Confidence Interval for F-statistic: ", boot_f_ci_obf$bca[4], " to "
```

```
Bootstrapped 95% Confidence Interval for F-statistic:  0.02164191  to  2.471062
```

- The results of this bootstrapped one-way ANOVA indicate that we should accept the null hypothesis that there are no differences between the `PaperType` groups on the `LingObf` variable. No post-hoc test needs to be conducted.

Hypothesis 2: References Hypothesis

Hypothesis 2 also consists of two variants. The first is that fraudulent research will contain more references than non-fraudulent research, functioning to make the research more costly to assess from outside readers (Markowitz & Hancock, 2016) or as an analogue to third-person pronoun usage in other linguistic studies of deception (Schmidt, 2022). The second is our adaptation of the first version which emphasizes the salience of the research group as the audience of this communicative style, similar to the logic of Hypothesis 1b. Specifically, the hypotheses are as follows:

Hypothesis 2a: Fraudulent research will contain more references than non-fraudulent research. [Replication]

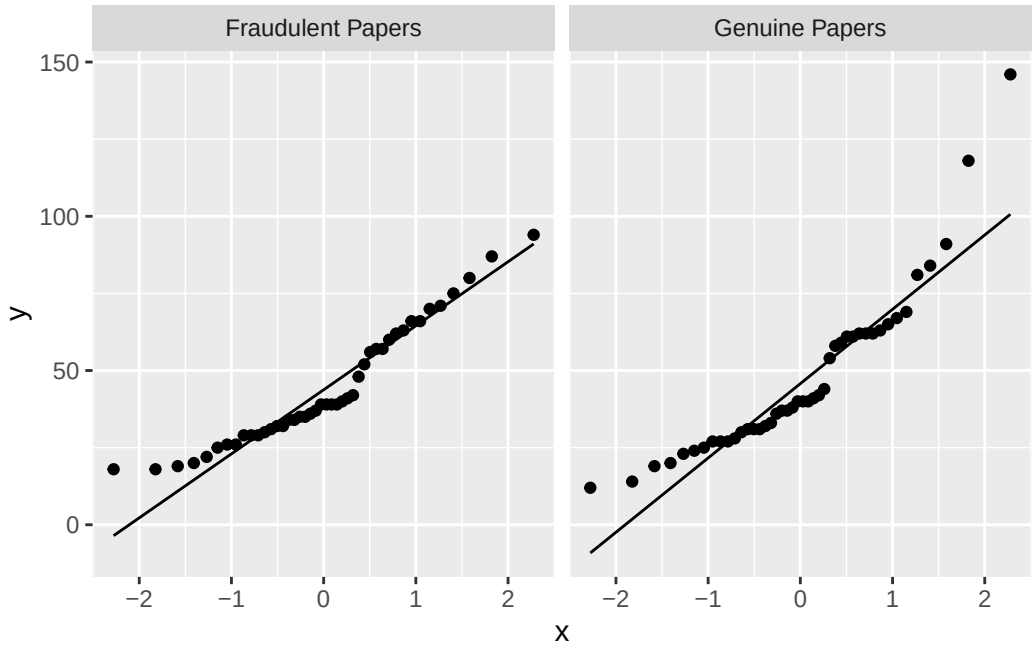
Hypothesis 2b: Single-author fraudulent research will contain more references than non-fraudulent research but fewer references than multi-author fraudulent research. [Novel]

Testing Hypothesis 2a

To test **Hypothesis 2a**, we will compare the mean number of `Refs` between the `FPaper` and `GPaper` groups using an independent-samples t-test.

1. To check the assumption of normality we will use the `ggplot2` package to create Q-Q plots for `Refs` within the `Genuine_or_Fraudulent` dichotomous variable. The Q-Q plots will be investigated visually. As references are technically count, rather than continuous, data, there is a higher likelihood that it will violate the assumption of normality. If this is the case we will instead run a Mann-Whitney U test to compare the two groups by their median `Refs`.

```
# To produce Q-Q plots for Refs within the FPaper and GPaper groups
ggplot(data, aes(sample = Refs)) +
  stat_qq() +
  facet_wrap(~ Genuine_or_Fraudulent) +
  geom_qq_line()
```



- The QQ-plots do not look promising, especially for genuine papers, which we noted likely contained outlier values before. We will test normality more formally with a Shapiro-Wilks test.

```
# Performing the Shapiro-Wilks test for normality within each group
sw_results_Refs_GorP <- data %>%
  group_by(Genuine_or_Fraudulent) %>%
  summarise(SW_test_result3 = list(shapiro.test(Refs)))
```

```r
# Displaying the results
sw_results_Refs_GorP$SW_test_result3[[1]]
```

```
    Shapiro-Wilk normality test

data:  Refs
W = 0.92403, p-value = 0.006505
```

```r
sw_results_Refs_GorP$SW_test_result3[[2]]
```

```
    Shapiro-Wilk normality test

data:  Refs
W = 0.86705, p-value = 0.0001223
```

- For both groups, the Shapiro-Wilks test for normality indicates that we should reject the null hypothesis that the data is normally distributed. Therefore, we will move right ahead with a Mann-Whitney U test, skipping Levene's test for homogeneity of variances.

2. Conducting the Mann-Whitney U test.

```r
# To conduct Mann-Whitney U test
hyp_two_a_mann_test <- wilcox.test(Refs ~ Genuine_or_Fraudulent, data = data)
```

```
Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
compute exact p-value with ties
```

```r
# Displaying the results
hyp_two_a_mann_test
```

```
    Wilcoxon rank sum test with continuity correction

data:  Refs by Genuine_or_Fraudulent
W = 934.5, p-value = 0.7829
alternative hypothesis: true location shift is not equal to 0
```

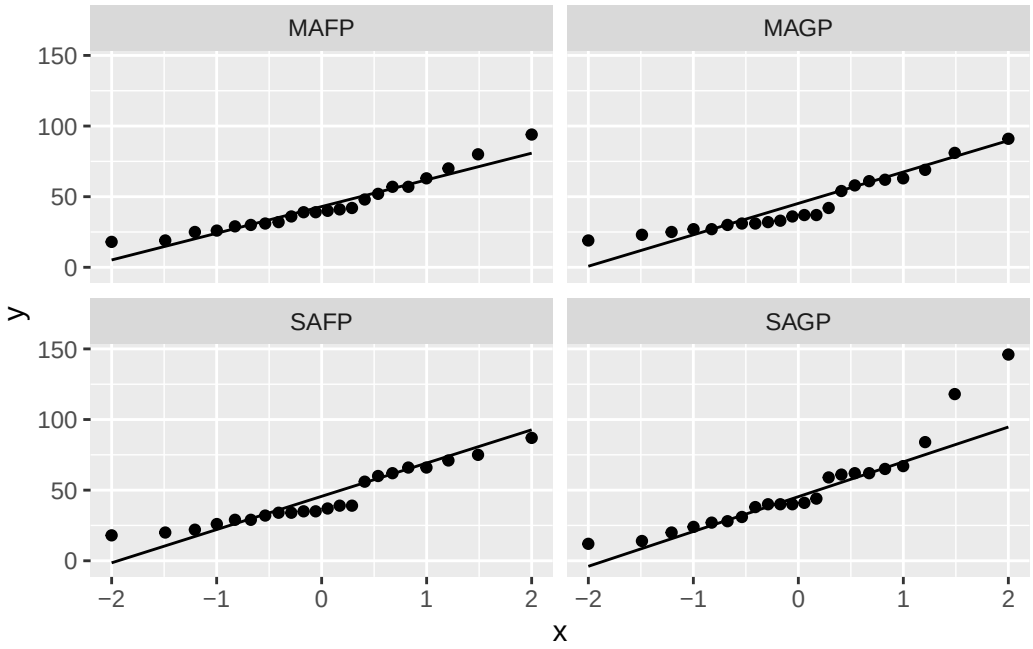
- The results of the Mann-Whitney U test indicate that we should accept the null hypothesis that the medians of the two groups are different. Although we mentioned in the data exploration section that we might want to remove the outlier before testing hypotheses with references as the outcome variable, the Mann-Whitney U test does not use means and should therefore not be biased due to outliers leverage on the statistics used. Therefore, it is not necessary to test this with the outlier removed.

Testing Hypothesis 2b

To test **Hypothesis 2b** we will conduct a one-way analysis of variance (ANOVA) to compare group means of `Refs` for SAFP, MAFP, SAGP, and MAGP. That is, we will determine whether SAFP contains more linguistic obfuscation than the genuine paper groups but less linguistic obfuscation than the MAFP. As references are technically count, rather than continuous, data, there is a higher likelihood that it will violate the assumption of normality, so in this case we will instead run a Kruskal-Wallis test to compare the groups by their median `Refs`.

1. To check the assumption of normality we will use the `ggplot2` package to create Q-Q plots for `Refs`. The Q-Q plots will be investigated visually. If the assumption of normality is violated we will use the non-parametric Kruskal-Wallis test instead.

```
# To produce Q-Q plots for Refs within the PaperType groups
ggplot(data, aes(sample = Refs)) +
  stat_qq() +
  facet_wrap(~ PaperType) +
  geom_qq_line()
```



- Again, the data do not look like they are normally distributed based on the QQ-plots, especially for SAGP. To test this formally, we conduct Shapiro-Wilks tests for normality within groups.

```
# Performing the Shapiro-Wilks test for normality within each group
sw_results_Refs_PT <- data %>%
  group_by(PaperType) %>%
  summarise(SW_test_result4 = list(shapiro.test(Refs)))

# Displaying the results
sw_results_Refs_PT$SW_test_result4[[1]]
```

```
    Shapiro-Wilk normality test

data:  Refs
W = 0.92665, p-value = 0.1045
```

```
sw_results_Refs_PT$SW_test_result4[[2]]
```

```
    Shapiro-Wilk normality test
```

```
data:  Refs
W = 0.8939, p-value = 0.02249
```

```
sw_results_Refs_PT$SW_test_result4[[3]]
```

```
    Shapiro-Wilk normality test

data:  Refs
W = 0.9078, p-value = 0.0427
```

```
sw_results_Refs_PT$SW_test_result4[[4]]
```

```
    Shapiro-Wilk normality test

data:  Refs
W = 0.86365, p-value = 0.005977
```

- Indeed, the Shapiro-Wilks tests (except for the MAFP group) indicate that we should reject the null hypothesis that the data are normally distributed. We will now skip past testing for homogeneity of variances to the Kruskal-Wallis test as a robust version of a one-way ANOVA to test if there are any differences between group medians for references.

2. Now to conduct the Kruskal-Wallis test.

```
# To conduct Kruskal-Wallis
hyp_two_b_krusk <- kruskal.test(Refs ~ PaperType, data = data)

# To print the results of the Kruskal-Wallis test
hyp_two_b_krusk
```

```
    Kruskal-Wallis rank sum test

data:  Refs by PaperType
Kruskal-Wallis chi-squared = 0.49984, df = 3, p-value = 0.9189
```

- The Kruskal-Wallis test indicates that we should accept the null hypothesis that there are no differences between the median `Refs` across groups. Similarly to **Hypothesis 2a**, this test should be robust to outliers, so there is no need to remove them here.

Hypothesis 3: Certainty

Hypothesis 3 investigates the use of certainty language in cases of scientific fraud. While a case study of Deidrik Stapel tended to use more certain language (Markowitz & Hancock, 2014), others have found less certainty language in retracted papers than non-retracted papers (Dehdarirad & Schirone, 2023). Thus, **Hypothesis 3** is meant to provide clarity regarding the use of certainty language in scientific fraud. Specifically, it is stated as follows:

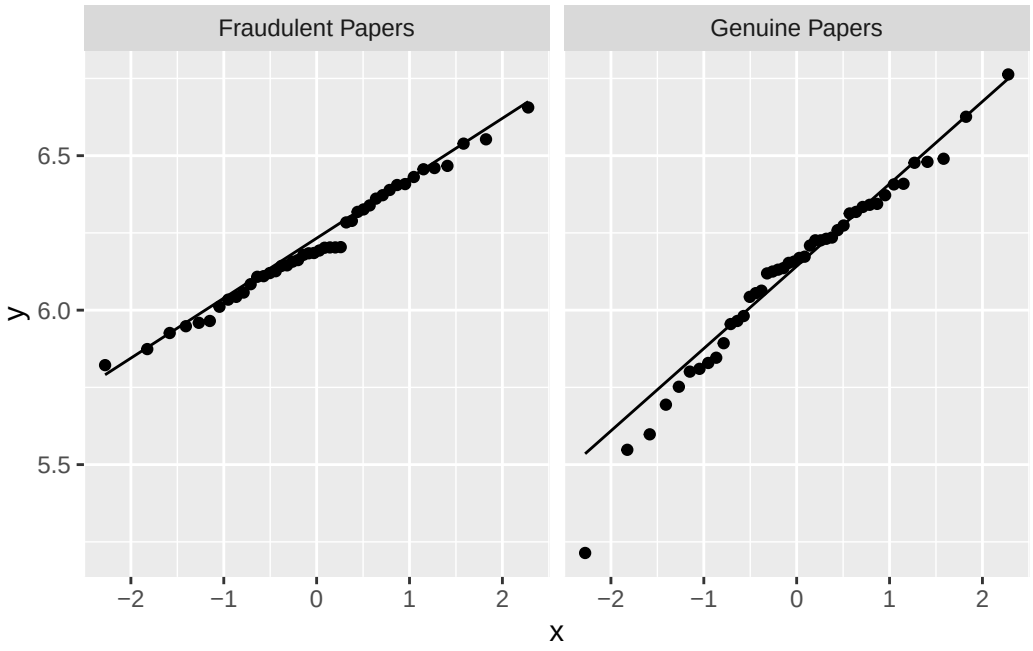
Hypothesis 3: Fraudulent research will contain less certainty than non-fraudulent research. [Replication]

Testing Hypothesis 3:

To test **Hypothesis 3**, we will compare the mean `CertSent` score between the `FPaper` and `GPaper` groups using an independent-samples t-test.

1. To check the assumption of normality we will use the `ggplot2` package to create Q-Q plots for `CertSent` within the `FPaper` and `GPaper` groups. The Q-Q plots will be investigated visually. If the assumption of normality is violated we will use a bootstrapped t-test to fit a more robust model.

```
# To produce Q-Q plots for CertSent within the FPaper and GPaper groups
ggplot(data, aes(sample = CertSent)) +
  stat_qq() +
  facet_wrap(~ Genuine_or_Fraudulent) +
  geom_qq_line()
```



- The QQ-plots seem to diverge from normality in the middle of the distribution (for fraudulent papers) and at the lower end of the distribution (for the genuine papers). I will assess non-normality more formally using Shapiro-Wilks tests within groups.

```
# Performing the Shapiro-Wilks test for normality within each group
sw_results_CertSent_GorF <- data %>%
  group_by(Genuine_or_Fraudulent) %>%
  summarise(SW_test_result5 = list(shapiro.test(CertSent)))

# Displaying the results
sw_results_CertSent_GorF$SW_test_result5[[1]]
```

```
    Shapiro-Wilk normality test

data:  CertSent
W = 0.98578, p-value = 0.8563
```

```
sw_results_CertSent_GorF$SW_test_result5[[2]]
```

```
    Shapiro-Wilk normality test
```

```
data:  CertSent
W = 0.96977, p-value = 0.2971
```

- The results from the Shapiro-Wilks test for normality indicate that we should accept the null hypothesis that the data are normally distributed within groups.
- Thus, we may assume normality for our t-test.

2. To check the assumption of homogeneity of variances, we will conduct Levene's test using the `car` package.

```
# To load the car package for Levene's test
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:boot':

    logit

The following object is masked from 'package:psych':

    logit

The following object is masked from 'package:dplyr':

    recode
```

```
# To conduct Levene's test for homogeneity of variances
hyp_three_levene_test <- leveneTest(CertSent ~ Genuine_or_Fraudulent, data = data)

# To display the Levene's test result
hyp_three_levene_test
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  1  4.0182 0.04816 *
      86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Levene's test for homogeneity of variance indicates that we should reject the null hypothesis that the variances of the two groups is equal. Thus, we will run a Welch's t-test, which does not assume equal variances.

3. Conducting the Welch's t-test.

```
# To conduct Welch's t-test
hyp_three_welch_t_test <- t.test(CertSent ~ Genuine_or_Fraudulent, data = data, var.equal

# To print the Welch's t-test result
hyp_three_welch_t_test
```

```
    Welch Two Sample t-test

data:  CertSent by Genuine_or_Fraudulent
t = 1.6329, df = 73.47, p-value = 0.1068
alternative hypothesis: true difference in means between group Fraudulent Papers and group G
95 percent confidence interval:
 -0.0193144  0.1945871
sample estimates:
mean in group Fraudulent Papers    mean in group Genuine Papers
                       6.213636                        6.126000
```

- Contrary to our hypothesis, Fraudulent Papers actually have higher mean certainty sentiment than Genuine Papers, although this result is not significant.
  - These results align with (Dehdarirad & Schirone, 2023).

Hypothesis 4:

Hypothesis 4 is a logical extension of the notion that to hide information you must first have control over related information flow. Based on this principle, **Hypothesis 4** is formulated as follows:

Hypothesis 4: For multi-author fraudulent papers, the fraudulent author will be the corresponding author more frequently than other authors. [Novel]

Hypothesis 4 rests on the following two assumptions:

- *First, all authors on a paper are equally likely to be the corresponding author.* Although this assumption oversimplifies norms of assigning scientists to be corresponding authors, in the absence of more information regarding, for example, author order, author responsibilities, or laboratory status, it is the most accurate prediction we can make regarding

the likelihood of an author being the corresponding author. That is, in the absence of other information it is the baseline prediction.

- *Second, if fraudulent authors to not attempt to control information flow, they are equally likely to be the corresponding as their coauthors are.* This may also oversimplify the norms of assigning scientists to be corresponding authors. For example, it is possible that fraudulent authors perform more data management, and it is possible that scientists that are responsible for data management are more likely to be corresponding authors. However, given the absence of this information (e.g., regarding research group norms), this assumption is reasonable.

These two assumptions allow us to make a prediction regarding the baseline expected frequency that fraudulent authors will be the corresponding authors of fraudulent research papers if they do not attempt to control information flow.

Testing Hypothesis 4:

To test **Hypothesis 4** we will conduct a binomial test to compare the observed likelihood that the fraudulent author is the corresponding author (i.e., `FraudCorrAuth`) to the expected probability given the assumption of equal likelihood for all authors and the average number of authors on each MAFP.

1. To attain the expected probabiltiy we will calculate the inverse of the mean number of authors on all MAFP where the corresponding author is known. To do this, we will first need to filter out other `PaperType` categories and MAFP cases where the corresponding author is unknown (i.e., `FraudCorrAuth` has a missing value) to make a new data frame, `hypothesis_four_df`. Then, we will calculate the mean `NumAuth` within this data frame and take its inverse.

```
# To create data frame that only includes MAFP with a score for FraudCorrAuth
hypothesis_four_df <- data[data$PaperType == 'MAFP' & data$FraudCorrAuth %in% c("0", "1"),

# To calculate the mean number of authors within the data frame
mean_MAFP_NumAuth_corrauth_known <- mean(hypothesis_four_df$NumAuth)

# To display the mean value
mean_MAFP_NumAuth_corrauth_known
```

```
[1] 3.125
```

```
# To calculate the inverse of mean_MAFP_NumAuth_corrauth_known
expected_prob_FraudCorrAuth <- 1 / mean_MAFP_NumAuth_corrauth_known
```

```
# To display the expected probability
expected_prob_FraudCorrAuth
```

```
[1] 0.32
```

- The mean number of authors in this subsample is 3.125, and the expected likelihood that the fraudulent author would be the corresponding author is therefore .32.

2. Next, we will conduct a binomial test to compare the observed likelihood that a corresponding author is the fraudulent author to the expected probability (`expected_prob_FraudCorrAuth`).

```
# To run the binomial test with the expected probability within the hypothesis four data f
hyp_four_binomial_test <- binom.test(sum(hypothesis_four_df$FraudCorrAuth), nrow(hypothesi

# To print the binomial test results
hyp_four_binomial_test
```

```
    Exact binomial test

data:  sum(hypothesis_four_df$FraudCorrAuth) and nrow(hypothesis_four_df)
number of successes = 4, number of trials = 8, p-value = 0.2776
alternative hypothesis: true probability of success is not equal to 0.32
95 percent confidence interval:
 0.1570128 0.8429872
sample estimates:
probability of success
                   0.5
```

- Although our observed likelihood is slightly higher than our expected probability, the hypothesis test indicates that we should accept the null hypothesis that there is no difference between them.
  - Of course, this test has very low power, with only 8 trials.

Saving Data

Now I will quickly export the main data frame to a csv file.

```
write.csv(data, "post_analysis_study_dataset.csv")
```

The data is now available as "post_analysis_study_dataset.csv".

Aggarwal, S. B., Chaitanya. (2022). *textstat: Calculate statistical features from text* [Python]. https://github.com/shivam5992/textstat

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1–47. https://www.researchgate.net/profile/Ryan-Boyd-8/publication/358725479_The_Development_and_Psychometric_Properties_of_LIWC-22/links/6210f62c4be28e145ca1e60b/The-Development-and-Psychometric-Properties-of-LIWC-22.pdf

Dehdarirad, T., & Schirone, M. (2023). Use of positive terms and certainty language in retracted and non-retracted articles: The case of biochemistry. *Journal of Information Science*, 1655515231176650. https://doi.org/10.1177/01655515231176650

Markowitz, D. M., & Hancock, J. T. (2014). Linguistic Traces of a Scientific Fraud: The Case of Diederik Stapel. *PLoS ONE*, *9*(8), e105937. https://doi.org/10.1371/journal.pone.0105937

Markowitz, D. M., & Hancock, J. T. (2016). Linguistic Obfuscation in Fraudulent Science. *Journal of Language and Social Psychology*, *35*(4), 435–445. https://doi.org/10.1177/0261927X15614605

Retraction Watch Database. (2023). http://retractiondatabase.org/RetractionSearch.aspx?

Rocklage, M. D., He, S., Rucker, D. D., & Nordgren, L. F. (2023). Beyond Sentiment: The Value and Measurement of Consumer Certainty in Language. *Journal of Marketing Research*, *60*(5), 870–888. https://doi.org/10.1177/00222437221134802

Schmidt, E. (2022). *Can linguistic features unmask fraudulent research? A study that builds an NLP classifier to distinguish retracted papers from non-retracted papers based on text and linguistic features.* [Master Thesis]. https://studenttheses.uu.nl/handle/20.500.12932/42411