WSP R setup and demographics code

Lizzie Jones[[1]](#footnote-1)

02/05/2021

# WSP - Initial data exploration

#### About R Markdowns

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>. To generate the document of all content, click the **Knit** button. To change the output (e.g. PDF, HTML) change the ‘output’ at the top to any of the outputs listerd here: <https://rmarkdown.rstudio.com/lesson-9.html>.

## Data cleaning

First I have conducted some data cleaning to identify any respondents or data points that need to be removed and explain why. First I converted the ‘TimeTaken’ column to a total number of seconds (SecsTaken) for easier to more easily investigate means and quantiles. I initially focussed on the fastest 10% of respondents across both surveys as they are most likely to have straightlined through the survey. I visually inspected the data, then used the ‘careless’ package to find evidence of straightliningm ‘even-odd’ consistencies, and intra-individual response variability (IRV), across the whole survey and within the multiple choice questions (particularly questions 4, 5, 13, 15, 16, 17, 22, 23, 24)

### Explore average time taken to complete questionnaire and check for straightlining  
all\_data$SecsTaken <- as.numeric(lubridate::seconds(all\_data$TimeTaken)) # Create numeric column of time taken (seconds)  
  
quantile(all\_data$SecsTaken, 0.1) # Fastest 10% of all respondents = completion in 188.9 seconds/ about 3 mins

## 10%   
## 188.9

quantile(all\_data$SecsTaken, 0.05) # Fastest 5% of all respondents = completion in 117.95 seconds/ about 2 mins

## 5%   
## 117.95

quantile(all\_data$SecsTaken, 0.025) # Fastest 2.5% of all respondents = completion in 70.975 seconds/ about 1.2 mins

## 2.5%   
## 70.975

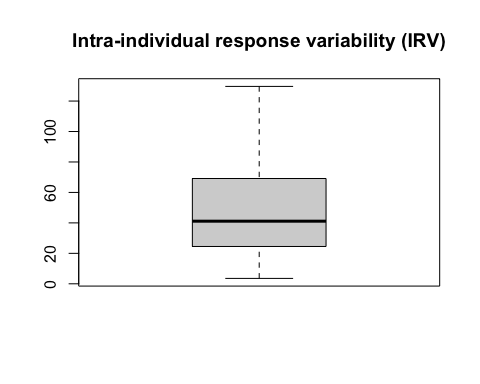
fastest\_10 <- subset(all\_data, SecsTaken < 188.8) # Sample of fastest 10% of all respondents  
fastest\_5 <- subset(all\_data, SecsTaken < 117.95) # Sample of fastest 5% of all respondents  
fastest\_2.5 <- subset(all\_data, SecsTaken < 70.975) # Sample of fastest 2.5% of all respondents  
  
summary(fastest\_5$SurveyType) # 96.07% of respondents in fastest 5% are from the NatRep sample

## NatRep Proactive   
## 171 7

summary(fastest\_2.5$SurveyType) # 100% of respondents in fastest 2.5% are from the NatRep sample

## NatRep Proactive   
## 89 0

### Checking fastest 5% of respondents  
# Checking the fastest 5% for straightlining across whole survey   
long\_fastest\_5 <- longstring(fastest\_5, avg = FALSE) # Identifies the longest string of identical consecutive responses for each respondent  
evenodd\_fastest\_5 <- evenodd(fastest\_5, rep(5,10)) # Calculates the even-odd consistency score  
irv\_fast\_5 <- irv(fastest\_5) # Calculates the intra-individual response variability (IRV)  
boxplot(irv\_fast\_5, main="Intra-individual response variability (IRV)")



# Checking the fastest 5% for straightlining within each set of mutliple choice questions

# Q6 habitat  
summary(all\_data$Q6\_habitat\_overallscore)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.4000 0.6000 0.5873 0.8000 1.0000

summary(fastest\_5$Q6\_habitat\_overallscore)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.6000 0.4146 0.6000 1.0000

# Overall knowledge score  
summary(all\_data$KnowledgeScore)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.100 3.600 3.579 5.000 7.000

summary(fastest\_5$KnowledgeScore)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 2.000 2.042 3.000 5.000

# NCI  
summary(all\_data$NCI)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 38.00 59.00 59.53 82.00 100.00

summary(fastest\_5$NCI)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 18.00 31.50 36.25 49.00 100.00

# Pro-cons behaviours  
summary(all\_data$ProCoBS)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 4.00 12.00 16.00 15.94 20.00 28.00 22

summary(fastest\_5$ProCoBS)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 4.00 8.00 12.00 12.55 16.00 28.00 3

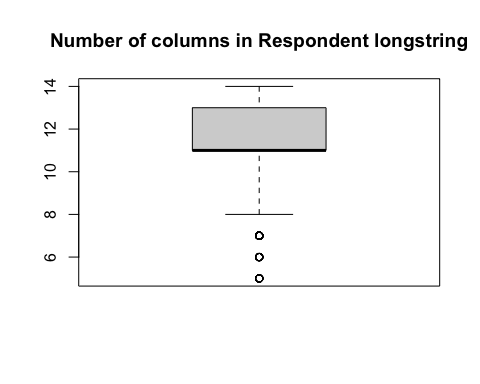
# Bird Interest Score  
summary(all\_data$BirdInterestScore)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.00 15.00 17.00 16.59 20.00 20.00

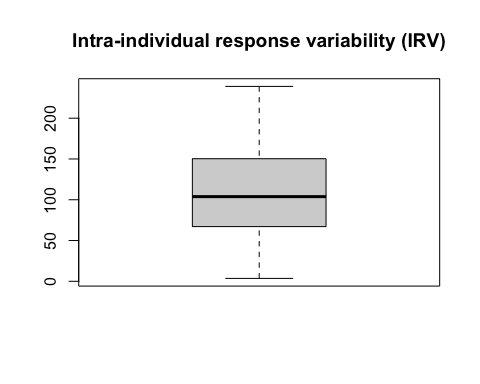
summary(fastest\_5$BirdInterestScore)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.00 11.00 12.00 12.63 14.00 20.00

##### Data cleaning using the 'Careless' package  
# Identifies the longest string of identical consecutive responses for each observation  
careless\_long <- longstring(all\_data, avg = FALSE)  
careless\_avg <- longstring(all\_data, avg = TRUE)  
boxplot(careless\_avg$longstr, main="Number of columns in Respondent longstring") #produce a boxplot of the longstring index



# Calculates the even-odd consistency score  
careless\_all <- evenodd(all\_data, rep(5,10))  
careless\_alldiag <- evenodd(all\_data, rep(5,10), diag = TRUE)  
  
# Calculates the intra-individual response variability (IRV)  
irv\_total <- irv(all\_data)  
boxplot(irv\_total, main="Intra-individual response variability (IRV)")



#calculate the irv over all items + calculate the irv for each quarter of the questionnaire  
irv\_split <- irv(all\_data, split = TRUE, num.split = 4)  
# boxplot(irv\_split$irv4) #produce a boxplot of the IRV for the fourth quarter

## Exploring Respondent demographics

Chart, treemap chart

Description automatically generatedThe distribution of gender and education is explored and compared between samples using stacked bar plots.Chart, bar chart

Description automatically generatedChart, bar chart, treemap chart

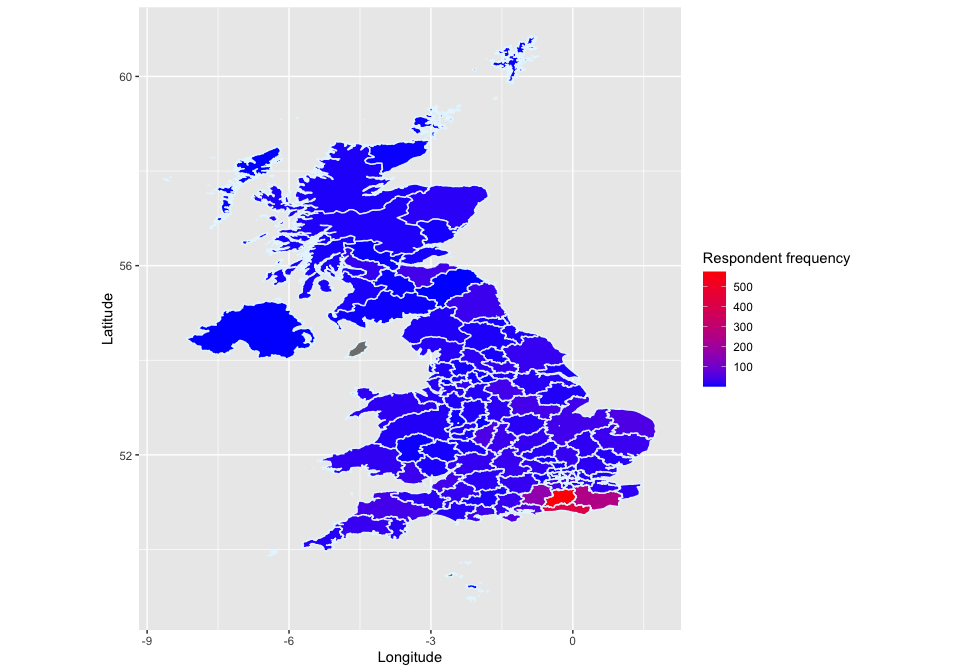
Description automatically generated

### Respondent demographics table

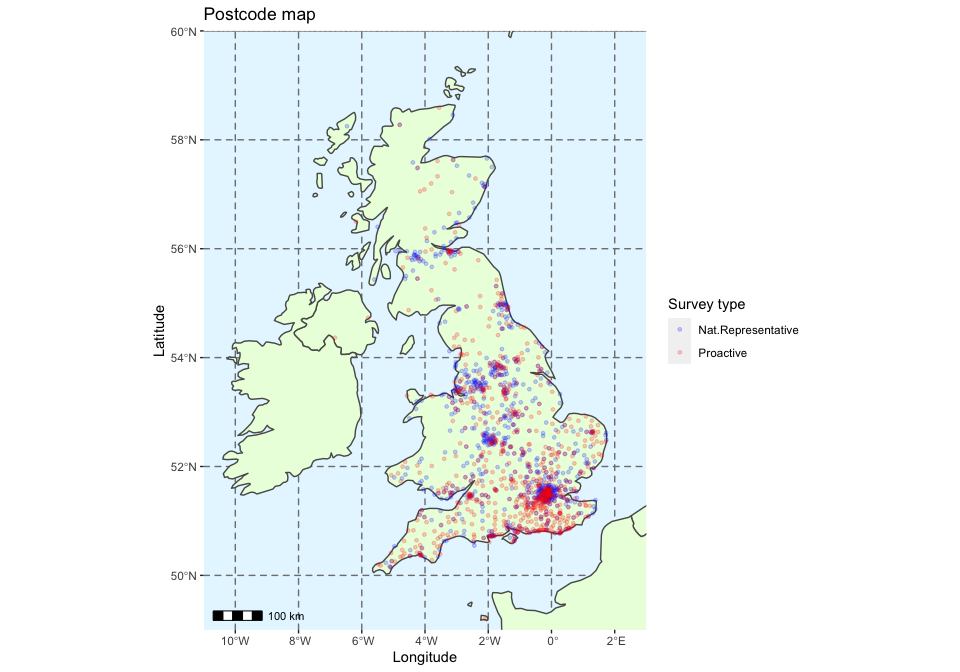
The table below (created using the package “table1”) outlines the demographic characteriscs of each of the two samples, and the overall demographics of all respondents across both samples. For each demographic variable the table provides a breakdown of the number of respondents within each level/group and the percentage.

### Respondent postcode mapping

Maps of respondent location using different methods: A. Map of first 1 or 2 alphabetical digits, (e.g. SW or N) for all participants with postcode boundaries, in which colour of area reflects density of participants per postcode region, and B. Map of first 4 digits of postcode (e.g., TN28), in which points are colour-coded according to survey type.



Map of first 2 digits of all postcodes (e.g., SW)



Map of first 4 digits of postcode (e.g., ), colour = survey type

1. University of Brighton, [l.jones4@brighton.ac.uk](mailto:l.jones4@brighton.ac.uk) [↑](#footnote-ref-1)