dataset B summary

Peder M. Isager

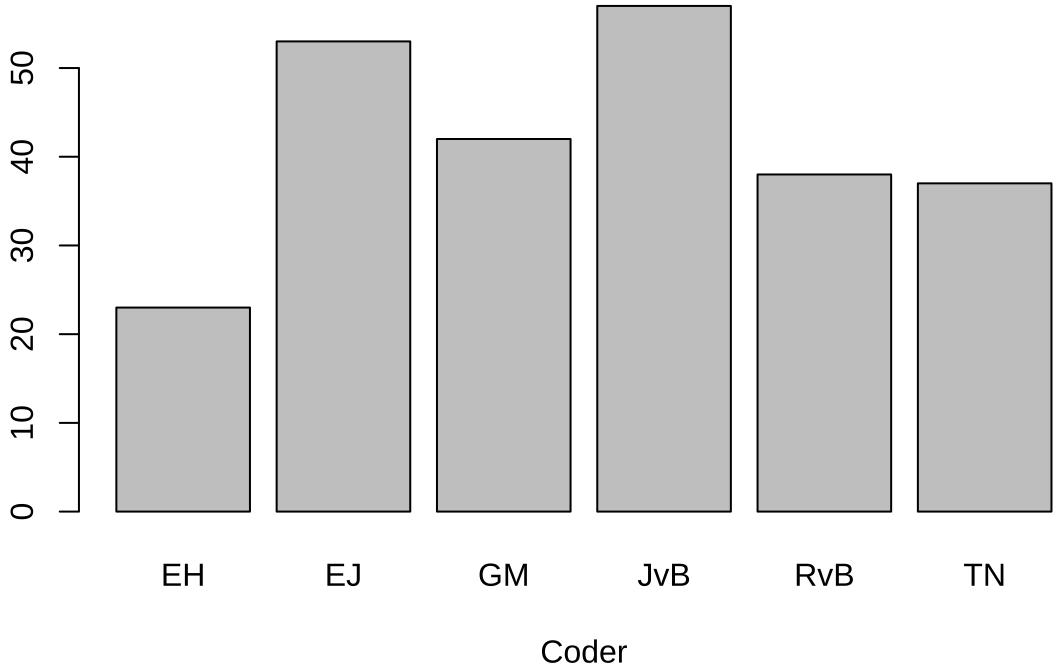
11/1/2019

```r
# Read dataset A

## Read the coded data
data.A.coded <- read.table(file = "../raw_data/dataset_A_coded.tsv", header = T, sep = "\t", quote = "\

## Read the full WoS info data
data.A.wos <- readRDS(file = "../raw_data/dataset_A_wos.rds")

## merge the two versions of the data by WOS number
data.A <- merge(data.A.coded, data.A.wos[, !names(data.A.wos) %in% c("AU", "TI", "PY", "DI")], by = "UT



# Wrangle dataset

## Filter out excluded rows
data.A$excluded[is.na(data.A$excluded)] <- 0
data.A.filt <- data.A[data.A$excluded != 1,]

## Reformat key columns
data.A.filt$study_number <- as.factor(data.A.filt$study_number)
data.A.filt$coder <- as.factor(data.A.filt$coder)
data.A.filt$resolver <- as.factor(data.A.filt$resolver)

## Calculate RV

data.A.filt$sample_bins <- cut(as.numeric(data.A.filt$sample_size), breaks = round(seq(1, max(as.numeri

data.A.filt$TC <- as.numeric(data.A.filt$TC)
data.A.filt$PY <- as.numeric(data.A.filt$PY)
data.A.filt$sample_size <- as.numeric(data.A.filt$sample_size)

current.year <- 2019
data.A.filt$RV <- (data.A.filt$TC / (current.year-data.A.filt$PY) ) / (data.A.filt$sample_size - 3)




# Sample 250 rows randomly from dataset A to generate dataset B

set.seed(11012019)  # Set seed to ensure reproducibility

sample.rows <- sample(x = nrow(data.A.filt), size = 250, replace = F)
data.B <- data.A.filt[sample.rows,]
```

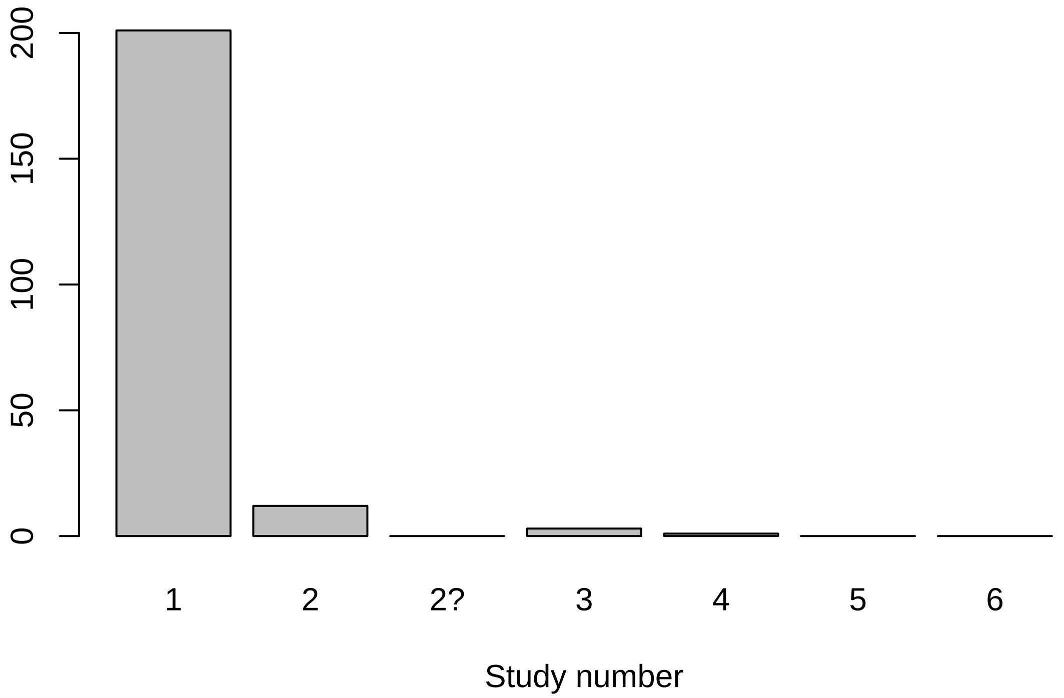
```r
# Summarize dataset B

## summary of key variables

key.vars <- c("PY", "study_number", "sample_size", "coder", "resolver", "excluded", "TC", "RV")

summary(data.B[, key.vars])
```

```
##        PY         study_number  sample_size     coder    resolver
##  Min.   :2009   1      :201   Min.   :  1.00   EH :23   AV  : 41
##  1st Qu.:2012   2      : 12   1st Qu.: 18.00   EJ :53   PI  : 33
##  Median :2015   3      :  3   Median : 24.00   GM :42   NA's:176
##  Mean   :2014   4      :  1   Mean   : 31.48   JvB:57
##  3rd Qu.:2017   2?     :  0   3rd Qu.: 37.00   RvB:38
##  Max.   :2019   (Other):  0   Max.   :202.00   TN :37
##                 NA's   : 33   NA's   :35
##     excluded       TC                RV
##  Min.   :0   Min.   :  0.00   Min.   :0.0000
##  1st Qu.:0   1st Qu.:  2.25   1st Qu.:0.0400
##  Median :0   Median :  9.00   Median :0.1082
##  Mean   :0   Mean   : 24.65   Mean   :   Inf
##  3rd Qu.:0   3rd Qu.: 28.00   3rd Qu.:0.2362
##  Max.   :0   Max.   :416.00   Max.   :   Inf
##                               NA's   :39
```

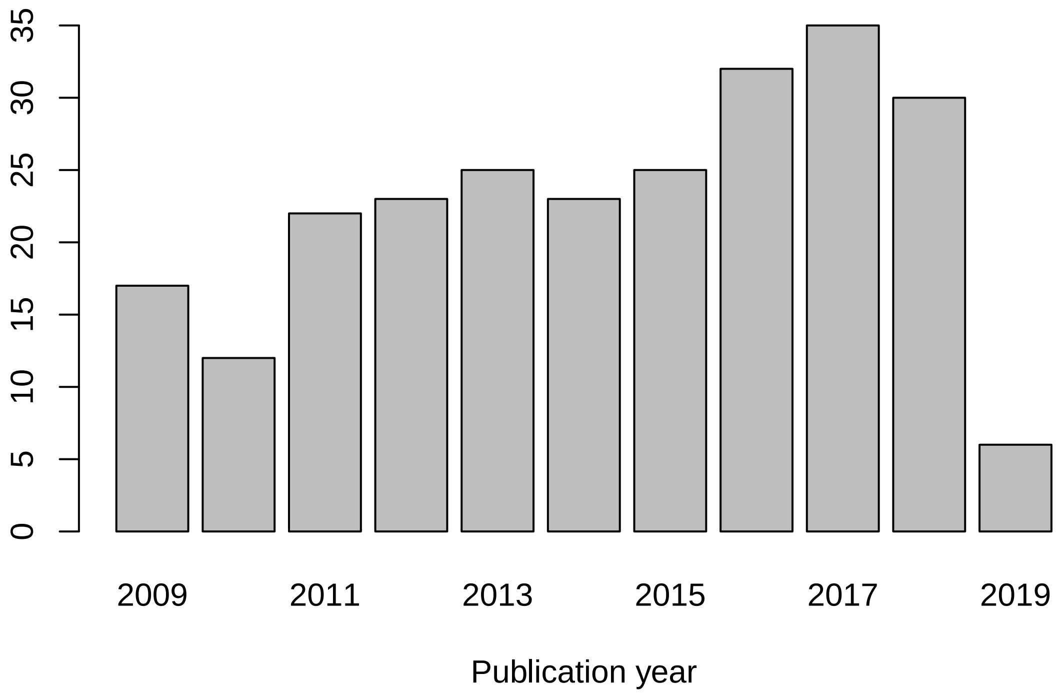
```r
## visualization of key variables

coder.freq <- table(data.B$coder)
barplot(coder.freq, xlab = "Coder")  # Plot frequency of coders
```



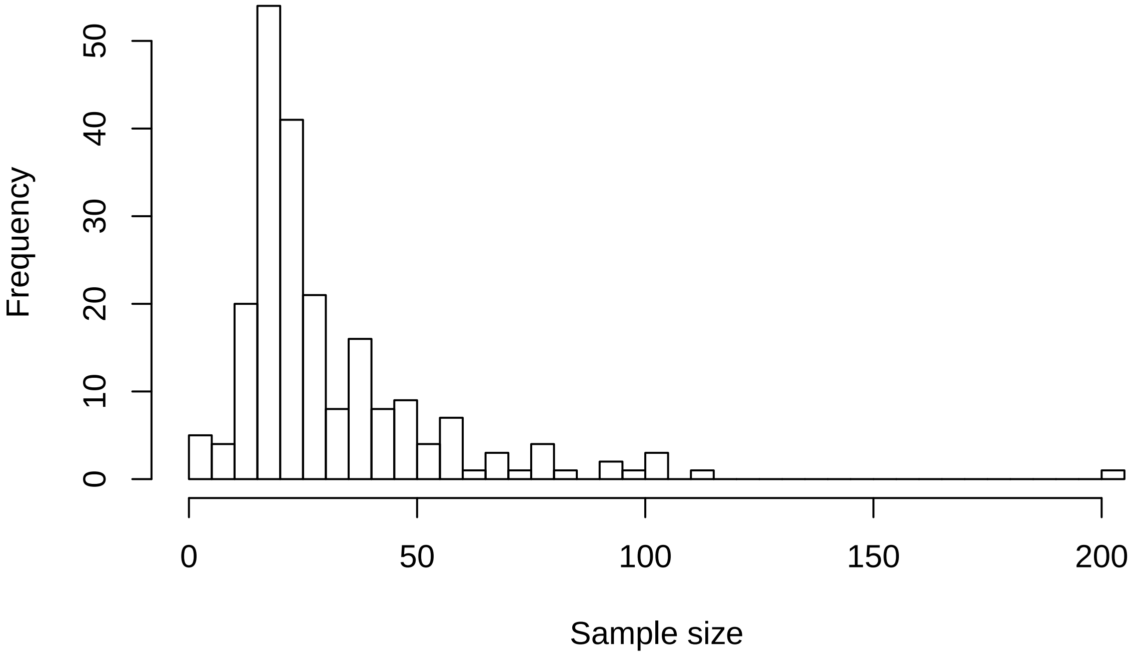
```
studyn.freq <- table(data.B$study_number)
barplot(studyn.freq, xlab = "Study number")  # Plot frequency of study numbers (first vs. second vs. fo
```



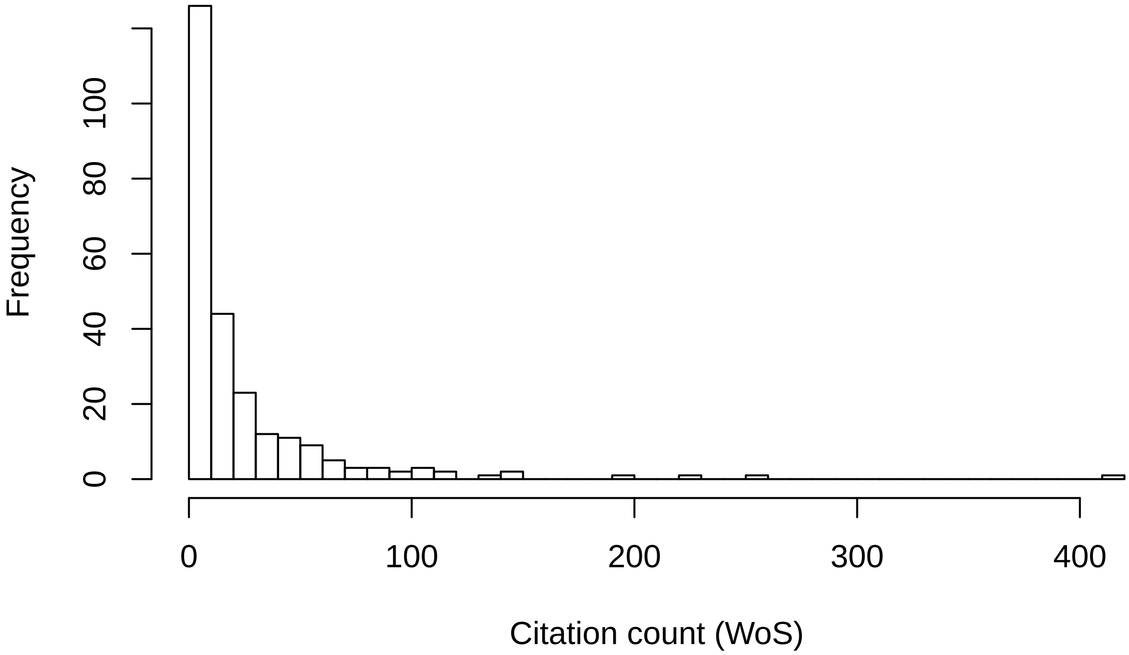
```
pubyear.freq <- table(data.B$PY)
barplot(pubyear.freq, xlab = "Publication year")  # Plot frequency of publication years
```



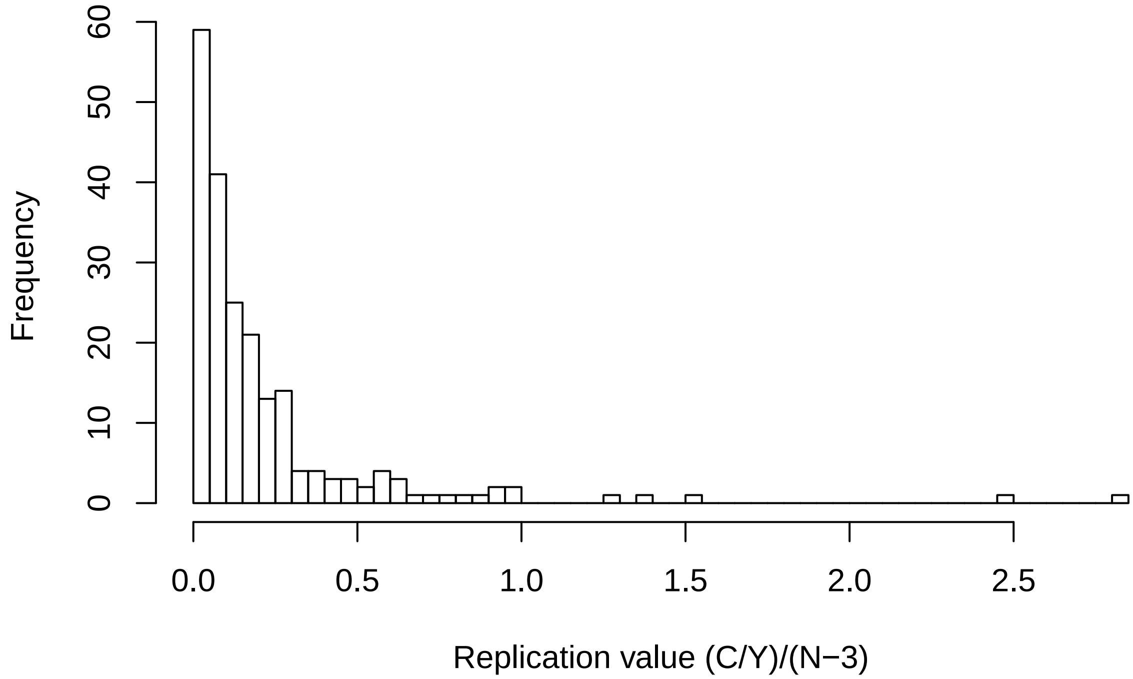
```
hist(data.B$sample_size, breaks = 50, xlab = "Sample size", main = "")  # Plot sample size distribution
```



```r
hist(data.B$TC, breaks = 50, xlab = "Citation count (WoS)", main = "")  # Plot citation count distribut
```



```
hist(data.B$RV, breaks = 50, xlab = "Replication value (C/Y)/(N-3)", main = "")  # Plot replication val
```


Frequency
0
10
20
30
40
50
60
0.0
0.5
1.0
1.5
2.0
2.5
Replication value (C/Y)/(N−3)