Tool for Creating Faculty Bibliographies from PubMed

Michael L. Bernauer
Jake L. Nash
Philip J. Kroth

May 10, 2019

```
require('lattice')
```

```
## Loading required package: lattice
```

```
require('dplyr')
```

```
## Loading required package: dplyr
```

```
##
## 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
```

```
require('tidyr')
```

```
## Loading required package: tidyr
```

```
# load faculty publications
p.dat = read.csv('./data/2015.csv') %>%
    mutate(year=2015) %>%
    union_all(read.csv('./data/2016.csv') %>%
                mutate(year=2016))

s.dat = read.csv('./data/self_reported_counts.csv') %>%
    filter(!is.na(count)) %>%
    filter(type == 'Article')
```

```
# preprocess faculty publication data
p.cnts = p.dat %>%
    group_by(department) %>%
    summarize(est.cnt = n_distinct(url,last_name,first_name,middle_initial)
              , auth.cnt = n_distinct(last_name, first_name, middle_initial)) %>%
    mutate(pub.per.auth = est.cnt/auth.cnt)

# aggregate self-reported counts
s.cnts = s.dat %>%
    group_by(department) %>%
    summarize(self.rep.cnt = sum(count))
```

```
# join self-reported data with estimates from pubmed
c.dat = s.cnts %>%
    left_join(p.cnts
              , by='department') %>%
    mutate(est.cnt = ifelse(is.na(est.cnt),0,est.cnt)) %>%
    mutate(diff = est.cnt - self.rep.cnt) %>%
    mutate(avg = (est.cnt+self.rep.cnt)/2)
```

```
options(knitr.kable.NA='-')
cap = "This table shows the total number of authors for each department in which
```

```
a publication was found in PubMed."
t1 = c.dat %>%
    mutate(department = stringr::str_to_title(department)) %>%
    rename(Department = department
           , `Self-Reported`=self.rep.cnt
           , `Estimated`=est.cnt
           , `Authors`=auth.cnt
           , `Pubs/Author`=pub.per.auth) %>%
    select(Department
           , `Authors`
           , `Self-Reported`
           , `Estimated`
           , `Pubs/Author`) %>%
    arrange(-`Authors`) %>%
    knitr::kable(digits=1, caption = cap)

t1
```

Table 1: This table shows the total number of authors for each department in which a publication was found in PubMed.

Department	Authors	Self-Reported	Estimated	Pubs/Author
Family & Community Medicine	43	154	148	3.4
Surgery	42	112	123	2.9
Emergency Medicine	37	66	93	2.5
Orthopaedics	31	36	81	2.6
Obstetrics & Gynecology	28	114	143	5.1
Anesthesiology	22	28	53	2.4
Neurosurgery	20	37	80	4.0
Neurology	16	37	80	5.0
Neurosciences	11	57	78	7.1
Health Sciences Library & Informatics Center	7	14	9	1.3
Dental Medicine	-	4	0	-

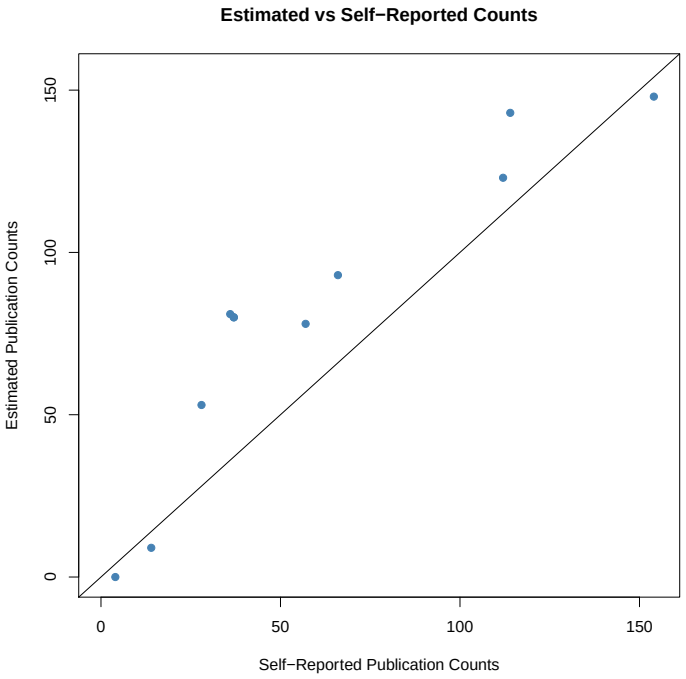
```
writeLines(t1,con='./tables/table.md')
```

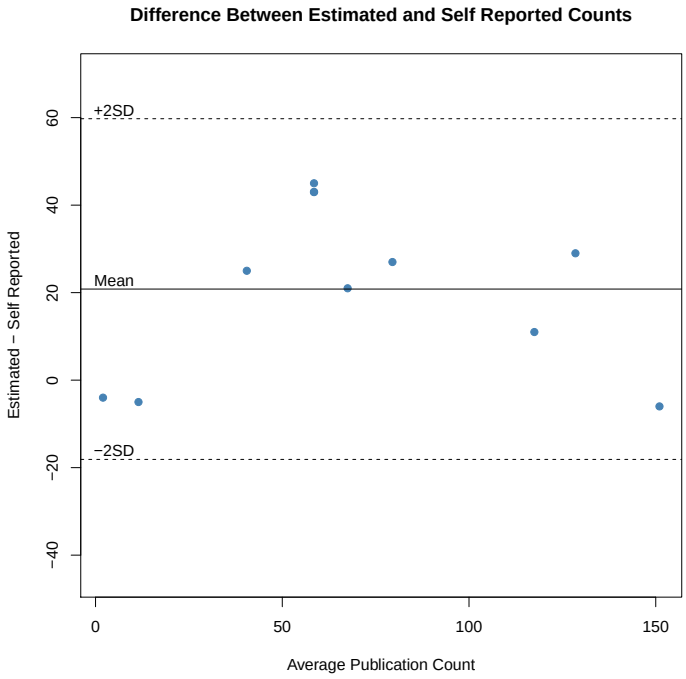
```
pdf(file='./figs/altman-bland.pdf', width=15, height=7.5)
par(mfrow=c(1,2))
plot(est.cnt ~ self.rep.cnt
     , data = c.dat
     , main = 'Estimated vs Self-Reported Counts'
     , xlab='Self-Reported Publication Counts'
     , xlim=c(0,155)
     , ylim=c(0,155)
     , pch=19
     , col='steelblue'
     , ylab='Estimated Publication Counts')
abline(a=0,b=1)

plot(diff ~ avg
     , data = c.dat
     , main = 'Difference Between Estimated and Self Reported Counts'
     , xlab = 'Average Publication Count'
     , ylab = 'Estimated - Self Reported'
     , col='steelblue'
     , pch=19
     , ylim=c(-45,70))
abline(h=mean(c.dat$diff)-2*sd(c.dat$diff), lty=2, col='black')
abline(h=mean(c.dat$diff), lty=1, col='black')
```

```
abline(h=mean(c.dat$diff)+2*sd(c.dat$diff), lty=2, col='black')
text(x=5,y=mean(c.dat$diff)+2+2*sd(c.dat$diff), '+2SD')
text(x=5,y=mean(c.dat$diff)+2, 'Mean')
text(x=5,y=mean(c.dat$diff)+2-2*sd(c.dat$diff), '-2SD')
dev.off()
```

```
## pdf
##   2
```






```
# check for normality
shapiro.test(c.dat$diff)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  c.dat$diff
## W = 0.89055, p-value = 0.1413
```

```
# test difference
t.test(c.dat$diff)
```

```
##
##  One Sample t-test
##
## data:  c.dat$diff
## t = 3.5468, df = 10, p-value = 0.005296
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   7.740081 33.896283
## sample estimates:
## mean of x
##  20.81818
```

```
tmp = p.dat %>%
    unique() %>%
    group_by(entrezuid) %>%
    summarize(auth.cnt = n()) %>%
    count(auth.cnt)
```

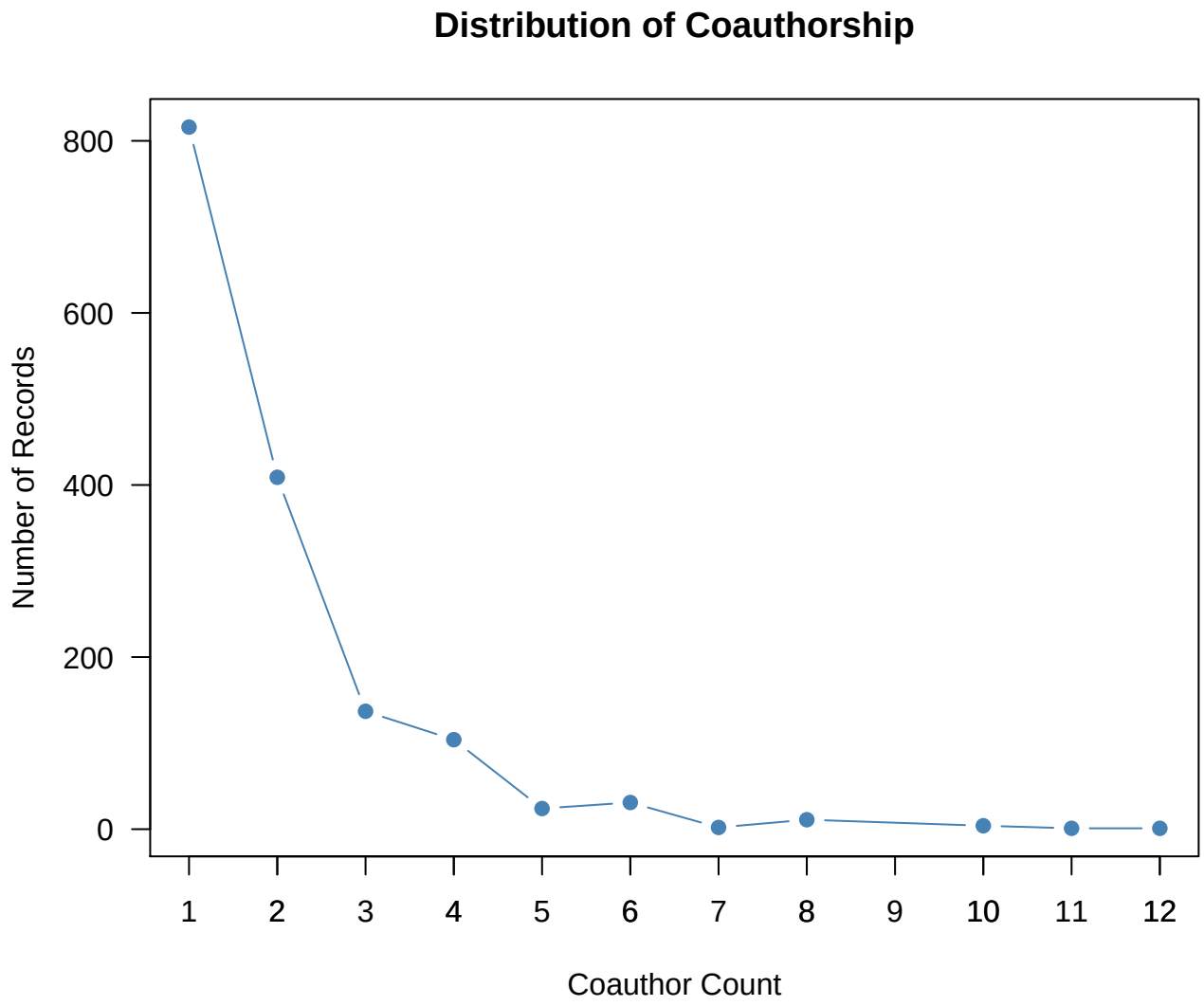


Figure 1: Coauthorship distribution

```
pdf(file='./figs/coauthor.pdf',width=7,height=6)
p = plot(n ~ auth.cnt
     , data = tmp
     , type='b'
     , pch=19
     , las=1
     , ylab='Number of Records'
     , main = 'Distribution of Coauthorship'
     , xlab='Coauthor Count'
     , col='steelblue')
axis(side=1,at=1:12)
dev.off()
```

```
## pdf
##   2
```