Proportion of Brazil as the affiliation of documents in SciELO Brazil

Our goal is to find the proportion of Brazil in the affiliations of documents belonging to the SciELO Brazil collection. Let d be a document, then the proportion we're looking for is:

$$p(d) = \frac{\text{number of affiliations of } d \text{ in Brazil}}{\text{total number of affiliations of } d}$$

We're going to study the $p(d)$ on an yearly basis, counting only the affiliations whose country we know.

Let's load from SciELO Analytics the CSV of documents affiliations in SciELO Brazil:

```r
# We shouldn't interpret Namibia (NA) as "not available"
doc_aff <- read.csv("tabs_bra/documents_affiliations.csv", na.strings = c())
dim(doc_aff) # Number of rows and columns
```

```
## [1] 804928     26
```

```r
as.data.frame(t(head(doc_aff, 1))) # First entry
```

	1
extraction.date	2018-09-13
study.unit	document
collection	scl
ISSN.SciELO	0100-879X
ISSN.s	0100-879X;1414-431X
title.at.SciELO	Brazilian Journal of Medical and Biological Research
title.thematic.areas	Biological Sciences;Health Sciences
title.is.agricultural.sciences	0
title.is.applied.social.sciences	0
title.is.biological.sciences	1
title.is.engineering	0
title.is.exact.and.earth.sciences	0
title.is.health.sciences	1
title.is.human.sciences	0
title.is.linguistics..letters.and.arts	0
title.is.multidisciplinary	0
title.current.status	current
document.publishing.ID..PID.SciELO.	S0100-879X1998000800006
document.publishing.year	1998
document.type	research-article
document.is.citable	1
document.affiliation.instituition	University of Gorakhpur
document.affiliation.country	
document.affiliation.country.ISO.3166	
document.affiliation.state	
document.affiliation.city	

R already simplifies the column names in some sense, replacing the whitespaces and special characters by a dot. We can see the names with `names(doc_aff)`.

Categorical fields are known as *factors*.

```r
class(doc_aff$document.type)
```

```
## [1] "factor"
```

```r
class(doc_aff$document.affiliation.country.ISO.3166)
```

```
## [1] "factor"
```

The *levels* of a factor are the values one *factor* vector can have.

```r
levels(doc_aff$document.type)
```

```
##  [1] "abstract"            "addendum"            "article-commentary"
##  [4] "book-review"         "brief-report"        "case-report"
##  [7] "correction"          "editorial"           "letter"
## [10] "news"                "press-release"       "rapid-communication"
## [13] "research-article"    "review-article"      "undefined"
```

```r
levels(doc_aff$document.affiliation.country.ISO.3166)
```

```
##   [1] ""   "AE" "AG" "AL" "AM" "AN" "AO" "AR" "AS" "AT" "AU" "AZ" "BA" "BB"
##  [15] "BD" "BE" "BF" "BG" "BH" "BI" "BJ" "BO" "BR" "BS" "BT" "BW" "BY" "CA"
##  [29] "CD" "CF" "CH" "CI" "CL" "CM" "CN" "CO" "CR" "CS" "CU" "CV" "CY" "CZ"
##  [43] "DE" "DK" "DO" "DZ" "EC" "EE" "EG" "ES" "ET" "FI" "FJ" "FR" "GA" "GB"
##  [57] "GD" "GE" "GF" "GH" "GN" "GP" "GR" "GT" "GW" "GY" "HK" "HN" "HR" "HT"
##  [71] "HU" "ID" "IE" "IL" "IN" "IQ" "IR" "IS" "IT" "JM" "JO" "JP" "KE" "KG"
##  [85] "KN" "KR" "KW" "KY" "KZ" "LA" "LB" "LK" "LR" "LT" "LU" "LV" "LY" "MA"
##  [99] "ME" "MG" "MI" "MK" "ML" "MM" "MN" "MT" "MU" "MW" "MX" "MY" "MZ" "NA"
## [113] "NE" "NG" "NI" "NL" "NO" "NP" "NZ" "OM" "PA" "PE" "PG" "PH" "PK" "PL"
## [127] "PR" "PS" "PT" "PY" "QA" "RO" "RS" "RU" "RW" "SA" "SC" "SD" "SE" "SG"
## [141] "SI" "SK" "SL" "SN" "SR" "SS" "SU" "SV" "SY" "TG" "TH" "TL" "TN" "TR"
## [155] "TT" "TW" "TZ" "UA" "UG" "US" "UY" "VE" "VN" "YE" "YU" "ZA" "ZM" "ZW"
```

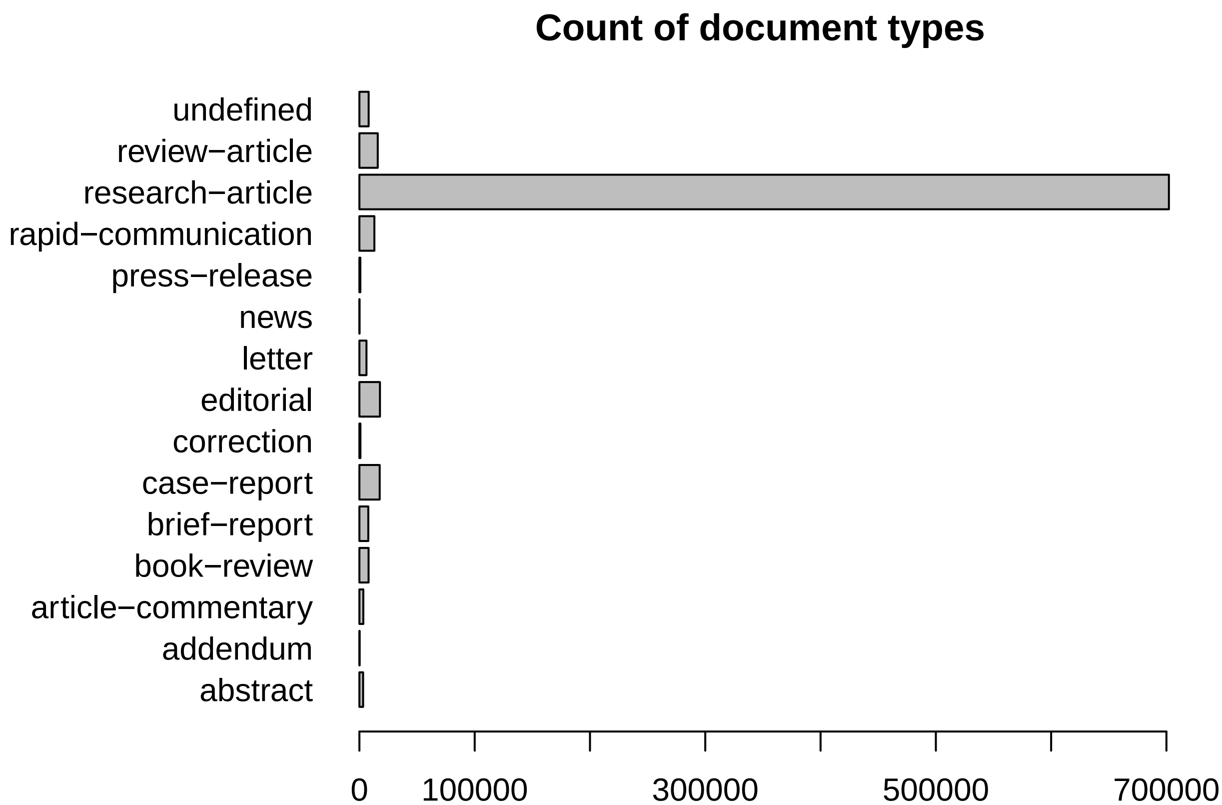
Most entries are research articles, we'll work only with this document type:

```r
options(scipen = 6) # Avoid scientific notation in plots
```

```r
as.data.frame(summary(doc_aff$document.type))
```

	summary(doc_aff$document.type)
abstract	3215
addendum	192
article-commentary	3396
book-review	7948
brief-report	7732
case-report	17602
correction	875
editorial	17827
letter	6122
news	111
press-release	839
rapid-communication	12995
research-article	702149
review-article	15902
undefined	8023

```r
par(mar = c(3, 9, 2, 2) + .1)
barplot(summary(doc_aff$document.type),
        horiz = TRUE,
        las = 1, # Horizontal labels
        main = "Count of document types")
```

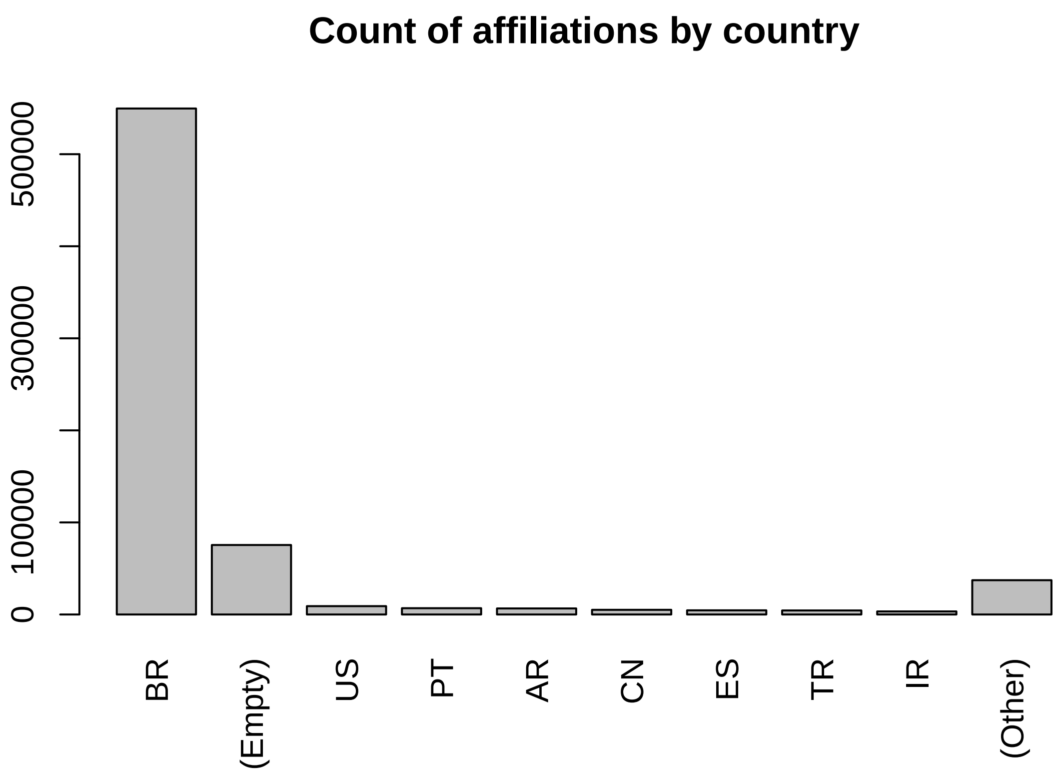



```r
articles <- doc_aff[doc_aff$document.type == "research-article",]
nrow(articles)
```

```
## [1] 702149
```

Most affiliation entries are from Brazil (that's somewhat expected for a Brazilian collection).

```r
aff_country_summary <- summary(articles$document.affiliation.country.ISO.3166,
                               maxsum = 10)
aff_country_summary_names <- replace(names(aff_country_summary),
                                     names(aff_country_summary) == "",
                                     "(Empty)")
acs_xmidpoints <- barplot(aff_country_summary,
                          axisnames = FALSE,
                          main = "Count of affiliations by country")
axis(1, at = acs_xmidpoints, las = 2,
        labels = aff_country_summary_names, xpd = TRUE,
        tick = FALSE)
```



Let's build a dataset with just four columns:

- One regarding the document publication year;
- One regarding to the PID, a way to identify an article;
- One logical, `TRUE` if an article have a Brazilian affiliation;
- One logical, `TRUE` if an article have a non-Brazilian affiliation.

We should remove the empty country entries, since they might belong to any country (Brazil or other). Using two columns should be cleaner to understand than merging the Brazilian/non-Brazilian affiliation as a single column.

```
dataset <- data.frame(
  articles$document.publishing.year,
  articles$document.publishing.ID..PID.SciELO.,
  articles$document.affiliation.country.ISO.3166 == "BR",
  grepl("[^B].|.[^R]", articles$document.affiliation.country.ISO.3166)
)
names(dataset) <- c("year", "pid", "br", "not_br")
dataset <- dataset[dataset$br | dataset$not_br,]
head(dataset)
```

	year	pid	br	not_br
624	1998	S0074-02761998000300014	TRUE	FALSE
2319	1998	S0102-76381998000400005	TRUE	FALSE
2321	1998	S0102-76381998000400003	TRUE	FALSE
2323	1998	S0102-76381998000400009	TRUE	FALSE
2333	1998	S0102-76381998000400004	TRUE	FALSE
2334	1998	S0102-76381998000400010	TRUE	FALSE

```r
nrow(dataset)
```

```
## [1] 626660
```

As all entries are either `br` or `not_br`, we just need to calculate the mean of `br` for each PID. We'll use `dplyr` to group that result by the PID.

```r
library(dplyr) # Masks intersect, setdiff, setequal, union, filter, lag
```

```r
proportions <- dataset %>% group_by(pid) %>% summarize(mean(br), max(year))
proportions <- proportions[c(2, 3)]
names(proportions) <- c("prop", "year")
nrow(proportions)
```

```
## [1] 284274
```

```r
head(proportions)
```

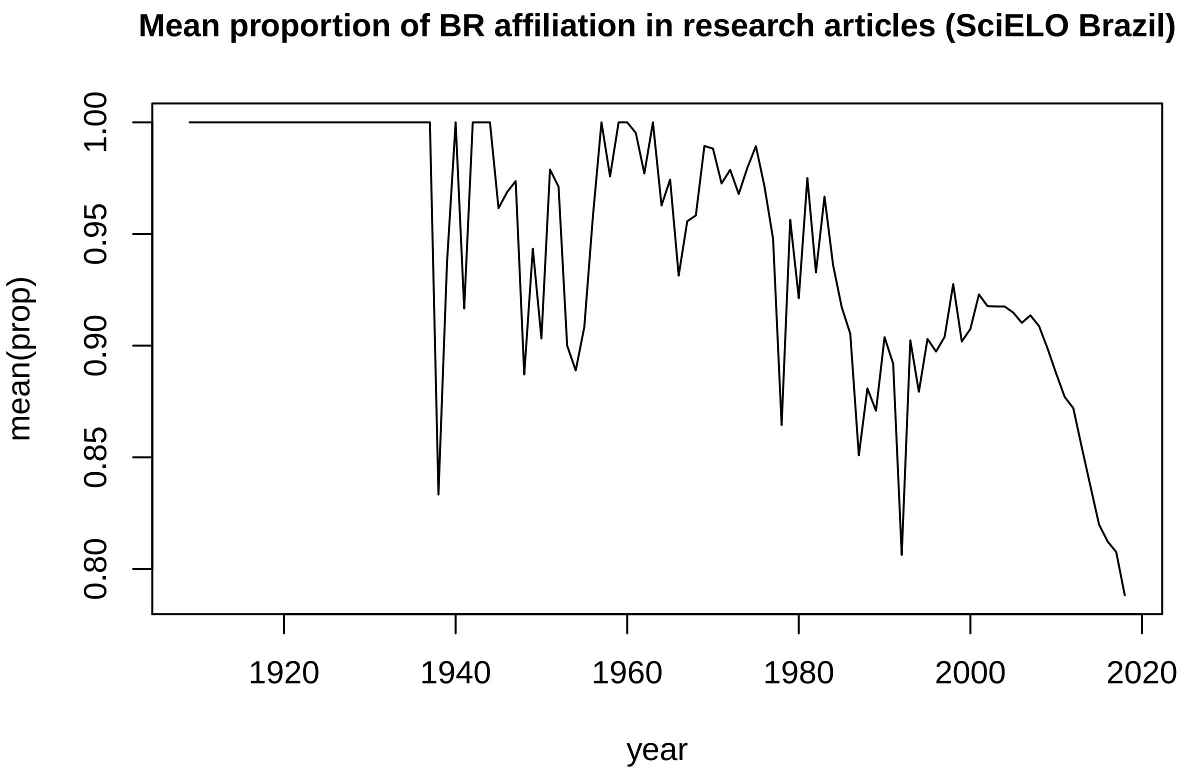
prop	year
1	1998
1	1998
0	1998
1	1998
1	1998
1	1998

Let's see the evolution of the mean of these proportions:

```r
mprops <- proportions %>% group_by(year) %>% summarize(mean(prop))
min(mprops$year, na.rm = TRUE) # Oldest document publication year
```

```
## [1] 1909
```

```r
plot(
  mprops,
  type = "l",
  cex.main = 1,
  main = paste("Mean proportion of BR affiliation",
               "in research articles (SciELO Brazil)",
               sep = " ")
)
```



The raw data:

```
library(kableExtra)
mprops_all_years <- merge(data.frame(year = 1909:2018), mprops, all.x = TRUE)
mprops_all_years$year = as.character(mprops_all_years$year)
kable(
  cbind(mprops_all_years[seq(from = 1, length = 22),],
        mprops_all_years[seq(from = 23, length = 22),],
        mprops_all_years[seq(from = 45, length = 22),],
        mprops_all_years[seq(from = 67, length = 22),],
        mprops_all_years[seq(from = 89, length = 22),]),
  digits = 5, format.args = list(nsmall = 5),
) %>%
  kable_styling(latex_options = "striped") %>%
  add_header_above(c("1909-1930" = 2, "1931-1952" = 2, "1953-1974" = 2,
                     "1975-1996" = 2, "1997-2018" = 2))
```

1909-1930		1931-1952		1953-1974		1975-1996		1997-2018	
year	mean(prop)	year	mean(prop)	year	mean(prop)	year	mean(prop)	year	mean(prop)
1909	1.00000	1931	1.00000	1953	0.90000	1975	0.98935	1997	0.90395
1910	NA	1932	1.00000	1954	0.88889	1976	0.97136	1998	0.92757
1911	NA	1933	1.00000	1955	0.90833	1977	0.94796	1999	0.90184
1912	NA	1934	1.00000	1956	0.95775	1978	0.86443	2000	0.90748
1913	NA	1935	NA	1957	1.00000	1979	0.95635	2001	0.92294
1914	NA	1936	1.00000	1958	0.97581	1980	0.92127	2002	0.91769
1915	NA	1937	1.00000	1959	1.00000	1981	0.97504	2003	0.91756
1916	NA	1938	0.83333	1960	1.00000	1982	0.93280	2004	0.91753
1917	1.00000	1939	0.93750	1961	0.99528	1983	0.96676	2005	0.91481
1918	1.00000	1940	1.00000	1962	0.97710	1984	0.93607	2006	0.91021
1919	NA	1941	0.91667	1963	1.00000	1985	0.91731	2007	0.91356
1920	NA	1942	1.00000	1964	0.96277	1986	0.90533	2008	0.90882
1921	NA	1943	1.00000	1965	0.97436	1987	0.85090	2009	0.89872
1922	1.00000	1944	1.00000	1966	0.93137	1988	0.88082	2010	0.88754
1923	1.00000	1945	0.96154	1967	0.95570	1989	0.87085	2011	0.87702
1924	1.00000	1946	0.96875	1968	0.95833	1990	0.90380	2012	0.87200
1925	1.00000	1947	0.97368	1969	0.98944	1991	0.89194	2013	0.85422
1926	1.00000	1948	0.88710	1970	0.98824	1992	0.80630	2014	0.83702
1927	1.00000	1949	0.94340	1971	0.97264	1993	0.90240	2015	0.81992
1928	1.00000	1950	0.90323	1972	0.97877	1994	0.87933	2016	0.81231
1929	1.00000	1951	0.97887	1973	0.96792	1995	0.90300	2017	0.80760
1930	NA	1952	0.97115	1974	0.97958	1996	0.89737	2018	0.78823

Is that significantly decreasing? To answer that, let's consider the linear regression slope, which should be negative, that is, the mean proportion should get lower when the year gets higher.

```r
regr <- lm(mean.prop. ~ year, data.frame(mprops))
summary(regr)
```

```
##
## Call:
## lm(formula = mean.prop. ~ year, data = data.frame(mprops))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.149550 -0.011874  0.005626  0.025461  0.058664
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.7169339  0.2630088   14.13   <2e-16 ***
## year        -0.0014108  0.0001336  -10.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03833 on 96 degrees of freedom
## Multiple R-squared:  0.5375, Adjusted R-squared:  0.5327
## F-statistic: 111.6 on 1 and 96 DF,  p-value: < 2.2e-16
```

The slope (the `year` estimate) is negative.

But is that negative for the 95%CI range?

```r
confint(regr, level = .95)
```

```
##                    2.5 %       97.5 %
## (Intercept)  3.194865642  4.239002162
## year        -0.001675865 -0.001145654
```

Yes, it's decreasing! The slope (`year`, last row of `confint` result) is negative for the entire 95% confidence interval.