

R

table(): クロス集計

■ Reference

- ・ <http://eau.uijin.com/stats/frequencies.html>
- ・ <https://www1.doshisha.ac.jp/~mjin/R/46/46.html>

例

■ サンプル・データ

- ・ <https://www.openintro.org/data/index.php?data=acs12>
 - ・ Results from the US Census American Community Survey, 2012
 - ・ 2012 年のアメリカ人口統計からの 2,000 のサンプル

■ データの取り込み

■ クロス集計 : table(データフレームのカラム , データフレームのカラム)

- ・ 人種と教育

■ クロス集計 : table(dataframe[,c(カラム番号を並べる)])

■ NA も集計に入れる : useNA = "ifany"

- ・ useNA = "always" とすると、縦横ですべて表示

■ 縦横の集計も加える : addmargins()

- ・ NA の数を確認 : is.na() で TRUE が NA の数

■ 比率で表示 : prop.table(table())

- ・ 全体の内訳の比率

■ 比率で表示 (四捨五入) : round(prop.table(table()),2)

- ・ 小数点以下 2 桁

■ 全体の比率ではなく、横で集計した比率 : margin = 1

- ・ この場合は、人種ごとに学歴の比率を出す
- ・ 白人の 73% は高卒
- ・ 大卒の比率はアジア系が一番高い
- ・ 大学院修了者の比率もアジア系が一番高い
 - ・ アジア系の 13% の人は大学院修了

■全体の比率ではなく、縦で集計した比率：margin = 2

- ・ 学歴ごとに、それを構成する人種の割合
- ・ 大卒の 83% は白人

モザイク・グラフ

vcd: Visualizing Categorical Data

```
install.packages("vcd")
```

<https://cran.r-project.org/web/packages/vcd/index.html>

<https://stats.stackexchange.com/questions/147863/how-to-interpret-the-residual-colors-on-a-mosaic-plot>

- ・ オプションで shade=T つけると、色がつく。
 - ・ 残差分析
 - ・ 青：期待値より観察値が多い
 - ・ 赤：期待値より観察値が少ない

結果をデータフレームに変換：as.data.frame()

■変換前

```
table(nns.bigram.df) %>% head()
```

ngram.all	0 years	05 euro	1 2019	1 euro	1 grandfather	1 hour
	327	209	23	418	295	159

■変換後

```
table(nns.bigram.df) %>% as.data.frame() %>% head()
```

	ngram.all	Freq
1	0 years	327
2	05 euro	209
3	1 2019	23
4	1 euro	418
5	1 grandfather	295
6	1 hour	159