Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

金田　拓

4 Basic statistical methods

✓ 4.5 Two vectors with counts　　pp. 111 - 113

- 今までのセクションでは実測値を扱ってきたが、整数である頻度カウントを扱う際にはまた違う手法が必要となる。そこで、オランダ語の動詞データ(auxiliaries)を用いて、規則性(Regularity)と助動詞(Aux)の選択を表にしてみる

```
> xt = xtabs(~ Aux + Regularity, data = auxiliaries)
> xt
```

- 割合を表にするためには、prop.table( )関数を用いる。2 つ目の引数が 1 なら行について、2 なら列について割合の計算を行う。

```
> prop.table(xt, 1)          #rows add up to 1
> prop.table(xt, 2)          #columns add up to 1
```

- 直接全体の合計で割れば、要素が全体に占める割合が求められる

```
> xt/sum(xt)
```

- 結果から、不規則動詞には hebben、規則動詞には zijn が多く用いられていることが見てとれる。モザイクプロットで観察すると、より差異がはっきりと観察できる。

```
> mosaicplot(xt, col = TRUE)
```

- 仮に以下のような観察結果が得られたとする。

```
> x = data.frame(irregular = c(100, 8, 30), regular = c(77, 6, 22))
> rownames(x) = c("hebben", "zijn", "zijnheb")
> x
```

結果を見る限り、割合は均等であるようだが、これらには本当に差が無いのか？統計的仮説検定を行う。

- 分割表を検定するには、χ二乗検定かフィッシャーの正確確率検定を行う
- χ二乗検定は以前にもやった通り

```
> chisq.test(xt)
```

- より詳細な情報を求める場合には、summary( )を用いる

```
> summary(xt)
```

- *p* 値が小さいのでやはり差がある。
- 作ったデータの方は、予想通り差が無い

```
> chisq.test(x)
```

- あまり調査するデータが大きくない時は、フィッシャーの正確確率検定の方が正確といえる。

```
> fisher.test(xt)
```

この例に関しては、フィッシャーの正確確率検定で得られた *p* 値がカイ二乗よりも低かった。