Corpus Statistics
2009/01/11
Baayen (2008: 128-138)

Toru Hitomi

5.1.3 Tables with counts: Correspondence analysis

主成分分析や因子分析 → two-way tables of measurements
コレスポンデンス(対応)分析 → two-way contingency tables(頻度データ)

- コレスポンデンス分析の手順
  - 2 種類の距離行列(列ごと、行ごと)を算出する。
    chi-squared distance(**χ 2 乗距離**)→行と列の近接性が相関を示している。
  - *"カイ二乗距離は、列の割合に基づく重み付きユークリッド距離と見ることもできる。" (Everitt, 2007)
  - 2 次元プロットを打つ。

e.g.① Ernestus et al. (2007): 中世フランス語において、統語構造のレジスター・通時的変容の研究

- 29 名の筆者、35 種類の 3 タグ連鎖(tag trigram)
- 2 種類のdata frame: oldFrench(タグ連鎖の頻度)とoldFrenchMeta(テキスト情報)
- 手順
  - CA 関数 corres.fnc() の割り当て
  - CA 関数の要約
    - eigenvalue (固有値) rates: 主成分分析の説明率と似た解釈
      *より高次の次元はあまり考慮されない。
      Eigenvalue rates:
      0.1704139 0.1326913 0.06854973 0.05852097 0.05394474 ...
      第一要因→X 軸 第二要因→Y 軸

      Factor 1

	coordinates	correlations	contributions
T30.16.00	−0.113	0.074	0.012
T00.31.51	−0.560	0.464	0.103
T16.00.31	−0.139	0.053	0.006

      ...
      Factor 2

	coordinates	correlations	contributions
T30.16.00	0.119	0.082	0.017
T00.31.51	0.205	0.062	0.018
T16.00.31	0.255	0.179	0.024

      ...
    - 座標、相関係数、寄与率
  - 作図する。
    パターン 1
    ```
    > plot(oldFrench.ca)          # text code をプロットしている。
    ```
    パターン 2(Figure 5.6, p. 132)
    ```
    > plot(oldFrench.ca, rlabels = oldFrenchMeta$Genre,      # ジャンルごとのプロットに置き換え
    + rcol = as.numeric(oldFrenchMeta$Genre), rcex = 0.5, extreme = 0.1, ccol = "blue")
    ```
  - 次に、prose のみに焦点化
    - data frame を prose のみに絞り込む。
    ```
    > prose = oldFrench[oldFrenchMeta$Genre =="prose"& !is.na(oldFrenchMeta$Year),]
    > proseinfo = oldFrenchMeta[oldFrenchMeta$Genre =="prose"& !is.na(oldFrenchMeta$Year),]
    ```
    *欠損値を含むベクトルで、欠損＝TRUE の値を返す。 "!"つきなので欠損＝FALSE を返す。

コメント [T1]: 筆者名、ジャンル、年号など

コメント [T2]: row color, ジャンルごとに色分け

コメント [T3]: column color,「タグ連鎖」の色づけ ccex はデフォルトで 1

コメント [T4]: 年号が与えられているもののみ抽出するため

- 年号の境目を作る。

```
> proseinfo$Period = as.factor(proseinfo$Year <= 1250)
```

＊ TRUE / FALSE を返す。

- CA 関数を割り当てる。

```
> corsup.fnc (prose.ca, bycol = F, supp = proseSup, font = 2, cex = 0.8,
+ labels = substr(rownames(proseSup), 1, 4))
```

- 作図する。

```
> plot(prose.ca, addcol = F, rcol = as.numeric(proseinfo$Period) + 1, rlabels = proseinfo$Year,
+ rcex = 0.7)
```

- さらに、年号不詳のデータも考慮する。
  - data frame に年齢不詳データを割り当てる。

```
> proseSup = oldFrench[oldFrenchMeta$Genre == "prose" & is.na (oldFrenchMeta$Year), ]
> corsup.fnc(prose.ca, bycol = F, supp = proseSup, font = 2, cex = 0.8,
+ labels = substr(rownames(proseSup), 1, 4))
```

- 年号がわかっているものと、不詳のものを別々に分析することが重要。

コメント [T5]: fragment number を取り除く関数。

コメント [T6]: default ではコラムの追加、ここでは、行の追加なので、bycol=F とする。

e.g.② data set, variationLijk, 接辞がついた語の頻度と発話者の変数との関係

- 発話者の変数：国、性別、教育レベル→8 水準
  - $\chi 2$ 乗検定をする。

```
> chisq.test(variationLijk)

  Pearson's Chi-squared test

data:  variationLijk
X-squared = 575.3482, df = 217, p-value < 2.2e-16
```

  - CA 関数の割り当て。

```
variationLijk.ca = corres.fnc(variationLijk)
```

  - 作図をする。(Figure 5.8, p. 135)

```
> plot(variationLijk.ca)
```

Corpus Statistics
2009/01/11
Baayen (2008: 128–138)

Toru Hitomi

5.1.4 Tables with distances: Multidimensional scaling

多次元尺度法(**Multidimensional scaling, MDS**):距離行列を考察し、データの構造を解析する。主成分分析と同じように data reduction 系の統計手法。コレスポンデンス分析は MDS の一つ。

- MDS の目的
  - "to provide a visual representation of the pattern of proximities (i.e., similarities or distances) among a set of objects." (Borgatti, 1997)[1]
  - "to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissimilarities (distances) between the investigated objects." (StatSoft, 2008)[2]

e.g. オランダ語の話し言葉コーパスにおいて、テクスト(話し方)の類似性と出生年・性別の関係性を調べたい。

コメント [T7]: cross-entropy(交差エントロピー)に基づいている。

- 165 人の話者
- data set, dutchSpeakersDistMeta; dutchSpeakerDist(), を用いる。
  - as.dist()を用いて、距離オブジェクトに変換する。

```
> dutchSpeakersDist.d = as.dist(dutchSpeakersDist)  #  デフォルトがユークリッド距離
```

	1	2	…	164
2	3.657			
3	3.748	3.738		
⋮	⋮	⋮	⋮	
165	3.755	3.890	…	3.419

  - 標準的な MDS 関数、cmdscale()、を割り当てる。

次元数

```
> dutchSpeakersDist.mds = cmdscale(dutchSpeakersDist.d, k = 3)
```

(出力はテキスト p. 136 を参照)

  - 発話者情報と合わせて、MDS の情報を data frame に組み込む。

```
> dat = data.frame(dutchSpeakersDist.mds, Sex = dutchSpeakersDistMeta$Sex,
+ Year = dutchSpeakersDistMeta$AgeYear, EduLevel = dutchSpeakersDistMeta$EduLevel)
> dat = dat[!is.na(dat$Year), ]
```

コメント [T8]: 年齢不詳のデータをはずすため。

  - 作図する。(Figure 5.9, p. 137)

```
> plot(dat$Year, dat$X1, xlab = "year of birth", ylab = "dimension 1", type = "p")
> lines(lowess(dat$Year, dat$X1))
> boxplot(dat$X3~dat$Sex, ylab = "dimension 3")
```

コメント [T9]: 平滑化曲線

  - プロットを解釈する。
    - 第一次元が年齢の効果を示している。
    - 第三次元に性別の差が示されている。

[1] http://www.analytictech.com/Borgatti/mds.htm
[2] http://www.statsoft.com/textbook/stmulsca.html

- 統計的な検定をする。
  - 第一次元と出生年の相関

```
> cor.test(dat$X1, dat$Year, method="sp")        # スピアマンの順位和相関係数
Spearman's rank correlation rho

data:  dat$X1 and dat$Year
S = 392556.7, p-value = 9.435e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho = 0.4561149
```

  - 第三次元における性別ごとの平均値の比較

```
> t.test(dat$X3~dat$Sex)       # T 検定
  Welch Two Sample t-test

  data:  dat$X3 by dat$Sex
  t = 2.1384, df = 155.156, p-value = 0.03405
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   0.008260503 0.208387229
  sample estimates:
  mean in group female    mean in group male
             0.04567817            -0.06264569
```

Corpus Statistics
2009/01/11
Baayen (2008: 128–138)

Toru Hitomi

資料 1

e.g 20 行 × 5 列の表→2 種類の距離行列

列の距離行列

	c1	c2	c3	c4	c5
c1					
c2					
c3					
c4					
c5					

行の距離行列

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2																				
3																				
4																				
5																				
6																				
7																				
8																				
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16																				
17																				
18																				
19																				
20																				