HW5_suggested_solutions

2025-05-11

```
swschool <- read.csv('/Users/Noppawee/Documents/NTHU/categorical_data/swiss_school.txt', sep="")
str(swschool)
```

```
## 'data.frame':    96 obs. of  3 variables:
##  $ level    : chr  "Aucune.formation" "Scolarité.obligatoire" "Formation.professionnelle" "Maturité"
##  $ community: chr  "Belmont" "Belmont" "Belmont" "Belmont" ...
##  $ Freq     : int  6 344 752 163 155 62 196 10 26 677 ...
```

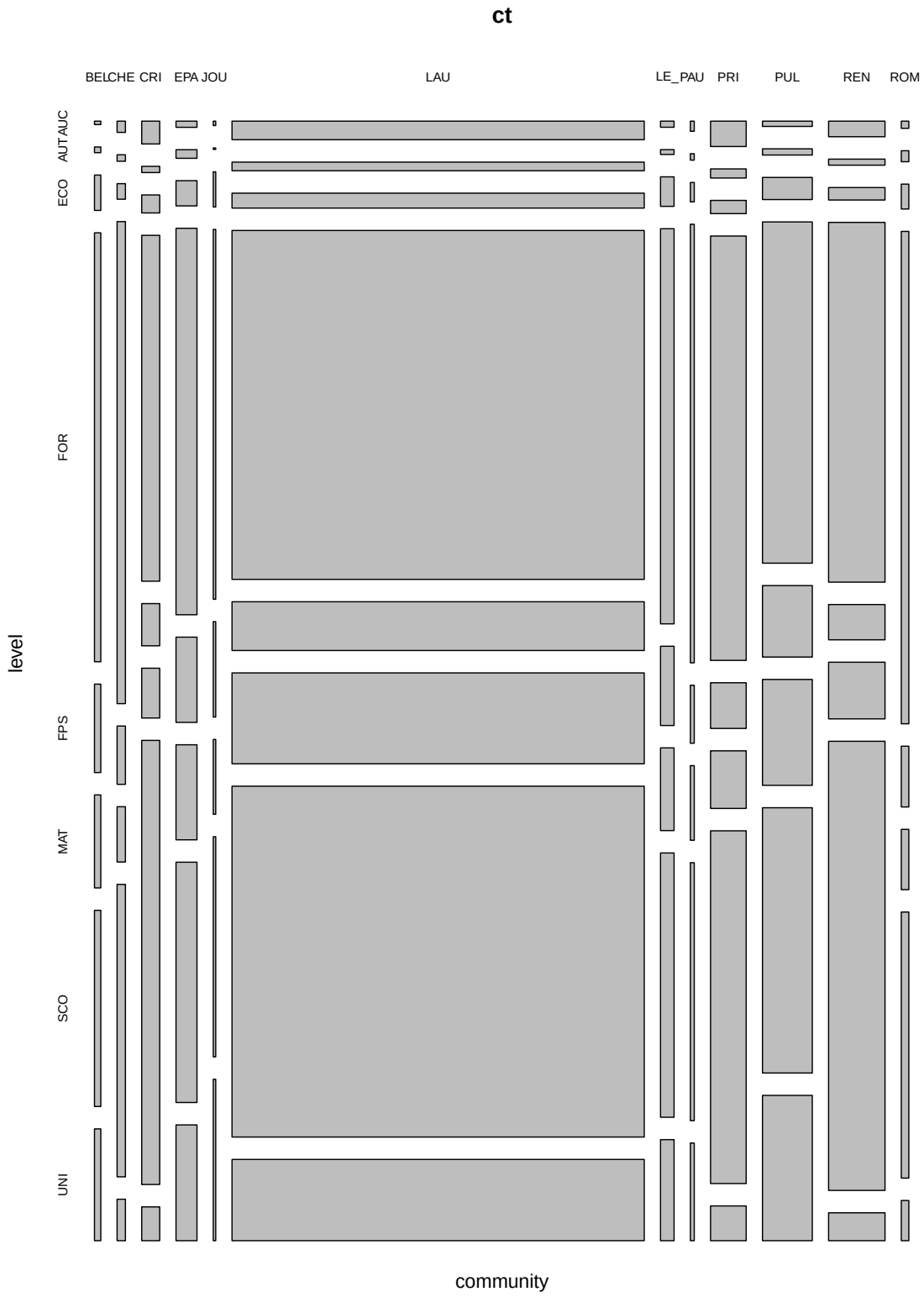
```
library(stringr)

data = swschool
for (i in 1:nrow(swschool)) {
if (data$level[i] == "Formation.professionnelle.supérieure") {
data$level[i] = "FPS"
}else{
data$level[i] = toupper(str_sub(data$level[i], 1,3))
}
data$community[i] = toupper(str_sub(data$community[i], 1,3))
}
```

1.

```
ct = xtabs(Freq ~ community + level, data)
ct
```

```
##           level
## community   AUC   AUT   ECO   FOR   FPS   MAT   SCO   UNI
##       BEL     6    10    62   752   155   163   344   196
##       CHE    26    15    36  1116   135   128   677    96
##       CRI   114    31    90  1729   211   249  2220   169
##       EPA    36    50   147  2253   497   554  1401   675
##       JOU     3     1    24   252    65    51   150   110
##       LAU  2126   990  1709 39941  5583 10405 40165  9302
##       LE_    23    18   111  1486   298   311   994   380
##       PAU    11     7    21   476    63    81   280   106
##       PRI   251    90   131  4200   452   570  3491   344
##       PUL    73    86   306  4721   989  1465  3670  2010
##       REN   244    95   195  5638   553   888  7039   437
##       ROM    15    23    52  1029   127   126   556    84
```

```r
mosaicplot(ct)
```



If the two variables were independent, the proportions of each education level across different communities would be similar. However, this plot shows that the distributions of education levels appear to vary

substantially among communities. For example, in the JOU or PUL communities, certain education levels (such as UNI) have notably high conditional probabilities. However, we still need to formally test for independence to make a valid statistical inference.

2.

We assume that the data was randomly collected.

This is a 2-dimensional contingency table data set, there are 12×8 cells and total count 169836. We could regard this as a multinomial sample with $12 \times 8 = 96$ categories and $N = 169836$, or we could regard each of the 96 counts as independent Poisson random variables.

We choose to model the 96 counts as independent Poisson random variables with cell (i, j) having mean μ_{ij}. We use a log link function to link μ_{ij} to the variables *community* and *education*.

If we use the model:

$$Y \sim community + education,$$

i.e.

$$\log(\mu_{ij}) = \alpha_0 + \sum_{i=1}^{11} \beta_i x_i + \sum_{j=1}^{7} \gamma_j y_j,$$

where $x_i's$ are dummy variables for community and $y_j's$ are dummy variables for *education*, then $\pi_{ij} = \pi_{i+}\pi_{+j}$ (see slide $5-5$), implying that *community* and *education* are independent.

If we use the model:

$$Y \sim community + education + community : education,$$

i.e.

$$\log(\mu_{ij}) = \alpha_0 + \sum_{i=1}^{11} \beta_i x_i + \sum_{j=1}^{7} \gamma_j y_j + \sum_{1 \le i \le 11, 1 \le j \le 7} \psi_{ij} x_i y_j,$$

then $\pi_{ij} \neq \pi_{i+}\pi_{+j}$, implying that *community* and *education* are not independent (see slide $5-6$).

The inferences about π_{ij} made by the Poisson log-linear framework will be the same as using a multinomial model (see slide $5-8$), and it is generally more convenient to use a Poisson log-linear model. However, bear in mind that inferences about the size of Poisson random variables may not be valid, this depends on how the data was collected.

3.

We want to test:

$$H_0 : education \text{ and } community \text{ are independent,}$$

$$H_1 : education \text{ and } community \text{ are not independent,}$$

i.e.

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j},$$
$$H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}.$$

Under H_0, $\log(\mu_{ij}) = \log(t\pi_{ij}) = \log(t\pi_{i+}\pi_{+j}) = \log(t) + \log(\pi_{i+}) + \log(\pi_{+j})$, therefore, the model

$$Y \sim community + education,$$

i.e.

$$\log(\mu_{ij}) = \alpha_0 + \sum_{i=1}^{11} \beta_i x_i + \sum_{j=1}^{7} \gamma_j y_j,$$

should fit the data. We fit this model:

```
indep_model = glm(Freq ~ community + level, data = data, family = poisson)
summary(indep_model)
```

```
##
## Call:
## glm(formula = Freq ~ community + level, family = poisson, data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.37079    0.03046 110.648  < 2e-16 ***
## communityCHE  0.27801    0.03227   8.616  < 2e-16 ***
## communityCRI  1.04778    0.02829  37.040  < 2e-16 ***
## communityEPA  1.20154    0.02776  43.284  < 2e-16 ***
## communityJOU -0.94514    0.04601 -20.543  < 2e-16 ***
## communityLAU  4.17894    0.02453 170.394  < 2e-16 ***
## communityLE_  0.76321    0.02947  25.896  < 2e-16 ***
## communityPAU -0.47953    0.03936 -12.183  < 2e-16 ***
## communityPRI  1.73080    0.02641  65.542  < 2e-16 ***
## communityPUL  2.06572    0.02584  79.956  < 2e-16 ***
## communityREN  2.19042    0.02566  85.347  < 2e-16 ***
## communityROM  0.17558    0.03301   5.320 1.04e-07 ***
## levelAUT     -0.72648    0.03237 -22.444  < 2e-16 ***
## levelECO     -0.01514    0.02623  -0.577    0.564
## levelFOR      3.07818    0.01890 162.857  < 2e-16 ***
## levelFPS      1.13703    0.02124  53.536  < 2e-16 ***
## levelMAT      1.63313    0.02020  80.829  < 2e-16 ***
## levelSCO      3.03634    0.01892 160.492  < 2e-16 ***
## levelUNI      1.55822    0.02033  76.635  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 587141.4  on 95  degrees of freedom
## Residual deviance:   5409.6  on 77  degrees of freedom
## AIC: 6152.7
##
## Number of Fisher Scoring iterations: 4
```

The model we fit has deviance 5409.6 on 77 degrees of freedom. The deviance for this model has asymptotic distribution χ^2_{77}, therefore, the p-value for this hypothesis test is approximately:

$$P(\chi^2_{77} \geq 5409.6) \approx 0,$$

we therefore reject H_0 and conclude that schooling and community are not independent. The model does not fit the data even though most of the predictors are highly significant.

4.

```
ct = xtabs(Freq ~ community + level, data)
summary(ct)
```

```
## Call: xtabs(formula = Freq ~ community + level, data = data)
## Number of cases in table: 169836
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 5260, df = 77, p-value = 0
```

The command *ct* also conducts the hypothesis test:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j},$$
$$H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}.$$

This test is based on the the χ^2 test statistic, which is $\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, this test is therefore different from the test in question 3, which is a deviance-based test (i.e. a likelihood-ratio test). The two tests lead to the same conclusion because for large data sets, Pearson's χ^2 is approximately equal to the deviance (see slide $3-33$).

5.

```
round(xtabs(residuals(indep_model) ~ community+level, data),3)
```

```
##           level
## community     AUC     AUT     ECO     FOR     FPS     MAT     SCO     UNI
##       BEL  -5.221  -1.146   5.385   4.631   6.123   1.130 -11.600   4.619
##       CHE  -2.131  -0.861  -0.303   9.257   1.361  -5.239  -4.482  -7.050
##       CRI   3.221  -1.502   0.900  -1.735  -3.063  -9.253  11.324 -12.810
##       EPA  -7.095   0.463   4.899   3.261  10.276   2.581 -14.492   9.381
##       JOU  -2.942  -2.354   3.335   0.405   4.476  -0.926  -5.977   6.716
##       LAU   5.081   2.314  -3.817  -6.582  -4.473   6.777   2.935   2.883
##       LE_  -5.738  -2.401   5.666   3.480   6.868  -0.484  -8.865   4.642
##       PAU  -1.783  -0.601   0.751   4.140   0.894  -1.195  -5.151   2.127
##       PRI   6.273   1.159  -2.506  10.289  -2.712  -9.933   1.179 -17.584
##       PUL -12.081  -2.477   5.033  -3.808   9.644   8.122 -16.789  24.872
##       REN  -1.011  -2.872  -3.995  -0.158  -9.617 -12.954  21.037 -26.248
##       ROM  -3.772   1.438   2.831   9.508   1.765  -4.086  -6.458  -6.954
```

The 5 largest (in absolute value) residuals have values: $-26.248, 24.872, 21.037, -17.584, -16.789$, they are the residuals for $REN-UNI, PUL-UNI, REN-SCO, PRI-UNI$ and $PUL-SCO$, respectively.

A residual that is large in absolute value indicates that that cell has a lot more (if sign is positive) or a lot fewer (if sign is negative) counts than would be expected under independence between the two variables.

For this data set:

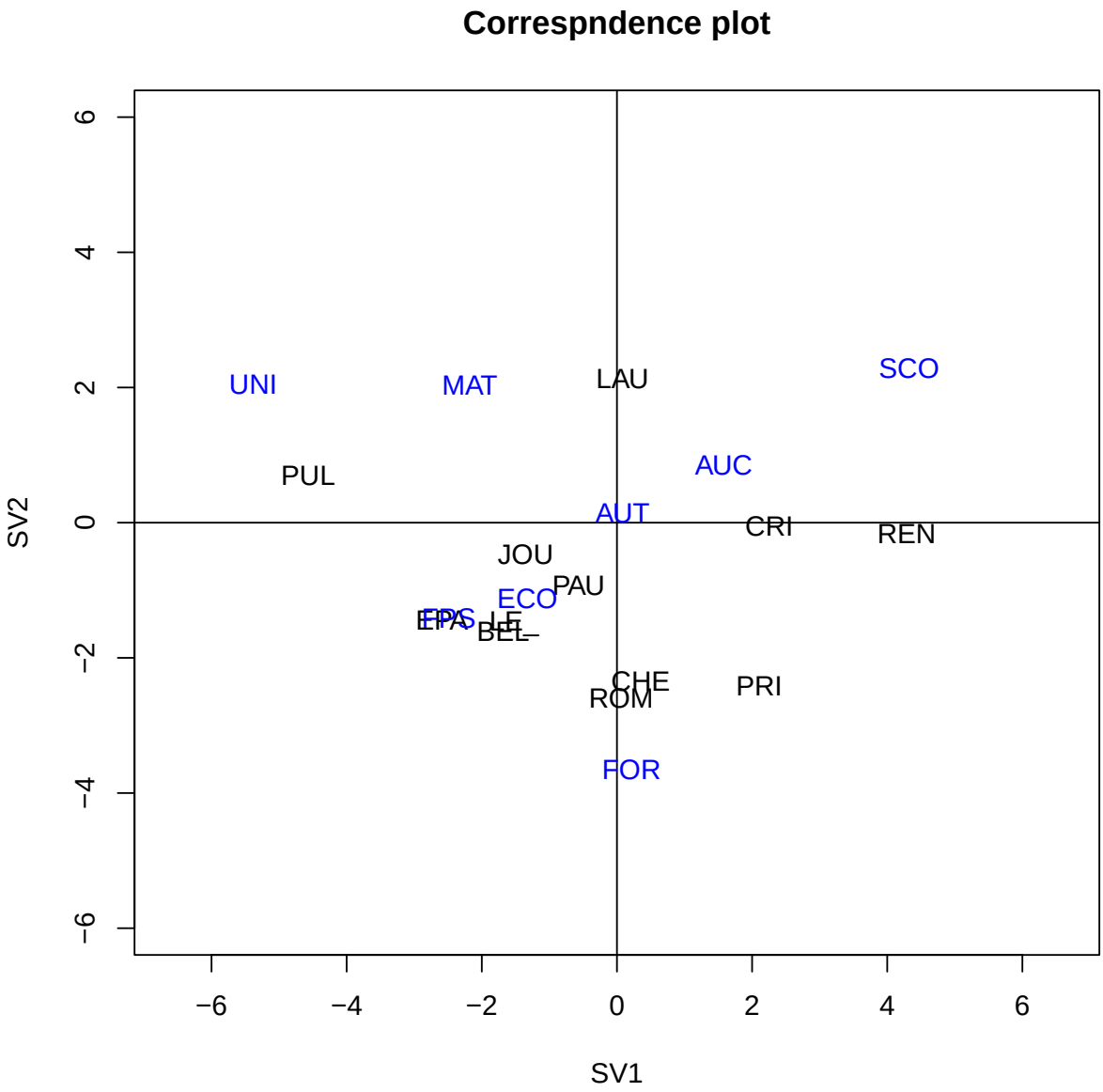
- the number of university-educated people from Renens is much much lower than would be expected under independence
- the number of university-educated people from Pully is much much higher than would be expected under independence
- the number of people with the education level of 'Scolarité obligatoire' from Renens is much much higher than would be expected under independence
- the number of university-educated people from Prilly is much much lower than would be expected under independence
- the number of people with the education level of 'Scolarité obligatoire' from Pully is much much lower than would be expected under independence

6.

The plot below reveals the following information:

- the distribution of the level of education Autre is similar to the overall education distribution, Autre is not particularly associated with any community (because it is quite close to the origin)
- the levels of education Scolarité obligatoire, Université / Haute école, Formation professionnelle all have distributions quite different from the typical education level distribution (because their distances from the origin is quite large)
- Université / Haute école and Pully are quite close to each other and far from the origin, therefore, there are many more people in Pully with highest education level Université / Haute école than there would be if community and education level were independent
- some pairs are negatively associated, they are diametrically opposite in the correspondence analysis diagram and (somewhat) far from the origin, these pairs are: Lausanne-Formation professionnelle, Belmont-Scolarité obligatoire, Université / Haute école - Prilly, so the entries in the contingency table corresponding to these pairs are lower than what would be expected if community and education level were independent
- the following communities have similar patterns of association with the education level, they could be merged: Jouxtens-Paudex, Le Mont-Belmont-Epalinges, and Cheseaux-Romanel, we can infer this from their proximities to each other in the correspondence analysis diagram
- Ecole professionnelle supérieure and Formation professionnelle supérieure have similar patterns of association with level of education, we infer this from how close they are to each other in the diagram

```
z <- xtabs(residuals(indep_model, type="pearson") ~ community+level, data)
svdz <- svd(z,2,2)
leftsv <- svdz$u %*% diag(sqrt(svdz$d[1:2]))
rightsv <- svdz$v %*% diag(sqrt(svdz$d[1:2]))
ll <- 1.1*max(abs(rightsv), abs(leftsv))
plot(rbind(leftsv,rightsv),asp=1,xlim=c(-ll,ll),ylim=c(-ll,ll),
main = "Correspndence plot",xlab="SV1",ylab="SV2",type="n")
abline(h=0,v=0)
text(leftsv,dimnames(z)[[1]], col = "black")
text(rightsv,dimnames(z)[[2]], col = "blue")
```


Correspndence plot
SV2
SV1
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
UNI
MAT
LAU
SCO
PUL
AUC
AUT
CRI
REN
JOU
PAU
ECO
EPS
LE
BEL
CHE
ROM
PRI
FOR