Discrete Data Analysis

A Friendly Guide to Visualising Categorical Data

Julian Hatwell

04 July, 2019

Introduction

About Me

- Over 15 years in enterprise systems, management information systems and BI in the for-profit education sector, UK and Singapore
- MSc (Distinction) Business Intelligence, Birmingham City University. Master's Dissertation: *An Association Rules Based Method for Imputing Missing Likert Scale Data*
- Research to PhD (in progress): *Designing Explanation Systems for Decision Forests*, Data Analytics and Artifical Intelligence research group, Birmingham City University

Academic Analytics

Monitor, Evaluate and Optimise

- Typical BI analysis of sales and marketing (student recruitment), financial, HR, operations, etc
- **student life-cycle**: recruitment to alumni services
- class room utilisation and exam planning
- lecturer management (part-time and adjunct)

Credits

- **Michael Friendly** is a pioneer in this field and has contributed to the development of modules and libraries for SAS and R.
- This book is *the bible*.
- Shorter, valuable tutorial on these topics:
  - **R> library("vcd")**
  - **R> vignette("vcd")**
- This link https://www.youtube.com/watch?v=qfNsoc7Tf60 for an in depth lecture, by Michael Friendly.

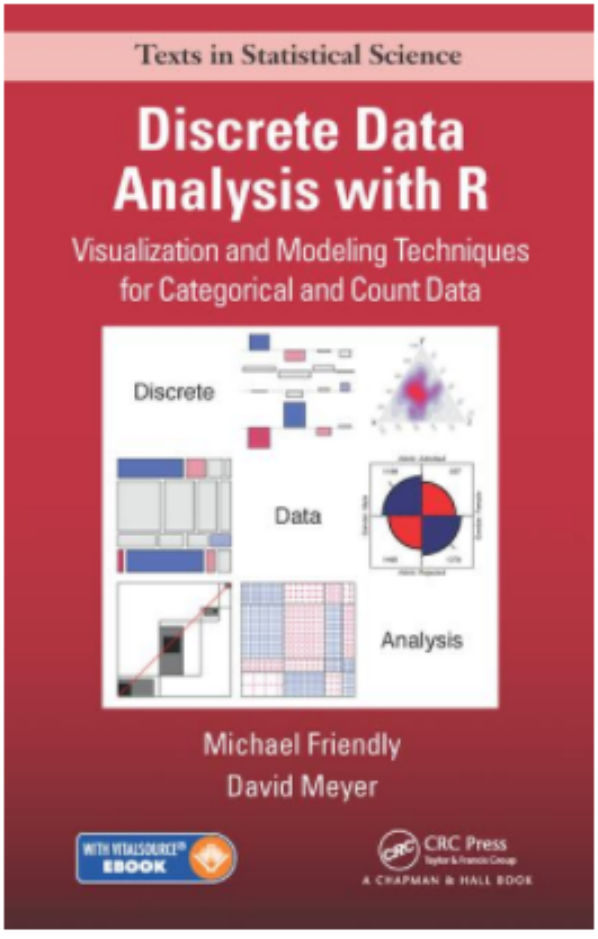


Figure 1: Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data

Visualising Categorical Data

Rigorous statistical analysis of categorical data is not only supported by visual tools, it is best performed visually.

The main theme of Michael Friendly's work.

Case Study

Analysis of Negative Trend in Student Outcomes

This is a synthetic data set, designed to replicate and exaggerate some problems found in a real life scenario

Academic life-cycle data for undergraduates. 2014 and 2015 intakes, graduating in 2017 and 2018 respectively. In 2018, rates of non-graduation (fails and withdrawals) were found to be significantly higher than the previous year.

An analysis was conducted to determine the factors associated with this trend.

Academic Data

```
## Total Students

## [1] 7193

##    year       faculty             hqual
##  2014:3192   fin :2129   dip       :2784
##  2015:4001   law :2188   dip-other: 616
##              mgmt:2876   hs        :2404
##                          hs-equiv :1389

##   outcome         grad
##  dist : 423   FALSE:1186
##  fail : 793   TRUE :6007
##  merit:1212
##  pass :4372
##  wthdr: 393
```

Limitations of Working with Tables Directly

```
##         year
## grad     2014 2015
##   FALSE  481  705
##   TRUE  2711 3296
```

Limitations of Working with Tables Directly

```
##         year
## grad     2014 2015
##   FALSE  481  705
##   TRUE  2711 3296
```

```
loddsratio(with(sts, table(grad, year)))
```

```
## log odds ratios for grad and year
##
## [1] -0.1869385
```

```
confint(loddsratio(with(sts, table(grad, year))))
```

```
##                                2.5 %     97.5 %
## FALSE:TRUE/2014:2015 -0.3134998 -0.0603772
```

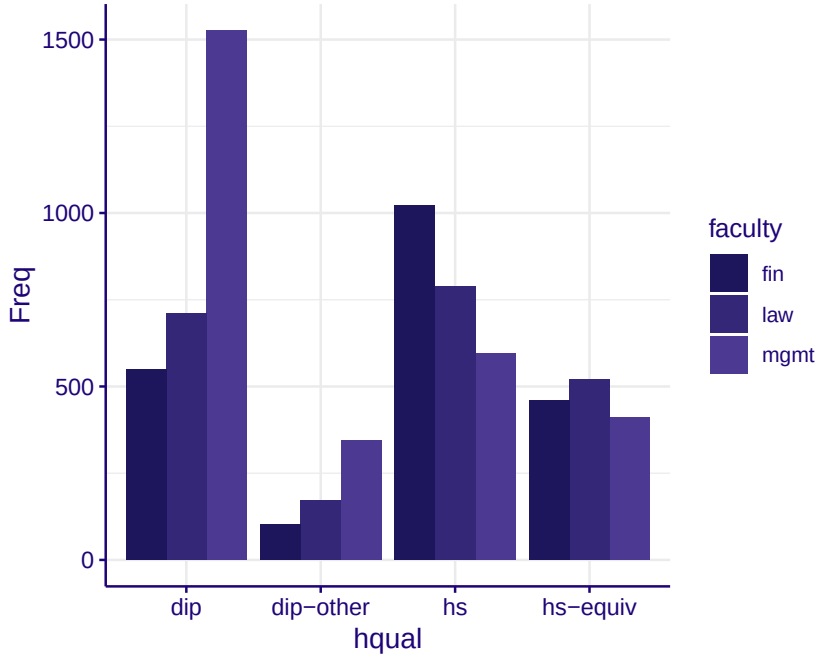
Limitations of Working with Tables Directly

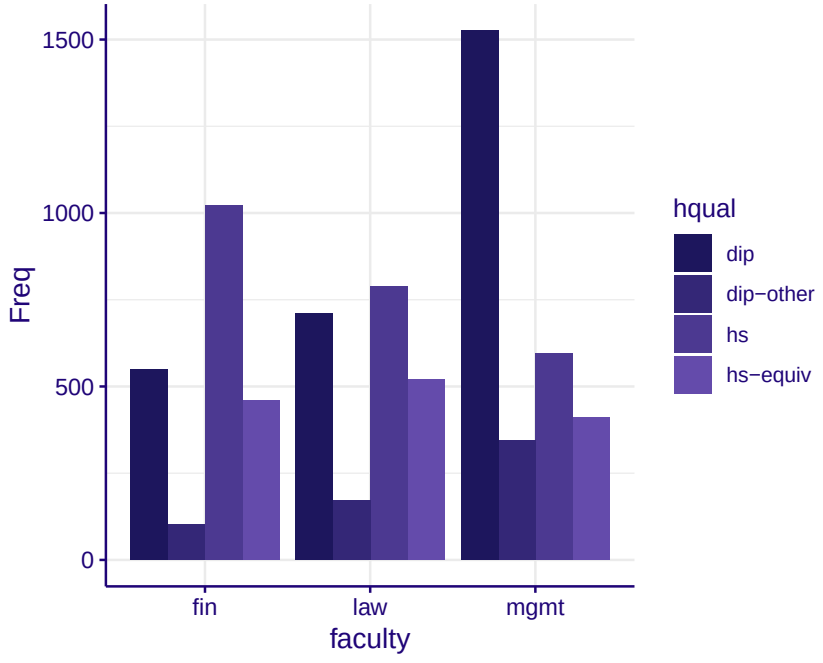
```
##         hqual
## faculty  dip dip-other   hs hs-equiv
##    fin   548       101 1021      459
##    law   710       171  788      519
##    mgmt 1526       344  595      411
```

Naive Approach: Barplots

Which is the right grouping variable?

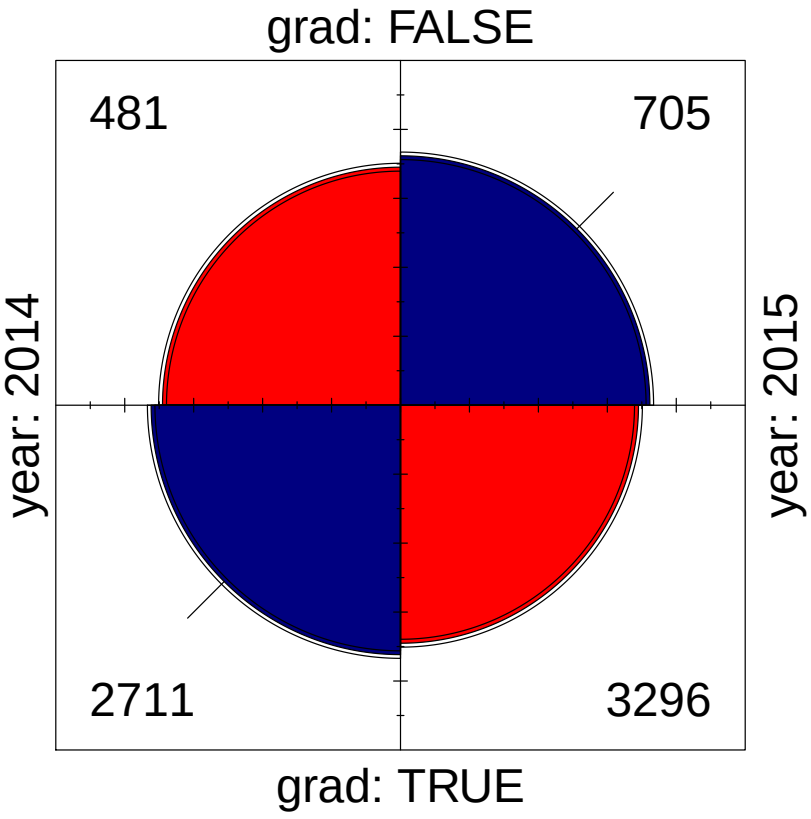
Can we succinctly describe the relationship?





Fourfold Plots

```
fourfold(with(sts, table(grad, year)))
```

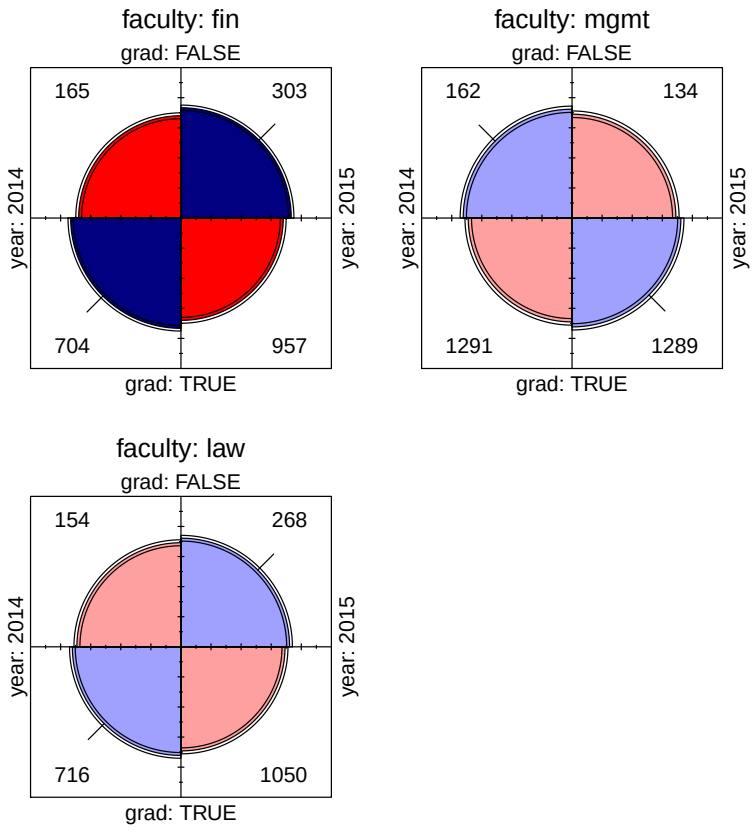


A specialised plot for 2×2 contingency tables which exposes the log odds ratio. A significant difference in the ratios shows up as non-overlapping CI.

Fourfold Plots

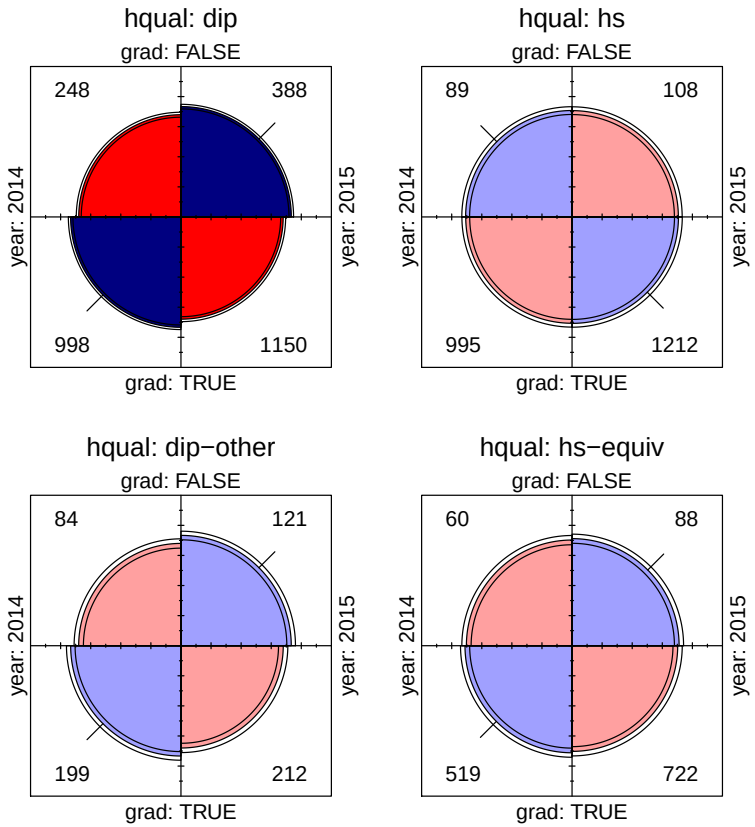
Fourfold plots are automatically stratified over the levels of a third variable, when given a $2 \times 2 \times k$ table.

```
fourfold(with(sts, table(grad, year, faculty)))
```



Fourfold Plots

```
fourfold(with(sts, table(grad, year, hqual)))
```



Correspondence Analysis

An unsupervised/clustering method. R Package "ca". per level associations.

Multiple/Joint ca, the mjca() function, operates simultaneously on two or more variables. mjca is harder to interpret and not shown here.

Simple ca, the ca() function, operates on just two variables.

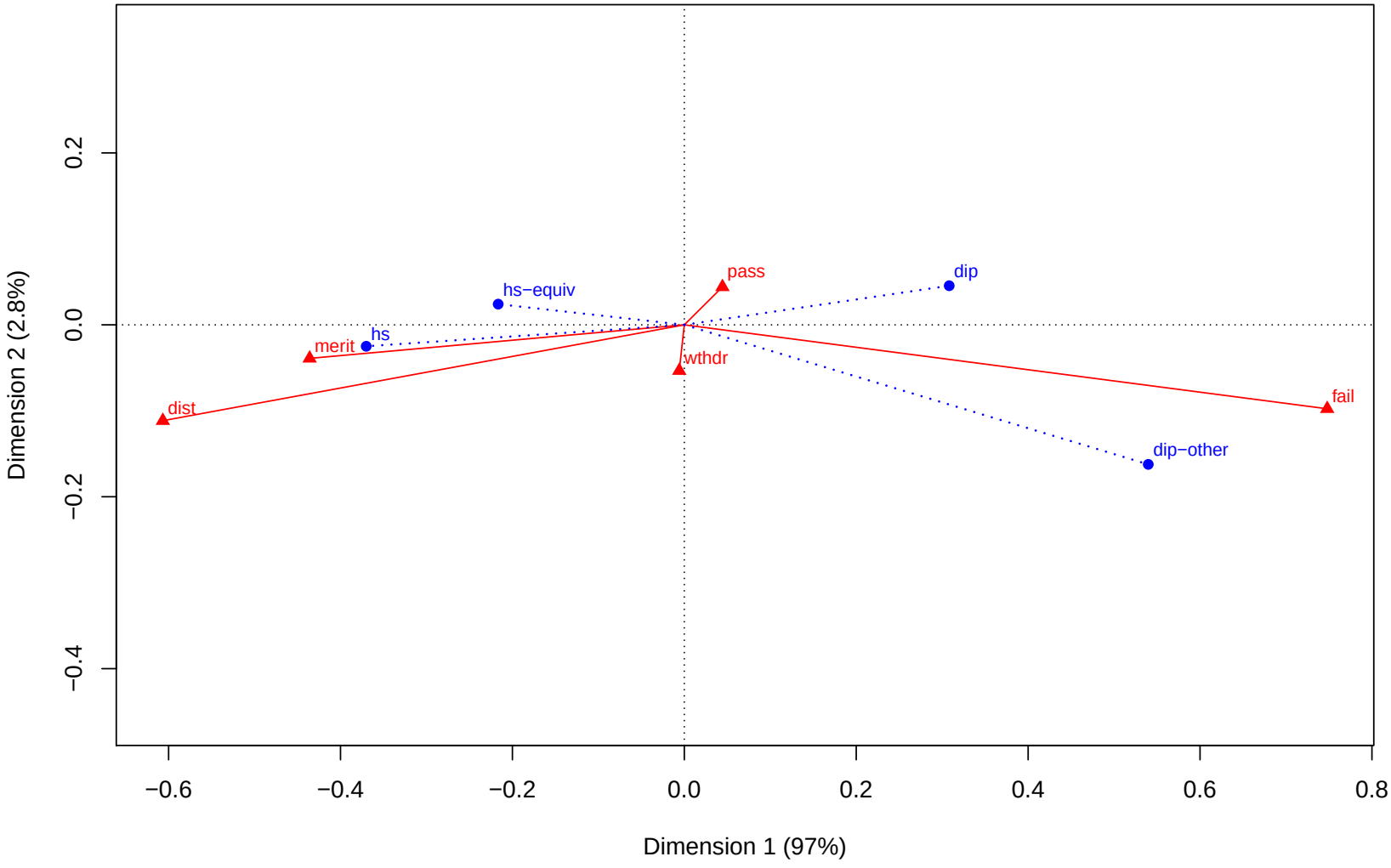
```
sts_ca_hq_out <- ca(with(sts, table(hqual, outcome)))

sts_ca_fc_out <- ca(with(sts, table(faculty, outcome)))
```

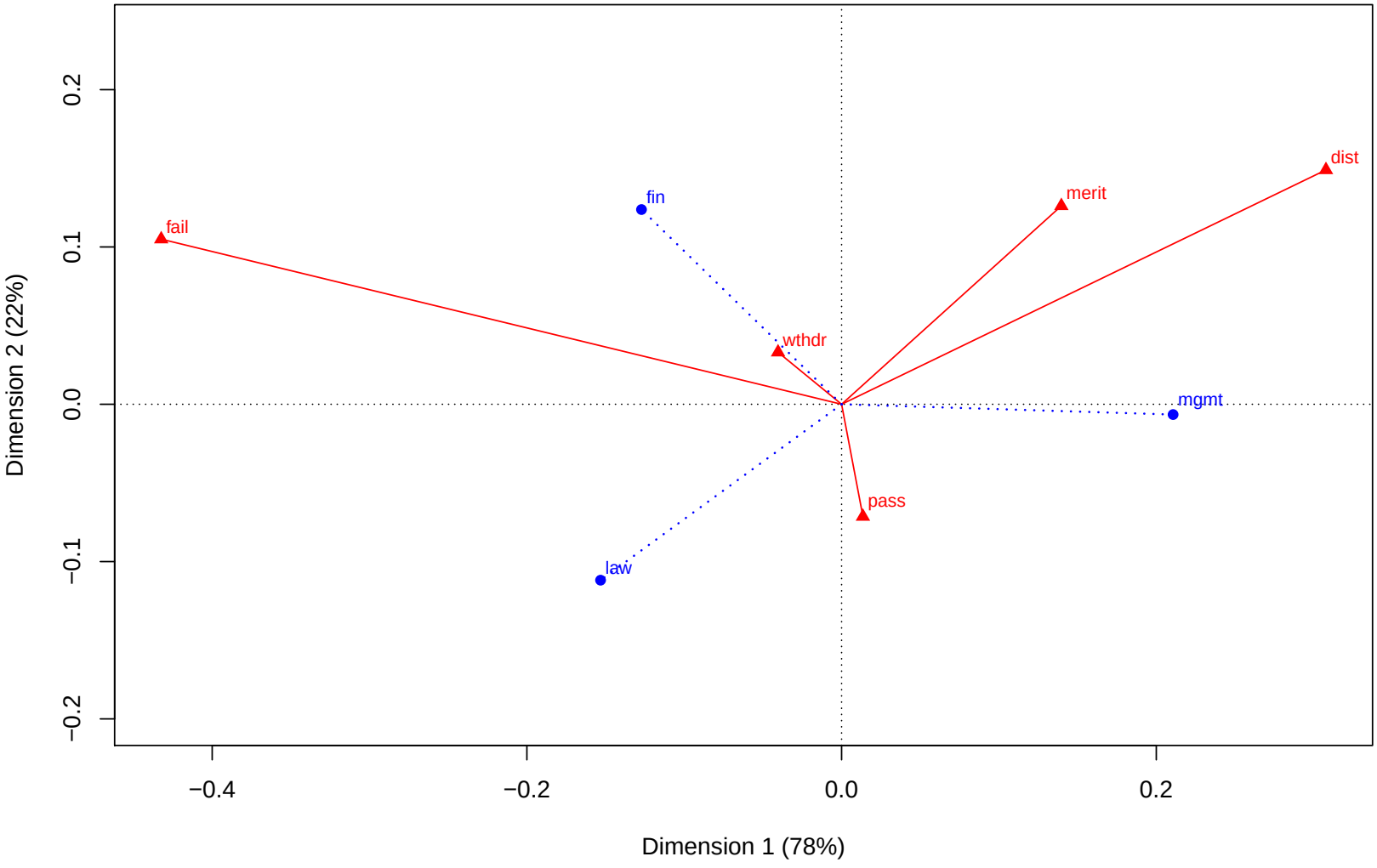
Correspondence Analysis Plot

```
# Generate the plot
res.ca <- plot(x)
# add some segments from the origin to make things clearer
segments(0, 0, res.ca$cols[, 1]
         , res.ca$cols[, 2]
         , col = "red", lwd = 1)
segments(0, 0, res.ca$rows[, 1]
         , res.ca$rows[, 2]
         , col = "blue", lwd = 1.5, lty = 3)
```

Correspondence Analysis



Correspondence Analysis



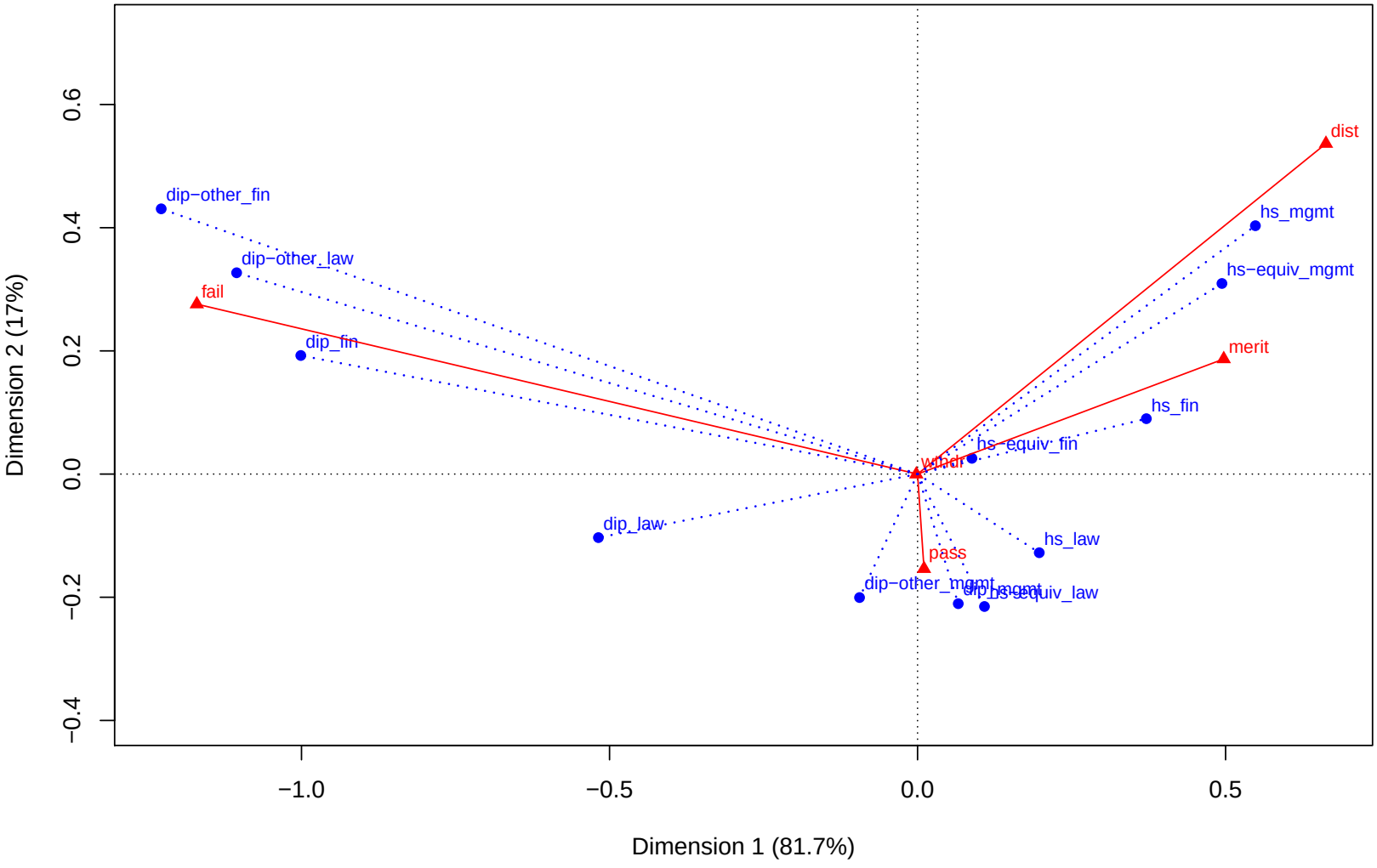
Extending Simple Correspondence Analysis

A pivot table creates rows for combos of > 1 variable.

```
structable(outcome ~ hqual + faculty, data = sts)
```

```
##                    outcome dist fail merit pass wthdr
## hqual     faculty
## dip        fin               2  227    10  280    29
##            law               6  168    32  463    41
##            mgmt             54   95   224 1077    76
## dip-other  fin               0   52     5   39     5
##            law               2   79     4   76    10
##            mgmt             10   36    36  239    23
## hs         fin              92   22   274  566    67
##            law              45   30   146  528    39
##            mgmt            108    8   176  272    31
## hs-equiv   fin              28   42    94  271    24
##            law              17   26    84  363    29
##            mgmt             59    8   127  198    19
```

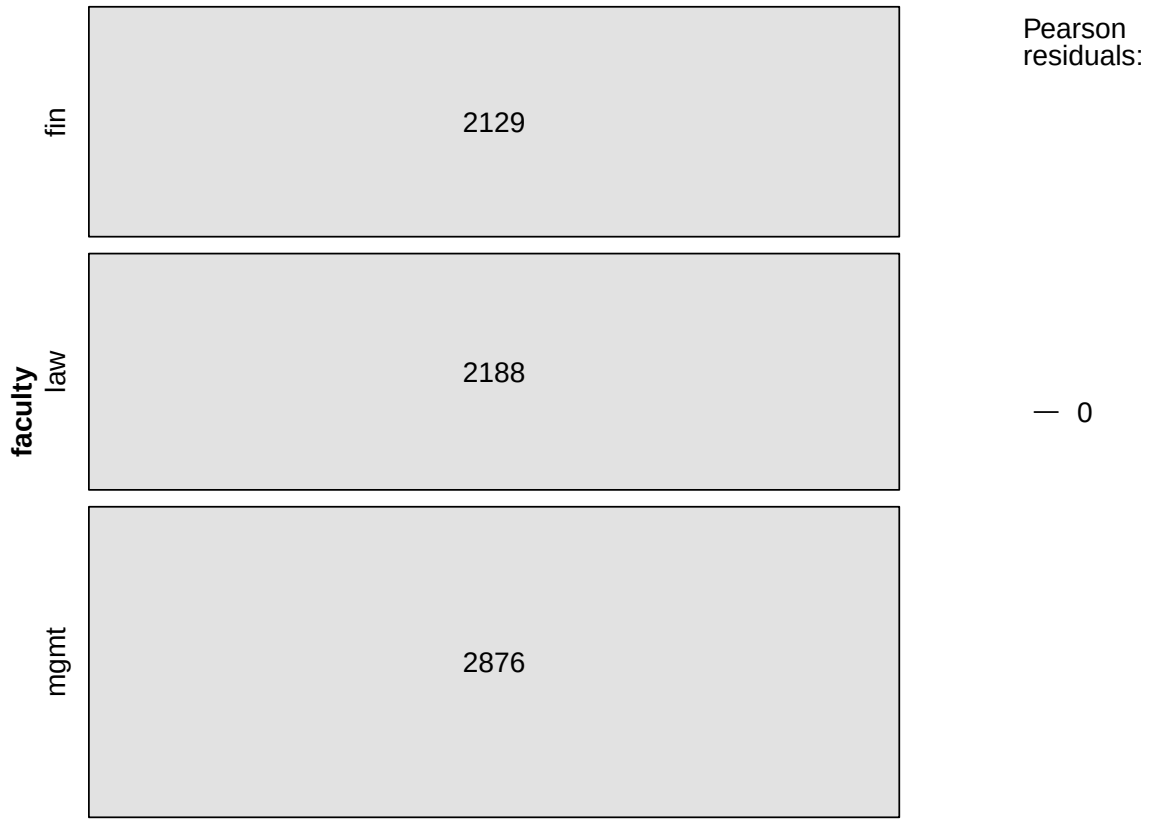
Correspondence Analysis



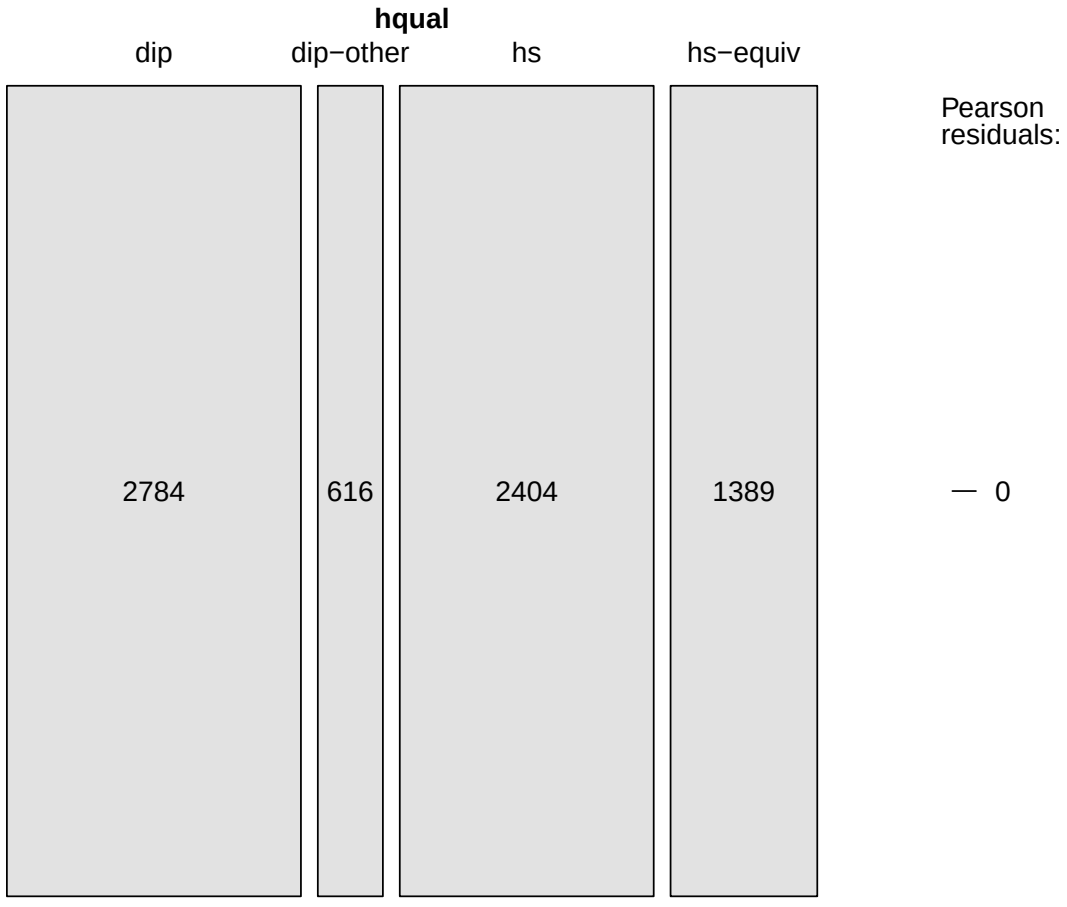
Introducing the Strucplot Framework

- Purpose built for handling tables - no dependent/independent variables.
- Respects the table structure given.
- Table is represented as rectangular tiles.
- Tile area $\propto$ cell count.
- Several Variants: mosaic, sieve, tile, assoc, spine, doubledecker
- Mosaic is the most versatile and supports higher dimensional data and models.
- Fourfold is also part of the strucplot framework but uniquely specialised for 2×2 contingency tables and log odds ratio tests.

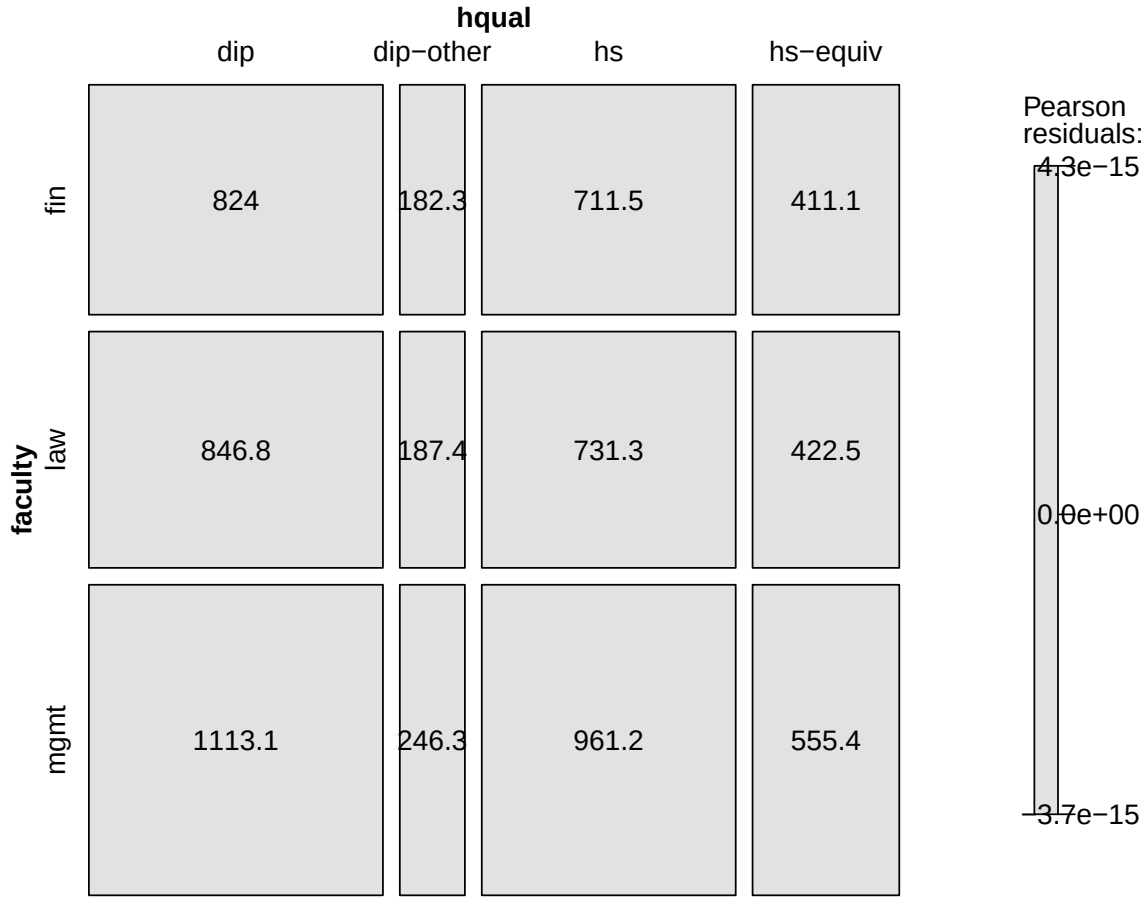
faculty Marginal Totals



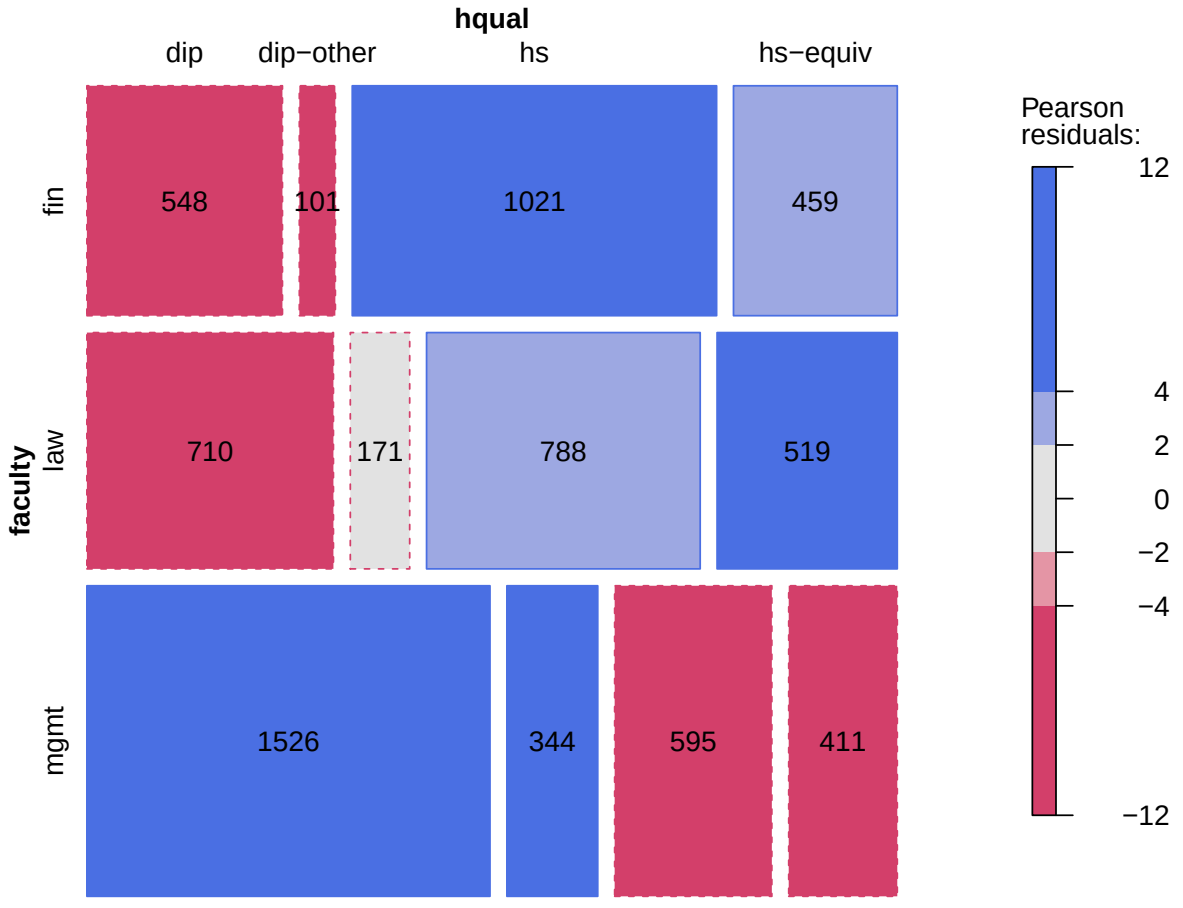
hqual Marginal Totals



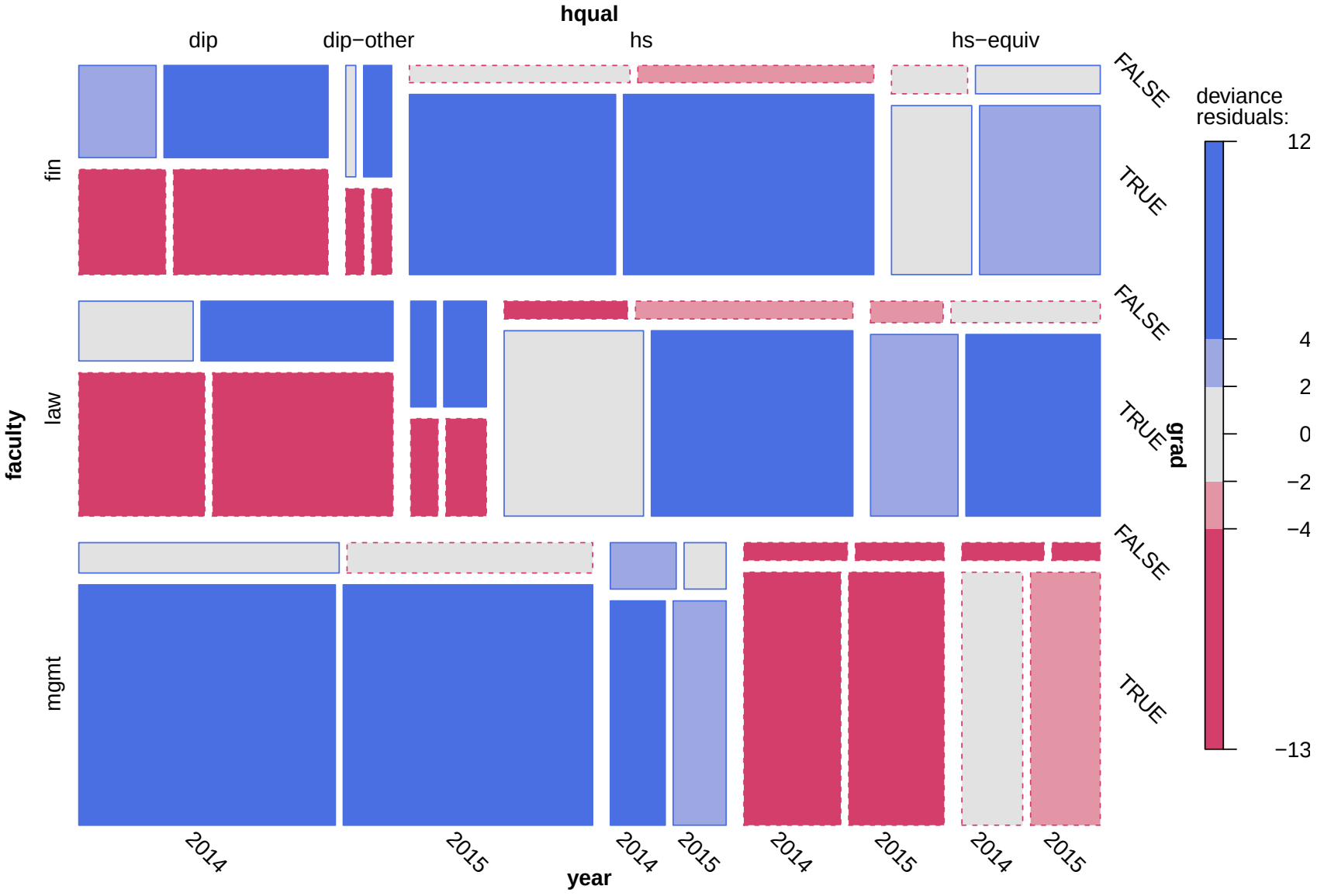
Expected Frequencies (Assume Independence)



Observed Frequencies



Scaling to Higher Dimensions



Example Code

```
mosaic(my_table
       # shading function
       , gp = shading_Friendly2
       # tile spacing
       , spacing = spacing_equal(sp = unit(0.4, "lines"))
       # label positioning
       , rot_labels = c(0, -45, -45, 90)
       , rot_varnames = c(0, -90, 0, 90)
       , offset_labels = c(0, 0.5, 0, 0)
       , offset_varnames = c(0, 1, 0, 0.5))
```

Loglinear Models

Two discrete variables:

$$A = \{A_1, \ \ldots, \ A_I\}, \ B = \{B_1, \ \ldots, \ B_J\}$$

Represented as an $I \times J$ contingency table. Under independence, each cell count n_{ij} in the table is assumed to be a Poisson distributed random variable:

$$n_{ij} \sim Pois(\frac{n_{i+} \cdot n_{+j}}{n_{++}})$$

$n_{i+} = \text{row total}, \ n_{+j} = \text{col total}, \ n_{++} = \text{table total}$

Loglinear Models

Rearranging, we get a model that is linear in all the logs:

$$\log(\frac{n_{i+} \cdot n_{+j}}{n_{++}}) = \log(n_{i+}) + \log(n_{+j}) - \log(n_{++})$$

This is usually represented as:

$$\mu + \lambda_i^A + \lambda_j^B$$

where μ is in "intercept" term that is a property of the total count and the λ terms are the "main" effects.

Under independence, the main effects can be calculated directly from the marginal totals.

Loglinear Models

Deviation from independence implies an interaction term:

$$\mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

To allow for estimation, we constrain:
$\sum_{i=1}^{I} \lambda_i^A = 0,\ \sum_{j=1}^{J} \lambda_i^B = 0,\ \sum_{i=1}^{I} \lambda_{ij}^{AB} = \sum_{j=1}^{J} \lambda_{ij}^{AB} = 0.$

This scales to more variables as follows:

$$\mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC} + \lambda_{ijk}^{ABC}$$

The model with all parameters (one for each table cell) included is referred to as the saturated model - we can always fit the data, but learn nothing.

Loglinear Models

When modelling, we seek the most parsimious, best fit. Find a model that is:

- as "good" as the saturated model - non-significant residuals
- does not use all the parameters / degrees of freedom

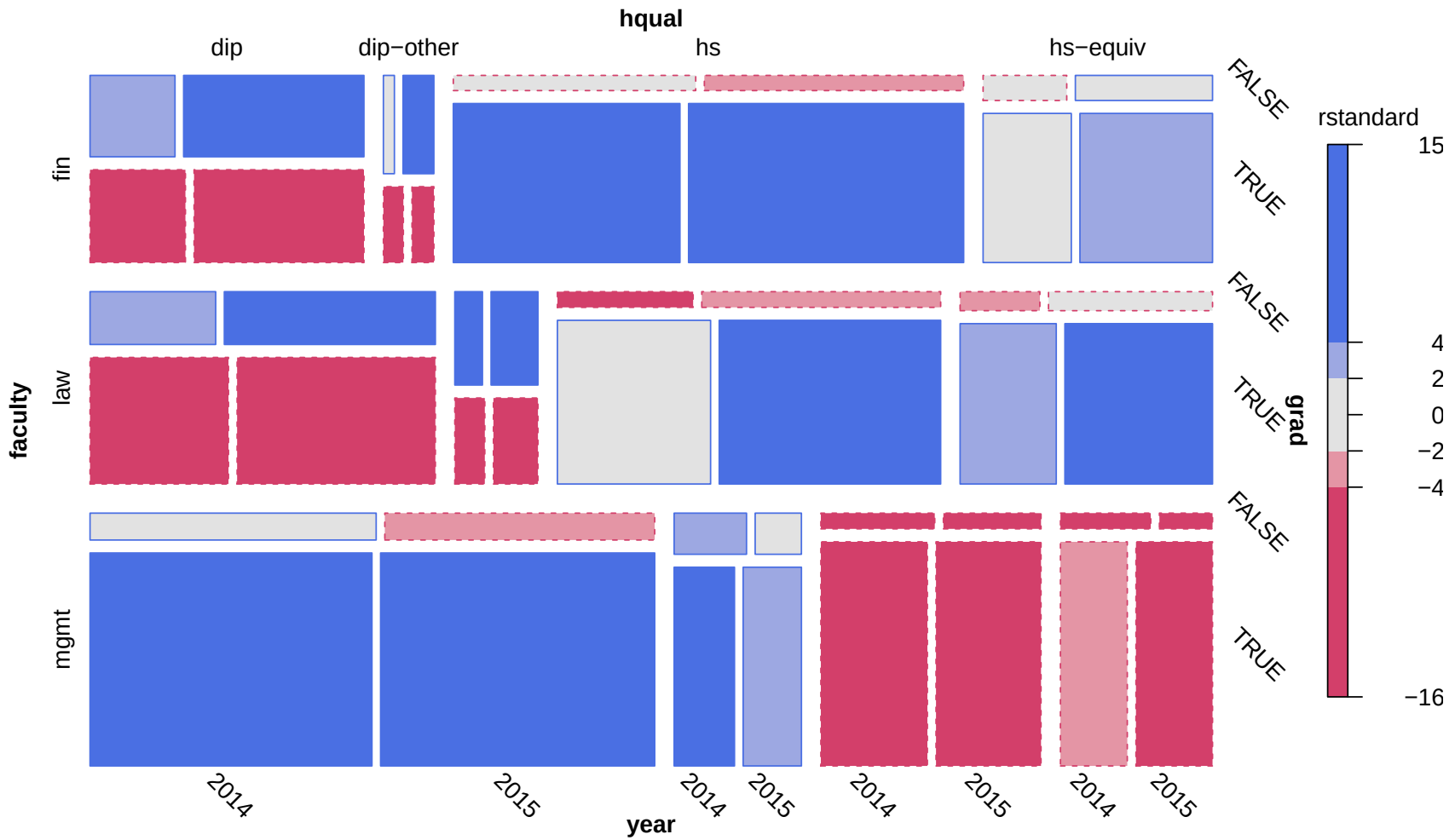
Recommended Methodology:

- generalised linear models; glm(family = poisson)
- add interaction terms one at a time and hierarchically, according to domain knowledge
- visual inspection of mosaic plots for significant residuals
- repeat until residuals are non-significant and the mosaic is "cleaned up"

Diagnostics and model selection with using anova and AIC/BIC.

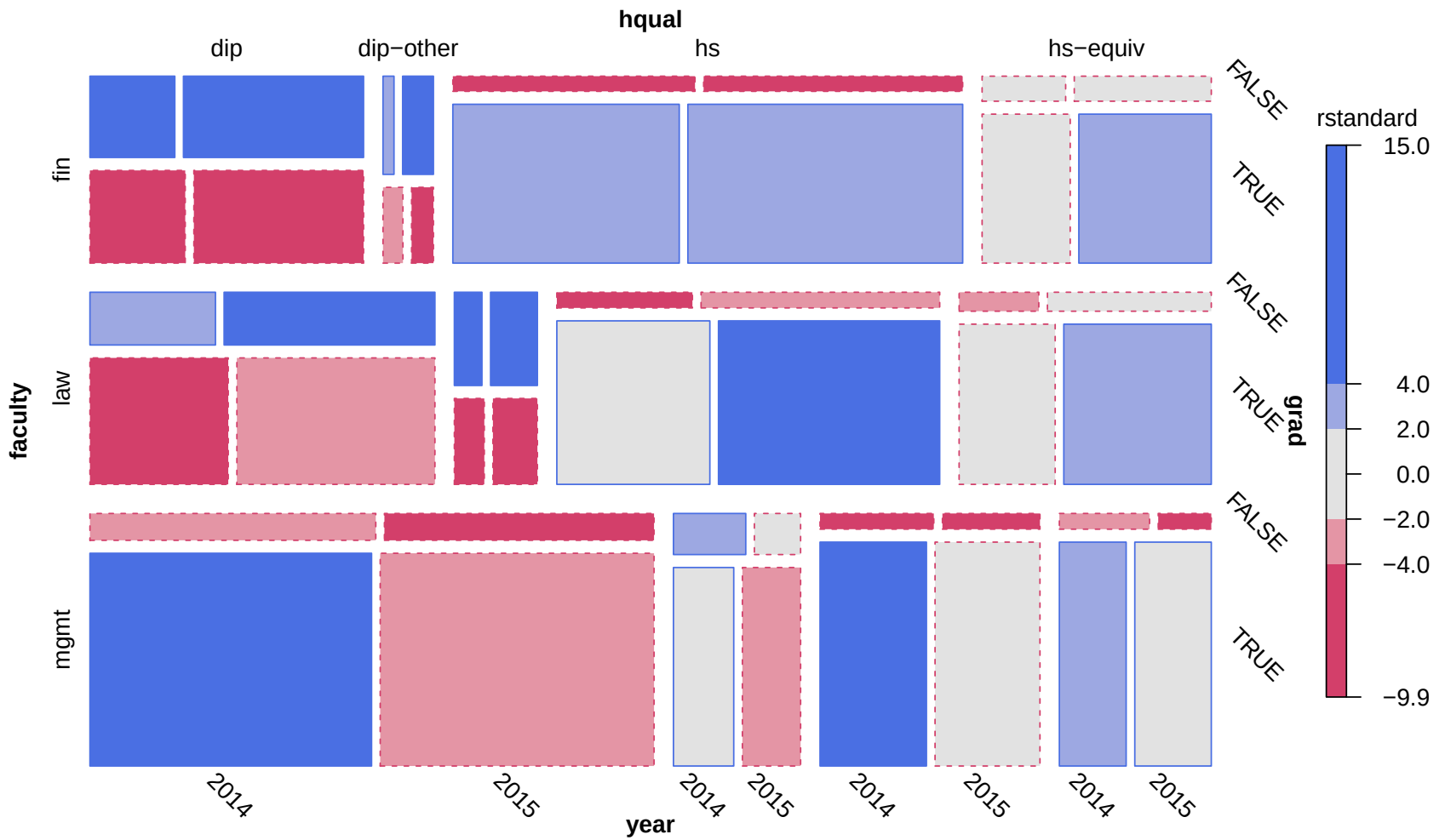
Null Model [F] [H] [G] [Y]

```
glm0 <- glm(Freq~faculty +
              hqual + grad + year
            , data = sts_df, family = poisson)
```



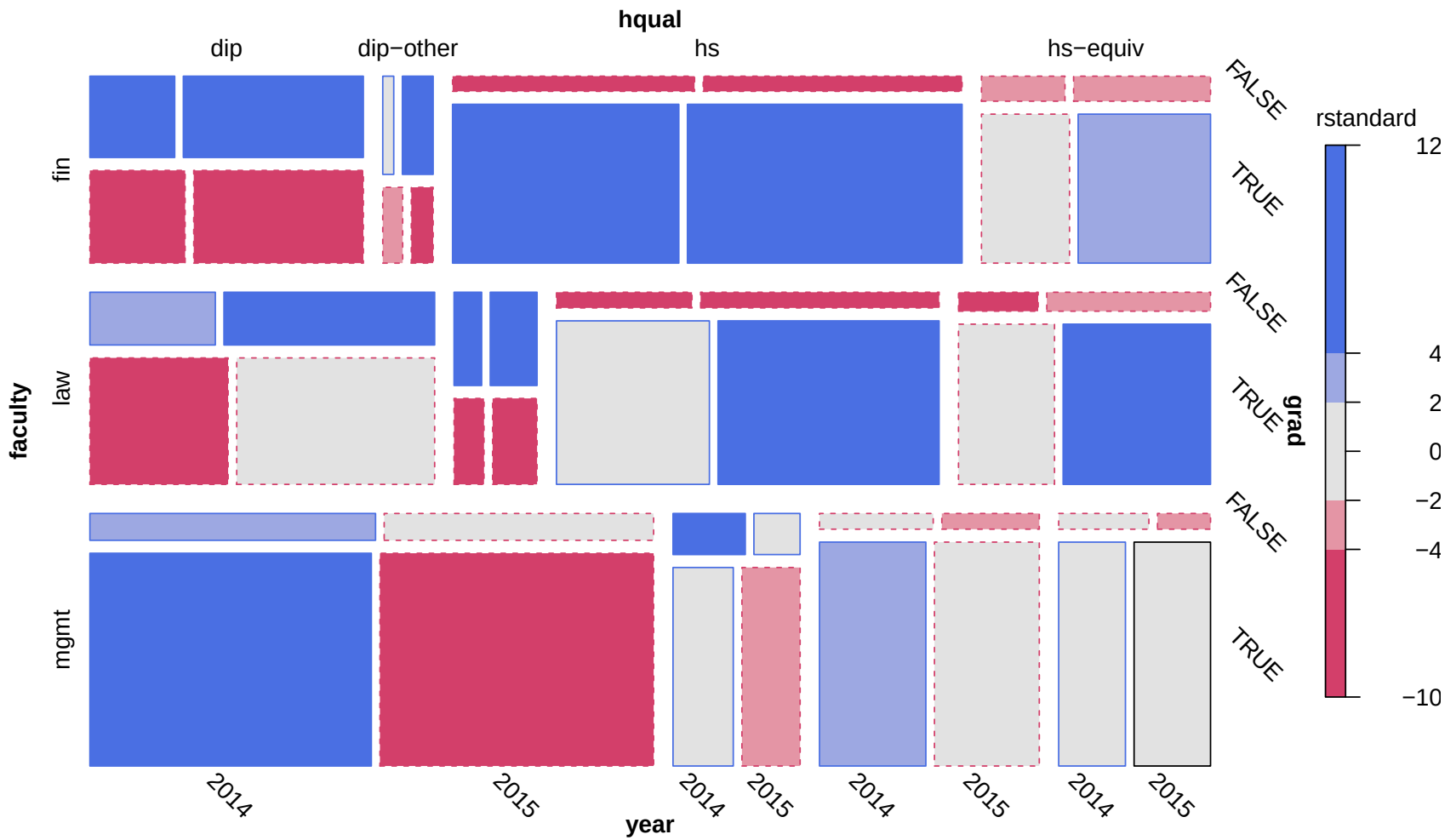
Model 1 [FH] [G] [Y]

```
glm1 <- glm(Freq~faculty *
              hqual + grad + year
            , data = sts_df, family = poisson)
```



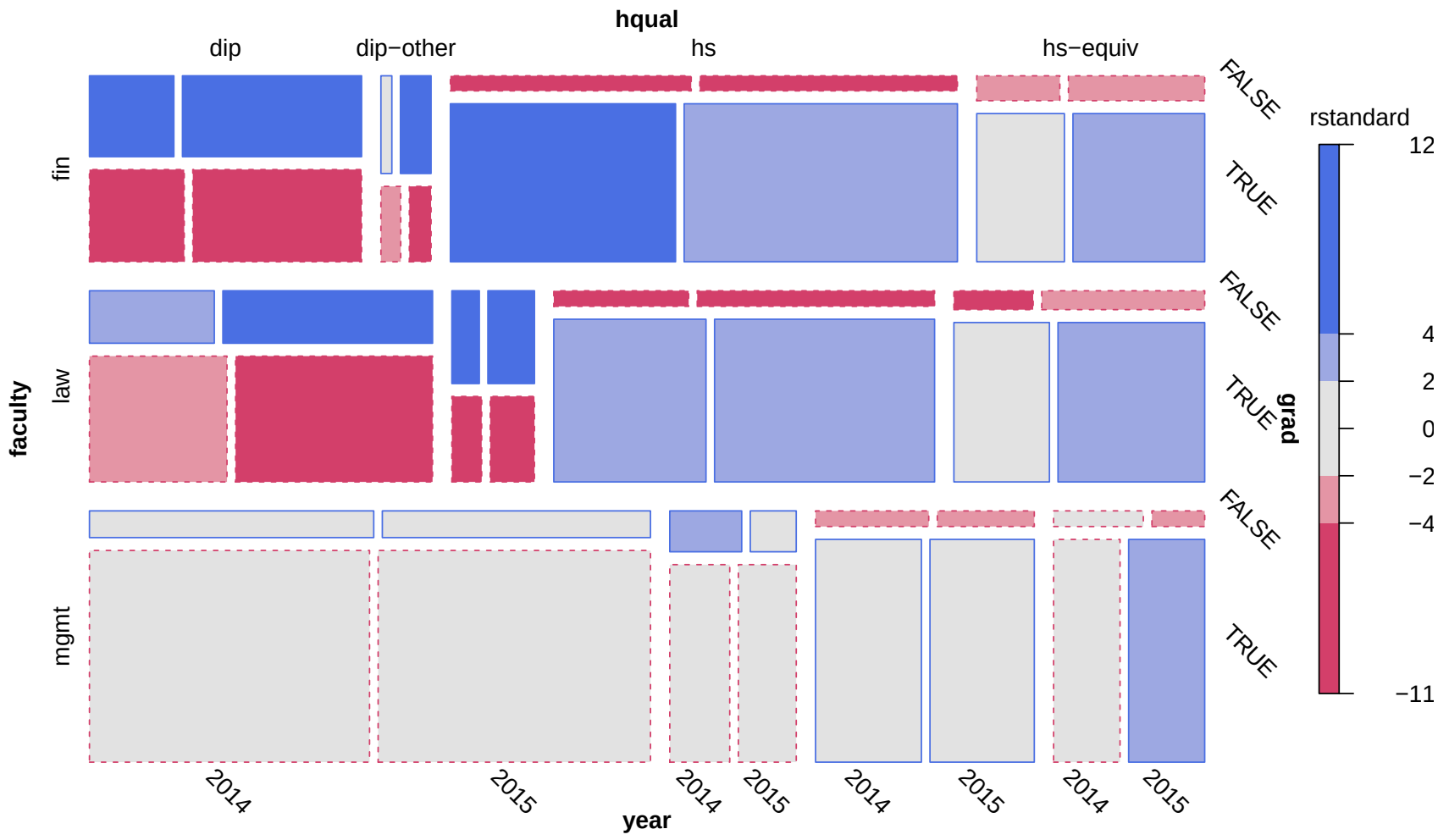
Model 2 [FH] [FG] [Y]

```
glm2 <- glm(Freq~faculty *
                (hqual + grad) + year
              , data = sts_df, family = poisson)
```



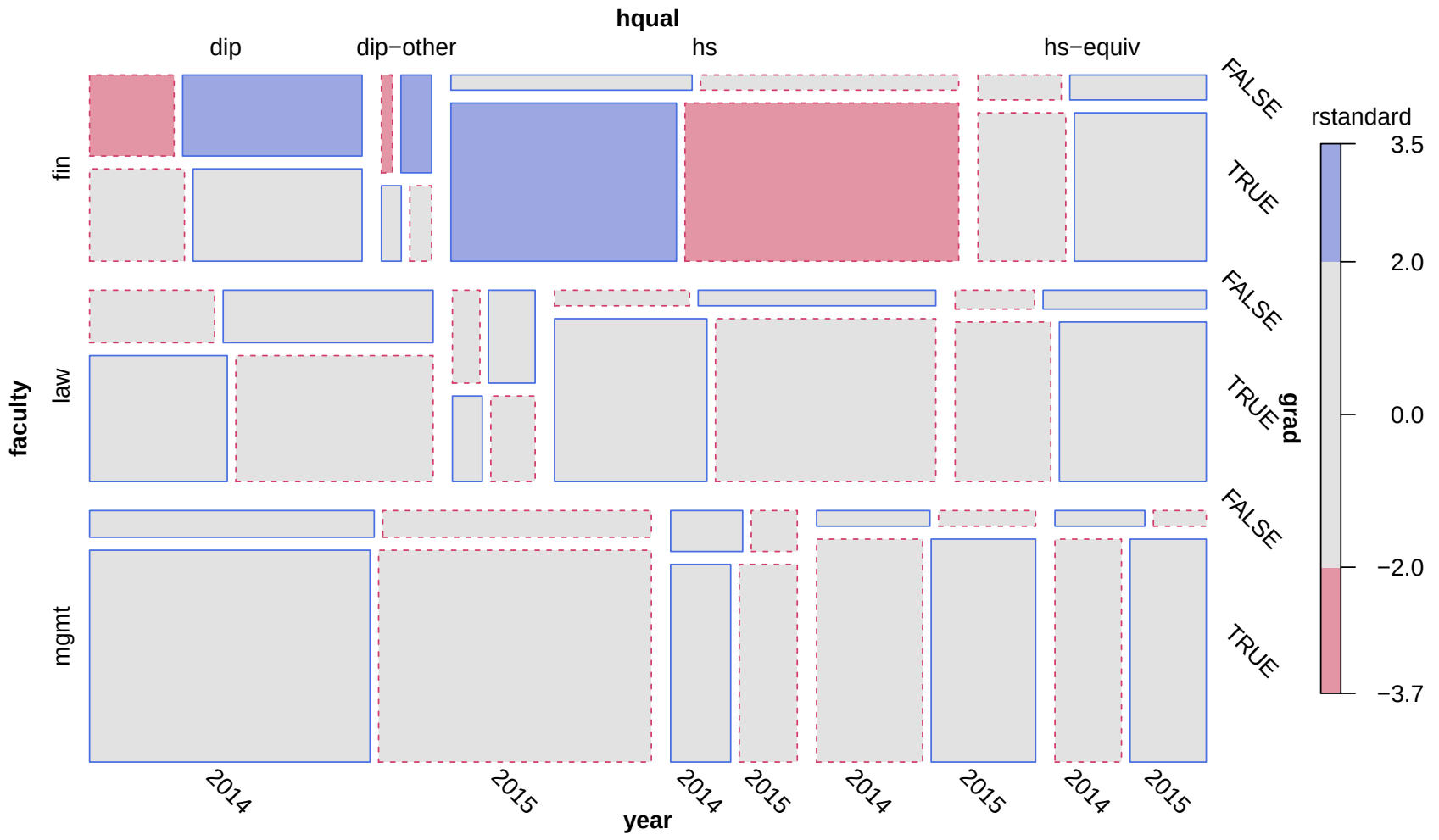
Model 3 [FH] [FG] [FY]

```
glm3 <- glm(Freq~faculty *
              (hqual + grad + year)
            , data = sts_df, family = poisson)
```



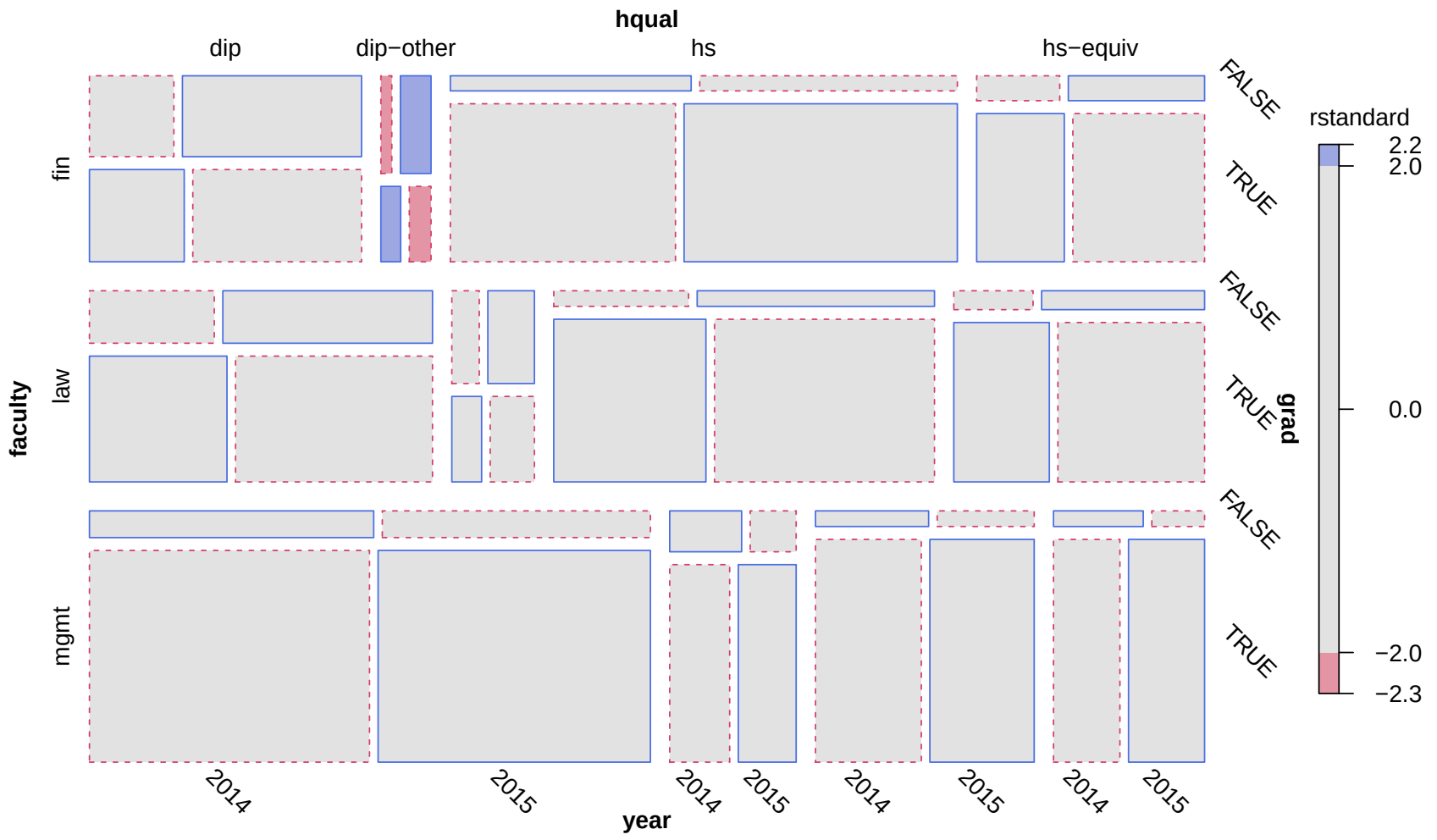
Model 4 [FHG] [FY]

```
glm4 <- glm(Freq~(faculty * hqual * grad) +
                 (faculty * year)
               , data = sts_df, family = poisson)
```



Model 5 [FHG] [FHY]

```
glm5 <- glm(Freq~(faculty * hqual * grad) +
              (faculty * hqual * year)
            , data = sts_df, family = poisson)
```



Case Study Conclusions

A policy change between 2014 and 2015 to increase student numbers led to lowering of entry level thresholds.

The foundation diploma appears not to have met the needs of so many additional students on the more challenging finance degree.

Recommend a review of curriculum and resources for the diploma to better support finance students, as well as interventions and support for students who are already one or two years into their course.

Diagnostic Tests

```
as.data.frame(anova(glm0, glm1, glm2
                    , glm3, glm4, glm5))
```

```
##   Resid. Df Resid. Dev Df  Deviance
## 1        40 1580.34522 NA        NA
## 2        34  884.04489  6 696.30033
## 3        32  738.68488  2 145.36001
## 4        30  664.95392  2  73.73096
## 5        21   37.10138  9 627.85254
## 6        12   13.88574  9  23.21564
```

Diagnostic Tests

```
LRstats(glm4)
```

```
## Likelihood summary table:
##          AIC    BIC LR Chisq Df Pr(>Chisq)
## glm4 391.54 442.06   37.101 21    0.01639 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```
LRstats(glm5)
```

```
## Likelihood summary table:
##          AIC    BIC LR Chisq Df Pr(>Chisq)
## glm5 386.32 453.69   13.886 12     0.3081
```

End