STAT 770 August 31 Lecture Part B

Illustrative R Data Analysis of a Simple Table

We import a clinical trial dataset, transform it from data-frame to table, and use it to ask various simple hypothesis testing questions to be covered formally in Chaps. 1-2 of Agresti.

Reading: in addition to Ch. 1 contingency table definitions, begin with **R** 'Getting Started' material from course web-page.

Dummy Variables and Discrete Predictors

Suppose $C = \{1, \ldots, m\}$ and n, $\{Z_a\}_{a=1}^n$ nonrandom

Data-frame: rows $N_{z,c} = \sum_{a=1}^n I_{[Z_a=z,\, X_a=c]},\ z,\ c)$
with row-index enumerating (z, c)

Now suppose $Z_a = (Z_{j,a},\ j = 1, \ldots, d) \in \mathcal{Z} \equiv I_1 \times \cdots \times I_d$

b^{th} **Dummy Variable for Z_j:** $(I_{[Z_{a,j}]} = b,\ a = 1, \ldots, n)$
column n-vector for each $j = 1, \ldots, d,\ b = 1, \ldots, I_j$
Use I_j n-vectors to account for categorical $Z_{j,a}$ in regression,
but just 1 vector $\{Z_{j,a}\}_{a=1}^n$ for numerical predictor $Z_{j,a}$

Tabular Data: $N_{z,c}$ entries in d-way table indexed $z = (z_1, \ldots, z_d)$

Access multicenter clinical trial data (Table 6.9, Agresti) in **R**:

```
> infect = read.table("http://users.stat.ufl.edu/~aa/cda/data/Infection.dat",
       header=T)           ### this option reads first line as column names

> t(infect)[,1:12]       ## first 12 columns of 16x4 data-frame
       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
center    1    1    2    2    3    3    4    4    5     5     6     6
treat     1    0    1    0    1    0    1    0    1     0     1     0
y        11   10   16   22   14    7    2    1    6     0     1     0
n        36   37   20   32   19   19   16   17   17    12    11    10

# "y" = treatmt resp = success, "treat" = indicator of experimental group
# Data: treatmt ctr identifiers, & counts of successes & failures
# "treat" is a "dummy column" for purpose of regression,
#  e.g. of   y counts/(y+n counts)    [or log, or logit]  versus  "treat"

### Alternative data presentation as a multi-way table
> infect.arr = array( data.matrix(infect[,3:4]), c(2,8,2), dimnames =
        list(c("Drug","Control"), 1:8, c("Success","Failure")) )
```

Questions to Address in R Data Analysis

- association overall between `y/n` and `treat`

- variability across clusters (centers) of association

- can centers be ignored with respect to treatment efficacy

Further R Steps in File `Rscript1.txt`

Step 1. Chi-squared Test of Row-column indep. in 2×2 table

Observed Table

	Succ	Fail
Drug	55	130
Control	47	143

Expected Table

	Succ	Fail
Drug	50.32	134.68
Control	51.68	138.32

$X^2 = \sum_{cell} \frac{(O-E)^2}{E}$ Corrected $= \sum_{cell} \frac{(|O-E|-0.5)^2}{E}$

Both referred to χ_1^2 table

With multiple (indep.) 2×2 tables can add their statistics and df. Other tests used if associations are likely in same direction.