STAT 770 August 31 Lecture Part A
Overview of Categorical Data Analysis

This lecture segment is an overview of data, models and formal setup we use under the heading of Categorical Data.

Reading for this week's lectures: Chap. 1 in Agresti's book, today in the 1st segment about kinds of data and basic models (Secs. 1.1, 1.2.1-1.2.3, 1.2.5), and in the 2nd segment using **R** about a small dataset illustrating techniques for data display and some questions we can ask and answer using simple **R** functions.

For this week's lectures, also see the historical material on contingency tables in **Ch. 16**, along with **Handouts 2 & 3** on the course web-page, respectively a link to a recorded lecture by Agresti about history and an article about how recent is the idea of analyzing cross-classified data in contingency tables.

Data and Models

Broadly, we will be studying Binomial and Multinomial count data, using parametric models to express occurrence probabilities in terms of parameter vectors θ and explanatory variables.

Data $(X_a, Z_a,\ \ a = 1, \ldots, n)$, a indexing experimental units

$X_a \in C$ $=$ discrete set of distinct **category** labels c

$Z_a \in \mathbb{R}^d$ vector of **explanatory variables** (usually discrete)

Models express $P(X_a = c,\, Z_a = z\, ;\, \theta)$ or $P(X_a = c \,|\, Z_a = z\, ;\, \theta)$

initially assume (X_a, Z_a) indep. identically distributed *(iid)*

$C,\ Z_a,\ \theta$ defined to reflect **structure connecting probabilities** for different (c, z)

Generality of this Formulation

(i) $C = \{0, 1\}$ or $\{\text{Failure, Success}\}$ or $T_a \in (b_j, b_{j+1}]$
X_a outcomes may be **ordinal** (ordered) or not
$c = (x_1, \ldots, x_r)$ may be **longitudinal, repeated measures**

(ii) $a = (i, j, k)$ may be multi-index, i, j, k indexing **factor levels**
equivalently Z_a may contain (i, j, k) coordinates
this is where multiway contingency tables come from

Next discuss data-frame versus contingency table
data representations

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

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 $z = (z_1, \ldots, z_d)$

Why Binomial and Multinomial ?

When Z_a are nonrandom: $N_{z,c} = \sum_{a=1}^{n} I_{[Z_a=z,\, X_a=c]} \sim \text{Binom}(n, p_{z,c})$
sum of iid binary r.v.'s, jointly distributed as Multinom(n, $\{p_{z,c}\}_{(z,c)}$)
since each a belongs to only one $(z,c) = (Z_a, X_a)$, with prob. $p_{z,c}$.

Unconditional parameterization, where $\theta = \{p_{z,c}\}_{(z,c)}$
and $N_{z+} = \sum_{c \in C} N_{z,c}$ is a random outcome

Sometimes sample data (*stratified*) fixing $N_{z,+} \equiv n_z$, so that
$(N_{z,c},\ c \in C) \sim \text{Multinom}(n_z,\ \{p_{c|z}\}_{c \in C})$
where $p_{c|z} = P(X_a = c \,|\, Z_a = z) = p_{z,c} / \sum_{k \in C} p_{z,k}$
and $\theta = \{p_{c|z}\}_{(z,c) \in \mathcal{Z} \times C}$ is **conditional parameterization**

Parameter Spaces and Statistical Questions

In unconditional parameterization, Categorical Statistics is about Multinomial Data with parameters $\{p_{z,c}\}$: in interesting cases parameters are restricted/shared to reflect tabular and regression structure.

Examples: (a) $\log(p_{i,c}) = \alpha_c + \beta_i$ or $\log(p_{c|z}) = \alpha_c + \beta' z$

(b) multiway extensions, similar models with (z, c) interactions,

(c) extensions reflecting longitudinal c's, or other *link functions* relating $p_{z,c}$'s to $E(N_{z,c})$'s

Questions: Tests and Conf. Int's for parameter components, Predictons of $N_{z,c}$ (*Classification*)

Sampling Design, Conditioning & Poisson

Some extensions condition on Marginals, e.g. Fisher Exact Test fixes m_1, n_1, n in Multinomial

	X=0	1	Tot
Z=0	N_{00}	N_{01}	n_0
1	N_{01}	N_{11}	n_1
Tot	m_0	m_1	n

Useful **distributional fact**: Multinom(n, $\{p_{z,c}\}$) dist'n for $\{N_{z,c}\}$ is equivalent to the conditional joint distribution of independent $N_{z,c} \sim$ Poisson($\lambda p_{z,c}$) r.v.'s given $\sum_{(z,c)} N_{z,c} = n$.

(A good self-contained exercise for review, not to be handed in.)

With this fact, conditioning in multinomial-data setting can be viewed as further conditioning on indep. Poisson underlying data.