Lecture 2

27.10.2019.

Recap: Discrete exponential families are toric varieties and are represented by data A, h where A is a $k \times r$ integer matrix with $\mathbb{1}$ in its row-span and h is a vector in $\mathbb{R}^n_{>0}$.

Example: X, Y binary. Independence model.

$$\mathcal{M}_{X \perp\!\!\!\perp Y} = \left\{ \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} : p_{00}p_{11} - p_{01}p_{10} = 0, \ p_{ij} \geq 0, \ \sum p_{ij} = 1 \right\}$$

$$= \left\{ \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} : \begin{pmatrix} p_{0+} \\ p_{1+} \end{pmatrix} (p_{+0} \ p_{+1}) \right\}.$$

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \qquad h = (1,1,1,1), \ \phi^{A,h} : \mathbb{R}^k \to \mathbb{R}^n$$

$$(\theta_1, \theta_2, \theta_3, \theta_4) \mapsto (\theta_1\theta_3, \theta_1\theta_4, \theta_2\theta_3, \theta_2\theta_4).$$

If $u = (u_{00}, u_{01}, u_{10}, u_{00}) \in \mathbb{N}^4$

$$Au = \begin{pmatrix} u_{0+} \\ u_{1+} \\ u_{+0} \\ u_{+1} \end{pmatrix}$$ captures the marginal counts of the contingency table

u_{00}	u_{01}	u_{0+}
u_{10}	u_{11}	u_{1+}
u_{+0}	u_{+1}	

Statistical tests for $\overbrace{\text{discrete exponential families}}^{\text{DEF}}$.

General set up: Consider a DEF $\mathcal{M}_{A,h} \subseteq \Delta_{r-1}$ and suppose we collect data $X^{(1)}, \dots, X^{(n)} \in [r]$ which are indep. and identic. distributed according to a distribution $p \in \operatorname{int}(\Delta_{r-1})$.
We would like to test

$$H_0 : p \in \mathcal{M}_{A,h} \quad \text{vs.} \quad p \notin \mathcal{M}_{A,h}.$$

Example: Data 326 homicide indictments in Florida.

Binary classif: $X = \text{Race} = \begin{Bmatrix} B, \\ W \end{Bmatrix}$, $Y = \text{Penalty} = \begin{Bmatrix} D \\ \cancel{D} \end{Bmatrix}$

Q: Were decisions of penalty made independent of race?

Race \ Penalty	D	$\cancel{D}$	Total
B	19	141	160
W	17	149	166
Total	36	290	326

Hypothesis testing problem: $H_0: p \in \mathcal{M}_{X \perp\!\!\!\perp Y}$ vs. $H_1: p \notin \mathcal{M}_{X \perp\!\!\!\perp Y}$.

Chi-squared test of indep: If H_0 is true $\Rightarrow p_{ij} = p_{i+} p_{+j}$

The expected number of occurrences of $\{X = i, Y = j\}$ is $n p_{i+} p_{+j}$

We can estimate the marginal probabilities by using

$$\hat{p}_{i+} = \frac{u_{i+}}{n} \qquad \hat{p}_{+j} = \frac{u_{j+}}{n}, \qquad \begin{aligned} u_{i+} &= u_{i0} + u_{i1} \\ u_{+j} &= u_{0j} + u_{1j} \end{aligned}$$

→ We can estimate the number of counts $n p_{i+} p_{+j}$ by

$$\hat{u}_{ij} = n \, \hat{p}_{i+} \, \hat{p}_{+j}$$

The X^2-statistic $\quad X^2(u) = \sum_{i=1}^{r}\sum_{j=1}^{r} \frac{(u_{ij} - \hat{u}_{ij})^2}{\hat{u}_{ij}}$

If H_0 is true we expect a small value for $X^2(u)$.
We reject H_0 is $X^2(u)$ is too large. How large?

$\left\{\right.$ The probability that $X^2(U)$ takes a value greater than or equal to $X^2(u)$ provided that H_0 is true.
p-value

Goal: Compute $P(X^2(U) \geq X^2(u))$

Asymptotic approach: Suppose $n \to \infty$ and use

Prop. 9.1.1: If the joint distribution of X is given by a distribution $p \in \mathcal{M}_{A,h}$, then

$$\lim_{n\to\infty} P(X_n^2(v) \geq t) = P(\chi^2_{df} \geq t) \quad \text{for all } t \geq 0,$$

where $df = r - 1 - \dim \mathcal{M}_{A,h}$ is the codim. or # of degrees of freedom of the model.
i.e. For a true distribution lying in the model, the Pearson X^2 statisti converges in distrib. to a χ^2_{df}.

Drawbacks: What if sample size is small and contingency table is sparse?

Exact Approach: Can we calculate this p-value some other way?

Denote by $[r]$ the outcome space. Assume $h=1$.

Let $\mathcal{T}(n) = \{ u \in \mathbb{N}^r : \sum_{i \in r} u_i = n \}$

<u>Def 1.1.10:</u> We call the vector Au the <u>minimal sufficient</u> statistics for the model $\mathcal{M}_A$, and the set of tables

$$\mathcal{F}(u) = \{ v \in \mathbb{N}^r : Av = Au \}$$

is called the fiber of the contingency table $u \in \mathcal{T}(n)$ w.r.t the model.

$1 \in \text{rowspan} \Rightarrow \quad xA = 1 \qquad x(Av) = x(Au)$

$1v = 1u \quad \Rightarrow \sum v_i = \sum u_i = n.$

<u>Prop 1:</u> If $p \in \mathcal{M}_A$, $p(x) = \theta_1^{a_{1x}} \cdots \theta_K^{a_{Kx}}$, $x \in [r]$, and $u \in \mathcal{T}(n)$, then

$$P(U=u) = \frac{n!}{\prod_{i \in r} u_i!} \boldsymbol{\theta}^{Au}$$

and the conditional probability $P(U=u \mid AU = Au)$ does not depend on p.

$$\begin{pmatrix} a_{11} & a_{1x} & a_{1r} \\ \vdots & & \\ a_{K1} & a_{Kx} & a_{Kr} \end{pmatrix} \begin{pmatrix} u_1 \\ \\ u_r \end{pmatrix}$$

pf.

$$P(U=u_i) = \frac{n!}{\prod_{i \in r} u_i!} \prod_{i \in r} p_i^{u_i} = \frac{n!}{\prod_{i \in r} u_i!} \prod_{i \in r} (\theta_1^{a_{1i}} \cdots \theta_K^{a_{Ki}})^{u_i}$$

$$= \frac{n!}{\prod_{i \in r} u_i!} \prod_{i \in r} \theta_1^{a_{1i} \cdot u_i} \cdots \theta_K^{a_{Ki} \cdot U_r} = \frac{n!}{\prod_{i \in r} u_i!} \prod_{i \in r} \theta_1^{\sum_i a_{1i} \cdot u_i} \cdots \theta_K^{\sum_i \theta_{Ki} \cdot u_i}$$

$$= \qquad\qquad = \square \, \boldsymbol{\theta}^{Au}$$

Moreover, $P(U=u \mid AU=Au) = \dfrac{P(U=u)}{P(AU=Au)}$

$$P(AU=Au) = \sum_{v \in \mathcal{F}(u)} \frac{n!}{\prod_{i \in r} v_i} \theta^{Av} = n!\, \theta^{Au} \sum_{v \in \mathcal{F}(u)} \left(\prod_{i \in r} v_i\right)^{-1}$$

$$\Rightarrow P(U=u \mid AU=Au) = \frac{1/\left(\prod_{i \in r} u_i!\right)}{\sum_{v \in \mathcal{F}(u)} 1/\left(\prod_{i \in r} v_i!\right)} \qquad \blacksquare$$

- Based on Prop 1, we generalize Fisher's exact test by computing the p-value

$$P(X^2(U) \geq X^2(u) \mid AU=Au)$$

Here $X^2(U) = \sum_{i \in} \frac{(U_i - \hat{u}_i)^2}{\hat{u}_i}$, $\hat{u}_i = n\hat{p}_i$ where

$\hat{p}_i$ is the MLE. $\downarrow$ Lecture 4.

$\hat{p}_i$ is the same for all tables in the fiber.

Hence the p-value is:

$$\frac{\sum_{v \in \mathcal{F}(u)} 1_{X^2(v) \geq X^2(u)} / \left(\prod_{i \in r} v_i!\right)}{\sum_{v \in \mathcal{F}(u)} 1/\left(\prod_{i \in r} v_i!\right)}$$

$\rightarrow$ Exact computation of this quantity is prohibitive. Thus we sample from elements in the fiber.