week 7 day 2

"Exact testing for model/data fit for log-linear models"
"Part three"
"Algebraic & Geometric Methods in Statistics"

Sonja Petrović
Created for Math/Stat 561

Feb 22, 2023.

Agenda

- Chapter 9 from our textbook: Fisher's exact test
- Part of chapter 8, as we may need the cone of sufficient statistics.

Goals

- LAST WEEK:
  - Understand hypotheses testing for model/data fit
- LAST LECTURE: we worked towards
  - What is a p-value for a goodness-of-fit test?
  - Asymptotic vs. exact tests
  - Fisher's test and example
- THIS LECTURE:
  - General goodness of fit test for log-linear models
  - Open problems and relation to projects!

Recap

Recap: definition of exact conditional tests

In an **exact** goodness-of-fit test, one uses the exact distribution of a GoF statistic... for example the X^2 statistic defined in lecture 10. The reference distribution is the *exact* distribution of that statistic on the set of tables with given fixed sufficient statistics (margins).

```
        gender
range    M  F Nb
  <=135K 8  1  4
  > 135K 2  9  2
```

```
        gender
range    M  F Nb
  <=135K 9  0  4
  > 135K 1 10  2
        gender
range    M  F Nb
  <=135K 9  1  3
  > 135K 1  9  3
```

Conclusion? Evidence in the data? Significance?

Definition [p-value]

Define the p-value for the GoF statistic.

- Read the beginning of Chapter 9. Section 9.1: Conditional inference.
  - We are *conditioning* on the row and column sums of the table.
  - These are sufficient statistics for the independence model.
  - This is a *general strategy*...

Interpret: what are all the possible tables? What is the probability of any given table?

	M	F	T/Nb	totals
$\leq$ 135K	?	?	?	**13**
$>$ 135K	?	?	?	**13**
totals	**10**	**10**	**6**	**26**

The general exact test for contingency tables [board lecture]

- Definition 9.1.3. - fiber
- p194: Problem 9.1.6. - understand the problem definition

Models with a design matrix

- $X_1, \ldots, X_k$ discrete random variables, $X_i \in \{1, \ldots, d_i\}$
- u = a k-way contingency table $u \in \mathbb{Z}_{\geq 0}^{d_1 \times \cdots \times d_k}$ [Draw a table!] Flatten u to vector.

Log-linear model

Sufficient statistics = marginals of u: $P_\theta(U = u) = \exp\{\langle Au, \theta\rangle - \psi(\theta)\}$.

Example $X_1 \perp\!\!\!\perp X_2$

$$\left[\begin{array}{cccc|cccc|c|cccc}
1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 1 & 1 & \cdots & 1 \\
\hline
1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & \cdots & 1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 & \cdots & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 & \cdots & 0 & 0 & \cdots & 1
\end{array}\right]_{(d_1+d_2)\times d_1 d_2} \cdot \begin{bmatrix} u_{11} \\ \vdots \\ u_{d_1 d_2} \end{bmatrix} = \begin{bmatrix} u_{1+} & \cdots & u_{+d_2} \end{bmatrix}.$$

Main ingredients for MCMC for exact testing of model/data fit

- Markov bases and Metropolis-Hastings - that is the start of Section 9.2.
  - include example 201-202 culminating with Proposition 9.2.10.
  - look out for Felix's talk in april!

A warning sign

include example. 8.2.2. nonexistent MLE!

Resources & License

- Quick summary notes about p-values that I wrote for Stat 514.
- Read about hypothesis tests for context of the model fitting tests in these lecture notes.
- This lesson from Penn State online offers a one-page summary of Fisher's exact test for 2×2 tables, as it was developed by Sir Fisher!
- Believe it or not, there is a great 2×2 example on Wikipedia, a page which actually contains a really good explanation for this one example.

This document is created for Math/Stat 561, Spring 2023.