Algebraic Statistics

Benjamin C. W. Brown
B.Brown@ed.ac.uk

26th October 2020

Contents I

Statistical Models

planetmath.org

A *statistical model* is usually parameterised by a function, called a *parametrisation*

$$\Theta \to \mathcal{P}, \quad \text{given by} \quad \theta \mapsto P_\theta, \quad \text{so that} \quad \mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

where Θ is the *parameter space.* Θ is usually a subset of $\mathbb{R}^n$.

McCullagh, 2002

This should be defined using category theory.

A contingency table contains counts obtained by cross-classifying observed cases according to two or more discrete criteria.

Example

TODO: Figure (Florida death sentences)

We ask whether the sentences were made independently of the defendant's race.

- Classify using two criteria with r and c levels, yields two random variables X and Y.
- Code outcomes as $[r] := \{1, \ldots, r\}$, and $[c] := \{1, \ldots, c\}$.

All information about X and Y is contained in the *joint probabilities*

$$p_{ij} = P(X = i; Y = j), \quad i \in [r], \ j \in [c].$$

- These in turn determine the *marginal probabilities*:

$$p_{i+} := \sum_{j=1}^{c} p_{ij} = P(X = i), \quad i \in [r],$$

$$p_{+j} := \sum_{i=1}^{r} p_{ij} = P(Y = j), \quad j \in [c].$$

Definition

Two random variables X and Y are *independent* if the joint probabilities factor as $p_{ij} = p_{i+} \cdot p_{+j}$, for all $i \in [r]$ and $j \in [c]$. Denote independence of X and Y by $X \perp\!\!\!\perp Y$.

Proposition

Two random variables X and Y are independent if and only if the $(r \times c)$-matrix $p = (p_{ij})$ has rank one.

Summary

Questions?