Why do we care about generators of toric ideals?

- $\mathbb{k}$ - field

$R = \mathbb{k}[x_1, \dots, x_n]$

- $A \in \mathbb{Z}^{d \times n}$ matrix

gives toric ideal

$$I_A = \langle \underbrace{x^u}_{= \prod_{i=1}^{n} x_i^{u_i}} - x^v : u, v \in \mathbb{N}^n, Au = Av \rangle$$

<u>Hypothesis testing in statistics.</u>

X_1, X_2 random variables taking finitely many values.

E.g. traits of individuals in a population

~~Good~~ Explicit examples: hair color, gender, ...

In a population there is a true distribution of these variables

Statistics wants to make inference about the true distribution from (small) samples

Example. Independence of X_1 and X_2

In hypothesis testing we want to refute hypotheses such as independence
(reject)

Argument goes like:

- Assume hypothesis
- Assess how likely it is to get the data that came out of the actual experiment.
- If the actual data looks very exceptional among data ~~coming~~ assuming the hypothesis.

Then either the hypothesis was wrong
(or you got unlucky)

In the independence example:

$X_1 \in [r] = \{1, \dots, r\}, \quad X_2 \in [S]$

Data:

$$u = \begin{array}{c|c} \begin{pmatrix} u_{11} & & \\ & \ddots & \\ & & u_{rs} \end{pmatrix} & \begin{array}{c} u_{1+} \\ \vdots \\ u_{r+} \end{array} \\ \hline u_{+1} \cdots u_{+s} & u_{++} \end{array}$$

of integers counts

u_{++}: sample size

Def:
The marginals of u are the row- and column sums

Under the assumption of in

Expect: $u_{ij} = u_{++} \frac{u_{i+} \cdot u_{+j}}{u_{++}^2}$

Can compare actual data u_{ij} to estimate under hypothesis $\frac{u_{i+} u_{+j}}{u_{++}}$ using any suitable distance like X^2 distance.

Get a value for $X^2(u)$, but is it big or small? What is the scale?

Fisher's idea: Compare to fake data v with the same marginals and thus the same estimation.

Marginal map is $\mathbb{Z}$-linear $u \to A \cdot u = (u_{\cdot +}, u_{+ \cdot})$

Def: $b \in \mathbb{Z}^d$. The fiber of b is

$$A^{-1}[b] = \{ v \in \mathbb{N}^k : Av = b \}$$

For Fisher's exact test need to evaluate X^2 on $A^{-1}[Au]$ where u actual data.

if real data is among 5% of most extreme X^2 values, reject hypothesis.

"The lady tasting tea" tells the story.

Typically the fiber is too large to be enumerated.

Diaconis-Sturmfels: Use a Markov chain to sampe uniformly from the fiber.

Need a set of elementary moves:

$\mathcal{M} \subseteq \ker_{\mathbb{Z}}(A)$ finite

Can you reach every point from every other point in the fiber using moves in $\mathcal{M}$ and not leaving the fiber.

$\mathcal{M} \ni m = m^+ - m^-$ gives a binomial

$x^{m^+} - x^{m^-} \in R$

$I_{\mathcal{M}} = \langle x^{m^+} - x^{m^-} : m \in \mathcal{M} \rangle$

Prop: $u, v \in \mathbb{N}^n$. There exists a walk

(*) $u = u_0, u_1, \dots, u_s = v$, $u_i \in \mathbb{N}^n$

$u_{i-1} - u_i \in \pm \mathcal{M}$

if and only if $x^u - x^v \in I_{\mathcal{M}}$

Rem: ~~Toric ideal generators~~: $x^u - x^v \in I_{\mathcal{M}} \Leftrightarrow Au = Av$.

$\mathcal{M}$ Markov basis $\Leftrightarrow$

Proof: if a walk exists.

$$X^u - X^v = \underbrace{X^u - X^{u_1}}_{\in I_{\mathcal{M}}} + X^{u_1} + \dots + X^{s-1} - X^v \in I_{\mathcal{M}}$$

$$X^w(X^{m^+} - X^{m^-})$$

on the other hand: $X^u - X^v \in I_{\mathcal{M}}$

Exercise: Show that

$$X^u - X^v = \sum_{m \in \mathcal{M}} X^{w_m}(X^{m^+} - X^{m^-})$$

Now compare coefficients.

X^u appears with positive sign on r.h.s.

$$X^u = \cancel{X^{w_i}(X^{m_i^+} - X^{m_i^-})}$$

$$X^{w_{m_i}} X^{m_i^+} \quad \text{then } X^{w_i} X^{m_i^-} = X^{u_1}$$

$u \to w_i m_i^-$ is the first step.

Either $\overset{\shortparallel}{v}$ or it cancels. Finite induction //

Prop. says if $\mathcal{M}$ connects any u, v with $Au = Av$

$I_A \subseteq I_{\mathcal{M}}$ and $\supseteq$ is clear since $\mathcal{M} \subseteq \ker(A)$