Katherine and Victoria,

Both your presentation and your final project paper were good, but I think it was probably not necessary to include some of the probability definitions in the first section (you never used them later). I was also hoping to see more about the connection between varieties and statistical models, especially in the paper. You did include the binomial probability models with $n = 3$, but another example like the $m \times n$ independence models or some other example would have been a better choice to round out the paper. You concentrated quite heavily on maximum likelihood estimation, and that is one very important side of the story. However, the examples you did of that did not really require any of the computational algebra techniques we discussed (see below, especially Comment 10), so they weren't the best examples to use to make your points.

Specific Comments:

1. In looking over things, I see that your history is not quite accurate here. The involvement of Persi Diaconis in this story is really through an article he wrote with Sturmfels quite a bit *before* the 2003 conference that led to the *Algebraic Statistics for Computational Biology* book. That article was "Algebraic algorithms for sampling from conditional distributions." Ann. Statist. 26 (1998), no. 1, 363397. The idea there was that one can use Gröbner techniques to produce Markov bases for sampling from the collection of contingency tables with given marginals to do hypothesis testing in situations where χ^2 distributions are not sufficiently accurate, but Fisher's Exact Test is unfeasible because there are too many tables to consider. (I discussed this briefly with Katherine the day of the Sulski Lecture.) The second main author of the *Algebraic Statistics for Computational Biology* is Lior Pachter, not Persi Diaconis.

2. This is mostly a writing point. When you want to define a term this way, it's really better to *paraphrase* than to use a direct quotation (especially if the quote doesn't fit with your sentence). Say something like: Let Y be a continuous random variable. We say $f(y)$ is the pdf of Y if $P(a \leq Y \leq b) = \int_a^b f(y)\, dy$. (and then give the reference).

3. MLE is really a standard probability or statistical technique (that is, not specific to algebraic statistics). The algebraic viewpoint gives other tools to compute maximum likelihood estimators.

4. The variable of interest for a binomial experiment is the *number of "successes"* in the n trials.

5. Statisticians would write $\hat{p} = \frac{y}{n}$. The notation $\hat{p}$ denotes the *estimator* for the *model parameter* p. (You need to keep those two things separate as you think about this!)

6. You're not really "expanding out" anything here. You are collecting the values of the binomial pmf into a vector.

7. The x_i are the different values of the pmf as functions of p and $y = 0, 1, 2, 3$. They are probabilities, not outcomes. The idea here is that you are writing the pmf in a different form, as a function of the model parameter p. The vectors you get all lie on the variety given at the bottom of the page. Note, though, that not every point of the variety is obtained since the $0 \leq x_i \leq 1$. The fancy way to say this is that the probability model is the intersection of the variety and the *probability simplex* in $\mathbf{R}^4$. The probability simplex is the set

$$\{(x_1, x_2, x_3, x_4) : 0 \leq x_i \leq 1, x_1 + x_2 + x_3 + x_4 = 1\}$$

8. This is really a general binomial probability model, including *all of* the different binomial distributions for different values of p. It's a parametrization of all possible binomial distributions.

9. I'm not sure what you meant here. It's definitely not true that a large sample implies independence in general. If you meant to say that if the Y_i are independent, you can handle even very large samples of data, then OK. But that's not exactly what you wrote.

10. The maximum likelihood estimators for μ and σ in the normal model can be derived using just multivariable calculus. Here's how it works. (I'm going to treat the variance σ^2 as a parameter in its own right, not think of it as the square of the SD.) Since

$$\ln(L(y_1, \ldots, y_n | \mu, \sigma^2)) = \frac{-n}{2} \ln(\sigma^2) - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} (y_i - \mu)^2 \right),$$

taking partial derivatives and setting them equal to zero, we get

$$\begin{aligned} \frac{\partial \ln(L)}{\partial \mu} &= \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} (y_i - \mu) \right) = 0 \\ \frac{\partial \ln(L)}{\partial \sigma^2} &= \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \left(\sum_{i=1}^{n} (y_i - \mu)^2 \right) = 0 \end{aligned}$$

The first equation implies the MLE for μ is:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}$$

(the sample mean). Substituting that into the second equation and solving for σ^2, we get

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1} (y_i - \overline{y})^2,$$

which is (one form of) the *the sample variance.* (It's the biased form of the estimator; the constant multiple

$$\frac{1}{n-1}\sum_{i=1}(y_i - \overline{y})^2$$

is the unbiased form.) You can check, using the Second Derivative Test for functions of two variables, that this is a maximum of the log-likelihood function L. Note that I didn't need to use any different techniques here for different n. This is a completely general argument and it works equally well whatever n is.

Final Project Presentation: 92 (A-)

Final Project Paper: 92 (A-)