MAT-322, Professor Mileti, 05/11/2021

A Survey on Algebraic Methods in Statistics

Mira Tellegen, Delaney Gaughan, and Lilith Hafner

Introduction:

In our paper, we are conducting a survey of intersections of algebra and statistics which does not rely on advanced statistics. In the field of theoretical mathematics, there is a body of research online surrounding the translation of traditional statistical methods and equations into formulations using Abstract Algebra ideas. For example, randomized experimental design in statistics can be understood by intersections of varieties, and statistical Fisher information can be understood as a Riemannian metric on statistical models. We collected a breadth of the current results in the field, and in the following paper, we touch on each intersection of algebra and statistics that we found academic writing on, exploring both the traditional statistics approach and the algebra-related approach to understanding. The first topic, which appeared across most of the papers we surveyed, concerns contingency tables.

2x2 Contingency Tables:

One of the first techniques learned in statistics is the use of 2x2 contingency tables, which can be applied to situations in which one is looking to determine whether or not two factors are related. In classic statistics, the following formula is used to calculate a z-score:

$$z = (x - \mu) \div \sigma$$

The z-score can be then compared to a normal distribution curve to determine the likelihood of the factors appearing related by random chance alone. The normal distribution curve is considered unconditional, meaning that it is not dependent on the selected sample and is applied generally (Aoki et al). In cases where there is a large sample size that is being generalized to a population of an unknown size, having an unconditional normal distribution curve is very helpful. However, the application of the normal distribution curve does not always fit exactly. For example, in cases where the sample size is very small, it is unlikely for the data to form a normal distribution, even if the two factors are not related.

In algebraic statistics, the joint probability function circumvents the z-score step entirely, and instead results in a direct probability that the two functions are related.

$$p(x,\ y) = \binom{n_1}{x} {p_1}^x (1 - p_1)^{n_1 - x} \binom{n_2}{x} {p_2}^y (1 - p_2)^{n_2 - y}$$

In this expression, we take the probability of their being *x* successes in n_1 trials for the first variable being tested, and the probability of their being *y* successes n_2 trials, and multiply them together to get the probability of both happening at once.

There are both similarities and differences in each approach. For example, both classic statistics and algebraic statistics utilize a null hypothesis from which the resulting statistic is compared. In classic, the null hypothesis for a two-variable contingency table is the idea that $p_1 = p_2$. In algebraic statistics, however, the variable used to determine statistical significance is **X**. The following function, an adaptation of the previous joint probability test, is known as Fisher's exact test, referred to as such because the significance is determined by hypergeometric distribution, and is thus conditional to the sample size. The conditional distribution of X given T is as follows:

$$P(X = x,\ T = t) = \frac{\binom{n_1}{x}\binom{n_2}{t-x} p_1{}^t (1-p_1)^{n-t}}{\binom{n_1 + n_2}{t} p_1{}^t (1-p_1)^{n-t}} = \frac{\binom{n_1}{x}\binom{n_2}{t-x}}{\binom{n}{t}} \quad .$$

Where $t = x + y$. As such, the equation simplifies the data into a product of a single variable *x*, to find it's conditional distribution. The equation thus relies on p_1, the probability of *x,* rather than fixing $p_1 = p_2$. This resulting distribution depends on X, and is, notably, not symmetric when $n_1 \neq n_2$. This helps make the resulting region of rejection for the null hypothesis unbiased.

However, for both, increasing the sample size can strengthen the conclusion. Because the algebraic statistics approach maps to a hypergeometric distribution, rather than to a binomial distribution, it is more applicable in different kinds of calculations.

Hypergeometric vs. Binomial Distribution

To delve more into the difference between binomial and hypergeometric distributions, the binomial distribution is often used to estimate the proportion of a larger population by drawing, without replacement, a smaller sample from the population. A cumulative binomial distribution is represented by the following:

$$P(X \leq k;\ n,\ p) \ = \ \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \ for\ k = 0, 1, ..., n \ .$$

In this model, *n* refers to the number of trials conducted, and *p* the success probability. Notice that *np* is the mean number of successes. The distribution shows the probability of there being *i* successes for all potential values of *i.* When *p*=0.5, the distribution is symmetric, since there is an equal likelihood of success and failure. The larger the value of *n,* the more the actual data will resemble the theoretical distribution based on *p.* Generally, binomial distributions are

recommended for situations where *n* is large, and the exact size of the population is unknown. This is helpful in cases where researchers want to generalize results to a larger population, such as in drug trials.

The binomial distribution is not always appropriate in cases in which the sample size is being selected from a population of a known size, or a smaller population (Krishnamoorthy). In those cases, a hypergeometric distribution is used. A hypergeometric distribution is used to determine the probability of getting a specific result from a known sample size. For example, suppose we had a collection of twelve tiles, half of which were yellow, half of which were green. If we were to draw four tiles at random, a hypergeometric distribution could tell us the probability that all of the tiles we drew were yellow.

For a single selection, the probability of achieving a specific outcome is given by

$$h(x;\ N,\ n,\ k)\ =\ [kC_x][N - kC_{n-x}]\ \div\ [NC_n]$$

and the cumulative distribution is given by

$$F(k\ n,\ M,\ N) = \sum_{i=L}^{k} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \quad L = max\{O,\ M - N + n\}.$$

Where $\binom{M}{i}$ signifies the number of ways *i* nondefective items can be selected from a sample size of *M,* and $\binom{N-M}{n-i}$ is the number of ways $n - i$ defective items can be selected from a sample size of $M - N$. The cumulative distribution gives us, for each *k,* the probability of observing that many nondefective items in a sample of size *n.*

Notice that this equation only works when the values of *M* and *N,* the total number of nondefective and defective items, respectively, are known. As such, while the hypergeometric distribution is more accurate for the cases where it applies, particularly with small sample sizes, the results do not map to a population of an unknown size.

While there are some cases in which the hypergeometric and binomial distributions line up, they can differ greatly. Notably, they are more similar in cases where sample size is large, when the binomial distribution is generally considered more accurate, than in cases in which the sample size is small.

Joint Probability Function:

When looking at a 2x2 contingency table, and further at larger tables, the total probability for all of the cells of the table is always going to be one. In that, we can define the different cells p_{ij}, where *i* corresponds to the first variable being tested, and *j* corresponds to the second, wherein there are two potential outcomes for each. Note that, since every outcome is going to fit into one of the cells of the contingency table,

$$\sum_{i,j=1}^{2} p_{ij} = 1 .$$

From there, it is possible to generate a joint probability function **X** for the table, wherein $\mathbf{X} = \{X_{11}, X_{12}, X_{21}, X_{22}\}$, given by

$$p(x) = \binom{n}{x_{11}, x_{12}, x_{21}, x_{22}} p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}} p_{22}^{x_{22}}$$

Where we define **x** as the frequency vector. **x** is a vector which shows the strength of the probability of each individual outcome on the table. These results can be further generalized to calculate the probability of functions on contingency tables with more than two variables in the same fashion.

Experimental Design:

"Experimental design is defined simply as the choice of sites, or observation points, at which to observe a response, or output" (Gibilisco 159) In choosing these points, we can view the task of selecting observation points as equivalent to selecting algebraic varieties.

One example of experimental design where algebra can prove extremely useful is population sampling in ecological statistics. In traditional statistics, to determine the population in an area, rather than counting every member of a species, statisticians approach census by random sampling: they may pick many random points across the whole habitat, or section off the habitat and sample only some sections, or sample randomly with more random points in some areas based on expected population density.

Algebraic varieties particularly apply in the last sampling process, when it comes to density estimates. Scientists may draw lines (transects) across the habitat, then measure the distance of certain members of the population from the transect. *Reconstruction* is the process of then finding functions which describe the populations' relationship to the transect, to determine population density. When scientists *reconstruct* to find polynomial functions which describe the population density in an area, they are performing interpolation to find polynomial functions and their corresponding varieties (Maruri-Aguilar 160).

"Interpolation is the construction of a function *f(x)* that coincides with observed data at *n* given observation points." For our points $\mathbf{D} = \{d_1, \ldots, d_n\} \in \mathbb{R}^k$, and our observed values $y_1, ..., y_n \in \mathbb{R}$, "we build a function such that $f(d_i) = y_i$ for $i = 1,...,n$" (Maruri-Aguilar 161). Another approach to interpolation by scholars Pistone and Wynn in 1996 builds "polynomial interpolators" by examining an isomorphism between $\phi : D \to \mathbb{R}$ where **D** represents our experimental design, and the quotient ring $\mathbb{R}[x_1, ..., x_k] / I(D)$ (Maruri-Aguilar 162). Our final

sampling comes from a selection of points from the resultant varieties of the polynomials we've found through interpolation.

In summary, using algebraic geometry, we can find random observation points by randomly sampling varieties, taken from polynomials we derive by interpolating over a research region (in the case of the example above). This provides a standardized method for randomly sampling in statistics using a basis of algebraic theory.

In Notari's article on the subject, he notes that the similarity in understanding between random sampling in experimental design and algebraic varieties is that "points in a cloud are moved towards the common point along straight lines," which we see with the concept of interpolation on the transects (Notari 201). However, he reveals several ongoing issues with this application of Algebraic Statistics in relation to confounding variables. In statistics, confounding variables are variables which aren't controlled or examined in the experiment, but which may affect the result. For example, measuring car speed with drivers and car varieties as your variables, ice on the road would be a confounding variable. Notari notes that "a satisfactory description of the aliasing structure of a design with replicated points" is currently not available, touching on a more complex subject in the article of when we have duplicate or near duplicate points (a concept common in statistics but which doesn't mesh well with the polynomial function approach).

The Fisher Information Metric:

"*In the 1940s Rao and Jeffreys observed that Fisher information can be seen as a Riemannian metric on a statistical model"* (Gibilisco et al.). In the following sections, we explicate this observation with definitions and examples. A Riemannian metric is a general

operation on a space with some requirements; Fisher information as an operation on statistical models as spaces satisfies the requirements of a Riemannian metric, so all results about Riemannian metrics in general automatically hold for Fisher information on a statistical model.

Riemannian metrics:

"Suppose for every point x in a manifold m, an inner product $<\cdot,\cdot>x$ is defined on a tangent space T_xM of M at x. Then the collection of all these inner products is called the Riemannian metric" (Weisstein). The expression $<\cdot,\cdot>x$ uses $\cdot$ as a placeholder for an unnamed function argument. In this case—as implied by the resemblance to inner product notation, and more directly stated elsewhere in the definition—those arguments are elements of the tangent space surrounding x, while x is a point in the manifold. In effect, this is a function with nonstandard syntax. Using standard function notation, we could write this as a function f(u,v,x), where u and v are names for the previously unnamed vector arguments. The criteria for inner products are that for all vectors u,v,w and scalars a:

(1) $< u+v, w > \;=\; < u, w > \;+\; < v, w >$

(2) $< au, v > \;=\; a < u, v >$

(3) $< u, v > \;=\; < v, u >$ and

(4) $< v, v > \;\geq\; 0$ and equal only if $v = 0$ (Renze et al.), that is, $v \neq 0 \Longrightarrow\; < v, v > \neq 0$

We can think of the dot product in $\mathbb{R}^2$ as a first example. The manifold is $\mathbb{R}^2$, the tangent space is R^2 for all x, and the inner product is defined by $< (x_1, y_1), (x_2, y_2) > x \;=\; x_1x_2 \;+\; y_1y_2$. We can now observe that the dot product satisfies these properties in $\mathbb{R}^2$. We get (1) and (2) from the distributive axioms of linear algebra, (3) follows from commutativity, and (4) arises from the fact that for all real a, $a^2 \geq 0$ and equal only if

$a = 0$. Our justification for the dot product serving as a riemannian metric on $\mathbb{R}^2$ relies on our field's commutativity and a property about squares in that field. We can thus generalize from the dot product on $\mathbb{R}^2$ to the dot product on k^n where *k* is a commutative field with for all *a* in *k*, $a^2 \geq 0$ and $a \neq 0 \implies a^2 \neq 0$.

Gorodski defines a Riemanian metric as "a family of smoothly varying inner products on the tangent spaces of a smooth manifold", notably including smoothness as an additional requirement. This precludes, for example, the family of functions $< u, v >_x = u \cdot v$ if *x* is on the rational lattice and $2u \cdot v$ otherwise, defined on the tangent spaces of R^2 around *x*.

Any inner product space, a vector space with an inner product operation, is an example of a Riemannian metric on a space. The metric is simply defined as $< u, v >_x = u \cdot v$ for all *x*. But we can find examples of Riemannian metrics on a space that are not also inner product spaces. For example, $< u, v >_x = (u \cdot v)(||x|| + 1)$ on $\mathbb{R}^2$ is smooth as it is the product of two smooth functions, and at each *x*, $< \cdot, \cdot >_x$ is an inner product because it is a positive scalar multiple of the dot product which is an inner product. To formally show this we can apply commutativity of multiplication in a *k*-algebra to prove each of the properties of an inner product remain satisfied when the product is multiplied by a positive constant.

Statistical model:

"A statistical model is a set of probability distributions on the sample space S. A parameterized statistical model is a parameter set together with a function $P : \Theta \to P(S)$, which assigns to each parameter point $\theta \in \Theta$ a probability distribution P_θ on *S*" (McCullagh). We will be working with parameterized statistical models. Notably these models are *not* probability distributions, but rather functions from parameters to probability distributions. This

allows us to describe a collection of potential distributions. For example, a statistical model of two coins parameterized by the odds of each landing on heads where $\Theta = [0,1]x[0,1]$, is P((a,b)) = {HH: ab, HT: a(1-b), TH: (1-a)b, TT: (1-a)(1-b)}.

Fisher information:

Fisher information is a function of a statistical model and a parameter that represents how accurately the parameter can be reconstructed from an observation. In other words it expresses how much information an observation carries about the parameter. It is defined by $I(\theta) = E_\theta \left[\frac{\partial^2 ln\, P(x|\theta)}{\partial \theta^2}\right]$ (Wolpert). The expression $\frac{\partial^2 ln\, P(x|\theta)}{\partial \theta^2}$ tells us how rapidly the probability of observing **X** falls off for slightly different θ, and the quantifier $E_\theta[...]$ gives us the expected value of that "statistic" for the given θ. I put "statistic" in quotes here because a formal statistic must only depend on **X**, not on θ, but this one does not.

Fisher information can be seen as a Riemannian metric on a statistical model:

We can express the Fisher information function as a Riemannian metric using matrix notation as $g_{ij}(\theta) = E_\theta \left[\frac{\partial^2 ln\, P(x|\theta)}{\partial \theta_i \partial \theta_j}\right]$, where *i* and *j* denote axes of θ. In function notation, $< u, v > \theta = \sum_i \sum_j u_i v_j E_\theta \left[\frac{\partial^2 ln\, P(x|\theta)}{\partial \theta_i \partial \theta_j}\right]$. This is a lot to think about. Let's start in the one dimensional case. Here, *g* is a 1x1 matrix with the element $E_\theta \left[\frac{\partial^2 ln\, P(x|\theta)}{\partial \theta^2}\right]$, or the inner product $< u, v > \theta = uv E_\theta \left[\frac{\partial^2 ln\, P(x|\theta)}{\partial \theta_i \partial \theta_j}\right]$; dot product scaled by Fisher information at θ. Provided the statistical model is smooth, Fisher information should be smooth. To ensure that Fischer information is positive, we have to add another constraint to the statistical model: the

parameterization must be identifiable, that is, distinct parameters give distinct probability distributions and equal probability distributions imply equal parameters. This means that every parameter value's probability distribution is distinct, and so each observation will bear some information about the parameter value.

In the multidimensional case, the Fisher information becomes a matrix, the Fisher information matrix, rather than a scalar, and can be used as a Riemannian metric whenever we have an identifiable probability distribution. Further explication of the multidimensional case is outside the scope of this paper, however.

Bibliography

Aoki, Satoshi, et al. *Markov Bases in Algebraic Statistics*. Springer New York, 2012.

This source by Aoki, similar to our paper, gives a full survey of current techniques in Algebraic Statistics. It is particularly helpful when it comes to algebraic applications in contingency tables, which is an essential building block of algebraic statistics.

Gibilisco, Paolo. *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, 2010.

In this source, Gibilisco covers how to use varieties in the process of randomly sampling in experimental design. He suggests that hyperplane arrangements, variety intersections, and polynomial interpolation can ease the process of experimental design.

Gorodski, Claudio. “Riemann Manifolds.” Universidade de São Paulo, https://www.ime.usp.br/~gorodski/teaching/mat5771-2016/ch1.pdf.

In this paper, Gorodski explains Riemann Manifolds and metrics, a tool for understanding manifold spaces, defined by smoothly varying inner products.

Hibi, Takayuki. “Harmony Of Grobner Bases And The Modern Industrial Society .” *The Second Crest-Sbm International Conference*.

In statistical applications to science, such as nuclear science, the statistical probability that a system, machine, or experiment is reliable can be crucial. In this paper, Hibi covers the concept of system reliability, and suggests that this concept fundamentally relates to networks and paths from node to node in a network.

Renze, John; Stover, Christopher; and Weisstein, Eric W. “Inner Product.” *Wolfram MathWorld*, mathworld.wolfram.com/InnerProduct.html.

This source provides a definition and breakdown of the concept of an inner product.

Krishnamoorthy, K. (2015). Handbook of Statistical Distributions with Applications (2nd ed.). Chapman and Hall/CRC. https://doi-org.grinnell.idm.oclc.org/10.1201/b19191

This book serves as a handbook for statistical formulas, and walks through the significance of binomial distributions and hypergeometric distributions.

Maruri-Aguilar, H.., and H. Wynn. “Generalised design: interpolation and statistical modelling over varieties.” *Algebraic and Geometric Methods in Statistics*, doi:10.1017/cbo9780511642401.012.

In this source, Maruri-Aguilar and Wynn cover experimental design and statistical regression from an algebra point of view. They give examples of how algebraic devices can help us understand experimental design.

McCullah, Peter. “What is a Statistical Model?” University of Chicago, http://www.stat.uchicago.edu/~pmcc/pubs/AOS023.pdf

In this document, McCullah explores the types and importance of statistical models, including examples and relations to regression and random experimental design.

Notari, R., and E. Riccomagno. “Replicated Measurements and Algebraic Statistics.” *Algebraic and Geometric Methods in Statistics*, pp. 187–202., doi:10.1017/cbo9780511642401.012.

In this source, Notari and Riccomagno cover experimental design and statistical regression from an algebra point of view. They walk through the use of varieties for randomized experiment design, using zero-dimensional varieties to construct statistical models, and using quotient rings on ideals in the process of linear regression.

Pistone, G. “Algebraic Varieties vs. Differentiable Manifolds in Statistical Models.” *Algebraic and Geometric Methods in Statistics*, pp. 341–366., doi:10.1017/cbo9780511642401.022.

In this source, Pistone (a prominent scholar in algebraic statistics), discusses Riemann geometry and Fisher information in examining the relation of affine geometry and toric varieties to statistical models.

Robbiano, Lorenzo. “Gröbner Bases and Statistics.” *Gröbner Bases and Applications*, 1998, pp. 179–204., doi:10.1017/cbo9780511565847.010.

Statistics How To, 18 Dec. 2020, www.statisticshowto.com/.

This site walks through traditional statistical methods and formulas, which provides a basis on which to build our survey of algebraic statistics. They cover z-scores, binomial distributions, and hypergeometric distributions in a traditional statistical context.

Weisstein, Eric W. "Riemannian Metric." From *MathWorld*--A Wolfram Web Resource. https://mathworld.wolfram.com/RiemannianMetric.html

This source provides a definition and breakdown of the concept of a Riemannian metric.

Wolpert, Robert L. “Fisher Information & Efficiency.” Department of Statistical Science, Duke University, https://www2.stat.duke.edu/courses/Spring16/sta532/lec/fish.pdf

In this document, Wolpert unpacks Fisher information, which is a key part of statistical modeling, which relates to the information about the parameter in the model.