ARITHMETIC STATISTICS COURSE NOTES

ROBERT J. LEMKE OLIVER

0. What is arithmetic statistics?

Arithmetic statistics is broadly concerned with questions about objects like:

- Prime numbers
- Polynomials over the integers, i.e. $f(x) \in \mathbb{Z}[x]$
- Number fields, i.e. $\mathbb{Q}(\alpha)$ where α is a root of an irreducible polynomial $f(x)$
- Diophantine equations, e.g. $y^2 = x^3 + x + 1$, where we are concerned with understanding the integral or rational solutions to the equation.

If one has a specific object in mind, most interesting questions have a purely concrete answer. A number is either prime or it isn't, a polynomial or number field either has a property or it doesn't, a diophantine equation either has a solution or it doesn't, etc. These specific questions and the tools used to answer them are often important and interesting, but arithmetic statistics is instead concerned with how these properties "typically" or "randomly" behave. For example, we might ask:

- How likely is a "random" integer to be prime? Or, how likely is a "random" prime to have a certain property (e.g., that it's congruent to 1 (mod 4))?
- How likely is a "random" integer polynomial to be irreducible? Among those that are irreducible, what is the most likely Galois group?
- How does factorization behave in a "random" number field?
- How many rational solutions should one expect to an equation $y^2 = x^3 + Ax + B$ for "random" integers A and B?

An important feature of all of these questions is that they're asking questions about the typical behavior of elements of *infinite* sets: there are infinitely many primes, infinitely many polynomials, infinitely many number fields, and infinitely many diophantine equations. This is important for the question to leave the realm of the concrete – if there were only finitely many primes, for example, we could understand *everything* about the primes just by studying that finite set – but it means that the notion of "random" is not really defined. Instead, the template for an arithmetic statistics question is something like:

For some large X, study the properties of the objects with "complexity" at most X, in particular the proportion that have a given property. Then let $X \to \infty$.

What "complexity" means varies from context to context, and for some problems, it's the subject of current research to decide on a good notion. The most important feature, however, is that there should be only finitely many objects with complexity at most X, for any X. This is necessary to ensure that the proportion asked for in the template question actually makes sense. This leads also to one of the other fundamental questions of arithmetic statistics: how many objects are there with complexity at most X, asymptotically as X tends to infinity?

Date: January 20, 2022.

With regard to our example problems, here are some of the typical notions of complexity used and the associated counting problem.

- Primes. Complexity is just the size of the prime. The counting problem is then to determine the number of primes below X. This is the subject of the prime number theorem, which asserts that the answer is asymptotically $X/\log X$.
- Polynomials. We usually split polynomials up first by their degree (i.e., we consider only polynomials of degree n for some n), and the complexity is typically measured by the largest absolute value of the coefficients $a_1, \dots, a_n$. The counting problem is to determine the number of polynomials of degree n all of whose coefficients are at most X in absolute value. (This counting problem is relatively straightforward, and for that reason, this will actually be the first class of object we consider.)
- Number fields. The usual notion of complexity is the *discriminant*, which we'll define in a few lectures. The counting problem is to determine the number of number fields with discriminant at most X. This is a very hard open problem on which we'll discuss current progress and ideas.
- Diophantine equations. The notion of complexity varies depending on the class of equation considered, but usually pays attention to the size of the coefficients. For equations of the form $y^2 = x^3 + Ax + B$, a reasonable notion of complexity would be $\max\{|A|, |B|\}$. As it happens, the notion that's more typically used is $\max\{4|A|^3, 27B^2\}$, for reasons we'll discuss later in the semester. A subtlety in this counting problem is that some of these equations are *isomorphic* in a particular sense, and one typically only wants to consider one element from a given isomorphism class. Thus, the counting problem is usually to determine the number of isomorphism classes with complexity at most X.

In this class, we will focus on problems related to polynomials, number fields, and diophantine equations. Prime numbers are interesting, and will play a key role throughout this course, but as I taught a course recently[1] on primes, I will not be focusing on these results. In particular, I may sometimes quote without proof certain results on the distribution of primes if the result is pertinent. I will be more than happy to explain why the result is true outside of class if you are interested.

It's now time to make these vague notions more concrete. We do so by considering integer polynomials and a classical theorem due to Hilbert. This theorem was proved in the 19th century but motivates current research, including a major conjecture proved in 2021!

[1] OK, four years ago.