IT'S ALL IN THE GENES! - STATISTICS AS A TOOL IN MODERN GENETICS

Mark Yang
University of Florida
Gainesville, FL

Statistics is the science of making conclusions under uncertainty. For example, can we prove that smoking causes cancer? No, because we see many smokers never develop cancer. The relation is uncertain at the individual level. Nevertheless, it is still useful if we can establish that data show smoking and lung cancer to be highly correlated. To reach this conclusion, we need statistics.

Humans have long known that characteristics can be passed down through generations, but exactly how this is done is less clear. Though we say "like father like son," we have also observed that some children are very different from their parents. Because of this uncertainty, geneticists have used statistics as a tool since the very beginning of this science. Mendel's law could be verified only in statistical sense. According to his theory, in a population with approximately 11% blue-eyed individuals, one brown-eyed parent and one blue-eyed parent will tend to produce brown-eyed offspring in the ratio of three to one, but the theory can be verified only by means of statistical studies of observed offspring.

Traditionally, genetics has been used in the process of plant and animal breeding so as to produce desirable characteristics in offspring. But many important traits that benefit humans, such as milk from a cow or sugar from sugar cane, are controlled by genes and the environment. Whether the success of agriculture is mainly due to the selection of seeds or due to fertilizer and pest-control is open to debate. (Similarly, there is a constant debate on whether the prolonging of human life is mainly due to medicine or due to the rise of living standard.) When all the factors are combined, how to select the most beneficial parents to breed the next generation is a challenging statistical problem. It turns out the best choice not only depends on the individuals, but also their parents, siblings, relatives and the different environments in which they live. Complex statistical models help sort out the environmental effects from the genetic, allowing sound decisions to be made on breeding. Thanks (partially) to statistical analysis, we can enjoy better and better food products every year.

One of the most important topics in modern genetics is to locate the gene that is responsible for a disease or a character. Once a disease gene is found, we may discover what is lacking due to the abnormal function of this gene and consequently have a clue as to how to treat the disease. Moreover, we may be able to replace the gene by gene therapy or genetic engineering. To find a gene is not easy. The human genome has 3,000,000,000 DNA codes and each gene has in the neighborhood of 10,000 codes. Locating a particular gene is like finding a one-yard long section in a 170-mile highway. Moreover, this particular section does not stand out as special if we are driving along that highway. We have to find it indirectly from its expression, called phenotype, in a living organism. Biochemists have found many markers in the human genome as well as that of other living organisms. These markers serve as landmarks like the mileage sticks along the highway. When the gene passes from parents to children, the markers adjacent to the gene pass with it. By tracing the family pedigree on the disease phenotype and the marker pattern, we may be able to identify the marker, which is adjacent to the gene, and consequently locate the gene. Unfortunately, the gene's expression is usually not clearcut. For example, many people with a diabetes gene may never develop diabetes and many with a normal gene develop diabetes. Obviously, statistics is needed in gene hunting under this kind of uncertainty. With intensive statistical effort, many disease genes have been found, such as those for cystic fibrosis and certain breast and colon cancers.

Everyone has heard of the human genome project. It is to lay out the 3,000,000,000 DNA codes in the whole human genome. We may say that the blue print of constructing a human being from a fertilized egg to a full-grown person is there. But to understand how it works is a challenge of the 21st century. How do these codes work together? We hope that one day we can decipher the whole genome function. Without this full understanding, any genetic engineering on human genome would be too risky. Now we know that the genome contains not only useful genes, but also a lot of useless codes called junk DNA. At this moment, we still have no sure method to identify the gene portion in the genome. There are rules to identify genes and there are exceptions to the rules. Due to this type of uncertainty, statistics plays a key role in human genome research.

Two mysteries that arouse human's deepest curiosity are the origin and the evolution of the universe and the origin and the evolution of life. Where did life originate? How did it get here? Is this process so likely that there should be many intelligent species like human beings, or it is so unlikely that we are the only intelligence in the universe? We have no answer yet, but we, possibly the only intelligence that knows to ask this question, have the obligation to find the answer. Several years ago, the journal *Science* did a survey asking leading scientists what will be the hottest research topic in the 21st century. The overwhelming answer was genetics. It seems that they have confidence that the mystery of life can be solved quicker than that of the universe. The genomes of living organisms provide a clue on how living organisms evolved in the past. Can we piece the puzzle together? There were so many random events in the past that no deterministic equations can provide us the answer. We need statistical help.

In summary, we see that statistics is useful in almost every branch of modern genetics. However, we do not want to over-emphasize its importance. Modern genetics can no longer be handled by a few disciplines. In addition to the traditional bio-medical sciences, chemistry, physics, computer and information sciences are all key players. Statistics is only one of them, but an indispensable one.

Further Readings:

Weir, Bruce, 1996, *Genetics Data Analysis II*, Sinauer Associates, Inc., Massachusetts.

Roff, Derek A., 1997, *Evolutionary Quantitative Genetics*, Chapman & Hall, N. Y.

Yang, Mark C. K., 2000, *Introduction to Statistical Methods in Modern Genetics*, Gordon & Breach Science Publishers, Amsterdam.