Statistical methods in laboratory and basic science research

Michael A. Newton, *Professor, University of Wisconsin, Madison, WI*

Keywords
statistics, probability, modeling, randomization, molecular biology, Luria Delbrück fluctuation test, comparative genomic hybridization, gene expression microarrays

Contents

Glossary

data: recordings from an experiment or investigation

inference: a conclusion drawn by reasoning from available information

likelihood: the probability (density) of the data as a function of the parameters

model: a collection of probability distributions for data

parameter: a quantity that indexes a statistical model

odds: the probability of an event divided by the probability of its complement

probability distribution: a representation of uncertainty in an unknown quantity

randomization: a technique of random rearrangement of measurements for the purpose of testing a hypothesis

statistic: something that can be computed from data

statistics: the field of inquiry concerned with obtaining, summarizing, processing, and drawing inferences from data

stochastic process: a probability distribution for a general outcome

Summary: To think statistically is to know that the measurements taken in an experiment are subject to systematic and random sources of variation, and that it is beneficial to base methods of data analysis on probabilistic models. Mathematical results from statistical theory indicate certain types of distributions that govern fluctuations in data, and some of these results are reviewed as they bear on statistical analysis in the laboratory and basic sciences. An example of statistical thinking to advance knowledge in molecular biology is described, as are some general strategies for statistical analysis that may be appropriate for a collaborating statistician. Four case studies demonstrate these concepts.

1 Introduction

Only by measurement does the experimentalist record features of the system that he or she is studying, be the system a population of insects growing in the laboratory or the network of biochemical events that cause a cell to divide or a tumor to grow. Measurements arise in a context within which their naked numerical form acquires the weightier status of information. The process of extracting information from numerical data is a central issue in the field of statistics generally and in applications of statistics to laboratory and basic sciences.

Were it known with certainty the numerical values of measurements that are about to be taken in some experiment, it would seem to be a waste of effort to take the measurements at all! Measurements are unpredictable. Even with a good understanding of the measurement process and the system being studied, one often acknowledges that precise recordings will exhibit unpredictable fluctuations caused by different sources of variation. In spite of these fluctuations, part of the variation may be more systematic, and repeated measurement may elucidate these systematic sources. Statistical methods are ways of processing numerical data for the purpose of drawing inferences about the system: inferences may be to estimate a parameter, test a hypothesis about the parameter, classify an experimental unit into one of several groups, assess the relationship between two factors, predict future measurements, or decide on one of several courses of action in an ongoing experiment.

Statistical methods become enacted during data analysis. The statistical approach to data analysis is founded on the premise that measurements are the realization of a stochastic process. This has a significant effect on the tone of deliberations because emphasis shifts immediately from the particular data in hand to the process by which the data arise. Indeed, many formal discussions distinguish the data which does occur, say x, from the stochastic process X: the function or rule which reveals the actual data $X(\omega) = x$ when the experiment is instantiated as one particular outcome ω amongst a universe of possibilities. This reversion from what actually is measured to what might be measured seems at first to complicate matters, but it is a necessary template for the theory of probability, and it provides a means to make precise quantitative statements about things that are intrinsically unpredictable. Of course it is not to say that the actual data x are ignored – far from it; rather, the significance of particular irregularities in x is gauged in part by the probabilities governing X.

This article considers elements of statistical thinking that arise in laboratory and basic sciences. The comments are informed primarily by the experience of being a research statistician who collaborates with biological scientists, and the emphasis is much more on statistics in molecular biology than statistics in the basic sciences generally. Some mathematical results from statistical theory described in the next section are followed by some comments on the role of statistics at different levels of investigation. This is followed by a discussion of data analysis strategies and then a series of four case studies in which statistical thinking has been helpful.

2 Theory: Universal Distributions

There is great diversity in the systems being studied in basic science laboratories. One of the contributions of statistical theory is to identify common structures present in a wide range of experiments — in particular, common features of the variation of certain measurements. The Poisson limit law is a good example. Suppose that the system under consideration is comprised of a large number n of experimental units, and each of these units provides a binary response to some query. For instance, millions of bacterial cells are growing in culture and one asks whether or not each cell has a particular genetic mutation at one locus in the genome. The total number Y of units which have one of the binary states may be a quantity of some interest as it may affect the experimental design to pinpoint the locus, for example. Under a wide range of conditions on the basic binary variables it is known that fluctuations in Y are well approximated by a Poisson distribution:

$$\text{Prob}(Y = y) = e^{-\lambda}\lambda^{y}/y! \qquad y = 0, 1, \ldots$$

where λ is the expected value of Y. Usually this result is presented in the special case where the binary variables are independent and identically distributed Bernoulli random variables. Then the sum Y has a Binomial distribution with parameters n and p, the common expectation of all the Bernoulli variables. With large n and small p, and $\lambda \approx np$, the Poisson approximation becomes valid. The assumptions of independence and common distribution of the binary variables are rather strict, and evidence has mounted that the Poisson approximation may work much more broadly. Indeed the Poisson clumping heuristic theory extends the result significantly; there can be quite complicated forms of dependence amongst the binary variables and still the Poisson limit holds. This is important since in many examples some dependence is expected. For instance, cell lineage effects will cause statistical dependence in the bacterial growth example.

The most important universal distributional result is the central limit theorem which concerns fluctuations in the arithmetic mean of a random sample. It provides conditions under which the sampling distribution is Gaussian (bell-curved) regardless of the nature of fluctuations in the variables which comprise the sample mean. Indeed the theory is really a collection of results dating back to the early work on probability by many including de Moivre, Laplace, and Gauss, and culminating with 20th century work by Polya, Lindeberg, Feller, Levy and others.

Other universal laws receive perhaps less attention but are still very important for making connections between diverse problems. When appropriately centered and scaled, for example, the largest observation in a large random sample must exhibit variations from one of exactly three well-characterized distributional forms, regardless of the sampling distribution of the data (This is sometimes called the extreme value trinity theorem, developed by E. J. Gumbel and others.) The Erdos-Renyi law and extensions of it concern the distribution of long head-rich runs in sequences of coin tosses, and this has found significant application in problems of matching biomolecular sequences. Universal long-range dependence structures have been identified in certain kinds of time-series measurements also. Often distributional forms arise as the stationary distribution of a Markov process characterizing random fluctuations in the system over time. For example the Gamma distribution is the stationary distribution of abundance when a population evolves stochastically according to certain constraints. Knowing these universal laws assists both experimental design and data analysis. They can be used for 'back-of-the-envelope' sample size calculations, and they can form the basis for more detailed modeling efforts.

3 The Role of Statistics

3.1 Exceptional Cases

A beautiful illustration of statistics in the service of the basic laboratory sciences is the work by S. E. Luria and M. Delbrück concerning heritable changes in bacteria. Before Luria and Delbrück's work in the 1940's it was well known that a bacterial culture exposed to a certain virus could readily die out, but that periodically there would emerge clones of resistant bacteria. Various explanations presented themselves. Possibly some of the bacterial cells adapt to the invading virus and survive to form a resistant colony. Contrary to this adaptation hypothesis is the mutation hypothesis, which has stood the test of time and which is a central element in modern bacteriology. The mutation hypothesis asserts that bacterial variants (i.e., mutants) arise during normal growth of the colony, and that certain mutants resistant to the virus may by chance exist in the culture prior to viral infection. If so, they emerge for observation simply by the process of selection after the virus has killed the sensitive cells; the virus itself does not effect an adaptation of the bacterial cells.

The ingenious experiment devised by Luria and Delbück to address the problem involved a comparison of the variance of resistant cell counts grown under different conditions. In one condition separate cultures each grew from a very small initial population size; in the control condition a single large colony was separated into a similar number of separate cultures. All cultures then were exposed to the bacteriophage (virus) and counts were made of the number of resistant cells in each culture. Regardless of how bacterial variants arise, one can argue that in cultures grown in the control condition (i.e., from subsets of a large colony) there should be Poisson variation in the number of resistant cells. On the adaptation hypothesis, this same level of variation is expected in cultures of the first type, however the mutation hypothesis predicts extra-Poisson variation. Cultures in which a resistant mutant arises early will present a very large number of resistant cells compared to cultures in which the mutant arises later. It was by comparing the variation in cell counts between these two conditions that evidence favoring the mutation hypothesis was derived.

Having a statistical argument be central to a major scientific advance is fascinating, especially for people dedicated to the study of statistics; but it seems that such elegant Luria-Delbrük-like case studies are the exception rather than the rule in the application of statistics in the laboratory and basic sciences. Certainly there are wonderful case studies — the formulation of the idea of a tumor suppressor gene was a fundamental advance in cancer research brought about by the work of A.G. Knudson in his statistical analysis of retinoblastoma; the ability to map genes such as those responsible for Huntington's disease and cystic fibrosis is based on statistical properties governing the transmission of DNA during meiosis; the work of Sewall Wright used the variance of phenotypes in different experimental crosses to estimate the number of genetic loci affecting the phenotype. What will be the next great advance?

3.2 Endemic Methods

The ordinary application of statistical thinking in scientific discourse is to characterize imperfect knowledge; it forms one step of many to compile, describe, and report experimental results. At this level statistical discourse is a basic language for dealing with intrinsic variability; it is endemic in the sense of being regularly occurring, and it affects numerous steps in experimental design and data analysis. As examples of statistical questions consider the following: Upon measuring a cell proliferation rate in two different conditions, are the observed rate differences more than one expects by chance alone? If not, then it may be justified to treat the two conditions as one. When measuring some property of a cell type by preparing cells at different liquid dilutions, how can the measurements be combined efficiently across dilutions? In studying the production of some chemical compound, how can one identify optimal settings of several factors that affect production? Related questions are addressed in the case studies described later.

Supporting the notion that statistical methods are endemic is the fact that basic statistical calculations are embedded in much of the operating software of modern laboratory equipment. For example, a flow cytometer is an important device to determine properties of cells by measuring scattered light and fluorescence of mobilized, fluid-suspended cells. Part of the sophisticated computations built into a cytometer is a statistical discriminant analysis to classify individual cells by features such as cell size or granularity. Statistical discriminant analysis is part of the electronic nose, a device to detect airborne scents for use in food quality testing and other applications.

Statistical calculations are also embedded in the basic protocols of many high-throughput laboratory methods. For example, the technique of comparative genomic hybridization measures DNA copy number variation in cancer cells by processing fluorescence image intensities from labeled tumor and normal DNA that have competitively hybridized to immobilized DNA on a glass slide. Intensity signals from the two sources are measured all along the genome, and statistical signal processing is used to decide when one channel is significantly stronger than the other. Further, DNA microarrays are now widely used to measure simultaneously the level of gene expression of thousands of genes. A large amount of raw image data constitutes the results of one measurement, and this is processed to create a single record for each gene by a series of statistical manipulations of the image data. For instance, with spotted cDNA microarrays, an algorithm is used to localize each spot, account for local background fluorescence, and normalize measurements across the microarray. Statistical methods are used in these cases, and elsewhere, because they can automatically process large amounts of raw data in a potentially meaningful way.

4 Statistical Strategies

Statistical thinking can be effective in laboratory and basic sciences, both at the level of experimental design and data analysis: the former is critical for drawing strong conclusions while the latter is necessary to disentangle technical and biological sources of variation so as to uncover relevant biological signals. This article emphasizes statistical strategies at the data analysis phase where the investigator has in hand a complicated set of numerical results bearing on his or her research hypothesis. Though technological issues are ever present, the sources of variation most significant in the examples considered here are biological sources, having to do with natural variation in the components of living systems.

Three stages of statistical thinking are evident in many successful applications: informal descriptive analysis, global hypothesis testing, and modeling. In fact these are overlapping and linked and one or another may be more important in a given case.

4.1 Descriptive Statistics

Typically the initial phase of data analysis is descriptive and one tends to avoid drawing inferences or making probabilistic statements. Good practice involves checking the integrity of the data, making summary tabulations, scatterplots, histograms, and other graphical representations. The intelligent use of descriptive statistics can be highly informative; alternative non-numerical summaries of data serve to highlight intrinsic structural features; and, though they may not directly address inference questions, they certainly reveal properties of the data that will guide the use of other statistical methods. The importance of good descriptive statistics cannot be understated; indeed, this area currently presents great challenges as the amount of raw numerical information which can be obtained and stored is becoming vast.

Beyond the special use of descriptive statistics in laboratory and basic sciences, most people are familiar with their widespread use in diverse domains such as sports, politics, and public health. Although in so many human endeavors one needs to record and summarize what has happened, it is important to recognize that statistical analysis involves much more than the recording and presentation of descriptive statistics – it provides a framework to interpret data and to draw statistical inference. Often, this next stage of analysis addresses the question of whether or not the data in hand exhibits more variation than what one might expect by chance alone. If so, then it is justifiable to look closely for patterns of systematic variation.

4.2 Randomization

Sir R.A. Fisher may have been the most influential figure in the development of statistics, and one of his great contributions was the randomization test. It addresses the question as to whether or not there is any interesting structure in the measurements. Generally the method is easy to describe and implement; it proceeds conditionally upon numerical data but unconditionally on the arrangement of the data and it is widely applicable. It concerns significance testing; but, being conditional, its conclusions are not highly sensitive to distributional assumptions on the data. As one part in the sequence of statistical methods used in a data analysis, randomization testing is helpful because it can serve to justify the decision to finalize an analysis or to move forward with a more sophisticated modeling strategy.

4.3 Modeling

If there is interesting structure in the data, as evidenced perhaps by careful graphical analysis or randomization testing, then a powerful approach to quantifying that structure and learning more about it is to fit a statistical model; in other words, to identify a joint probability distribution for the measurements. Probability models are the basic elements of statistics: like particles in physics or cells in biology. Formally, a model is a collection of probability distributions. Models describe relationships among measurements, not exact connections, but probabilistic dependencies. Models relate measurements of different kinds, for example by describing how the expected value of a response measure is affected by changes in a set of predictors. Models can relate measurements in terms of unobserved, latent, or hypothetical constructs. Basically, models come in many shapes and sizes: there are linear models, generalized linear models, graphical models, Markov random field models, survival models, time-series models, random effects models, hierarchical models, semiparametric models, to name a few general classes.

Fitting the model identifies one member of the collection which captures better than the others patterns in the observed measurements. Central to the business of model fitting is another concept developed by Sir R.A. Fisher, the likelihood function. Mathematically, the domain of the likelihood function is the set of parameters that index the model, and the function evaluates to the joint probability (or probability density) of the observed data. Selecting the distribution having the highest likelihood is one approach to model fitting. There are other methods of model fitting, but likelihood methods are compelling and are used in some of the case studies described next.

Identifying an adequate model takes some care as one usually must trade off complexity and goodness of fit. Models that enforce as little structure as possible (e.g. nonparametric, semiparametric) tend to be favored when the data carry a lot of information, but their flexibility is a drawback in some problems, especially where subject-area knowledge can be encoded as a simple parametric form.

When unobservable factors are modeled together with the observable data, it is often effective to use the conditional probability distribution of the unknowns given the knowns to generate inferences. This is an essential feature of the Bayesian approach to statistical inference. The approach is compelling for several reasons, not the least of which is that all inferential computations reduce to integrating probability distributions in one way or another. Thus, for a given analysis it is usually more clear by this approach what needs to be calculated, and modern computational methods are making the implementation much more routine. In Bayesian analysis the interpretation of probability is extended from the classical focus on relative frequency to a more general measure of imperfect knowledge.

5 Case Studies

5.1 Microbial Biodiversity and Conditional Inference

An interesting statistical problem arose in a study by some University of Wisconsin biologists of bacterioplankton diversity in a certain lake ecosystem. M.M. Fisher and her colleagues (including the author) used a technique from molecular biology to probe lake-water samples and measure their composition with respect to different species of bacteria. Cruder measurement techniques had been in use to identify the overall amount of bacteria in a sample, but these techniques did not allow one to know how much diversity the sample contained. Understanding the effects of pollution or agricultural run-off would be enhanced if these more refined measurements could be taken.

Very roughly, the data from one lake-water sample could be summarized as a vector of binary variables each indicating presence or absence of a certain bacterial species in the sample. In one component of the experiment, lake-water samples from $m = 15$ different conditions were assayed and each gave a binary vector of length $n = 82$. There were $n = 82$ different bacterial species identified in this component of the experiment, across all samples. In fact the $m = 15$ samples arose from different environmental conditions with respect to levels of chemicals such as nitrogen and phosphorus and sampled at different times relative to certain treatments.

Naturally the bacterial composition profiles exhibited variation, and it was expected that some of this variation was measurement error and some represented underlying fluctuations in the bacterioplankton populations. To quantify the variation between samples, Fisher and her colleagues computed the Sorensen index for each pair of samples (i, j):

$$C_{i,j} = 2\frac{\sum_{k=1}^{n} X_{i,k}X_{j,k}}{[(\sum_k X_{i,k}) + (\sum_k X_{j,k})]}$$

where $X_{i,k}$ is the Bernoulli trial indicating presence of species k in sample i. In other words, $C_{i,j}$ is the number of species that are in common between the two samples divided by the average total number of species present in the two samples, and it ranges between 0 and 1.

For the purposes of reporting it was important to calibrate these indices by some means. There was also the problem that $m = 15$ profiles gave rise to 105 different pairwise comparisons, and so it was not clear how best to report significant findings. Using some elementary statistical theory in combination with conditional inference provided an effective solution.

Suppose that the Bernoulli trials $X_{i,k}$ are independent among all samples and all loci, but that the probability $p_k = P[X_{i,k} = 1]$ could depend on the species k but not the sample i. In other words the null hypothesis is that the whole profiles are identically distributed random vectors, meaning that chance alone caused differences from one sample to the next. Some species could be more prevalent than others across all samples because the probability p_k could depend on the species. To test the global null hypothesis of identically distributed samples, one still has to deal with the many nuisance parameters $p_1, \ldots, p_n$. Note that for any species k, the joint distribution of $X_{1,k}, \ldots, X_{m,k}$ conditional on the sum $X_{\cdot,k} = \sum_{i=1}^{m} X_{i,k}$ is a discrete uniform distribution not involving any nuisance parameters. Therefore, the null conditional distribution is fully specified and can be realized, for example, by permutation.

Viewing the raw data as an $m \times n$ matrix $X = (X_{i,j})$, repeatedly generate hypothetical data sets X^* by randomly shuffling the elements of each column of X. From every such simulated data set X^*, compute and store all 105 Sorenson indices;

a total of $B = 9999$ X^* matrices might be simulated. From this calculation one has a large sample from the null distribution of various quantities, such as the maximum or minimum index across all 105 pairwise indices. Any pair of samples (i, j) whose observed index $C_{i,j}$ was in the upper tail the distribution of the maximum index may be said to exhibit more similar profiles than expected by chance alone. Analogously, any pair whose index was in the lower tail of the distribution of the minimum is unusually different. Specifically, the $p-$value for one pair is the proportion of samples having at least as extreme a score as the observed score, taken relative to the full set of $B = 10000$ scores. In this case study the statistical contribution was simply to calibrate the Sorensen indices by randomization.

5.2 Comparative Genomic Hybridization and Mixture Modeling

Comparative genomic hybridization (CGH) is used by oncologists to probe for aberrations in cancer cell genomes — aberrations in the number of copies of DNA at any one of many locations in the genome. As with many molecular biological techniques, CGH relies on measuring fluorescence of differently labeled DNAs, and it is based on hybridization to sort out a complex pool of molecules. The protocols involve some sophisticated image analysis to process the fluorescent signals and arrive at a determination at each genetic locus as to whether that locus is deleted, amplified, or normal in the tumor cell DNA. Eschewing the technological issues, let us consider the downstream data analysis problem of processing these discrete CGH profiles from a study of bladder cancer development.

As an example, consider the small CGH experiment involving $n = 6$ bladder cancer tumors analyzed by C.A. Reznikoff, T.R. Yeager, and colleagues (including the author) at UW Madison. Cells from each tumor were prepared and compared to normal cells using CGH; this resulted in a profile for each tumor that recorded whether there was deletion, amplification, or normal copy number along the whole genome at a certain resolution. To be manageable and easily comparable to other studies, the data were summarized to a rather course resolution at the level of the chromosome arm. (Recall that the nonacrocentric chromosomes have a long and short arm. In the present study data from two arms was available on 18 nonacrocentric chromosomes and data from a single arm was available on the other 5.) Thus the data may be arranged as $X_{i,j}$ where i indicates the tumor and j indicates the chromosome arm. There is interest in both the loss data and the gain data, but for simplicity let us discuss only loss information here, so $X_{i,j}$ is a Bernoulli trial indicating whether or not there is deletion of DNA on chromosome arm j in tumor i.

A simple barplot was used to graph the 41 summary counts $X_{\cdot,j} = \sum_i X_{i,j}$. High rates of deletion were observed on some of the arms, though their average value was $55/246 = 0.22$. A test of the null hypothesis that all Bernoulli $X_{i,j}$ had the same success probability is a way to measure the variation and to assess the significance of the high counts. The null hypothesis means that all losses are sporadic. The problem seems similar to the biodiversity problem discussed above, but there is an important difference which invalidates the use of a simple permutation procedure. The trials $X_{i,j}$ and $X_{i,k}$ are positively correlated if j and k index two arms of the same chromosome. This correlation is caused by the physical process by which deletion happens; i.e. some deletions involve the loss of a whole chromosome and thus the loss indicators are not independent. The form of dependence amongst the $X_{i,j}$ means that their conditional distribution given simple row or column totals is not exchangeable, and one is therefore not justified to shuffle the loss indicators to calibrate a test. Statistical modeling provides a solution.

Envision a partially observable random variable $Z_{i,c}$ for each tumor i and each nonacrocentric chromosome $c = (j, k)$. This Bernoulli trial indicates whether or not the measurements $X_{i,j}$ and $X_{i,k}$ are linked. If $Z_{i,c} = 0$, then assume $X_{i,j}$ and $X_{i,k}$ are conditionally independent records of deletion. On the other hand if $Z_{i,c} = 1$ assume that the arms j and k are linked and so either the whole chromosome is deleted $X_{i,j} = X_{i,k} = 1$ or the whole chromosome is retained $X_{i,j} = X_{i,k} = 0$. By the way that CGH data are recorded, one can observe $Z_{i,c}$ if either $X_{i,j} = 1$ or $X_{i,k} = 1$, but it is unobservable when no deletions occur. Clearly the introduction of these latent variables induces statistical dependence amongst the observed counts $\{X_{\cdot,j}\}$.

A statistical model is a complete description of the joint probability distribution of the data. The model is set up on a general alternative hypothesis and used to derive a likelihood ratio test of the sporadic loss null hypothesis. The *hot-cold* model asserts that some of the arms are *hot* in that their marginal loss rate β is higher than the marginal loss rate α of other *cold* arms. Unlinked loss events on a hot arm are Bernoulli trials with success probability β, and unlinked loss events on a cold arm have rate α. One way to deal with linked losses is to say that the common Bernoulli trial is hot if and only if either of the linked arms is hot. There is a parameter τ which is the probability that two arms on the same chromosome are linked (i.e. it is the success probability of $Z_{i,c}$). All that remains to fully specify the joint distribution of data is the location of hot and cold arms. With so many arms (41) compared to the amount of data, it seems quite reasonable to treat these locations as random effects. In other words, a given arm is hot with probability η, and otherwise it is cold. The hot-cold model has four parameters $\theta = (\alpha, \beta, \tau, \eta)$ that characterize the rate of deletion for cold and hot arms, the rate of linkage, and the overall rate of hot arms.

The likelihood function is a product across the $n = 6$ tumors and the 23 chromosomes:

$$L(\theta) = \prod_i \prod_c P\left(\text{data}_{i,c}\right)$$

The marginal probability of $\text{data}_{i,c}$ is a mixture over the latent $Z_{i,c}$ and a mixture over the possible patterns of hot and cold arms for that chromosome. Take, for example, the record $(0,1)$ from one nonacrocentric chromosome in one tumor; i.e. one arm is deleted and one is retained. The probability of this record according to our model is

$$\bar{\tau}\left[\bar{\alpha}\alpha\bar{\eta}^2 + \bar{\alpha}\beta\eta\bar{\eta} + \bar{\beta}\alpha\eta\bar{\eta} + \bar{\beta}\beta\eta^2\right]$$

using the notation $\bar{x} = 1 - x$. The rationale for this mixture formula is that having observed a discrepant result between the two arms it must be that they are unlinked, and hence $Z_{i,c} = 0$ which happens with probability $\bar{\tau}$. The four additional terms indicate the probability of $(0,1)$ under the four arrangements of hot and cold. Evidently there is no simple formula for the maximum likelihood estimator $\hat{\theta}$, though it can be found by straightforward numerical methods.

The null hypothesis of sporadic loss is equivalent to the assertion that the hot and cold rates are equal; i.e. $\alpha = \beta$. Interestingly, in this case the rate η is unidentified so its value is immaterial to the null-restricted likelihood $L_0(\theta)$. Using the Yeager-Reznikoff data it is estimated that $\hat{\theta}_0 = 0.22$ and $\hat{\tau}_0 = 0.28$ on the null hypothesis. The maximized value of $\log L_0$ is -125.3. With the full model the estimates are $\hat{\theta} = 0.05$,

$\hat{\beta} = 0.38$, $\hat{\tau} = 0.22$, and $\hat{\eta} = 0.47$, with a maximized loglikelihood of -120.5. Thus the observed value of the generalized likelihood ratio statistic is $-120.5 + 125.3 = 4.8$.

Statistical theory suggests that likelihood ratio statistic enables a powerful test of the null hypothesis of sporadic deletion. Often the this statistic is calibrated using Wilk's result that negative twice the log ratio converges in distribution to a chi-square random variable, but the validity of this asymptotic approximation is questionable in the present model. A potentially more accurate approximation is one based on bootstrap sampling under the null. This form of bootstrapping is a natural model-based alternative to a permutation procedure, and is justified when additional features such as dependence are accounted for in the analysis. Bootstrapping under the null amounts to simulating data sets according to the model in which $\alpha = \beta = 0.22$ and $\tau = 0.28$, and then recomputing the likelihood ratio statistic for each simulated data set. Contrary to permutation, the bootstrap simulation is not conditional on statistics from the data. It is approximate because it presumes knowledge of the parameters in the composite null. The intensive optimization which must accompany each simulated data set limits the size of the test sets. In a sample of 249 bootstrap data sets, none had as extreme a loglikelihood ratio statistic as the observed data. Thus a p-value of 1/250 can be reported, and it may be concluded that there is some evidence against the sporadic loss hypothesis.

Having gone to the trouble to structure elements of this interesting data analysis problem into a hierarchical statistical model, it is worthwhile to see what other conclusions can be drawn. Very often statisticians get hung up on hypothesis testing and do not focus the analysis beyond an initial question of structure versus no structure in the data. In the present model each chromosome arm is either hot or cold; the test asked if hot is different from cold, and assuming that it is one can assay each arm for the posterior probability that it is hot given the data. The posterior probability is proportional to η times the probability of the data on that arm given that the arm is hot. The calculation is complicated somewhat by the possible linkage with the opposing arm, but still it is routine to evaluate the resulting posterior probability. Only chromosome arm 8p, which had presented 5 losses out of 6 tumors, had a posterior probability exceeding 0.95 of being hot. In this case and elsewhere the processing of data from raw counts to posterior probabilities has the effect of reducing variation and highlighting interesting structure in the data.

5.3 Mouse Mutagenesis, Randomization Testing and Modeling

It is rewarding to be involved in a sustained long-term effort with a large laboratory to monitor, design, and analyze a complex experiment. Such has been the case in work with W.F. Dove's lab studying the normal and neoplastic growth of intestinal epithelium in a mouse model of human colon cancer. Central to the experiment is the MIN mouse, a mouse harboring a nonsense mutation in one copy of the Adenometous Polyposis Coli (Apc) gene, and consequently a mouse that spontaneously develops **m**ultiple **i**ntestinal **n**eoplasia. An inherited defect in the human APC gene predisposes carriers to the highly penetrant disorder familial adenometous polyposis (FAP) characterized by the development in the intestine of a large number of polyps. As colon cancers arise from polyps, FAP patients are at a much increased risk for colon cancer. Importantly, the APC gene is known to be defective in colon cancer cells from most sporadic cases of the disease, and thus APC is a central player in the molecular biology of colon cancer. Of interest to Dove's lab are other as yet unknown players that modify the effect of APC.

From a statistical perspective, the important features of the mutagenesis experiments performed in Dove's lab may be summarized rather succinctly. A measurement $X_{i,j}$ is made on the jth offspring of animal i. Typically, the measurement is a count of the number of intestinal tumors carried by the animal when it is sacrificed at three months of age. The animal (i, j) is a member of a pedigree generated by various experimental breeding schemes, but often the dependence structure induced by the pedigree can be treated in the following very simple way. The parent i may or may not carry an allele $*$. If it does not, then the tumor count $X_{i,j}$ of offspring j arises from a tumor count distribution $f(x)$ typical of the underlying genetic background. On the other hand, the parent might carry $*$ in which case $*$ is transmitted randomly to about half of its offspring by Mendelian segregation. A MIN animal j that also carries $*$ presents tumors counts $X_{i,j}$ that follow a modified distribution $g(x)$. For instance $*$ might tend to cause a reduction in tumor count and so $g(x)$ would be shifted to the left of $f(x)$ in some way. Because the tumor count phenotype is quantitative and exhibits a large amount of variation, and because the experiment involves a large number of potential carriers i, each with a modest to large number n_i of offspring, a statistical approach to the data analysis provides an efficient methodology for deriving inferences as the experiment proceeds.

A first question to consider when analyzing a branch of the experiment is whether or not there is any evidence against the null hypothesis $H_0 : f = g$. Basically, H_0 asserts that there is no modifier gene in the pedigree. Here conditional inference and randomization are critical. By conditioning on the set of measurements, or, in other words, their empirical distribution, what remains random is their assignment into subkindreds i. Permutation is used to shuffle observed tumor counts so that each parent i is assigned a random subset of size n_i drawn without replacement from the whole sample. A likelihood ratio statistic is calibrated by this permutation. Sometimes the set of offspring of i is further arranged in subfamilies, and then the permutation shuffles whole subfamilies to preserve within-subfamily correlation.

Again, randomization testing provides the first phase of statistical inference. With evidence that an allele $*$ may be modifying the phenotype distribution, one may go further with more detailed analyses. For example one may estimate the modified distribution or assay the posterior probability that the allele $*$ is segregating in subkindred i. This latter measure has become a very useful summary of the set of data taken on animal i, and it provides a means for directing subsequent experimental crosses. In the simplest case, the probability of data $x_i = \{x_{i,j}\}$ from parent i is

$$p(x_i) = \begin{cases} \prod_{j=1}^{n_i} f(x_{i,j}) & \text{if no } * \text{ in } i \\ 2^{-n_i} \prod_{j=1}^{n_i} [f(x_{i,j}) + g(x_{i,j})] & \text{if } * \text{ in } i \end{cases}$$

and so the posterior probability of $*$ in i is computed from Bayes rule as soon as estimates of $f(x)$ and $g(x)$ are in place. An obvious model for tumor counts has f and g being Poisson distributions, but these turn out to be inadequate. Parametric analysis based on a generalization, the negative binomial distribution has proven to work well. Nonparametric analysis based on stochastic ordering of f and g is another approach that relies on fewer modeling assumptions. In either case, maximum likelihood may be used to estimate f and g.

The likelihood function here has the form

$$L(f,g) = \prod_i \left\{ \frac{1}{2} \prod_{j=1}^{n_i} f(x_{i,j}) + \frac{1}{2} \prod_{j=1}^{n_i} \frac{f(x_{i,j}) + g(x_{i,j})}{2} \right\}.$$

In order to fit the model and to evaluate the likelihood ratio statistic mentioned above, computational methods are needed to maximize $L(f,g)$.

The mutagenesis case study requires randomization to test for the presence of modifier genes and also modeling to quantify information in the data about likely carriers of important alleles.

5.4 Gene Expression Data Analysis: Hierarchical Modeling

The recent advent of DNA microarray technology is allowing unparalleled investigation into the molecular biology of the cell, and, simultaneously, is creating intriguing new statistical data analysis problems. The following data structure is not uncommon. A gene expression measurement $X_{i,j}$ is obtained for gene i in sample j, where $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, and $m >> n$. The value $X_{i,j}$ is a surrogate for the abundance of messenger RNA transcribed by gene i within cells j. In one experiment from M.N. Gould's laboratory, the cells are rat mammary epithelial cells taken from rats of different inbred strains which vary in their susceptibility to mammary cancer. In another experiment with D. Jarrard, the cells are cultured prostate epithelial cells sampled at different stages of proliferation. A basic problem is to understand patterns of differential gene expression among the different conditions that characterize the contributing cells.

Both technological and biological sources of variation affect the measured abundance $X_{i,j}$, but in a well calibrated experiment one can hope to compare the measurements from different samples j (i.e. different microarrays). Suppose that $j = 1, 2, \ldots, n_1$ index measurements from cells of one condition, called A and $j = n_1 + 1, \ldots, n$ index measurements of the second condition B. To assert differential expression of gene i between conditions A and B is to make the inference that the observed differences between $D_A = \{X_{i,1}, \ldots, X_{i,n_1}\}$ and $D_B = \{X_{i,n_1+1}, \ldots, X_{i,n}\}$ correspond only to chance fluctuation and do not represent underlying shifts in expression level.

Some authors measure the group difference using a t-statistic for gene i, perhaps after the data are transformed to the logarithmic scale. Differential expression is concluded if the observed statistic lies in a tail of the t-distribution. This straightforward approach treats all the genes separately and does not directly address the question, "what genes are most differentially expressed." In part, the testing scenario is concerned with the rather extreme straw-man hypothesis that no genes are differentially expressed. Of course what makes the cells different is the expression of their genes and so we can reject the global hypothesis out of hand. Accepting that some genes are differentially expressed and some are not, the statistical problem can be approached differently by using a discrete mixture model. The problem is to estimate the set of differentially expressed genes and to compute for each gene the posterior probability that it is differentially expressed. An advantage of this approach is that information can be shared among genes, which is helpful because the number of samples n is typically so small compared with the number of genes m.

Consider that there is a latent variable Z_i for gene i which indicates whether or not the gene is differentially expressed between conditions A and B. A proportion θ of the genes are differentially expressed; i.e. $\theta = P(Z_i = 1)$. A model parameterizes the probability of data $P(D_A, D_B|Z_i)$ for each gene and gene-specific inference is based on the posterior probability $P(Z_i = 1|D_A, D_B)$ that is evaluated using estimates of the model parameters. More specifically, the model $P(D_A, D_B|Z_i)$ may be structured as a mixture over the unknown levels of expression μ_A and μ_B; i.e. $\mu_A = E(X_{i,j})$ for j in the first group of samples and μ_B is similarly defined for the second group of samples. A Gamma model for $X_{i,j}$ given μ often fits well and is both analytically and computationally efficient. If there is equivalent expression, $Z_i = 0$, and $\mu_A = \mu_B$, and this common, unknown, gene-specific mean is presumed to fluctuate across the set of genes according to some distribution $p(\mu)$. Positive correlation within D_A and D_B is induced by integrating with respect to $p(\mu)$. On the other hand if there is differential expression, one assumes that independently the two conditions A and B draw means μ_A and μ_B from $p(\mu)$. Thus the modeling is hierarchical, with three basic stages: measurements $X_{i,j}$ given mean values; differential expression Z_i of the mean values; and level of the mean values μ_A, μ_B.

Randomization could be considered in this case study, but, as a matter of global evaluation, it is testing a hypothesis that one can reject out of hand. Statistical modeling allows the investigator to quantify the evidence for differential expression of each gene and to estimate the fraction of differentially expressed genes. Contrary to the other case studies, in this one the analysis includes a large number of unobservable variables that are linked through the statistical model.

6 Closing Remarks

Statistical thinking has for a very long time had an important role to play in the basic and laboratory sciences. One reason is that the theory of mathematical statistics and probability demonstrates connections between diverse data analysis problems owing to common features of measurement variation. Also, the methods of statistical analysis provide ways to extract information from data and include quantitative assessments of uncertainty. The use of statistical methods continues to be important because technological advances have dramatically increased the amount of data that investigators can obtain and store, and challenging data analysis problems are ever present. Certainly in biological problems improvements in measurement technology do not obviate the need for statistical thinking because natural sources of variation cannot be dampened out. Statistical modeling may now have a greater potential since quite sophisticated models can be entertained and fit thanks to advances in computing.

Statistical research has given investigators an impressive array of tools to process data. Yet, theoretical research has been much less concerned with wholesale strategies for data analysis than it has been concerned with optimal properties of particular steps, and so it can be difficult for the user of statistical methods to know how to proceed. The most important thing, perhaps, is that the statistician becomes immersed in the context of a problem. This may assure an effective data analysis strategy and that something gets done which is helpful.

Acknowledgements: I thank Dr. X. Lin for giving me the opportunity to contribute to EOLSS, and my many collaborators without whom this work could not have taken form.

Bibliography

P. Armitage, 1952. The statistical theory of bacterial populations subject to mutation. *Journal of the Royal Statistical Society. Ser. B (Methodological)*, **14** (1), 1-40. [*This is an excellent introduction to the statistical issues surrounding the Luria Dellbruck fluctuation test.*]

J. Besag and P. Clifford, 1989, Generalized Monte Carlo significance tests, *Biometrika*, **76**, 633–642. [*This paper clearly describes the structure and application of Monte Carlo tests both generally and as applied in several spatial problems.*

G.E.P. Box, W.G. Hunter, and G.S. Hunter. *Statistics for experimenters*, Wiley: New York, 1978. [*This book is a classic introduction to statistical issues in both the design and analysis of experiments affected by multiple sources of variation. Emphasis is on linear statistical inference.*]

W.F. Dove, R.T. Cormier, K.A. Gould, R.B. Halberg, A.J. Merritt, M.A. Newton, and A.R. Shoemaker (1998). The intestinal epithelium and its neoplasms: genetic, cellular, and tissue interactions. *Phil. Trans. R. Soc. Lond. B*, **353**, 915-923. [*This review paper summarizes the state of knowledge regarding the cellular dynamics of the intestinal epithelium; emphasis is on mouse experiments to understand the factors affecting cancer growth in this tissue.*]

M.M. Fisher, J.L. Klug, G. Lauster, M.A. Newton, and E.W. Triplett, 2000. Effects of resources and trophic interactions on freshwater bacterioplankton. *Microbial Ecology*, 40:125-138. [*This research article describes Dr. Fisher's work characterizing bacterial biodiversity in a certain freshwater environment.*]

M.A. Newton, C.M. Kendziorski, C.R. Richmond, F.R. Blattner, and K.W. Tsui, 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology.* **8** (1), 37-52. [*This research article introduces the empirical Bayes statistical methodology for the analysis of high-throughput -gene expression data.*]

M. A. Newton, T. Yeager, C.A. Reznikoff, 1999. A statistical analysis of cancer genome variation. In, *Statistics in Genetics, IMA Volumes in Mathematics and its Applications*, M.E. Halloran and S. Geisser (eds), **112**, 223-236, New York:Springer. [*This article reviews the two-step instability-selection probability model as it applies to data measuring genomic aberrations in cancer. Emphasis is on data from chromosome-based comparative genomic hybridizations taken in a bladder cancer study.*]

J. Pickands III, 1975. Statistical inference using extreme order statistics. *The Annals of Statistics*, **3**, 119-131. [*This is an example of work which considers the extreme-value trinity theorem.*]

S. Sarkar, 1991. Haldane's solution of the Luria-Delbrück distribution. Reprinted in, *Perspectives on Genetics*, pp 199-203, J.F. Crow and W.F. Dove editors. The University of Wisconsin Press, 2000. [*This excellent review of research on the Luria-Delbruck problem emphasizes both statistical and biological aspects of the problem.*]

T.R. Yeager, S. DeVries, D.F. Jarrard, C. Kao, S.Y. Nakada, T.D. Moon, R. Bruskewitz, W.M. Stadler, L.F. Meisner, K.W. Gilchrist, M.A. Newton, F.M. Waldman, and C.A. Reznikoff (1998). Overcoming cellular senescence in human cancer pathogenesis. *Genes and Development*, **12**, 163–174. [*This research article describes experiments with human cancer cell lines and manipulations aiming to understand rate-limiting factors affecting cell division.*]