Lady Tasting Tea paper

BE.104 Spring 2005

Sueann Lee

The title of the book undoubtedly first caught my attention. What is this book about, and what is it doing in my list of books for a class on toxicology and chemicals? Going on, I found that the first paragraph of the book is all about an anecdote of how a lady insisted that she could tell the difference between tea where milk was added first, and tea where milk was added last. How can we tell? This brought out one of the main characters in the huge, brilliant cast of the book -- R.A. Fisher, and his method of hypothesis testing. For such a minor problem, one might doubt the necessity of invoking statistical methods to tell differences between hypothesis, but upon further reflections it should dawn upon us that many problems in science involves hypotheses that need to be validated by experimental data, and it may be more difficult to reach the conclusions then we think given that no experiment is ever 'ideal'. The 20th century saw a huge increase in scientific research in fields ranging from agriculture to medicine to different fields of engineering. As scientists came up with ever more hypotheses, it became increasingly evident that science is inextricably linked with mathematics. As modern medicine emerge, statistics steadily attached itself to the work of medicine and ultimately came to sit in judgment on it, literally, gaining public acceptance as ‘the way’ for sound science.

What is statistics? The word ‘statistics’ seem to have two meanings – the use of numbers to describe whole patterns of activity, and the analytic methods that allow manipulations of such data. Author David Salsburg believes, rightfully, that the public is not fully aware of the degree to which recent developments in statistics impact the way we perceive the world. True, I concede that before taking this course and reading this book, I looked at statistics as a tool that is, although useful, rather dispensable, being not at the heart of the scientific question. I also tended to think of statistics as a rather negative tool that is easily abused by media and propaganda, used by people who are intent on using a tool without considering the implications or liabilities of using that tool. What I missed is that statistics is essential for deciphering useful conclusive data amongst random scatter. The idea that “whatever we measure is really part of a random scatter” is central to all science and one of the central ideas that started off the statistical revolution beginning with Pearson and carried forward by generations after generations of brilliant statisticians and mathematicians.

The book established the importance of statistics through many examples of revolutions in fields of science and technology brought about by statistical models and methods. Fisher, in his early days before he was an established statistician, dug through ninety years of experimentation that was, for the most part, “a mess of confusion and vast troves of unpublished and useless data”, and with original statistical methods distilled out his “Studies on Crop Variation”. The early statisticians had numerous other applications for their novel methods: Pearson examined historical issues such as lengths of reigns of kings, and “Jewish-Gentile Relationships”, concluding the racial theories of the Nazis to be sheer nonsense. Florence Nightingale once wrote that, "To understand God's thoughts, we must study statistics, for these are the measure of his purpose." The famous nurse, I found from the book, was a self-taught statistician who invented the pie chart to assert to army authorities how deaths due to untreated wounded soldiers contributed significantly to overall casualties. I cannot highlight all the important, humorous or interesting examples the book gave for historical uses of statistics by the great minds, demonstrating the importance of statistics. Without statistics, we cannot establish whether an experiment showed anything, or compare between sets of data with any degree of certainty. Through a series of vignettes about the individuals who asked questions and devised methods for quantitative manipulations and analytical interpretations of statistical data, Salsburg introduced to the reader the quiet but profound revolution that may be unnoticed. As he pointed out, the twentieth century saw the fading of a deterministic outlook and the rise of a statistical and probabilistic way of looking at the world

Many of the short chapters of the book were devoted to stories of individuals, highlighting the ‘great men’ and their ‘great ideas’ and ‘great questions’. This was a savory way to follow the evolution of the statistical development – the mean and the standard deviation, parameters that we so know and love, are mathematical abstractions that are the brainchildren of Karl Pearson, one of the first great and stubborn statisticians. Another great and stubborn statistician is Fisher, who we know from the common hypothesis testing. Fisher contributed much more to the field of statistics: he introduced randomized controlled experiments, analysis of covariance, the concept of degrees of freedom, as well as correcting many of his older rival, Pearson’s, mistakes. How statistics transformed from the deterministic worldview in the nineteenth century is clearly portrayed in the book through the philosophical “Fisher versus Pearson”: Pearson viewed statistical distributions as describing the actual collections of data eh would analyze. For Fisher, the true distribution is an abstract mathematical formula, and collected data can only be used to estimate the parameters. Such historical descriptions contributed to the emerging clarity with regards to the meanings of such concepts for the non-mathematician reader.

The book allowed us to savor the style and personality of statistical heroes and about heroic ideas while covering most of the interesting histories of these heroes. All who have used Student's t-test can now meet Gosset (who I now learn is the real identity of 'Students', and why is that so) working in the brewery on ways to make fermentation consistent. A nice presentation of the history behind the theory of distributions of extremes also illustrates the contribution of Tippet, Fisher and Gumbel. Interesting sections on work such as Kolmogorov's on probability theory and axioms of probability, or Bliss's on probit analysis allowed the reader not only to learn about statistical methods in their development, but also ask oneself the important questions like "what is probability in real life?"

The book brings to live each piece of development. As we learn about Fisher's development of 'p-value' significance testing, of Pitman establishing the robustness of non-parametric methods, of Neyman and the younger Pearson developing hypothesis testing, we experience vicariously the joy of discovery and witness how more widely applicable methods emerged from solutions to particular problems. The methods of science are often developed to answer a question in hand, and it seems that with special insight found in these geniuses generalizations can ensue.

The author added all sorts of little life stories, anecdotes, quotations in a series of historical and biographical sketches such that the lives of the contributors to the field and the unfolding of probabilistic and statistical ideas were presented in an intimate way. The reader may feel as if he was present whenever something relevant to statistics had happened. Salsburg places the contributors in the context of his times and situations such that we may understand the historical background. He described how the anti-intellectual atmospheres in the Nazi and Soviet regimes had a stifling effect on the quest for statistical understanding. In Germany, many were forced to flee the anti-Semitic terror, while in the Soviet Union, communist dogma prevented the brilliant work of her scientists from being fully appreciated. The life stories and personal and intellectual history were often touching, such that there is a feel of an epic novel involved, and reminding one of the funny reality mix-up of mathematics and the real, physical world.

One story that carries across multiple chapters is that of the antagonism between Pearson, the 20th century survivor of 19th century authoritative mathematicians and founder of *Biometrika*, and Fisher, the genius founder of modern statistical methods. Later in the book, Fisher excoriated Neyman and young Pearson in a historical echo of the way Pearson treated Fisher, a good illustration that perhaps not all great minds are from great humans.

The various tales herein are ultimately unified in a single theme: the conversion of science from observational natural history into rigorously defined statistical models of data collection and analysis – exactly what the title 'How statistics revolutionized Science in the Twentieth Century" meant. The book, on a deeper level, is a forceful assertion that statistics has arrived as a theoretical and applied science that experienced a profound shift in basic philosophy. It was very interesting indeed to keep pondering on such 'problems of philosophy' that were brought up through the discussions on the great minds of the field. I felt like I gained a better general idea of the complex thoughts behind "statistical thinking". The reader is ever reminded of how even medical statistics has an abstruse theoretical base. The big picture is sketched from the ground-level probability distributions that Pearson described around the turn of the century, to things truly abstract such as Heisenberg's Uncertainty Principle – concepts we may not, at first glance, associate with medical research concerning yes or no answers.

The author has his own interesting issues on various issues as well. For example, there was a short discussion on the "intent to treat" rules of evaluating every patient assigned to a treatment group, even if they changed treatments. This was indeed a very strange method of analysis, and without the clear explanation and discussion I would not have understood at all the justifications.

The book is clear in its reconstruction of the interlocking histories of superior minds grappling with theoretical and applied problems, following one thread at a time so that the book is easy to read. However, there is one major problem with the book: its lack of equations. Although this is obviously the authors intent, such that the lay person may not be intimidated, the book's more appropriate audience should be those with at least some statistical background. If you are not interested(at least slightly) in statistics, why should you be interested in the book? Multiple concepts were attempted to be explained in words, but mathematics just gets even more abstract at times with only words. If I had not been introduced to some basic methods of statistics in class, I wonder if I would have understood a single thing. Simply telling the reader who developed maximum likelihood and t-statistics is not very helpful if the reader doesn't conceptually understand what they mean. It would certainly have helped if the book is more mathematically or philosophically rigorous, or if I had taken one more semester of statistics. There were numerous technical terms that were insufficiently explained, and narration on the development of the great ideas of statistics does little to help the reader grasp these mathematical concepts. Other examples of seemingly unsupported assertions include calling the law of cause and effect "a vague notion that will not withstand the battering of pure reason" – but what does "pure reason" happens to be? Some of the basic questions about statistics are not brought up. It is almost as if you were being taught how the Bohr atom evolved into the electron cloud model into wave functions of the atom, without even knowing what exactly is an atom. Presenting some mathematics would definitely aid the curious mind to take awe of the beautiful, sinuous mathematics Overall, in a vivid and clear style, Salsburg delivered an almost personal acquaintance with men and women of the past, their personalities, revolutions in the subject, and real world problems. The presentation of the statistical revolution and its relationship with science was lucid, but introduction of equations into the book would substantially improve the enjoyment that can be elicited from the book.

After reviewing the book, one can easily see the wide applications of statistics in science. Here I shall review a study on the effects of an environmental contaminant on public health, which has made use of statistical methods.

The study, published as a paper in March 2001 in the journal Environmental Geochemistry and Health, investigates the influence of environmental PCB contamination on public health in the city of Serpukhov, in the Moscow region of Russia. Chlorinated biphenyls(PCBs) appear to be among the most dangerous contaminants of the ecosystem that includes the urban areas and lands which are actively used for vegetable production. In the study, special attention to the environmental situation in Serpukhov has been focused on a condenser factor, where, from 1967 until 1987, condensers were manufactured using mixtures of PCBs with chlorine contents up to 75% of saturation. The amount of PCBs used annually during the 20 years was about 1.4 tons per year. Previous studies has shown that PCB concentrations in the soils and vegetables grown on the soils, and also in surface waters located within 2km from the plant were much higher than maximum permissible levels(Bobovnikova et al, 1993, 2000). Furthermore, extremely high concentrations of PCBs in breast milk of women working at the plant and/or living next to the factory has been observed(Bobovnikova et al., 1993). This route of exposure for nursing infants could continue for many years after the mother's PCB exposure due to the long half-life of PCBs in the body. Because of the great potential for bioaccumulation of PCBs and the actions of sediments in aquatic systems as a reservoir, and their toxicity and carcinogenicity for humans, there is a health risk for the Serpukhov population from exposure to PCBs.

PCBs are known to be persistent in the environment and toxic to animals and humans. The adverse effects of exposure to PCBs range from skin irritation; acne; rashes; nasal irritation; lung irritation; and liver, stomach, kidney and thyroid damage to damaged reproduction, growth and development; cancer; and death. The International Agency for research on Cancer and the US Environmental Protection Agency both consider PCBs to be probable human carcinogens. In addition, PCBs have been identified as neurotoxins. The aim of the study was to analyze human health in Serpukhov City, Russia, and to assess the potential risk to the local population from exposure to industrial pollution. Their main objective was to estimate the contribution of a separate environmental factor(soil contamination with PCBs) to the total level of illness of the city.

In general, the study has been well designed, conforming to principles mentioned in the book about experimental design(such as Fisher's ideas presented in the beginning of the book) for furthering knowledge. In order for the region to be analyzed differentially, it was subdivided into five districts, each served by a separate medial clinic. The basic demographic parameters(birth rate, mortality and natural increase) are available officially, so data on morbidity for 10 nosology groups within a 6 year period were counted per 100,000 persons per year. The ecological situation in Serpukhov have also been evaluated. Soil contamination data for 1991-1999 were used to calculate the concentrations for PCBs absorbed into the food chains.

The overall goal of establishing whether there is potential correlation between PCB contamination and various diseases is broad, so the study rightly divided into 2 parts: 1) comparative analysis of human health indexes; 2) calculation of cancer risk from soil contamination with PCBs. The first part is further divided into a statistical analysis of morbidity data, and a special cohort research with two groups of school children.

In order to compare the morbidity amongst the different areas of the city, a special data set was created where different areas of the city, regarded as observations, were characterized by 10 parameters on 10 classes of the diseases regarded as variables. This allowed for easy comparison across the 5 districts and also for different diseases.

The researchers rightly pointed out that to make a valid comparative analysis using the morbidity data, there are two factors whose influence has to be taken into account: dynamics of morbidity parameters over time, and compatibility of the data from the different clinics. The first issue is addressed through a significance test with a one-factor analysis of variance of disease level(discussed below). The problem of compatibility can be expected to exist in practice, and as swn by Andrews et al.(1972), the corresponding statistics of small sample processing can be improved by using robust estimates. The authors purported to have applied these in their calculations, but how it was applied was not mentioned so one can only accept the authors words with faith. If they are correct, robust estimates are useful, as demonstrated by George Box. Box developed the idea of robustness, where methods can still be useful even if the conditions for assumptions about the distributional properties of the data in the statistical methods do not hold.

The whole set of data as divided into two 3-year sets. For each disease, computed ranks characterize the morbidity level in investigated areas by an arithmetic mean of the ranks of the two 3-year sets. This minimizes possible distortions connected to the character of the statistical record keeping in different clinics.

As stated in the Lady Tasting Tea, variations in data are inherent and need to be properly analyzed for meaningful results to be obtained. Variations were indeed addressed in this study: by comparing the ranks with the illness parameter across different clinics for each disease and doing a significance test. The p values for each disease were all found to be small(except for respiratory diseases), and the R^2 value quite large(again except for respiratory diseases). The significance of the significance level and regression is to validate the assumption of keeping the factor of the clinic to be constant. What this demonstrates is that the significance of the factor, 'clinic', for all nosology groups – the fluctuations in the disease level from year to year are smaller than the distinctions caused by the clinic location and hence by the area in which people live. In other words, we can safely compare regional differences in morbidity which will presumably be due to the environment as part of the hypothesis.

One area of the city, the area served by clinic 2, is much more contaminated than the others. To establish whether there is a linkage, the next step of the analysis was a comparison of means for each group of diseases in order to show up significant differences in values. The researchers used the multiple comparison method of the GLM procedure, and results showed that for a number of diseases the maximum levels of sickness are concentrated in clinic 2, while the minimum levels are located in clinic 1, which is characterized by the lowest environmental contamination. Thus, the hypothesis that the difference can be attributed to ecological situation is plausible. Having a plausible hypothesis to test upon is very important.

To test this the researchers did another study, where they also cleverly removed other confounds such as occupational exposure by selecting for the right subjects. A special cohort research within two groups of school children has been carried out. The school children have been chosen to eliminate a possible influence of contaminant exposure through occupational activity. The children were also chosen to exclude the influence of family occupational activities on their health, and children who lived in the destined areas less than 7 years were excluded. One group of children lived in the area mostly contaminated with PCBs(served by Clinic 2) and the second(control) group lived in areas mostly clear of PCBs. An assumption has been made that the adverse effects of exposure to other contaminants are considered negligible for the two groups. One potential flaw of this study may be that it never justified this assumption.

The number of disease cases for 3 years, the occurrences per 100 persons, and the standard errors have been calculated using the corresponding estimations of the parameters for the dichotomous variables. The authors estimated the standard errors assuming a binomial distribution, which is reasonable given that these diseases are rare events. However, it was not clear to the reader why they used the 'estimation based on a Poisson distribution to calculate intensive parameters.' The application of two distribution models was confusing to the reader. Other than that, the application of statistics for this analysis is clear and sound. Student's t-test was used to calculate significance levels of the differences in occurrence for the two groups. It appears that for p values less than 0.002, the authors considered the difference significant. The application of Student's t-test is most suitable and demonstrates the level of chronic diseases in the first research group is higher than the second control group, by different ratios(total level of acute disease: 1.6 times) depending on the disease, the trend holding for acute respiratory diseases and acute disease of digestive organs as well as acute allergy diseases. However, the authors did not note that fairly high variance was found for respiratory diseases in different years, which may undermine the validity of the correlation for respiratory diseases and exposure. The level of chronic diseases was also found to be significantly higher for the first research group. This demonstrated the power of statistical analysis.

The second part of the research dealt with the calculation of cancer risk from soil contamination with PCBs. Previous sampling (Bobovnikova et al., 2000) have shown that

PCB concentrations in the soils of many private houses and gardens in the served by Clinic 2 are much higher than the maximum permissible level. The PCBs are known to enter the human body mostly by the inhalation pathway and with foodstuffs containing fat and to a lesser degree wit vegetables. PCB accumulation in the vegetables grown in the area have been previously estimated as a range for several vegetables. An assumption was made in the present study that a large portion of the locally produced vegetables is consumed by the residents of the Clinic 2 area and makes up a large share of their diet. The validity of this assumption was not addressed, however.

While the authors acknowledged that data available are not adequate for a complete statistical analysis over time of the relation between exposure to PCBs and the rate of occurrence of diseases, and no dose reconstruction study over time to determine the amount of exposure during the years when the plant was operating, nevertheless they have undertaken an analysis to compare calculated risk and observed occurrence of cancer. This is justified since PCBs have a fairly long residence time in the environment and present PCB concentrations in the environment is known. Therefore, the cross-sectional-retrospective study may be conducted looking at their current exposure. The analysis has been limited to an estimate of the increased risk due to dosage now being received from consumption of vegetables produced in contaminated soils.

To estimate the pollution of vegetable products grown in private units, they have assumed that the food concentrations can be calculated by multiplying the soil concentrations by a factor of 0.05-0.3. However, they did not include errors for this number, so does that mean this range is a 95% confidence interval? This is unclear. However, this assumption may be justified, since the calculated PCB concentration is “more or less commensurate with the real data from vegetable sampling” according to the author.

Another assumption that was made in their calculations is that the people living in private houses and/or having vegetable gardens in the Clinic 2 area consume about 75% of daily diet products grown on contaminated lands. The range of daily consumption of PCB for the group is thus calculated from the above data. Similarly, the possible daily consumption of PCB for the population living in ‘the most ecologically favorable’ area served by clinic 1 was also calculated. The above calculations are reasonably followed although some errors are missing; however, it may be too difficult to provide such errors since they may be rough estimates and how they obtained them are unknown.

The cancer risk from PCBs exposure has been calculated using the average mass of an individual as 65 kg nad the average life span as 70 years. However, it was not clear how these were used in the calculations and how variance in the many parameters were taken care of. True, as Fisher said, randomizing in the populations will take care of a lot of these problems, but there is inherent bias that is not taken care of. People living in different areas are likely to have different socio-economic statuses that are confounding factors. This problem of bias is inherent in the entire study, although the investigators have already tried their best to minimize them. On a side note, it may help if the authors also calculated attributable risk of cancer, to present a clearer picture with this statistic.

The results are that cancer cases per 1000,000 persons per year in clinic 2 shows more than a 100 percent excess when compared to the other 4 clinics. It is hypothesized that the excess is at least in large part due to exposure to PCBs. The effect of continued exposure through consumption of locally grown food is much smaller. Note that it is still not possible to 100% establish exposure to PCB as THE cause for cancer/diseases since such statistical tests cannot prove an hypothesis in such a way. However, this does add substantially to evidence of a correlation between exposure to PCBs and diseases.

In conclusion, the study was well thought-out and used many statistical tools appropriately. The comparative study allowed for a quantitative comparison of risk by high levels of exposure to PCBs. Comparison of morbidity between two groups of schoolchildren served by clinic 2 but with difference in residence in high or low PCB contaminated areas further factored out differences in clinics and in the overall area attributes such that overall social and residential conditions in the two groups are more similar(although power is decreased by decrease in sample size). Thus differences in morbidity may be attributed to the differences in exposure to PCBs. Cancer was also found to be in excess in the area served by clinic 2 by calculating the PCB dose exposure average for each area and the individual risk and prevalence of the disease for each population. This also indicated a significant increase in cancer rate. Thus, the study is conclusive in adding evidence to the hypothesis of adverse health effects of environmental PCB exposure.