МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2005
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2005
Proceedings of the Thirty Fourth Spring Conference of
the Union of Bulgarian Mathematicians
Borovets, April 6–9, 2005

A HISTORY OF STATISTICS IN THE LAST 100 YEARS

George Dimakos, Tsigoni Anastasia

In this article, we follow the history of statistics in the last 100 years. We begin with the work of Karl Pearson and the concept of the significance test. Next, we pass to Student and his distribution and we reach R.A. Fisher's concept of likelihood which is presented and we discuss the influence of Fisher's books. Sampling is next as it emerged from the works of Neyman and Pearson. Decision theory and Wald's ideas come next, with the extension of Savage. Recent developments are also treated, such as computational methods, Bayesian methods, sequential analysis, stochastic processes and experimental design.

Suppose that each number of replicated observations can fall into one of the n classes and that, according to (same) theory, E_i is the number expected in class i. Then a are observed that fall into that class, Pearson proposed using $\sum_i \frac{(O_i - E_i)}{E_i}$, to test the adequacy of the theory or, as we say nowadays, to test the hypothesis. Large values of k^2 would lead one to suspect the theory. In the last decade of the 19^{th} century, two English statisticians, Karl Pearson (1857–1936) and his student G. U. Yule (1871–1951), did further (research) work in showing how to use statistics to come to definite conclusions about the relationships between several quantities. Pearson obtained the approximate distribution of χ^2 when the theory was true, the null hypothesis. The approximation, especially with minor changes, has turned out to be remarkably accurate. It is usual to select a small probability a, to use this distribution to calculate x^2(a) satisfying $p(x^2 > x^2(a)) = a$ and eject the null hypothesis if the criterion $\sum_i \frac{(O_i - E_i)}{E_i}$ exceeded $x^2(a)$. Thus Pearson produced a typical example of a tail-area significance test, in which a is rejected if the tail of the null distribution beyond the observed value is less than a prescribed, small number. In 1893, Pearson developed the so called χ-square statistics as a way of measuring the relationship between two quantities. χ-square has many useful properties, for example, under many circumstances it is additive. It can easily incorporate composite hypothesis wherein the E_i are only partially specified. Today Pearson's test is greatly used and widely accepted as sound by almost all theories. In 1908 Student, the pseudonym of W. S. Gosset obtained another distribution that was to endure. Suppose that n replicated observations k_i, measurements now, not counts, were, on some theory, supposed to have meaning M. It had been common since the time of Gauss, to use the arithmetic meaning of the data $\overline{x}$, and compare it with μ by the criterion

$\frac{\sqrt{n}(\overline{x}-\mu)}{\sigma}$, where σ is the standard deviation of an individual, so that $\frac{\sigma}{\sqrt{n}}$ is that of $\overline{x}$. With large samples this was known to be normally, distributed, with zero meaning if the theory was true, and unit standard deviation. In the agricultural work that concerned Student, there are two difficulties: first, the samples were small, second σ was unknown. The latter could be overcome by estimating σ^2 by $s^2 = \sum_i \frac{(x_i - x_i)^2}{n-1}$ and replacing σ by s obtaining the criterion today almost always denoted by $t = \frac{\sqrt{n}(\overline{x}-\mu)}{s}$. The difficulty Student faced was that now, the distribution of t was unknown, especially for small n. He found the exact distribution though the proof was not rigorous and he had a stroke of luck. But with the distribution, he could construct a tail-area significance test, just like Pearson's, a test that is widely used today. Student's paper was important because it inaugurates the field of small sample statistics, in which the earlier asympotics ($u \to \infty$) were replaced by exact probabilities. Since then considerable mathematical skill has gone into the determination of exact distributions. In 1915, R. A. Fisher obtained the distribution of the correlation coefficient using sophisticated, geometrical ideas. There are historians who believe that events are determined by individuals. In the 1920s in Statistics, this belief was proved to be true, with R. A. Fisher utterly changing the subject. In rapid succession he published numerous papers and two books, all containing brilliant, operational ideas. One of these was likelihood. The model that is commonly employed in modern statistics supposes the data x are governed by a probability distribution, which is known apart from a finite set of real numbers measure, there is a probable density $p(x/\theta)$, where p is known, x will be a finite set of real numbers, thus with respect to some dominating measure these are the observations, but the parameter θ remains unknown, congruency inferences are to be made about it as function of x for fixed θ, p is positive, its total integral is one and $\int_A p(x/\theta)dx$ is the probability that x belongs to a set A. As a function of x for fixed θ, p is positive but its integrals $\int_\omega p(x/\theta)d\theta$ have no meaning. This function is the like hood ($f(\theta)$, given x). The value of θ that minimizes this function, denoted $\hat{\theta}(x)$, or $\hat{\theta}$ is the maximum likelihood (m.h.) estimate of θ. Fisher showed that has many desirable properties: it is typically superior to other methods of estimation and has an easily-calculated variance and asymptotic distribution. Fisher also introduced the enduring concepts of sufficiency and efficiency.

Concurrently with these theoretical advances, Fisher wrote a marvelous book, *Statistical Methods for Research Workers*, in which fairly simple, yet efficient, methods were put forward for problems of small samples. This book was widely appreciated by the scientists for whom it was intended and the techniques came into general use of it. The book lays less happily with mathematicians who found the explanations imprecise and regretted the lack of theorems. In fact, an industry developed out of trying to interpret Fisher rigorously. Except on of the matter of fiducial probability, he was always correct. The technique of least squares had been widely used since the time of Gauss. Fisher extended the idea by breaking up the variation presented in the data into additive components. For example, in an agricultural trial involving varieties and treatments, variations due to variety, to treatments and to evidential errors could be separated. The first two could

be compared with the last by a distribution and a test he introduced, which are now notationally named after him, the F-distribution and test. By this means, the apparent differences in treatments or varieties could be assessed as real, or accountable by technique error. This analysis of variance technique, mathematically based on a Pythagorean breakdown of sums of squares, stimulated many mathematical extensions of least squares. The distributions χ^2, t and F were all known to earlier writers. It was their contexts, and appreciations of their importance that were new.

Fisher's second book was also remarkable for the novelty of its ideas. The Design of Experiments after a brilliant chapter about a lady tasting tea, deals with the problem of how scientists should experiment, demonstrating that it is more efficient to vary all the factors at a time. The designs were enthusiastically adopted in agriculture and later in biology and medicine, but largely ignored by physical scientists who relied on big experiments. Mathematicians delighted in Fisher's designs because they used known, combinatorial results and stimulated new ones. For example, R.C. Bose achieved to be on the front page of the *New York Times* by constructing two orthogonal Latin squares of Side 22, thus proving Euler's conjecture false. The analysis of variance was basic to the interpretation of the data resulting from such experiments, the balances in the designs being reflected in the orthogonal ties of the components of variance. By 1935 the face of statistics had been dramatically changed by Fisher. A year before, a statistician from Poland, Neyman, Rad investigated a different type of experimentation, sampling from a population , and had established optimum ways of stratifying. He came to England and continued to produce important work, both on en his own and in co-operation with E. Pearson, Karl's son. It had never been felt adequate merely to provide a point estimate $f(x)$ of a parameter θ, at least one requiring a measure of variability, such as Fisher had given for the m.l. estimate, although this was only approximate. The new concept of a confidence interval appeared to provide a better resolution. Confining ourselves to the real line for ease of exposition, although the idea is more general, $\{t_1(x),\ t_2(x)\}$ is a confidence interval at level α for θ if

$$p(t_1(x) < \theta < t_2(x)/\theta) = \alpha \quad \text{for all} \quad \theta \tag{1}$$

The probability here is derived from $p(x/\theta)$ above. Notice that in (1), θ is fixed and x is the random element to which the probability refers. The interpretation of (1) is that the random interval includes the parameter with probability α whatever the value of θ Confidence intervals are widely used though, as we will see, some statisticians judge them to be flawed. Their construction stimulated much mathematics, in particular, in the requirement that (1) hold for all θ, the existence of similar regions or sets having the same probability for all θ.

Neyman and Pearson developed a theory of tail-area significance tests. The research was stimulated by asking whether popular tests, like χ^2 and t, were the best that could be found and generally what was meant by an optimum test and how could it be constructed.

Answers were found which depended on a different attitude to inference from that adopted by most statisticians at the time, including Fisher.

The new concept was to look at inductive behavior, rather than inductive inference: to emphasize action, rather than thought. They referred to the actions of rejecting or accepting a hypothesis. Basic concepts were the two types of error, rejecting a hypothesis when it was true, and accepting it when false. In terms of reduction of errors a theory

of optimum tests was developed. Tests based on t and F were found to be optimum but this result did as much to bolster the new theory as it did to enhance the two tests were well-established.

It only needs a slight change of language to present this theory of tests as part of decision theory. A. Wald working in United States, extended the idea in his book *Statistical Decision Functions* (1950). Herein we encounter a basic theorem that lies at the heart of modern statistics and is new to the 20^{th} century. Let us look at it first in Wald's form. Take the same parametric model, $p(x/\theta)$, for data x und parameter θ, but include also decisions d. The task is to observe x and use the result to select d. This requires a decision function $\delta(x)$ mapping x into d. Wald supposed that d and θ could be related by a loss function $L(d, \theta)$ that measured the loss in choosing d when θ obtains (for example, the loss in rejecting when the hypothesis $\theta = \theta_0$ is true).

A natural quantity to consider is the expected loss using decision function $\delta(x)$, namely

$$\int L(\delta(x), \theta)p(x)\theta)dx = R(\delta, \theta) \tag{2}$$

called the risk function. This cannot be minimized since θ is unknown but can be used to eliminate δ_2 in favor of δ_1 if $R(\delta_1, \theta) \leq R(\delta_2, \theta)$ for all θ, with strict inequality for some θ. This partial ordering leads to a complete class of decision functions, namely those δ_2 for which no such δ_1 exists. Wald found the complete class.

Let $p(\theta)$ be a density for θ that is, it is positive and integrates to one. Then we may calculate an expected risk

$$\int R(\delta, \theta)p(\theta)d\theta. \tag{3}$$

Wald's result, loosely stated, is that the complete class is the class of δ's that minimize (3) as $p(\theta)$ ranges over all densities. Such are called Bayes solutions, and the result is that the Bayes class and the complete class are essentially the same. The small gap between the two gives rise to difficulties that will be considered later. The reason for the terminology is that if the definition of R in (2) is inserted into (3) and the orders of integration reversed, we need to minimize $\iiint L(\delta(x), \theta)p(\theta/x)d\theta p(x)dx$. This can be done for each x separately by selecting to minimize

$$\int L(d, \theta)p(\theta/x)d\theta \tag{4}$$

Here $p(\theta/x)$ is the density of θ, given x, obtained from $p(x/\theta)$ and $p(\theta)$ by Bayes theorem, $p(\theta/x) = p(x/\theta)\dfrac{p(\theta)}{p(x)}$ and $p(x) = \int p(x/\theta)p(\theta)d\theta$ is the marginal density of x, (4) is the expected loss, where the expectation is over θ, not x as in (2). Savage in his book *The foundations of statistics* (1954), significantly extended Wald's result. He took a very general and abstract position involving decisions d and unknown states of the world θ.

With axioms that reflect reasonable requirements of decision-making (for example, if you prefer d_1 to d_2, both $\theta = \theta_1$ and when $\theta = \theta_2$, then you prefer d_1 to d_2 when you are uncertain whether θ is θ_1 or θ_2) he demonstrated the existence of numerical measures of uncertainty obeying the rules of the probability calculus, the existence of a loss function (though be called it utility function) and the optimality of minimizing expected loss (4).

Notice that probability, loss and expectation emerge as deductions, not as part of the system of axioms. He set out with a view to providing a sound, mathematical basis for all standard, statistical practice. In fact, his work showed that much of this practice was unsound, though it was a while before this was appreciated.

The appreciation came about largely because of work by Bitnbaum in 1962 on the likelihood principle. The minimization of expected loss (4) is equivalent by Bayes to minimization of $\int L(d,\theta)p(x/\theta)p(\theta)d\theta$ in which the data only enter through the likelihood function $p(x/\theta)$. Hence the likelihood principle which says that data sets x_1, x_2 with the same likelihood should result in the same decisions or inferences. Tail-area significance tests and confidence intervals (1) violate the principle because they involve integration over x, which (4) does not. Clearly one can produce situations in which the likelihood is the same for two values of x but which differ when other values are included and used in the required integration. Birnbaum justified the principle on grounds of conditionality and sufficiency, two ideas commonly accepted by statisticians. It owed nothing to Savage's approach. Savage theorem was not entirely new. It had been given by Ramsey in 1926, developed in an inferential context by Jeffreys in 1939 and over a period of years by de Finetti, beginning in the 1930s. Its implications split statisticians in two groups that still exist today. Those that accept the theorem as relevant are termed Bayesians, because of the basic role played by Bayes theorem. They claim that probability is the only sound measure of uncertainty both for data and parameter. Thus a tail area significance test should be replaced by the calculation of the probability that the null hypothesis is true, given the data. A confidence interval (1) should be replaced by the distribution of θ, $p(\theta/x)$. Non-Bayesians object to the use of a probability for θ, $p(\theta)$.

Confidence intervals and posterior distributions provide two solutions to the central, formal problem of mathematical statistics, that of saying something about a parameter θ on the basis of data k. Fisher proposed a third method: $f(x,\theta)$ is called a pivot if its distribution and is the same for all θ. This if is $N(\theta,1)$ then $x-\theta$ is a pivot having an $N(0,1)$ distribution. With a pivot $p(f(x,\theta)<t)$ is a number which can be found for any t. For suitable f, $f(x,\theta)<t$ can be written $\theta<g(x,t)$ and so $p(\theta<g(x,t))$ is found. This is called the fiducial distribution for θ, given x. The method is now not highly regarded because of inconsistencies and the existence of many pivots, leading to ambiguity in the results. It is the only idea of Fisher's that time has proved to be seriously flawed. To go back in time in 1936 Hotelling had been studying the relationships between two sets of quantities and, using linearity and normality, considered how they could best be expressed Over the years these ideas have led to the extensive development of multivariate statistics. Most of this has been based on the normal distribution, which is easily the simplest to handle of distributions of high dimensionality. Impressive distributional and inferential results have been obtained. More recently, less formal methods, such as multidimensional scaling, which Dave had suggested for appreciating that it exist between large numbers of quantities.

Historians are wary of discussing recent events because they are too close to them for a broad appreciation. This is true of statistics, and the account of developments in the last 40 years must necessarily be a list of separate advances. One exception is the effect modern computers have had. At first, they merely enabled statisticians to handle bigger data sets, using the models that had been adequate with desk calculators. More

recently, new models that exploit the vastly increased scope of numerical work have been introduced. Gauss's linear model had a vector y of observations, normally distributed with mean μ. In addition there were covariates $x_1, x_2, \ldots, x_\mu$ of which a linear function $n = \sum \beta_i x_i$ was supposed to influence y. In fact μ and n were the same. Generalized, linear models introduce a link function g that connects these by $n = g(\mu)$ and extends normality to include any member of the exponential family. Inference for these models is performed using methods based on some form of likelihood. Many different modifications of likelihood have been suggested which yield to modern maximalization routines.

This likelihood development has proceeded despite a surprising demonstration by Stein in 1961 that, even in the simplest, normal case where m.I. and least squares coincide, it is unsound. Treating the estimation of the normal mean as a decision function δ, he showed that, in dimensions greater than two, a decision function better than m.I. in the sense of the partial ordering described above, could be found. Technically m.I., is not a member of the complete class. Progress has been made on improvements to m.I., for example by empirical Bayes methods which introduce $p(\theta)$ and then attempt to estimate it. Completely Bayesian methods have produced results. Nevertheless, likelihood methods that ignore these matters have been widely and successfully used. The mathematical difficulty here is that m.I. is Bayesian but only in a wide sense with $p(\theta)$ equal to a constant. Such a function cannot, over an unbounded space, integrate to one. This often prevents m.I. belonging to the complete class. Another way of looking at the problem is through finitely – additive measures, in place of the sigma-additive measures of standard probability calculus.

Advances in computational techniques have enabled Bayesian methods to develop swiftly. Within that paradigm it is usually clear what to do. Inference about a quantity is expressed through its probability distribution, given all the information available, and is found using the probability calculus. Decisions are made by minimizing expected utility. Whilst there are situations that yield to mathematical analysis, in many one has to be content with approximations. Recently, ingenious numerical methods, like Gibbs sampling, have provided answers. These do not give a general solution but only results for a given data set, but this is exactly what an applied statistician requires. Bayesian methods have the additional advantage that nuisance parameters are eliminated by integration, a marginal density being obtained by integration of the full density. Integration is similarly amenable to numerical methods.

Besides his work on decisions, Wald was the first to develop sequential analysis. Earlier work had dealt with the case where data collection came first, followed by analysis. Wald supposed that data came in sequence and was paralleled by analyses that included the option of stopping collection and reaching a decision. In the case of two derisions he developed a workable analysis, but the general case proved much harder. A reason for this is that in sequential cases the space of possible data sets is much wider than with a fixed size of sample, so that optimal results are much harder to find. The difficulty is reduced if the likelihood principle is accepted since, under wide conditions, the likelihood is the same whether the data size is fixed or determined sequentially. In Bayesian analyses, the probability $p(\theta/x)$ is continually updated by Bayes theorem as the data come in, that is, as x changes.

Another context in which sequential ideas abound is inference for stochastic processes.

In its simplest case, the real values x_1, x_2, ..., x_n refer to observations gathered in time, or space, indicated by the suffixes. Generalizations to vectors and continuous time are possible. Typically the observations are equally spaced in time and the origin of time is supposed fo be irrelevant – the stationary case. Interest then centers on the covariance v_s between x_t and x_{t-s} which, by stationarity, does not depend on t v_s is called the autocovariance. Alternatively analysis can take place in the frequency domain utilizing the spectral density function, the Fourier transform $f(\omega) = \sum_{-\infty}^{+\infty} v_s \exp(i\omega s)$.

A natural requirement for a stochastic process is to predict its development, in its simplest form, to predict $\mathrm{x}_{\eta+1}$ One way to do this is by a linear expression $\sum_{i=1}^{n} \beta_i x_i$.

Adapting terminology from electrical engineering this is called a filter. Decisions may also be introduced, reflecting control that may be exerted on the process so that $x_{\eta+1}$ can be as near as possible to a target value. An important innovation is the Kalman filter that generalizes least-squares to prediction and control. Prediction is not confined to processes. Many statisticians feel that parameters are often nothing more than convenient constructs and the main purpose should be to predict future data y from past data x. Within the Bayesian framework one requires $p(y/x) = \int p(y/\theta)p(\theta/x)d\theta$, assuming that y and x are independent, given θ.

The development of models for stochastic processes has been fruitful. The older, autoregressive AR and moving-average MA models combine to produce an ARMA model $\sum_{j=0}^{p} a_j x_{t-i} = \sum_{j=0}^{q} b_j e_{t-j}$, where the process x_t has an autoregressive component of order p that equates to a moving average of random errors of order q.

An alternative system is the dynamic, linear model DLM in which there is an observation equation $x_t = F_t\theta_t + v_t$ where v_t is an error usually supposed normal, F_t is a known operator and θ_t describes the state of the process. This satisfies $\theta_t = G_t\theta_{t-1} + w_t$, where w_t is error and G_t is another known operator. Here Bayesian methods provide convenient updating. We have seen how important parameters are in most inferences. Attempts have been made to avoid the deficiency that parameters impose of restricting the class of probabilities under consideration. These lead to non parametric methods and envisage the class of all, or nearly all, distributions.

While some useful ideas have been developed, the theory and applications are limited. The mathematical reason for this partial failure lies in the fact that it is difficult to make constructive statements about a wide class of distributions, With $\theta \in R$, $p(\theta < t)$, for all t, enables all uncertainties about a finite vector θ to be described, and is easy to construct. A comparable expression when θ is of infinite dimension is not available, despite the existence of much theory concerning such spaces. Theory without such constructive ability, is useless for applications.

There is another way around the restriction parameters impose, and which has been more successful. Take a parametric model $p(x/\theta)$, with a loss function $L(d, \theta)$ if decision is involved and $p(\theta)$ is working within the Bayesian framework. Then develop an inference or decision procedure and its properties. It is possible to ask how the procedure and its properties are affected by departure from any of the three elements above. One would

like the procedure to be little affected by small departures, when it is termed robust. Thus the t-test is robust, the F-test is less so. Many robust procedures have been found. Some procedures are sensitive to outliers. These are observations that are not typical. Hardly any data set is completely free of outliers. They have a large effect on a mean but not on a median. The exponential family, which is popular because of the existence of sufficient statistics, does not allow outliers because the "tails" die away exponentially, rather the polynomially, fast. In many cases, an analysis based on a t-distribution, not of the exponential family, is more robust than one dependent on normality, which is.

The field of experimental design has developed rapidly. New designs have been found, often using mathematically interesting combinatorial concepts. Other designs have been developed for specific purposes, like finding the maximum of a function, so that a production process can be run at optimal conditions.

Statistics today are very different from what they were a century ago. Statisticians still face real problems in the collection and presentation of data, but sophisticated methods of analysis are available that did not exist in the 19^{th} century.

These methods have enabled statistical inference and decision to enter almost every field of human endeavour.

REFERENCES

[1] H. Behnke e.a. *Grundzüge der Mathematik*, Band V, Götingen, 1968.
[2] H. Eves. An Intioduction to the History of Mathematics, Sixth Ed. Orland 1926.
[3] A. King. C. B. Read Pathways Probability, NY, 1963.
[4] A. N. Kolmogorov et al. Mathematics of the 20 Century, Birkhäuser 1996.

George Dimakos
Tsigoni Anastasia
University of Athens
Greece

ИСТОРИЯ НА СТАТИСТИКАТА ПРЕЗ ПОСЛЕДНИТЕ 100 ГОДИНИ

Георги Димакос, Цигони Анастасиа

Статията е опит да се проследи развитието на математическата статистика през 20 век във възможно най-сбито изложение. Авторите изхождат от теорията на Уолд за вземане на решения и от разделението на статистиците на "байесианци" и "небайесианци", оказало се особено плодотворно за развитието на статистиката както в теоретичен, така и в приложен план.