COMMON ORIGINS OF STATISTICAL INFERENCE AND CALCULUS

Ralph deLaubenfels

Department of Mathematics, Ohio State University, Columbus, Ohio, 43210

delau@math.ohio-state.edu; (614) 592-5330

ABSTRACT. It is argued that calculus and statistical inference share many origins in classical Greek mathematics and philosophy. While neither of these disciplines can be said to have been formulated by the classical Greeks, they performed the necessary, and arguably most difficult and important, first step of identifying the fundamental and significant problems that were ultimately addressed by both subjects. I preface my arguments with summaries of statistical inference and calculus, and with a description of the formative years of each of these subjects.

Key Words and Phrases: history of statistics, history of mathematics, classical origins

I. INTRODUCTION.

There are many legacies of classical Greek philosophy, mathematics, and science in modern science and mathematics. This paper focuses on a style of problem considered for the first time by the classical Greeks, that is prerequisite and fundamental to both calculus and statistical inference.

For the sake of maximizing accessibility, standardizing terminology, and easy reference, and to emphasize that the author is fully aware that mathematics and statistics are different disciplines, I have included brief summaries of statistical inference and calculus (Sections II and III, respectively). Anyone familiar with these subjects may skip these summaries with no discomfort.

Presented next are sections on the formative years of statistical inference, in the 18th century and early 19th century (Section IV), and calculus, in the late 17th century through the middle of the 19th century. These similar formative years correspond to analogous resolutions of the relationship with classical Greek mathematics: on the one hand, supplementing geometry with algebra, on the other hand, matching the rigor.

What statistical inference and calculus–or more generally, the major sub-area of mathematical research known as analysis–have in common is a style of intellectual endeavor, both in the problems addressed and the solutions arrived at. Both deal with pervasive, endemic uncertainty. The elements of a sequence are rarely the limit, and a sample is rarely a large part of the population; and yet the elements of a sequence and a sample are the objects we can really get our hands on, while the limit and the population are the objects of interest. A sort of philosophical balance is necessary to be constantly having to use results that we know are imperfect, while pursuing a perfection that we visualize but cannot obtain explicitly.

I argue in the last section that the intellectual style just referred to was introduced by the classical Greeks, in many aspects of their mathematics and philosophy; in particular, in their study of motion, irrational magnitudes, geometry, continuous models, and the approximation of the material by the immaterial, as in Plato's parable of the cave.

Regarding the relationship between statistical inference and calculus, it should be mentioned that it is the limiting process, underlying calculus, that provides the ultimate vindication of most statistical techniques; e.g., for an estimator, that it converges to the parameter of interest it is meant to estimate, as the sample size goes to infinity (it is then known as a *consistent* estimator). Without the limiting process, even the meaning of a probability, in a practical sense, is not clear. Take for example the relatively tangible relative frequency definition of probability; e.g., if one draws at random from a black box containing seven red marbles and three yellow marbles, the probability of getting a red marble is 70%. The person drawing the marble has only red or yellow as an outcome. The 70% is experienced only as the limit of averages of the relative number of red marbles, after repeated drawings (with replacement).

But the symbiosis works both ways. Many students in undergraduate mathematics classes would be surprised to learn that the Laplace transform and Gaussian elimination both arose from solving statistics problems: the Laplace transform was created (by Laplace) for the central limit theorem and Gaussian elimination (by Gauss) for the best least-squares fit to data.

"Statistics" will be short for "statistical inference," as opposed to descriptive statistics, the mere exposition of data. "Introductory" will mean topics in standard introductory classes. "Greek" will be short for "classical Greek," meaning approximately 500 B.C. to 400 A.D. "Calculus" is perhaps

more accurately labelled "analysis"; I have chosen the more familiar synonym, again for maximum accessibility.

II. SUMMARY OF INTRODUCTORY STATISTICS.

The key duality in statistics is population versus sample. A *population* is a set of things we care about. A *parameter* is a measurement of the population. A *sample* is a subset of the population, that we may explicitly study or measure. A *statistic* is a function of the sample.

For example, a population could be all human beings born in the twenty-first century. A parameter could be average birthweight of those people (including ones not yet born). A sample could be 100 randomly chosen people, with the obvious statistic of the average birthweight of those 100 people.

Statistics is using a statistic to make an intelligent guess about a parameter. A population is usually very large; as in the example above, it might have no theoretical limit, and should be thought of as infinite. A sample is usually small, for practical reasons, such as the intrusive nature of sampling and the damage that might be done by measurement; for example, if the parameter is durability of cars, the statistic might involve doing reckless things to each car in a sample. The small size of the sample relative to the population makes statistics a very daunting, arguably impossible, endeavor. Measuring how good a statistical guess is, or even defining such goodness appropriately, is an interesting challenge by itself.

The statistical guess, or measure of how good the guess is, usually has the form of a probability statement. Commonly (parametric statistics) a certain probability distribution for the population is assumed, with only a few parameters missing to give us all information about the distribution.

Note that this inference can be seen as inverse reasoning, from effect to cause, as when Sherlock Holmes estimates the height (parameter) of a suspect from the spacing of footprints (statistic); here the population is all possible sets of footprints.

Statistical guesses traditionally have one of the following forms:

1. estimation;
2. confidence interval;
3. hypothesis test.

Estimation means choosing a particular statistic, called an estimator, to estimate a parameter. Sometimes the choice of estimator seems obvious, as with estimating average of a population with average of a sample, but sometimes great cleverness or subtlety is required, as with choosing the straight line "closest" to a set of points in the plane. Minimizing the sum of the *squares*, of the differences between the actual y values and the predicted y values, turns out to be the best technique, in many ways (see deLaubenfels (2006)); minimizing this definition of "closeness" is called *least squares* regression.

As with this geometrically challenging example, the difficulty might lie with how one measures the desirability of an estimator.

Confidence intervals add a margin of error to the single number produced by the estimator; thus we now have a range of values in which we hope the parameter falls. Associated with this interval is a measure of our confidence that the parameter lies within the interval.

The two primary goals when constructing a confidence interval are making the width small and the confidence large. Intuitively, these goals are in opposition, in the sense that the confidence decreases as the width decreases; think of manna falling randomly from heaven–the larger the net you hold up, the larger the percent of the manna you catch.

A hypothesis test partitions all possible values of the parameter into two sets: a *null* hypothesis H_0 and an *alternative* hypothesis H_a, and specifies a ("decision") rule for deciding when to reject the null hypothesis. The null hypothesis is a default, something we must give convincing evidence against before rejecting. The most popular example of a null hypothesis comes from our legal system: the statement "the defendant is not guilty" (more precisely, not proven guilty).

The extent to which the null hypothesis is a default, corresponding to the decisiveness of a rejection, is measured by what is called the *significance* of our decision rule:

Rather than give here precise definitions (see (Casella and Berger, 2002)) of confidence and significance, I will restrict my presentation to a common special case that unifies the three forms of guessing above. Perturb an estimator into a function $T(\vec{x}, \theta)$ of both data $\vec{x}$ and the parameter θ that has, for any θ, a fixed and very familiar probability distribution, one for which tables have been carefully constructed and preserved (T is sometimes called a *pivot*). For a particularly popular example, if θ is the mean of a population with known population standard deviation σ, define

$$T(\vec{x}, \theta) \equiv \frac{\sqrt{n}(\overline{x} - \theta)}{\sigma},$$

where $\overline{x}$ is the sample mean and n is the sample size. The *central limit theorem*, asserting convergence of T to the standard normal distribution, Z, as n goes to infinity, explains the popularity of both T and Z.

For $0 \leq \alpha \leq 1$, T a pivot, writing P for probability and replacing $\vec{x}$ with $\vec{X}$ to indicate the random variable of all possible samples, use an appropriate table to find c_1, c_2 such that

$$P(c_1 \leq T(\vec{X}, \theta) \leq c_2) = 1 - \alpha.$$

Given data $\vec{x}$,

$$\{\theta \,|\, c_1 \leq T(\vec{x}, \theta) \leq c_2\}$$

will be a confidence interval (or at least, a confidence set) of confidence level $(1 - \alpha)$. The hypothesis test

$$H_0 : \theta = \theta_0 \;\; H_a : \theta \neq \theta_0$$

with the decision rule

$$\text{“ accept } H_0 \text{(more correctly, fail to reject } H_0\text{) if and only if } c_1 \leq T(\vec{x}, \theta_0) \leq c_2 \text{ ”}$$

will have significance level α.

The foregoing has all been *classical* or *frequentist* statistics. Popular misconceptions about the information conveyed by confidence intervals and hypothesis tests should be mentioned.

Given a confidence interval, call it $[a, b]$, say of confidence level 99%, for the parameter θ, it is not correct to say “there is a 99% chance that θ is between a and b,” because θ is a fixed number, not a random variable; it either is or isn't in the interval $[a, b]$. Similarly, if we reject the null hypothesis H_0 in the hypothesis test

$$H_0 : \theta = \theta_0 \;\; H_a : \theta \neq \theta_0$$

with significance level 1%, it is not correct to say “the probability that θ equals θ_0 is at most 1%.” The information conveyed by this rejection is more indirect: it is saying that, if we assumed θ equalled θ_0, the probability of getting the data we did, or data more extreme, is at most 1%. For the null hypothesis “the defendant is not guilty of murder,” rejection is saying “the probability of getting the evidence presented by the prosecution (e.g., blood on hands, overheard stated intention to commit murder), if the defendant were innocent, is at most 1%.”

The intuition here is that we don't want to believe that unlikely things have happened. A more conspicuous focus on this desire is in perhaps the most popular choice of estimator, the *maximum likelihood estimator* (MLE). This estimator chooses the value of the parameter θ that maximizes the probability of the data observed; thus we are making what happened as likely as possible, choosing causes that maximize the probability of the effects.

It should be mentioned here that the preference just described for likely things happening is how the scientific method of fitting a model to data is done. A model is chosen that makes what we observe most likely. See (Weinert, 2010) for an interesting discussion of the role of statistics (from a Bayesian point of view, although it could equivalently be done from a frequentist point of view) in choosing scientific models.

Although it is not normally mentioned in introductory statistics courses, for both historical and intellectual completeness, I should mention the alternative approach to statistics known as *Bayesian* statistics. Here the parameter θ (more accurately, our uncertainty about θ) is treated as a random variable. It is assigned a *prior* probability distribution; after data $\vec{x}$ is collected, the *posterior*

distribution, $(\theta|\vec{x})$, θ given $\vec{x}$, may be calculated, reflecting a change in our beliefs about θ because of the data.

For example, suppose a coin is flipped ten times, to try to get a clue about the parameter θ defined to be the probability of getting heads on each flip. If we got all heads in those ten flips, the frequentist MLE for θ is 1; that is, a frequentist conclusion might be that the coin is two-headed. A Bayesian might glance at the coin and notice there is a tails on one side and a heads on the other, and, lacking any other information about θ, assign a uniform prior distribution. Then a calculation, that I would rather not go into here, shows that the posterior distribution is given by

$$P(a \leq \theta \leq b) = b^{11} - a^{11} \quad (0 \leq a \leq b \leq 1).$$

Notice that values of θ near one are now strongly favored, as expected from the string of heads.

The Bayesian approach allows one to make bonafide probability statements about parameters with the Bayesian analogues of confidence intervals and hypothesis tests. But the frequentist approach has the satisfying modernist flavor of purely data-driven results (except for the choice of probabilty model for the population).

III. SUMMARY OF INTRODUCTORY CALCULUS.

Here is the usual order of appearance of the panoply of "freshman calculus."

1. Derivative: the instantaneous rate of change of a function.

2. Integration: initially, area between two curves, but the rigorous definition quickly implies numerous other quantities of interest that may be obtained by integration, such as volume, mass, work, etc.

3. (Infinite) sequences and sums.

4. Multivariable versions of the above.

Calculus is an activity performed on functions, thus the traditional "precalculus" class is primarily about functions. A (real) function f is a rule that assigns, to each point x in a subset of the real line called the domain a unique real number $f(x)$. Equivalently, a function (or its graph Γ) is a subset of the Cartesian plane $\{(x, y) \,|\, x \text{ and } y \text{ are real }\}$ such that, for every x in the domain, there is precisely one y such that (x, y) is in Γ.

The most important idea from precalculus is *average rate of change* of a function f. Given two values $x_1 < x_2$ in the domain of f, the average rate of change of f over the interval $[x_1, x_2]$ is

$$\frac{(f(x_1) - f(x_2))}{(x_2 - x_1)} \equiv \frac{\Delta y}{\Delta x},$$

the *slope* of the line between the two points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ on the graph of f. "Δ" stands for "difference," Δy and Δx are affectionately known as "rise" and "run," respectively, and slope is English usage, as in the slope of a hillside with height $y = f(x)$, a distance x from a fixed reference point.

What is difficult to communicate in the traditional rush to applications is that the *limit* is the unifying concept of calculus. To say that a sequence $\{x_n\}_{n=1}^{\infty} \equiv \{x_1, x_2, x_3, ...\}$ converges to a limit ℓ, denoted $x_n \to \ell$, is to say that we are guaranteed arbitrarily good approximations of ℓ by a member x_n of the sequence; more reassuringly, to guarantee a specified accuracy of approximation, we need only go sufficiently far in the sequence; for any n sufficiently large, we will be certain that x_n approximates ℓ to within that specified accuracy.

Limits, hence calculus, are about approximations; not only specifying how to make them, but, more importantly, controlling their accuracy.

Limits are implicit in many statements and constructions: when we say

$$\pi = 3.1415... \quad \text{(beware the dots)},$$

we mean that π is the limit of a sequence of rational numbers that begins

$$3, 3.1, 3.14, 3.141, 3.1415, ...$$

Limits allow us to pass to what we need (such as the area of a disc) from what can be explicitly calculated or (as the Greeks said) constructed (such as the areas of polygons inscribed in a disc). For derivatives, we need the limits of slopes of straight lines: using the Leibniz notation (see Section V) for the derivative $\frac{dy}{dx}$ of the function $y = f(x)$, we have

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x};$$

the instantaneous rate of change is the limit of the average rates of change, over intervals whose length shrinks to zero. For integration, we need the limits of (finite) sums of areas of rectangles (see the end of Section V for the heights and bases of the rectangles). For infinite sums, we need the limits of finite sums; for the special case of sums of x^k, powers of a variable x of increasing degree k, we obtain approximation by polynomials of a wide class of functions. What we need often cannot even be *defined* without a limiting process; see the attempts described in Section V to define the derivative.

The limit concept is fundamental in any practical use of data. Unless one is merely counting, there is no such thing as an exact measurement. The Heisenburg uncertainty principle implies that this is true regardless of the precision of your measuring instrument. The statement "I am five feet, eight inches tall" is not an exact statement, it only places me within a range of values, probably between 67.5 and 68.5 inches. It seems traditional to pretend two decimal places represents sufficient precision, as in "pi equals 3.14," but, even ignoring the issue of units involved, surely it is easy to imagine scenarios where more precision is desired–say, retinal laser surgery. Errors of approximation can easily blow up after they appear in calculations; for example, an error of one-tenth of an inch in the width of a rectangle of length 600 feet becomes an error of five feet squared in the area. It is highly desirable to have a method of getting arbitrarily good approximations; this is what a limit provides.

The limit of a function f at a point a, $\lim_{x\to a} f(x)$, may be defined with sequences:

$$\lim_{x\to a} f(x) = \ell \quad \text{if and only if} \quad f(x_n) \to \ell \quad \text{whenever } x_n \to a.$$

More directly, here is the "$\epsilon - \delta$" definition of $\lim_{x\to a} f(x) = \ell$:

$$\text{for all } \epsilon > 0 \text{ there exists } \delta > 0 \text{ such that } |f(x) - \ell| < \epsilon \text{ when } |x - a| < \delta.$$

Continuity, another very fundamental idea from calculus, is defined in terms of limits: the function f is continuous at a if $\lim_{x\to a} f(x) = f(a)$. Continuity intuitively means no jumps or similar surprises. An immediately applicable example of this intuition is the intermediate value theorem: if f is continuous on $[a, b]$ and ℓ is between $f(a)$ and $f(b)$, then $f(c) = \ell$, for some c in (a, b).

IV. THE SOLIDIFICATION OF INTRODUCTORY STATISTICS, EARLY EIGHTEENTH CENTURY TO EARLY NINETEENTH CENTURY.

One could argue that statistical inference has been done unconsciously, by most of our species for most of its history. When trying to understand illness, the creation of the world, or any other mysterious but important phenomenon, we instinctively choose the model that makes the observed data most likely, analogously to an MLE; when deciding guilt or innocence, we are performing a hypothesis test, by estimating if the evidence presented is too unlikely, when assuming innocence; a prediction of crop yield, or any other numerical estimate, we know is worth more if it has a margin of error placed around it. In a sense, we are all performing statistics without knowing it; but we often do it very badly, and might not have any way of evaluating how good a job we're doing.

But these reflexes are to statistical inference what verbal notions such as "few" and "many" are to counting and actual arithmetic. What is called statistical inference begins with quantification of these intuitive ideas, via explicit definitions and formalizations of technique, so that we can understand what we're doing and evaluate it.

It may come as a surprise that statistics as an intellectual discipline existed at all during the period of this section. It is easy to get the impression, especially when beginning graduate study in the subject, that statistics is about one hundred years old. Particularly misleading is the common

use of "classical" as being synonymous with "frequentist," especially as formulated by Fisher and his contemporaries, in the basic partitioning of statistics into two camps of varying partisanship. The excellent textbook Casella and Berger (2002) introduces Bayesian statistics with the following sentence, at the beginning of Section 7.2.3, top of page 324: "The Bayesian approach to statistics is fundamentally different from the classical approach [frequentist, Fisher et al.] that we have been taking." Consider also, on page 2 of Lehmann and Casella (1999) "frequentist (the classical approach of averaging over repeated experiments),...". The following more detailed exposition may be found in Stuart and Ord (1991, p. 1197, lines 15–18): "...the *frequentist* paradigm, sometimes known as the *classical* approach, which has been the dominant school of statistical thought for most of this [the twentieth] century. However, the *Bayesian* viewpoint has grown in popularity since the 1950s, ..." (italics theirs). In that same book, under "classical inference" in the index, one finds "*see* Frequentist inference."

Thus we seem to see frequentist statistics, beginning in the late nineteenth or early twentieth century, as the Parthenon of statistics, with Bayesian statistics a new avant garde approach arising in the lifetime of the currently middle-aged. Some discomfort with this modernist conception of statistics may occur if one hears of Thomas Bayes, 1701–1761, but this coincidence of names could be explained as another esoteric joke, analogous to the statistician formerly known as Student, and his distributions. A superficial "Google" search of Thomas Bayes yields "British mathematician and Presbyterian minister," not really the model of a modern statistician.

In apparent support of this view of statistics' chronology, Stigler (1999), in Chapter 8, proposes to "advance and defend the claim that mathematical statistics began in 1933."

I will briefly demonstrate that most of the objects that now appear in an introductory statistics course were in standard use in the eighteenth century.

Let's begin with the central limit theorem, really a probability result, but so fundamental to large-sample statistical inference that it may be considered on the interface of probability and statistics. DeMoivre proved it for binomial distributions in 1733 (see Hald (1998, Sec. 2.3, p. 17)). Laplace proved the general case in 1810 (Hald (1998, Sec. 17.2, p. 307)). See also Kendall and Plackett (1977, pp. 101–104), for a description of Daniel Bernoulli's work on the central limit theorem, 1770–1771.

In keeping with the theme of this paper, it is of interest here to note that the normal distribution, which eventually grew to such prominence in statistical modelling and inference, first appeared as a limit of binomial distributions (the goal being to approximate binomial probabilities, for a large number of trials) (see Hald (2003, Ch. 24)).

Least-squares regression (not by that name) was introduced by Legendre in 1805 (see Stigler (1986, p. 55) and Hald (1998, Sec. 6.4, p. 118)) and Gauss in 1809 (see Stigler (1986, p. 140) and Hald (1998, Sec. 18.1, p. 351)). Gauss asserted, probably correctly, that he had gotten his results as early as 1795. Gauss's formulation included probability distributions on the parameters and the error. Throughout the second half of the eighteenth century, the same idea of choosing parameters that minimize error was presented, with either the sum or the maximum of the absolute deviations replacing the sum of the squares of the deviations, in other words, L^1 or L^∞ replacing L^2. See Hald (1998, Chap. 6), Stigler (1986, Chap. 1), and deLaubenfels (2006), and the references therein.

Lambert in 1760 solves a maximum likelihood equation to get the MLE (maximum likelihood estimator) for a population mean μ, in a probability distribution $p(x) = f(x - \mu)$, with the function f specified (see Hald (1998, pp. 81–82)). Daniel Bernoulli also constructs an MLE in a 1778 paper, although there is reason to believe he wrote up the results as early as 1769 (see Pearson and Kendall (1970, pp. 155–172) and Hald (1998, pp. 84–85); the former reference includes a response by Euler criticizing the MLE as being arbitrary). Lambert gives no motivation for using the MLE, while Daniel Bernoulli calls his motivation "metaphysical rather than mathematical." Laplace's "Principle" of inverse probability,

$$p(\theta|\vec{x}) \propto p(\vec{x}|\theta),$$

for a probability density or mass function p, data $\vec{x}$, parameter θ, which follows by Bayes' theorem from his uniform formulation of parameter spaces (see Stigler (1986, pp. 102–105)) was stated in 1774 for θ discrete and in 1786 for θ continuous, although the continuous formulation was used by

him in both 1774 and 1781 papers (see Hald (1998, pp. 160–161)). Besides launching a general theory of Bayesian statistical inference, note that Laplace's principle makes maximizing the posterior probability equivalent to choosing the MLE. Gauss's presentation of least-squares regression, mentioned above, included the observation that, for a normal distribution on the error, the least-squares estimators are MLEs.

Credible sets, the Bayesian analogues of confidence intervals, are virtually automatic once one has explicit representations of posterior distributions. But authentic frequentist confidence intervals also appear in the eighteenth century. Lagrange in 1776 uses the asymptotic normality of $(p - \hat{p})$, for a binomial parameter p, sample proportion $\hat{p}$, with mean zero and variance $\frac{\hat{p}(1-\hat{p})}{n}$, to assert, for any positive t, that "the probability that the value of p is enclosed between the limits"

$$\hat{p} \pm t\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

equals $(\Phi(t) - \Phi(-t))$, where Φ is the cumulative distribution function for the standard normal variable Z (see Hald (1998, p. 23)). The interval just given is a (frequentist) confidence interval for p, with confidence level $(\Phi(t) - \Phi(-t))$. Laplace also gets this confidence interval from the central limit theorem in Laplace (1812), and comments there that the same result is obtained by the Bayesian approach of treating p as a random variable with uniform prior distribution, and using the asymptotic normality of the posterior distribution, as he did in 1774 (see Hald (1998, pp. 24–25)).

Arbuthnot in 1710 tested the null hypothesis of male and female births in London being equally likely; he called this hypothesis "Chance" (see Kendall and Plackett (1977, pp. 30–34)). The data consisted of there being more male births than female births each year for 82 years in a row. The probability of this data, given the null hypothesis, is $(.5)^{82}$, sufficiently small for him to conclude that "it is Art, not Chance, that governs." That was a frequentist hypothesis test (Stigler (1986, pp. 225-226), Hald (1998, p. 65)). Other statisticians, such as Nicholas Bernoulli, suggested that the hypothesized "chance" (that is, the binomial parameter p defined to be the probability that a birth in London is male) could be something other than .5 ((Stigler, 1986, p. 226), (Hald, 1998, pp. 65–66)). Laplace, in 1781, addressed a similar problem in Paris, but with Bayesian techniques: given $251,527$ male births and $241,945$ female births from 1745–1770, he calculated a posterior probability

$$P(q \leq .5 \,|\, \text{ data }) = 1.1521 \times 10^{-42} \;\; (q \equiv \text{ probability that a birth in Paris is male}),$$

concluding that male births were more likely than female births. Even more extreme numbers gave a similar conclusion in London. He also tested the null hypothesis of male births in London being less likely than male births in Paris, getting a posterior probability of

$$P(q > p \,|\, \text{ data }) = \frac{1}{410,458},$$

and concluding that London was more prone to male births than Paris. Interestingly, when a similar comparison to Paris and Naples gave a posterior probability of $\frac{1}{100}$, he said it "is not sufficiently extreme for an irrevocable pronouncement" (Stigler, 1986, pp. 134–135)). It is stimulating to one's historical imagination to contrast this with the modern habit, at least in many circles, of a significance level of .01 always representing sufficiently strong data for rejecting a null hypothesis.

Speaking of historical imagination, Arbuthnot's 1710 paper was titled "An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes." Here we have two subjects a prudent modern speaker or instructor might be hesitant to address: anything as openly religious as "Divine Providence" and formulating birth gender as a binomial problem–which gender do I choose as "success" in the traditional binomial counting of "successes"?

V. THE SOLIDIFICATION OF INTRODUCTORY CALCULUS, LATE SEVENTEENTH CENTURY TO MIDDLE NINETEENTH CENTURY.

Although special cases of calculus techniques (integral calculations by Archimedes and Wallis, optimization via the first derivative by Fermat, for example; see (Edwards, 1979), (Struik, 1987), (Resnikoff and Wells, 1984), or (Stillwell, 1989)) already existed, the invention of calculus is credited, independently, to Newton and Leibniz in the late seventeenth century primarily because of their

identification and general application of the fundamental theorem of calculus, recognizing the inverse nature of integration and differentiation.

Newton defined only time derivatives, or fluxions, because of his focus on motion, particularly arising from gravitation; he wrote $\dot{x}$ for the derivative (with respect to time) of x. Leibniz introduced the more general (instantaneous) rate of change $\frac{dy}{dx}$ of one variable, y, with respect to another, x. Newton considered slopes of tangent lines $\frac{dy}{dx}$; but he wrote that slope as $\frac{\dot{y}}{\dot{x}}$.

Many of the computational techniques currently taught in introductory calculus were developed by Newton and Leibniz and their contemporaries; see (Edwards, 1979) and (Stillwell, 1989, Chapters 8 and 9).

The foundations of calculus–in particular, an unambiguous definition of the derivative–as taught in a week or two in modern introductory calculus classes, took about a century and a half to develop.

Newton treated the concept of motion as a first principle; but, as evidenced by the paradoxes involving motion that frustrated Greek mathematicians (see the last section of this paper), any explanation based on motion needed a good deal of fundamental elaboration. Leibniz based his explanations on infinitesimals, quantities infinitely small but nonzero; this produces good intuition, as we try to communicate in introductory calculus classes, so that the derivative $\frac{dy}{dx}$ looks like slope $\frac{\Delta y}{\Delta x}$, and the integral $\int f(x)\,dx$ looks like a sum $\sum f(x)\Delta x$ (Leibniz introduced the integral sign $\int$, as a script S suggesting "sum"). There is reason to believe that Leibniz saw infinitesimals as an intuitive guide rather than a well-defined concept (see Edwards, pp. 264–265), but many of his enthusiasts took their existence seriously.

These ambiguities–are we dividing by zero or not?–in the definition of derivative justifiably led to many criticisms of calculus. The most famous are due to Bishop Berkeley, especially his witty and extensive 1734 essay "The Analyst, or A Discourse Addressed to an Infidel Mathematician." Here are two quotes that I think give the flavor of his essay. Referring to the "evanescent increments" presented to explain fluxions or infinitesimals, he wrote "May we not call them the ghosts of departed quantities?" Elsewhere he wrote "He who can digest a second or third fluxion, a second or third difference, need not, methinks, be squeamish about any point in divinity." (Edwards, 1979, pp. 293–5) The philosopher Hobbes wrote of some calculus techniques "to understand this for sense, it is not required that a man should be a geometrician [often synonymous, at the time, for mathematician] or logician, but that he should be mad." (Stillwell, 1989, p. 103).

It is of interest to the theme of this paper that one of the rebuttals to "The Analyst" was the second of three papers by Thomas Bayes, of Bayesian statistics fame, "An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians against the objections of the Author of the Analyst." According to (Smith, 1980), Bayes' paper was "...not unlike Cauchy's treatment of limits."

A very interesting summary of eighteenth-century mathematicians' attempts to establish foundations for calculus appears in (Grabiner, 1981, pp. 32–36), illustrated by how they would explain why the derivative of x^2 is $2x$.

Two evolutions of Greek thought occurred simultaneously during the time period of this section. There was an increase in rigor, not matching the Greeks until Cauchy or Weierstrass, in the nineteenth century. At the same time, geometry was supplemented by algebra. The interaction of geometry and algebra may be considered central to many of the achievement of mathematics since the seventeenth century. Geometry provides intuitive pictures, while algebra provides precision. The main barrier to the further advance of Greek mathematics was their inability to reconcile numbers (which they chose to be only rational, because they had no rigorous algebraic method for defining irrationals) with geometry, where irrational magnitudes appeared routinely (see Section VI).

An example of the Greek-inspired geometrical outlook of the seventeenth century is that calculus was performed on curves rather than functions. Although Leibniz introduced the term "function," Euler in 1748 (see (Edwards, 1979, bottom of p. 269)) was the first to define functions in a way similar to the present day, and make them the recipients of calculus.

Continuity in the eighteenth century was used in a verbal sense, as in continuity of definition: able to be described by only one formula (see the first few sections of (Edwards, 1979, Chapter 11)).

In 1754 d'Alembert made an important step towards avoiding the ambiguities of infinitesimals, by emphasizing the ratios of increasingly small things, rather than the small things individually. Thus we see some notion of the derivative as a limit of ratios, except that limit had not been clearly defined (Edwards, 1979, p. 295).

Lagrange, in 1797, took a different approach to defining derivatives of all orders as coefficients of the Taylor series: if

$$f(x) = \sum_{k=0}^{\infty} a_k (x-a)^k,$$

then, for any nonnegative integer n, the nth derivative $f^{(n)}(a)$ of f, at the number a, is by (Lagrange's) *definition* $n!a_n$. (Edwards, p. 296)

This is an equivalent definition of f's derivatives, *if* f has a Taylor series centered at a; Cauchy pointed out that not all functions have Taylor series.

The idea of limit, needed for a rigorous definition of a continuous function, derivative, and integral, did not appear in its present form until the early nineteenth century, due independently to Cauchy (1821) and Bolzano (1817), at which time they also defined continuity rigorously. Throughout the 1820s Cauchy also used the limit to define derivatives and integrals. The literal "$\epsilon - \delta$" definition of limit mentioned at the end of Section III, finally completing the replacement of geometry with algebra, did not appear until Weierstrass in the middle of the 19th century.

One final twist was needed to get the "Riemann integral" taught in introductory calculus. Here is a quick description of the limiting process referred to in Section III to get $\int_a^b f(x)\,dx$, the integral of a function f over an interval $[a,b]$. For an arbitrary nonnegative integer n, partition $[a,b]$ into subintervals $[x_0,x_1],[x_1,x_2],...,[x_{n-1},x_n]$, where $a = x_0 < x_1 < x_2 ... < x_n = b$. The collection of subintervals, or, equivalently, the partition points $x_0, x_1, ..., x_n$ is called a *partition* P, of $[a,b]$. The *norm* of P, $\|P\|$, is the maximum width $|x_k - x_{k-1}|, 1 \le k \le n$, of the subintervals. Then the Cauchy integral is

$$\lim_{\|P\|\to 0} \sum_{k=1}^{n} f(x_{k-1})(x_k - x_{k-1});$$

notice that, for f nonnegative, $\sum_{k=1}^{n} f(x_{k-1})(x_k - x_{k-1})$ is the area of a collection of rectangles approximating the graph of f, with the height of each rectangle given by the value of f at the left endpoint. The generalization due to Riemann, in his 1854 Habilitation, allows one to use any point $\overline{x}_k$ in the base $[x_{k-1}, x_k]$ to get the height of the rectangle:

$$\lim_{\|P\|\to 0} \sum_{k=1}^{n} f(\overline{x}_k)(x_k - x_{k-1});$$

(Riemann) integrability means the limit exists and is the same, regardless of the partitions and points $\overline{x}$ chosen (see (Edwards, 1979, page 323)).

VI. COMMON ORIGINS IN CLASSICAL GREEK MATHEMATICS AND PHILOSOPHY.

This section presents the primary thesis of this paper: that many origins of both statistical inference and calculus may be found in certain aspects of classical Greek mathematics and philosophy. This is not meant to imply these ideas *are* statistics or calculus, but that statistical inference and calculus share them as common ancestors, so that calculus and statistics may be considered supportive cousins.

The fundamental and necessary first step in the development of both statistics and calculus was the willingness to deal with unremitting uncertainty. This uncertainty is inherent in both disciplines.

Because the origins of calculus in Greek mathematics are well documented (see (Edwards, 1979), (Anglin and Lambek, 1995), (Struik, 1987), (Resnikoff and Wells, 1984), (Stillwell, 1989)), the emphasis of this section will be on statistics. Also, the emphasis will be negative–on problems unsolved rather than successful techniques such as the method of exhaustion and "The Method" of Archimedes (see (Struik, 1987, pp. 46–7)).

Here is an outline of my argument.

1. Historically, Greek mathematicians were the first to deal with the style of uncertainty that statistics and calculus deal with.

2. Closely correlated with this uncertainty was their dealing with continuous models.

3. They were the first to not only make approximations but methodically estimate the error in these approximations, in particular controlling error with inequalities involving the quantity being estimated (see Resnikoff and Wells (1984, Chapter 4)).

4. Platonism, roughly, the philosophy that there is a perfect immaterial world, which our observed material world can only approximate or be shadows or reflections of, nicely describes the relationship of statistics to probability, a sample to a population, or a statistic to a parameter, and, more generally, the incorrigible and always potentially infinite error and uncertainty involved in statistical inference.

In the limiting process of calculus, the elements of the converging sequence can similarly remain qualitatively entirely different from the limit of the sequence; e.g., a sequence of rational numbers converging to an irrational number, or a sequence of polygons converging to a disc.

5. Although the Greeks did not develop probability, they understood how essential it was to knowledge or truth, as illustrated by Aristotle's question of whether the statement "there will be a sea battle tomorrow" is true or false ((Anglin and Lambek, 1995, p. 68)).

Analogously, Aristotle spoke of the need to understand motion, continuity, and the infinite, all precursors to calculus (see (Edwards, 1979, p. 86)).

For both calculus and statistics, we thus have the Greeks taking the essential first step of identifying the right problems.

Now I will briefly elaborate on some of the points above.

Unlike probability, where both discrete and continuous models appear routinely, statistics seems fundamentally continuous, both literally in the sense that a continuous parameter space is usually the most natural, and subjectively, in the elusiveness of the quantities sought. Calculus, I think it is clear, is fundamentally about continuous processes.

At least as early as the statement "all is water" ascribed to Thales (624–547 B.C.), historically the first real mathematician in the sense of addressing "why" in addition to "how," many Greeks had a continuous model for the world (not all; Pythagoras (570–500 B.C.) stated "all is number," meaning natural or at most rational numbers, a discrete model). Besides their belief in truth as an end in itself, the Greeks' great interest in geometry led to considering continuous phenomena, and thence to encountering systemic uncertainty. Where, exactly, are you on a line?

Irreconcilable confusion was exacerbated by their insistence on allowing as "numbers" only rational numbers (This perspective persisted even with many mathematicians up through the end of the nineteenth century; consider this quote from Kronecker, in 1886: "The integer numbers were made by God, everything else is the work of man." See (Struik, 1987, p. 162).

The earliest Pythagoreans were forced into intimacy with this discord between their continuous geometry and their discrete arithmetic merely by drawing the simplest right triangle, with both legs of length one, hence a hypotenuse of length $\sqrt{2}$, not rational (see (Struik, 1987, p. 42), (Stillwell, 1989, Sec. 1.5), (Anglin and Lambek, 1995, Chap. 10)).

Probably the Greeks' most famous encounter with uncertainty is Zeno's paradoxes (see (Anglin and Lambek, 1995, p. 55)). Here is one perhaps less well known than the tortoise that could never be overtaken: an arrow in flight, at a fixed instant, travels no distance, therefore has no motion. As with the tortoise, Zeno's surely ironic conclusion is that motion is an illusion. As with other paradoxes of the time, the seeming contradiction comes from the conflict between continuous (geometry; specifically, length) and discrete (numbers and time) models.

This inconsistency in modelling time as discrete and space as continuous arguably still exists. A space or length analogue of the idea of an "instant" of time does not seem to have as popular a usage; "atom" was originally meant to be such an analogue, but has lost its indivisibility with the model of electrons circling a nucleus.

The style that the problems above share with statistics and calculus is the pervasiveness of the uncertainty. Perhaps especially motivating, with their sense of incompleteness, are those problems that weren't solved at the time, such as Zeno's paradoxes. These problems set the stage for both statistics and calculus. They led to the practical philosophical orientation that approximation is necessary and error and imperfection are inevitable; hence the desire for models that are not too sensitive to imperfect measurements or assumptions, denoted "robust" in statistics and "well posed" in mathematics. Statistics and calculus represent different, but closely related, responses to that outlook: calculus with limits, statistics with inference from samples.

Speaking more literally of philosophy, (Dale, 1999) suggests, in footnote 3 of Chapter 1, that Plato's parable of the cave could be considered the first example of an inverse problem. More generally, according to Platonism the material world that we see consists merely of shadows of an unseen immaterial world of "forms." See (Anglin and Lambek, 1995, pp. 67–68 and elsewhere), for a quick description of Platonism. For example, the idea, or "form" of a circle is (only) *approximated* by a drawing of a circle, an earring, or an Olympics symbol. Reasoning from these material approximations to a conception of the immaterial idea is inverse reasoning or inference, similar to the inverse nature of statistics, reasoning, as Laplace described it, from effect to cause.

More directly to the point, any continuous model must be considered in the realm of the immaterial. Any information achieved materially (that is, data) is finite, and the set of all possible such information is therefore countable. Continuous objects are uncountable. For example, the support of a continuous random variable is a union of intervals of real numbers, an uncountable set. The data that we actually calculate or work with is decimal approximations of real numbers, a countable set.

The acknowledgment of the existence of approximation and error is equivalent to accepting the presence of uncertainty. A sign of this equivalence is the development of a "theory of errors" ((Hald, 1998, Chap. 5)) in the foundations of statistics that developed in the eighteenth century. A key step in the development of statistical inference is Thomas Simpson's decomposition, in 1755, of a measurement into

$$\text{measurement} \;=\; (\text{true value}) + (\text{error})$$

(see (Eisenhart, 1961), (Plackett, 1972), and (Stigler, 1986, p. 90)).

Note that the extension from approximations of a desired quantity to error bounds on that quantity, alluded to in point 3, is very analogous to the important step in statistics of putting a margin of error around an estimator of a parameter, to form either confidence or credible (depending on one's statistical orientation) intervals.

(Grabiner, 1981) notes that Cauchy's use of inequalities was essential to the rigorization of calculus; the "$\epsilon - \delta$" definition of limit at the end of Section III consists entirely of inequalities.

Early origins of probability may be found in games of chance, including interesting precursors of dice; see (David, 1998). I think Aristotle's comment in point 5 shows that the Greeks understood more than others of the time how fundamental probability is. See (Kendall and Plackett, 1977, pp. 1–14), for a discussion of the classical Greeks and probability.

The intimate relationship between calculus and statistics, supported by this common ancestry, is important, because, as argued in the Introducton, the ultimate justification for statistical constructions and techniques, frequentist or Bayesian, really lies in their asymptotic behavior, that is, their convergence, the calculus response to uncertainty, to the desired quantity.

VII. CONCLUDING REMARKS.

Besides wanting to encourage interaction between statisticians and analysts, I would like to speak briefly about the usual practical rationale for studying history, not repeating mistakes of the past.

I argue that origins of the sort I've described, of a people (Greek mathematicians and philosophers) identifying problems whose solutions were often left to their intellectual descendants (modern statisticians and mathematicians) is the sort of historical connection that is particularly valuable. Here is an example. Greek mathematicians eventually got the area of a circle (see Archimedes' proof, (Anglin and Lambek, 1995, pp. 98–100)); much earlier phases of the analysis began with upper and

lower approximations, by, respectively, circumscribing and inscribing regular polygons. In an early phase of this reasoning, the sophist Antiphon (about 425 B.C.) asserted that a circle actually *is* the inscribed polygon, at least for sufficiently many sides. This was closely related to Zeno's paradoxes, sharing the same confusion about continuous versus discrete; in this case, Antiphon was visualizing space as being discrete, so that the true nature of the circle—smooth, instantaneous changes of direction–was degraded into something subjectively quite different, the finite number of abrupt changes of direction of a polygon (see (Anglin and Lambek, 1995, pp. 60–61)). This equating of an approximation with the real thing is analogous to treating 99% probability (or the more indirect confidence or significance) as being equivalent to certainty; analogously in using mathematics, I refer again to the common acceptance of two decimal places as precision.

REFERENCES

Anglin, W. S., and Lambek, J. (1995), *The Heritage of Thales,* Springer, New York.

Casella, G., and Berger, R. (2002), *Statistical Inference* (2nd ed.), Duxbury.

Dale, A. I. (1999), *A History of Inverse Probability* (2nd ed.), Springer, New York.

David, F. N.(1998), *Games, God and Gambling*, Dover, New York.

R. deLaubenfels (2006), The Victory of Least Squares and Orthogonality in Statistics, The American Statistician 60, 315–321.

Edwards, Jr., C. H. (1979), *The Historical Development of the Calculus,* Springer-Verlag, New York.

Eisenhart, C. (1961), Boscovich and the combination of observations, in *Roger Joseph Boscovich,* ed. L. L. Whyte, Allen and Unwin, London, 200–212; see also (Kendall and Plackett 1977, pp. 88–100).

Grabiner, J.V. (1981), *Origins of Cauchy's Rigorous Calculus*, The MIT Press, Cambridge, Massachusetts, and London, England.

Hald, A. (1998), *History of Mathematical Statistics, From 1750–1930,* Wiley Series in Probability and Statistics, New York, NY.

Hald, A. (2003), *History of Probability and Statistics and Their Applications before 1750,* Wiley.

Hald, A. (2007), *A History of Parametric Statistical Inference From Bernoulli to Fisher, 1713–1935,* Springer.

Kendall, M. G., and Plackett, R. L., eds. (1977), *Studies in the History of Statistics and Probability,*(Vol. 2) London: Griffin.

Laplace, P. S. (1812), *Theorie Analytique des Probabilites*, Courcier. Paris.

Lehmann, E. L. and Casella, G. (1999), *Theory of Point Estimation,* (2nd ed.), Springer, New York.

Pearson, E. S., and Kendall, M. G., eds. (1970), *Studies in the History of Statistics and Probability,* (Vol 1) London: Charles Griffin.

Plackett, R. L. (1972), The discovery of the method of least squares, *Biometrika* 59, 239–251; see also (Kendall and Plackett 1977, pp. 279–291).

Resnikoff, H. L., and Wells, R. O., Jr. (1984), *Mathematics in Civilization,* (2nd ed.), Dover, New York.

Smith, G.C. (1980), Thomas Bayes and fluxions, Hist. Math. 7 (1980), 379–388.

Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty before 1900,* the Belknap Press of Harvard University Press, Cambridge, MA.

Stigler, S. M. (1999), *Statistics on the Table: The History of Statistical Concepts and Methods,* Harvard University Press, Cambridge, MA.

Stillwell, J. (1989), *Mathematics and its History,* Springer-Verlag, New York.

Struik, D. J. (1987), *A Concise History of Mathematics* (4th ed.), Dover, Ontario.

Stuart, A., and Ord, J. K. (1991), *Kendall's Advanced Theory of Statistics* Vol. 2, Oxford University Press, New York.

Weinert, F. (2010), The role of probability arguments in the history of science, Studies in History and Philosophy of Science 41, 95–104.