EARLY HISTORY OF THE INNER PRODUCT IN STATISTICS

ABSTRACT. This thesis outlines the history of least squares, analysis of variance, the general linear model, conditional probability, covariance, and sufficiency, up to the point where measure theory or social science dominates. The role of the inner product in developing these ideas is emphasized. A detailed expository appendix shows how fundamental results in these subjects can be unified and clarified as deriving from the inner product.

INTRODUCTION AND RATIONALIZATION

In the grand tradition of academics desperately justifying their right to exist, it is appropriate to begin by trying to convince the reader that I am doing something worthwhile. A quick defense of the study of history is the cliché, "He who does not know the past is doomed to repeat it," a chilling threat for those at all familiar with history. What I wish to defend is studying the history of ideas.

From the point of view of research, one gets clues about what techniques may work in the future by seeing what worked in the past. Research is like digging for gold. You choose a spot to dig in without knowing there is anything there; any clues about which places are more likely to have anything worthwhile buried are welcome. It is also true that techniques that *failed* in the past might turn out to be useful in the future, either because the problems being attacked are different, or because the previously unsuccesful technique can now be supplemented by something that will make it more useful.

A rigidity of thought can be a consequence of growing up seeing only slick, synthesized, predigested versions of a difficult idea. Sometimes it is good for the imagination to see things done the hard way. More generally, seeing things done a different way than has become customary is always desirable, to break the ubiquitous habit of doing things because other people do them, the worst possible reason.

One also can learn, from studying the histories of ideas, how worthwhile intellectual interaction can be. Here I must hasten to avoid the impression of a banal social cliché. By "interaction" I do not mean the intellectually vacuous cheerleading and blind conformity known as "teamwork" in pseudo-intellectual social circles. I mean regular, clear communication and discussion of ideas, with blunt criticisms ("negative feedback"—bad for teambuilding) always essential. Studying history drives home how much even the most intelligent and creative researchers "stand on the shoulders of giants," to paraphrase Einstein. One can also see how much time is wasted by people being unaware of previous work, hence duplicating it.

Studying the history of an idea has particular value for pedagogy. It should help any instructor appreciate the struggles of his/her students when he/she sees how confused the most intelligent people have been trying to develop and understand an idea. It might also give us clues about our students' most likely misconceptions and confusions, to see the misconceptions and confusions of scientists first encountering an idea.

I'll quickly mention here one lesson that I see from even a cursory examination of the history of ideas. I have to ask why such intelligent people had so much trouble understanding ideas that can now be taught in a few weeks. The difference is the theoretical background that students can now bring to a subject. The message I see is that requirements for enrollment in a class or acquisition of a degree should be ample, chosen carefully and taken seriously; in particular, they should not be discarded frivolously.

I would summarize the value of history in general by saying that, to understand any object, you need to understand its origins. I am taking as an axiom, in the context of this academic document for an academic credential in an academic department of statistics, that it is worthwhile to understand fundamental ideas in statistics.

Since I am not applying for funding, I will now cease salesmanship and get down to business.

This thesis is a partial history of one idea: the inner product, as used in statistics. The inner product is arguably as fundamental in statistics as it is in differential equations and physics, where there appears to be more cognizance of its role. In basic statistics there are two inner products often used,

$$\langle X, Y \rangle_1 \equiv E(XY), \quad \text{and} \quad \langle X, Y \rangle_2 \equiv \text{Cov}\,(X, Y)$$

(the latter is actually a *pre*-inner product, since there are random variables whose variance is zero). Since $\langle \cdot \rangle_1$ is also covariance, when restricted to random variables of mean zero, to feel like statisticians rather than mathematicians we can visualize either of these as covariance. Note that the standard inner product in $\mathbf{R}^n$,

$$\left\langle \vec{a}, \vec{b} \right\rangle \equiv \sum_{k=1}^{n} a_k b_k$$

is a constant multiple of $\langle \cdot \rangle_1$, on the probability space (Ω, P),

$$\Omega \equiv \{1, 2, ..., n\}, \quad P(\{k\}) \equiv \frac{1}{n}, \quad 1 \leq k \leq n,$$

with $X(k) \equiv a_k, \quad Y(k) \equiv b_k, 1 \leq k \leq n$.

In the appendices, after presenting sufficient material on inner products for this thesis, I try to indicate how analysis of variance, the general linear model, conditional probability, sufficiency, covariance, and variance reduction are all expressions of this one idea. Of course, people developing these ideas did not realize their unity, nor initially have access to the idea of an inner product. Looking back at the history of these ideas is like looking at a picture of your now-grown child as a baby; you see the beginnings of a personality that you didn't see then, because it was nascent. The evolution of species is another example of participants in the development of a pattern being unaware of the pattern.

I will treat separately some history of (a) least squares and analysis of variance, (b) the general linear model, (c) conditional probability and covariance, and (d) sufficiency. I will not go into depth on any of them; my goal is to trace some aspects of these lines of inquiry that led separately to the same unifying idea. Where possible, I will give references to more detailed histories. For the sake of intellectual accessibility, in particular to avoid both measure theory and social science, I will restrict myself to subjects whose study began before the second half of the nineteenth century.

There is a surprising shortage of studies of the history of statistics. For studying the history of statistics before 1930, I recommend [H], a thorough study emphasizing ideas over personalities. For those particularly interested in the application of statistics to social science, I recommend [St2], although there is again a truncation, this time at 1900. The best reference for more recent history of statistics is the notes at the ends of chapters in [Le], where critical comments about both frequentist and Bayesian statistics are made.

I. ORIGINS OF LEAST SQUARES AND ANALYSIS OF VARIANCE

If statistics is defined to be "decision making under uncertainty" (from the Ohio State Department of Statistics webpage, www.stat.ohio-state.edu/news/emphasis.html), perhaps the simplest example of statistics would be measuring the same thing repeatedly, but getting different results. For example, when I weigh myself, my scale might say 195 pounds. Another weighing gives 190, another 194. There is definitely uncertainty; my decision might be what answer to give to an authority figure asking the question, "How much do you weigh?"

For the weighing example, it might now seem natural to take the average, $193 = \frac{1}{3}(190+194+195)$ pounds as my estimate. This is an example of a *combination of observations;* the average, which also happens to be the best least-squares solution of the inconsistent system of equations

$$195 = a, \;\; 190 = a, \;\; 194 = a, \tag{1.1}$$

is the solution of the equation obtained by adding together the three equations.

Thus, to study the history of least squares it is natural to begin with the combination of observations, that is, algebraic manipulations of possibly contradictory data. According to [St2, p. 11], as late as the nineteenth century, much of statistics was given this name.

The general question here is, what to do with an inconsistent set of equations? If we believe that it *would* be consistent if only our measurements and model were accurate, what do we give as an estimate of the solution we would obtain in that ideal scenario?

The simple special case above, taking the average $\overline{y} \equiv \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$ of contradictory measurements of the same quantity a,

$$y_1 = a, \;\; y_2 = a, \;\; \ldots, \;\; y_n = a, \tag{1.2}$$

was used as early as the late sixteenth century by the astronomer Tycho Brahe (see [Pl1, pp. 122–124]). There is reason to believe that some idea of the average being more accurate than individual measurements existed at the end of the seventeenth century (see [Pl1, p. 124]). Roger Cotes went so far as to use a weighted average; this result was published in 1722, but had to have been written earlier, since he died in 1716; see [E, p. 202], [Pe-K, p. 155], and [H, pp. 93–94].

To an eighteenth-century scientist, combining observations was not necessarily considered a good thing to do. If the observations were made under much different circumstances, it was generally felt that they shouldn't be combined, that observations with large error would contaminate the more accurate measurements.

This is not an unreasonable concern. Suppose, in my weighing example, I had two scales, one accurate, the other always subtracting ten pounds from my weight. Averaging the good scale's measurements with the bad would certainly damage the accurate data; here I'm disregarding benefits to my self-image resulting from this particular inaccuracy.

Even for the simplest special case of an average, it was sometimes felt that one measurement might be as good as the average of several measurements. Here is a quote from Thomas Simpson, 1755 ([St2, p. 90], [H, p. 35], [Pl-K, p. 155]) " ... some persons, of considerable note, have been of opinion, and even publickly [sic] maintained, that one single observation, taken with due care, was as much to be relied on as the Mean [sic] of a great number."

Many problems of the eighteenth century yielded inconsistent sets of equations much more complex than the three equations given by my imperfect scale (1.1). The cartographer and astronomer Tobias Mayer, in 1750, studied the libration of the moon. This refers to the fact that the moon does not show precisely the same face to the earth at all times; about sixty percent of the moon's face is visible to us at some time. His observations led to 27 equations in 3 variables. The mathematician Leonhard Euler, in 1749, studied the effects of Jupiter and Saturn on each other's motions; more precisely, Jupiter being larger, he studied the effect of Jupiter on Saturn's motions. This "three-body problem" (Jupiter, Saturn, and the Sun) is immensely more complex than the "two-body problem" of calculating Saturn's motion due only to the influence of the Sun. Euler came up with 75 equations in 8 variables.

Mayer first used "the method of selected points" ([H, p. 95]): choosing 3 (not too similar) of the 27 equations, and solving those for the 3 variables. He was aware of the disadvantages of this method; it is desirable to involve all the data, and there are $\binom{27}{3} = 2925$ possible choices of 3 equations.

Mayer then broke his 27 equations into 3 groups of 9 equations each, and, within each group, added the equations together. He then solved the resulting system of 3 equations in 3 variables. Because Euler noticed a periodicity of 59 years in 6 of the 8 variables, he was willing to combine equations that were a multiple of 59 years apart, to solve for the 2 variables that were not periodic. But he was unwilling to combine other observations, partly because of their dissimilarity; the observations were spread out over the period 1582–1745. Mayer's observations were made between April 1748 and March 1749.

Laplace, also studying Jupiter and Saturn in 1787, had a system of 24 equations with 4 variables. He obtained 4 equations in the same variables with a much more elaborate combination of observations. One equation was the sum of all 24 equations; another

$$(1^{st} + 2^{nd} + \cdots + 12^{th}) - (13^{th} + 14^{th} + \cdots + 24^{th});$$

another

$$(3^{rd} + 4^{th} + 10^{th} + 11^{th} + 17^{th} + 18^{th} + 23^{rd} + 24^{th}) - (1^{st} + 7^{th} + 14^{th} + 20^{th});$$

and, finally,

$$(2^{nd} + 8^{th} + 9^{th} + 15^{th} + 16^{th} + 21^{st} + 22^{nd}) - (5^{th} + 6^{th} + 12^{th} + 13^{th} + 19^{th}).$$

See [St2, pp. 31–39] for a detailed discussion, including speculation on why Laplace chose these particular linear combinations.

The technique of Mayer and Laplace was sometimes called the "method of averages," and was often the method of choice for dealing with inconsistent linear systems, until least squares (see [H, pp. 107–108] and [F2]).

All these combinations of observations can be considered statistical techniques, in that they give an estimate of a parameter that is definitely uncertain. The choice of *which* combination to use was subjective, and did not seem to have a uniform motivation.

Note that the method of averages inevitably involves a loss of information. For example, if we replace the two equations

$$x = y, \quad y = z$$

with their sum

$$x + y = y + z, \quad \text{equivalent to} \quad x = z,$$

then we have lost the information $x = y$.

What I consider the most important step in developing the technique of least squares is using minimization of error, meaning a measurement of the distance between the model and the data, as an explicit criterion for choosing the combination of observations. This was first done by the Jesuit Roger Boscovich, in 1757.

Boscovich was estimating the ellipticity of the earth by measuring meridian arcs at different latitudes. He had five equations with two variables to consider. In his initial analysis, in 1755, he solved each of the ten ($\binom{5}{2}$) pairs of equations, and combined them in various ways; sort of a *tour de force* of the method of averages used by Mayer and Laplace. By 1757, he had developed the following technique for his 1755 data. For simplicity and familiarity, let's write his equations in the form

$$y_i = a + bx_i, \quad 1 \le i \le n, \tag{1.3}$$

where (x_i, y_i) are data, a and b are parameters to be estimated, and (for this data) $n = 5$. The error in the i^{th} observation, in whatever choice of a and b is made, is $(y_i - (a + bx_i))$, $1 \le i \le n$. Boscovich wanted a and b to satisfy

$$(1)\ \sum_{i=1}^{n}(y_i - (a + bx_i)) = 0 \quad \text{and} \quad (2)\ \sum_{i=1}^{n}|y_i - (a + bx_i)| \text{ is minimized .}$$

Notice how close to a least-squares best fit this is: merely replace $|y_i - (a + bx_i)|$ with $(y_i - (a + bx_i))^2$ in (2); that is, least squares demands that

$$(3)\ \sum_{i=1}^{n}(y_i - (a + bx_i))^2 \text{ is minimized .}$$

In fact, (1) follows automatically from (3). The only difference between (2) above and least squares is in how we measure the distance between two points in $\mathbf{R}^n$;

$$\|\vec{x} - \vec{y}\|_1 \equiv \sum_{k=1}^{n} |x_k - y_k| \quad \text{(mean deviation)}$$

or

$$\|\vec{x} - \vec{y}\|_2 \equiv \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad \text{(standard deviation)}.$$

It is also of interest to consider the one-dimensional analog with which we began this chapter: getting an approximate solution (a choice of parameter a) of (1.2).

The analogues of Boscovich's conditions are

$$(1') \ \sum_{i=1}^{n} (y_i - a) = 0 \quad \text{and} \quad (2') \ \sum_{i=1}^{n} |(y_i - a)| \text{ is minimized,}$$

and least squares is now

$$(3') \ \sum_{i=1}^{n} (y_i - a)^2 \text{ is minimized.}$$

Both $(1')$ and $(3')$ yield the average of the data $y_1, y_2, \ldots, y_n$, while $(2')$ gives us the median of the data.

At least at first glance, there is no reason why one could not generalize (2) of Boscovich's method by replacing the usual least-squares estimator of the parameter $\vec{\beta}$ in

$$\vec{Y} = X\vec{\beta}$$

with one that minimizes

$$\sum_{k=1}^{n} |Y_k - (X\vec{\beta})_k|$$

(instead of $\sum_{k=1}^{n} (Y_k - (X\vec{\beta})_k)^2$).

Boscovich used Newton's "geometric" style of argument, which means drawing pictures in lieu of an analytic argument. See [E] for a description.

Laplace in 1789 gave an analytic presentation of Boscovich's technique, calling it the "Method of Situation." In 1799 he generalized Boscovich's method by throwing in weights; that is, $(y_i - (a + bx_i))$ is replaced with $w_i(y_i - (a + bx_i))$, $i = 1, 2, \ldots, n$, for appropriate weights w_i (see [St2, pp. 51–55] and [H, Chapter 6.7, pp. 112–115]).

Euler in 1749 and Lambert in 1765 introduced an L^∞ analogue of (2) above, that is, searching for a and b that minimize

$$\max{}_{1 \le i \le n} |y_i - (a + bx_i)|;$$

a popular shorthand for this is "minimax." See [Sh]. Laplace gave an explicit technique for this in 1783 (see [Pl2], also [H, Chapter 6.6, p. 108]).

Before getting to least squares, the method that "won" historically, I want to emphasize the similarity between minimax, Boscovich's method, and least squares. The only difference is in the choice of norm, in measuring total error from the errors $e_k, 1 \le k \le n$, in each observation:

$$\max{}_{1 \le k \le n} |e_k|, \quad \sum_{k=1}^{n} |e_k|, \quad \text{or} \ \sum_{k=1}^{n} |e_k|^2.$$

The mathematician Legendre, in 1805, near the end of a memoir studying comets, stated "But it is possible to reduce them [the errors] further by seeking the *minimum* of the sum of the squares of the quantities ... " ([St2, p. 57]). In an appendix, after an introduction including " ... there is no one [principle] more general,

more exact, and more easy to apply than that which we have made use of in the preceding researches, and which consists in making the sum of the squares of the errors a *minimum*"([H, p. 119]; in both quotes the italics are due to Legendre). He then applied this technique to the measurements of meridian arcs, replacing (2) above with (3).

More generally, in this appendix, he considered the general linear system

$$y_i = \sum_{j=1}^{m} x_{ij}\beta_j \quad (1 \le i \le n) \tag{1.4}$$

where the x_{ij}s are given and the β_js are to be estimated; in matrix language this is

$$\vec{y} = X\vec{\beta} \quad (\vec{y} \in \mathbf{R}^n,\, \vec{\beta} \in \mathbf{R}^m, X \text{ an } n \times m \text{ matrix}), \tag{1.5}$$

and derived the normal equations, whose solution, call it $\hat{\beta}$, minimizes

$$\|\vec{y} - X\vec{\beta}\|_2^2 \equiv \sum_{i=1}^{n} \left(\vec{y} - X\vec{\beta}\right)_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j \right)^2$$

by taking derivatives $\frac{d}{d\beta_j}$ and setting them equal to zero.

It should be mentioned here that having all the partial derivatives equal to zero is not sufficient to guarantee that we have a minimum; to guarantee even a local minimum, it must be checked that the matrix of second derivatives is strictly negative definite. In this case, the matrix of second derivatives equals $-2X^TX$, which we now know is strictly negative definite if X is full rank. See Theorem A2.1 for a simpler and more intuitive inner-product approach to this minimization problem.

Legendre then mentions the special case of $m = 1$ (see (3′) above), and shows that the best least-squares estimator in that case, minimizing $\sum_{k=1}^{n}(y_i - \beta)^2$, is $\hat{\beta} = \overline{y}$.

Legendre is the first person to publicly discover the least-squares technique. He then fell victim to what many mathematicians during Gauss's lifetime suffered, Gauss's "secret notebook," a collection of results Gauss obtained but didn't publish. After Legendre published his least-squares results, Gauss asserted that he had discovered the result in 1795, without publishing it. After Legendre protested, Gauss produced witnesses to the fact that Gauss had discussed it soon after 1795. See [Pl2] for a discussion, including interesting correspondence from the time.

What is still missing, to make least squares the statistical technique it is now, is a probability distribution.

The self-taught mathematician Thomas Simpson was the first ([St2, p. 90], [E], [Pl2]) to decompose an observation, X, into the true value, θ, plus an error ϵ :

$$X = \theta + \epsilon. \tag{1.6}$$

This opened the possibility of applying probability to the estimate of the true value, by giving the error a probability distribution. He presented this decomposition in a 1755 letter to the Earl of Macclesfield titled "On the Advantage of Taking the Mean of a Number of Observations, in practical Astronomy." Thomas Bayes, of Bayesian inference fame, cast doubt on Simpson's optimism in asserting that the error of the average would always shrink (in a convergence-in-probability sense) by taking more observations, pointing out that if, for example, your measuring device always overestimates what it measures by two units, then that same overestimate will appear in the average. Probably in part due to this criticism, Simpson, in a 1757 revision of his letter of 1755, added hypotheses that essentially mean that the expected value of the error is zero, and that the errors are bounded above and below (see [St2, p. 95], [H, p. 37]).

The decomposition (1.6), with a probability distribution on the errors, was also developed, probably independently, by Lambert in 1760, where it is believed that the term "theory of errors" was introduced (see [H, Chapter 5.1, pp. 79–83], [F1, p. 79] and [Sh]), Lagrange in 1776 (see [F1, p. 79], and [Sh]), and Daniel Bernoulli in 1778 (see [F1, p. 79] and [K1]). Both Lambert and Bernoulli looked for a maximum likelihood estimator of θ; see [F1, pp. 79–80] for details.

Laplace, between 1772 and 1781, gave a methodical presentation of "inverse probability"—inferring causes from effects. Part of this was the decomposition (1.6), with a probability density ϕ on the error ϵ, that Laplace called a "curve of errors." He asked only that ϕ be symmetric and decreasing to zero to the right of its axis of symmetry. His goal was to choose an estimator of θ from observed values X (a "choice of mean" of the observed values) that minimized the posterior expected loss, with absolute-value loss function and uniform prior. He showed that this is the median of the posterior distribution of θ. His favorite choice of error curve (after trying many others—see [H, pp. 87–88]) was (for some constant m) the double exponential

$$\phi(x) = \frac{m}{2} e^{-m|x|} \ \ (x \in \mathbf{R}).$$

With this error curve, he found great difficulty in calculating the desired median. In modern terminology, $\pi(\theta) \equiv 1$, so that the posterior distribution,

$$\pi(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)\pi(\theta)}{\int f(\vec{x}|\theta)\pi(\theta)\,d\theta} = \frac{f(\vec{x}|\theta)}{\int f(\vec{x}|\theta)\,d\theta}$$

is proportional to

$$f(\vec{x}|\theta) = \prod_{k=1}^{n} f(x_k|\theta) = \prod_{k=1}^{n} \phi(x_k - \theta) = \left(\frac{m}{2}\right)^n \exp\left(-m \sum_{k=1}^{n} |x_k - \theta|\right). \tag{1.7}$$

The median, call it $\mathrm{med}(\theta)$, would be the real number that satisfies

$$\int_{-\infty}^{\mathrm{med}\ (\theta)} \exp\left(-m \sum_{k=1}^{n} |x_k - \theta|\right) d\theta = \int_{\mathrm{med}\ (\theta)}^{\infty} \exp\left(-m \sum_{k=1}^{n} |x_k - \theta|\right) d\theta.$$

Even for $n = 3$, Laplace found this difficult; see [H, Chapter 10.4, pp. 171–176] and [St2, pp. 113–117]. He was also frustrated by the fact that, when he could solve for $\mathrm{med}(\theta)$, it was not the average of the data, which he felt it should be.

Gauss in 1809 took a different approach to (1.6). First, for a probability distribution on the error ϵ, he wished to find the mode, rather than the median, of the posterior distribution of θ, for a uniform prior; in frequentist language, he was looking for the MLE—maximum likelihood estimator. Second, rather than starting with an error curve and finding the best estimator, he started by assuming that the best estimator—in this case, the MLE—should be the average $\overline{x}$ of the data x. He then showed that that implied that the probability distribution on ϵ must be normal. He wrote this as the density ϕ having the form

$$\phi(\Delta) = \frac{h}{\sqrt{\pi}} \exp(-h^2\Delta^2),$$

where h was a measure of precision (equal to $\frac{1}{\sqrt{2}\sigma}$). He also showed the converse: if we assume ϕ has a normal distribution, then it follows that $\overline{x}$ is the MLE, since, as with (1.7), the distribution of the posterior distribution of $\theta, \pi(\theta|\vec{x})$, is now (compare with (1.7)) proportional to

$$\exp\left(-h^2 \sum_{k=1}^{n} (x_k - \theta)^2\right), \tag{1.8}$$

which is maximized by choosing $\theta = \overline{x}$. More precisely, Gauss noted that

$$\sum_{k=1}^{n} (x_k - \theta)^2 = \sum_{k=1}^{n} (x_k - \overline{x})^2 + n(\theta - \overline{x})^2,$$

so that the posterior distribution of θ is proportional to

$$\exp\left(-nh^2(\theta - \overline{x})^2\right), \tag{1.9}$$

a normal distribution with precision $\sqrt{n}h$ and mean $\overline{x}$.

More generally, if different observations are assumed normal, but with different precisions h_i, that is,

$$\phi_i(x_i) = \frac{h_i}{\sqrt{\pi}} \exp(-h_i^2 x_i^2) \quad (1 \le i \le n), \tag{1.10}$$

then Gauss showed that the MLE is the weighted average

$$\hat{\theta} = \frac{\sum_{i=1}^n h_i^2 x_i}{\sum_{i=1}^n h_i^2},$$

and that the posterior distribution of θ is normal with precision $\sqrt{\sum_{i=1}^n h_i^2}$ and mean $\hat{\theta}$, by decomposing $\sum_{k=1}^n h_k^2 (x_k - \theta)^2$ as we did above.

Gauss also found the same $\hat{\beta}$, the best least-squares solution of (1.5), that Legendre did. His original contribution here (not counting things done earlier but not published) was the probabilistic setting: he put independent normal distributions on the errors $\epsilon_i \equiv (Y_i - (X\vec{\beta})_i)$, found the posterior distribution of $\hat{\beta}$, given independent uniform prior distributions on β_i, and showed that $\hat{\beta}$ was the MLE of $\vec{\beta}$.

Since this represents a major turning point in the development of the general linear model, and I cannot discuss Gauss's work in detail without doubling the length of this section, I will put off this essential synthesis of (1.5) and (1.6) to the next.

Towards the end of 1810, Laplace saw Gauss's work, and realized that his (Laplace's) work described above, with the double exponential error curve, would work much better with a normal error curve. The posterior distribution of θ is now proportional to (1.9) instead of (1.7). The distribution (1.9), being symmetric and unimodal (in θ), has median, mean, and mode all equal to $\overline{x}$. Thus, $\overline{x}$ is not only the MLE, it minimizes the posterior expected loss with respect to the absolute-value loss function (in fact, also with respect to the squared loss function). Laplace could give a much better rationale for using a normal error curve, since he had recently proven the central theorem for sums of independent, identically distributed random variables with finite variance.

It is interesting to note that, if Laplace had been, like Gauss, looking for the MLE of θ, he could have found it easily even with his original double exponential error curve: the mode of the posterior distribution in (1.7) is $\hat{\theta} =$ (median of $x_1, \ldots, x_n$).

If only out of patriotism, I should mention here that an American mathematician, R. Adrain, in solving a surveying problem, independently developed least squares from a probabilistic point of view in 1808; see [H, pp. 368–373] or [St3].

This chapter discusses two developments that coalesce at the conclusion of the chapter. The first is purely algebraic: how to estimate the solution of a linear system (the correct theoretical model) that has been transformed into an inconsistent system by errors in either measurement or choice of model. The second is the basic general linear model or one-way analysis-of-variance model (1.6). The first development culminated in the least-squares technique, the second in placing a normal distribution on the error, as the methods of choice. I'd like to briefly discuss why these became the preferred methods.

As mentioned just before introducing Legendre, in estimating the solution of a linear system, three ways to measure the total error to be minimized, from the individual errors $e_k, 1 \le k \le n$, arose:

$$\max_{1 \le k \le n} |e_k|, \quad \sum_{k=1}^n |e_k|, \quad \text{or} \quad \sum_{k=1}^n |e_k|^2.$$

For shorthand, call them maximum error, absolute error, and squared error, respectively.

It is not surprising that maximum error and absolute error appeared first. These are more natural and simple to describe. Maximum error is like the strength of a chain: only as strong as the weakest link. The difference between absolute error and squared error is like the difference between mean deviation ("average distance from the mean") and standard deviation ("square root of the average of the squared distance from the mean"). The latter definition sounds suspiciously awkward. The fact that it takes so many more words to define standard deviation should make one suspicious of its use. More generally, the squaring in squared

error looks artificial; why are we exaggerating the effect of larger errors by squaring (the fact that the mean, the minimizer of squared-error loss with one parameter, is sensitive to extreme values, is a reflection of this)?

Historically, squared error came to be used instead of maximum or absolute error, because the calculations are much easier; compare Laplace's work in [St2, pp. 51–55] and [H, Chapter 6.7, pp. 112–115], for absolute, and in [Pl2] and [H, Chapter 6.6, p. 108], for maximum, to the least-squares technique. See [E, pp. 209–210].

But this begs the question. What is it about squared error that makes it easier to use than maximum or absolute error? The answer is that it is a norm that comes from an inner product:

$$\sum_{k=1}^{n} |e_k|^2 = \langle \vec{e}, \vec{e} \rangle \quad \text{where} \quad \langle \vec{x}, \vec{y} \rangle \equiv \sum_{k=1}^{n} x_k y_k.$$

For such norms, closed, convex sets are not only guaranteed to have a unique point of minimum norm (not true for maximum and absolute error), but also the concept of orthogonality (inner product equal to zero) gives us an intuitive and straightforward way to obtain that point; see Theorem A1.3(d), Remark A1.4, and Theorem A2.1.

Standard deviation has the same advantage over mean deviation. Standard deviation also has a more qualitative desirability in its relation to the normal distribution, hence to asymptotics. This leads us to the second historical development, the choice of the normal distribution for the error term in (1.6).

The normal distribution is easy to work with because it has an inner-product structure. The joint density for n independent, standard-normal random variables $\vec{Z} \equiv (Z_1, Z_2, \ldots, Z_n)$ is

$$f(\vec{x}) = c \exp\left(-\frac{1}{2} \|\vec{x}\|^2 \right),$$

where $\|\vec{x}\|$ is a norm from the inner product defined above. More generally, if $\vec{Y}$ is a multivariate normal with covariance Σ, then

$$f_{\vec{Y}}(\vec{y}) = c \exp\left(-\frac{1}{2} \|\vec{y}\|_\Sigma^2 \right),$$

where

$$\|\vec{y}\|_\Sigma^2 \equiv \langle \vec{y}, \vec{y} \rangle_\Sigma, \quad \langle \vec{x}, \vec{w} \rangle_\Sigma \equiv \left\langle \Sigma^{-1} \vec{x}, \vec{w} \right\rangle,$$

and $\langle \cdot, \cdot \rangle$ is the usual inner product above.

This inner-product structure also explains why the standard deviation, an inner product measurement, is so fundamental to the normal distribution.

The hidden idea in both these developments is the inner product.

II. INTERMEDIATE HISTORY OF LEAST SQUARES AND THE GENERAL LINEAR MODEL

In the previous section, we have seen how statisticians looking (unfortunately not in matrix language) at the general linear model

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon} \tag{2.1}$$

focused, at the beginning of the nineteenth century, on the *best least-squares estimator* $\hat{\beta}$ of $\vec{\beta}$, defined to be a vector that minimizes

$$\|\vec{Y} - X\vec{\beta}\|^2 = \|\vec{\epsilon}\|^2 \equiv \sum_{k=1}^{n} \epsilon_k^2.$$

Here $\vec{Y} \in \mathbf{R}^n, \vec{\beta} \in \mathbf{R}^m$, X is an $n \times m$ matrix, and $\{\epsilon_k\}_{k=1}^n$ is a sequence of independent, identically distributed random variables. We will assume X has rank m.

Legendre (1805) and Gauss (allegedly 1795; not publicly until 1809) independently introduced the best least-squares estimator, and showed that being a best least-squares estimator is equivalent to being a solution of the normal equations

$$X^T\vec{Y} = X^T X\vec{\beta}. \tag{2.2}$$

It is interesting that Gaussian elimination, to this day the most popular method of solving linear equations, whether done in matrix notation, by computer or by hand, was developed (by Gauss, of course) to solve the normal equations. Gauss also in 1809 showed that $\hat{\beta}$ is the maximum likelihood estimator of β if and only if ϵ_k is normal ($1 \le k \le n$).

Another sort of minimization, among a class of unbiased estimators of $\vec{\beta}$, is to look for the estimator of minimum variance. It is surprising that, in the class of linear estimators (shorthand for linear combinations of the data $\vec{Y}$), $\hat{\beta}$ is the unbiased estimator of minimum variance (for simplicity, I am assuming X is full rank). More precisely, Gauss showed in 1821 that, for $1 \le k \le n$, the unbiased linear estimator of β_k of minimum variance is $\hat{\beta}_k$. The result is often called the Gauss-Markov theorem because Neyman, unaware of Gauss's proof, believed Markov, whose proof appeared in 1912, was the first to prove it.

Thus $\hat{\beta}$ is optimal in at least two ways, minimizing the norm of the error and the variance, and, for ϵ_k normal, is also optimal in being a maximum likelihood estimator.

In 1823, Gauss generalized this to the following: for any $\vec{c} \in \mathbf{R}^m$,

$$\left\langle \vec{c}, \hat{\beta} \right\rangle \equiv \sum_{k=1}^{m} c_k \hat{\beta}_k$$

is the unbiased linear estimator of $\langle \vec{c}, \beta \rangle$ of minimum variance. Neyman and David (see [D-N]) believed they were proving this for the first time in 1938.

Laplace produced an asymptotic version of the "Gauss-Markov" theorem in 1811, although the proof was arguably not complete until 1816; see [H, Chapters 20.6 and 20.7]. By "asymptotic version" I mean he showed that, when ϵ_k has finite variance ($1 \le k \le n$), then, for any linear function $\left\langle \vec{c}, \vec{\beta} \right\rangle$ of $\vec{\beta}$, the variance of the limiting distribution, as $n \to \infty$, of $\left\langle \vec{c}, \hat{\beta} \right\rangle$ is less than or equal to the variance of the limiting distribution of any unbiased linear estimator of $\left\langle \vec{c}, \vec{\beta} \right\rangle$.

Many other popular results in the general linear model or analysis of variance are due to Gauss in 1823. He showed that $s^2 \equiv \frac{1}{n-m}(\vec{Y} - X\hat{\beta})$ is an unbiased estimator of the variance of ϵ_k, and found the variances of $\hat{\beta}, \left\langle \vec{c}, \hat{\beta} \right\rangle$, and s^2. As with the "Gauss-Markov" theorem, Laplace produced asymptotic analogues of these variance formulas in 1811.

Gauss in 1823 also considered the effect on $\hat{\beta}$ of making another observation Y_{n+1}. Specifically, in (2.1), replace

$$\vec{Y} \text{ with } \begin{bmatrix} \vec{Y} \\ Y_{n+1} \end{bmatrix}, \ X \text{ with } \begin{bmatrix} X \\ h^T \end{bmatrix}, \ \vec{\epsilon} \text{ with } \begin{bmatrix} \vec{\epsilon} \\ \epsilon_{n+1} \end{bmatrix}, \tag{2.3}$$

where $Y_{n+1} \in \mathbf{R}, h \in \mathbf{R}^m$, and ϵ_{n+1} is a random variable independent of $\vec{\epsilon}$, with the same distribution as $\epsilon_k, 1 \le k \le n$. It would be inefficient to recalculate $\hat{\beta}$ from scratch. Gauss sought a simple formula that would calculate the new (using $Y_1, Y_2, ..., Y_{n+1}$) $\hat{\beta}$ from the old (using $Y_1, Y_2, ..., Y_n$) $\hat{\beta}$, along with other information, such as $(X^T X)^{-1}$, that would be likely to be available.

Using the inner-product definition of the least-squares estimator of $\vec{\beta}$ in (2.1), it is a straightforward calculation to derive Gauss's result. To be consistent with his theorem as stated in [H, p. 480], let $b \equiv \hat{\beta}$, the least-squares estimator of $\vec{\beta}$ in (2.1), and denote by $\hat{b}$ the new least-squares estimator after another observation, as in (2.3). Also let $e \equiv (Y_{n+1} - h^T b), c \equiv (X^T X)^{-1} h, k \equiv (1 + h^T c)^{-1}$. Then for all $\vec{\beta} \in \mathbf{R}^m$,

$$\begin{aligned}
0 &= \left\langle \begin{bmatrix} \vec{Y} \\ Y_{n+1} \end{bmatrix} - \begin{bmatrix} X \\ h^T \end{bmatrix} \hat{b}, \begin{bmatrix} X \\ h^T \end{bmatrix} \vec{\beta} \right\rangle_{\mathbf{R}^{n+1}} \\
&= \left\langle \begin{bmatrix} \vec{Y} \\ Y_{n+1} \end{bmatrix} - \begin{bmatrix} X \\ h^T \end{bmatrix} (b + (\hat{b} - b)), \begin{bmatrix} X \\ h^T \end{bmatrix} \vec{\beta} \right\rangle_{\mathbf{R}^{n+1}} \\
&= \left\langle \vec{Y} - Xb - X(\hat{b} - b), X\vec{\beta} \right\rangle_{\mathbf{R}^n} + \left((Y_{n+1} - h^T b) - h^T (\hat{b} - b) \right) (h^T \vec{\beta}) \\
&= - \left\langle X(\hat{b} - b), X\vec{\beta} \right\rangle_{\mathbf{R}^n} + (e - h^T (\hat{b} - b)) \left\langle h, \vec{\beta} \right\rangle_{\mathbf{R}^n} \\
&= \left\langle -(X^T X)^{-1} (\hat{b} - b) + (e - h^T (\hat{b} - b)) h, \vec{\beta} \right\rangle_{\mathbf{R}^n}.
\end{aligned}$$

Thus

$$(X^T X)(\hat{b} - b) = (e - h^T (\hat{b} - b)) h,$$

so that

$$(\hat{b} - b) = (e - h^T (\hat{b} - b))(X^T X)^{-1} h \equiv (e - h^T (\hat{b} - b)) c. \tag{2.4}$$

This makes it natural to conjecture that $(\hat{b} - b) = \alpha c$, for some real α, which is determined by plugging into (2.4):

$$\alpha c = (e - h^T (\alpha c)) c = (e - \alpha h^T c) c,$$

so that $\alpha = (e - \alpha h^T c)$; the solution of this linear equation is $\alpha = (1 + h^T c)^{-1} e \equiv ke$; thus,

$$\hat{b} = b + kec. \tag{2.5}$$

Gauss also obtained expressions for the variance of the new $\hat{\beta}$ in terms of k and c above and the variance of the old $\hat{\beta}$. (2.5) was independently obtained by Cochran ([C]) in 1938. Plackett ([Pl3]) in 1950 generalized (2.5) to arbitrarily many new observations $Y_{n+1}, Y_{n+2}, ..., Y_{n+s}$; see [H, pp. 482–483].

Similar considerations of efficiency arise when adding another component, β_{m+1}, to $\vec{\beta}$ in (2.1). In general, all the components of $\hat{\beta}, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_m$ would have to be recalculated. In terms of X in (2.1), Gauss's addition of an extra observation Y_{n+1} corresponds to adding another row to X, whereas adding the extra parameter β_{m+1} corresponds to adding another column.

It would be nice if $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_m$ were unchanged by the addition of β_{m+1} to (2.1), leaving only $\hat{\beta}_{m+1}$ to calculate. This may be shown to be equivalent to the columns of X being *orthonormal*, meaning both orthogonal (inner product of two different columns equals zero) and of norm one (inner product of a column with itself equals one). The sufficiency is clear:

$$\hat{\beta} = (X^T X)^{-1} X^T \vec{Y} = X^T \vec{Y},$$

when the columns of X are orthonormal, so that, for $1 \le k \le m + 1$, $\hat{\beta}_k$ is the inner product of the k^{th} column of X (the k^{th} row of X^T) with $\vec{Y}$.

Thus, given (2.1) for arbitrary X, the goal is to rewrite (2.1) in such a way that X is replaced by a matrix with orthonormal columns. The real work is in achieving orthogonality; an orthogonal set is changed to an orthonormal set merely by dividing each vector by its norm.

Cauchy, in 1835, was the first to consider this sort of orthogonalization problem, although, analogous to Legendre's treatment of least squares, without the probabilistic setting. Many mathematicians and

statisticians worked on this problem throughout the nineteenth century; see [H, Chapter 25]. The modern simple and geometrically intuitive method known as Gram-Schmidt orthogonalization (see the Appendix, proof of Corollary A1.7) did not appear until 1907.

An interesting intermediate technique, due to Chebyshev and Gram, is the use of *orthogonal polynomials*, or, more generally, orthogonal functions. Consider the usual choice of X in (2.1):

$$X = [\vec{x}^0, \vec{x}^1, ..., \vec{x}^{m-1}]$$

where, for fixed $x_1, x_2, ..., x_n, 0 \leq k < m$, I define $\vec{x}^k$ to be the vector $[x_1^k, x_2^k, ..., x_n^k]^T$. For f and g arbitrary functions from the real line to itself, define the inner product

$$\langle f, g \rangle_{\vec{x}} \equiv \sum_{i=1}^{n} f(x_i) g(x_i).$$

Then, for $0 \leq j, k < m$, the usual $\mathbf{R}^n$ inner product of the j^{th} and k^{th} columns of X equals

$$\langle f, g \rangle_{\vec{x}}, \quad \text{for} \quad f(x) \equiv x^{j-1}, g(x) \equiv x^{k-1}.$$

Thus we may think of the columns of X as being the polynomials $1, x, x^2, ..., x^{m-1}$, and orthogonalizing the columns of X in $\mathbf{R}^n$ is equivalent to orthogonalizing the usual basis for $(m-1)^{st}$ degree polynomials with respect to the inner product $\langle \cdot \rangle_{\vec{x}}$. The resulting (after, say, Gram-Schmidt orthogonalization applied to $1, x, x^2, ..., x^{m-1}$) polynomials are called (a set of) *orthogonal polynomials*.

To summarize: every different choice of points $x_1, x_2, ..., x_n$ produces a different inner product $\langle \cdot \rangle_{\vec{x}}$, and every different inner product produces a different set of orthogonal polynomials. These have come to have an immense range of applications, much wider than statistics. Certain particular choices of inner product have produced particular orthogonal polynomials famous enough to be given a particular name. Among other things, it enables one to do with regular polynomials the same things one does with trigonometric polynomials $1, e^{i\theta}, e^{2i\theta}, ...$, where orthogonality already exists, and one can immediately write down (automatically orthogonal) Fourier series.

Laplace, in 1816, was performing orthogonalizations very similar to that ultimately named "Gram-Schmidt," in the process of determining the asymptotic distributions of $\hat{\beta}$ (see [H, Chapter 20.7]).

On a personal note, I think it is interesting that Laplace and Gauss, despite their extensive mutual intellectual stimulation, never met.

I should make it clear that all the results in this section were derived without matrices. This author is amazed that the matrix representation (2.1) of the general linear model did not appear until 1935 ([A]); a complete algebra of matrices was produced by Cayley in 1858 (not to mention the use of matrices by Chinese mathematicians to solve systems of equations in the Han dynasty, about 200 B.C.–200 A.D.). Perhaps even more surprising, and inhibiting of progress, was the lack of awareness of Gauss's 1821–1823 statistical work.

III. ORIGINS OF CONDITIONAL PROBABILITY AND COVARIANCE

Conditional probability, especially in multiplication laws, is part of the earliest formulations of probability. I have included covariance in the same section as conditional probability because of the intimacy of their relationship; see Appendix III, Propositions A3.10 and A3.14.

One could argue that any statistical inference is conditional probability: finding the probability of a cause given an effect, what Laplace called "inverse probability." Since my goal is much more humble than a complete history of statistics, I must be selective in my choice of topics in conditional probability. The earliest explicit appearance of conditional probability as statistical inference would be Bayesian inference, as done by Thomas Bayes.

Bayes was a Presbyterian minister in Tunbridge Wells, a suburb of London. His interests were theology, mathematics, and statistics, in that order. He published only two papers during his lifetime, *Divine Benevolence: Or, An Attempt to prove that the Principal End Of the Divine Providence and Government is the Happiness of his Creatures,* in 1731, and *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians against the objections of the Author of the Analyst,* in 1736. "The Analyst" was a paper by Bishop Berkeley criticizing (probably with much justification) the foundations of calculus. According to [S], Bayes' paper was "...not unlike Cauchy's treatment of limits,..."

No other papers were published by Bayes during his lifetime, although I have mentioned in Section I his 1756 criticism of Simpson's too-enthusiastic praise for taking the average of measurements. Bayes' results on the posterior distribution of the parameter θ in a binomial (n, θ) distribution when the prior distribution is uniform, in *An Essay towards solving a Problem in the Doctrine of Chances* (hereafter referred to as "the *Essay*," or "his *Essay*"), was not published until 1764, three years after his death, by Richard Price, another Presbyterian minister-mathematician-statistician. It is believed Bayes got the results in his *Essay* between 1746 and 1749 (see [D1] and [D2]).

If only because of widespread misunderstandings of the part of the *Essay* known as "Bayes' theorem," it is of interest to describe it in some detail. Bayes visualized tossing a ball randomly onto a rectangular table; for convenience I'll make the table the unit square $[0,1] \times [0,1]$ in the Cartesian plane. Draw a vertical line through the point where the ball landed. Now toss the ball randomly onto the table n times, and let U_n be the number of times the ball landed to the left of the vertical line. If θ is defined to be the distance from the vertical line to the y axis, that is, the x coordinate of the point where the ball originally landed, then

$$U_n \sim \text{ binom } (n, \theta).$$

Assuming a uniform prior distribution on θ, it is now a familiar calculation to get the posterior distribution; for $0 < \theta_1 < \theta_2 < 1$:

$$\begin{aligned} P(\theta_1 < \theta < \theta_2 \,|\, U_n = k) &= \frac{P(U_n = k,\ \theta_1 < \theta < \theta_2)}{P(U_n = k)} = \frac{\int_{\theta_1}^{\theta_2} P(U_n = k \,|\, \theta)\, d\theta}{\int_0^1 P(U_n = k \,|\, \theta)\, d\theta} = \frac{\int_{\theta_1}^{\theta_2} \binom{n}{k} \theta^k (1-\theta)^{n-k}\, d\theta}{\int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k}\, d\theta} \\ &= (n+1) \int_{\theta_1}^{\theta_2} \binom{n}{k} \theta^k (1-\theta)^{n-k}\, d\theta = \int_{\theta_1}^{\theta_2} \beta(k+1, n-k+1) \theta^k (1-\theta)^{n-k}\, d\theta; \end{aligned} \tag{3.1}$$

that is, the posterior distribution of θ, given $U_n = k$, is beta $(k+1, n-k+1)$.

This is a sophisticated version of (a special case of) what we now call "Bayes' theorem," involving as it does continuous distributions.

Among the many criticisms of Bayes' work, perhaps the most well-known is that of the quintessential criticizer, Fisher. He argued that choosing the uniform distribution for θ, as representing initial ignorance on the part of the observer, was arbitrary, because one could reparametrize by replacing θ with $\psi \equiv g(\theta)$, for some injective g, e.g., $\psi \equiv \theta^2$, and ψ would then fail to have a uniform distribution, although the observer would be identically ignorant of ψ.

As pointed out by [Mo] and [Ed], and [St2], this criticism is based on a misunderstanding of Bayes' paper. Bayes was not assuming a uniform distribution for θ, but for U_n (unconditional on θ); that is, $P(U_n = k) = \frac{1}{n+1}$, for $k = 0, 1, ..., n+1$. This is clear from the following quote from his paper, which I have copied from [H, pp. 142–3].

"And that the same rule is the proper one to be used in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it, seems to appear from the following consideration; viz. that concerning such an event I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another. For, on this account, I may justly reason concerning it as if its probability had been at first unfixed, and then determined in such a manner as to give no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another."

Note that, for a uniform prior distribution on θ, the calculation in (3.1) contains the conclusion that then U_n is uniform.

Conversely, assuming U_n is uniform, for arbitrary n, uniquely determines the prior distribution on θ. If μ is the prior distribution on θ, then, for $0 \le k \le n$,

$$\frac{1}{n+1} = P(U_n = k) = \int_0^1 P(U_n = k \mid \theta)\, d\mu(\theta) = \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k}\, d\mu(\theta);$$

in particular (it's interesting that this is sufficient to force μ to be uniform),

$$\frac{1}{n+1} = P(U_n = n) = \int_0^1 \theta^n\, d\mu(\theta) \ (n = 0, 1, 2, ...).$$

Since μ has finite support, specifying all its moments uniquely determines μ. In other words, the uniform prior on θ, not $g(\theta)$, is the only choice of prior that makes the marginal distribution of U_n uniform.

The detailed analysis in Bayes' paper leading to (3.1) was preceded by his personal construction of the foundations of probability. It is interesting that he defines probability in terms of expectation, reversing the order of definitions usually given prior to him, and in a way anticipating the modern definition of conditional probability which begins with conditional expectation.

I will quote here his propositions regarding conditional probability; see [H, Chapters 8.2–4], for all the propositions.

Prop. 3 The probability that two subsequent events will both happen is a ratio compounded of the probability of the first, and the probability of the second on supposition the first happens.

Prop. 5 If there be two subsequent events, the probability of the 2d $\frac{b}{N}$ and the probability of both together $\frac{P}{N}$, and it being 1st discovered that the 2d event has happened, from hence I guess that the 1st event has also happened, the probability I am in the right is $\frac{P}{b}$.

In other words, if A and B are events, with A preceding B, Prop. 3 says

$$P(A \cap B) = P(B|A)P(A),$$

while Prop. 5 says

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

It seems curious that the same result, except that past and future are interchanged, is stated as two different results. To the modern Bayesian, to the extent that I understand such a creature ($\sim$ "probability represents the uncertainty of the observer"), this distinction would be meaningless. And yet the Heisenberg uncertainty principle seems to imply an unavoidable distinction between past and future, stating, in effect, a sort of absolute uncertainty about the future, regardless of the expertise and knowledge of the observer.

I suppose, according to some sort of Taoist principle of balance, every thoughtful person needs a salesman. Richard Price was an effective publicist for Bayes' results, finding and stating significance that Bayes would probably have been too humble to assert. But I think it is a tragedy that Price wrote his own introduction to the *Essay*, and published it in place of Bayes'; Bayes' introduction to his *Essay*, along with any other clue about what he thought of his results, seems to be lost to the world. Fisher used the fact that Bayes didn't try to publish his *Essay* as evidence of its (according to Fisher) intellectual invalidity. Stigler ([St2, p.

130] speculates that Bayes was inhibited by his lack of useful techniques for calculating the beta distribution integrals

$$\int_a^b x^\alpha (1-x)^\beta \, dx$$

needed for posterior probabilities of the parameter θ.

Almost certainly unaware of the work of Bayes, Laplace, beginning in 1774, developed a theory of *inverse probability:* reasoning from effect to cause, as opposed to direct probability: reasoning from cause to effect. Bernoulli, in *Ars Conjectandi* (1713) also made this distinction, using the law of large numbers, with the usual example of an urn full of white and black balls: direct (which he called "an *a priori* determination") being knowing the number of white and black balls, and determining the probability of drawing a white ball; indirect (which he called "an *a posteriori* determination") being counting the relative frequency of white balls in many drawings.

Laplace's work included, in 1814, a general statement of what is now called Bayes' theorem,

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^n P(B|A_j)P(A_j)}. \tag{3.2}$$

Laplace's method of choice was to set up his analysis so that each cause A_j is equally likely; in other words, put a uniform prior distribution on $\{A_j\}_{j=1}^n$, so that (3.2) becomes

$$P(A_k|B) = \frac{P(B|A_k)}{\sum_{j=1}^n P(B|A_j)}. \tag{3.3}$$

This uniform prior was not meant as an arbitrary assumption; Laplace merely meant that, if the initial formulation of the problem did not permit a uniform prior, the causes A_j should be further broken down until we did have a partition into equally likely causes.

This formulation has been called the "principle of insufficient reason"([H, p. 143, and top of p. 159]), or sometimes the "indifference principle." Laplace also used nonuniform prior distributions; see [St2, pp. 135–136].

Laplace's 1774 *principle of inverse probabilty*, copied from [H, p. 160] or [St2, p. 102], is the following.

"If an event can be produced by a number n of different causes, the probabilities of the existence of these causes given the event are to each other as the probabilities of the event given the causes, and the probability of the existence of each of them is equal to the probability of the event given that cause divided by the sum of all the probabilities of the event given each of the causes."

In other words, for A_k and B as in (3.3), as follows from (3.3),

$$\frac{P(A_i|B)}{P(A_j|B)} = \frac{P(B|A_i)}{P(B|A_j)}.$$

We have discussed in Section I Laplace's use of the double exponential as an "error curve," that is, a distribution for the error ϵ in (1.6). [St2, pp. 113–117] shows how a fundamental misunderstanding of conditional probability on the part of Laplace led to mistakes in his calculation of the median of the posterior distribution of ϵ. He essentially mistook the proportionality of $p(x|y)$ (the conditional density) and $p(x, y)$ (the joint density) for equality.

Laplace also did extensive and very clever work making the approximations of the beta probability integral needed, as described in our discussion of Bayes' work, for estimating posterior probabilities of a binomial parameter. See [H, Chapter 10.3].

Notions of covariance, correlation, and dependence were not expressed explicitly until Galton introduced them in 1877, although Bravais in 1846, and Gauss and Laplace earlier, had correlation appearing implicitly in their representations of normal distributions. These ideas had not arisen naturally, because the observations being studied—astronomical or geodetic measurements—tended to be automatically independent. In Galton's measurements of biological and social phenomena, especially given his focus on inheritance, correlation and dependence are crucial. The development, beginning mathematically with Edgeworth in 1892, was awkward, partly because it was done by a new British school of statisticians who were often unaware of

the work of Laplace and Gauss or other mathematicians from the European continent; for example, it was not realized at first that regression is a special case of the general linear model. See [L], [H, Chapters 26.4 and 26.5], or [St2, Part Three] for details.

Simple (two-dimensional) orthogonality arguments are sufficient to see how covariance and correlation arise naturally in regression. Let X and Y be random variables, and let W be the space of affine (often misstated as "linear") functions of X, that is,

$$W \equiv \{\alpha + \beta X \mid \alpha, \beta \in \mathbf{R}\}.$$

In Appendix II, Theorem A2.3, is a very short construction of the point in W that minimizes $E\left((Y - (\alpha + \beta X))^2\right)$, the natural measurement of the distance from Y to W :

$$P(Y) = \mu_Y + \frac{\text{Cov } (X, Y)}{\sigma_X^2}(X - \mu_X),$$

where P is the orthogonal projection onto W. This implies that, if we standardize X and Y,

$$\tilde{X} \equiv \frac{(X - \mu_X)}{\sigma_X}, \quad \tilde{Y} \equiv \frac{(Y - \mu_Y)}{\sigma_Y},$$

then

$$P\left(\tilde{Y}\right) = \rho \tilde{X},$$

where ρ is the correlation of X and Y. If d is the distance from $\tilde{Y}$ to W, then the Pythagorean theorem implies that

$$1 = E\left(\tilde{Y}^2\right) = d^2 + E\left((\rho \tilde{X})^2\right) = d^2 + \rho^2.$$

Thus ρ is measuring how close $\tilde{Y}$ is to W, a geometric picture of the linear (actually, affine) relationship between Y and X.

For fitting a straight line to data $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, let the probability space $\Omega \equiv \{1, 2, ..., n\}$, $P(\{k\}) \equiv \frac{1}{n}, Y(k) \equiv y_k, X(k) \equiv x_k, 1 \leq k \leq n$; then we get the familiar expressions

$$\text{Cov } (X, Y) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y}), \quad \sigma_X^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2, \quad \rho = \frac{\text{Cov } (X, Y)}{\sigma_X \sigma_Y}.$$

The math needed for this sort of analysis existed at least as early as 1907.

Galton made no distinction between data about physical parameters such as height and data about "natural abilities" such as intelligence; his interest was eugenics in the most complete sense. Statistics had come a long way, in a little over a century, from the hesitation of Euler and Bayes about manipulating inaccurate or meaningless data.

IV. ORIGINS OF SUFFICIENCY

This is one of those ideas that, with twenty-twenty hindsight, seems straightforward. Any random variable can be considered data reduction; why not continue reducing the data by applying another measurable function to the original random variable? The constraint here is that we don't want to lose information about an underlying parameter, whatever "losing information" means.

This history is a very short story, because it is predicated on a much longer story, the development of the concept of conditional probability with respect to a random variable or sigma algebra (see Section III, which itself is unrealistically short because of my avoidance of both measure theory and social science).

Sufficiency (the idea, not the name, which appeared a year later) was invented by Fisher in 1920, in the process of rebutting a statement by an astronomer, A. S. Eddington, who asserted the superiority of the sample mean deviation

$$\sigma_1 \equiv \sqrt{\frac{\pi}{2}} \left(\frac{1}{n} \sum_{k=1}^{n} |X_k - \overline{X}| \right)$$

over the sample standard deviation

$$\sigma_2 \equiv \sqrt{\frac{1}{n} \sum_{k=1}^{n} (X_k - \overline{X})^2}.$$

Fisher showed that, for a random sample from a normal distribution, the variance of σ_2 is less than σ_1. He also showed that, for a random sample from a double exponential distribution, Laplace's original distribution of choice, the variance of σ_1 is less than the variance of σ_2. This is a reminder of the Hilbert-space nature of the normal distribution, commented on at the end of Section I.

As with choosing squared error over absolute error, in the seminal studies of the analysis of variance (1.6) discussed in Section I, this expresses a preference for the L^2 norm, coming from a pre-inner product,

$$\left\langle \vec{X}, \vec{Y} \right\rangle \equiv \frac{1}{n} \sum_{k=1}^{n} (X_k - \overline{X})(Y_k - \overline{Y}),$$

over the L^1 norm, which does not. Fisher also showed that, for any natural number p not equal to 2, for a normal distribution the L^p norm estimator

$$\sigma_p \equiv c_p \left(\frac{1}{n} \sum_{k=1}^{n} |X_k - \overline{X}|^p \right)^{\frac{1}{p}},$$

where c_p is chosen to make the estimator unbiased, has larger variance than σ_2.

This much was done by Gauss in 1816; see [H, Chapter 21.1] for details. What is historically significant is that Fisher went on to consider the joint distribution of (σ_1, σ_2), and showed "*For a given value of* σ_2, *the distribution of* σ_1 *is independent of* σ." He also showed this to be true with σ_1 replaced by $\sigma_p, p \in \mathbf{N}$. He concluded "*The whole of the information respecting* σ, *which a sample provides, is summed up in the value of* σ_2." The italics are Fisher's. Even if, as with the current author, one is uncertain what the definition of "information" is, it does seem plausible to assert that a distribution that is independent of σ contains no information about σ.

Note how similar the competition between σ_1 and σ_2 is to the choice of method in Section I for measuring the total error of observation,

$$\sum_{i=1}^{n} |e_i| \quad \text{versus} \quad \sum_{i=1}^{n} e_i^2.$$

My rationale for including this subject, in a history that allegedly stops somewhere in the nineteenth century, is the fact, pointed out by Stigler ([St1]), that Laplace began a very similar investigation in 1818. Laplace was considering the special case of the linear model

$$y_i = x_i \beta + \epsilon_i, \;\; 1 \leq i \leq n,$$

for fixed x_i, and, analogous to Fisher, was comparing two estimators of β, one, call it β_1, that minimizes the L^1 norm

$$\sum_{i=1}^{n} |y_i - x_i\beta|,$$

versus another, call it β_2, that minimizes the L^2 norm

$$\sum_{i=1}^{n} (y_i - x_i\beta)^2.$$

Laplace first compared asymptotic variances of β_1 and β_2, finding necessary and sufficient conditions on the density of the population sampled from for β_1 to be preferable to β_2. Then, like Fisher one hundred years later, Laplace considered the joint distribution of (β_1, β_2). Again in terms of the density, he found the number C that minimizes the asymptotic variance of

$$\beta_2 - (\beta_2 - \beta_1)C.$$

He showed that, for a normal distribution, no such C exists; that is, β_2 is a better estimator than any other unbiased linear combination of β_1 and β_2.

It is not surprising that Laplace didn't anticipate Fisher in his (Fisher's) final step, of looking at the distribution of β_1 given β_2, because conditional probability was far from being understood. Although Fisher preferred to describe sufficiency in the quasi-intuitive language of "information," his earliest descriptions of sufficiency, including the 1920 paper we have quoted from in which he hasn't introduced the word sufficiency, include the modern precise definition in terms of conditional probability; see [St1] and [H, Chapter 28.5], for quotes from Fisher.

APPENDICES

These appendices are meant to show how all the subjects whose early history I've outlined are based on the idea of an inner product, and how their presentation, both definitions and proofs, can be unified, clarified and simplified with this intuitive, geometric concept. Appendix I presents all the results about inner products needed. What I believe are simple proofs are given, with the exception of the spectral theorem, which requires, at least for infinite-dimensional vector spaces, more sophisticated math than might appear at the undergraduate level. The other appendices give proofs of a large portion of a master's degree statistics curriculum that are based on, and simplified by, the results in Appendix I. Appendix II presents basic results in analysis of variance and the general linear model. Appendix III is about conditional expectation and independence. This leads naturally to Appendix IV, on sufficiency, and a very short Appendix V, on variance shrinking.

On a personal note, these appendices also represent some of my attempts to understand, in as clear, simple and unified a manner as possible, results that were presented in classes I took. Although probably not original in the sense of never having been done before, all of the proofs in Appendices II–V, with the exception of the Cramer-Rao theorem, I have derived independently. In the language of my pedagogical youth, I have tried throughout my second graduate career to use the "discovery" method of learning (also known as "reinventing the wheel").

AI. INNER PRODUCT SPACES

I'll present real inner product spaces initially; the best operator theory is done on complex inner product spaces, but the extension is not difficult.

Everything you need to know about inner product spaces can be done in two dimensions. In $\mathbf{R}^2$, define the inner product of two vectors $\vec{x} \equiv (x_1, x_2), \vec{y} \equiv (y_1, y_2)$ by

$$\langle \vec{x}, \vec{y} \rangle \equiv (x_1 y_1 + x_2 y_2).$$

Define also the norm of $\vec{x}$ by

$$\|\vec{x}\| \equiv \sqrt{x_1^2 + x_2^2},$$

the length, via the Pythagorean theorem, of the traditional arrow representing $\vec{x}$, going from the origin $(0, 0)$ to $\vec{x}$. Note that $\langle \vec{x}, \vec{x} \rangle = \|\vec{x}\|^2$.

Some trigonometry shows that

$$\langle \vec{x}, \vec{y} \rangle = \|\vec{x}\| \|\vec{y}\| \cos\theta,$$

where θ is the angle between (the arrows representing) $\vec{x}$ and $\vec{y}$. This is a surprising sort of result, relating geometry (something you can draw) to algebra (something you can calculate). Algebra has the advantage of precision, while geometry has the advantage of providing a picture, more conceptually intuitive and profound than words or numbers.

Everything you might want to know about the pair of vectors $(\vec{x}, \vec{y})$; their lengths and the angle between them; is contained in three inner products, of $\vec{x}$ with $\vec{y}$, of $\vec{x}$ with $\vec{x}$ and of $\vec{y}$ with $\vec{y}$.

Of particular interest is the case $\theta = \frac{\pi}{2}$: it follows that $\vec{x}$ and $\vec{y}$ are orthogonal (perpendicular) if and only if their inner product is zero. Let's write this as $\vec{x} \perp \vec{y}$.

Now consider the following approximation problem: given a line W through the origin and a point $\vec{x}$, probably not in W, find the point in W closest to $\vec{x}$; that is, we want a point $w_0 \in W$ such that

$$\|w - \vec{x}\| \geq \|w_0 - \vec{x}\|, \quad \text{for all } w \in W.$$

It might be believable that we get w_0 by drawing a line from $\vec{x}$ to W that is perpendicular to W; that is, w_0 satisfies

$$(\vec{x} - w_0) \perp w, \quad \text{or,} \quad \langle \vec{x}, w \rangle = \langle w_0, w \rangle \ \ \forall w \in W.$$

Here are some natural extensions of our inner product on $\mathbf{R}^2$. On $\mathbf{R}^n$, define

$$\langle \vec{x}, \vec{y} \rangle \equiv \sum_{k=1}^{n} x_k y_k.$$

Or we could throw on weights: for $w_k \geq 0$ $(k = 1, 2, ..., n)$, define

$$\langle \vec{x}, \vec{y} \rangle \equiv \sum_{k=1}^{n} w_k x_k y_k.$$

Leaping into infinite dimensions, define, for random variables $X, Y \in L^2(\Omega, P)$ (that is, of finite variance),

$$\langle X, Y \rangle_1 \equiv E(XY) = \int_\Omega (XY)(\omega)\, dP(\omega),$$

or

$$\langle X, Y \rangle_2 \equiv \mathrm{Cov}(X, Y) \equiv E\left((X - E(X))(Y - E(Y))\right).$$

As with any abstraction, it is worthwhile to filter out the irrelevant details, and identify the useful idea. Here is all we need for an inner product.

Definitions A1.1. Suppose V is a real vector space. A *pre-inner product* $\langle \cdot \rangle$ on V is a nonnegative, symmetric, bilinear map from $V \times V \to \mathbf{R}$; this means

(1) for all $y \in V$, the map $x \mapsto \langle x, y \rangle : V \to \mathbf{R}$ is linear;

(2) $\langle x, y \rangle = \langle y, x \rangle$, for all $x, y \in V$; and

(3) for all $x \in V, \langle x, x \rangle \geq 0$.

If, in addition to (1)–(3),

(4) $\langle x, x\rangle = 0$ only when x is the zero vector,

then $\langle\cdot\rangle$ is an *inner product* on V.

The pair $(V, \langle\cdot\rangle)$ is then a (*pre-*) *inner product space*.

The *norm* of x is

$$\|x\| \equiv \sqrt{\langle x, x\rangle}.$$

The vectors x and y are said to be *perpendicular*, or *orthogonal*, written $x \perp y$, if

$$\langle x, y\rangle = 0.$$

As a challenge, let's see how much can be done without anything resembling calculus.

Definition A1.2. If W is a subspace of V, and $x \in V$, then the *orthogonal projection* of x onto W, written $P_W(x)$, is the set of vectors $w_0 \in W$ such that

$$(x - w_0) \perp w, \quad \forall w \in W.$$

More generally, if W is an affine space, that is, a translation of a subspace, $w_0 \in P_W(x)$ if $(x - w_0) \perp (w - w_0), \quad \forall w \in W$. For simplicity, I will restrict myself to subspaces W.

Here are some useful facts.

Theorem A1.3. *Suppose $(V, \langle\cdot\rangle)$ is a pre-inner product space and $x, y \in V$.*

(a) *(Cauchy inequality) $|\langle x, y\rangle| \le \|x\|\|y\|$; equality occurs if and only if there exists real α such that $\|x + \alpha y\| = 0$.*

(b) *(Pythagorean theorem) $x \perp y$ if and only if $\|x + y\|^2 = \|x\|^2 + \|y\|^2$.*

(c) *$x \perp y$ if and only if $\|x + \alpha y\| \ge \|x\|$, for all real α.*

(d) *(best approximation) If W is a subspace of V and $P_W(x)$ is nonempty, then*

$$\|x - z\| \le \|x - w\|, \forall w \in W, z \in P_W(x).$$

If w is not in $P_W(x)$, then

$$\|x - z\| < \|x - w\|.$$

(e) *If $z \in P_W(x)$, then $P_W(x) = \{z + y \,|\, \|y\| = 0\}$.*

(f) *(parallelogram law)*

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

(g) *(triangle inequality) $\|x + y\| \le \|x\| + \|y\|$.*

Proof: (a) The "trick" here is to consider the nonnegative quadratic function $t \mapsto \|x + ty\|^2$: for any real t,

$$\|x + ty\|^2 \equiv \langle x + ty, x + ty\rangle = \langle x, x\rangle + 2t\langle x, y\rangle + t^2\langle y, y\rangle = \|y\|^2\left(t^2 + 2t\frac{\langle x, y\rangle}{\|y\|^2}\right) + \|x\|^2$$
$$= \|y\|^2\left(t + \frac{\langle x, y\rangle}{\|y\|^2}\right)^2 + \left(\|x\|^2 - \frac{(\langle x, y\rangle)^2}{\|y\|^2}\right).$$

Setting $t = -\frac{\langle x,y\rangle}{\|y\|^2}$ implies that

$$\left(\|x\|^2 - \frac{(\langle x, y\rangle)^2}{\|y\|^2}\right) = \|x + \left(-\frac{\langle x, y\rangle}{\|y\|^2}\right)y\|^2 \le \|x + ty\|^2, \ \forall t \in \mathbf{R}. \qquad (*)$$

The equality in (*) implies the Cauchy inequality, and the fact that $|\langle x, y\rangle| = \|x\|\|y\|$ implies $\|x + \alpha y\| = 0$ for some real α; the converse follows from the inequality in (*).

(b)
$$\|x+y\|^2 \equiv \langle (x+y),(x+y)\rangle = \langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle = \|x\|^2 + \|y\|^2 + 2\langle x,y\rangle,$$
which equals $\|x\|^2 + \|y\|^2$ if and only if $\langle x,y\rangle = 0$.
(c) From the calculations in the proof of (a),
$$\|x+\alpha y\|^2 = \|y\|^2\left(\alpha + \frac{\langle x,y\rangle}{\|y\|^2}\right)^2 + \left(\|x\|^2 - \frac{(\langle x,y\rangle)^2}{\|y\|^2}\right).$$
If $\langle x,y\rangle = 0$, then
$$\|x+\alpha y\|^2 = \|x\|^2 + \alpha^2\|y\|^2 \geq \|x\|.$$
Otherwise, we have
$$\|x - \frac{\langle x,y\rangle}{\|y\|^2}y\|^2 = \left(\|x\|^2 - \frac{(\langle x,y\rangle)^2}{\|y\|^2}\right) < \|x\|^2,$$
proving the converse.
(d) By the Pythagorean theorem,
$$\|x-w\|^2 = \|x-z\|^2 + \|z-w\|^2 \geq \|x-z\|^2,$$
with equality occurring only if $\|z-w\|^2 = 0$, which implies that, for any $w' \in W$,
$$\langle x-w,w'\rangle = \langle x-z,w'\rangle + \langle z-w,w'\rangle = \langle z-w,w'\rangle \leq \|z-w\|\|w'\| = 0;$$
that is, $w \in P_W(x)$.
(e) The argument at the conclusion of the proof of (d) shows that $P_W(x)$ contains $\{z+y \,|\, \|y\| = 0\}$. Conversely, suppose $w \in P_W(x)$. Then by (d) and the Pythagorean theorem,
$$\|x-w\|^2 = \|x-z\|^2 = \|x-w\|^2 + \|w-z\|^2,$$
so that $w = z + (w-z)$, with $\|(w-z)\| = 0$.
(f) Calculate:
$$\begin{aligned}\|x+y\|^2 + \|x-y\|^2 &= (\langle x+y,x+y\rangle) + (\langle x-y,x-y\rangle) = (\langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle) + (\langle x,x\rangle - 2\langle x,y\rangle + \langle y,y\rangle)\\ &= 2(\langle x,x\rangle + \langle y,y\rangle) = 2\left(\|x\|^2 + \|y\|^2\right).\end{aligned}$$
(g) As before,
$$\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x,y\rangle = (\|x\| + \|y\|)^2 + 2(\langle x,y\rangle - \|x\|\|y\|) \leq (\|x\| + \|y\|)^2,$$
by the Cauchy inequality. □

Remark A1.4. The value of Theorem A1.3(d) cannot be overemphasized. We shall see later that $P_W(x)$ exists, for any x, when W is closed (see Definitions A1.12; this includes W finite-dimensional). Thus we are not only guaranteed a best approximation from W, which is unique when we are in an inner product space, we are given a simple characterization, which leads to simple constructions; see, for example, Theorems A1.6 and A2.1.

Definition A1.5. An *orthogonal set* is a set of vectors any pair of which are orthogonal.

Infinite-dimensional analogues of the following theorem are straightforward with a little topology (mainly infinite sums), but I'll restrict attention to finite dimensions.

Theorem A1.6. *Suppose W is an n-dimensional subspace of the pre-inner product space V and $\{w_1, w_2, ..., w_n\}$ is an orthogonal set of nontrivial vectors. Then for any $x \in V$,*

$$\sum_{k=1}^{n} \frac{\langle x, w_k \rangle}{\|w_k\|^2} w_k \in P_W(x).$$

Proof: Linear algebra shows that $\{w_1, w_2, ..., w_n\}$ is a basis for W. Check the definition of $P_W(x)$: if $w \in W$, then there exist $\alpha_1, ..., \alpha_n$ such that $w = \sum_{j=1}^{n} \alpha_j w_j$, thus, by orthogonality and the definition of norm,

$$\begin{aligned}\left\langle x - \sum_{k=1}^{n} \frac{\langle x, w_k \rangle}{\|w_k\|^2} w_k, w \right\rangle &= \sum_{j=1}^{n} \alpha_j \langle x, w_j \rangle - \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \frac{\langle x, w_k \rangle}{\|w_k\|^2} \langle w_k, w_j \rangle \\ &= \sum_{j=1}^{n} \alpha_j \langle x, w_j \rangle - \sum_{j=1}^{n} \alpha_j \frac{\langle x, w_j \rangle}{\|w_j\|^2} \langle w_j, w_j \rangle = 0.\end{aligned}$$

□

Corollary A1.7. *If W is a finite-dimensional subspace of an inner product space V, then $P_W(x)$ exists for any $x \in V$.*

Proof: Let $\{v_1, v_2, ..., v_n\}$ be a basis for the subspace W. Apply Gram-Schmidt orthogonalization, using Theorem A1.6:

$$w_1 \equiv v_1, \;\; w_k \equiv v_k - P_{\text{span } (w_1, w_2, ..., w_{k-1})}(v_k), 1 < k \leq n.$$

Then $\{w_1, w_2, ..., w_n\}$ is an orthogonal basis for W, and Theorem A1.6 guarantees that $P_W(x)$ exists for any x. □

Again to avoid minor topology, for now I'll just state the uniqueness part of an existence and uniqueness theorem.

Proposition A1.8. *Suppose W is a convex subset of the pre-inner product space $(V, \langle \cdot \rangle)$. If x and y are both elements of W of minimum norm, then $\|x - y\| = 0$.*

Proof: Convexity implies that $\frac{1}{2}(x + y) \in W$. By the parallelogram law,

$$\|x - y\|^2 = 2(\|x\|^2 + \|y\|^2) - \|x + y\|^2 = 2\|x\|^2 + 2\|y\|^2 - 4\|\frac{1}{2}(x + y)\|^2 \leq 2\|x\|^2 + 2\|y\|^2 - 4\|y\|^2 = 0.$$

□

Definition A1.9. If W is a subset of the inner product space V, then the *orthogonal complement* of W, $W^{\perp}$, is the set of all vectors in V that are orthogonal to W, that is,

$$W^{\perp} \equiv \{v \in V \,|\, v \perp w \;\; \forall w \in W\}.$$

For a finite-dimensional affine space, we can characterize its element of minimum norm; see Corollary A.15 for a major extension.

Proposition A1.10. *If $(V, \langle \cdot \rangle)$ is an inner product space, $x \in V$ and W is a finite dimensional subspace of V, then*

$$(x + W) \cap W^{\perp} = \{z\},$$

a set containing a single point, and z is of minimum norm in $(x + W)$.

Proof: Since W is a subspace, $(x + W) = (x - W)$. By Theorem A1.6, $P_W(x)$ exists; by its definition,

$$(x - W) \cap W^{\perp} = \{x - P_W(x)\},$$

which is unique by Theorem A1.3(e). By Theorem A1.3(d),

$$\|(x - P_W(x)\| \leq \|x - w\| \; \forall w \in W.$$

This is saying that $z \equiv (x - P_W(x))$ is the element of minimum norm in $(x - W)$. □

The following proposition says that two vectors in an inner product space are uniquely determined by their inner products.

Proposition A1.11. *If $(V, \langle\cdot\rangle)$ is an inner product space, $x, y \in V$, and $\langle x, v\rangle = \langle y, v\rangle$ for all $v \in V$, then $x = y$.*

Proof: $\|x - y\|^2 = \langle x - y, x - y\rangle = \langle x, x - y\rangle - \langle y, x - y\rangle = 0$, by hypothesis. $\square$

To get some operator theory, and a complete description of when orthogonal projections exist, we need some mathematical analysis; that is, a notion of convergence.

Definitions A1.12. The inner product space $(H, \langle\cdot\rangle)$ is a *Hilbert space* if it is complete w.r.t. its norm; that is, whenever $\{x_k\}_{k=1}^{\infty}$ is a sequence in H that is *Cauchy*, meaning

$$\lim_{n,m\to\infty} \|x_m - x_n\| = 0,$$

then there exists $x \in H$ such that $\{x_k\}_{k=1}^{\infty}$ converges to x, meaning

$$\lim_{n\to\infty} \|x_n - x\| = 0.$$

A subset, W, of the Hilbert space H, is *closed* if the limit of every convergent sequence from W is in W.

Here's the existence part corresponding to the uniqueness result in Proposition A1.8.

Theorem A1.13. *Suppose W is a closed, convex subset of the Hilbert space $(H, \langle\cdot\rangle)$. Then there exists a point in W of minimum norm; that is, $x \in W$ such that*

$$\|x\| \le \|w\| \ \ \forall w \in W.$$

Proof: Since the norm is a continuous map from H into $[0, \infty)$, there exists $d \ge 0$ such that

$$d = \ \min\{\|w\| \mid w \in W\}.$$

Let $\{x_k\}_{k=1}^{\infty}$ be a sequence from W such that

$$\|x_k\| \to \|d\| \quad \text{as} \quad k \to \infty.$$

The parallelogram law, applied as with uniqueness, will show that $\{x_k\}_{k=1}^{\infty}$ is Cauchy:

$$\|x_m - x_n\|^2 = 2(\|x_m\|^2 + \|x_n\|^2) - \|x_m + x_n\|^2 = 2\|x_m\|^2 + 2\|x_n\|^2 - 4\|\frac{1}{2}(x_m + x_n)\|^2 \le 2\|x_m\|^2 + 2\|x_n\|^2 - 4d^2,$$

which converges to 0 as $n, m \to \infty$. Since W is closed, there exists $x \in W$ such that $x_k \to x$ as $k \to \infty$, so that $\|x\| = \lim_{k\to\infty} \|x_k\| = d$, as desired. $\square$

Now it is easy to characterize subspaces W for which $P_W(x)$ exists, for all x.

Theorem A1.14. *Suppose W is a subspace of the Hilbert space H. Then the following are equivalent.*

(a) *$P_W(x)$ exists, for all $x \in H$.*

(b) *For all $x \in H$ there exists $w_0 \in W$ such that*

$$\|x - w_0\| \le \|x - w\| \ \ \forall w \in W.$$

(c) *W is closed.*

Proof: (a) $\to$ (b) is Theorem A1.3(d). (c) $\to$ (b) is Theorem A1.13 applied to $x - W$.
(b) $\to$ (c). If W is not closed, then there exists $x \in H$ that is in the closure of W but is not in W. This means that, for any $w_1 \in W$, there exists $w_2 \in W$ such that $\|x - w_2\| < \|x - w_1\|$, but there exists no $w \in W$ such that $\|x - w\| = 0$. Thus w_0 as in (b) is impossible.
(b) $\to$ (a). For any real α, $w \in W$,

$$\|(x - w_0) + \alpha w\| = \|x - (w_0 - \alpha w)\| \ge \|(x - w_0)\|,$$

since $(w_0 - \alpha w) \in W$. Theorem A1.3(c) implies that $(x - w_0) \perp w$. $\square$

The same proof now extends Proposition A1.10.

Corollary A1.15. *Proposition A1.10 holds with "finite-dimensional" replaced by "closed."*

Theorem A1.14 also provides a novel way of showing that finite-dimensional subspaces are closed.

Corollary A1.16. *Any finite-dimensional subspace of a Hilbert space is closed.*

Proof: Corollary A1.7 and Theorem A1.14. □

Definition A1.17. Suppose $(V, \langle\cdot\rangle_V)$ and $(W, \langle\cdot\rangle_W)$ are inner product spaces and $T : V \to W$ is linear. Then the *null space* of T, $\mathcal{N}(T)$, is $\{x \in V \,|\, Tx = 0\}$, and the *range space* of T, $\mathcal{R}(T)$, is $\{Tx \in W \,|\, x \in V\}$.

The following could be considered an existence analogue of Proposition A1.11.

Proposition A1.18 (Riesz's lemma). *Suppose $(H, \langle\cdot\rangle)$ is a Hilbert space and $\phi : H \to \mathbf{R}$ is linear and continuous w.r.t $\|\cdot\|$. Then there exists a unique $y \in H$ such that*

$$\phi(x) = \langle y, x\rangle$$

for all $x \in H$.

Proof: Let $W \equiv \mathcal{N}(\phi)$. W is closed: if $w_n \to x \in H$, then $\phi(w_n) = 0$, for all n, hence by continuity of ϕ, $\phi(x) = \lim_{n\to\infty} \phi(w_n) = 0$, so that $x \in W$.

If ϕ is trivial, let $y \equiv \vec{0}$. Otherwise, there exists $z \in H$ such that $\phi(z)$ is not zero. Let

$$y \equiv \frac{\phi(z - P_W(z))}{\|z - P_W(z)\|^2}(z - P_W(z)).$$

Then $y \in W^\perp$ and $\phi(y) = \|y\|^2$.

Note that $W^\perp$ is one-dimensional: if $x_1, x_2 \in W^\perp$, then, since the range of ϕ is one-dimensional, there exists real $\alpha_j, j = 1, 2$, such that

$$0 = \alpha_1\phi(x_1) + \alpha_2\phi(x_2) = \phi(\alpha_1 x_1 + \alpha_2 x_2);$$

that is, $(\alpha_1 x_1 + \alpha_2 x_2) \in W$, so that

$$\alpha_1 x_1 + \alpha_2 x_2 = P_W(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 P_W(x_1) + \alpha_2 P_W(x_2) = 0.$$

Thus, for any $x \in H$, since y and $(x - P_W(x)) \in W^\perp$ and $P_W(x) \in W$, Theorem A1.6 implies that

$$\phi(x) = \phi(x - P_W(x)) + \phi(P_W(x)) = \phi(\frac{\langle y, (x - P_W(x))\rangle}{\|y\|^2} y) = \frac{\langle y, (x - P_W(x))\rangle}{\|y\|^2}\phi(y) = \frac{\langle y, x\rangle}{\|y\|^2}\|y\|^2 = \langle y, x\rangle .$$

Uniqueness follows from Proposition A1.11. □

This enables us to define the *adjoint* of an operator.

Definition A1.19. Suppose $(H, \langle\cdot\rangle_H)$ and $(K, \langle\cdot\rangle_K)$ are Hilbert spaces and $T : H \to K$ is linear and continuous w.r.t. $\|\cdot\|$.

The *adjoint*, $T^* : K \to H$ is defined by

$$\langle T^* y, x\rangle_H \equiv \langle y, Tx\rangle_K , \ \forall x \in H, y \in K.$$

Proposition A1.18 applied to $\phi_y(x) \equiv \langle y, Tx\rangle_K$ implies that T^* is well defined and linear. Note that $T^{**} \equiv (T^*)^* = T$.

An $n \times m$ matrix A represents an operator, call it T_A, from $\mathbf{R}^m$ to $\mathbf{R}^n$ via matrix multiplication

$$T_A\vec{x} \equiv A\vec{x}.$$

A calculation shows that $(T_A)^* = T_{A^T}$, where A^T means the transpose of A.

Proposition A1.20. *If T is as in Definition A1.19, then*

$$\mathcal{N}(T) = [\mathcal{R}(T^*)]^\perp = \mathcal{N}(T^*T).$$

Proof: By Proposition A1.11,

$$x \in \mathcal{N}(T) \iff \langle y, Tx\rangle_K = 0 \ \forall y \in K \iff \langle T^*y, x\rangle_H = 0 \ \forall y \in K \iff x \in (\mathcal{R}(T^*))^\perp .$$

It is clear that $\mathcal{N}(T) \subseteq \mathcal{N}(T^*T)$. Conversely, suppose $x \in \mathcal{N}(T^*T)$. Then $Tx \in \mathcal{N}(T^*) = [\mathcal{R}(T)]^\perp$. Thus

$$Tx \in [\mathcal{R}(T)] \cap [\mathcal{R}(T)]^\perp .$$

For *any* subspace W, $W \cap W^\perp$ is trivial. Thus $Tx = \vec{0}$, or $x \in \mathcal{N}(T)$, as desired. □

Definition A1.21. If V is an inner product space, $P : V \to V$ and there exists a subspace W such that $P(x) = P_W(x)$, for all $x \in V$, then P is an *orthogonal projection.*.

It is useful to have algebraic characterizations of orthogonal projections.

Proposition A1.22. *Suppose P is a linear, continuous map from a Hilbert space $(H, \langle\cdot\rangle)$ to itself.*

(1) *P is an orthogonal projection if and only if $P^2 = P = P^*$.*

(2) *If P is an orthogonal projection, then $x \in \mathcal{R}(P)$ if and only if $Px = x$.*

(3) *If P_1 and P_2 are orthogonal projections then $\mathcal{R}(P_2) \subseteq \mathcal{R}(P_1)$ if and only if $P_1P_2 = P_2$.*

Proof: (2) Suppose W is a subspace and $P \equiv P_W$. Since $P_W w = w$ for all $w \in W$, $W \subseteq \mathcal{R}(P)$. Clearly $P_W x \in W$ for all $x \in H$. Thus $W = \mathcal{R}(P)$.

If $x \in \mathcal{R}(P)$, then $x \in W$, thus $Px = P_W x = x$. Conversely, if x is not in W, then since $Px \in W$, Px cannot equal x.

(1) If $P = P_W$, for W a subspace, then by (2), for any $x \in H$, $P^2x = P(Px) = Px$, since $Px \in W = \mathcal{R}(P)$. Thus $P^2 = P$. For any $x, y \in H$,

$$\begin{aligned}\langle Px, y\rangle &= \langle Px - x, y\rangle + \langle x, y\rangle = (\langle Px - x, Py\rangle + \langle Px - x, y - Py\rangle) + (\langle x, Py\rangle + \langle x, y - Py\rangle) \\ &= (\langle Px - x, y - Py\rangle) + (\langle x, Py\rangle + \langle x, y - Py\rangle) = \langle Px, y - Py\rangle + \langle x, Py\rangle = \langle x, Py\rangle ;\end{aligned}$$

that is, $P = P^*$.

Conversely, suppose $P : H \to H$ satisfies $P^2 = P = P^*$. Let $W \equiv \mathcal{R}(P)$. To see that $P = P_W$, fix $x \in H$. Note first that, by definition of $\mathcal{R}$, $Px \in W$. For any $y \in H$,

$$\langle x - Px, Py\rangle = \langle P(x - Px), y\rangle = \left\langle Px - P^2x, y\right\rangle = \langle Px - Px, y\rangle = \left\langle \vec{0}, y\right\rangle = 0;$$

that is, $(x - Px) \perp Py$, for all $y \in H$. Since $W = \{Py \,|\, y \in H\}$, this shows that $Px = P_W(x)$, as desired.

(3) Suppose $\mathcal{R}(P_2) \subseteq \mathcal{R}(P_1)$. If $x \in H$, then $P_2x \in \mathcal{R}(P_2)$, hence $P_2x \in \mathcal{R}(P_1)$, so by (2) $P_1P_2x = P_1(P_2x) = P_2x$. Conversely, if $P_1P_2 = P_2$, suppose $x \in \mathcal{R}(P_2)$. By (2), $P_2x = x$, thus $P_1x = P_1P_2x = P_2x = x$, thus, again by (2), $x \in \mathcal{R}(P_1)$. This is saying that $\mathcal{R}(P_2) \subseteq \mathcal{R}(P_1)$. □

Definitions A1.23. Let T be a linear, continuous map from a Hilbert space to itself.

(1) T is *symmetric* if $T = T^*$.

(2) T is *unitary* if T is invertible and $T^{-1} = T^*$.

In particular, note that an orthogonal projection is symmetric, by Proposition A1.22(1). Note also that, by the comments after Definition A1.19, a matrix is symmetric (that is, equals its transpose) if and only if the corresponding operator (defined by matrix multiplication) is symmetric.

Being unitary is equivalent to preserving the inner product: T is unitary if and only if it is surjective and

$$\langle T\vec{x}, T\vec{y}\rangle = \langle \vec{x}, \vec{y}\rangle ,$$

for all $\vec{x}, \vec{y}$ in the Hilbert space.

Here is arguably the most famous result in operator theory, but stated only in finite dimensions.

Theorem A1.24 (Spectral Theorem). *If A is a symmetric matrix, then there exists unitary U and diagonal D such that*

$$A = UDU^{-1}.$$

AII. GENERAL LINEAR MODEL AND ANALYSIS OF VARIANCE

Throughout, $\vec{y} \in \mathbf{R}^n, \vec{Y} = (Y_1, Y_2, ..., Y_n)$ is a random vector. Unless stated otherwise, in this section $\langle \cdot \rangle$ will be the usual inner product on $\mathbf{R}^n$,

$$\langle \vec{x}, \vec{y} \rangle \equiv \sum_{j=1}^{n} x_j y_j.$$

Also denote by A^T the transpose of a matrix A. A calculation shows that $A^T = A^*$, from Definition A1.19, where we still denote by A the operator that maps $\vec{x} \to A\vec{x}$.

The extension to *weighted* least-squares, etc., should be clear by using

$$\langle \vec{x}, \vec{y} \rangle \equiv \sum_{j=1}^{n} \omega_j x_j y_j,$$

for weights $\omega_j, j = 1, 2, ..., n$.

The linear system

$$\vec{y} = X\vec{\beta} \quad (\vec{y} \in \mathbf{R}^n, \vec{\beta} \in \mathbf{R}^m, X \text{ an } n \times m \text{ matrix }),$$

that one obtains from measurements, will probably be inconsistent. Consistency, when it exists, is an unstable equilibrium; for example, consider the system

$$1 = \beta, \ 1 + \epsilon = \beta$$

($\vec{y} = \begin{bmatrix} 1 \\ 1+\epsilon \end{bmatrix}, X = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \vec{\beta} = \beta$). For any ϵ not equal to zero this is inconsistent, yet it is consistent when $\epsilon = 0$.

We need a more constructive response than the assertion "there is no solution." What might seem natural, in this imperfect world, is to be satisfied with finding $\vec{\beta}$ that minimizes the distance between $\vec{y}$ and $X\vec{\beta}$, or the norm of $(\vec{y} - X\vec{\beta})$. Minimization becomes possible, and not too hard, with a norm that comes from an inner product.

The following is a corollary of Proposition A1.20.

Theorem A2.1 (Normal Equations). *Suppose X is an $n \times m$ matrix, $\vec{y} \in \mathbf{R}^n$ and $\vec{\beta} \in \mathbf{R}^m$. Then $\vec{\beta}_0$ minimizes*

$$\|\vec{y} - X\vec{\beta}\|$$

if and only if $\vec{\beta}_0$ is a solution of

$$X^T\vec{y} = X^T X\vec{\beta}.$$

Proof: By Theorem A1.3(d), $\vec{\beta}_0$ minimizes $\|\vec{y} - X\vec{\beta}\|$ if and only if $X\vec{\beta}_0 = P_W(\vec{y})$, where $W = \mathcal{R}(X)$, and we denote by $P_W(\vec{y})$ the unique vector in the set $P_W(\vec{y})$. By definition of P_W and Proposition A1.20, this is equivalent to

$$(\vec{y} - X\vec{\beta}_0) \in [\mathcal{R}(X)]^{\perp} = \mathcal{N}(X^*) = \mathcal{N}(X^T),$$

which is the same as saying that

$$X^T(\vec{y} - X\vec{\beta}_0) = 0.$$

□

Remarks A2.2. In particular, if $X^T X$ is invertible (this is equivalent to X having rank m), then the unique minimizer of $\|\vec{y} - X\beta\|$; that is, the best least-squares solution of

$$\vec{y} = X\vec{\beta},$$

is given by

$$\hat{\beta} = (X^T X)^{-1} X^T \vec{y}.$$

Note that $X\hat{\beta} = X(X^T X)^{-1} X\vec{y}$ is the orthogonal projection of $\vec{y}$ onto $\mathcal{R}(X)$.

For minimizing $\sum_{j=1}^{n}(y_j - (\alpha + \beta x_j))^2$ (best least-squares straight line fitting some data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$), we are minimizing

$$\|\vec{y} - X \begin{bmatrix} \alpha \\ \beta \end{bmatrix}\|, \quad \text{where} \quad X \equiv \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix}.$$

Since $X^T X = \begin{bmatrix} n & \sum_j x_j \\ \sum_j x_j & \sum_j x_j^2 \end{bmatrix}$ and $X^T \vec{y} = \begin{bmatrix} \sum_j y_j \\ \sum_j x_j y_j \end{bmatrix}$, the normal equations become

$$\overline{y} = \alpha + \beta \overline{x}, \quad \overline{(xy)} = \alpha \overline{x} + \beta \overline{(x^2)},$$

which is quickly solved as

$$\beta = \frac{\text{Cov } (\vec{x}, \vec{y})}{\text{Var } (\vec{x})}, \alpha = \overline{y} - \beta \overline{x}.$$

But this special case can be derived almost instantly by using Theorem A1.6 instead of Proposition A1.20. More generally, we can quickly show

Theorem A2.3. *(best affine approximation) Suppose X and Y are random variables of finite variance. Then the affine function of X, $(\alpha + \beta X)$, that minimizes*

$$E((Y - (\alpha + \beta X))^2)$$

is given by $\beta = \frac{\text{Cov } (X,Y)}{\text{Var } (X)}, \alpha = E(Y) - \beta E(X)$.

Proof: With respect to the inner product

$$\langle Z, W \rangle \equiv E(ZW),$$

$\{1, X - E(X)\}$ is an orthogonal subset of the 2 dimensional subspace $W \equiv \{\alpha + \beta X \mid \alpha, \beta \in \mathbf{R}\}$ of whatever L^2 space we construct X and Y to act on. By Theorem A1.3(d) and A1.6, our desired affine approximation is given by

$$\frac{\langle Y, 1 \rangle}{\|1\|^2} 1 + \frac{\langle Y, X - E(X) \rangle}{\|X - E(X)\|^2}(X - E(X)) = E(Y) + \frac{\text{Cov } (X, Y)}{\text{Var } (X)}(X - E(X)).$$

□

For the special case in Remarks A2.2 (best least-squares straight line fit), choose the probability space $\Omega \equiv \{1, 2, ..., n\}$, with measure $P(\{k\}) \equiv \frac{1}{n}, Y(k) \equiv y_k, X(k) \equiv x_k, k = 1, 2, ..., n$.

Consider, for X a fixed $n \times m$ matrix,

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon}, \tag{A2.4}$$

where $\text{Cov}(\vec{\epsilon}) = \sigma^2 I_n$ and $\vec{\beta} \in \mathbf{R}^m$ is the parameter(s) to be estimated. When X has rank m, which is equivalent to (XX^T) being invertible, Theorem A2.1 implies that

$$\hat{\beta} \equiv (X^T X)^{-1} X^T \vec{Y}$$

minimizes $\|\vec{\epsilon}\|$.

The Gauss-Markov theorem gives another sense in which $\hat{\beta}$ is "best." "Linear estimator" will mean a linear combination of $\{Y_1, Y_2, ..., Y_n\}$. In the following, it is not hard to modify the proof for X not full rank, to show that, if $c^T \vec{\beta}$ is estimable, then $c^T \hat{\beta}$ is the unbiased linear estimator of minimum variance.

Theorem A2.5 (Gauss-Markov). *Let $X, \vec{\beta}, \hat{\beta}$ be as in (A2.4).*

(1) *There exists an unbiased linear estimator of β_j, for all $1 \leq j \leq n$, if and only if the rank of X is m if and only if $(X^T X)$ is invertible.*

(2) *Then, for $1 \leq j \leq m$, $\hat{\beta}_j$ is the unbiased linear estimator of β_j of minimum variance.*

Proof: (1) If $(X^T X)$ is invertible, then

$$E(\hat{\beta}) = (X^T X)^{-1} X^T E(\vec{Y}) = (X^T X)^{-1} X^T X \vec{\beta} = \vec{\beta},$$

thus $\hat{\beta}_j$ is an unbiased estimator of β_j, for all j.

Conversely, for $1 \leq j \leq m$, suppose

$$W_j \equiv \sum_{i=1}^{n} d_{ij} Y_i \equiv \left\langle \vec{d}_j, \vec{Y} \right\rangle$$

is an unbiased estimator of β_j. Writing $\vec{e}_j$ for the column vector in $\mathbf{R}^m$ that is one in the j^{th} component, zero elsewhere,

$$\left\langle \vec{e}_j, \vec{\beta} \right\rangle = \beta_j = E(W_j) = \left\langle \vec{d}_j, E(\vec{Y}) \right\rangle = \left\langle \vec{d}_j, X\vec{\beta} \right\rangle = \left\langle X^T \vec{d}_j, \vec{\beta} \right\rangle,$$

for any $\vec{\beta} \in \mathbf{R}^m$. By Proposition A1.11, $\vec{e}_j = X^T \vec{d}_j$, for $1 \leq j \leq m$. This implies that $\mathcal{R}(X^T) = \mathbf{R}^m$, which is equivalent to the rank of X^T, which equals the rank of X, being m.

The equivalence with $(X^T X)$ being invertible follows from Proposition A1.20:

$$\mathcal{R}(X^T) = \mathbf{R}^m \iff \mathcal{N}(X) = \{\vec{0}\} \iff \mathcal{N}(X^T X) = \{\vec{0}\}.$$

(2) Fix j between 1 and m. Define

$$\mathcal{U}_j \equiv \{\vec{d} \in \mathbf{R}^n \mid X^T \vec{d} = \vec{e}_j\}.$$

The argument in (1) (and its converse) shows that $W = \left\langle \vec{d}, \vec{Y} \right\rangle$ is an unbiased estimator of β_j if and only if $\vec{d} \in \mathcal{U}_j$. For such W,

$$\text{Var } (W) = \text{ Cov } (\vec{d}_j^T \vec{Y}) = \vec{d}_j^T \text{ Cov } (\vec{Y}) \vec{d}_j = \vec{d}_j^T \sigma^2 I_n \vec{d}_j = \sigma^2 \|\vec{d}_j\|^2.$$

Thus we want $\vec{d}_j \in \mathcal{U}_j$ of minimum norm. Since $\mathcal{U}_j$ is a translate of $\mathcal{N}(X^T)$, by Proposition A1.10 the element in $\mathcal{U}_j$ of minimum norm is the unique point in

$$\mathcal{U}_j \bigcap \left[\mathcal{N}(X^T)\right]^{\perp} = \mathcal{U}_j \bigcap \mathcal{R}(X),$$

by Proposition A1.20. That is, this desired point

$$\vec{d}_j = X\vec{z}_j, \quad \text{for some } z_j \in \mathbf{R}^n,$$

so that

$$(X^T X)\vec{z}_j = X^T \vec{d}_j = \vec{e}_j \to \vec{z}_j = (X^T X)^{-1} \vec{e}_j \to \vec{d}_j = X(X^T X)^{-1} \vec{e}_j,$$

and

$$W_j \equiv \left\langle \vec{d}_j, \vec{Y} \right\rangle = \left\langle X(X^T X)^{-1} \vec{e}_j, \vec{Y} \right\rangle = \left\langle \vec{e}_j, (X^T X)^{-1} X^T \vec{Y} \right\rangle = \left\langle \vec{e}_j, \hat{\beta} \right\rangle = \hat{\beta}_j,$$

as desired. $\square$

Lemma A2.6. *Suppose A and B are $n \times n$ matrices and $AB^T = 0$. Then*

(a) $A^T\vec{y} \perp B^T\vec{z}, \forall \vec{y}, \vec{z} \in \mathbf{R}^n$.

(b) $Cov(A\vec{Y}, B\vec{Y}) = 0$, *if* $Cov(\vec{Y}) = cI_n$.

Proof: (a) $\left\langle A^T\vec{y}, B^T\vec{z}\right\rangle = \left\langle \vec{y}, AB^T\vec{z}\right\rangle = 0$.
(b) $\text{Cov}(A\vec{Y}, B\vec{Y}) = A \text{ Cov } (\vec{Y})B^T = A(cI_n)B^T = cAB^T = 0$. $\square$

In the following, note that, by Proposition A1.22, the hypotheses on P_j are equivalent to $P_j^2 = P_j = P_j^T$, for $j = 1, 2$, and $P_1P_2 = P_2$.

Corollary A2.7. *Suppose P_1, P_2 are orthogonal projections and $\mathcal{R}(P_2) \subseteq \mathcal{R}(P_1)$. Then*

(a) *for all $\vec{y} \in \mathbf{R}^n, (\vec{y} - P_1\vec{y}) \perp (P_1\vec{y} - P_2\vec{y}), P_2\vec{y} \perp (P_1\vec{y} - P_2\vec{y})$, and $(\vec{y} - P_1\vec{y}) \perp P_1\vec{y}$;*

(b) $Cov((\vec{Y} - P_1\vec{Y}), (P_1\vec{Y} - P_2\vec{Y})) = 0 = \ Cov\ (P_2\vec{Y}, (P_1\vec{Y} - P_2\vec{Y})) = \ Cov\ ((\vec{Y} - P_1\vec{Y}, P_1\vec{Y})$, *if* $Cov(\vec{Y}) = cI_n$; *and*

(c) $(\vec{Y} - P_1\vec{Y})$ *and* $(P_1\vec{Y} - P_2\vec{Y})$, $P_2\vec{Y}$ *and* $(P_1\vec{Y} - P_2\vec{Y})$, *and* $(\vec{Y} - P_1\vec{Y})$ *and* $P_1\vec{Y}$ *are independent if* $\vec{Y} \sim N(\vec{\mu}, \sigma^2 I_n)$.

Proof: (a) and (b) are immediate applications of Lemma A2.6, then (c) follows from the fact that covariance implies independence, for members of a multivariate normal family. I'll do the calculation for

$$A^T = A \equiv (I_n - P_1), B^T = B \equiv (P_1 - P_2),$$

and leave it to the reader (if he/she exists) to do the other two calculations.

$$(I_n - P_1)(P_1 - P_2)^T = (I_n - P_1)(P_1 - P_2) = (P_1 - P_2) - (P_1^2 - P_1P_2) = (P_1 - P_2) - (P_1 - P_2) = 0.$$

$\square$

Here is an application of the spectral theorem.

Proposition A2.8. *Suppose W is a subspace of $\mathbf{R}^n, P : \mathbf{R}^n \to W$ is the orthogonal projection onto W, $P(\vec{\mu}) = \vec{0}$ and $\vec{Y} \sim N(\vec{\mu}, \sigma^2 I_n)$. Then*

$$\frac{1}{\sigma^2}\|P\vec{Y}\|^2 \sim \chi_k^2,$$

where $k \equiv$ dim (W).

Proof: By the spectral theorem, there exists unitary $n \times n$ U such that

$$UPU^{-1} = D \equiv \begin{bmatrix} I_k & 0_{n-k} \\ 0_k & 0_{n-k} \end{bmatrix}.$$

Cov $(U(\vec{Y} - \vec{\mu})) = U$ Cov $(\vec{Y})U^T = U\sigma^2 I_n U^T = \sigma^2 UU^T = \sigma^2 I_n$, since U is unitary. Again using the fact that U is unitary, since $P(\vec{\mu}) = \vec{0}$,

$$\|P\vec{Y}\|^2 = \|P(\vec{Y} - \vec{\mu})\|^2 = \|UP(\vec{Y} - \vec{\mu})\|^2 = \|UPU^{-1}U(\vec{Y} - \vec{\mu})\|^2$$

$$= \|DU(\vec{Y} - \vec{\mu})\|^2 = \sum_{j=1}^{k} \left[U(\vec{Y} - \vec{\mu})\right]_j^2 \sim \sigma^2\chi_k^2,$$

since $\text{Cov}(U(\vec{Y} - \vec{\mu})) = \sigma^2 I_n$ and $\vec{Y}$ is normal. $\square$

Remark A2.9. Without normality, the same proof up to the mention of χ^2 shows that

$$E\left(\frac{1}{k}\|P\vec{Y}\|^2\right) = \sigma^2;$$

that is, $\frac{1}{k}\|P\vec{Y}\|^2$ is an unbiased estimator of σ^2.

ANOVA TABLES and other giddy delights. The pedagogical fashion seems to be to present these as a table of mind-numbing, random, alphabet soups–a different one for every new model. Yet there are really only simple and intuitive consequences of orthogonal projections going on here. They always have the following form.

Theorem A2.10. *Suppose P_1 and P_2 are orthogonal projections on $\mathbf{R}^n$, with $\vec{\mu} \in \mathcal{R}(P_2) \subseteq \mathcal{R}(P_1)$, and $\vec{Y} \sim N(\vec{\mu}, \sigma^2 I_n)$. Then*

(1)
$$\|\vec{Y} - P_2\vec{Y}\|^2 = \|\vec{Y} - P_1\vec{Y}\|^2 + \|P_1\vec{Y} - P_2\vec{Y}\|^2,$$

(2) *$(\vec{Y} - P_1\vec{Y})$ and $(P_1\vec{Y} - P_2\vec{Y})$, and $P_1\vec{Y}$ and $(\vec{Y} - P_1\vec{Y})$ are independent; and*

(3)
$$\frac{1}{\sigma^2}\|\vec{Y} - P_1\vec{Y}\|^2 \sim \chi^2_{k_1}, \quad \frac{1}{\sigma^2}\|P_1\vec{Y} - P_2\vec{Y}\|^2 \sim \chi^2_{k_2}, \quad \text{and} \quad \frac{1}{\sigma^2}\|\vec{Y} - P_2\vec{Y}\|^2 \sim \chi^2_{k_1+k_2},$$
where $k_1 \equiv n -$ rank $(P_1), k_2 \equiv$ rank $(P_1) -$ rank (P_2).

Proof: (1) is the Pythagorean theorem and Corollary A2.7(a). (2) is Corollary A2.7(c). For (3), we may apply Proposition A2.8, since by Proposition A1.22(2) $P_j(\vec{\mu}) = \vec{\mu}$, for $j = 1, 2$, thus
$$(I_n - P_1)(\vec{\mu}) = \vec{0} = (P_1 - P_2)(\vec{\mu}) = (I_n - P_2)(\vec{\mu}),$$
to conclude that $\|\vec{Y} - P_1\vec{Y}\|^2 \sim \sigma^2\chi^2_{\text{rank }(I-P_1)}$, etc. The degrees follow from the orthogonality of Corollary A2.7(a), which implies that
$$\text{rank }(P_1) = \text{ rank }(P_2) + \text{ rank }(P_2 - P_1),$$
and
$$n = \text{ rank }(I_n) = \text{ rank }(P_1) + \text{ rank }(I_n - P_1).$$
□

There is a natural one-to-one correspondence between "models" and subspaces (more generally, affine subsets) of the data space $\mathbf{R}^n$, or, equivalently, orthogonal projections on $\mathbf{R}^n$. A subspace of $\mathbf{R}^n$ corresponds to where the data $\vec{Y}$ "should" be, according to the model.

For everyone's favorite example, fitting $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ to a straight line, the subspace that $\vec{Y}$ is being projected onto is the span of $\{(1, 1, ..., 1), \vec{x}\}$.

In general, P_1 is the "full model" and P_2 is the "reduced model," sometimes the null hypothesis of a hypothesis test. I'll illustrate this with two special cases, then mention a popular general case. Throughout, for simplicity, I will assume X is full rank.

CASE 1: general linear model. This is (A2.4), with $\vec{\epsilon} \sim N(\vec{0}, \sigma^2 I_n)$.

We already know that $\hat{\beta}$ is the best least-squares solution of $\vec{Y} = X\vec{\beta}$, and satisfies the minimum variance condition of the Gauss-Markov theorem. Now a short calculation, which I skip because it has nothing to do with inner products, shows that $\hat{\beta}$ is the MLE of $\vec{\beta}$, and the MLE of σ^2 is $\frac{1}{n}\|\vec{Y} - X\hat{\beta}\|^2$.

For this particular "Anova table," I need to assume that the rank of X is m and the first column of X is $\vec{1}$. Note that this includes polynomial regression.

Apply Theorem A2.10 with
$$P_1 \equiv X(X^TX)^{-1}X^T (\text{so that } P_1\vec{Y} = X\hat{\beta}), \quad P_2 \equiv \frac{1}{n}J_n,$$
where J_n is the $n \times n$ matrix consisting entirely of ones, to get the following ingredients for confidence intervals, t tests and F tests.

Corollary A2.11.

(1) *$\hat{\beta}$ and $\|\vec{Y} - X\hat{\beta}\|^2$ are independent.*

(2) *$\hat{\beta} \sim N(\vec{\beta}, (X^TX)^{-1}\sigma^2)$.*

(3) *$\frac{1}{\sigma^2}\|\vec{Y} - X\hat{\beta}\|^2 \sim \chi^2_{n-k}$, $k \equiv$ rank (X).*

(4) *If $\vec{\beta}_j = 0, j = 2, 3, ..., m$, then $\|\vec{Y} - X\hat{\beta}\|^2$ and $\|X\hat{\beta} - \overline{Y}\|^2$ are independent,*
$$\sum_{j=1}^n (Y_j - \overline{Y})^2 = \|\vec{Y} - \overline{Y}\|^2 = \|\vec{Y} - X\hat{\beta}\|^2 + \|X\hat{\beta} - \overline{Y}\|^2 = \sum_{j=1}^n (Y_j - (X\hat{\beta})_j)^2 + \sum_{j=1}^n ((X\hat{\beta})_j - \overline{Y})^2,$$
with $\frac{1}{\sigma^2}\|X\vec{\beta} - \overline{Y}\|^2 \sim \chi^2_{k-1}$, $\frac{1}{\sigma^2}\|\vec{Y} - \overline{Y}\|^2 \sim \chi^2_{n-1}$.

Proof: (1) follows from Lemma A2.6(b) and the fact that $\hat{\beta}$ and $(\vec{Y} - X\hat{\beta})$ are members of a multivariate normal family, since

$$\begin{aligned}(X^TX)^{-1}X^T\left(1 - X(X^TX)^{-1}X^T\right)^T &= (X^TX)^{-1}X^T(1 - X(X^TX)^{-1}X^T) \\ &= (X^TX)^{-1}X^T - (X^TX)^{-1}X^TX(X^TX)^{-1}X^T) \\ &= (X^TX)^{-1}X^T - (X^TX)^{-1}X^T = 0.\end{aligned}$$

Note that $\vec{Y} \sim N(X\vec{\beta}, \sigma^2 I_n)$. Thus (2) follows from $\hat{\beta} = (X^TX)^{-1}X^T\vec{Y}$. Assertions (3) and (4) follow from Theorem A2.10 and Proposition A1.22, since

$$P_1(X\vec{\beta}) = X(X^TX)^{-1}X^TX\vec{\beta} = X\vec{\beta},$$

and, under the hypotheses of (4),

$$P_2(X\vec{\beta}) = \beta_1 P_2(\text{ first column of } X\ ) = \beta_1\vec{1} = X\vec{\beta}.$$

□

In simple linear regression, the general linear model with

$$X = \begin{bmatrix} 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 \\ x_1 & x_2 & x_3 & \cdot & \cdot & \cdot & x_n \end{bmatrix}^T,$$

we have

$$(P_1\vec{Y})_k = (X\vec{\beta})_k = \overline{Y} + (x_k - \overline{x})\frac{\text{Cov }(\vec{x}, \vec{y})}{\text{Var }(\vec{x})}, \quad 1 \le k \le n,$$

the orthogonal projection onto the span of $\{(1, 1, ..., 1), \vec{x}\}$.

Then

$$X\hat{\beta} - \overline{Y} = P_1\vec{Y} - P_2\vec{Y} = \frac{\text{Cov }(\vec{x}, \vec{y})}{\text{Var }(\vec{x})}(\vec{x} - \overline{x}),$$

so that, in (4) of Corollary A2.11, $k = 2$ and

$$\sum_{j=1}^{n}(Y_j - \overline{Y})^2 = \sum_{j=1}^{n}(Y_j - (\hat{\beta}_1 + \hat{\beta}_2 x_j))^2 + \frac{(\text{ Cov }(\vec{x}, \vec{y}))^2}{\text{Var }(\vec{x})}.$$

CASE 2: one-way analysis of variance. For another example of an ANOVA table, one-way analysis of variance is

$$Y_{i,j} = \theta_i + \epsilon_{ij} \ \ 1 \le i \le k, 1 \le j \le n_i.$$

$\epsilon_{ij} \sim N(0, \sigma^2)$ for all i, j, and $\{\epsilon_{ij}\}_{i,j}$ independent.

The only confusion here is terminology. $Y_{i,j}$ is not a vector, so we are not in the form of (A2.4), unless we rewrite $Y_{i,j}$ as a vector. There is more than one way to do this, so it is ambiguous to pretend that one-way analysis of variance, as written above, is in the form (A2.4).

Let $n \equiv \sum_{i=1}^{k} n_i$. If we describe our projections as maps from $\{(i, j) \,|\, 1 \le i \le k, 1 \le j \le n_i\}$ into itself, then we have

$$(P_1(\{Y_{ij}\})_{\ell,m} \equiv \overline{Y}_\ell \equiv \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} Y_{\ell j}, \qquad (P_2(\{Y_{ij}\})_{\ell,m} \equiv \overline{Y} \equiv \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} \qquad 1 \le \ell \le k, 1 \le m \le n_\ell.$$

When we want to represent projections by matrices, we will denote by $\vec{Y}$ the vector in $\mathbf{R}^n$

$$(Y_{11}, Y_{12}, ..., Y_{1n_1}, Y_{21}, Y_{22}, ..., Y_{2n_2}, ..., Y_{k1}, ..., Y_{kn_k}).$$

Then

$$P_2 = \frac{1}{n} J_n$$

(notation as in CASE 1), the orthogonal projection of $\vec{Y}$ onto the space of constant vectors, and

$$P_1 = \begin{bmatrix} \frac{1}{n_1} J_{n_1} & 0_{n_2} & \cdot & \cdot & 0_{n_k} \\ 0_{n_1} & \frac{1}{n_2} J_{n_2} & 0_{n_3} & \cdot & 0_{n_k} \\ \cdot & \cdot & & & \\ & & & & \frac{1}{n_k} J_{n_k} \end{bmatrix},$$

the orthogonal projection onto

$$\{(Y_{11}, Y_{12}, ..., Y_{1n_1}, Y_{21}, Y_{22}, ..., Y_{2n_2}, ..., Y_{k1}, ..., Y_{kn_k}) \,|\, Y_{ij} = Y_{i\ell},\ 1 \le i \le k, 1 \le j, \ell \le n_i\}.$$

Applying Theorem A2.10 to these choices of P_1 and P_2 gives us the following, again setting things up for confidence intervals, t tests and F tests.

Corollary A2.12.

(1) *For $1 \le i \le k, \overline{Y}_i$ is independent of $\sum_{ij}(Y_{ij} - \overline{Y}_i)^2$.*

(2) *For $1 \le i \le k, \overline{Y}_i \sim N(\theta_i, \frac{\sigma^2}{n_i})$.*

(3) *$\frac{1}{\sigma^2}\sum_{ij}(Y_{ij} - \overline{Y}_i)^2 \sim \chi^2_{n-k}$.*

(4) *If $\theta_i = \theta_\ell$, for $1 \le i, \ell \le k$, then $\sum_i n_i(\overline{Y}_i - \overline{Y})^2$ and $\sum_{i,j}(Y_{i,j} - \overline{Y}_i)^2$ are independent,*

$$\sum_{i,j}(Y_{i,j} - \overline{Y})^2 = \sum_{ij}(Y_{i,j} - \overline{Y}_i)^2 + \sum_i n_i(\overline{Y}_i - \overline{Y})^2$$

with $\frac{1}{\sigma^2}\sum_i n_i(\overline{Y}_i - \overline{Y})^2 \sim \chi^2_{k-1}$ and $\frac{1}{\sigma^2}\sum_{i,j}(Y_{i,j} - \overline{Y})^2 \sim \chi^2_{n-1}$.

Proof: Here $\vec{Y} \sim N(\vec{\mu}, \sigma^2 I_n)$, where $\mu_\ell = \theta_1$ for $1 \le \ell \le n_1$, $\mu_\ell = \theta_2$ for $n_1 + 1 \le \ell \le n_2$, etc. If we wrote $\vec{\mu}$ as an element of $\{(i,j) \,|\, 1 \le i \le k, 1 \le j \le n_i\}$,

$$\vec{\mu}_{ij} = \theta_i \quad 1 \le i \le k, 1 \le j \le n_i.$$

Either way, it should be clear that $P_1(\vec{\mu}) = \vec{\mu}$, and, for $\theta_i = \theta_\ell$, for $1 \le i, \ell \le k$, so that $\vec{\mu} = \vec{1}$, $P_2(\vec{\mu}) = \vec{\mu}$. Thus we may apply Theorem A2.10, to obtain assertions (1), (3) and (4). Assertion (2) is standard, from independence and normality of the ϵ_{ij} $\square$

A GENERAL CASE. A popular very general choice of reduced model is to satisfy the null hypothesis

$$C^T\vec{\beta} = \vec{d},$$

for C a $k \times m$ matrix of rank k, $\vec{d} \in \mathbf{R}^k$, so that P_2 is the orthogonal projection onto $\{X\vec{\beta} \,|\, C^T\vec{\beta} = \vec{d}\}$.

As before, let $\hat{\beta} \equiv (X^TX)^{-1}X^T\vec{Y}$, $P_1\vec{Y} \equiv X\hat{\beta}$, and let β_0 be such that $P_2\vec{Y} = X\beta_0$. Here's a straightforward orthogonality proof of what's needed for hypothesis tests, confidence intervals, etc.

Proposition A2.13.

(1) $\beta_0 = \hat{\beta} - (X^TX)^{-1}C(C^T(X^TX)^{-1}C)^{-1}(C^T\hat{\beta} - \vec{d})$.

(2) $\|X\hat{\beta} - X\beta_0\|^2 \equiv \|P_1\vec{Y} - P_2\vec{Y}\|^2 = (C^T\hat{\beta} - \vec{d})^T(C^T(X^TX)^{-1}C)^{-1}(C^T\hat{\beta} - \vec{d})$.

Proof: (1) We have $(\vec{Y} - X\beta_0) \perp \{X\vec{\beta} \,|\, C^T\vec{\beta} = \vec{0}\}$, thus

$$0 = \left\langle \vec{Y} - X\beta_0, X\vec{\beta} \right\rangle = \left\langle X^T(\vec{Y} - X\beta_0), \vec{\beta} \right\rangle \ \ \forall \vec{\beta} \in \mathcal{N}(C^T);$$

that is,

$$X^T(\vec{Y} - X\beta_0) \in \left[\mathcal{N}(C^T)\right]^\perp = \mathcal{R}(C)$$

by Proposition A1.20. Thus there exists $\vec{z}$ such that $X^T(\vec{Y} - X\beta_0) = C\vec{z}$, so that we may solve for β_0 :

$$\beta_0 = (X^TX)^{-1}(X^T\vec{Y} - C\vec{z}) = \hat{\beta} - (X^TX)^{-1}C\vec{z}, \qquad (*)$$

so that

$$\vec{d} = C^T\beta_0 = C^T\hat{\beta} - C^T(X^TX)^{-1}C\vec{z},$$

thus

$$\vec{z} = (C^T(X^TX)^{-1}C)^{-1}(C^T\hat{\beta} - \vec{d});$$

plugging this into (*) gives us (1).
(2) By (1),

$$\begin{aligned}
\|X\hat{\beta} - X\beta_0\|^2 &= \|X(X^TX)^{-1}C(C^T(X^TX)^{-1}C)^{-1}(C^T\hat{\beta} - \vec{d})\|^2 \\
&= \left(X(X^TX)^{-1}C(C^T(X^TX)^{-1}C)^{-1}(C^T\hat{\beta} - \vec{d})\right)^T \left(X(X^TX)^{-1}C(C^T(X^TX)^{-1}C)^{-1}(C^T\hat{\beta} - \vec{d})\right), \\
&= (C^T\hat{\beta} - \vec{d})^T(C^T(X^TX)^{-1}C)^{-1}C^T(X^TX)^{-1}X^TX(X^TX)^{-1}C(C^T(X^TX)^{-1}C)^{-1}(C^T\hat{\beta} - \vec{d}),
\end{aligned}$$

which becomes (2) after cancellation. $\square$

AIII. CONDITIONAL EXPECTATION AND PROBABILITY

Throughout this section use the inner product

$$\langle W, Z\rangle \equiv E(WZ),$$

although the definitions that follow could also be done with covariance (see Proposition A3.10).

I will state all results for random variables. To extend to random vectors, replace the inner product above with

$$\left\langle \vec{W}, \vec{Z} \right\rangle \equiv \sum_{k=1}^{n} E(W_k Z_k),$$

on the inner product space $L^2(\Omega, \mathcal{F}, P; \mathbf{R}^n)$, measurable functions $\vec{Z} : (\Omega, \mathcal{F}) \to \mathbf{R}^n$ such that $\sum_{k=1}^n E(Z_k^2) < \infty$.

Definitions A3.1. If $X \in L^2(\Omega, \mathcal{F}, P)$, and $\mathcal{G}$ is a sigma algebra contained in $\mathcal{F}$, then

$$E(X|\mathcal{G}) \equiv P_{L^2(\Omega,\mathcal{G},P)}(X). \tag{A3.1a}$$

This is equivalent to saying that $E(X|\mathcal{G}) \in L^2(\Omega, \mathcal{G}, P)$ and

$$E(XW) = E\left(E(X|\mathcal{G})W\right) \quad \forall W \in L^2(\Omega, \mathcal{G}, P).$$

By Theorem A1.14, this projection is guaranteed to exist.

Note that Definition A3.1 can be extended to $X \in L^1(\Omega, \mathcal{F}, P)$ by replacing $W \in L^2(\Omega, \mathcal{G}, P)$ with $W \in L^\infty(\Omega, \mathcal{G}, P)$.

The famous equality

$$E(E(X|\mathcal{G}_1)|\mathcal{G}_2) = E(E(X|\mathcal{G}_2)|\mathcal{G}_1) = E(X|\mathcal{G}_1),$$

when $\mathcal{G}_1 \subseteq \mathcal{G}_2$, follows immediately, with this definition, from a simple fact about orthogonal projections (see Proposition A1.22):

$$P_1 P_2 = P_2 P_1 = P_2,$$

if the range of P_2 (the space being projected onto) is contained in the range of P_1.

This definition also fits our intuition about conditional probability, a shrinking of the universe of possibilities, from $L^2(\Omega, \mathcal{F}, P)$ to the subspace $L^2(\Omega, \mathcal{G}, P)$.

If $Y : (\Omega, \mathcal{F}) \to \mathbf{R}$ is measurable, then

$$E(X|Y) \equiv E(X|\mathcal{F}(Y)), \tag{A3.1b}$$

where $\mathcal{F}(Y)$ is the sigma algebra generated by Y.

The function from $\mathbf{R} \to \mathbf{R}$

$$g(y) \equiv E(X|Y = y) \tag{A3.1c}$$

satisfies $g(Y) = E(X|Y)$; as usual, this is defined only almost everywhere w.r.t. the measure on $\mathbf{R}$ induced by Y. The existence of such a function is guaranteed by the following, a quick consequence of the Radon-Nikodym theorem, which itself is proven most easily by Riesz's lemma (Proposition A1.18).

Lemma A3.2 ([Ch, p. 299]). *If Z is a random variable measurable w.r.t. $\mathcal{F}(W)$, then there exists Borel-measurable ϕ such that $Z = \phi(W)$.*

I would like to characterize when conditioning on one random variable is stronger than conditioning on another. In the following, $L^2(\mathcal{F}(T))$ is shorthand for $L^2\left(\Omega, \mathcal{F}(T), P\right)$ and P_T means the orthogonal projection onto $L^2(\mathcal{F}(T))$.

Corollary A3.3. *If T and S are random variables, then the following are equivalent.*

(a) $E(k(T)|S) = k(T)$, *when k is Borel measurable and $k(T)$ has finite variance.*

(b) $\mathcal{F}(T) \subseteq \mathcal{F}(S)$.

(c) $L^2(\mathcal{F}(T)) \subseteq L^2(\mathcal{F}(S))$.

(d) $P_T P_S = P_T$.

(e) *there exists Borel measurable g such that $T = g(S)$.*

If T has finite variance, then all of the above are equivalent to

(f) $E(T|S) = T$.

Proof: The equivalence of (b), (c), and (d) seems clear. (e) $\rightarrow$ (b) is clear from the definition of $\mathcal{F}(T)$ and $\mathcal{F}(S)$.
(c) $\iff$ (a). Lemma A3.2 implies that

$$L^2(\Omega, \mathcal{F}(W), P) = \{g(W) \,|\, g : \mathbf{R} \rightarrow \mathbf{R} \text{ is Borel measurable, } g(W) \in L^2(\Omega, \mathcal{F}, P)\}.$$

Thus both (a) and (c) are equivalent to: for all k such that $k(T) \in L^2$ there exists Borel g such that $k(T) = g(S)$; $g(s) = E(k(T)|S = s)$.)
(a) $\rightarrow$ (e). For $n \in \mathbf{N}$, define $k_n \equiv 1_{[-n,n]}$; for all n, there exists g_n such that $k_n(T) = g_n(S)$. For $|T(\omega)| \leq n$, $T(\omega) = k_n(T(\omega)) = g_n(S(\omega))$, so define

$$g(S(\omega)) \equiv g_n(S(\omega)), \quad \text{for} \quad |T(\omega)| \leq n.$$

Then g is Borel measurable and $g(S) = T$.
(e) $\iff$ (f), when T has finite variance, is clear. $\square$

All three of the definitions generalize to conditional probabilities:

$$P(A|\cdot) \equiv E(1_A|\cdot), \tag{A3.1d}$$

for $A \subseteq \mathcal{F}$.

Finally, for $A, B \in \mathcal{F}$, $P(A|B)$ is the value of $E(1_A|1_B)$ on B :

$$P(A|B) \equiv [E(1_A|1_B)](\omega) \quad (\omega \in B). \tag{A3.1e}$$

Note that

$$E(X1_B(Y)) = E\left(E(X|Y)1_B(Y)\right) = E\left((y \mapsto E(X|Y = y)1_B(y))\,(Y)\right) = \int_B E(X|Y = y)\, dP(Y \leq y) \tag{A3.4}$$

for all Borel B; in particular,

$$P(A \cap (Y \in B)) = \int_B P(A|Y = y)\, dP(Y \leq y), \tag{A3.5}$$

for all $A \in \mathcal{F}$, Borel B.

It is desirable to have constructive versions of these definitions.

Lemma A3.6. *If*

$$\nu(A) \equiv \int_A g\, d\mu,$$

where μ is the sum of continuous and discrete finite measures on the real line, then

$$g(x) = \lim_{n \rightarrow \infty} \frac{\nu([x, x + \frac{1}{n}])}{\mu([x, x + \frac{1}{n}])} \;\; a.e.(\mu)$$

for x in the support of g.

Proof: If $d\mu = f\,dm$, where m is Lebesgue measure, then

$$\frac{\nu([x, x+\frac{1}{n}])}{\mu([x, x+\frac{1}{n}])} = \frac{n\int_x^{x+\frac{1}{n}} g(t)f(t)\,dm(t)}{n\int_x^{x+\frac{1}{n}} f(t)\,dm(t)};$$

the numerator converges to $g(x)f(x)$, the denominator to $f(x)$ a.e. (m).

If μ is discrete, then

$$\frac{\nu([x, x+\frac{1}{n}])}{\mu([x, x+\frac{1}{n}])} = \frac{g(x)\mu(x) + \sum_{y\in(x,x+\frac{1}{n}]}\nu(y)}{\mu(x) + \sum_{y\in(x,x+\frac{1}{n}]}\mu(y)}.$$

Since $\sum_{y\in\mathbf{R}}\nu(y) < \infty$, dominated convergence implies that

$$\lim_{n\to\infty}\sum_{y\in(x,x+\frac{1}{n}]}\nu(y) = 0;$$

the same argument holds for $\lim_{n\to\infty}\sum_{y\in(x,x+\frac{1}{n})}\mu(y)$. $\square$

Corollary A3.7. *For $A \in \mathcal{F}$,*

$$P(A|Y=y) = \lim_{n\to\infty}\frac{P(A\cap(Y\in[y, y+\frac{1}{n}]))}{P(Y\in[y, y+\frac{1}{n}])}$$

almost everywhere w.r.t. the measure induced by Y, if Y is the sum of a continuous and discrete random variable.

Proof: Let $\nu(B) \equiv P(A\cap(Y\in B)), \mu(B) \equiv P(Y\in B)$, and apply Lemma A3.6 and (A3.5). $\square$

Now we may justify the "expectation" terminology.

Proposition A3.8. *Suppose $A \mapsto P(X\in A|Y=y)$ defines a Borel measure on $\mathbf{R}$, for all real y. Then*

$$E(X|Y=y) = \int_{\mathbf{R}} x\,dF_{X|Y=y}(x),\ \ \forall y\in\mathbf{R},$$

where $F_{X|Y=y}(x) \equiv P(X\le x|Y=y)$.

Proof: Let's denote by F_Y the measure induced by Y, and by $F_{(X,Y)}$ the measure induced by (X,Y).

For A, B Borel subsets of the real line, by (A3.5),

$$\int_{A\times B} dF_{X|Y=y}(x)\,dF_Y(y) = \int_B\int_A dF_{X|Y=y}(x)\,dF_Y(y) = \int_B P(X\in A|Y=y)\,dF_Y(y) = P(X\in A, Y\in B)$$
$$= \int_{A\times B} dF_{(X,Y)}(x,y).$$

In the usual way, since the Borel sigma algebra on the plane is generated by $\{A\times B \mid A, B \text{ are Borel subsets of } \mathbf{R}\}$, this implies that

$$\int_{\mathbf{R}^2} h(x,y)\,dF_{X|Y=y}(x)\,dF_Y(y) = \int_{\mathbf{R}^2} h(x,y)\,dF_{(X,Y)}(x,y),$$

for all Borel $h : \mathbf{R}^2 \to \mathbf{R}$.

In particular, for any Borel $g : \mathbf{R}\to\mathbf{R}$ such that $g(Y)\in L^2(\Omega,\mathcal{F},P)$, if

$$k(y) \equiv \int_{\mathbf{R}} x\,dF_{X|Y=y}(x),$$

then

$$\langle k(Y), g(Y)\rangle \equiv E(k(Y)g(Y)) = \int_{\mathbf{R}} k(y)g(y)\, dF_Y(y) = \int_{\mathbf{R}}\int_{\mathbf{R}} xg(y)\, dF_{X|Y=y}\, dF_Y(y)$$
$$= \int_{\mathbf{R}^2} xg(y)\, dF_{(X,Y)}(x,y) = E(Xg(Y)) \equiv \langle X, g(Y)\rangle\,;$$

this is saying that

$$k(Y) = P_{L^2(\Omega,\mathcal{F}(Y),P}(X),$$

as desired. □

This proposition justifies the terminology

$$\text{Var }(X|\mathcal{G}) \equiv E(X^2|\mathcal{G}) - [E(X|\mathcal{G})]^2\,, \tag{A3.9}$$

with the obvious extensions to $\text{Var}(X|Y)$ and $\text{Var}(X|Y=y)$, since, for X, Y as in Proposition A3.8, the variance of the distribution defined by the cdf $F_{X|Y=y}$, from the proof, equals

$$\text{Var }(X|Y=y) = \int_{\mathbf{R}} x^2\, dF_{X|Y=y}(x) - \left[\int_{\mathbf{R}} x\, dF_{X|Y=y}(x)\right]^2.$$

Conditional expectation may be defined in terms of covariance.

Proposition A3.10. *For $X, \mathcal{G}$ as in Definition A3.1, with respect to the inner product*

$$\langle W, Z\rangle_2 \equiv \ Cov\ (W, Z),$$

$E(X|\mathcal{G})$ is that element of $P_{L^2(\Omega,\mathcal{G},P)}(X)$ whose expectation is $E(X)$.

Proof: Let Y be the element of $P_{L^2(\Omega,\mathcal{G},P)}(X)$ whose expectation is $E(X)$. By definition, for all $W \in L^2(\Omega,\mathcal{G},P)$,

$$E(XW) - E(X)E(W) = \text{ Cov }(X,W) = \text{ Cov }(Y,W) = E(YW) - E(Y)E(W) = E(YW) - E(X)E(W),$$

thus $E(XW) = E(YW)$. □

Here is a famous result that follows quickly and naturally from the Pythagorean theorem.

Proposition A3.11. *If X and Y are random variables and X has finite variance, then*

$$Var\ (X) = \ Var\ (E(X|Y)) + E(\ Var\ (X|Y)).$$

Proof: By the definition of the orthogonal projection, with respect to both $\langle\cdot\rangle$ and $\langle W, Z\rangle_2 \equiv \text{ Cov }(W, Z)$ (see Definition A3.1 and Proposition A3.10), and the Pythagorean theorem,

$$E\left((X - E(X|Y))^2\right) = E(X^2) - E\left((E(X|Y))^2\right)$$

and

$$\text{Var }(X - E(X|Y)) = \text{ Var }(X) - \text{ Var }(E(X|Y))\,.$$

Since $E(X) = E\left(E(X|Y)\right)$, it follows that $\text{Var}(X - E(X|Y)) = E\left((X - E(X|Y))^2\right)$, thus

$$\text{Var }(X) - \text{ Var }(E(X|Y)) = E(X^2) - E\left((E(X|Y)^2\right) = E\left(E(X^2|Y)\right) - E\left((E(X|Y)^2\right)$$
$$= E\left[E(X^2|Y) - (E(X|Y))^2\right] \equiv E\,(\ \text{Var }(X|Y))\,.$$

□

The construction of Proposition A3.8 becomes even more explicit when combined with everyone's favorite wishful thinking.

Proposition A3.12. *If (X, Y) is jointly continuous (discrete), then for y in the support of Y,*

$$dP(X \le x | Y = y) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)}\, dx,$$

where, for any random vector W, f_W is the density (mass) function for W.

Proof: I think this is clear when X and Y are discrete. Suppose (X, Y) is jointly continuous. For $A \subseteq \mathbf{R}$ Borel measurable, Corollary A3.7, dominated convergence, and standard results about differentiating integrals, in that order, imply that, for almost all y,

$$\begin{aligned} P(X \in A | Y = y) &= \lim_{n\to\infty} \frac{P(X \in A, (Y \in [y, y + \frac{1}{n}]))}{P(Y \in [y, y + \frac{1}{n}])} = \lim_{n\to\infty} \frac{n \int_A \int_y^{y+\frac{1}{n}} f_{(X,Y)}(x, t)\, dt\, dx}{n \int_y^{y+\frac{1}{n}} f_Y(s)\, ds} \\ &= \frac{\int_A \left[\lim_{n\to\infty} n \int_y^{y+\frac{1}{n}} f_{(X,Y)}(x, t)\, dt\right] dx}{\lim_{n\to\infty} n \int_y^{y+\frac{1}{n}} f_Y(s)\, ds} = \frac{\int_A f_{(X,Y)}(x, y)\, dx}{f_Y(y)} = \int_A \left[\frac{f_{(X,Y)}(x, y)}{f_Y(y)}\right] dx. \end{aligned}$$

□

Remark A3.13. The formula in Proposition A3.12 is *not* sufficient (this will be seen to be a pun) for treating important examples of conditional probability. The definition of sufficiency, a fundamental concept in statistics, involves $E(\vec{X}|Y)$, where $Y = T(\vec{X})$, a function of $\vec{X}$. When $\vec{X}$ is continuous, it is highly unlikely that $(\vec{X}, Y)$ will be jointly continuous. For a very simple one-dimensional case, consider $Y = X^2$, for X continuous with support equal to $\mathbf{R}$. Then (X, Y) has support equal to the parabola $\{(x, y) \,|\, y = x^2\}$, thus is neither discrete nor continuous. For any fixed, positive y, the distribution induced by the cdf

$$F_{X|Y=y} \equiv P(X \le x | Y = y)$$

is supported on $\{-\sqrt{y}, \sqrt{y}\}$; if (X, Y) were jointly continuous, Proposition A3.12 would imply that $F_{X|Y=y}$ is continuous.

I have often found it necessary to use Corollary A3.7 to construct the distribution for $X|Y = y$, or Proposition A3.8 for $E(X|Y)$.

Independence can be characterized both in terms of conditional expectation and covariance.

Proposition A3.14. *The following are equivalent, for random variables X and Y.*

(a) *X and Y are independent.*

(b) *$Cov(h(X), g(Y)) = 0$, for all Borel $h, g : \mathbf{R} \to \mathbf{R}$ such that $h(X)$ and $g(Y)$ have finite variance.*

(c) *$E(h(X)|Y) = E(h(X))$, for h as in (b).*

Proof: (a) → (b). Independence of X and Y implies that $h(X)$ and $g(Y)$ are independent, which implies that their covariance is zero.
(b) ⟺ (c). By Proposition A3.10, $E(h(X)|Y) = E(h(X))$ if and only if, for all g as in (b), $\mathrm{Cov}(h(X), g(Y)) = \mathrm{Cov}\,(E(h(X)), g(Y))$, which equals zero.
(b) → (a). For any real s, t, since $\mathrm{Cov}(e^{itX}, e^{isY}) = 0$, it follows that $E(e^{itX}e^{isY}) = E(e^{itX})E(e^{isY})$; that is,

$$\int_{\mathbf{R}^2} e^{i(tx+sy)}\, dF_{(X,Y)}(x, y) = \left(\int_{\mathbf{R}} e^{itx}\, dF_X(x)\right)\left(\int_{\mathbf{R}} e^{isy}\, dF_Y(y)\right) = \int_{\mathbf{R}^2} e^{i(tx+sy)}\, dF_X(x)\, dF_Y(y).$$

By uniqueness of Fourier-Stieltjes transforms, this implies that

$$dF_{(X,Y)} = dF_X(x)\, dF_Y(y),$$

which implies independence. □

Remark A3.15. (b) is a statement about orthogonality, w.r.t. to the covariance pre-inner product, of subspaces, and gives perspective about the relationship between covariance and independence: X and Y are independent if the subspaces $L^2(\Omega, \mathcal{F}(X))$ and $L^2(\Omega, \mathcal{F}(Y))$ are orthogonal, while orthogonality of the individual random variables X and Y is equivalent to X and Y being uncorrelated.

AIV. SUFFICIENCY

Throughout this section, $\vec{X} \equiv (X_1, X_2, ..., X_n)$ is a random sample whose distribution depends on a parameter θ and E_θ denotes expected value given the parameter θ; that is, for any Borel measurable $h : \mathbf{R}^n \to \mathbf{R}$,

$$E_\theta(h(\vec{X})) = \int_{\mathbf{R}^n} h(\vec{x})\, dP_\theta(X_k \le x_k, 1 \le k \le n).$$

Also T will always be a statistic $T(\vec{X})$. For simplicity of notation, T and θ will be assumed one-dimensional; it should be clear how to replace integrals over $\mathbf{R}$ with integrals over $\mathbf{R}^m$, etc., to extend to m-dimensional T and θ.

Definition A4.1. The statistic T is *sufficient* for θ if $E_\theta(k(\vec{X})\,|T)$ is constant w.r.t. θ, for all Borel measurable $k : \mathbf{R}^n \to \mathbf{R}$ such that $k(\vec{X})$ has finite variance.

This is saying (see the previous section) that the orthogonal projection onto

$$\{g(T)\,|\,g \text{ is Borel measurable, } g(T) \text{ has finite variance }\}$$

is unaffected by θ. Since changing θ changes the inner product

$$\left\langle f(\vec{X}), h(\vec{X}) \right\rangle_\theta \equiv E_\theta(f(\vec{X})h(\vec{X})),$$

hence changes orthogonality, this is a strong statement.

Theorem A4.2 (Factorization Theorem). *Suppose that either*

(1) $\vec{X}$ *is discrete, or*

(2) $\vec{X}$ *is continuous and* T *is continuous with density function* $t \mapsto f_T(t|\theta)$ *continuous on its support. Then the following are equivalent.*

(a) T *is sufficient for* θ.

(b) *For all* θ *there exists continuous* g_θ *and Borel* h *such that*

$$f(\vec{x}|\theta) = g_\theta(T(\vec{x}))h(\vec{x}) \;\; a.e.$$

(c) $\frac{f(\vec{x}|\theta)}{f_T(T(\vec{x})|\theta)}$ *is constant w.r.t.* θ.

Proof: First let's assume hypothesis (2).
(a) $\to$ (c). For $\vec{x}$ in the support of $\vec{X}, \epsilon > 0$, denote by

$$B_\epsilon(\vec{x}) \equiv \{\vec{y} \in \mathbf{R}^n \,|\, \|\vec{y} - \vec{x}\| < \epsilon\},$$

and define

$$k_{\vec{x},\epsilon} \equiv 1_{B_\epsilon(\vec{x})}.$$

By definition of conditional expectation,

$$\int_{B_\epsilon(\vec{x})} f(\vec{y}|\theta)\, d\vec{y} = E_\theta(k_{\vec{x},\epsilon}(\vec{X})) = E_\theta\left(E(k_{\vec{x},\epsilon}(\vec{X})|T)\right) = \int E(k_{\vec{x},\epsilon}(\vec{X})|T = t) f_T(t|\theta)\, dt$$
$$= \int_{T(B_\epsilon(\vec{x}))} E(k_{\vec{x},\epsilon}(\vec{X})|T = t) f_T(t|\theta)\, dt.$$

By the intermediate-value theorem, there exists $t_{\vec{x},\epsilon,\theta} \in T(B_\epsilon(\vec{x}))$ such that the last integral equals

$$f_T(t_{\vec{x},\epsilon,\theta}|\theta) \int_{T(B_\epsilon(\vec{x}))} E(k_{\vec{x},\epsilon}(\vec{X})|T = t)\, dt.$$

Now let ϵ go to zero in the equality

$$\frac{1}{\epsilon}\int_{B_\epsilon(\vec{x})} f(\vec{y}|\theta)\,d\vec{y} = f_T(t_{\vec{x},\epsilon,\theta}|\theta)\left[\frac{1}{\epsilon}\int_{T(B_\epsilon(\vec{x}))} E(k_{\vec{x},\epsilon}(\vec{X})|T=t)\,dt\right].$$

By the usual differentiation-of-integral result, we have

$$\frac{1}{\epsilon}\int_{B_\epsilon(\vec{x})} f(\vec{y}|\theta)\,d\vec{y} \to f(\vec{x}|\theta) \quad \text{a. e. .}$$

By continuity of T, $t_{\vec{x},\epsilon,\theta} \to T(\vec{x})$ as $\epsilon \to 0$, hence by continuity of f_T, $f_T(t_{\vec{x},\epsilon,\theta}) \to f_T(T(\vec{x})|\theta)$. This implies that

$$h(\vec{x}) \equiv \lim_{\epsilon\to 0}\left[\frac{1}{\epsilon}\int_{T(B_\epsilon(\vec{x}))} E(k_{\vec{x},\epsilon}(\vec{X})|T=t)\,dt\right]$$

exists a. e., with

$$f(\vec{x}|\theta) = f_T(T(\vec{x})|\theta)h(\vec{x}),$$

giving us (c).
(b) $\to$ (a). By Corollary A3.7, for A Borel measurable, t in the support of T,

$$\begin{aligned} P_\theta(\vec{X}\in A|T=t) &= \lim_{n\to\infty}\frac{P_\theta((\vec{X}\in A)\cap(T\in[t,t+\frac{1}{n}]))}{P_\theta(T\in[t,t+\frac{1}{n}])} \\ &= \lim_{n\to\infty}\frac{\int_{A\cap T^{-1}([t,t+\frac{1}{n}])} g_\theta(T(x))h(x)\,dx}{\int_{T^{-1}([t,t+\frac{1}{n}])} g_\theta(T(y))h(y)\,dy} \\ &= \lim_{n\to\infty}\frac{g_\theta(T(x_{t,n}))\int_{A\cap T^{-1}([t,t+\frac{1}{n}])} h(x)\,dx}{g(T(y_{t,n}))\int_{T^{-1}([t,t+\frac{1}{n}])} h(y)\,dy}, \end{aligned}$$

for $x_{t,n}, y_{t,n} \in T^{-1}([t,t+\frac{1}{n}])$, by the intermediate-value theorem. By continuity of g_θ,

$$\lim_{n\to\infty} g_\theta(T(x_{t,n})) = g_\theta(t) = \lim_{n\to\infty} g_\theta(T(y_{t,n})).$$

Thus

$$P_\theta(\vec{X}\in A|T=t) = \lim_{n\to\infty}\frac{\int_{A\cap T^{-1}([t,t+\frac{1}{n}])} h(x)\,dx}{\int_{T^{-1}([t,t+\frac{1}{n}])} h(y)\,dy},$$

which is constant w.r.t. θ. In the usual measure-theoretic way, this extends to $E_\theta(k(\vec{X})|T)$ being constant w.r.t. θ, for any Borel measurable k such that $k(T)$ has finite variance.
(c) $\to$ (b)is clear by letting $g_\theta(t) \equiv f_T(t|\theta), h(\vec{x}) \equiv \frac{f(\vec{x}|\theta)}{f_T(T(\vec{x})|\theta))}$.

Under hypothesis (1), for $\vec{x}$ in the support of $\vec{X}$,

$$\begin{aligned} f(\vec{x}|\theta) &= P_\theta(\vec{X}=\vec{x}) = P_\theta((\vec{X}=x)\cap(T=T(\vec{x}))) = P_\theta(\vec{X}=\vec{x}|T=T(\vec{x}))P_\theta(T=T(\vec{x})) \\ &= P_\theta(\vec{X}=\vec{x}|T=T(\vec{x}))f_T(T(\vec{x})|\theta), \end{aligned}$$

so that (a) and (c) are equivalent. Clearly (c) $\to$ (b), while if (b) holds, then

$$\frac{f(\vec{x}|\theta)}{f_T(T(\vec{x})|\theta)} = \frac{P(\vec{X}=\vec{x})}{\sum_{T(\vec{y})=T(\vec{x})} P(\vec{X}=\vec{y})} = \frac{h(\vec{x})}{\sum_{T(\vec{y})=T(\vec{x})} h(\vec{y})},$$

which is constant w.r.t. θ. $\square$

Remark A4.3. Information and geometry. When one is asked what sufficiency "means," the socially appropriate response is something like "the sufficient statistic contains all the information about θ from the data." Aside from the Fisher information number, this author has seen no definition of information, so it is natural to ask what information is contained in that quoted speech about information.

The Likelihood Principle says that (c) of the Factorization Theorem implies that $\vec{x}$ and $T(\vec{x})$ contain the same information about θ. Thus if we accept the Likelihood Principle as a meaningful axiom, then the quoted speech above acquires meaning. See also Remark A4.5.

There is also an interesting geometric interpretation of sufficiency. Consider first the case where $\vec{X}$ is discrete, so that, for $\vec{x}$ in the support of $\vec{X}$,

$$E_\theta\left(1_{\{\vec{x}\}}(\vec{X})|T\right) = \frac{P_\theta(\vec{X}=\vec{x})}{P_\theta(T=T(\vec{x}))}1_{\{T(\vec{x})\}}(T).$$

Denoting by $\|\cdot\|_\theta$ the Hilbert-space norm

$$\|Z\|_\theta^2 \equiv \langle Z,Z\rangle_\theta\,,\quad \langle Z,W\rangle_\theta \equiv E_\theta(ZW),$$

we have

$$\frac{\|1_{\{\vec{x}\}}(\vec{X})\|_\theta^2}{\|E_\theta\left(1_{\{\vec{x}\}}(\vec{X})|T\right)\|_\theta^2} = \left[\frac{P_\theta(T=T(\vec{x}))}{P_\theta(\vec{X}=\vec{x})}\right]^2 \frac{\|1_{\{\vec{x}\}}(\vec{X})\|_\theta^2}{\|1_{\{T(\vec{x})\}}(T)\|_\theta^2} = \left[\frac{P_\theta(T=T(\vec{x}))}{P_\theta(\vec{X}=\vec{x})}\right].$$

Thus sufficiency is equivalent to

$$\frac{\|1_{\{\vec{x}\}}(\vec{X})\|_\theta}{\|E_\theta\left(1_{\{\vec{x}\}}(\vec{X})|T\right)\|_\theta}$$

being constant w.r.t. θ. This is saying that the right triangles formed by $\vec{0}, 1_{\{\vec{x}\}}(\vec{X})$, and $E_\theta\left(1_{\{\vec{x}\}}(\vec{X})|T\right)$ remain similar, as θ runs through the parameter space.

For $\vec{X}, T$ as in hypothesis (2) of Theorem A4.2, there is a similar result, with $B_\epsilon(\vec{x})$, for $\epsilon>0, \vec{x}$ in the support of $\vec{X}$, terminology as in the proof of Theorem A4.2, replacing $\vec{x}$. Since T is being assumed sufficient, the θ in E_θ may be removed.

It is not clear to me if

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\|E(1_{B_\epsilon(\vec{x})}(\vec{X})|T)\|^2 \qquad (*)$$

exists (although by the Pythagorean theorem,

$$\frac{1}{\epsilon}\|E(1_{B_\epsilon(\vec{x})}(\vec{X})|T)\|^2 \le \frac{1}{\epsilon}\|1_{B_\epsilon(\vec{x})}(\vec{X})\|^2,$$

which converges to $f(\vec{x}|\theta)$ for almost all $\vec{x}$), so I will restrict attention to sequences $\epsilon_k\to 0$ for which (*) does converge.

Proposition. *If T is sufficient, $\vec{x}$ is in the support of $\vec{X}$, $\epsilon_k\to 0$, and*

$$\lim_{k\to\infty}\frac{1}{\epsilon_k}\|E(1_{B_{\epsilon_k}(\vec{x})}(\vec{X})|T)\|_\theta^2$$

exists, for some θ, then

$$\lim_{k\to\infty}\frac{\|1_{B_{\epsilon_k}(\vec{x})}(\vec{X})\|_\theta^2}{\|E(1_{B_{\epsilon_k}(\vec{x})}(\vec{X})|T)\|_\theta^2}$$

exists, for all θ, and is a constant w.r.t. θ.

Proof: Define

$$g_{\vec{x},\epsilon}(t) \equiv E\left(1_{B_\epsilon(\vec{x})}(\vec{X})|T=t\right).$$

From the proof of Theorem A4.2(a) $\to$ (c),

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{T(B_\epsilon(\vec{x}))} g_{\vec{x},\epsilon}(t)\,dt = \frac{f(\vec{x}|\theta)}{f_T(T(\vec{x})|\theta)} \quad \text{and} \quad \lim_{\epsilon\to 0}\frac{1}{\epsilon}\|1_{B_\epsilon(\vec{x})}(\vec{X})\|^2 = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{B_\epsilon(\vec{x})} f(\vec{y}|\theta)\,d\vec{y} = f(\vec{x}|\theta) \ \text{ a.e. },$$

so that

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\|1_{B_\epsilon(\vec{x})}(\vec{X})\|_\theta^2 = f_T(T(\vec{x})|\theta)\lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{T(B_\epsilon(\vec{x}))} g_{\vec{x},\epsilon}(t)\,dt.$$

Also, arguing as in the proof of Theorem A4.2(a) $\to$ (c),

$$\lim_{k\to\infty}\frac{1}{\epsilon_k}\int_{T(B_{\epsilon_k}(\vec{x}))} \left(g_{\vec{x},\epsilon_k}(t)\right)^2\,dt$$

exists, with

$$\begin{aligned}\lim_{k\to\infty}\frac{1}{\epsilon_k}\|E(1_{B_{\epsilon_k}(\vec{x})}(\vec{X})|T)\|_\theta^2 &= \lim_{k\to\infty}\frac{1}{\epsilon_k}\int_{T\left(B_{\epsilon_k}(\vec{x})\right)} \left(g_{\vec{x},\epsilon}(t)\right)^2\, f_T(t|\theta)\,dt \\ &= f_T(T(\vec{x})|\theta)\lim_{k\to\infty}\frac{1}{\epsilon_k}\int_{T(B_{\epsilon_k}(\vec{x}))} \left(g_{\vec{x},\epsilon_k}(t)\right)^2\,dt.\end{aligned}$$

Thus

$$\lim_{k\to\infty}\frac{\|1_{B_{\epsilon_k}(\vec{x})}(\vec{X})\|_\theta^2}{\|E(1_{B_{\epsilon_k}(\vec{x})}(\vec{X})|T)\|_\theta^2} = \lim_{k\to\infty}\frac{\frac{1}{\epsilon_k}\|1_{B_{\epsilon_k}(\vec{x})}(\vec{X})\|_\theta^2}{\frac{1}{\epsilon_k}\|E(1_{B_{\epsilon_k}(\vec{x})}(\vec{X})|T)\|_\theta^2} = \frac{\lim_{k\to\infty}\frac{1}{\epsilon_k}\int_{T(B_{\epsilon_k}(\vec{x}))} g_{\vec{x},\epsilon_k}(t)\,dt}{\lim_{k\to\infty}\frac{1}{\epsilon_k}\int_{T(B_{\epsilon_k}(\vec{x}))}\left(g_{\vec{x},\epsilon_k}(t)\right)^2\,dt},$$

clearly constant w.r.t. θ. $\square$

Theorem A4.4. *Suppose $\vec{X}$ and T are as in Theorem A4.2 and there exists continuous $k : Im(T) \to \mathbf{R}^n$ such that*

(1) *$T(k(z)) = z$ for all $z \in Im(T)$, and*

(2) *for all $\vec{x} \in \mathbf{R}^n$, the map $\theta \mapsto f(k(T(\vec{x}))|\theta)$ is not identically zero on $\{\theta \,|\, f(\vec{x}|\theta)$ is nonzero $\}$.*

Consider

(3)$_{\vec{x},\vec{y}}$ *$\theta \mapsto \frac{f(\vec{x}|\theta)}{f(\vec{y}|\theta)}$ constant on $\{\theta \,|\, f(\vec{y}|\theta)$ is nonzero $\}$.*

Then T is sufficient if and only if $((T(\vec{x}) = T(\vec{y})) \to (3)_{\vec{x},\vec{y}})$.

Proof: Suppose T is sufficient. Then by Theorem A4.2, $f(\vec{x}|\theta) = f_T(T(\vec{x})|\theta)h(\vec{x})$, thus if $T(\vec{x}) = T(\vec{y})$, we have

$$\frac{f(\vec{x}|\theta)}{f(\vec{y}|\theta)} = \frac{h(\vec{x})}{h(\vec{y})},$$

so that $(3)_{\vec{x},\vec{y}}$ holds.

Conversely, suppose $((T(\vec{x}) = T(\vec{y})) \to (3)_{\vec{x},\vec{y}})$. Define

$$g_\theta(z) \equiv f(k(z)|\theta) \ \ (z \in Im(T)).$$

Since $T(k(T(\vec{x}))) = T(\vec{x})$,

$$\theta \mapsto \frac{f(k(T(\vec{x}))|\theta)}{f(\vec{x}|\theta)} \quad \text{is constant} \qquad (*)$$

on $\{\theta \,|\, f(\vec{x}|\theta)$ is nonzero $\}$. By (2), there exists θ_0 such that both $f(\vec{x}|\theta_0)$ and $f(k(T(\vec{x}))|\theta_0)$ are nonzero. By (*), $f(k(T(\vec{x}))|\theta)$ is *never* zero on $\{\theta \,|\, f(\vec{x}|\theta)$ is nonzero $\}$. Thus, for $\vec{x}, \theta$ such that $f(\vec{x}|\theta)$ is nonzero, we can define

$$h(\vec{x}) \equiv \frac{f(\vec{x}|\theta)}{f(k(T(\vec{x}))|\theta)}.$$

By definition,

$$g_\theta(T(\vec{x}))h(\vec{x}) = f(\vec{x}|\theta)$$

for such $\vec{x}, \theta$; defining $h(\vec{x}) \equiv 0$ when $f(\vec{x}|\theta) = 0$ now implies, by the Factorization Theorem, that T is sufficient. $\square$

Remark A4.5. Note how Theorem A4.4 is consistent with the Likelihood Principle, which states that $\vec{x}$ and $\vec{y}$ contain the same information about θ if $(3)_{\vec{x},\vec{y}}$ holds. Theorem A4.4 is then saying that if T is sufficient and $T(\vec{x}) = T(\vec{y})$, then $\vec{x}$ and $\vec{y}$ contain the same information about θ; thus there is no gained information, after knowing $T(\vec{x})$, in knowing $\vec{x}$.

Theorem A4.6 (Basu's Theorem). *If T is a complete, sufficient statistic and S is an ancillary statistic, then S and T are independent for all θ.*

Proof: Fix h, an arbitrary Borel measurable function such that $h(S)$ has finite variance. Define, for $y \in \mathbf{R}$,

$$g(y) \equiv E_\theta(h(S)|T = y) - E_\theta(h(S)).$$

Note that, by definition of sufficiency and ancillarity, g is independent of θ. For all θ,

$$E_\theta(g(T)) = E_\theta\left(E_\theta(h(S)|T)\right) - E_\theta(h(S)) = 0,$$

by definition of conditional expectation. By definition of completeness, $g = 0$ almost everywhere w.r.t. the measure induced by T, thus, for all θ,

$$E_\theta(h(S)|T) = E_\theta(h(S)).$$

By Proposition A3.14, S and T are independent, for any θ. $\square$

Definition A4.7. The sufficient statistic T is *minimal* if, for any sufficient statistic S, there exists Borel measurable g such that $g(S) = T$.

See Corollary A3.3 for equivalent definitions. Of particular interest, using (e) $\iff$ (c) of Corollary A3.3, is the fact that minimality is equivalent to a statement about subspaces: $L^2(\mathcal{F}(T)) \subseteq L^2(\mathcal{F}(S))$, for any sufficient statistic S.

"Minimal" is actually a misnomer here. "Minimal" means there is no further reduction of data possible, without losing information about θ; equivalently, that there exists no sufficient S such that $L^2(\mathcal{F}(S))$ is properly contained in $L^2(\mathcal{F}(T))$. The "minimal" sufficient statistic T is actually a *minimum*, meaning that it represents a reduction of data from any other sufficient statistic. If we define the natural partial order of inclusion, $W \leq Z$ if $L^2(\mathcal{F}(W)) \subseteq L^2(\mathcal{F}(Z))$, then "minimal" means there exists no sufficient S strictly less than T, while *minimum* (what T should be called) means that $T \leq S$, for all sufficient S.

Proposition A4.8. *If there exists a minimal sufficient statistic for θ and T is complete and sufficient, then T is minimal sufficient.*

Proof: Let R be a minimal sufficient statistic. Fix Borel k such that $k(T)$ has finite variance. For any θ,

$$E_\theta\left(k(T) - E(k(T)|R)\right) = 0,$$

by definition of conditional expectation. But $E(k(T)|R)$ is a Borel function of R, which, by minimality, is a Borel function of T. Thus $(k(T) - E(k(T)|R))$ is a Borel function of T whose expectation is zero, for all θ, thus by completeness

$$k(T) = E(k(T)|R).$$

By Corollary A3.3, there exists Borel h such that $T = h(R)$.

If S is any sufficient statistic, then by minimality there exists Borel ℓ such that $R = \ell(S)$. If $g \equiv h \circ \ell$, then

$$T = h(R) = (h \circ \ell)(S) \equiv g(S),$$

as desired. $\square$

Finally, let's return to the terminology of Theorem A4.4.

Proposition A4.9. *Suppose there exists a minimal sufficient statistic for θ, and for all $\vec{x}, \vec{y} \in \mathbf{R}^n$,*

$$(T(\vec{x}) = T(\vec{y})) \iff ((3)_{\vec{x},\vec{y}}) .$$

Then T is minimal sufficient.

Proof: By Theorem A4.4 T is sufficient. Let W be the hypothesized minimal sufficient statistic. If $W(\vec{x}) = W(\vec{y})$, then by Theorem A4.4, $(3)_{\vec{x},\vec{y}}$ holds, thus, by the hypothesis on T, $T(\vec{x}) = T(\vec{y})$. This means that

$$h(W(\vec{x})) \equiv T(\vec{x}), \ \ \forall \vec{x} \in \mathbf{R}^n$$

unambiguously defines h such that $h(W) = T$. Since W is minimal, there exists Borel measurable k such that $W = k(T)$. Thus h is the inverse of k, hence is Borel measurable.

Now let S be an arbitrary sufficient statistic. There exists Borel measurable ℓ such that $W = \ell(S)$, thus, if $g \equiv h \circ \ell$, then $T = g(S)$, as desired. $\square$

AV. VARIANCE SHRINKING

$\vec{X}, T$, and θ are as in the previous section.

The following was first proven by the mathematician Frechet ([Fr]), in 1943 (see [Le, p. 236]).

Theorem A5.1 (Cramer-Rao inequality). *Suppose $f(x|\theta)$ is a density or mass function for $\vec{X}$, $W : \mathbf{R}^n \to \mathbf{R}$ is Borel measurable, such that*

$$\frac{d}{d\theta}\int_{\mathbf{R}} W(\vec{x})f(\vec{x}|\theta)\,d\vec{x} = \int_{\mathbf{R}} \frac{\partial}{\partial\theta}W(\vec{x})f(\vec{x}|\theta)\,d\vec{x}, \quad \text{and} \quad \frac{d}{d\theta}\int_{\mathbf{R}} f(\vec{x}|\theta)\,d\vec{x} = \int_{\mathbf{R}} \frac{\partial}{\partial\theta}f(\vec{x}|\theta)\,d\vec{x}.$$

Then

$$Var_\theta W(\vec{X}) \geq \frac{\left(\frac{d}{d\theta}E_\theta(W(\vec{X}))\right)^2}{E_\theta\left(\left(\frac{d}{d\theta}\ln(f(\vec{X}|\theta))\right)^2\right)},$$

with equality occurring if and only if there exist $a(\theta), b(\theta)$ such that

$$W(\vec{X}) = a(\theta) + b(\theta)\frac{d}{d\theta}\ln(f(\vec{X}|\theta)).$$

Proof: Apply the Cauchy inequality, with the covariance pre-inner product on $L^2(\mathbf{R}^n, f(\vec{x}|\theta)\,d\vec{x})$,

$$\langle S, T\rangle_\theta \equiv \int_{\mathbf{R}^n} (S(\vec{x}) - E_\theta(S))\,(T(\vec{x}) - E_\theta(T))\,f(\vec{x}|\theta)\,d\vec{x}, \quad E_\theta(R) \equiv \int_{\mathbf{R}^n} R(\vec{x})f(\vec{x}|\theta)\,d\vec{x},$$

with $T(\vec{x}) \equiv \frac{d}{d\theta}\ln(f(\vec{x}|\theta))$. Note first that

$$E_\theta(T) = \int_{\mathbf{R}^n} \frac{\partial}{\partial\theta}f(\vec{x}|\theta)\,d\vec{x} = \frac{d}{d\theta}\int_{\mathbf{R}^n} f(\vec{x}|\theta)\,d\vec{x} = \frac{d}{d\theta}(1) = 0,$$

thus $\|T\|_\theta^2 = E_\theta\left(\left(\frac{d}{d\theta}\ln(f(\vec{X}|\theta))\right)^2\right)$, and $\langle S, T\rangle_\theta = \int_{\mathbf{R}^n} S(\vec{x})T(\vec{x})f(\vec{x}|\theta)\,d\vec{x}$, for any $S \in L^2(\mathbf{R}^n, f(\vec{x}|\theta)\,d\vec{x})$.

$$\begin{aligned}\frac{d}{d\theta}E_\theta W(\vec{X}) &= \frac{d}{d\theta}\int_{\mathbf{R}^n} W(\vec{x})f(\vec{x}|\theta)\,d\vec{x} = \int_{\mathbf{R}^n}\frac{\partial}{\partial\theta}W(\vec{x})f(\vec{x}|\theta)\,d\vec{x}\\ &= \int_{\mathbf{R}^n} W(\vec{x})\left[\frac{\partial}{\partial\theta}\ln(f(\vec{x}|\theta))\right]f(\vec{x}|\theta)\,d\vec{x} \qquad (*)\\ &= \langle W, T\rangle_\theta \leq \|W\|_\theta\|T\|_\theta = \sqrt{\mathrm{Var}_\theta(W(\vec{X}))E_\theta\left(\left(\frac{d}{d\theta}\ln(f(\vec{X}|\theta))\right)^2\right)},\end{aligned}$$

giving the inequality.

Again by the Cauchy inequality, and (*), equality occurs only if there exists a constant $b(\theta)$ such that

$$0 = \|W - b(\theta)T\|_\theta^2 = \mathrm{Var}_\theta(W(\vec{X}) - b(\theta)T(\vec{X})),$$

which is equivalent to the existence of another constant $a(\theta)$ such that

$$W(\vec{X}) - b(\theta)T(\vec{X}) = a(\theta).$$

□

Proposition A5.2. *Suppose X and Y are random variables, and X has finite variance.*

(a) $Var(E(X|Y)) \leq Var\,(X)$.

(b) $Var(E(X|Y)) = Var\,(X) \iff E(X|Y) = X \iff$ *there exists Borel measurable g such that* $X = g(Y)$.

Proof: By the Pythagorean theorem, w.r.t. the inner product $\langle W, Z\rangle \equiv \text{Cov}\,(W, Z)$,

$$\text{Var}\,(X) - \text{Var}\,(E(X|Y)) = \text{Var}\,(X - E(X|Y)).$$

This gives (a); also it implies that

$$\text{Var}\,(E(X|Y)) = \text{Var}\,(X) \iff \text{Var}\,(X - E(X|Y)) = 0 \iff X - E(X|Y) = E\,(X - E(X|Y)) = 0$$
$$\iff E(X|Y) = X \iff X = g(Y),$$

by definition of $E(X|Y)$ and Lemma A3.2. □

Since $E(E(X|Y)) = E(X)$ by definition, we immediately get the following. Note that $E_\theta(W|T)$ is independent of θ, for T sufficient and W a function of $\vec{X}$.

Corollary A5.3 (Rao-Blackwell). *Suppose W is an unbiased estimator of $\tau(\theta)$ and T is sufficient for θ. Then $E(W|T)$ is an unbiased estimator of $\tau(\theta)$ such that*

(1) $Var(E(W|T)) \leq Var\,(W)$; *and*

(2) $Var(E(W|T)) < Var\,(W)$, *unless* $E(W|T) = W$.

In other words, conditioning cannot increase variance, and, unless it leaves the entire random variable unchanged, will strictly decrease variance.

Proposition A5.4. *If a UMVUE exists, it is unique.*

Proof: Let $\mathcal{W}$ be the set of all unbiased estimators of finite variance, with pre-inner product $\langle W, Z\rangle \equiv \text{Cov}\,(W, Z)$. Since $\mathcal{W}$ is convex, and the UMVUE is the element of minimum norm in $\mathcal{W}$, Proposition A1.8 implies that, if X and Y are UMVUEs, then $\text{Var}(X - Y) = 0$. This implies that $X - Y = E(X - Y) = 0$, since X and Y are both unbiased, hence have equal expectations. □

Proposition A5.5. *If W is unbiased, then W is the UMVUE if and only if $Cov(W, U) = 0$ whenever U is an unbiased estimator of zero.*

Proof: Let $\mathcal{U}$ be the set of all unbiased estimators of zero. The set of all unbiased estimators equals $W + \mathcal{U}$, the set of random variables of the form $W + U$, $U \in \mathcal{U}$. Thus, w.r.t. the covariance inner product,

$$W \text{ is the UMVUE} \iff \|W\| \leq \|W + U\| \forall U \in \mathcal{U} \iff \|W\| \leq \|W + \alpha U\|\ \forall U \in \mathcal{U}, \alpha \in \mathbf{R},$$

since $\mathcal{U}$ is a vector space. By Proposition A1.3(c), the last equivalence is equivalent to $W \perp U, \forall U \in \mathcal{U}$, which by definition means $\text{Cov}(W, U) = 0$. □

The following is an immediate consequence of the definition of completeness.

Proposition A5.6 (Lehmann-Scheffe). *If T is complete and sufficient, then for any $\theta \mapsto \tau(\theta)$*

$$\{\ \textit{unbiased estimators of } \tau(\theta)\} \bigcap \{g(T) \mid g \textit{ is Borel measurable and } \ Var_\theta(g(T)) < \infty\ \ \forall\theta\}$$

is at most one point.

Then Rao-Blackwell quickly gives the following.

Proposition A5.7. *If T is complete and sufficient and g is a Borel measurable function, then $g(T)$ is the UMVUE of $E_\theta(g(T))$.*

Proof: For any unbiased estimator S, Rao-Blackwell implies that $\text{Var}_\theta(E(S|T)) \leq \text{Var}_\theta(S)$. By Rao-Blackwell and Lehmann-Scheffe, $E(S|T) = g(T)$. □

REFERENCES

[A] Aitken, A. C., *On least squares and linear combination of observations,* Proc. Royal Soc. of Edinb. 55 (1935), 42–48.
[Ch] Chung, K.L., A Course in Probability Theory, second edition, Academic Press, 1974.
[C] Cochran, W. G., *The omission or addition of an independent variate in multiple linear regression,* J. R. Statist. Soc. Suppl. 5 (1938), 171–176.
[D1] Dale, A.I., *A newly discovered result of Thomas Bayes,* Arch. Hist. Ex. Sci. 35 (1986), 101–113.
[D2] Dale, A.I., A History of Inverse Probability. From Thomas Bayes to Karl Pearson, Springer, New York, 1991.
[D-N] David, F. N. and Neyman, J., *Extension of the Markoff theorem on least squares,* Statist. Res. Mem. 2 (1938), 105–116.
[Ed] Edwards,A. W. F., *Commentary on the arguments of Thomas Bayes,* Scand. J. Stat. 5 (1978), 116–118.
[E] Eisenhart, C., *Boscovich and the combination of observations,* "Roger Joseph Boscovich," ed. L. L. Whyte, Allen and Unwin, London (1961), 200–212; see also [K-Pl], 88–100.
[F1] Farebrother, R. W., *The statistical estimation of the standard linear model, 1756–1853*, Proc. of the First International Tampere Seminar on Linear Statistical Models and Their Applications, eds. T. Pukkila and S. Puntanen, Dept. of Mathematical Sciences, Univ. of Tampere, 1985, 77–99.
[F2] Farebrother, R. W., *The historical development of the method of averages, 1750–1987,*Unpublished manuscript, Dept. of Econometrics, University of Manchester, UK, 1988.
[Fr] Fréchet, M., *Sur l'extension de certaines évaluations statistiques de petis échantillons,* Internat. Statis. Rev. 11 (1943), 182–205.
[H] Hald, A., A History of Mathematical Statistics, From 1750–1930, Wiley Series in Probability and Statistics, New York, NY, 1998.
[K1] Kendall, M. G., *Daniel Bernoulli on maximum likelihood*, "Studies in the History of Probability and Statistics," vol. 1. London: Charles Griffin, 155–156.
[K2] Kendall, M. G., *Where shall the history of statistics begin?*, Biometrika 47 (1960), 447–449; see also [PeE-K], 45–46.
[K-Pl] Kendall, M. G., and Plackett, R. L., eds. 1977, "Studies in the History of Statistics and Probability," vol. 2. London: Griffin.
[L] Lancaster, H. O., *Development of the notion of statistical dependence,* sixth New Zealand Mathematics Colloquium, Wellington, 17–19 May, 1971; see [K-Pl], 293–308.
[Le] Le Cam, L. M. and Yang, G. L., Asymptotics in Statistics: Some Basic Concepts, second edition, Springer, New York, 2000.
[Mo] Molina, E. C., *Bayes' theorem*, Ann. Math. Stat. 2 (1931), 23–37.
[PeE-K] Pearson, E. S., and Kendall, M. G., eds. 1970, "Studies in the History of Statistics and Probability," vol 1. London: Charles Griffin.
[PeK] Pearson, K., *Notes on the history of correlation,* Biometrika 13 (1920), 25–45; see also [PeE-K], 185–205.
[Pl1] Plackett, R. L., *The principle of the arithmetic mean,* Biometrika 45 (1958), 130–135; see also [PeE-K], 21–126.
[Pl2] Plackett, R. L., *The discovery of the method of least squares,* Biometrika 59 (1972), 239–251; see also [K-Pl], 279–291.
[Pl3] Plackett, R. L, *Some theorems in least squares,* Biometrika 37 (1950), 149–157.
[Se] Seal, H. L., *The historical development of the Gauss linear model,* Biometrika 54 (1967), 1–24; see also [PeE-K], 207–230.
[Sh] Sheynin, O. B., *Origin of the theory of errors, Nature, Lond.* 211 (1966), 1003–1004.
[S] Smith, G.C., *Thomas Bayes and fluxions,* Hist. Math. 7 (1980), 379–388.
[St1] Stigler, S. M., *Laplace, Fisher, and the discovery of the concept of sufficiency,* Biometrika 60 (1973), 439–445; see also [K-Pl], 271–277.
[St2] Stigler, S. M., The History of Statistics: The Measurement of Uncertainty before 1900, the Belknap Press of Harvard University Press, Cambridge, MA, 1986.
[St3] Stigler, S. M., American Contributions to Mathematical Statistics in the Nineteenth Century (2 vols.), New York: Arno Press, 1980.