2003

# Gauss' method of least squares: an historically-based introduction

Belinda B. Brand
*Louisiana State University and Agricultural and Mechanical College*, bbrand@lsu.edu

# GAUSS' METHOD OF LEAST SQUARES: AN HISTORICALLY-BASED INTRODUCTION

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

in

The Department of Mathematics

by
Belinda B. Brand
B.S. in Education, L.S.U., 1973
August 2003

# Acknowledgments

This work was motivated by the work of Gauss brought to my attention by Professor James J. Madden, Ph.D. It is a pleasure to thank Dr. Madden for being generous with his time and guidance.

This thesis is dedicated to my dear husband Jude, my mother Gwen Bruton, and my son Steve Massengale for their unbounded support and encouragement, and for giving me the courage to take the road less traveled.

# Table of Contents

# List of Figures

# Abstract

This work presents Gauss' justification of the method of least squares, following the treatment given by Gauss himself in *Theoria Combinationis Observationum Erroribus Minimus Obnoxiae,* where the main idea is to show that the least squares estimate is the unbiased linear estimate of minimum variance. (Actually, we present Gauss' argument both in his terminology and translated into matrix terminology.) We show how this contrasts with Gauss' earlier justification in *Theoria Motus Corporum Coelestium*, which was based on the assumption of a normal distribution of errors and yielded the estimate of maximum likelihood. We present as a background the development from scratch of all the probability theory needed, albeit we have not treated explicitly all the needed measure theory.

# Chapter 1. Introduction

Science is the study of natural phenomena. We were all taught this definition, or something similar to it in school. We were also cautioned that there were very few "scientific truths." If we were very lucky, we were taught the methods that scientists use to study the world around us in a laboratory class. Part of the scientific method that we learn is to design an experiment, and collect data. Then, we are told that the collection of data is a process inherently full of error. So if there are few "truths" and our methods of study are rife with error, how does a scientist make any intelligent conclusions about reality?

It is an empirical fact that errors, in many cases, are distributed according to some simple law. It is the study of error and its relationship to this distribution curve that leads a scientist to some certainty, more exactly, a high probability of certainty of these observations. We can, with certainty, describe what will happen when we drop a pencil. It will fall to the earth, or the nearest surface. Can we, with absolute certainty describe the speed at which it will fall, or the exact time it takes to fall a given distance? No. Can we with a high probability of success describe the speed and time of the fall? Yes, we can, and using modern methods of statistics, we can also predict with high probability the speed and time of the next fall of the pencil from the same height. If we are able to measure these two quantities as accurately as possible, then we can hope for a small difference between the observations we make and the predicted values of what we measure. The actual error may be large or small depending on whether the wind blows, or if conditions are nearly perfect. These events point to the random nature of errors in measurements.

The treatment of errors as random variables was a major step forward in bringing the study of probability away from games of chance toward the study of quantitative inference. Simpson and Bayes, two eighteenth-century mathematicians, began this transition in thinking by "studying the errors of observations rather than the observations themselves" [16, pg. 100]. It was Laplace who revolutionized the study of applying probability to inference which associates the set of errors to a distribution curve. The curve may then be used to infer the value of a variable whose direct measurement is impossible.

The concept is a simple one, but it was, nonetheless, groundbreaking. Let $e$ be an error, $P$ the point observed, and $O$ the observation. It is easy to surmise that if $O = P + e$, then $P = O - e$. If one fixes the value of $P$, and allows $e$ to be a random variable, then $e = O - P$. From this assumption, $e$ may be linked to a probability function, and the value of $P$ may be inferred from minimizing $e$ [16, pg. 101].

In this survey, the elements of randomness and probability theory will be explored, and special emphasis will be given to those topics related to the method of least squares developed by Gauss in the late eighteenth century. The first part provides a review of the basic language of probability theory. The second is devoted to Gauss' theory of errors. Finally, the Theory of Minimum Variance as Gauss presented it will be summarized, and a modern treatment of his justification will be considered.

# Chapter 2. Probability Theory

The mathematical meaning of probability is a measure of sets in an abstract space of events. This idea was established by Andrei Kolmogorov who wrote a definitive work on the mathematical treatment of probability. It is essential in applying probability to real-life situations that the space of events be identified in great detail for the problem being studied. Prior to Kolmogorov's treatment of probability, it was common to collect a large amount of data then determine the probability distribution. There are many problems inherent in this approach, not the least of which is its inaccuracy when applied to statistical analysis. For instance, it is common when applying statistical models to observational data to have more than one mathematical model "fit the data" [15, pg. 293].

It is the assignment of a particular probability distribution obtained from the set of outcomes with "sufficient exactitude" that allows us to calculate the probabilities needed to do statistical analysis of the data of an experiment [15, pg. 302]. R.A. Fisher proposed making the sample space to be the set of all permutations of random assignments that could be made between experimental subjects, thereby making all the events equally probable. The use of computers has made it possible for the use of randomized controlled experiments to become the standard in scientific inquiry [15, pg. 303].

Essential to probability theory is the concept of a random experiment. In simple terms, a random experiment is one whose outcome is uncertain. The more precise definition of a random experiment is one consisting of the following three elements, each of which we will consider in more detail.

1. **A sample space** $S$**:** *the set of all outcomes of a random experiment*

2. **An event space $E$:** *the set of all events that can occur when the experiment is performed. Each event is a subset of the sample space.*

3. **A probability measure,** $P(\cdot) : E \longmapsto \mathbb{R}$ : *a function that assigns to each event $A \in E$ a real number called the "probability" of the event. This function must satisfy certain axioms discussed later.*

In the next sections, we will examine these essential elements of random experiments. We will then look at the concept of a probability space.

## 2.1 The Sample Space

The discussion in this section is informal. See Section 1.4 where we introduce the precise, logical framework that we will be using in subsequent considerations.

The concept of a sample space is a very important one when considering random experiments. The sample space should be exhaustive in its description, and the elements of the sample space should be mutually exclusive. Each element of the sample space has a conceptual meaning typically corresponding to the most detailed and specific description of an outcome that one would ever want in the experiment. For example, if an experiment is designed to study the toss of a single coin, then most people would characterize the sample space to be $S = \{H, T\}$, where $H$ corresponds to a head, and $T$ corresponds to a tail. However, this may not be the only way to define the sample space depending on the way the experiment is defined. A more thorough look at the experiment might convince someone to include more outcomes of this simple experiment, such as $S = \{H, T, \text{ the coin lands on its edge, anything else happens to the coin}\}$. It is important to note that the sample space for any experiment may be defined in a number of ways, as the next example will show.

**Example 2.1.1.** An experiment is done to study the toss of a pair of dice. What is chosen to use as a sample space will depend on the purposes and intent of the experiment. Suppose a student simply wishes to see if the total number of spots that land face-up on a throw of the dice are equally likely. The student plans to gather evidence by rolling the dice repeatedly and tallying the outcomes. A suitable sample space for this experiment would be $S_1 = \{2, 3, 4, 5, \cdots, 11, 12\}$.

Another student might design an experiment to determine the same thing, but defines the sample space to be all the combinations $\{a, b\}$ of the number of dots on each die without any designation of each die as "first" or "second." For example, in this sample space, there is no distinction between $\{1, 2\}$ and $\{2, 1\}$ since each of these elements are themselves sets. This sample space would look like the following:

$$S_2 = \{\{1,1\},\{1,2\},\{1,3\},\cdots,\{1,6\},\{2,2\},\{2,3\},\cdots,\{2,6\},$$
$$\{3,3\},\{3,4\},\{3,5\},\{3,6\},\{4,4\},\{4,5\},\{4,6\},\{5,5\},\{5,6\},\{6,6\}\}.$$

A third student designs an experiment using two different colors of dice so that he is able to define his ordered pair as $(a, b)$ where $a$ is the number of dots on a blue die, and $b$ is the number of dots on a red die. This sample space would have all 36 combinations of ordered pairs of the numbers 1 through 6 and would be defined as follows:

$$S_3 = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),\cdots\}.$$

A fourth, and most thoughtful student, decides to define the experiment in a similar way as Student 2, in that the sample space makes no distinction as to "first die" or "second die". In addition, this student adds a rule in the design of the experiment that both die must land on the coffee table in order to be counted.

The addition of this rule might expand the sample space $S_2$ to include events such as $A_1 =$ one die falls off the table or $A_2 =$ both dice fall off the table. □

It is apparent from this example that the sample space depends on the purpose and intent of the experimenter since it is a "mental model" for the experiment itself. All of the examples previously discussed in this section have been examples of finite sample spaces. There are many examples of infinite sample spaces. We will look at three.

**Example 2.1.2.** There is a floor tiled with square tiles. A student tosses a coin onto the floor. There are an infinite number of positions the coin could take on the floor, or even on a single tile of the floor. There are an infinite number of elements in this sample space. □

**Example 2.1.3.** *The Game of Pig*: In the game of Pig, two players take turns tossing a single six-sided die. Each player repeatedly rolls the die until a 1 is rolled, or the player calls a hold. If the player rolls a 1, his turn is lost, and he scores no points. If a player rolls a 2, 3, 4, 5, or 6, the points are added to the total for his turn, and his turn continues. If the player calls a hold before a 1 is rolled, the points for that turn are added to the player's total. The first player to reach a score of 100 wins the game.

If the experiment is defined as the length of a move or turn in this game, the sample space is infinite, and very easy to describe. If a player rolls a 1 on his first roll, then we will assign a 1 as an element of the sample space. If the player rolls a 1 on his second roll of the die, then there is an element of 2 in the sample space. This could continue on to infinity, so the sample space $S = \{1, 2, 3, \cdots\}$. □

**Example 2.1.4.** *Buffon's Needle Problem*: A floor is marked with equidistant parallel lines (planks). A needle is tossed onto the floor. The experiment is designed

to study whether the needle lands on one of the parallel lines of the floor. We wish to simplify the problem. Let's assume that each of the parallel lines, $l_1$ and $l_2$ are of width $h$ apart, and the needle has length $d$. The needle could land perfectly parallel to the parallel lines on the floor, perfectly perpendicular to the lines, or at some angle $\theta$ to the lines as shown in the figure.

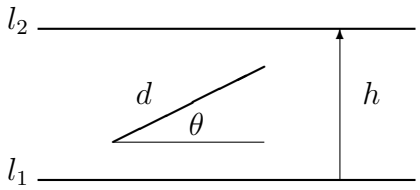



FIGURE 2.1. Buffon's Needle Problem

There are several choices for the sample space of this problem. The natural choice of sample spaces would be the finest, that one consisting of the exact location of the needle in terms of the perpendicular distance from each wall, $d_1$ and $d_2$ along with the angle $\theta$ as shown in the figure. The elements of this sample space would look like the ordered triple $(d_1, d_2, \theta)$.

We can simplify this sample space by arguing that the distance from a wall perpendicular to $l_1$ in the figure is not needed to describe the position of the needle within two parallel lines. This would eliminate the need for the quantity $d_2$, and would cause the elements of this sample space to consist of the ordered pairs $(d_1, \theta)$.

If we only wish to find the probability that the needle crosses one of the parallel lines on the floor, it isn't necessary to find the perpendicular distance of the point of the needle from a parallel wall, but only the distance of the needle from a parallel line. If we take this distance to be $s$, then the sample space need only consist of the ordered pair $(s, \theta)$.

Finally, once we realize that whether the needle crosses a parallel line on the floor consists only of its "effective height" described using the expression $d \sin \theta$

where $d$ is the length of the needle, then the sample space can consist of the possible values of $\theta$. Now, it is evident that the orientation of the needle would be the same whether the angle is measured from the line $l_1$ or $l_2$. The sample space now consists of $S = \{\theta, 0 \leq \theta \leq \pi\}$, an infinite sample space. □

The examples above not only allow us to see the relationship between sample spaces and the experiment, but they also gives us some insight into how the choice of sample spaces can ease the solution to a problem in probability. For instance, when trying to determine whether the possible sums that result on the toss of two die are equally likely, the probability of each sum is easily determined by using the finest sample space, the one in which each element in the sample space is equally likely. In Buffon's Needle problem, the solution is simplified by taking the coarsest sample space.

In summary, the important aspects of a sample space are the following:

- The sample space is a mental or conceptual model of the results of a random experiment.
- Picking the sample space sets the terms of the rest of the experiment. One may always choose to think about an outcome in more general terms than that defined by the sample space since that outcome will still be a partition of the sample space.
- One should choose the sample space that will best serve the intent of the experiment.

## 2.2 The Event Space

The event space consists of all subsets of the sample space $S$. As such, it contains the empty set, the entire sample space itself, countable intersections,

unions, and complements of all events. The enumeration of all the elements of the event space is a lengthy process, and is not an exercise that is necessary in this context. For example, the event space for the roll of a single die consists of $2^6 = 64$ elements. It is much more helpful to describe "an event". An event can best be described as a *property* that an outcome might or might not have. An example of an event when throwing two dice would be "throwing an even number of spots," a well-described property that one might study. It is important to remember that an event is *not* the result of a single trial of an experiment, but the collection of favorable results that one defines before the experiment is done.

## 2.3 Probability Measure

A probability measure is a function $P(\cdot) : E \to R$ that assigns a real value to an event $A$ representing the probability of the occurrence of that event. For a finite sample space $S = \{s_1, s_2, \cdots, s_n\}$, a space with $n$ equally likely elements, the probability of an event $A \in E$ occurring is defined by $P(A) = \frac{|A|}{|S|} = \frac{|A|}{n}$ where $|A|$ is the number of elements of $A$. Going back to the example of flipping a single coin, the number of elements in the sample space is four, written $|S| = 4$, and consists of the set $S = \{\emptyset, H, T, H \text{ or } T\}$. Most people would consider $P(\emptyset) = 0$ since the only ways one would not flip a head or a tail would be for the coin to land on its edge, a very unlikely occurrence, or if one lost the coin somehow in the process of flipping. Experiments show that $P(H) \approx .50$, and likewise for $P(T)$. We also know that $P(H \text{ or } T) \approx 1$. This simple example may be useful in understanding the three axioms that must be satisfied by a probability measure.

1. $0 \leq P(A) \leq 1$ for any event $A \in S$.

2. $P(S) = 1$ also written $\sum_{A \in S} P(A) = 1$. In simple terms, the probability of the entire sample space is equal to 100%.

3. If $A_1, A_2, ...$ are disjoint events, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$, meaning the probability of the countable union of any number of disjoint events is the sum of the probabilities of the individual events.

Some useful properties of probability measures are results of the three axioms given below [12, pg. 3].

1. $P(A^c) = 1 - \Pr(A)$.
2. $P(A \cap B) \leq \min(\Pr(A), \Pr(B))$.
3. $P(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Let's revisit some of the previous examples for the infinite sample spaces.

**Example 2.3.1.** *Throwing a coin onto a square-tiled floor:* We want to calculate the probability of a coin thrown onto a floor made of square tiles landing totally within a tile. Let the length of a side of the square be $s$ and the radius of the coin be $r$. The experiment would not make sense for any values of the variables for which $s \leq r$ so we will make the definition that $s > r$. The coin will land totally within the boundaries of the tile if the center of the coin lands within a smaller area of the tile with side equal to $s - 2r$. The probability that the coin will land totally with the tile is expressed as $P = \frac{(s-2r)^2}{s^2}$ This expression will be defined everywhere since we specified that $s > r$. □

**Example 2.3.2.** *Buffon's needle problem* We will use the variables as defined in Example 4 where $h$ is the width of the "planks" of the floor, and $d$ is the length of the needle. If the needle lands with $\theta = 0$, then the probability $P = 0$ of the needle landing on a line. If the needle lands with $\theta = \frac{\pi}{2}$, then $P = 1 - \frac{h-d}{h} = \frac{d}{h}$. When the needle lands at a particular angle other than 0 or $\frac{\pi}{2}$, then the effective height

(perpendicular to the parallel line) is $d \sin\theta$. The probability of the needle landing on a line can now be expressed as $P = \frac{d \sin\theta}{h} = \sin\theta \cdot \frac{d}{h}$. for a particular value $\theta$. Since the function $f(x) = \sin x$ is continuous everywhere, this probability measure is defined at all particular values $0 \leq x \leq \pi$. The derivation of the average value of $f(x) = \sin x$ in the interval $[0, \pi]$. is as follows:

$$\begin{aligned}
\text{Average value} &= \frac{1}{\pi - 0} \int_0^{\pi} \sin x \, dx \\
&= \frac{2}{\pi - 0} \int_0^{\frac{\pi}{2}} \sin x \, dx \\
&= \frac{2}{\pi}(-\cos x \,|_0^{\frac{\pi}{2}}) \\
&= \frac{2}{\pi}(0 + 1) = \frac{2}{\pi}.
\end{aligned}$$

The total probability that the needle lands on a line is

$$\begin{aligned}
P = \text{the average value of } \sin x \cdot \frac{d}{h} \\
= \frac{2}{\pi} \cdot \frac{d}{h}
\end{aligned}$$

for all $x$ in the interval $[0, \pi]$. □

## 2.4 Probability Spaces, Subspaces, and Product Spaces

In this section, we want to give a more formal treatment to the relationship between sample spaces, event spaces and probability distributions. These three elements are essential to a probability space whose definition is below.

**Definition 2.4.1.** A probability space is a set $S$, equipped with measure $\mu$ such that $\mu(S) = 1$.[1]

[1] Using measure theory, this space is a triple $(S, E, \mu)$ consisting of the sample space $S$ of outcomes, a $\sigma$-algebra $E$ of sets in $S$, and the probability measure $\mu$ on $(S, E)$ with axioms and conditions mentioned in the previous sections.

Any experiment will have an associated probability space. Specific events will be described in the probability space and each event will have an associated probability measure. These descriptions will be adapted to the individual experiment as discussed earlier.

There are some interesting operations on a probability space that we will consider. These operations are those which form subspaces, partitions and products of the sample space $S$. First, we will consider the formation of subspaces.

Subspaces of $S$ are formed from imposing a *condition* on a sample space. Conditional probability is used to find the probability of an event $A$ occurring once event $B$ has taken place. This is denoted by $P(A|B)$, read "the probability of $A$ given $B$," with the definition given below [8, pg. 8].

**Definition 2.4.2.** If $P(B) > 0$ then the *conditional probability* that $A$ occurs given that $B$ occurs is defined to be

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

It is easy to see how a subspace is formed with conditional probability using the discrete example of the throw of two dice. We will set the condition $B$ as the "throw of a sum of five dots." Next, we will define the sample space as $S = \{(1,4),(2,3),(3,2),(4,1)\}$, the ways in which a value of five can be thrown on two standard dice. In other words, a new sample space has been formed from the original sample space which we described earlier in Example 2.1.1. The condition we imposed induced this new sample space with new probability distributions. To determine the probability in this new sample space that a 1 and a 4 are thrown, it is easy to see that this probability is $1/2$.

Now, we will turn our attention to partitions of the sample space.

**Definition 2.4.3.** *A* partition *of a space $S$, such as the sample space, consists of a set of non-empty subsets $\{A, B, C, \cdots\}$ satisfying the following properties:*

- $A \cup B \cup C \cup \cdots = S$
- *the intersection of every pair of distinct subsets of $S$ is empty.*

It is an *equivalence relation* that partitions the sample space, and every partition has an associated equivalence relation. Simply, an equivalence relation $\sim$ is an operation on $S$ having the reflexive, symmetric, and transitive properties. If an equivalence relation $\sim$ is determined by a function, $f$, then for elements of the sample space, $a_1$ and $a_2$, $a_1 \sim a_2 \Leftrightarrow f(a_1) = f(a_2)$.

The following example will give us a systematic way of seeing the relationships between the first three sample spaces in Example 1. Let's look at a partition of a sample space consisting of the fair throws of two distinct die each with three faces:

**Example 2.4.4.** The sample space $S = \begin{bmatrix} (1,1),(1,2),(1,3) \\ (2,1),(2,2),(2,3) \\ (3,1),(3,2),(3,3) \end{bmatrix}$. We will define a function $F$ as follows:

$$\begin{bmatrix} (1,1),(1,2),(1,3) \\ (2,1),(2,2),(2,3) \\ (3,1),(3,2),(3,3) \end{bmatrix} \xrightarrow{F} \begin{bmatrix} \{1\},\{2\},\{3\} \\ \{1,2\},\{2,3\} \\ \{1,3\} \end{bmatrix}.$$

The function $F$ maps the ordered pairs of the sample space $S$ to the sets corresponding to these pairs. For example, the ordered pairs $(2,3)$ and $(3,2)$ both map to the set $\{2,3\}$.

Effectively, the function $F$ partitions all the combinations of the sample space into permutations of $S$. We will then define a function $G$ that maps the results of

$F$ as follows:

$$S = \begin{bmatrix} (1,1),(1,2),(1,3) \\ (2,1),(2,2),(2,3) \\ (3,1),(3,2),(3,3) \end{bmatrix} \overset{F}{\rightarrow} F(S) = \begin{bmatrix} \{1\},\{2\},\{3\} \\ \{1,2\},\{2,3\} \\ \{1,3\} \end{bmatrix} \overset{G}{\rightarrow} \begin{bmatrix} 2,3,4,5,6 \end{bmatrix}.$$

The function $G$ partitions the elements of $F(S)$ into the sums of the elements of the sample space. It is important to note that the sample space $S$ consists of equally likely outcomes while $F(S)$ consists of sets that are not equally likely outcomes of the experiment. It should be evident that the elements of $G(F(S))$ are not equally likely outcomes of the experiment since there are more than one pair of numbers whose sums are 3, 4, or 5 in this experiment. This fact will be important in our discussions of probability distribution functions that will be explored in later sections. □

The product space is an important concept since it would be a factor in experiments involving repeated trials. An example this idea is the repeated throw of a single three-faced die similar to that described in the above section. The sample space would consist of all the possible ordered pairs formed by the Cartesian cross $\{1,2,3\} \times \{1,2,3\}$. The probabilities of each of the outcomes would be associated with the combinatorial product described by $\mu(A_1 \times A_2) = \mu(A_1)\mu(A_2)$, where $A_1$ would be the results of the first throw and $A_2$ would be the results of the second throw [5, pg. 118]. In general, the result of $n$ trials would be a point in $\mathbb{R}^n$-space with probability measure equal to the product of the probabilities of the individual $n$ trials.

## 2.5 Random Variables

We begin this section with a definition of a random variable, then continue the discussion by investigating the types of random variables along with some examples.

**Definition 2.5.1.** A random variable $X$ is a measurable [2] function of elementary events on a sample space $S$ [4, pg. 199].

Random variables may be of two types:

1. Discrete random variables: random variables $X$ having at most a countable number of possible values. Recall that some infinite sets are countable, for example, $\mathbb{N}$, the set of natural numbers, and $\mathbb{Q}$, the set of rational numbers.

2. Continuous random variables: random variables $X$ having a continuous probability density function at every real number $x$. This statement implies that for random variable $X$ and real numbers $a, b$ for which $a < b$, $P(a \leq X \leq b) = P(a < X < b)$, and $P(X = x) = 0$ for all real numbers $x$.

A random variable $X : S \to \mathbb{R}$ on a discrete sample space $S$ maps the sample space $S$ to the real number line so that for every $s \in S, X(s) \in \mathbb{R}$ [11, pg. 18]. The random variable $X$ determines a probability space in which the set of real numbers itself is the sample space. The real numbers themselves are obviously elements of this sample space, and intervals of the real number line are events [2, pg. 35].

By defining a variable $X$ as a random variable, the study of the outcomes of an experiment may be redefined according to the real number system rather than the sample space itself. In Example 1 referring to the toss of a single die,

[2] The measureability requirement is a technical requirement that ensures that the integrals mentioned all make sense. In this thesis, we are not going to devote attention to determining the most general measure-theoretical assumptions that suffice to assure logical consistency. Like Gauss, we deal with probability distributions whose continuity can generally be assumed.

the number of spots on each face of the die corresponds to a real number. This means $X(\text{one spot}) = 1$, $X(\text{two spots}) = 2$, and the sample space becomes $S = \{1, 2, 3, 4, 5, 6\}$.

**Example 2.5.2.** If an experiment is done where a coin is tossed until the first head comes up, then the sample space is $S = \{\emptyset, H, TH, TTH, TTTH, ...\}$. Then the random variable $X$ may be defined as the number of tails that are tossed until the first head is tossed. In this way, $X(H) = 0$, $X(TH) = 1$, $X(TTH) = 2$, and so on [12, pg. 16]. □

We will consider functions of random variables and how to describe them. For $X$ and $Y$ with density functions $f_X(x)$ and $f_Y(y)$ respectively, we wish to describe in what sense $X + Y$ is a random variable. Let $S_1$ be the sample space for $X$, and $S_2$ be the sample space for $Y$. Since $X$ and $Y$ are measurable functions on $S_1$ and $S_2$, respectively, then $X + Y$ is a measurable function on $S_1 \times S_2$. Thus, $X + Y$ is a new random variable [5, pg 114]. In fact, for any continuous function $U$ of a finitely-many random variables, a new random variable is determined [5, pg. 114]. With this in mind, it is clear that for a real constant $\lambda$, the function $U = \lambda X$ is also a random variable.

## 2.6 Probability Functions and Random Variables

Although we already defined the sample space as a measurable space with measure 1, probability theory has a specialized language for talking about the measures. This section discusses some of the typical formulations.

The probability function of random variables is known by many other names. Some common ones are probability mass function, distribution function, frequency function, and density function. Most texts reserve the term "density function"

when working with continuous random variables [12, pg. 19]. The definitions of the frequency function are different for discrete and continuous random variables.

### 2.6.1 Probability Functions of Discrete Random Variables

For discrete random variables, the frequency function is $\phi(x) = P(X = x)$, where $x = a_1, a_2, \cdots$ for some real-valued $a_i$. We let $p_i$ denote the probability that $X$ assumes the value $a_i$. A graphical representation of the probability distribution of $X$ can be obtained in the Cartesian coordinate plane by marking the points $(a_i, p_i)$. If one connects the abscissa of each of the points with a vertical segment of the length of the ordinate to the point itself, this representation would constitute a graph of the probability function [3] $\phi(x)$ as shown below [2, pg. 37].

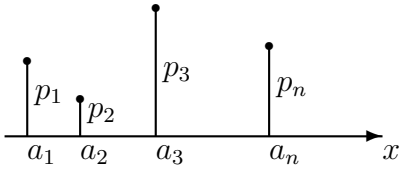



FIGURE 2.2. Probability Distribution of Discrete Random Variable

Recall that every real number $x$ is an elementary event in this probability space induced by the random variable $X$. It is important to describe the probability function $\phi(x)$ when $x \neq a_1, a_2, \cdots$. The event described earlier as $\{a < X < b\}$ is the union of all disjoint events in which $X = a_i$ for all values of $i = 1, 2, \cdots$ in which $\{a < a_i < b\}$. Since this event is the probability of a union of subevents, then $P(a < X < b) = \sum_{a<x<b} \phi(x)$. This sum can be represented by the sum of all the heights of the vertical lines between chosen values of $a$ and $b$ on the graph of $\phi(x)$ in the figure above [2, pg. 38].

[3] If the sample space is a discrete subset of $\mathbb{R}$, then the probability function $\phi(x)$ is actually a probability measure.

### 2.6.2 Probability Functions of Continuous Random Variables

Recall that a random variable $X$ determines a probability space on $\mathbb{R}$. We will define a continuous probability space as follows [11, pg. 46].

**Definition 2.6.1.** A continuous probability space in $\mathbb{R}$ consists of a sample space $S$ and a function $f(x) : S \to \mathbb{R}$ such that

$$f(x) \geq 0 \text{ for each } x \in S \text{ and } \int_S f(x)\,dx = 1.$$

A probability density function will satisfy both these conditions over the whole sample space. This function $f$ determines a distribution on fixed interval as follows [11, pg. 160].

**Definition 2.6.2.** A density function on a continuous random variable $X$ is a function with the property that $P\{a < X < b\} = \int_b^a f(x)\,dx$.

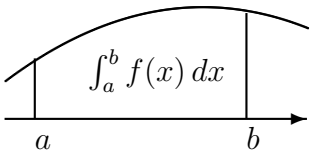



FIGURE 2.3. Density Function on a Continuous Random Variable

The concept of the integral in this definition is analogous to a classic Riemann sum, and the area under the curve $\phi(x)$ is the probability that the continuous random variable has a value in the interval $(a, b)$. The definition of continuous random variables tells us that the probability of $X$ assuming a particular real value $x$ is zero. From this, we can see that we might have for every point $x \in S$, $P(x) = 0$, but will still have $P(A) \neq 0$ for some event A.

### 2.6.3 Cumulative Distribution Function

The concept of a cumulative distribution function is one that we will use in subsequent proofs. We will begin with the definition, and then explore the relationship between a density function and its cumulative distribution function.

**Definition 2.6.3.** The cumulative distribution function, $F(x)$ on $\mathbb{R}$ is defined by

$$F(x) = P(-\infty, x] = P(X \leq x),$$

that is, the accumulated probability of all numbers less than or equal to x [11, pg. 156].

Clearly, as the value of $x$ increases, the probability that $X \leq x$ grows continually. For a discrete random variable having $\mathbb{R}$ as its sample space, the graph of the distribution function is a collection of vertical lines at each $x$ with non-zero probability. The cumulative distribution function for this discrete case would be a step function.

The case we need to explore more fully is that of the continuous probability space. We need to know the relationship between a continuous density function and its cumulative distribution function. Simply, this relationship is given by the expression below [11, pg. 160].

$$F(x) = P(-\infty, x] = \int_{-\infty}^{x} f(t)\, dt.$$

The similarity of this relationship to the Fundamental Theorem of Calculus is readily noticeable. The Fundamental Theorem requires that $f$ be continuous within a bounded interval. What is given here as an improper integral can be an extension of the theorem when we require the integral of $f$ over all of $\mathbb{R}$ to be finite, which, by definition, it is. Also, we need to address the fact that not all density functions are continuous, and would not be differentiable at every point. So, we will let $F(x)$

be the cumulative distribution function of the continuous distribution with density function $f(x)$. Then for all values $x$ where $f$ is continuous, then [9, pg. 61]

$$\frac{d}{dx}F(x) = f(x).$$

At worst, the cumulative distribution function $F$ is differentiable except at a small number of values of $x$.[4] The function $f$ is also the integral of its derivative, as shown in the definition [11, pg 161].

### 2.6.4 Joint Density and Cumulative Distribution Functions

Recall the fact that measurable functions of random variables are themselves considered random variables [cf. Section 2.5]. Let $X$ and $Y$ be continuous random variables. For these random variables, consider the joint random variable $\overline{X} = (X, Y)$, with density function $f_{\overline{X}}(x, y)$ on $\mathbb{R}^2$ satisfying

$$\int_a^b \int_c^d f_{\overline{X}}(x, y)\, dy\, dx = P(a \leq X \leq b, c \leq Y \leq d). \tag{2.6.1}$$

When $X$ and $Y$ are independent random variables, the joint distribution function is easy to describe. In fact, $X$ and $Y$ are independent if and only if the following holds:

$$P(a \leq x \leq b, c \leq y \leq d) = P(a \leq x \leq b) \cdot P(c \leq y \leq d),$$

and the joint distribution function

$$f_{\overline{X}}(x, y) = f_X(x) \cdot f_Y(y).$$

In other words, the joint distribution function for $\overline{X}$ is the product of the individual distribution functions for $X$ and $Y$ [9, pg. 143]. Now, we define the *joint cumulative distribution function* [9, pg. 165].

[4] Recall from elementary analysis that a function $f$ is Riemann integrable even though $f$ may have a countable number of discontinuities.

**Definition 2.6.4.** Let $X$ and $Y$ be continuous, independent random variables, and let $\overline{X} = (X, Y)$. Then the *joint cumulative distribution function* of $\overline{X}$ is defined by $F(x, y) = P(X < x, Y \leq y)$, and satisfies the equations

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(t_1, t_2)\, dt_1\, dt_2,$$

where $f = f_{\overline{X}}$, and

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

Let $Z = X + Y$. We know from above that $Z$ is a random variable. We define a function $F(z) = P(X + Y \leq z)$ as the cumulative density function for $z$. For any real number $z$, we know $F(z)$ is the probability that $(x, y)$ lies below the line $x + y = z$. $F(z)$ is given by the integral

$$F(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y)\, dy\, dx.$$

If $X$ and $Y$ are independent, then

$$\begin{aligned} F(z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y)\, dy\, dx \\ &= \int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{z-x} f_Y(y)\, dy\, dx \\ &= \int_{-\infty}^{\infty} f_X(x) F_Y(z - x)\, dx, \end{aligned}$$

where $F_Y$ is the cumulative distribution function for $Y$. By the above definition, when we differentiate $F$ with respect to $z$, then we will obtain the density function $f_Z(z)$ [11, pg. 318].

$$\begin{aligned} F'(z) &= \frac{d}{dz} \int_{-\infty}^{\infty} f_X(x) \cdot F_Y(z - x)\, dx \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx. \end{aligned}$$

This is the formula for the convolution, $f_X * f_Y$. Therefore, the density function of $X + Y$ is the convolution of $f_X$ and $f_Y$. By completing the proof once again for

$x = z - y$, we have an equivalent expression in terms of $y$ given below:

$$f_X * f_Y = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)dx = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)dy.$$

## 2.7 Expected Value

When presented with a large set of data, we are not usually interested in individual values found in the set. It is important, however, to describe the trends found in the set of numbers [9, pg. 225]. One such description is *expected value.*

### 2.7.1 The Discrete and Continuous Cases

**Definition 2.7.1.** Let $X$ be a discrete random variable having a sample space $S$ and a probability function $\phi(x)$. The *expected value*, $E(X)$ is defined as follows:

$$\mu = E(X) = \sum_{x \in S} x\, \phi(x),$$

provided $\sum_{x \in S} |x| \phi(x)$ exists [9, pg. 226]. If this sum does not converge absolutely, then $X$ has no expected value. The expected value is often called the *mean,* and is frequently denoted $\mu$ when the reference is unambiguous.

**Example 2.7.2.** Let's consider the expected value of three tosses of a single coin. We will define the random variable $X$ to be the number of tails that result from the three tosses. The possible values of $X$ are 0, 1, 2, and 3. We have the "tree" of possibilities shown in the figure.

We need to determine the theoretical frequency of each value of $X$. When $X = 0$, no tails are tossed. Using the tree above, we can see this happens with a frequency of $\frac{1}{8}$. For $X = 1$, there are three paths through the tree in which only one tail is tossed, and the frequency of this value is $\frac{3}{8}$, and likewise for $X = 2$. The frequency for $X = 3$ is $\frac{1}{8}$. Applying the definition of expected value, we have the calculation $\mu = E(X) = 0\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) = \frac{3}{2}$. □

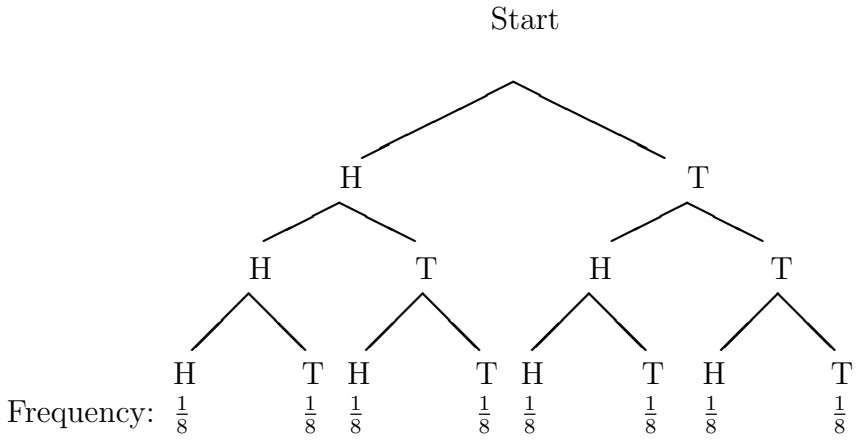



FIGURE 2.4. Probability Tree for Three Tosses of a Coin

We now recall the definition of expected value of a function of a random variable which is a generalization of this idea. Let $X$ be a discrete random variable with sample space $S$ and a distribution function $\phi(x)$. If $f : S \mapsto \mathbb{R}$ is a function on $S$, then the expected value of $E(f(X))$ is defined as follows

$$E(f) = \sum_{x \in S} f(x)\phi(x),$$

provided the sum converges absolutely [9, pg. 230]. It is easy to see that in this case, the expected value is defined in a manner similar as before. To restate this idea, we can say if $X$ and $Y$ are two random variables and $Y$ can be written as a function of $X$, then one can compute the expected value of $Y$ using the distribution function of $X$ [9, pg. 230].

As the definition below will demonstrate, the expected value of a continuous random variable is analogous to that for the discrete case.

**Definition 2.7.3.** Let $X$ be a real-valued continuous random variable with density function $\phi(x)$. The *expected value* $E(X)$ is defined by the expression

$$E(X) = \int_{-\infty}^{\infty} x\phi(x)\, dx,$$

provided the integral $\int_{-\infty}^{\infty} |x|\, f(x)\, dx$ is finite[9, pg. 268].

### 2.7.2 Additive and Multiplicative Properties of Expected Value

In this section, we demonstrate that $E$ behaves linearly when applied to linear functions of independent random variables. The relevance of this behavior is that Gauss assumes linearity of his approximation functions, or uses linear approximations for nonlinear functions. In addition, Gauss derives some properties of expected value in §12 and §13 using an argument involving implicit functions.[5] We will derive the important conclusions by different methods.

Let $X$ and $Y$ be random variables. The expected value of $X + Y$ with joint density function $\phi(x, y)$ is given by the following expression.

$$E(X + Y) = \int_{-\infty}^{\infty} (x + y)\phi(x, y)\, dx\, dy.$$

Now, we have an important lemma and three corollaries:

**Lemma 2.7.4.** Let the independent random variables $X$ and $Y$ have probability distribution functions $\phi_X(x)$ and $\phi_Y(y)$. Let $a$ and $b$ be real constants, and let $f$ and $g$ be functions of $X$ and $Y$. Then the expected value $E$ is determined by

$$E(af + bg) = aE(f) + bE(g).$$

[5]The justification Gauss gives for these arguments is not a rigorous proof, but merely an outline of the main ideas [7, pg. 222].

*Proof.* For the random variable $X + Y$ we have the joint density function $\phi(x, y)$. The expected value of $X + Y$ is given by

$$\begin{aligned}
E(af + bg) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(af(x) + bg(y)) \cdot \phi_X(x)\phi_Y(y)dy\,dx \\
&= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)\phi_X(x)\phi_Y(y)dy\,dx + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y)\phi_X(x)\phi_Y(y)dx\,dy \\
&= a\int_{-\infty}^{\infty} f(x)\phi_X(x)\left[\underbrace{\int_{-\infty}^{\infty}\phi_Y(y)dy}_{=1}\right]dx + b\int_{-\infty}^{\infty} g(y)\phi_Y(y)\left[\underbrace{\int_{-\infty}^{\infty}\phi_X(x)\,dx}_{=1}\right]dy \\
&= a\int_{-\infty}^{\infty} f(x)\cdot\phi_X(x)\,dx + b\int_{-\infty}^{\infty} g(y)\cdot\phi_Y(y)dy \\
&= aE(f) + bE(g).
\end{aligned}$$

□

**Corollary 2.7.5.** *If $X$ and $Y$ are independent, random variables, then*

$$E(aX + bY) = aE(X) + bE(Y).$$

*Proof.* Let $f$ and $g$ be the identity function. Then

$$E(af + bg) = E(aX + bY) = aE(X) + bE(Y).$$

□

It is easy to see that the results of this theorem can be extended to a finite number of random variables.

**Corollary 2.7.6.** *If $X_1, X_2, \cdots, X_n$ are random variables, and $\lambda_1, \lambda_2, \cdots, \lambda_n$ are real constants, then*

$$E(\sum_i \lambda_i x_i) = \sum_i \lambda_i E(x_i).$$

*Proof.* We offer a proof by induction. For $i = 1$, $E(\lambda_1 x_1) = \lambda_1 E(x_1)$ by Lemma 2.7.4. Assume the hypothesis for $i = 1, 2, \cdots, n$. We want to show it is true for $i = n + 1$.

$$E(\sum_{i=1}^{n+1} \lambda_i x_i) = E(\sum_{i=1}^{n} \lambda_i xi + \lambda_{n+1} x_{n+1}).$$

Again, by Lemma 2.7.4, we know this is

$$E(\sum_{i=1}^{n} \lambda_i x_i) + \lambda_{n+1} E(x_{n+1}).$$

From the induction hypothesis, we have

$$\sum_{i=1}^{n} \lambda_i E(x_i) + \lambda_{n+1} E(x_{n+1}) = \sum_{i=1}^{n+1} \lambda_i E(x_i).$$

□

**Corollary 2.7.7.** *Let the independent random variables $X$ and $Y$ have probability distribution functions $\phi_X(x)$ and $\phi_Y(y)$. Let $f$ and $g$ be functions of $X$ and $Y$. Then $E(f \cdot g) = E(f) \cdot E(g)$.*

*Proof.*

$$\begin{aligned} E(f \cdot g) &= \int\int f(x)g(y)\phi_X(x)\phi_Y(y)\,dx\,dy \\ &= \left(\int f(x)\phi_X(x)dx\right)\left(\int g(y)\phi_Y(y)dy\right) \\ &= E(f) \cdot E(g). \end{aligned}$$

□

Now, from Gauss' §14, there is another corollary:

**Corollary 2.7.8.** *Suppose $X_1, \cdots, X_k$ are independent random variables, and $f(X_1, \cdots, X_k)$ is a sum of terms of the form $aX_1^{n_1}X_2^{n_2}\cdots X_k^{n_k}$. Then $E(f)$ is the sum of the corresponding terms $aE(X_1^{n_1})E(X_2^{n_2})\cdots E(X_k^{n_k})$.*

*Proof.* This is an immediate consequence of the preceding corollaries. □

## 2.8 Variance and Standard Deviation

In this section, we will look at the important statistical properties of variance and standard deviation. As a concept, we can think about variance as the "dispersion" of a distribution of a random variable.

**Definition 2.8.1.** Let $X$ be a random variable, continuous or discrete, with $\mu = E(X)$ Then the *variance* of $X$, denoted $V(X)$ is

$$V(X) = E((x - \mu)^2),$$

where $\phi$ is the distribution function of $X$ [9, pg. 268], and $\mu = E(X)$ is the expected value of $X$.

**Definition 2.8.2.** The *standard deviation* of $X$, denoted by $SD(X)$ or by $\sigma$ is defined as

$$\sigma = SD(X) = \sqrt{V(X)}.$$

Note: Occasionally, the variance is denoted $\sigma^2$.

There are some useful properties of variance to be considered [9, pg. 272].

**Lemma 2.8.3.** If $X$ is any random variable with $E(X) = \mu$, then:

1. If $c$ is any constant, then $V(cX) = c^2V(X)$.

   *Proof.* By definition of $E$, we know $E(cX) = cE(X)$. Now, we have

   $$\begin{aligned} V(cX) &= E(cX - c\mu)^2) = E(c^2(X - \mu)^2) \\ &= c^2E(X - \mu)^2 = c^2V(X). \end{aligned}$$

   □

2. If $c$ is any constant, then $V(X + c) = V(X)$.

*Proof.* First, we find $E(X+c) = E(X) + E(c) = E(X) = \mu$. Next, we write

$$\begin{aligned} V(X+c) &= E(X+c-\mu)^2 \\ &= E(x^2 - 2\mu x + \mu^2) + \underbrace{E(2cx - 2c\mu + c^2)}_{\mu} \\ &= E((X-\mu)^2 = V(X). \end{aligned}$$

□

3. $V(X) = E(X^2) - \mu^2$

   *Proof.*

$$\begin{aligned} V(X) &= E((X-\mu)^2) \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu \underbrace{E(X)}_{0} + \mu^2 \\ &= E(X)^2 - \mu^2. \end{aligned}$$

   □

4. For any $a$ and $b$ (constants), $V(aX+b) = a^2V(X)$ and $SD(aX+b) = |a|\, SD(X)$.

   *Proof.* From item #2, we can write $V(aX+b) = V(aX)$. Then from item #1, we know $V(aX) = a^2V(X)$. The expression for the standard deviation follows from this result. □

**Example 2.8.4.** Let's once again consider the roll of a single die to illustrate the calculation of expected value, variance and standard deviation [9, pg. 257]. Let the

random variable $X$ be the number that turns up on the die, and each number is equally likely to turn up. The calculation of the expected value is

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{7}{2}.$$

We will use the table below to simplify our calculations of variance.

| $x$ | $m(x)$ | $x - \frac{7}{2}$ | $\left(x - \frac{7}{2}\right)^2$ |
|---|---|---|---|
| 1 | 1/6 | 5/2 | 25/4 |
| 2 | 1/6 | 3/2 | 9/4 |
| 3 | 1/6 | 1/2 | 1/4 |
| 4 | 1/6 | 1/2 | 1/4 |
| 5 | 1/6 | 3/2 | 9/4 |
| 6 | 1/6 | 5/2 | 25/4 |

FIGURE 2.5. Calculation of Variance

Now, we can easily calculate $V(X)$ as follows:

$$\sigma^2 = V(X) = \frac{1}{6}\left(\frac{25}{4} + \frac{9}{4} + \frac{1}{4} + \frac{1}{4} + \frac{9}{4} + \frac{25}{4}\right) = \frac{35}{12}.$$

It follows easily that the standard deviation, $\sigma$ is

$$\sigma = SD(X) = \sqrt{35/12} \approx 1.707.$$

□

Next, we will derive a lemma that plays a critical role in Gauss's 1820's justification of the method of least squares. Note here that Gauss uses the notation $m^2 = \int_{\mathbb{R}} x^2\phi(x)dx$

**Lemma 2.8.5.** Assume $x_1, x_2, \cdots, x_k$ are independent random variables each of which has mean value 0, *i.e.*, $E(x_i) = 0$. Let the distribution functions for the $x_i$'s be $\phi_i(x_i)$. Let $F : \mathbb{R}^k \mapsto \mathbb{R}$ be a linear function,

$$y = F(x_1, x_2, \cdots, x_k) = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k.$$

Then, the variance of $y$ is $V(y) = \sum_{i=1}^{k} \lambda_i^2 m_i^2$.

*Proof.* As we showed in Corollary 2.7.6 $E(F(x_1, x_2, \cdots, x_k)) = 0$. Now we calculate the variance of $y$. In the following set of equations, let $x = (x_1, x_2, \cdots, x_k)$, $\Phi(x) := \prod_{i=1}^{k} \phi_i(x_i)$, and $dx = \prod_{i=1}^{k} dx_i$.

$$\begin{aligned} V(y) &= \int_{\mathbb{R}^k} y^2 \phi_1(x_1)\phi_2(x_2)\cdots\phi_k(x_k)\, dx_1 \cdots dx_k \\ &= \int_{\mathbb{R}^k} (\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_k x_k)^2 \Phi(x)\, dx \\ &= \int_{\mathbb{R}^k} \left( \sum_{i=1}^{k} \lambda_i^2 x_i^2 + \sum_{j \neq i} \lambda_i \lambda_j x_i x_j \right) \Phi(x)\, dx. \end{aligned}$$

From earlier sections we know that

$$\int_{\mathbb{R}^k} \lambda_i^2 x_i^2 \Phi(x) dx = \underbrace{\left( \prod_{j \neq i} \int_{\mathbb{R}} \phi_j(x_j) dx_j \right)}_{1} \int_{\mathbb{R}} \lambda_i^2 x_i^2 \phi_i(x_i) dx_i = \lambda^2 m_i^2,$$

where $m_i^2 := \int_{\mathbb{R}} x_i^2\, \phi_i(x_i)\, dx_i$ and

$$\int_{\mathbb{R}^k} x_i x_j \Phi(x)\, dx = \underbrace{\int_{\mathbb{R}} x_i \phi_i x_i dx_i}_{0} \underbrace{\int_{\mathbb{R}} x_j \phi_j(x_j) dx_j}_{0} \prod_{l \neq i,j} \int_{\mathbb{R}} \phi_l(x_l) dx_l = 0.$$

Therefore

$$V(y) = \sum_{i=1}^{k} \lambda_i^2 m_i^2. \tag{2.8.1}$$

□

This result is used in Gauss' treatment of least squares. See the end of Section 3.1, and also see Section 4.4.

# Chapter 3. Gauss' Treatment of Error

The significance of the *Theoria Combinationis* goes beyond the Method of Least Squares. Gauss' treatment of errors as random variables is itself an important contribution to the study of mathematical statistics. The new concepts found in his work related to this treatment include [7, pg. 223]:

1. The treatment of errors as random variables.

2. A Chebyshev-like inequality

3. The convergence of the sample mean and the variance.

In this part, these concepts will be explored as Gauss presented them with some adaptations to modern notation and practice.

## 3.1 Treating Errors With Probability Theory

In the first eight sections of Gauss' work, he discusses the nature of error and its properties. In an historical sense the discussion is worth review. What is important in this part of the work is that he uses all the mechanisms from the previous summary of probability theory to describe errors. In §18, Gauss has a discussion of how to find the variance and standard deviation of a linear function of errors, a topic that will be important to the proof of the Theorem of Minimum Variance. We will consider this topic in the latter part of this section.

Gauss uses the quantity $x$ to denote the error in an observation. He discusses the two types of errors, the constant error associated most often with the calibration of an instrument, and the random error which is always present. Gauss commonly assumes there is no constant error since this type of error would be eliminated by

a careful scientific observer. He begins his study of error with two assumptions [7, pg. 5]:

1. Random errors of measurements of the same type lie within fixed limits.

2. All errors within these limits are possible, but not necessarily with equal likelihood.

Gauss introduces the function $\phi(x)$ with essentially the same meaning as in our discussion of density functions 2.6.2. Some of the properties of this function as explained by Gauss are [7, pg. 7]:

- For continuous errors, the probability of an error lying within a very small interval $(x, x + dx)$ is approximately $\phi(x)dx$.

- If constant error is eliminated, small errors are more likely to occur that large ones, and $\phi(x)$ will be greatest for $x = 0$, and will decrease when $|x|$ is very large.

- We assume that positive and negative errors of the same magnitude are equally likely, so $\phi(-x) = \phi(x)$ (This also implies no constant error).

- We assume that the distribution of errors is given by an integrable function $\phi : \mathbb{R} \mapsto [0, \infty)$. Simply, this means, if $P(a \leq x \leq b)$ denotes the probability that the error of a given observation lies between two values $a$ and $b$, then

$$P(a \leq x \leq b) = \int_a^b \phi(x)\, dx.$$

- Since $\phi(x)$ is a density function, it is known that

$$\int_{-\infty}^{\infty} \phi(x)\, dx = 1.$$

Gauss defines $k = \int_{-\infty}^{\infty} x\,\phi(x)\,dx$ where $k$ is the mean (center) value of all errors $x$. He calls this quantity the "constant part of the error." This agrees with modern terminology. Today, we refer to $k$ as the mean of the distribution of $\phi$. By the Law of Large Numbers, if numerous measurements are made and the errors recorded, then the average of the errors will be close to $k$ or "expected value" of $x$ (see Section 2.7.1.) In many cases, we assume this value to be 0.

Suppose the value of $k$ is known and is not zero. It would then be possible to correct each observation by eliminating the constant error $k$. We will let $x'$ be the corrected observation, so $x' = x - k$ and its probability $\phi'(x') = \phi(x)$. This leads us to the following calculation:

$$\begin{aligned} E(x-k) &= E(x) - E(k) \quad \text{by linearity} \\ &= E(x) - k \\ &= k - k = 0. \end{aligned}$$

This equation shows that errors in the corrected observations have no constant part. The value of $\int x\,\phi(x)\,dx = k$ also indicates the presence or absence of constant error and its magnitude.

Gauss next introduces what he calls the "mean-square error" of $x$ This is defined by

$$m^2 = \int_{-\infty}^{\infty} x^2\,\phi(x)\,dx = 0, \text{ and } m = \sqrt{m^2}.$$

Note that if $k = 0$, then $m^2$ is the variance of the observations and $m$ is the standard deviation as defined in Section 2.8. We will assume throughout this work, as Gauss did, that for all the measurements taken that the constant part of the error is 0, and the variance will be denoted by $m^2$ and the standard deviation by $m$.

It is Gauss' conclusion that the class of observations in the set having the smallest mean-square value is the one having the most precision or the best estimate of values. He discussed his choice of $m^2$ as an indicator of least error as being an arbitrary one. Admittedly, there are other ways to measure the variability of errors, such as the expression $\int_{-\infty}^{\infty} |x|\phi(x)\,dx$ chosen for the same purpose by Laplace. He points out that his chosen convention leads to results that are "distinguished by their wonderful simplicity and generality." Some specific reasons for his choice are listed below.

- The function $m^2$ is always positive. Also, it is the simplest power function with this property.
- The function is differentiable and integrable unlike the absolute value function.
- The function is approximates the average value in cases where large numbers of observations are being considered, and is simple to use when considering smaller numbers of observations.[6]

Gauss points out the similarity between errors in observations and the results of a game of chance. As in a game of chance, we may put an arbitrary bet $V(x)$ on each outcome $x$. If we do that, then we may calculate the expected value as $\int_S V(x)\,\phi(x)\,dx$, where the integral is taken over the sample space $S$. In calculating $m^2$, it is as if each error costs a value equal to the square of the error.

The quantity which Gauss calls the *mean error to be feared*, $m = \sqrt{m^2}$, is in modern terms the standard deviation when $k = 0$ [cf. Section 2.8]. When $k \neq 0$, the standard deviation is then defined by correcting the mean error $m$ by subtracting

[6]Even today, textbooks admit not having a good intuitive justification for using this method to measure precision.

the constant error $k$, and we will let $m'$ be the standard deviation. Now we have the calculation of $m'^2$ that follows:

$$\begin{aligned} m'^2 &= \int x'^2\,\phi'(x')\,dx = \int (x-k)^2\,\phi(x)\,dx \\ &= \int x^2\,\phi(x)\,dx - 2k\int x\,\phi(x)\,dx + k^2\int \phi(x)\,dx \\ &= m^2 - 2k^2 + k^2 \\ &= m^2 - k^2. \end{aligned}$$

In §18 of Gauss, there is an important result that will be needed to show his justification of the Method of Least Squares. Suppose that we have independent random variables $\{e, e', e'', \cdots\}$ that are the errors of some observations we have taken, all with standard deviation 1 and expected value 0. We have the linear function of the total errors given by

$$E = \lambda e + \lambda' e' + \cdots,$$

Now, recall from Equation 2.8.1, that the variance $M^2$ is

$$M^2 = \sum_{i=1}^{k} \lambda_i^2 e_i^2.$$

and the expression of the variance of a linear function of errors is in this case

$$M^2 = \|\lambda\| = \sum_{i=1}^{k} \lambda_i^2. \tag{3.1.1}$$

## 3.2 The Chebyshev-like Inequality

In this section, we will deal with §9-10 in Gauss' work. Chebyshev's Inequality is a simple statement that points out the fact that a small variance makes large deviations from the mean improbable [4, pg. 219].[7] What Gauss is demonstrating

[7]The statement of the theorem is, "Let $X$ be a random variable with mean $\mu = E(X)$ and variance $\sigma^2$. Then for any $\epsilon > 0, P|x - \mu \geq t \leq \frac{\sigma^2}{\epsilon^2}$."

in these three examples, and the theorem that follows is the relationship between $\lambda$ standard deviations of the mean, and the probability that all the errors lie within the range of $-\lambda m$ to $\lambda m$.

Recall from above that the value $m$ is the "mean error to be feared" (standard deviation). Gauss defines a function $\mu := \mu(\lambda) = \int_{-\lambda m}^{\lambda m} \phi(x)\, dx$ to be the probability that an error $x$ is within the limits of $-\lambda m$ to $\lambda m$. He also defines a number $\rho$ by the condition $\mu = \int_{-\rho}^{\rho} \phi(x)\, dx = 1/2$, and then investigates three examples. It will be helpful to note that the quantity $\lambda$ is a multiple of the standard deviation, that which Gauss calls the "mean error to be feared."

**Example 3.2.1.** In this case, we will assume the probability density function, $\phi(x)$ is uniform on the interval $[a, -a]$, and the constant error $k = 0$. Suppose all errors lie between the values of $a$ and $-a$, and that the errors between these two limits are equally probable. The graph of this density function $\phi$ is represented below: Since $\phi(x)$ is constant,

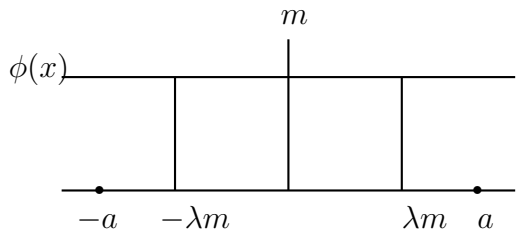


FIGURE 3.6. Uniform Density Function

$$\int_{-a}^{a} \phi(x)\, dx = \phi(x) \int_{-a}^{a} dx = 2a\, \phi(x).$$

Now, since all errors are within the limits of $-a$ and $a$, then $2a\phi(x) = 1$, which implies that $\phi(x) = \frac{1}{2a}$.

Next, we need to calculate the standard deviation, $m = \sqrt{m^2}$. We will let the quantity $a$ be a non-negative real number, and recall that $\phi(x) = \frac{1}{2a}$ is a constant.

Now we have

$$\begin{aligned} m^2 &= \int_{-a}^{a} x^2\,\phi(x)\,dx \\ &= \frac{1}{2a}\int_{-a}^{a} x^2\,dx \\ &= \frac{1}{2a}\left[\frac{1}{3}x^3\Big|_{-a}^{a}\right] \\ &= \frac{1}{6a}\left(2a^3\right) \\ &= \frac{a^2}{3} \\ \text{and,}\quad m &= \frac{a}{\sqrt{3}}. \end{aligned}$$

We turn our attention to the function $\mu(\lambda)$. We will set the limits of integration between $-\lambda m$ and $\lambda m$. The value of $\mu(\lambda)$ will indicate the probability that the error $x$ will be between the limits chosen. For $m = \frac{a}{\sqrt{3}}$, we have

$$\begin{aligned} \mu(\lambda) &= \int_{-\lambda m}^{\lambda m} \phi(x)\,dx = \int_{-\lambda\frac{a}{\sqrt{3}}}^{\lambda\frac{a}{\sqrt{3}}} \phi(x)\,dx \\ &= \frac{1}{2a}\left[x\Big|_{-\lambda\frac{a}{\sqrt{3}}}^{\lambda\frac{a}{\sqrt{3}}}\right] = \frac{1}{2a}\left(\frac{2a\lambda}{\sqrt{3}}\right) = \frac{\lambda}{\sqrt{3}}. \end{aligned}$$

For any distribution of errors $\phi$, Gauss defines the probable error to be that number $\rho := \lambda m$ where $\lambda$ is such that $\mu(\lambda) = 1/2$. In this case, $\mu = \frac{1}{2} = \frac{\lambda}{\sqrt{3}}$, so $\lambda = \frac{\sqrt{3}}{2}$ and $\rho \approx .8660m$.

**Example 3.2.2.** In this example, we will consider the constant error $k = 0$ and the probability distribution function $\phi$ to be a tent-shaped distribution, linear on the intervals $[-a, 0]$ and $[0, a]$ with a maximum at 0 as shown in the figure below: Since on $[0, a]$, $\phi(x) = \frac{a-x}{a^2}$. Also, since $\phi(x) = \phi(-x)$, we will consider the function

$$\mu = \int_{-a}^{a} \phi(x)\,dx = 2\int_{0}^{a} \left(\frac{a-x}{a^2}\right)\,dx = 2\int_{0}^{a} \left(\frac{1}{a} - \frac{x}{a^2}\right)\,dx.$$

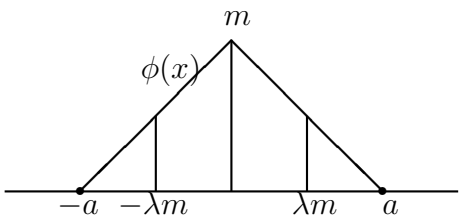



FIGURE 3.7. Tent-shaped Distribution

Now, we will calculate the standard deviation $m$.

$$\begin{aligned} m^2 &= \int_{-a}^{a} x^2\,\phi(x)\,dx = 2\int_0^a x^2\left(\frac{a-x}{a^2}\right)\,dx \\ &= 2\left[\int_0^a \frac{x^2}{a}\,dx - \int_0^a \frac{x^3}{a^2}\,dx\right] \\ &= 2\left[\frac{1}{3a}x^3\Big|_0^a - \frac{1}{4a^2}x^4\Big|_0^a\right] \\ &= \left[\frac{a^2}{3} - \frac{a^2}{4}\right] \\ &= 2\left[\frac{a^2}{12}\right] = \frac{a^2}{6}, \\ \text{so, } m &= \frac{a}{\sqrt{6}}. \end{aligned}$$

Now, we will calculate $\mu(\lambda)$.

$$\mu(\lambda) = 2\int_0^{\lambda m}\left(\frac{1}{a} - \frac{x}{a^2}\right)\,dx = \frac{2}{a}\left(x\Big|_0^{\lambda\frac{a}{\sqrt{6}}}\right) - \frac{2}{2a^2}\left(x^2\Big|_0^{\lambda\frac{a}{\sqrt{6}}}\right) = \lambda\sqrt{\frac{2}{3}} - \frac{1}{6}\lambda^2.$$

Solving for $\lambda$ in terms of $\mu$:

$$\lambda^2 - 2\sqrt{6}\lambda - 6\mu = 0.$$

Using the quadratic formula, we have

$$\lambda = \sqrt{6} - \sqrt{6 - 6\mu}.$$

For $0 \le x \le a$ and $\mu = \frac{1}{2}$, we make the appropriate substitutions, and

$$\lambda = \sqrt{6} - \sqrt{6-3} = \sqrt{6} - \sqrt{3}.$$

Finally, we have $\rho = \lambda m \approx .7174m$.

**Example 3.2.3.** In this example, the constant error $k = 0$ and $\phi$ is the normal distribution defined by

$$\phi(x) = \frac{e^{-x^2/h^2}}{h\sqrt{\pi}}.$$

In this case, Gauss tells us that $m = h\sqrt{1/2}$. To see this, he might have calculated $m$ as follows:

$$\begin{aligned} m^2 &= \frac{1}{h\sqrt{\pi}} \int_{-\infty}^{\infty} x^2\, e^{-x^2/h^2}\, dx \\ &= \frac{h^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} u^2\, e^{-u^2}\, du \qquad \text{(substituting } hu = x\text{).} \end{aligned} \tag{3.2.1}$$

Now, using integration by parts with $U = u$ and $dv = (-2u)e^{-u^2}$, we have

$$\begin{aligned} \int u^2 e^{-u^2}\, du &= -\frac{1}{2} \int u(-2u)e^{-u^2}\, du \\ &= -\frac{1}{2} \left( ue^{-u^2} - \int e^{-u^2}\, du \right). \end{aligned}$$

From this, we have

$$\int_{-a}^{a} u^2\, e^{-u^2}\, du = \frac{1}{2} \int_{-a}^{a} e^{-u^2}\, du,$$

hence,

$$\int_{-\infty}^{\infty} u^2\, e^{-u^2}\, du = \frac{1}{2} \int_{-\infty}^{\infty} e^{-u^2} = \frac{\sqrt{\pi}}{2}, \tag{3.2.2}$$

by the well-known formula $\int_{-\infty}^{\infty} e^{-u^2} = \sqrt{\pi}$. Combining (1) and (2), we get $m^2 = h^2/2$, which is what we sought to show.

Gauss uses the results of these examples to introduce the following theorem.

**Theorem 3.2.4.** *If the function $\phi(x)$ is non-increasing as $|x|$ increases, then*

- $\lambda \leq \sqrt{3}$, *whenever* $\mu(\lambda) < \frac{2}{3}$;
- $\lambda \leq \frac{2}{3\sqrt{1-\mu(\lambda)}}$, *whenever* $\mu(\lambda) > \frac{2}{3}$;
- *For* $\mu(\lambda) = \frac{2}{3}$, *the two bounds coincide, and* $\lambda \leq \frac{2}{\sqrt{3}}$.

*Proof.* Let $y = \int_{-x}^{x} \phi(x)\,dx$ be the probability an error is bounded by the values $-x$ and $x$. Let $x = g(y)$. Differentiating this equation with respect to $y$, we see that

$$1 = g'(y)dy.$$

Applying the Fundamental Theorem of Calculus,

$$\begin{aligned} g'(y) &= \frac{1}{\phi(x) - \phi(-x)} \\ &= \frac{1}{\phi(x) + \phi(x)}. \end{aligned}$$

Now, we know $g'(y)$ is non-decreasing from $y = 0$ to $y = 1$, thus $g''(y)$ is always non-negative. By the Product Rule, we have

$$d(yg'(y)) = g'(y)dy + y(g''(y)),$$

and

$$d(yg'(y)) - g'(y)dy = y(g''(y)).$$

Integrating both sides,

$$yg'(y) - g(y) = \int yg''(y)\,dy > 0 \text{ from the original assumption.}$$

Since

$$yg'(y) - g(y) > 0, \text{ and } yg'(y) > g(y),$$

then

$$1 > \frac{g(y)}{yg'(y)},$$

and therefore,

$$0 < 1 - \frac{g(y)}{yg'(y)} < 1.$$

Let $f$ be the value of $1 - \frac{g(y)}{yg'(y)}$ when $y = \mu$, and recall $g(y) = x = \lambda m$. Now, we have

$$\begin{aligned} f &= 1 - \frac{g(\mu)}{\mu g'(\mu)} \\ &= 1 - \frac{\lambda m}{\mu g'(\mu)}. \end{aligned}$$

Solving for $g'(\mu)$, we have

$$g'(\mu) = \frac{\lambda m}{(1-f)\mu}.$$

Consider the function $F$, so that

$$F(y) = \frac{\lambda m}{(1-f)\mu}(y - \mu f),$$

and by the Chain Rule,

$$d(F(y)) = F'(y)dy.$$

Now,

$$F(\mu) = \frac{\lambda m}{(1-f)\mu}(\mu - \mu f) = \lambda m = g(\mu),$$

and

$$F'(\mu) = \frac{\lambda m}{(1-f)\mu} = g'(\mu).$$

We know $g'(y)$ is non-decreasing with respect to an increase in $y$, and $F'(y)$ is a constant, so

$$g'(y) - F'(y) = \frac{d(g(y) - F(y))}{dy}$$

is also positive when $y > \mu$ and negative when $y < \mu$. This implies $g(y) - F(y)$ is always positive, so $g(y) - F(y) > 0$. Thus, $|g(y)| > |F(y)|$ for a positive-valued $F(y)$ (*i.e.*, $y = \mu f$ to $y = 1$). Hence,

$$\int_{\mu f}^{1} F^2 y \, dy < \int_{\mu f}^{1} g^2(y)dy < \int_{0}^{1} g^2(y)dy = m^2.$$

Next, we will calculate the value of the integral.

$$\begin{aligned}
\int_{\mu f}^{1} F^2(y)\,dy &= \int_{\mu f}^{1} \frac{\lambda^2 m^2}{(1-f)^2\mu^2}(y-\mu f)^2\,dy \\
&= \frac{\lambda^2 m^2}{(1-f)^2\mu^2}\left[\int_{\mu f}^{1}(y-\mu f)^2\,dy\right] \\
&= \frac{\lambda^2 m^2}{(1-f)^2\mu^2}\left[\frac{1}{3}y^3\Big|_{\mu f}^{1} - \mu f y^2\Big|_{\mu f}^{1} + \mu^2 f^2 y\Big|_{\mu f}^{1}\right] \\
&= \frac{\lambda^2 m^2}{(1-f)^2\mu^2}\left[\frac{1}{3} - \frac{1}{3}\mu^3 f^3 - \mu f + \mu^2 f^2\right] \\
&= \frac{\lambda^2 m^2}{3(1-f)^2\mu^2}\left[1 - 3\mu f + 3\mu^2 f - \mu^3 f^3\right] \\
&= \frac{\lambda^2 m^2}{3(1-f)^2\mu^2}(1-\mu f)^3 \leq 1.
\end{aligned}$$

This implies that for $0 \leq f < 1$,

$$\lambda^2 \leq \frac{3\mu^2(1-f)^2}{(1-\mu f)^3}. \tag{3.2.3}$$

Set the function $G(f)$ such that

$$G(f) = \frac{3\mu^2(1-f)^2}{(1-\mu f)^3}.$$

We want to maximize G, so we find $G'(f)$ as follows using the Product Rule.

$$\begin{aligned}
G'(f) &= \left[3\mu^2(-2)(1-f)\cdot\frac{1}{(1-\mu f)^3} + 3\mu^2(1-f)^2\cdot\frac{3\mu}{(1-\mu f)^4}\right]df \\
&= \left[\frac{-6\mu^2(1-f)}{(1-\mu f)^3} + \frac{9\mu^3(1-f)^2}{(1-\mu f)^4}\right]df \\
&= -3\mu^2(1-f)\left[\frac{2(1-\mu f)}{(1-\mu f)^4} - \frac{3\mu(1-f)}{(1-\mu f)^4}\right]df \\
&= \frac{-3\mu^2(1-f)}{(1-\mu f)^4}\left[2 - 2\mu f - 3\mu + 3\mu f\right]df. \\
&= \frac{-3\mu^2(1-f)}{(1-\mu f)^4}\left[2 - 3\mu + \mu f\right]df.
\end{aligned}$$

We set $G'(f) = 0$, then

$$0 = \frac{3\mu^2(1-f)}{(1-\mu f)^4} \text{ or } 0 = 2 - \mu + \mu f. \tag{3.2.4}$$

*Case 1:* Recall that the function $f$ has non-increasing values between 0 and 1, so the value of $f$ is at a maximum when $f = 0$. When $\mu < \frac{2}{3}$, and $f = 0$ then from 3.2.4 we know $G'(0) = 3\mu^2$. From 3.2.3 and $f = 0$, we have

$$\lambda^2 \leq 3\mu^2 \Rightarrow \lambda \leq \mu\sqrt{3} \text{ for } \mu < \frac{2}{3}.$$

*Case 2:* For $\mu > \frac{2}{3}$, and referring to 3.2.4, the value of $G$ is at a maximum for

$$G'(0) = 0 = 2 - 3\mu + \mu f \Rightarrow f = 3 - \frac{2}{\mu}.$$

Substituting again into 3.2.3, we have

$$\begin{aligned}
\lambda^2 &\leq \frac{3\mu^2(1 - 3 + \frac{2}{\mu})^2}{(1 - \mu(3 - \frac{2}{\mu}))^3} \\
&\leq \frac{3\mu^2(-2 + \frac{2}{\mu})^2}{(1 - 3\mu + 2)^3} \\
&\leq \frac{3\mu^2(4 - \frac{8}{\mu} + \frac{4}{\mu^2})}{(3 - 3\mu)^3} \\
&\leq \frac{12(\mu^2 - 2\mu + 1)}{27(1 - \mu)^3} \\
&\leq \frac{4(\mu - 1)^2}{9(1 - \mu)(\mu - 1)^2} = \frac{4}{9(1 - \mu)}.
\end{aligned}$$

And finally,

$$\lambda \leq \frac{2}{3\sqrt{1 - \mu}}.$$

*Case 3:* For $\mu = \frac{2}{3}$, then $f = 0$, and

$$\lambda^2 \leq 3\mu^2 \Rightarrow \lambda^2 \leq \frac{4}{3} \Rightarrow \lambda \leq \frac{2}{\sqrt{3}}.$$

□

## 3.3 Rate of Convergence of Sample Mean and Variance

In §15-16, Gauss takes a look at a method for determining the precision of estimates of errors. Let $e_1, e_2, \cdots, e_n$ be mutually independent random errors with

the same probability distribution (and, of course, no constant part). Let

$$y := \frac{e_1^2 + e_2^2 + \cdots + e_n^2}{n}.$$

Then, the expected value of $y$ is $m^2$. Now, we know the error in $y$ is $y - m^2$, and the variance of $y$ is calculated from 2.7.8 as

$$V(y) = E(y - m^2)^2.$$

We want to show that as the value of $n$ increases toward infinity, then a random value of $y$ does not vary significantly from its mean value, $m^2$

*Proof.* Let the $e_i$'s have no constant part, and the errors be taken from observations of the same class. This fact indicates that the probabilities of the individual $e_i$'s are represented by the same function that we will call $\phi(e)$.

We wish to compute the variance of $y$, *i.e.*, the expected value of

$$U = \left( \frac{e_1^2 + e_2^2 + \cdots + e_n^2}{n} - m^2 \right)^2. \tag{3.3.1}$$

Let e be one of the $e_i$, and let $p^4 = \int e^4 \phi(e) de$. Then, the expected value of a term like $\frac{e^4}{n^2}$ is

$$\int e^4 \phi(e)\, de = \frac{1}{n^2} \int e^4 \phi(e)\, de = \frac{p^4}{n^2}.$$

For a term like $\frac{2e_i^2 e_j^2}{n^2}$ is

$$\begin{aligned} \int \frac{2e_i^2 e_j^2}{n^2} \phi_i(e_i)\phi_j(e_j)\, d\overline{e} &= \frac{2}{n^2} \int e_i^2 e_j^2 \phi_i(e_i)\phi_j(e_j)\, d\overline{e} \\ &= \frac{2}{n^2} \int e_i^2 \phi_i(e_i) de \int e_j^2 \phi_j(e_j)\, de \\ &= \frac{2m^4}{n^2}, \end{aligned}$$

since all the $e$'s are independent and have the same expected value. The expected value of the function $U$ is the sum of the expected values of the individual terms

as proved in 2.7.4. Calculating the right side of 3.3.1 and substituting the expected values for the different types of terms will produce the following for every $n$:

$$\begin{aligned} E(U) &= (\frac{1}{n}\sum e_i^2 - m^2)^2 = (\frac{1}{n}\sum e_i^2)^2 - \frac{2m^2}{n}\sum e_i^2 + m^4 \\ &= \frac{1}{n^2}(\sum e_i^4 + \sum_{i\neq j} e_i^2 e_j^2) - \frac{2m^2}{n} + m^4. \end{aligned}$$

Now, apply the expected values we calculated for the types of terms we have, and we get

$$\begin{aligned} &= \frac{1}{n^2}(np^4 + n(n-1)m^4 - 2m^4 + m^4 \\ &= \frac{1}{n}(p^4 + (n-1)m^4) - m^4 \\ &= \frac{p^4 - m^4}{n}. \end{aligned}$$

It is clear that as $n$ increases, the value of the standard deviation decreases, and these two values $y$ and $m^2$ would be expected to be nearly equal.

Thus, with a sufficiently large number of mutually independent, random errors, $e_i$, the "mean error to be feared" by

$$\sqrt{\frac{p^4 - m^4}{n}}.$$

When the errors have a constant part, then the expected value is best approximated by the arithmetic mean of the errors, or

$$\frac{e_1 + e_2 + \cdots + e_n}{n} = m.$$

By a calculation similar to the one above, we know the variance of this function is $\frac{m^2}{n}$ and the "mean error to be feared" to be

$$\frac{m}{\sqrt{n}}.$$

It is clear that as $n$ becomes larger, the value of $\frac{m}{\sqrt{n}}$ comes closer to the expected value $m$. The significance of this expression can be used to see the relationship between precision of an estimate and the number of errors measured. For example, to get twice the precision, then four times the number of measurements need to be taken.

□

# Chapter 4. Method of Least Squares

## 4.1 Introduction

According to Stigler, Gauss' 1809 derivation of the method of least squares was the "most influential of his statistical works" [16, pg. 157]. Gauss' discovery of the method of least squares in the late 18th century was indeed a pivotal event in the study of mathematical statistics, since it irrevocably linked the errors of observations to probability theory. We begin this part with some of the history of this problem. Then we will summarize the mathematical development of this model taken from the publication of Gauss himself in his *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae,* published in 1821.

The eighteenth-century antecedents to Gauss' publication were numerous, and in some cases controversial. An example mathematically similar to the problem scientists then faced would be determining the line of best fit for a set of three or more data points in the plane. In general, no line passes through all the data points, yet the data points must somehow be combined to yield the slope and the $y$-intercept of the best linear approximation to the data. Gauss' contemporaries were concerned with "combinations of observations," of a type made in the study of astronomy and navigation [16, pg. 5].[8] Legendre published a paper on the method of least squares in 1805.[9] His treatment, however, lacked a "formal consideration of

[8] In the mid-1700's both Mayer and Euler worked with 21 inconsistent equations of three unknowns, and Euler with 75 inconsistent equations of eight unknowns. Euler would choose a small number of equations, and solve for the unknowns, but would only accept the results when they yielded very similar results believing that the combination of many equations would only result in larger errors. Mayer, on the other hand, cleverly combined the 21 into three, and solved, taking the statisticians view that random errors would eventually cancel themselves out [16, pg. 28].

[9] A priority argument ensued between Gauss and Legendre. Gauss was able to show that although he had not published his argument until 1809, he had written colleagues concerning its use before the publication by Legendre [16, pgs. 145–146].

probability and its relationship to least squares," making it impossible to determine the accuracy of the method when applied to real observations [16, pg. 139].

Gauss's treatment of the method of least squares had its basis in probability theory. In his 1809 publication, he assumed that errors obeyed a normal distribution. Stigler accuses Gauss of being circular in this treatment of least squares, because he assumed the normal distribution of errors, deduced the method, and then used the fact that instances of the method were in common use to justify the normality assumption [16, pgs. 140–143]. The circularity goes away if the 1809 treatment is viewed as a justification for using the method only when errors are known to be normally distributed. In his 1821 work, Gauss abandoned the use of the normal error function and presents an argument "making use of mathematical probability to assess uncertainty and make inferences" to justify the method [16, pg. 158].

In this part we will summarize §19-21 of Gauss's *Theoria Combinationis*. Here Gauss is dealing with the following situation. Several observable quantities are dependent upon unknown parameters in a known way. Observations of these quantities subject to errors of known variance have been made. What is our best estimate of the unknown parameters? In more detail, true values $V_1, V_2, \cdots, V_m$ of some physical constants are unknown. We attempt to measure the $V_i$ not by observing them directly, but by observing other quantities $U_1, U_2, \cdots, U_n$ that depend on the $V_i$ by known functions $F_j : \mathbb{R}^m \mapsto \mathbb{R}$ such that $U_j = F_j(V_1, V_2, \cdots, V_m), \quad (j = 1, \cdots, n)$. Suppose that numerous observations of the $U_j$ have been made.[10] Which

[10]Note that the $F_j's$ are not necessarily different functions. If a single function is repeated many times among the $F_j$, it simply corresponds to multiple observations of the same type. Also, if the $F_j$'s are linear, they need not be independent, so the observations "overdetermine" the $V_i$. For example, we might have 20 $F$'s and only five variables $V$.

estimates of $V_1, V_2, \cdots, V_m$ agree best with the observations? Gauss' treatment culminates with the following theorem, whose statement we quote exactly [7, pg. 45].

**Theorem 4.1.1.** *The most reliable values of the unknowns are those that minimize the sum of the ... squares of the differences between the observed and the computed values of the quantities* $V_1, V_2, \cdots$ .

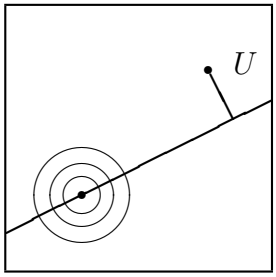
Up to this point, we treated the $V_i$'s as unknown constants. Now we wish to conceive of each as being a fixed value of a variable $v_i$. Any other values for the $v_i$ will yield values for observation variables $u_j = F_j(v_1, \cdots, v_m)$. For ease of notation, we will make the following abbreviations, $v := (v_1, v_2, \cdots, v_m)$ and $u := (u_1, u_2, \cdots, u_n)$. Now as $v$ varies over its entire domain in $\mathbb{R}^m$ (parameter space), $u = F(v)$ moves about in $u$-space (observation space), sweeping out a subset, *i.e.*, the range of the function $F$ which we will call range $F$. Assuming error distributions of the observations are known, any choice of values $V'$ for $v$ results in a probability distribution for the observed quantities. In other words, the observations will cluster around the point $F(V')$.[11] Or again, if $V'$ is fixed, then the probability distribution for the observation will have a peak at $F(V')$. The best estimate of $V'$ is that which places the observed data at the point of greatest probability density. Gauss' least squares says that this point occurs precisely for the value of $v$ that minimizes

$$\sum_{j=1}^{n} (u_j - F_j(v_1, v_2, \cdots, v_m))^2.$$

Let's look at a graphic representation of the problem. In $u$-space, the true observations corresponding to different parameter values are found in range $F$, which is depicted as a line in the figure. An observation $U$ has been made. It is

[11] Note that a vector $u$ is a single observation from several real measurements. When we say that observations "cluster" around $F(V')$, we mean that if the compound measurement was taken repeatedly, then the positions of the corresponding vectors would be near $F(V')$.

the point in $u$-space depicted near range $F$ and labeled $U$.[12] Now, as stated above, if some specific value $V'$ of $v$ is chosen, then there is a corresponding expectation for where the observation vectors will tend to lie. This is depicted by means of the circles around the unlabeled point in range $F$.[13]




$u$-space

FIGURE 4.8. Estimation of Error with Same Standard Deviation

If other values $V'', V''', \cdots$ of $v$ are examined, the peak moves to different locations. Now, what is the best estimate for the true values of the parameters? In his early work, Gauss appears to have reasoned as follows. We want to move the peak to the position that gives maximum likelihood to observations like the one we actually took. Assuming that the probability falls off with the radial distance from $F(V)$), if the peak is moved to the closest point in range $F$ to $U$, then $U$ will be at the point of greatest density.

Gauss' justification in *Theoria Combinationis* was completely different, and did not depend on normality. In the next section and in the sections to follow, we present this later argument.

---

[12] $U$ does not lie in range $F$ because errors have pushed it off. That is, if observations were error-free, $U$ would lie in range $F$.

[13] As we've pictured things, the level sets for the probability distribution are circles. The exact shape of these level sets depends on the distribution of the errors. In the case where each component measurement is normally distributed about its true value, and all the component measurements have the same standard deviation, then circles are indeed the shape of the level sets. Moreover, it is necessary for the errors of each $v_i$ to satisfy the normality assumption in order for these to be circles [5, pg. 78].

## 4.2 A Key Lemma

In this section, we present the results of Gauss' §20 in the same order as Gauss, and using very similar notation. We will give a second presentation of the same results in the next section using modern language. This is a key lemma upon which the justification of least squares in *Theoria Combinationis* is based. We will give the justification itself in Section 4.4.

In preparation, we will let $V, V', V'', \cdots$ be $\pi$ affine functions of $\rho$ unknowns $\mathbf{x} = (x, y, z, \cdots)$. Let $L, L', L'', \cdots$ be independent observations of the $V$'s. Gauss assumes the system is nondegenerate and dismisses all but the overdetermined case of the system in which $\pi > \rho$. We let the errors in the observations be

$$v := \frac{(V - L)}{\sqrt{p}}, v' = \frac{(V' - L')}{\sqrt{p'}}, \cdots , \tag{4.2.1}$$

where the $p$'s are the weights of the mean errors of the observations. As a result, the errors have the same variance [13, pg. 459].

Now, the $v, v', v'', \cdots$ are $\pi$ affine linear functions of the $\rho$ unknowns $\mathbf{x} = (x, y, z, \cdots)$, where the coefficients $a, b, c, \cdots$ are known.

$$\begin{aligned} v &= ax + by + cz + \cdots + l \\ v' &= a'x + b'y + c'z + \cdots + l' \\ v'' &= a''x + b''y + c''z + \cdots + l'' \\ \vdots &= \vdots \quad \vdots \quad \vdots \quad . \end{aligned} \tag{4.2.2}$$

where the $l, l', l'', \cdots$ are constants arising from the original equations and from the dilation of the coordinate space that occurred in Equation 4.2.1. Note that this system describes a mapping $F$ from $\mathbb{R}^\rho$ to $\mathbb{R}^\pi$. We are going to look for an affine linear mapping $G$ from $\mathbb{R}^\pi$ to $\mathbb{R}^\rho$ such that:

1. $G \circ F$ is the identity on $\mathbb{R}^\rho$

2. $G$ satisfies an optimality condition, described as follows:

   Suppose $g(v, v', v'', \cdots)$ is the first component of $G$. Then $g(v, v', v'', \cdots) = \kappa v + \kappa' v' + \kappa'' v'' + \cdots + k$. We want $\sum \kappa^2$, to be as small as possible, and we want similar conditions for the other components.

Assume that $G$ satisfies condition 1. If $\kappa, \kappa', \kappa'', \cdots$ are the coefficients of the $x$-coordinate of $G$, then a relation of the form

$$\kappa v + \kappa' v' + \kappa'' v'' + \cdots = x - k \quad (k \text{ is a constant,})$$

holds for all $x, y, z, \cdots$ (here we are viewing $v, v', v'', \cdots$ as functions of $x, y, z, \cdots$). Condition 2 demands that we find the $G$ such that

$$\kappa^2 + \kappa'^2 + \kappa''^2 + \cdots$$

is as small as possible.

Define the following functions of $v, v', v'', \cdots$.

$$\begin{aligned} T &= av + a'v' + a''v'' = \sum_i^{\pi} a_i v_i \\ T' &= bv + b'v' + b''v'' + \cdots = \sum_i^{\pi} b_i v_i \\ T'' &= cv + c'v' + c''v'' + \cdots = \sum_i^{\pi} c_i v_i \\ \vdots &= \vdots \quad \vdots \quad \vdots \quad . \end{aligned} \tag{4.2.3}$$

If we view $v, v', v'', \cdots$ as quantities depending on $x, y, z, \cdots$, then we see that $T, T', T'', \cdots$ are also functions of $x, y, z, \cdots$.

$$\begin{aligned} T &= \sum av = x\sum aa + y\sum ab + z\sum ac + \cdots + \sum al \\ T' &= \sum bv = x\sum ab + y\sum bb + z\sum bc + \cdots + \sum bl \\ T'' &= \sum cv = x\sum ac + y\sum bc + z\sum cc + \cdots + \sum cl \\ \vdots &= \vdots \quad \vdots \quad \vdots \quad , \end{aligned}$$

where $\sum av = av + a'v' + a''v'' + \cdots$ and $\sum aa = aa + a'a' + a''a'' + \cdots$, and so on. We will simplify the notation by using [aa] to denote $\sum aa$, [ab] to denote $\sum ab$ and so on as Gauss does. Now we have the system described below:

$$\begin{aligned} T &= [aa]x + [ab]y + [ac]z + \cdots + [al] \\ T' &= [ab]x + [bb]y + [bc]z + \cdots + [bl] \\ T'' &= [ac]x + [bc]y + [cc]z + \cdots + [cl] \\ \vdots &= \vdots \quad \vdots \quad \vdots \quad . \end{aligned} \tag{4.2.4}$$

The number of unknowns, $x, y, z, \cdots$ is now the same as the number of $T_i$'s, and since the original system 4.2.2 is nondegenerate, we can solve for the $x, y, z, \cdots$ in terms of the $T, T', T'', \cdots$ by elimination. Again, for ease of notation, we will use $[\alpha\alpha]$ and similar symbols to denote the coefficients we derive by elimination. Now, the solutions for the unknowns have the form

$$\begin{aligned} x &= A + [\alpha\alpha]T + [\alpha\beta]T' + [\alpha\gamma]T'' + \cdots \\ y &= B + [\alpha\beta]T + [\beta\beta]T' + [\beta\gamma]T'' + \cdots \\ z &= C + [\alpha\gamma]T + [\beta\gamma]T' + [\gamma\gamma]T'' + \cdots \\ \vdots &= \vdots \quad \vdots \quad \vdots \quad , \end{aligned} \tag{4.2.5}$$

where $A$, $B$, etc., are constants (this is the inverse matrix to the matrix in 4.2.4). Now, we substitute the $T_i$'s from Equation 4.2.3 back in the equation, and let

$\alpha = a[\alpha\alpha] = \sum_i \alpha_i$, to give us

$$\begin{aligned}
x - A &= [\alpha\alpha][av] + [\alpha\beta][bv], +[\alpha\gamma]cv + \cdots = \alpha v + \alpha' v' + \alpha'' v'' + \cdots \\
y - B &= [\alpha\beta][av] + [\beta\beta][bv], +[\beta\gamma]cv + \cdots = \beta v + \beta' v' + \beta'' v'' + \cdots \\
z - C &= [\alpha\gamma][av] + [\beta\gamma][bv], +[\gamma\gamma]cv + \cdots = \gamma v + \gamma' v' + \gamma'' v'' + \cdots \\
\vdots &= \vdots \qquad \vdots \qquad \vdots \qquad\qquad ,
\end{aligned}$$

and so on. Let us look at the first line of this system (similar considerations apply to subsequent lines). From the system above, we get the following relation, but for all $x, y, z, \cdots$

$$x - A = \alpha v + \alpha' v' + \alpha'' v'' \cdots \tag{4.2.6}$$

The $\alpha, \alpha', \alpha'', \cdots$ thus give us a set of coefficients of (the first row of) the system we seek satisfying condition 1. Consider any other set of coefficients $\kappa, \kappa', \kappa'', \cdots$. For these we have

$$x - k = \kappa v + \kappa' v' + \kappa'' v''. \tag{4.2.7}$$

Subtracting 4.2.7 and 4.2.6, we have

$$A - k = (\kappa - \alpha)v + (\kappa' - \alpha')v' + (\kappa'' - \alpha'')v'' + \cdots . \tag{4.2.8}$$

The left side is constant. The right side depends on $x, y, z, \cdots$. This implies

$$\begin{aligned}
(\kappa - \alpha)a + (\kappa' - \alpha')a' + (\kappa'' - \alpha'')a'' + \cdots = 0 \\
(\kappa - \alpha)b + (\kappa' - \alpha')b' + (\kappa'' - \alpha'')b'' + \cdots = 0 \\
(\kappa - \alpha)c + (\kappa' - \alpha')c' + (\kappa'' - \alpha'')c'' + \cdots = 0 \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad\qquad .
\end{aligned}$$

Then, by post-multiplying by $[\alpha\alpha], [\alpha\beta], [\alpha\gamma], \cdots$ in turn and adding, we have

$$(\kappa - \alpha)\alpha + (\kappa' - \alpha')\alpha' + (\kappa'' - \alpha'')\alpha'' + \cdots = 0. \tag{4.2.9}$$

Let $\overline{\kappa} := (\kappa, \kappa', \kappa'', \cdots)$ and similarly, $\overline{\alpha} := (\alpha, \alpha', \alpha'', \cdots)$. Then from 4.2.9, we have

$$(\overline{\kappa} - \overline{\alpha})\overline{\alpha} = 0. \tag{4.2.10}$$

Square the equation $\overline{\kappa} = \overline{\alpha} + (\overline{\kappa} - \overline{\alpha})$, we get

$$\begin{aligned} \overline{\kappa\kappa} &= (\overline{\alpha} + (\overline{\kappa} - \overline{\alpha})) \cdot (\overline{\alpha} + (\overline{\kappa} - \overline{\alpha})) \qquad (4.2.11) \\ &= \overline{\alpha\alpha} + \underbrace{2\overline{\alpha} \cdot (\overline{\kappa} - \overline{\alpha})}_{0} + (\overline{\kappa} - \overline{\alpha}) \cdot (\overline{\kappa} - \overline{\alpha}) \quad (\text{ from } 4.2.10). \end{aligned}$$

It is clear, therefore, that the sum of the $\overline{\kappa} \cdot \overline{\kappa}$ is at a minimum when $\overline{\kappa} = \overline{\alpha}$ which is the condition we set out to establish, since $(\overline{\kappa} - \overline{\alpha}) \cdot (\overline{\kappa} - \overline{\alpha})$ is strictly positive if $\overline{\kappa} \neq \overline{\alpha}$.

## 4.3 A Modern Look at Least Squares

In this section we will follow Gauss's proof using modern matrix notation. In an historical sense, it gives interesting insight into Gauss's methods. In a more practical sense, it gives a method to find a least squares solution of the overdetermined case similar to that shown in the previous section.

We begin with the assumption that observable quantities $V_1, \cdots, V_\pi$ are affine linear functions of parameters $x_1, \cdots, x_\rho$ such that

$$V_i = b_{1i}x_1 + \cdots + b_{i\rho}x_\rho + c_i \qquad b_{ij}, c_i \in \mathbb{R}. \tag{4.3.1}$$

We are envisioning a situation in which we know the values of all the $b_{ij}$ and $c_i$. We measure the $V_i$ in an attempt to infer values of the $x_i$.

Assume that an observation has been made, giving us values $L_i \in \mathbb{R}$ for the $V_i$. We switch to a new coordinate system in observation space where the point labeled by $L$ in the original system is at the origin, and where the coordinate axes have been dilated so that the variance of all observables are the same. The new

coordinates are

$$v_i := (V_i - L_i)/\sqrt{p_i}. \tag{4.3.2}$$

Now, we rewrite 4.3.1 in the new coordinate system, getting

$$v = Ax + l. \tag{4.3.3}$$

The $x_i$ are the original coordinates in parameter space. The $l_i$ are constants that have absorbed whatever constants were in the original equations as well as constants arising from the substitution in 4.3.2.

**Example 4.3.1.** Suppose $x$ is a single parameter, and $V_i = b_i x + c_i$. Then

$$\begin{aligned} \frac{V_i - L_i}{\sqrt{p_i}} &= \frac{b_i x + c_i - L_i}{\sqrt{p_i}} \\ &= \frac{b_i x}{\sqrt{p_i}} + \underbrace{\frac{c_i - L_i}{\sqrt{p_i}}}_{l_i}. \end{aligned}$$

Note that we cannot easily identify the $l_i$ with anything that has a concrete meaning. □

**Lemma 4.3.2.** *Suppose $A$ is a $\pi \times \rho, \pi > \rho$ matrix of rank $\rho$. Then there is a $\rho \times \pi$ matrix $K$ such that the following holds,*

$$\forall x \in \mathbb{R}^\rho \quad KAx = x,$$

*and among all such matrices the matrix $E = (A^T A)^{-1} A^T$ has rows of minimum norm.*

*Proof.* Since $A$ is a $\pi \times \rho$ matrix, $\pi > \rho$, and rank$A = \rho$, the $\rho \times \rho$ matrix $A^T A$ is invertible. Let $D$ denote its inverse. Then

$$\forall x, \;\; x = DA^T Ax.$$

Thus, $E := DA^T$ and satisfies the first condition of our lemma:

$$\forall x \;\; EAx = x. \tag{4.3.4}$$

Now, the optimality condition is that the quantities $\|K_i\|^2 = K_{ii}^2 + \cdots + K_{i\pi}^2$ should be as small as possible, where $K_i = K_{i1}, \cdots, K_{i\pi}$ denotes the $i$th row of $K$. This is equivalent to demanding that the sum of the diagonal entries of $KK^T$ should be as small as possible.

Take, then, any solution $K$ such that $\forall x, \; KAx = x$. Subtracting, we get

$$\forall x, (K - E)Ax = 0. \tag{4.3.5}$$

Thus $(K - E)A$ is the zero matrix. Multiplying on the right by $D^T$ and noting that $AD^T = E$, we get $(K - E)E^T = 0$. Finally

$$\begin{aligned} KK^T &= (E + (K - E))(E + (K - E))^T = (E + (K - E))(E^T + (K - E)^T) \\ &= EE^T + E(K - E)^T + (K - E)E^T + (K - E)(K - E)^T \\ &= EE^T + ((K - E)E^T)^T + (K - E)E^T + ((K - E)(K - E)^T \\ &= EE^T + (K - E)(K + E)^T. \end{aligned}$$

This shows that the solution $E$ is in fact the optimal one, since if $(K - E)$ has any non-zero entries, then $(K - E)(K + E)^T$ will have some strictly positive entries on its diagonal. □

Returning to equation 4.3.3, our lemma shows immediately that

$$G(v) := E(Ax + l) - El$$

is the left inverse to the function $F(x) = Ax + l$, (*i.e.,* $G \circ F(x) = x$), and among all linear left inverses, the non-consistent part of $G$ is optimal.

## 4.4 Gauss' Justification of Method of Least Squares

What is the relevance of this lemma to the justification of least squares? If we know that observables $v \in \mathbb{R}^\pi$ are affine linear functions $v = v(x)$ of $x \in \mathbb{R}^\rho$, we may seek a linear function from $\mathbb{R}^\pi$ to $\mathbb{R}^\rho$ that does the best job possible in recovering the values of the parameters that were behind our observation. Of course, we need to deal with the fact that the observation might include some error which prevents it from being of the form $v(x)$ for any $x$. Now, when we attempt to recover the parameter values, it's not unreasonable that we should seek a single function, once and for all, to be used for every observation (provided that we are dealing with the same function $v$).

Our observations are dispersed due to error. According to Gauss, this error represents a loss, as if in a hopeless game of chance where every play loses [7, pg. 9]. Error, therefore, is to be minimized. If we view the loss due to an error as proportional to its square, then the expected[14] loss is the variance of the error. Why should we choose the square? Ultimately, the choice is arbitrary. Granting this, the observer will choose instruments that are both free from bias and have minimum variation. Similarly, a method of estimating the true values of a parameter (our matrix $K$) will be most desirable when it amplifies the variance of the observations as little as possible. Now in §18, Gauss analyzes how the variance of $G(x) := \sum \lambda_i x_i$, a linear function of a random vector $x$ whose components have known variance, depends on the coefficients,$\lambda_i$ [cf. Equations 2.8.1 and 3.1.1]. He shows that the variance, in fact, is given by $\sum \lambda_i^2 V(x_i)$, where $V(x_i)$ is the variance of

[14] And here we mean expected in the proper probabilistic sense, that is, the long-term average cost of the losses.

$x_i$. So to minimize the sum of the squares of the coefficients in the estimator is to minimize the variance.

It is not immediately obvious how the estimates that result from using the optimal $K$ are related to the least squares estimate. Even though we've minimized some squares in our choice of $E$, these are not the squares that we minimize when using the method of least squares. Our task, now, is to show that the estimate from Section 4.2 is in fact the least squares estimate.

Returning to the set-up in 4.2, in $v$-space, our observations have given the (approximate) equations

$$v = 0, v' = 0, v'' = 0 \cdots .$$

In other words, we've chosen a coordinate system in which our measurement is at the origin (though we admit that the possible true values of the observables for different parameters do not include the origin). We have determined that

$$x - A = \alpha v + \alpha' v' + \alpha'' v'' + \cdots ,$$

so $x = A$ is our estimate for the parameters. Let

$$\begin{aligned} \Omega &= v^2 + v'^2 + v''^2 + \cdots \\ &= \frac{(V(x, y, z, \cdots) - L)^2}{p} + \frac{(V'(x, y, z, \cdots) - L')^2}{p'} + \cdots . \end{aligned}$$

Least squares picks the parameter values that minimize $\Omega$, where all the partials $\frac{\partial \Omega}{\partial x}, \frac{\partial \Omega}{\partial y}, \frac{\partial \Omega}{\partial z}, \cdots$ vanish. Now,

$$\begin{aligned} \frac{\partial \Omega}{\partial x} &= \frac{\partial}{\partial x}(v^2 + v'^2 + v''^2 + \cdots) \\ &= 2v\frac{\partial v}{\partial x} + 2v'\frac{\partial v'}{\partial x} + \cdots \\ &= 2va + 2va' + \cdots \\ &= 2T, \end{aligned}$$

where $T$ is as in 4.2.3. Similarly

$$\begin{aligned}\frac{\partial \Omega}{\partial y} &= \frac{\partial}{\partial y}(v^2 + v'^2 + v''^2 + \cdots) \\ &= 2v\frac{\partial v}{\partial y} + 2v'\frac{\partial v'}{\partial y} + \cdots \\ &= 2vb + 2vb' + \cdots \\ &= 2T'.\end{aligned}$$

So the values of $x, y, z, \cdots$ that minimize $\Omega$ are those for which $T = 0, T' = 0, T'' = 0, \cdots$ . But $(A, B, C, \cdots)$ solves the system $T = 0, T' = 0, T'' = 0, \cdots$ . as is clear from 4.2.5.

# References


[1] Bean, Michael A. 2001. Probability: The Science of Uncertainty with Applications to Investments, Insurance, and Engineering. Pacific Grove: Brooks/Cole.

[2] Brunk, H. D. 1960. Introduction to Mathematical Statistics. Boston: Ginn and Company.

[3] Dorrie, H. 1965. 100 Great Problems of Elementary Mathematics: Their History and Solutions. New York: Dover Press.

[4] Feller, William. 1950. An Introduction to Probability Theory and Its Applications, Vol. 1, Second Edition. New York: John Wiley & Sons, Inc.

[5] Feller, William. 1966. An Introduction to Probability Theory and Its Applications, Vol. 2, Second Corrected Printing. New York: John Wiley & Sons, Inc.

[6] Freedman, David and Pisani, Robert and Purves, Roger. 1998. Statistics, 3rd Edition. New York: W. W. Norton & Company.

[7] Gauss, Carl Friedrich, Translated by G. W. Stewart. 1995. Theory of the Combination of Observations Least Subject to Errors: Part One, Part Two, Supplement. Philadelphia: Society for Industrial and Applied Mathematics.

[8] Grimmett, G. R. and Stirzaker, D. R. 1982. Probability and Random Processes. Oxford: Clarendon Press.

[9] Grinstead, Charles M. and Snell, J. Laurie. 1997. Introduction to Probability, Second Revised Edition. Providence: American Mathematical Society.

[10] Hacking, Ian. 1975. The Emergence of Probability. Cambridge: Cambridge University Press.

[11] Kelly, Douglas C. 1994. Introduction to Probability. New York: Macmillan Publishing Company.

[12] Knight, Keith. 2000. Mathematical Statistics. Boca Raton: Chapman and Hall/CRC.

[13] Plackett, R. L. 1949. A Historical Note on the Method of Least Squares. Biometrika. 36:458–460.

[14] Plackett, Robin L. 1972. The Discovery of the Method of Least Squares. Biometrika. 59:239–251.

[15] Salsburg, David. 2001. The Lady Tasting Tea, How Statistics Revolutionized Science in the Twentieth Century. New York: Henry Holt and Company, LLC.

[16] Stigler, Stephen M. 1986. The History of Statistics, The Measurement of Uncertainty before 1900. Cambridge: The Belknap Press of Harvard University Press.

[17] van der Waerden, B. L. 1969. Mathematical Statistics. London: George Allen & Unwin Ltd.

# Vita

Belinda Bruton Brand was born December 23, 1951 in Clifton, Texas, and spent her childhood years in Baton Rouge, Louisiana. She finished her undergraduate degree in education from Louisiana State University in May 1973. After a career of 26 years in secondary math and science education, she came to Louisiana State University in June 2001 to pursue graduate studies in mathematics. She is currently a candidate for the degree of Master of Science, which will be awarded in August 2003.