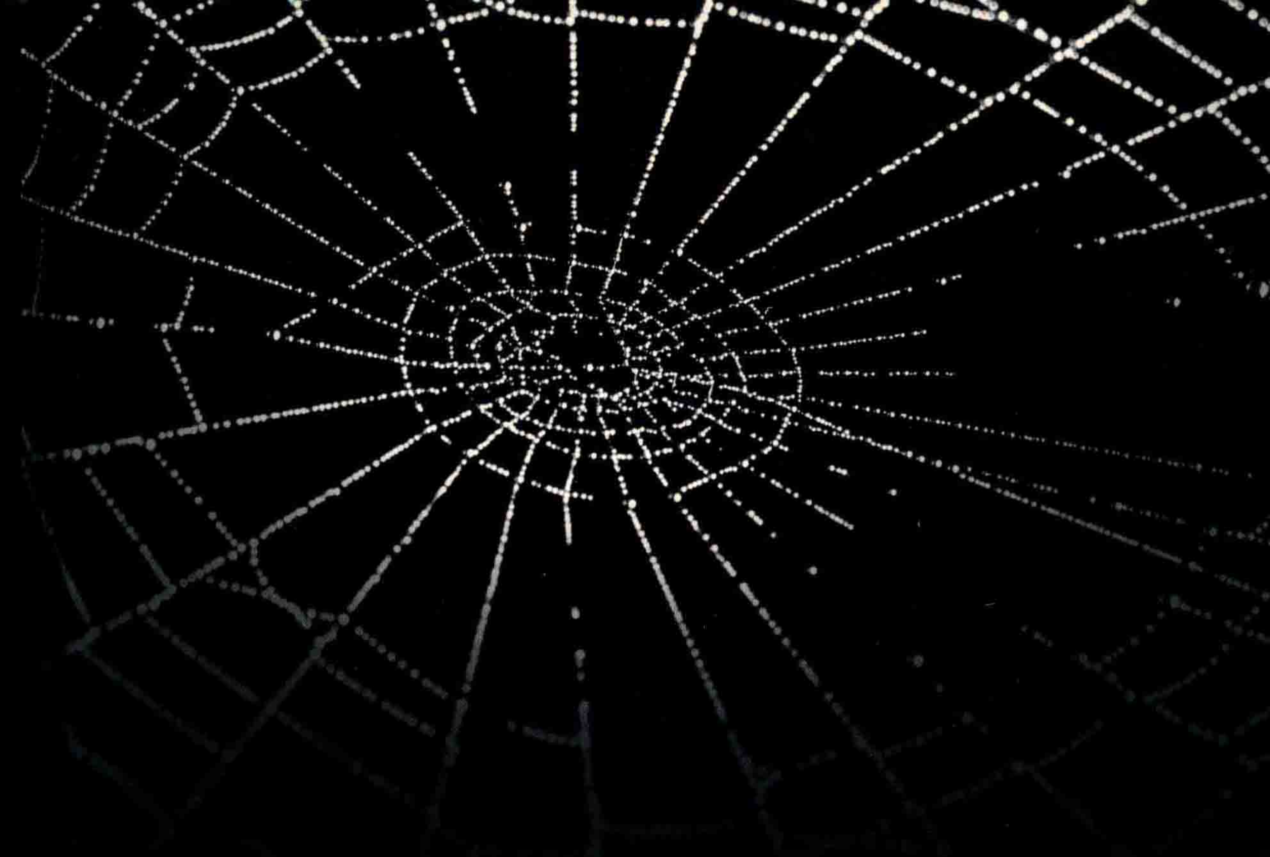

Complex Webs

Anticipating the Improbable

Bruce J. West
and **Paolo Grigolini**

CAMBRIDGE

Complex Webs

Anticipating the Improbable

BRUCE J. WEST
Army Research Office, USA

PAOLO GRIGOLINI
University of North Texas

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521113663

First published 2011

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

ISBN 978-0-521-11366-3 Hardback

Complex Webs

Complex Webs synthesizes modern mathematical developments with a broad range of complex network applications of interest to the engineer and system scientist, presenting the common principles, algorithms, and tools governing network behavior, dynamics, and complexity. The authors investigate multiple mathematical approaches to inverse power laws and expose the myth of normal statistics to describe natural and man-made networks. Richly illustrated throughout with real-world examples including cell phone use, accessing the Internet, failure of power grids, measures of health and disease, distribution of wealth, and many other familiar phenomena from physiology, bioengineering, biophysics, and informational and social networks, this book makes thought-provoking reading. With explanations of phenomena, diagrams, end-of-chapter problems, and worked examples, it is ideal for advanced undergraduate and graduate students in engineering and the life, social, and physical sciences. It is also a perfect introduction for researchers who are interested in this exciting new way of viewing dynamic networks.

Bruce J. West is Chief Scientist Mathematics with the Information Science Directorate at the Army Research Office, a position he has held for the last 10 years. After receiving his Ph.D. in Physics from the University of Rochester in 1970, he was Associate Director of a small private research institute (La Jolla Institute) for almost 20 years and a Professor at the University of North Texas for a decade. His research interests are in the nonlinear dynamics of complex networks. He has over 350 scientific publications, including 11 books and 8,500 citations, and he has received multiple academic and government awards for his research and publications.

Paolo Grigolini is currently a Professor in the Physics Department and the Center for Nonlinear Science at the University of North Texas. He is an internationally recognized theorist interested in the foundations of quantum mechanics, including wave-function collapse and the influence of classical chaos on quantum systems. His other research interests include the foundations of statistical physics, biophysical problems such as DNA sequencing, and the network science of human decision making and cognition.

Preface

The Italian engineer turned social scientist Vilfredo Pareto was the first investigator to determine that the income in western society followed a law that was fundamentally unfair. He was not making a value judgement about the poor and uneducated or about the rich and pampered; rather, he was interpreting the empirical finding that in 1894 the distribution of income in western societies was not "normal," but instead the number of people with a given income decreased as a power of the level of income. On bi-logarithmic graph paper this income distribution graphs as a straight-line segment of negative slope and is called an inverse power law. He interpreted his findings as meaning that a stable society has an intrinsic imbalance resulting from its complex nature, with the wealthy having a disproportionate fraction of the available wealth. Since then staggeringly many phenomena from biology, botany, economics, medicine, physics, physiology, psychology, in short every traditional discipline, have been found to involve complex phenomena that manifest inverse power-law behavior. These empirical laws were explained in the last half of the twentieth century as resulting from the complexity of the underlying phenomena.

As the twentieth century closed and the twenty-first century opened, a new understanding of the empirical inverse power laws emerged. This new understanding was based on the connectedness of the elements within the underlying phenomena and the supporting web structure. The idea of networks became pervasive as attention was drawn to society's reliance on sewers and the electric grid, cell phones and the Internet, banks and global stock markets, roads and rail lines, and the multitude of other human-engineered webbings that interconnect and support society. In parallel with the studies of social phenomena came new insight into the distribution in size and frequency of earthquakes and volcanic eruptions, global temperature anomalies and solar flares, river tributaries and a variety of other natural phenomena that have eluded exact description by the physical sciences. Moreover, the inverse power laws cropped up in unexpected places such as in heart rates, stride intervals and breathing, letter writing and emails, cities and wars, heart attacks and strokes; the inverse power law is apparently ubiquitous.

The synthesis of complexity and networks emphasizes the need for a new kind of scientific understanding, namely a grasp of how things work that exceeds the traditional mechanistic approach taken by science ever since Newton introduced gravity to explain planetary orbits and why things fall. The historical scientific approach reveals the workings of the two-body problem, but when three or more bodies interact in this way the strategy breaks down; chaos typically takes over and a different kind of thinking is required. This book is about how this new way of thinking has struggled to overcome the shackles of the "normal" distribution and the domination of the mean and standard deviation within the traditional disciplines.

The final result of our studies, ruminations and collaborations is not a standard textbook, although there are extended explanations of phenomena, diagrams, problems and worked-out examples. Instead this book has many characteristics associated with more idiosyncratic monographs; including well-labeled speculations, arguments that are meant to provoke response and stimulate thinking, and connections made to modern research discoveries that are usually denied to elementary texts. We have labored to get the mathematics right while not disregarding the language and the spirit of scientific discovery that is usually scrubbed from more traditional texts.

A number of people assisted us in preparing the present book. Gerhard Werner contributed directly by reading a draft of the manuscript and providing extensive feedback on how to improve its readability, suggestions that we took seriously. Our past students and present collaborators have contributed less directly to the manuscript and more directly to the research on which various sections of the book are based. The contributions of P. Allegrini, G. Aquino, F. Barbi, M. Bianucci, M. Bologna, M. Ignaccolo, M. Latka, A. Rocco and N. Scafetta are extensively referenced throughout, indicating their contributions to our fundamental theoretical understanding of the ubiquity of empirical inverse power-law distributions in complex webs.

Contents

1 Webs

The science of complex webs, also known as network science, is an exciting area of contemporary research, overarching the traditional scientific disciplines of biology, economics, physics, sociology and the other compartments of knowledge found in any college catalog. The transportation grids of planes, highways and railroads, the economic meshes of global finance and stock markets, the social webs of terrorism, governments, and businesses as well as churches, mosques, and synagogs, the physical lattices of telephones, the Internet, earthquakes, and global warming, in addition to the biological networks of gene regulation, the human body, clusters of neurons and food webs, share a number of apparently universal properties as the webs become increasingly complex. This conclusion is shared by the recent report *Network Science* [23] published under the auspices of the National Academy of Science. The terms networks and network science have become popular tags for these various areas of investigation, but we prefer the image of a web rather than the abstraction of a network, so we use the term web more often than the synonyms network, mesh, net, lattice, grille or fret. Colloquially, the term web entails the notion of entanglement that the name network does not share. Perhaps it is just the idea of the spider ensnaring its prey that appeals to our darker sides.

Whatever the intellectual material is called, this book is not about the research that has been done to understand complex webs, at least not in the sense of a monograph. We have attempted to put selected portions of that research into a pedagogic and often informal context, one that highlights the limitations of the more traditional descriptions of these areas. In this regard we are obligated to discuss the state of the art regarding a broad sweep of complex phenomena from a variety of academic disciplines. Sometimes properly setting the stage requires a historical approach and other times the historical view is replaced with personal perspectives, but with either approach we do not leave the reader alone to make sense of what can be difficult material. So we begin by illuminating the basic assumptions that often go unexamined in science.

1.1 The myth of normalcy

Natural philosophers metamorphosed into modern scientists in part by developing a passion for the quantifiable. But this belief in the virtue of numbers did not come about easily. In the time of Galileo Galilei (1564–1642), who was followed on his death by the birth of Sir Isaac Newton (1642–1727), experiments did not yield the kind of reproducible results we are accustomed to accepting today. Every freshman physics course has a laboratory experiment on measurement error that is intended to make students familiar with the fact that experiments are never exactly reproducible; there is always experimental error. But this pedagogic exercise is quickly forgotten, even when well learned. What is lost in the student's adjustment to college culture is that this is probably the most important experiment done during that first year. The implications of the uncertainty in scientific investigation extend far beyond the physics laboratory and are worthy of comments regarding their significance.

Most people recognize that they do not completely control their lives; whether it is the uncertainty in the economy and how it will affect a job, the unexpected illness that disrupts the planned vacation, or the death of one near and dear, all these things are beyond one's control. But there is solace in the belief, born of the industrial revolution, that we can control our destiny if only we had enough money, or sufficient prestige and, most recently, if we had an adequate amount of information. This is the legacy of science from the nineteenth and twentieth centuries, that the world can be controlled if only we more completely understood and could activate the levers of power. But is this true? Can we transfer the ideas of predictability and controllability from science to our everyday lives? Do the human sciences of sociology and psychology have laws in the same way that physics does?

In order to answer these and other similar questions it is necessary to understand how scientists have traditionally treated variability and uncertainty in the physical sciences. We begin with a focus on the physical sciences because physics was historically the first to develop the notion of quantification of physical laws and to construct the underlying mathematical infrastructure that enabled the physicist to view one physical phenomenon after another through the same lens and thereby achieve a fundamental level of understanding. One example of this unity of perspective is the atomic theory of matter, which enables us to explain much of what we see in the world, from the sun on our face to the rain wetting our clothes or the rainbow on the horizon, all from the same point of view. But the details are not so readily available.

We cannot predict exactly when a rain shower will begin, how long it will last, or how much ground it will cover. We might understand the basic phenomenon at the microscopic level and predict exactly the size and properties of molecules, but that does not establish the same level of certainty at the macroscopic level where water molecules fall as rain. The smallest seemingly unimportant microscopic variation is amplified by means of nonlinear interactions into a macroscopic uncertainty that is completely unpredictable. Therefore it appears to us that in order to develop defenses against the vagaries of life it is necessary to understand how science treats uncertainty and explore the limitations of those treatments. Most importantly, it is necessary to understand what science got right and what it got wrong. For that we go back to the freshman physics laboratory experiment on measurements.

Each time a measurement is made a certain amount of estimation is required, whether it is estimating the markings on a ruler or the alignment of a pointer on some gauge. If one measures a quantity q a given number of times, N say, then instead of having a single quantity Q the measurements yield a collection of quantities $Q_1, Q_2, \ldots, Q_N$. Such a collection is often called an ensemble and the challenge is to establish the best representation of the ensemble of measurements. Simpson, of "Simpson's rule" fame in the calculus, was the first scientist to recommend in print [30] that all the measurements taken in an experiment ought to be utilized in the determination of a quantity, not just those considered to be the most reliable, as was the custom in the seventeenth century. He was the first to recognize that the observed discrepancies between successively measured events follow a pattern that is characteristic of the ensemble of measurements. His observations were the forerunner to the *law of frequency of errors*, which asserts that there exists a relationship between the magnitude of an error and how many times it occurs in an ensemble of experimental results. Of course, the notion of an error implies that there is an exact value that the measurement is attempting to discern and that the variability in the data is a consequence of mistakes being made, resulting in deviations from the exact value, that is, in errors.

This notion of a correct value is an intriguing one in that it makes an implicit assumption about the nature of the world. Judges do not allow questions of the form "Have you stopped beating your wife?" because implicit in the question is the idea that the person had beaten his wife in the past. Therefore either answer, yes or no, confirms that the prisoner has beaten his wife in the past, which is, presumably, the question to be determined. Such leading questions are disallowed from the courtroom but are the bread and butter of science. Scientists are clever people and consequently they have raised the leading question to the level of hypothesis and turned the tables on their critics by asking "Have you measured the best value of this experimentally observed phenomenon?" Of course, either answer reinforces the idea of a best value. So what is this mysterious best value?

To answer this question we need to distinguish between statistics and probability; statistics has to do with measurements and data, whereas probability has to do with the mathematical theory of those measurements. Statistics arise because on the one hand individual results of experiments change in unpredictable ways and, on the other hand, the average values of long data sequences show remarkable stability. It is this statistical regularity that suggests the existence of a best value and hints at a mathematical model of the body of empirical data [8]. We point this out because it is not difficult to become confused over meaning in a discussion on the probability associated with a statistical process. The probability is a mathematical construct intended to represent the manner in which the fluctuating data are distributed over the range of possible values. Statistics represent the real world; probability represents one possible abstraction of that world that attempts to make quantitative deductions from the statistics. The novice should take note that the definition of probability is not universally accepted by the mathematical community. One camp interprets probability theory as a theory of *degrees of reasonable belief* and is completely disassociated from statistics in that a probability can be associated with any proposition, even one that is not reproducible. The second camp, with various degrees of subtlety, interprets probability theory in terms of the relative frequency of the occurrence of an event out of the universe of possible events. This second definition of probability is the one used throughout science and is adopted below.

Consider the relative number of times N_j a measurement error of a given magnitude occurs in a population of a given (large) size N; the relative frequency of occurrence of any particular error in this population is

$$p_j = \frac{N_j}{N}. \tag{1.1}$$

Here j indexes the above measurements into M different bins, where typically $N \gg M$. The relative number of occurrences provides an estimate of the probability that a measurement of this size will occur in further experiments. From this ensemble of N independent measurements we can define an average value,

$$\overline{Q} = \sum_{j=1}^{M} Q_j p_j, \tag{1.2}$$

with the average of a variable being denoted by the overbar and the N measurements put into M bins of equal size. The mean value $\overline{Q}$ is often thought to be an adequate characterization of the measurement and thus an operational definition of the experimental variable is associated with $\overline{Q}$. Simpson was the first to suggest that the mean value be accepted as the *best* value for the measured quantity. He further proposed that an isosceles triangle be used to represent the theoretical distribution in the measurements around the mean value. Of course, we now know that using the isosceles triangle as a measure of variability is wrong, but don't judge Simpson too harshly, after all, he was willing to put his reputation on the line and speculate on the possible solution to a very difficult scientific problem in his time and he got the principle right even if he got the equation wrong.

Subsequently, it was accepted that to be more quantitative one should examine the degree of variation of the measured value away from its average or "true" value. The magnitude of this variation is defined not by Simpson's isosceles triangles but by the standard deviation σ or the variance σ^2 of the measurements,

$$\sigma^2 \equiv \sum_{j=1}^{M} \left(Q_j - \overline{Q}\right)^2 p_j, \tag{1.3}$$

which, using the definition of the average and the normalization condition for the probability

$$\sum_{j=1}^{M} p_j = 1, \tag{1.4}$$

reduces to

$$\sigma^2 = \overline{Q^2} - \overline{Q}^2. \tag{1.5}$$

These equations are probably the most famous in statistics and form the basis of virtually every empirical theory of the physical, social and life sciences that uses discrete data sets.

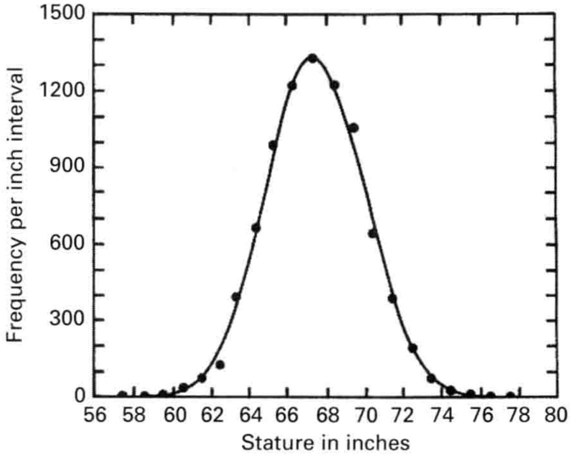

Figure 1.1. The dots denote the relative frequency of the heights of adult males in the British Isles [41]. The solid curve is the normal distribution with the same mean and variance as that of the data points.

In the continuum limit, that is, the limit in which the number of independent observations of a web variable approaches infinity, the characteristics of any measured quantity are specified by means of a distribution function. From this perspective any particular measurement has little or no meaning in itself; only the collection of measurements, the ensemble, has a scientific interpretation that is manifest through the distribution function. The distribution function is also called the probability density and it associates a probability with the occurrence of an event in the neighborhood of a measurement of a given size. For example in Figure 1.1 is depicted the frequency of occurrence of adult males of a given height in the general population of the British Isles. From this distribution it is clear that the likelihood of encountering a male six feet in height on your trip to Britain is substantial and the probability of meeting someone more than ten feet tall is zero.

Quantitatively, the probability of meeting someone with a height Q in the interval $(q, q + \Delta q)$ is given by the product of the distribution function and the size of the interval $P(q)\Delta q$. The solid curve in Figure 1.1 is given by a mathematical expression for the functional form of $P(q)$. Such a bell-shaped curve, whether from measurements of heights or from errors, is described by the well-known distribution of Gauss, and is also known as the normal distribution.

Half a century after Simpson's work the polymath Johann Carl Friedrich Gauss (1777–1855) [12] systematically investigated the properties of measurement errors and in so doing set the course of experimental science for the next two centuries. Gauss postulated that if each observation in a sequence of measurements $Q_1, Q_2, \ldots, Q_j, \ldots, Q_N$ was truly independent of all of the others then the deviation from the average value is a random variable

$$\xi_j = Q_j - \overline{Q}, \tag{1.6}$$

so that ξ_j and ξ_k are statistically independent of one another if $j \neq k$. This definition has the virtue of defining the average error in the measurement process to be zero,

$$\overline{\xi} = \sum_{j=1}^{N} \xi_j p_j = 0, \tag{1.7}$$

implying that the average value is the best representation of the data. Gauss determined that the variance defined by (1.3) in terms of the error (1.6) takes the form

$$\sigma^2 \equiv \sum_{j=1}^{N} \xi_j^2 p_j \tag{1.8}$$

and can be used to measure how well the average characterizes the ensemble of measurements. Note that it is not necessary to introduce p_j for the following argument and, although its introduction would not change the presentation in any substantial way, the discussion is somewhat simpler without it.

Gauss used the statistical independence of the measured quantities to prove that the average value gave their best representation and that, with a couple of physically reasonable assumptions, the associated statistical distribution was normal, an unfortunate name that had not been introduced at that time. We present a modified version of his arguments here to lay bare the requirements of normalcy. The probability I of obtaining a value in the interval $(Q, Q + \Delta Q)$ in any measurement is given by

$$I = P(Q)\Delta Q \tag{1.9}$$

and in a sequence of N measurements the data are replaced with the deviations from the average, that is, by the errors $\xi_1, \ldots, \xi_N$, allowing us to segment the range of values into N intervals,

$$P(\xi_j)\Delta\xi_j \equiv \text{ probability of observing the deviation } \xi_j. \tag{1.10}$$

In (1.10) the probability of making the N independent measurements in the ensemble together with the property that the probability of the occurrence of any two independent events is given by the product of their individual probabilities, and assuming $\Delta\xi_j = \Delta\xi$ for all j, is

$$I = \prod_{j=1}^{N} P(\xi_j)\Delta\xi_j = P(\xi_1)P(\xi_2)\ldots P(\xi_N)\Delta\xi^N. \tag{1.11}$$

According to Gauss the estimation of the value for Q appears *plausible* if the ensemble of measurements resulting in $\overline{Q}$ is the most *probable*. Thus, Q is determined in such a way that the probability I is a maximum for $Q = \overline{Q}$. To determine this form of the probability density we impose the condition

$$\frac{d \ln I}{d\overline{Q}} = 0 \tag{1.12}$$

and use (1.11) to obtain

$$\frac{d \ln I}{d\overline{Q}} = \sum_{j=1}^{N} \frac{d\xi_j}{d\overline{Q}} \frac{\partial \ln I}{\partial \xi_j} = -\sum_{j=1}^{N} \frac{\partial \ln P(\xi_j)}{\partial \xi_j}, \tag{1.13}$$

where we have used the fact that

$$\frac{d\xi_j}{d\overline{Q}} = -1$$

for all j. The constraint (1.12) applied to (1.13) is the mathematical rendition of the desirability of having the average value as the most probable value of the measured variable.

We now solve (1.13) subject to the constraint

$$\sum_{j=1}^{N} \xi_j = 0 \tag{1.14}$$

by assuming that the jth derivative of the logarithm of the probability density can be expanded as a polynomial in the random error

$$\frac{\partial \ln P\left(\xi_j\right)}{\partial \xi_j} = \sum_{k=0}^{\infty} C_k \xi_j^k, \tag{1.15}$$

where the set of constants $\{C_k\}$ is determined by the equation of constraint

$$-\sum_{j=1}^{N}\sum_{k=0}^{\infty} C_k \xi_j^k = 0. \tag{1.16}$$

All the coefficients in (1.16) vanish except $k = 1$, since by definition the fluctuations satisfy the constraint equation (1.14) so the coefficient $C_1 \neq 0$ satisfies the constraint. Thus, we obtain the equation for the probability density

$$\frac{\partial \ln P(\xi_j)}{\partial \xi_j} = C_1 \xi_j, \tag{1.17}$$

which integrates to

$$P(\xi_j) \propto \exp\left(\frac{C_1}{2}\xi_j^2\right). \tag{1.18}$$

The first thing to notice about this solution is that its extreme value occurs at $\xi_j = 0$, that is, at $Q_j = \overline{Q}$ as required. For this to be a maximum as Gauss required and Simpson speculated, the constant must be negative, $C_1 < 0$, so that the second derivative of P at the extremum is positive. With a negative constant the function decreases symmetrically to zero on either side, allowing the function to be normalized,

$$\int_{-\infty}^{\infty} P(\xi_j) d\xi_j = 1, \tag{1.19}$$

and because of this normalization the function can be interpreted as a probability density. Moreover, we can calculate the variance to be

$$\sigma_j^2 = \int_{-\infty}^{\infty} \xi_j^2 P(\xi_j) d\xi_j, \tag{1.20}$$

allowing us to express the normalized probability density as