FROM MEASUREMENT ERRORS TO NORMAL DISTRIBUTIONS: A BRIEF HISTORY AND ITS PEDAGOGICAL IMPLICATIONS

Ilhan M. Izmirli
George Mason University
iizmirl2@gmu.edu

ABSTRACT

Normal distributions, arguably the most pervasive constructs in statistics, provide us with very important data distribution patterns. As such, they are among the most fundamental concepts introduced in a basic statistics course. Not only are they immensely useful thanks to the Central Limit Theorem derived by Laplace in 1778, but they also afford a very convenient way of establishing the idea of continuous distributions without using calculus.

The normal distributions have, of course, vastly significant applications and numerous examples are given in introductory texts of those that are related to such wide-ranging topics from anatomy to finance. One very important application, however, usually goes unmentioned: physical quantities that are expected to be the sum of several independent processes (such as measurement errors), often have a distribution that is approximately normal.

Ironically, the development of the normal distributions can, in fact, be traced back to the formal study of errors, starting with Roger Cotes' (1682 – 1716), and continuing on with many distinguished scholars such as Daniel Bernoulli (1700 – 1782), Thomas Bayes (1701 – 1761), Thomas Simpson (1710 – 1761), Johann Heinrich Lambert (1728 – 1777), Joseph Louis Lagrange (1736 – 1813), Pierre-Simon Laplace (1749 – 1827), culminating finally in the works of Carl Friedrich Gauss (1777 – 1855).

We claim that this historical progression has important pedagogical implications. In fact, we will conclude our paper by the assertion that it is much more natural and pedagogically much sounder to introduce the concept of normality following this historical progression.

Another approach to the historical development of the normal distribution would be to follow Abraham de Moivre (1667 – 1754) and introduce it as an approximation to the binomial distribution as articulated in his 1733 paper ***Approximatio ad summam terminorum binomii $(a+b)^n$ in seriem expansi***. This path, which is equally illuminating and edifying, will be pursued in a different paper.

For considerations of space, the historical development given in this paper omits most of the derivational details. For a much more detailed approach with mathematical derivations see Stahl (2006).

1. A BRIEF HISTORY OF MEASUREMENT ERRORS

A meticulous analysis of measurement errors was first necessitated by the existence of several distinct numerical estimates of the same quantity in astronomical observations (Placket 1958), and dates all the way back to Hipparchus (190BCE – 120BCE) and Ptolemy (90CE – 168CE).

Some centuries later, the idea of taking repetitive measurements of the same quantity was incorporated into the practice of astronomy by Tycho Brahe (1546 – 1601): after taking several repeated measurements of a quantity, one would somehow use them to arrive at a single numerical value.

There was, however, no consensus among scholars of the time either on how many such measurements should be taken or on how to obtain a representative value. If

$x_1, x_2, \ldots, x_n$ were the measures, astronomers sometimes used the number that would minimize the function

$$f(x) = (x - x_1)^2 + (x - x_2)^2 + \cdots + (x - x_n)^2$$

i.e., the mean of these observations. At other times, they used the number that would minimize the function

$$f(x) = |x - x_1| + |x - x_2| + \cdots + |x - x_n|$$

i.e., the median of these observations.[1]

Sometimes the representative value was obtained by inexplicable even rather enigmatic methods (Plackett 1988). For example, Kepler, who had obtained the values

$$134^0\ 23'\ 39'',\ 134^0\ 27'\ 37'',\ 134^0\ 23'\ 18'',\ 134^0\ 29'48''$$

for the right ascension of Mars, used the representative value of $134^0\ 24'33''$, which is neither the mean ($134^0\ 26'\ 5.5''$) nor the median ($134^0\ 25'\ 38''$) of these numbers (Donahue 1992).

Galileo (1564-1642) noted in his famous ***Dialogue Concerning the Two Chief Systems of the World—Ptolemaic and Copernican*** published in 1632 that all measurements would be hampered with errors (Hald 1990). He went on to claim that these errors would be distributed symmetrically about zero and that small errors would occur more frequently than large errors

> " . . . it is plausible that the observers are more likely to have erred little than much . . . " [Drake 1967, p. 308].

a clear harbinger of the bell shaped curve.

The beginnings of a formal study of the theory of errors may be traced back to Roger Cotes' ***Opera Miscellanea***. In this work that was published posthumously in 1722, Roger Cotes made the following suggestion (given here in modern language and notation): If $x_1, x_2, \ldots, x_n$ are values obtained from subsequent observations, and $w_1, w_2, \ldots, w_n$ are the weights reciprocally proportional to the displacements which may arise from the errors in the single observations, as the representative value one should choose the center of gravity, which is the weighted average (Cotes 1722).

This, in fact, can be thought of as the early stages of the method of least squares: We want to represent the data $x_1, x_2, \ldots, x_n$, by the x value that minimizes the function

$$f(x) = w_1(x - x_1)^2 + w_2(x - x_2)^2 + \cdots + w_n(x - x_n)^2$$

Since we have

$$f'(x) = 2w_1(x - x_1) + 2w_2(x - x_2) + \cdots + 2w_n(x - x_n)$$

solving the equation $f'(x) = 0$, we obtain

$$x = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n}$$

In a memoir prepared in 1755 (printed 1756), where the axiom that positive and negative errors are equally probable was posited, Thomas Simpson discussed several possible distributions of error. He first considered the uniform distribution and then the discrete and the continuous symmetric triangular distribution (Simpson 1756).

[1] In fact, the eventual pronouncement to favor mean over the median had an unassailable role on the evolution of the normal distribution.

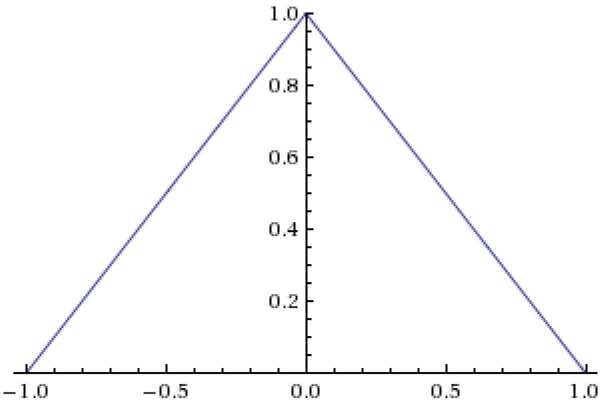

Figure 1. The Symmetric Triangular Distribution of Simpson on $[-1, 1]$: $f(x) = 1 - |x|$

In 1751, the Croatian astronomer/physicist/mathematician Ruder Boškovic (1711 – 1787), together with Christopher Maire (1697-1767), an English Jesuit, began to measure an arc of two degrees between Rome and Rimini, and later wrote a book to describe this process. In this 1755 work, ***De Litteraria expeditione per pontificiam ditionem ad dimetiendos duos meridiani gradus a PP. Maire et Boscovicli***, Boškovic proposed that the true value of a series of observations would be that which minimizes the sum of absolute errors; of course, in modern terminology this proposed value would be the median of the observations.

Johann Heinrich Lambert in his 1765 book ***Anlage zur Architectonic*** proposed the semi-ellipse as a distribution of errors:

$$f(x) = \frac{1}{2}\sqrt{1 - x^2}$$

where $-1 \leq x \leq 1$.

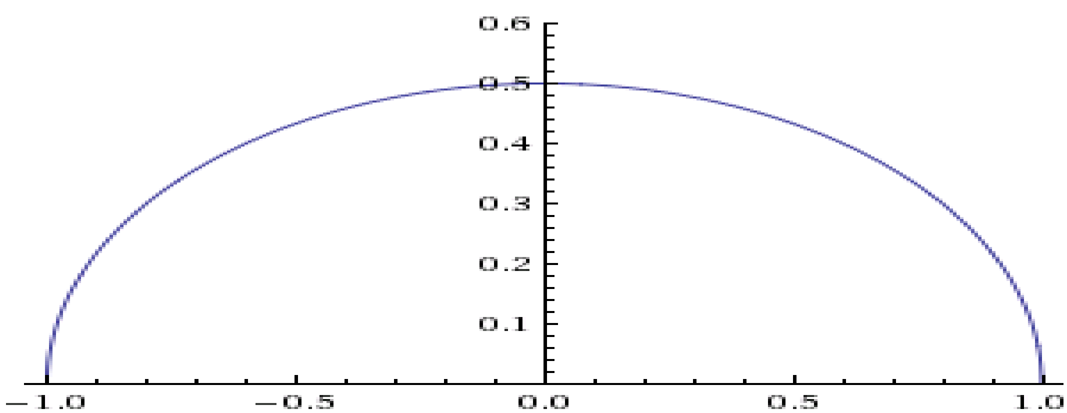

Figure 2. Graph of $f(x) = \frac{1}{2}\sqrt{1 - x^2}$, Lambert's Error Distribution

Pierre-Simon Laplace (1774) made the first attempt to deduce a rule for the distribution of errors using probabilistic principles, and claimed that the frequency of an error could be expressed as an exponential function of its magnitude once its sign was disregarded. He also stipulated that this function must be symmetric in x and monotone decreasing for $x > 0$.

This distribution is now known as the ***Laplace distribution***. A random variable has a Laplace distribution with parameters μ and ν if its probability density function is

$$f(x|\mu,\nu) = \frac{1}{2\nu}\exp\left(-\frac{|x-\mu|}{\nu}\right)$$
$$= \frac{1}{2\nu}\begin{cases}\exp\left(-\frac{\mu-x}{\nu}\right) & if\ x < \mu \\ \exp\left(-\frac{x-\mu}{\nu}\right) & if\ x \geq \mu\end{cases}$$

Here, μ is the mean and $2\nu^2$ is the variance.

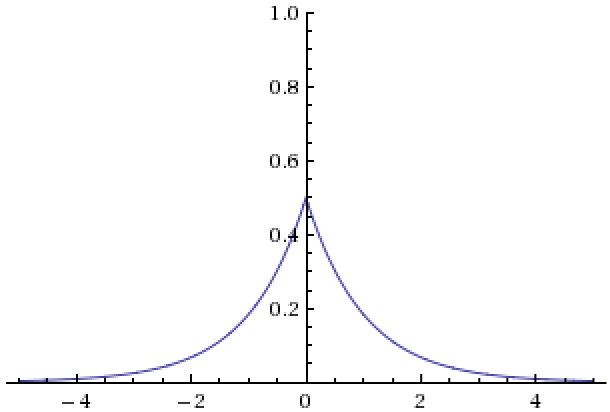

Figure 3. Graph of $f(x) = \frac{1}{2}e^{-|x|}$, the Laplace Distribution with $\nu = 1$ and $\mu = 0$

If $\mu = 0$, for $x \geq \mu = 0,$ this reduces to an exponential distribution scaled by $\frac{1}{2}$.

Whereas the normal distribution is expressed in terms of the squared difference from the mean, as the expression for the above defined density function indicates, the Laplace density is expressed in terms of the absolute difference from the mean. Consequently, the Laplace distribution has fatter tails than the normal distribution. Moreover, it is not differentiable at $x = 0$, but there is no indication that Laplace was in any way disturbed by this fact. In fact, as we shall see, his second proposed curve has a more drastic singularity at $x = 0$.

Three years later, possibly unhappy with the first distribution he suggested, Laplace proposed an alternative curve (See Gillispie 1979 for a reprint of his article):

$$f(x) = \frac{1}{2\alpha}\ln\left(\frac{\alpha}{|x|}\right)$$

for $-\alpha \leq x \leq \alpha$, where α is the supremum of all the possible errors. This function has a discontinuity at $x = 0$.

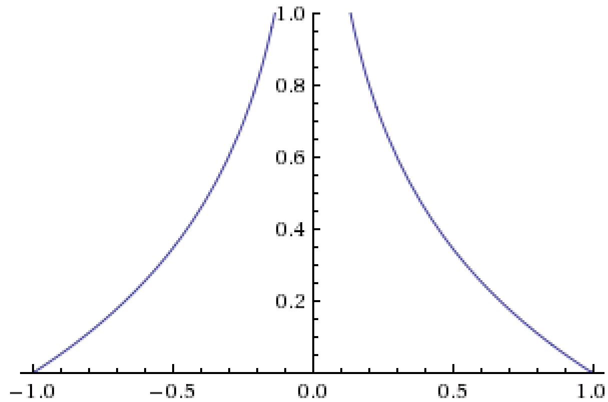

Figure 4. Graph of Laplace's Second Error Function with $\alpha = 1, f(x) = \frac{1}{2}\ln\left(\frac{1}{|x|}\right)$.

For more details see Bidwell (1923).

Shortly thereafter, in 1777, Daniel Bernoulli wrote in a paper titled ***Dijudicatio maxime probabilis plurium observationam discrepantium atque versimillima induction inde formanda*** (The most probable choice between several discrepant observations and the formation therefrom of the most likely induction)

> … the observations are added together and the sum divided by the number of observations; the quotient is then accepted as the true value of the required quantity, until better and more certain information is obtained. In this way, if the several observations can be considered as having, as it were, the same weight, the center of gravity is accepted as the true position of the object under investigation.

He also notes that errors are not equally likely:

> But is it right to hold that the several observations are of the same weight or moment or equally prone to any and every error? Are errors of some degrees as easy to make as others of as many minutes? Is there everywhere the same probability? Such an assertion would be quite absurd, which is undoubtedly the reason why astronomers prefer to reject completely observations which they judge to be too wide of the truth, while retaining the rest and, indeed, assigning to them the same reliability.

For more details see the translation of this paper by Allen (1961).

Based on these, Bernoulli suggested the semi-ellipse as such a curve, which, following a scaling argument, he then replaced with a semicircle.

The famous French mathematician and physicist Joseph Louis Lagrange was also interested in the study of error distributions and in 1781 suggested two functions for this purpose. The first is now known as the ***raised cosine distribution***. The raised cosine distribution is a continuous probability distribution supported on the interval $[\mu - \theta, \mu + \theta]$ with probability density function

$$f(x|\mu,\theta) = \begin{cases} \frac{1}{2\theta}\left(1 + \cos\left(\frac{x-\mu}{\theta}\pi\right)\right) & if\ \mu - \theta \le x \le \mu + \theta \\ 0 & otherwise \end{cases}$$

Its mean is μ and its variance is $\left(\frac{\pi^2-6}{3\pi^2}\right)\theta^2$

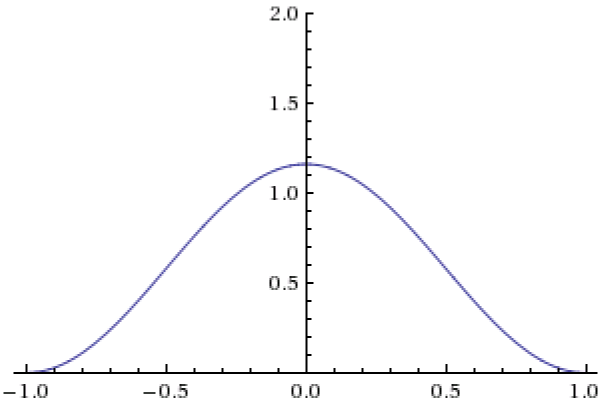

Figure 5. Graph of the Standard Raised Cosine Distribution ($\mu = 0$ and $\theta = 1$), $f(x) = \frac{1}{2}(1 + \cos \pi x)$

Lagrange's second suggestion was the ***logarithmic distribution***, a discrete distribution with probability mass function

$$f(k) = -\frac{1}{\ln(1-p)} \frac{p^k}{k}$$

for $k \geq 1$, and $0 < p < 1$, which would be a decreasing sequence of points from the maximum value of $-\frac{p}{\ln(1-p)}$ attained at $k = 1$.

Since

$$-\ln(1-p) = p + \frac{p^2}{2} + \frac{p^3}{3} + \cdots$$

it is easy to show that

$$\sum_{k=1}^{\infty} f(k) = 1$$

Moreover, clearly,

$$\mu = -\frac{1}{\ln(1-p)} \frac{p}{1-p}$$

and

$$\sigma^2 = \frac{-p(p + \ln(1-p))}{(1-p)^2 ln^2(1-p)}$$

The next important development had its roots in a celestial event, namely the discovery of Ceres, claimed to be a new planet by its discoverer the Italian priest and astronomer Giuseppe Piazzi (1746 – 1826). Inopportunely, before enough observations were made to establish its orbit precisely, Ceres disappeared behind the sun and was not expected to reemerge for an extended period of time. Not willing to wait that long, astronomers started suggesting areas of the sky to be searched. Gauss proposed an area of the sky quite different from those suggested by the other astronomers, which in fact, turned out to be the right area. See Teets and Whitehead (1999) for more details.

Gauss explained that he used the least squares criterion to locate the orbit that best fit the observations (Davis, 1963). For, Gauss claimed, a theory of measurement errors must be based on the following three assumptions:

1. Small errors are more likely than large errors.
2. For any real number ϵ the likelihood of errors of magnitudes $-\epsilon$ and ϵ are equal.

3. In case of repeated measurements of the same quantity, the most likely value of the quantity being measured is the average of these repeated measurements.

On the basis of these assumptions and using nothing beyond elementary calculus, Gauss proved that the probability density function for the error (that is, the error function) should be

$$f(x) = \frac{C}{\sqrt{\pi}} e^{-C^2x^2}$$

with the constant $C > 0$. This, of course, is the renowned ***normal distribution*** [2] (or the bell curve, o the Gaussian curve) with mean at zero and standard deviation $\frac{1}{C\sqrt{2}}$.

Actually, this result was implicitly stated in the Central Limit Theorem published by Laplace in 1810 (Gillispie 1979), which implied that if the error curve of a single observation is symmetric, then the error curve of the sum of several observations was indeed approximated by a Gaussian curve. Hence if one assumes that the error involved in an individual observation is the aggregate of a large number of errors, then this theorem predicts that the random error that occurs in that individual observation is controlled by a Gaussian curve: error, in modern terminology, is approximately normally distributed.

2. CONCLUSION

As the research in the field of mathematics education unambiguously and unmistakably demonstrates, most modern pedagogues soundly endorse the use of historical development of concepts in mathematics classes; they agree that this approach contributes immensely to students' appreciation of mathematics as well as their comprehension of theories and/or models of varying degrees of complexity. BUT NOT IN STAT.

That the inclusion of the historic development of ideas has significant pedagogical benefits in the teaching of mathematics is elaborated, for instance, in Wilson and Chavuot (2000). They claim that this approach to teaching

> … sharpens problem-solving skills, lays a foundation for better understanding, helps students make mathematical connections, and highlights the interaction between mathematics and society (Wilson & Chauvot 2000, p. 642).

Jankvist (2009) mentions that history as a pedagogical tool can give new perspectives of the material and can be of patent assistance in the cognitive difficulties students may encounter as they learn a particular mathematical topic.

We should also assert that integrating the history of the topic into classroom lectures has an added utility: it humanizes the subject matter. In other words, it puts a human face to mathematics which otherwise would be

> … closed, dead, emotionless and all discovered (Bidwell 1993, p. 461).

In addition to the benefits specified above, Jankvist (2009) identifies more gains that can be had by using the history of mathematics: increased motivation- through generating interest and excitement and decreased intimidation - and the realization that mathematics is a human creation. In fact, as expounded in Marshall and Rich,

---

[2] The use of the term normal is probably due to the influence of Karl E. Pearson who said in 1920

> Many years ago [in 1893] I called the Laplace-Gaussian curve the *normal* curve, which name, while it avoids the international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another *abnormal* (Pearson 1920).

> … history has a vital role to play in today's mathematics classrooms. It allows students and teachers to think and talk about mathematics in meaningful ways. It demythologizes mathematics by showing that it is the creation of human beings. History enriches the mathematics curriculum. It deepens and broadens the knowledge that students construct in mathematics class (Marshall and Rich 2000, p. 706).

Moreover, this pedagogical approach helps students understand that any theory in any study area including one that is as esoteric and arcane as mathematics, indeed goes through many false starts, progresses and regresses before it eventually reaches an acceptable level that is mutually agreed upon by the community of mathematicians. It also attests that many different people from many diverse and dissimilar cultures contribute to it. So, students get to understand that no theorem or result is born in the polished form it is presented in the textbooks, and no theorem or result is in the realm of a few geniuses who never make mistakes. Indeed, scientific progress is a collective effort that moves along a path, parts of which are marked with erroneous turns.

Clearly, the above mentioned problems related to mathematics classes are analogously prominent in statistics classes as well. There also we need to help students develop their problem solving skills. Statistics is also conceived as closed, dead, emotionless, all discovered, and intimidating by most students. It is, in parts, also perceived to be esoteric and arcane. A majority of students tend to think that a few intelligent people came up with these ideas and possibly except for a few simple applications, the future progress of the subject matter remains in the domain of those people.

Thus, I hypothesize that all the above arguments made in favor of incorporating historic development of concepts into mathematics classes are equally valid for statistics classes as well: students in statistics classes would benefit equally well from an infusion of historic development of processes into these courses.

It would contribute immensely to students' appreciation of statistics and their comprehension of concepts of varying degrees of complexity. It can be used to sharpen problem-solving skills, to help students make statistical connections, and to show the interaction between statistics and society to paraphrase Wilson & Chauvot (2000). It can be used as a tool to give new perspectives of the material. It can be used to humanize the subject matter. It helps with increased motivation- through generating interest and excitement and decreased intimidation – and with the realization that statistics is a human creation. To paraphrase Marshall and Rich (2000), it allows students and teachers to think and talk about statistics in meaningful ways. It demythologizes statistics by showing that it is the creation of human beings.

If we look at the particular example given in this paper, we can easily see that we can use this pedagogical approach to show students that ideas in statistics were not created in a vacuum by a few intelligent people. They were and still are created collectively by many scholars from many different cultural and economic backgrounds as means to answer societal and scientific needs of the times. They go through many false starts, progresses and regresses before they eventually reach an acceptable level that is mutually agreed upon by the community of statisticians.

Yet, most of us do not make a conscientious effort to include the historical development of the topics we cover in our classes. Major points made for this omission are that such a coverage is not directly related to the basic ideas we want to convey to the students or that we do not have enough time for such "digressions" in our already loaded schedules. However such customary (and by now quite predictable) objections are rather insubstantial and can easily be refuted, for example, by the present paper, a discussion which should not take more than one lecture and is an excellent example of the "real-life" uses of the normal distribution.

REFERENCES

Allen C. G. 1961. (Translator) Dijudicatio maxime probabilis plurium observationam discrepantium atque versimillima induction inde formanda. *Acta Acad. Sci. Imp. Petrop.,* **1** (1777) 3–23 by Daniel Bernoulli. English translation: The most probable choice between several discrepant observations and the formation therefrom of the most likely induction. *Biometrica*, **48** (1961) 1–18.

Bidwell, J. K. 1993. Humanize Your Classroom with The History of Mathematics. *Mathematics Teacher*, 86, 461-464.

Cotes, R. 1722. Aestimatio Errorum in Mixta Mathesi, per Variationes Partium Trianguli Plani et Sphaerici. In *Opera Miscellanea* Cambridge.

Donahue, W. H. 1992. (Translator). *New Astronomy* by Johannes Kepler. Cambridge: Cambridge University Press.

Drake, S. 1967. (Translator). *Dialogue Concerning the Two Chief World Systems—Ptolemaic and Copernican* by Galileo Galilei. (Second Edition.) Berkeley: Univ. California Press.

Davis, H. C. 1963. (Translator). *Theoria Motus Corporum Celestium*. Hamburg, Perthes et Besser, 1809, by C. F. Gauss. Translated as *Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections.* Boston: Little, Brown, 1857. (Reprint New York: Dover 1963.)

Gillispie C. 1979. Mémoires inédits ou anonymes de Laplace sur la théorie des erreurs, les polynomes de Legendre, et la philosophie des probabilités. *Révue d'histoire des sciences* **32** (1979), 223–279. pp. 228–256.

Hald A. 1990. *A History of Probability and Statistics and Their Applications before 1750*. New York: John Wiley & Sons.

Haverhals, Nick and Matt Roscoe. 2010. The history of mathematics as a pedagogical tool: Teaching the integral of the secant via Mercator's projection. *The Montana Mathematics Enthusiast*, ISSN 1551-3440, Vol. 7, nos.2&3, pp.339-368

Jankvist, U. T. 2009. A characterization of the "whys" and "hows" of using history in mathematics education. *Educational Studies in Mathematics, 71*(3), 235 – 261.

Johnson, Norman Lloyd, Adrienne W Kemp and Samuel Kotz. 2005. *Univariate discrete distributions* (Third Edition). New York: John Wiley & Sons.

Laplace, P. S. 1774. Mémoire sur la probabilité des causes par les évènements. *Mémoires de l'Académie Royale des Sciences Présentés par Divers Savants*, 6, (1774) 621–656.

Laplace, P. S. 1809. Mémoire sur les approximations des formules qui sont fonctions de très grand nombres et sur leur applications aux probabilités. *Mémoires de l'Academie des sciences de Paris*, 1809, pp. 353–415, 559–565.

Marshall, G. L. and B. S. Rich. 2000. The Role of History in a Mathematics Class. *Mathematics Teacher*, 93 (8), 704-706.

Pearson, Karl. 1920. *The Science of Man: its Needs and its Prospects*. Cambridge: Cambridge University Press.

Placket, R. L. 1958. The principle of the arithmetic mean, *Biometrika* **45** (1958), 130–135.

Plackett, R. L. 1988. Data Analysis before 1750, *International Statistical Review*, **56** (1988), No. 2, 181–195.

Simpson, Thomas. 1756. A Letter to the Right Honourable George Macclesfield, President of the Royal Society, on the Advantage of taking the Mean, of a Number of Observations, in Practical Astronomy. *Phil. Trans.* **49** (1756), 82–93.

Stahl, Saul. 2006. The Evolution of the Normal Distribution. *Mathematics Magazine*. Vol. 79, No. 2, April 2006, pp 96-113.

Teets, D. and K.Whitehead. 1999. The discovery of Ceres: how Gauss became famous. *Mathematics Magazine*. **72**:2 (1999), 83–93.

Wilson, Edwin Bidwell. 1923. First and second laws of error. *Journal of the American Statistical Association*, 18 (143), 841-851

Wilson, P. S. and J. B Chauvot. 2000. Who? How? What? A Strategy for Using History to Teach Mathematics. *Mathematics Teacher*, 93(8), 642-645.