Why Two Sigma? A Theoretical Justification

Hung T. Nguyen[1], Vladik Kreinovich[2], Gennady N. Solopchenko[3], and Chin-Wang Tao[4]

[1] Department of Mathematical Sciences, New Mexico State University, Las Cruces, NM 88003, USA, email hunguyen@nmsu.edu
[2] Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA, email vladik@cs.utep.edu
[3] Information-Measurement Department, State Technical University of St. Petersburg, Politechnicheskaya 29, St. Petersburg, 195251 Russia, email sol@study.ssl.stu.neva.ru
[4] Department of Electrical Engineering, National I-Lan Institute of Technology, 260 I-Lan, Taiwan, email cwtao@mail.ilantech.edu.tw

Abstract. For a normal distribution, the probability density $\rho(x)$ is everywhere positive, so in principle, all real numbers are possible. In reality, the probability that a random variable is far away from the mean is so small that this possibility can be often safely ignored. Usually, a small real number k is picked (e.g., 2 or 3); then, with a probability $P_0(k) \approx 1$ (depending on k), the normally distributed random variable with mean a and standard deviation σ belongs to the interval $\mathbf{a} = [a - k \cdot \sigma, a + k \cdot \sigma]$.

The actual error distribution may be non-Gaussian; hence, the probability $P(k)$ that a random variable belongs to $\mathbf{a}$ differs from $P_0(k)$. It is desirable to select k for which the dependence of $P_0(k)$ on the distribution is the smallest possible. Empirically, this dependence is the smallest for $k \in [1.5, 2.5]$. In this paper, we give a theoretical explanation for this empirical result.

1 Formulation of the Problem

For many measuring instruments, the measurement error is normally distributed; see, e.g., [10,11]. This known empirical fact has a good theoretical explanation (see, e.g., [2–4,13]; see also [14] pp. 2.17, 6.5, 9.8, and references therein): Usually, the manufacturers of the measuring instruments have made their best effort to eliminate the major sources of measurement error. The resulting measurement error comes from a variety of small independent error sources, and thus, can be described as a sum of a large number of small independent random variables. According to the central limit theorem, such a sum, under reasonable conditions, converges to normal distribution. Thus, if there are sufficiently many small random components, the resulting error distribution is indeed close to normal.

For a normal distribution, the probability density $\rho(x)$ is positive for all x, so in principle, all real numbers are possible. In reality, however, the probability of a random variable to be far away from the mean is so small that in many practical applications, this possibility can be safely ignored. So, the values of a normally distributed random variable are located, with a reasonably high probability, within a finite interval. To implement this idea, in practice, usually, a small real number k is picked (typically, $k = 2$ or $k = 3$). Then, with a probability $P_0(k) \approx 1$ (depending on k), the values of a normally distributed random variable with mean a and standard deviation σ belong to the interval $[a - k \cdot \sigma, a + k \cdot \sigma]$. For a normal distribution, this probability does not depend on a and σ, only on k. For $k = 2$, we have $P_0(k) \approx 0.95$; for $k = 3$, we have $P_0(k) \approx 0.999$; for $k = 6$, we have $P_0(k) \approx 1 - 10^{-6}$, etc.

For a *normal* distribution, we can pick an arbitrary k and get the interval which contains all the values with the corresponding probability $P_0(k)$. In many real-life situations, however, the actual error distribution is close to Gaussian but not exactly normal [10,11]. The deviation from a Gaussian distribution can be characterized by one or several parameters ε (so that Gaussian distribution corresponds to $\varepsilon = 0$). For a non-Gaussian distribution characterized by a parameter ε, the probability $P(k, \varepsilon)$ that a random variable belongs to the interval $\mathbf{a} = [a - k \cdot \sigma, a + k \cdot \sigma]$ is, in general, different from $P_0(k)$. It is therefore desirable to select k for which the dependence of $P(k, \varepsilon)$ on ε is the smallest possible, i.e., for which we can guarantee that $P(\eta \in \mathbf{a}) \approx P_0(k)$ irrespective of whether η is normally distributed or not.

The empirical analysis of actual probability distributions of different measuring instruments show that the smallest possible dependence occurs when k is between 1.5 and 2.5 [10]. This empirical fact is an important part of measurement practice, but until now, it has not been theoretically explained – not because we get a difficult-to-solve precisely formulated statistical problem, but because, due to uncertainty, this informal problem is very difficult to formalize in precise terms.

In this paper, we show how this problem can be formalized, and we show that this formalization indeed justifies the empirical choice of $k \in [1.5, 2.5]$.

2 Selecting a Class of Non-Gaussian Probability Distributions

2.1 Selecting Distributions Can Be Reduced to Selecting Functions

In principle, there can be many different probability distributions which are close to Gaussian. For different possible deviations from the Gaussian distribution, different values of k may be the least sensitive to the corresponding deviations. Therefore, before we start looking for the optimal value of k, we must first select a reasonable class of such deviations.

In many practical situations, there is no reason why a positive error value x should be more probable or less probable than the corresponding negative value $-x$, so we can assume that these probabilities coincide, and the probability distribution is symmetric w.r.t. $x \to -x$ (i.e., the corresponding probability density function $\rho(x)$ is even).

In particular, we will consider symmetric Gaussian distributions, i.e., Gaussian distributions with zero mean. It is known that an arbitrary distribution of this type, with an arbitrary standard deviation σ, can be obtained from a "standard" Gaussian distribution (with zero mean and unit standard deviation) by a linear transformation $f(z) = \sigma \cdot z$. In other words, if ζ is a random variable which is distributed according to the standard Gaussian distribution, then the variable $\eta = f(\zeta) = \sigma \cdot \zeta$ is distributed according to the Gaussian law with zero mean and standard deviation σ.

One can show that non-Gaussian distributions can be obtained in a similar manner, but with possibly non-linear increasing functions $f(z)$. Indeed, an arbitrary probability distribution can be described by its cumulative distribution function (cdf) $F(x) = P(\eta \le x)$. Let $F_0(x) = P(\zeta \le x)$ denote a cdf which corresponds to the Gaussian distribution. Let us show that by choosing an appropriate function $f(z)$, we can make $\eta = f(\zeta)$ have the desired cdf $F(x)$. Indeed, since we are only considering increasing functions $f(z)$, the inequality $f(\zeta) \le x$ is equivalent to $\zeta \le f^{-1}(x)$, where $f^{-1}(x)$ is the function which is inverse to $f(x)$ (i.e., $f^{-1}(x) = z$ if and only if $f(z) = x$). Thus, to guarantee that $P(f(\zeta) \le x) = F(x)$ for all x, we must guarantee that $P(\zeta \le f^{-1}(x)) = F(x)$. Since ζ is distributed according to the standard Gaussian distribution, we have $P(\zeta \le f^{-1}(x)) = F_0(f^{-1}(x))$. Thus, we must guarantee that for every x, we have $F(x) = F_0(f^{-1}(x))$. If we denote $z = f^{-1}(x)$, then we have $x = f(z)$, and the desired equality takes the form $F(f(z)) = F_0(z)$, hence we can take $f(z) = F^{-1}(F_0(z))$. So, every distribution can indeed be described as $\eta = f(\zeta)$ for an appropriate function $f(z)$.

Since we only consider symmetric distribution, these increasing functions $f(z)$ have to be odd: $f(-z) = -f(z)$. Hence, $f(0) = 0$, $f(z) \ge 0$ for $z \ge 0$, and to reconstruct the entire function $f(z)$, it is sufficient to know its values for $z \ge 0$.

In view of this representation of a arbitrary probability distribution by a transformation function $f(z)$, instead of selecting a class of probability distributions, we can select a class of functions $f(z)$. Then, we will be able to use, as new random variables, combinations $\eta = f(\zeta)$, where ζ has a standard Gaussian distribution and $f(z)$ is one of the selected functions.

The question is: How to select the "best" (most appropriate) functions $f(z)$?

2.2 Best In What Sense?

What do we mean by "the best"? It is not so difficult to come up with different criteria for choosing a functions $f(z)$:

- We may want to choose the function $f(z)$ for which the average distance $D(f)$ between the resulting probability distribution and the actual empirical distributions of measuring instruments is the smallest possible.
- We may also want to choose the function $f(z)$ for which the *average computation time* $C(f)$ of some statistical processing algorithms is the smallest (average in the same of some reasonable probability distribution on the set of all problems).

At first glance, the situation seems hopeless: it is difficult to feasibly estimate these numerical criteria even for a single function $f(z)$, so it may look like we therefore cannot undertake an even more ambitious task of finding the *optimal* function $f(z)$. Hopefully, the situation is not as hopeless as it may seem, because there is a symmetry-based formalism (actively used in the foundations of fuzzy, neural, genetic computations, see, e.g., [8]) which will enable us to find the optimal function $f(z)$.

2.3 We Must Choose a Family of Functions

If we simply replace the original measurement unit by a new unit which is C times smaller, then all the numerical values of the measurement error η get multiplied by C. Thus, if the function $f(z)$ (which describes the original probability distribution) is a reasonable transformation function, then the function $C \cdot f(z)$ which corresponds to the same distribution expressed in the new units is also reasonable. Thus, with every function $f(z)$, all the functions $C \cdot f(z)$ should be selected as well, the whole *family* of functions $\{C \cdot f(z)\}$ (characterized by a parameter $C > 0$) must be selected.

Thus, instead of selecting the "best" (more appropriate) *functions*, we should talk about selecting the best *families*.

In the following text, we will denote families of functions by caligraphic capital letters, such as $\mathcal{F}$, $\mathcal{F}_i$, $\mathcal{G}$, etc.

2.4 An Optimality Criterion Can Be Non-Numeric

Traditionally, optimality criteria are *numerical*, i.e., to every family $\mathcal{F}$, we assign some value $J(\mathcal{F})$ expressing its quality, and choose a family for which this value is minimal (i.e., when $J(\mathcal{F}) \le J(\mathcal{G})$ for every other alternative $\mathcal{G}$). However, it is not necessary to restrict ourselves to such numeric criteria only.

For example, if we have several different families $\mathcal{F}$ that have the same average distance $D(\mathcal{F})$, we can choose between them the one that has the minimal computational time $C(\mathcal{F})$. In this case, the actual criterion that we use to compare two families is not numeric, but more complicated: A family $\mathcal{F}_1$ is better than the family $\mathcal{F}_2$ if and only if either $D(\mathcal{F}_1) < D(\mathcal{F}_2)$, or $D(\mathcal{F}_1) = D(\mathcal{F}_2)$ and $C(\mathcal{F}_1) < C(\mathcal{F}_2)$.

The only thing that a criterion *must* do is to allow us, for every pair of families $(\mathcal{F}_1, \mathcal{F}_2)$, to make one of the following conclusions:

- the first family is better with respect to this criterion (we'll denote it by $\mathcal{F}_1 \succ \mathcal{F}_2$, or $\mathcal{F}_2 \prec \mathcal{F}_1$);
- with respect to the given criterion, the second family is better ($\mathcal{F}_2 \succ \mathcal{F}_1$);
- with respect to this criterion, the two families have the same quality (we'll denote it by $\mathcal{F}_1 \sim \mathcal{F}_2$);
- this criterion does not allow us to compare the two families.

Of course, it is necessary to demand that these choices be consistent. For example, if $\mathcal{F}_1 \succ \mathcal{F}_2$ and $\mathcal{F}_2 \succ \mathcal{F}_3$ then $\mathcal{F}_1 \succ \mathcal{F}_3$.

2.5 Optimality Criterion Must Be Final

A natural demand is that this criterion must choose a *unique* optimal family (i.e., a family that is better with respect to this criterion than any other family). The reason for this demand is very simple.

If a criterion *does not choose* any family at all, then it is of no use.

If *several* different families are the best according to this criterion, then we still have the problem of choosing the best among them. Therefore we need some additional criterion for that choice, like in the above example: If several families $\mathcal{F}_1, \mathcal{F}_2, \ldots$ turn out to have the same average distance ($D(\mathcal{F}_1) = D(\mathcal{F}_2) = \ldots$), we can choose among them a family with minimal computation time ($C(\mathcal{F}_i) \to \min$).

So what we actually do in this case is abandon that criterion for which there were several "best" families, and consider a new "composite" criterion instead: $\mathcal{F}_1$ is better than $\mathcal{F}_2$ according to this new criterion if either it was better according to the old criterion, or they had the same quality according to the old criterion and $\mathcal{F}_1$ is better than $\mathcal{F}_2$ according to the additional criterion.

In other words, if a criterion does not allow us to choose a unique best family, it means that this criterion is not final, we'll have to modify it until we come to a final criterion that will have that property.

2.6 The Criterion Must Not Change If We Change the Measuring Unit Corresponding to the Original Gaussian Distribution

The exact mathematical form of a function $f(z)$ depends on the exact choice of units for measuring the original normally distributed variable ζ. If we replace this unit by a new unit that is λ times larger, then the same physical value that was previously described by a numerical value ζ will now be described, in the new units, by a new numerical value $\widetilde{\zeta} = \zeta/\lambda$.

How will the expression for $f(z)$ change if we use the new units? In terms of $\widetilde{\zeta}$, we have $\zeta = \lambda \cdot \widetilde{\zeta}$. Thus, the variable η which was originally represented by a function $f(\zeta)$, will be described, in the new units, as $f\left(\lambda \cdot \widetilde{\zeta}\right)$, i.e., as $\widetilde{f}\left(\widetilde{\zeta}\right)$, where $\widetilde{f}(z) = f(\lambda \cdot z)$.

There is no reason why one choice of a unit should be preferable to another. Therefore, it is reasonable to assume that the relative quality of different families should not change if we simply change the units, i.e., if the family $\mathcal{F}$ is better than a family $\mathcal{G}$, then the transformed family $\widetilde{\mathcal{F}}$ should also be better than the family $\widetilde{\mathcal{G}}$.

We are now ready for the formal definitions.

2.7 Definitions and the Main Result

Definition 1. *Let $f(z)$ be a differentiable strictly increasing function from real numbers to non-negative real numbers. By a family that corresponds to this function $f(z)$, we mean a family of all functions of the type $\widetilde{f}(z) = C \cdot f(z)$, where $C > 0$ is an arbitrary positive real number. (Two families are considered equal if they coincide, i.e., consist of the same functions.)*

In the following text, we will denote the set of all possible families by Φ.

Definition 2. *By an optimality criterion, we mean a consistent pair $\langle \prec, \sim \rangle$ of relations on the set Φ of all alternatives which satisfies the following conditions, for every $\mathcal{F}, \mathcal{G}, \mathcal{H} \in \Phi$:*

1. *if $\mathcal{F} \prec \mathcal{G}$ and $\mathcal{G} \prec \mathcal{H}$ then $\mathcal{F} \prec \mathcal{H}$;*
2. *$\mathcal{F} \sim \mathcal{F}$;*
3. *if $\mathcal{F} \sim \mathcal{G}$ then $\mathcal{G} \sim \mathcal{F}$;*
4. *if $\mathcal{F} \sim \mathcal{G}$ and $\mathcal{G} \sim \mathcal{H}$ then $\mathcal{F} \sim \mathcal{H}$;*
5. *if $\mathcal{F} \prec \mathcal{G}$ and $\mathcal{G} \sim \mathcal{H}$ then $\mathcal{F} \prec \mathcal{H}$;*
6. *if $\mathcal{F} \sim \mathcal{G}$ and $\mathcal{G} \prec \mathcal{H}$ then $\mathcal{F} \prec \mathcal{H}$;*
7. *if $\mathcal{F} \prec \mathcal{G}$ then $\mathcal{G} \not\prec \mathcal{F}$ and $\mathcal{F} \not\sim \mathcal{G}$.*

Comment. The intended meaning of these relations is as follows:

- $\mathcal{F} \prec \mathcal{G}$ means that with respect to a given criterion, $\mathcal{G}$ is better than $\mathcal{F}$;
- $\mathcal{F} \sim \mathcal{G}$ means that with respect to a given criterion, $\mathcal{F}$ and $\mathcal{G}$ are of the same quality.

Under this interpretation, conditions *1.-7.* have simple intuitive meaning; e.g., the condition *1.* means that if $\mathcal{G}$ is better than $\mathcal{F}$, and $\mathcal{H}$ is better than $\mathcal{G}$, then $\mathcal{H}$ is better than $\mathcal{F}$.

Definition 3.

- *We say that an alternative $\mathcal{F}$ is optimal (or best) with respect to a criterion $\langle \prec, \sim \rangle$ if for every other alternative $\mathcal{G}$ either $\mathcal{F} \succ \mathcal{G}$ or $\mathcal{F} \sim \mathcal{G}$.*
- *We say that a criterion is final if there exists an optimal alternative, and this optimal alternative is unique.*

Definition 4. *Let $\lambda > 0$ be a positive real number.*

- *By a* λ*-rescaling of a function* $f(x)$ *we mean a function* $\widetilde{f}(x) = f(\lambda \cdot x)$.
- *By a* λ*-rescaling* $R_\lambda(\mathcal{F})$ *of a family of functions* $\mathcal{F}$ *we mean the family consisting of* λ*-rescalings of all functions from* $\mathcal{F}$.

Definition 5. *We say that an optimality criterion on* Φ *is unit-invariant if for every two families* $\mathcal{F}$ *and* $\mathcal{G}$ *and for every number* $\lambda > 0$, *the following two conditions are true:*

i) if $\mathcal{F}$ *is better than* $\mathcal{G}$ *in the sense of this criterion (i.e.,* $\mathcal{F} \succ \mathcal{G}$*), then* $R_\lambda(\mathcal{F}) \succ R_\lambda(\mathcal{G})$*;*
ii) if $\mathcal{F}$ *is equivalent to* $\mathcal{G}$ *in the sense of this criterion (i.e.,* $\mathcal{F} \sim \mathcal{G}$*), then* $R_\lambda(\mathcal{F}) \sim R_\lambda(\mathcal{G})$.

Theorem 1. *If a family* $\mathcal{F}$ *is optimal in the sense of some optimality criterion that is final and unit-invariant, then every function* $f(z)$ *from this family* $\mathcal{F}$ *has the form* $C \cdot z^\alpha$ *for some real numbers* C *and* α.

Comment. For the convenience of the readers, all the proofs are placed in the last section.

Since $f(z)$ is an odd function, we can therefore conclude that the corresponding random variable η can be described as $\eta = \mathrm{sign}(\zeta) \cdot |\zeta|^\alpha$, where ζ is a standard Gaussian random variable, i.e., a normally distributed random variable with zero mean and unit standard deviation. This is indeed a good description for empirical distributions of measurement error [10].

Gaussian variables correspond to $\alpha = 1$; so, since we are interested in distributions which are close to Gaussian, we should consider α close to 1, i.e., $\alpha = 1 + \varepsilon$ for some small ε.

3 Selecting the Optimal Value of k

Let us consider the class of probability distributions described in the previous section. We want to find k for which the dependence of $P(k, \varepsilon)$ on ε is the smallest. Since empirical distributions are close to normal, we have $\varepsilon \approx 0$. For $\varepsilon \approx 0$, we can neglect quadratic and higher order terms in the dependence of $P(k, \varepsilon)$ on ε, and conclude that $P(k, \varepsilon) \approx P_0(k) + \varepsilon \cdot P_1(k)$, where

$$P_1(k) = \frac{\partial P(k, \varepsilon)}{\partial \varepsilon}_{|\varepsilon=0}. \tag{1}$$

Thus, this dependence is the smallest if and only if the absolute value $|P_1(k)|$ of the coefficient $P_1(k)$ is the smallest possible. It turns out $|P_1(k)|$ achieves its smallest value $|P_1(k)| = 0$ for some k which is indeed close to the interval $[1.5, 2.5]$, thus justifying the above empirical fact:

Definition 6. *We say that the value k is the least sensitive to the possible non-Gaussian character of the probability distribution if for this k, the expression $|P_1(k)|$, where $P_1(k)$ is determined by the formula (1), attains the smallest possible value.*

The formulation of the result uses the Euler constant

$$\gamma = \lim_{n\to\infty}\left(1 + \frac{1}{2} + \ldots + \frac{1}{n} - \ln(n)\right) \approx 0.577.$$

Theorem 2. *The value $k = \dfrac{e}{\sqrt{2}\cdot e^{\gamma/2}} \approx 1.44$ is the least sensitive to the possible non-Gaussian character of the probability distribution, and for this k, we have $|P_1(k)| = 0$.*

Comment. For this value k, $P_0(k) \approx 0.85$, so at least 85% of the values of the random variable lie in the interval $[a - k\cdot\sigma, a + k\cdot\sigma]$. This value is in good accordance with common sense, namely, with the 20-80 "Pareto" law, according to which:

- 20% of the people drink 80% of the beer,
- 20% of the researchers write 80% of all papers etc.

This number is also in good accordance with the experimental fact that from each 25 rules typically discovered by a data mining system, approximately 20 (i.e., about 80%) are already known (see, e.g., [7]). It is probably worth mentioning that in [12], we give an alternative explanation of this same fact – by using fuzzy logic techniques (see, e.g., [6,9]) instead of probabilities.

4 Proofs

4.1 Proof of Theorem 1

This proof is based on the following lemma:

Lemma. *If an optimality criterion is final and unit-invariant, then the optimal family $\mathcal{F}_{\rm opt}$ is also unit-invariant, i.e., $R_\lambda(\mathcal{F}_{\rm opt}) = \mathcal{F}_{\rm opt}$ for every number λ.*

Proof of the Lemma. Since the optimality criterion is final, there exists a unique family $\mathcal{F}_{\rm opt}$ that is optimal with respect to this criterion, i.e., for every other $\mathcal{F}$, either $\mathcal{F}_{\rm opt} \succ \mathcal{F}$, or $\mathcal{F}_{\rm opt} \sim \mathcal{F}$.

To prove that $\mathcal{F}_{\rm opt} = R_\lambda(\mathcal{F}_{\rm opt})$, we will first show that the re-scaled family $R_\lambda(\mathcal{F}_{\rm opt})$ is also optimal, i.e., that for every family $\mathcal{F}$: either $R_\lambda(\mathcal{F}_{\rm opt}) \succ \mathcal{F}$, or $R_\lambda(\mathcal{F}_{\rm opt}) \sim \mathcal{F}$.

If we prove this optimality, then the desired equality will follow from the fact that our optimality criterion is final and therefore, there is only one optimal family (so, since the families $\mathcal{F}_{\rm opt}$ and $R_\lambda(\mathcal{F}_{\rm opt})$ are both optimal, they must be the same family).

Let us show that $R_\lambda(\mathcal{F}_{\rm opt})$ is indeed optimal. How can we, e.g., prove that $R_\lambda(\mathcal{F}_{\rm opt}) \succ \mathcal{F}$? Since the optimality criterion is unit-invariant, the desired relation is equivalent to $\mathcal{F}_{\rm opt} \succ R_{\lambda^{-1}}(\mathcal{F})$. Similarly, the relation $R_\lambda(\mathcal{F}_{\rm opt}) \sim \mathcal{F}$ is equivalent to $\mathcal{F}_{\rm opt} \sim R_{\lambda^{-1}}(\mathcal{F})$.

These two equivalences allow us to complete the proof of the lemma. Indeed, since $\mathcal{F}_{\rm opt}$ is optimal, we have one of the two possibilities: either $\mathcal{F}_{\rm opt} \succ R_{\lambda^{-1}}(\mathcal{F})$, or $\mathcal{F}_{\rm opt} \sim R_{\lambda^{-1}}(\mathcal{F})$. In the first case, we have $R_\lambda(\mathcal{F}_{\rm opt}) \succ \mathcal{F}$; in the second case, we have $R_\lambda(\mathcal{F}_{\rm opt}) \sim \mathcal{F}$.

Thus, whatever family $\mathcal{F}$ we take, we always have either $R_\lambda(\mathcal{F}_{\rm opt}) \succ \mathcal{F}$, or $R_\lambda(\mathcal{F}_{\rm opt}) \sim \mathcal{F}$. Hence, $R_\lambda(\mathcal{F}_{\rm opt})$ is indeed optimal and thence, $R_\lambda(\mathcal{F}_{\rm opt}) = \mathcal{F}_{\rm opt}$. The lemma is proven.

Let us now prove the theorem. Since the criterion is final, there exists an optimal family $\mathcal{F}_{\rm opt} = \{C \cdot f(z)\}$. Due to the lemma, the optimal family is unit-invariant.

From unit-invariance, it follows that for every λ, there exists a real number $A(\lambda)$ for which $f(\lambda \cdot z) = A(\lambda) \cdot f(z)$. Since the function $f(z)$ is differentiable, we can conclude that the ratio $A(\lambda) = f(\lambda \cdot z)/f(z)$ is differentiable as well. Thus, we can differentiate both sides of the above equation with respect to λ, and substitute $\lambda = 1$. As a result, we get the following differential equation for the unknown function $f(z)$:

$$z \cdot \frac{df}{dz} = \alpha \cdot f,$$

where by α, we denoted the value of the derivative $dA/d\lambda$ taken at $\lambda = 1$. Moving terms dz and z to the right-hand side and all the term containing f to the left-hand side, we conclude that

$$\frac{df}{f} = \alpha \cdot \frac{dz}{z}.$$

Integrating both sides of this equation, we conclude that $\ln(f) = \alpha \cdot \ln(z) + C$ for some constant C, and therefore, that $f(z) = \text{const} \cdot z^\alpha$. The theorem is proven.

4.2 Proof of Theorem 2

As we have shown in Section 2, for each ε, the corresponding random variable η can be described as $\eta = \text{sign}(\zeta) \cdot |\zeta|^{1+\varepsilon}$, where ζ is a standard Gaussian random variable, i.e., a normally distributed random variable with zero mean and unit standard deviation. For this random variable, the mean is equal to $E_\varepsilon(\eta) = 0$. Let $\sigma(\varepsilon) = \sqrt{E_\varepsilon(\eta^2)}$ denote its standard deviation. Then, the probability $P(k, \varepsilon)$ is equal to the probability that $\eta \in [-k \cdot \sigma(\varepsilon), k \cdot \sigma(\varepsilon)]$, i.e., to the probability that $|\eta| \le k \cdot \sigma(\varepsilon)$.

Since $|\eta| = |\zeta|^{1+\varepsilon}$, the probability $P(k,\varepsilon)$ is equal to the probability that $|\zeta|^{1+\varepsilon} \le k \cdot \sigma(\varepsilon)$, i.e., that for a standard Gaussian random variable ζ, we have $|\zeta| \le B(\varepsilon)$, where we denoted $B(\varepsilon) = (k \cdot \sigma(\varepsilon))^{1/(1+\varepsilon)}$. In other words, $P(k,\varepsilon) = F_{\rm erf}(B(\varepsilon))$, where we denoted

$$F_{\rm erf}(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-z}^{z} \exp\left(-\frac{t^2}{2}\right)\,{\rm d}t. \tag{2}$$

Due to the chain rule, $P_1(k) = F'(B(0)) \cdot B'(0)$, where $F'_{\rm erf}$ and B' denote derivatives with respect to ε. The function $F_{\rm erf}(z)$ is a strictly increasing function of z, with $F'_{\rm erf}(z) > 0$ for all z. Hence, $P_1(k) = 0$ if and only if $B'(0) = 0$.

By definition, $B(\varepsilon) = (k \cdot \sigma(\varepsilon))^{1/(1+\varepsilon)} = \exp(b(\varepsilon))$, where we denoted

$$b(\varepsilon) = \frac{\ln(k) + \ln(\sigma(\varepsilon))}{1+\varepsilon}. \tag{3}$$

Therefore, $B'(\varepsilon) = \exp(b(\varepsilon)) \cdot b'(\varepsilon)$. The first factor in this product is always positive, so $B'(0) = 0$ if and only if $b'(0) = 0$. Let use the equation $b'(0) = 0$ to determine the desired value k. Differentiating the above expression for $b(\varepsilon)$ and substituting $\varepsilon = 0$, we conclude that $(\ln(\sigma(\varepsilon))' - \ln(k) - \ln(\sigma(\varepsilon)) = 0$, i.e., $\sigma'(0)/\sigma(0) - \ln(k) - \ln(\sigma(0)) = 0$. For $\varepsilon = 0$, we have a standard Gaussian distribution, for which $\sigma(0) = 1$. Thus, the above equation takes the form $\sigma'(0) - \ln(k) = 0$, hence $\ln(k) = \sigma'(0)$ and

$$k = \exp(\sigma'(0)). \tag{4}$$

To complete our proof, let us find the explicit expression for $\sigma'(0)$. By definition,

$$\sigma^2(\varepsilon) = E_\varepsilon(\eta^2) = E_0(|\zeta|^{2+2\varepsilon}) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} |\zeta|^{2+2\varepsilon} \cdot \exp\left(-\frac{\zeta^2}{2}\right)\,{\rm d}\zeta. \tag{5}$$

Since negative and positive values ζ lead to an equal contribution to this integral, we can conclude that

$$\sigma^2(\varepsilon) = \frac{2}{\sqrt{2\pi}} \cdot \int_0^{\infty} \zeta^{2+2\varepsilon} \cdot \exp\left(-\frac{\zeta^2}{2}\right)\,{\rm d}\zeta =$$

$$\sqrt{\frac{2}{\pi}} \cdot \int_0^{\infty} \zeta^{2+2\varepsilon} \cdot \exp\left(-\frac{\zeta^2}{2}\right)\,{\rm d}\zeta. \tag{6}$$

To simplify this integral, we introduce a new variable $z = \zeta^2/2$; then $\zeta = 2^{1/2} \cdot z^{1/2}$, ${\rm d}\zeta = (\sqrt{2}/2) \cdot z^{-1/2} \cdot {\rm d}z$, and hence,

$$\sigma^2(\varepsilon) = \frac{2^{1+\varepsilon}}{\sqrt{\pi}} \cdot \int_0^{\infty} z^{1/2+\varepsilon} \cdot \exp(-z)\,{\rm d}z. \tag{7}$$

By definition of a gamma function

$$\Gamma(n) = \int_0^{\infty} z^{n-1} e^{-z}\, \mathrm{d}z \qquad (8)$$

(see, e.g., [1], p. 350; [5], Appendix A, Table 17), we thus have

$$\sigma^2(\varepsilon) = \frac{1}{\sqrt{\pi}} \cdot 2^{1+\varepsilon} \cdot \Gamma\left(\frac{3}{2} + \varepsilon\right). \qquad (9)$$

For $\varepsilon = 0$, we have $\sigma(0) = 1$ and $\Gamma(3/2) = \sqrt{\pi}/2$ (see [1]), so this equality clearly holds.

Differentiating both sides of the equality (9) with respect to ε, we conclude that

$$2\sigma(\varepsilon) \cdot \sigma'(\varepsilon) = \frac{1}{\sqrt{\pi}} \cdot \left(2^{1+\varepsilon} \cdot \ln(2) \cdot \Gamma\left(\frac{3}{2} + \varepsilon\right) + 2^{1+\varepsilon} \cdot \Gamma'\left(\frac{3}{2} + \varepsilon\right)\right). \qquad (10)$$

Substituting $\varepsilon = 0$, taking into consideration that $\Gamma(3/2) = \sqrt{\pi}/2$, and dividing both sides of the resulting equality by 2, we conclude that

$$\sigma'(0) = \frac{\ln(2)}{2} + \frac{\Gamma'(3/2)}{\sqrt{\pi}}. \qquad (11)$$

To compute $\Gamma'(3/2)$, we can use the following known equality (see, e.g., [5]):

$$\Gamma(z) \cdot \Gamma\left(z + \frac{1}{2}\right) = (2\pi)^{1/2} \cdot 2^{1/2-2z} \cdot \Gamma(2z); \qquad (12)$$

hence,

$$\Gamma\left(z + \frac{1}{2}\right) = \frac{\Gamma(2z)}{\Gamma(z)} \cdot 2 \cdot \sqrt{\pi} \cdot 2^{-2z}. \qquad (13)$$

In particular, for $z = 1 + \varepsilon$, we get

$$\Gamma\left(\frac{3}{2} + \varepsilon\right) = \frac{\Gamma(2 + 2\varepsilon)}{\Gamma(1 + \varepsilon)} \cdot \frac{\sqrt{\pi}}{2} \cdot 2^{-2\varepsilon}. \qquad (14)$$

One of the main properties of a gamma function is that $\Gamma(n+1) = n \cdot \Gamma(n)$; hence $\Gamma(2 + 2\varepsilon) = (1 + 2\varepsilon) \cdot \Gamma(1 + 2\varepsilon)$, and the equation (13) takes the form:

$$\Gamma\left(\frac{3}{2} + \varepsilon\right) = \frac{\Gamma(1 + 2\varepsilon) \cdot (1 + 2\varepsilon)}{\Gamma(1 + \varepsilon)} \cdot \frac{\sqrt{\pi}}{2} \cdot 2^{-2\varepsilon}. \qquad (15)$$

It is known [1] that $\Gamma'(1)$ is equal to $-\gamma$, where γ is the Euler's constant. Thus, for small ε, $\Gamma(1 + \varepsilon) = 1 - \gamma \cdot \varepsilon + o(\varepsilon)$, $\Gamma(1 + 2\varepsilon) = 1 - 2\gamma \cdot \varepsilon + o(\varepsilon)$, and $2^{-2\varepsilon} = e^{-2\varepsilon \cdot \ln(2)} = 1 - 2\ln(2) \cdot \varepsilon + o(\varepsilon)$. Hence, the equation (15) takes the form:

$$\Gamma\left(\frac{3}{2}+\varepsilon\right) = \frac{(1-2\gamma\cdot\varepsilon)\cdot(1+2\varepsilon)}{(1-\gamma\cdot\varepsilon)}\cdot\frac{\sqrt{\pi}}{2}\cdot(1-2\ln(2)\cdot\varepsilon)+o(\varepsilon) =$$

$$(1+\varepsilon\cdot(2-\gamma-2\ln(2))\cdot\frac{\sqrt{\pi}}{2}+o(\varepsilon). \tag{16}$$

Thus,

$$\Gamma'\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}\cdot(2-\gamma-2\ln(2)) = \sqrt{\pi}\cdot\left(1-\frac{\gamma}{2}-\ln(2)\right). \tag{17}$$

Substituting (17) into (11), we conclude that

$$\sigma'(0) = 1-\frac{\gamma}{2}-\frac{\ln(2)}{2}. \tag{18}$$

From the formula (4), we can now get the desired expression for k. The theorem is proven.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, by NSF grants No. DUE-9750858 and CDA-9522207, by the United Space Alliance, grant No. NAS 9-20000 (PWO C0C67713A6), by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number F49620-95-1-0518, and by the National Security Agency under Grant No. MDA904-98-1-0561.

References

1. Beyer, W. H. (1991) CRC Standard Mathematical Tables and Formulae, CRC Press, Boca Raton, FL.
2. Clifford, A. A. (1973) Multivariate Error Analysis, Wiley, New York.
3. Fuller, W. A. (1987) Measurement Error Models, Wiley, New York.
4. H. G. Hecht, H. G. (1990) Mathematics in Chemistry. An Introduction to Modern Methods, Prentice Hall, Englewood Cliffs, NJ.
5. Itô, K., ed. (1993) Encyclopedic Dictionary of Mathematics, MIT Press, Cambridge, MA.
6. Klir, G., Yuan, B. (1995) Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice Hall, Upper Saddle River, NJ.
7. Kruse, R., Borgelt, C., and Nauck, D. (1999) Fuzzy data analysis: challenges and perspectives, Proceedings of the 8th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'99), Seoul, Korea, August 22–25, 1999, **3**, 1211–1216.

8. Nguyen, H. T., Kreinovich, V. (1997) Applications of Continuous Mathematics to Computer Science, Kluwer, Dordrecht.
9. Nguyen, H. T., Walker, E. A. (1999) First Course in Fuzzy Logic, CRC Press, Boca Raton, FL.
10. Novitskii, P. V., Zograph, I. A. (1991) Estimating the Measurement Errors, Energoatomizdat, Leningrad (in Russian).
11. Orlov, A. I. (1991) How often are the observations normal? Industrial Laboratory **57**, No. 7, 770–772.
12. Osegueda, R. A., Ferregut, C., Kreinovich, V., Seetharami, S., Schulte, H. (2000) Fuzzy (granular) levels of quality, with applications to data mining and to structural integrity of aerospace structures, Proceedings of the 19th International Conference of the North American Fuzzy Information Society NAFIPS'2000, Atlanta, Georgia, July 13–15, 2000 (to appear).
13. Rabinovich, S. (1993) Measurement Errors: Theory and Practice, American Institute of Physics, New York.
14. Wadsworth, H. M. Jr., ed. (1990) Handbook of Statistical Methods for Engineers and Scientists, McGraw-Hill Publishing Co., New York.