Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2013, Article ID 131424, 4 pages
http://dx.doi.org/10.1155/2013/131424

Research Article

The Probability of a Confidence Interval Based on Minimal Estimates of the Mean and the Standard Deviation

Louis M. Houston

The Louisiana Accelerator Center, The University of Louisiana at Lafayette, LA 70504-4210, USA

Correspondence should be addressed to Louis M. Houston; houston@louisiana.edu

Received 22 January 2013; Revised 9 May 2013; Accepted 11 May 2013

Academic Editor: Jin L. Kuang

Using two measurements, we produce an estimate of the mean and the sample standard deviation. We construct a confidence interval with these parameters and compute the probability of the confidence interval by using the cumulative distribution function and averaging over the parameters. The probability is in the form of an integral that we compare to a computer simulation.

1. Introduction

A confidence interval is an interval in which a measurement falls with a given probability [1]. This paper addresses the question of defining the probability of a confidence interval that can be constructed when only a minimal number of measurements are known.

The problem has a couple of significant applications. In geophysics, time-lapse 3D seismic monitoring is used to monitor oil and gas reservoirs before and after production. Because of the high cost of 3D seismic monitoring, a minimal-effort time-lapse 3D seismic monitoring approach was proposed by Houston and Kinsland [2]. This approach proposes a minimal measurement interval based on preliminary measurements. It implies successively smaller seismic surveys as monitoring continues and knowledge of the behavior of the reservoir grows. The success of the minimal-effort method is based on the probability that the reservoir is detected by a seismic survey. The minimal-effort method is based on a minimal number of measurements, which connects it to the topic of this paper.

Another significant application of a confidence interval based on a minimal number of measurements is the problem of cancer treatment based on the irradiation of a tumor [3]. The purpose of irradiation is to destroy the tumor, but targeting the tumor can be problematic because tumors often exhibit small random motions within the body. Because of random tumor motion, the result of radiation treatment often includes the destruction of healthy tissue. Clearly, it is beneficial that cancer radiation treatment incorporates minimal intervals. The success of the cancer treatment is based on the probability that the tumor is irradiated while using a minimal irradiation interval. This success hinges on the ability of the radiation treatment to incorporate preliminary measurements that are often minimal. It is upon this basis that the cancer treatment problem connects to the topic of this paper.

In this paper, we construct a confidence interval based on minimal estimates of the mean and the standard deviation. Explicitly, we use two measurements to specify an interval that contains a subsequent measurement with a given probability. The mathematical effort includes the derivation of a specific probability for the confidence interval. The probability is computed using the cumulative distribution function, and this probability is averaged over both the estimate of the mean and the sample standard deviation to yield a specific value. The results are compared to a computer simulation that estimates the probability based on frequency.

2. Probability Based on the Cumulative Distribution Function

The cumulative distribution function [4] is given as

$$F\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2}\,dt, \tag{1}$$

where x is a measurement, μ is the mean, and σ is the standard deviation. Let $v = (t-\mu)/\sqrt{2}\sigma$. So, $dv = dt/\sqrt{2}\sigma$, and (1) becomes

$$\begin{aligned} F\left(x;\mu,\sigma^2\right) &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{(x-\mu)/\sqrt{2}\sigma} e^{-v^2}dv \\ &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{0} e^{-v^2}dv + \frac{1}{\sqrt{\pi}}\int_{0}^{(x-\mu)/\sqrt{2}\sigma} e^{-v^2}dv \\ &= \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right), \end{aligned} \tag{2}$$

where erf is the error function [5]. We know that the cumulative distribution function designates probability as

$$F\left(x;\mu,\sigma^2\right) = P\left(X \le x\right). \tag{3}$$

Therefore, we can write

$$\begin{aligned} P\left(\overline{x} - ns \le x \le \overline{x} + ns\right) = {} & F\left(\overline{x} + ns;\mu,\sigma^2\right) \\ & - F\left(\overline{x} - ns;\mu,\sigma^2\right), \end{aligned} \tag{4}$$

where $\overline{x}$ is the estimate of the mean given as

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

and s is the sample standard deviation given as

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}. \tag{6}$$

This allows us to write

$$\begin{aligned} P\left(\overline{x} - ns \le x \le \overline{x} + ns\right) = {} & \frac{1}{2}\operatorname{erf}\left(\frac{\overline{x} + ns - \mu}{\sigma\sqrt{2}}\right) \\ & + \frac{1}{2}\operatorname{erf}\left(\frac{ns + \mu - \overline{x}}{\sigma\sqrt{2}}\right). \end{aligned} \tag{7}$$

Let $N = 2$. This implies that

$$\begin{aligned} \overline{x} &= \frac{x_1 + x_2}{2}, \\ s &= \sqrt{\left(x_1 - \frac{\left(x_1 + x_2\right)}{2}\right)^2 + \left(x_2 - \frac{\left(x_1 + x_2\right)}{2}\right)^2} \\ &= \sqrt{\left(\frac{2x_1 - x_1 - x_2}{2}\right)^2 + \left(\frac{2x_2 - x_1 - x_2}{2}\right)^2} \\ &= \sqrt{\frac{\left(x_1 - x_2\right)^2}{4} + \frac{\left(x_2 - x_1\right)^2}{4}} \\ &= \frac{\left|x_1 - x_2\right|}{\sqrt{2}}. \end{aligned} \tag{8}$$

So we have

$$\begin{aligned} & P\left(\frac{\left(x_1 + x_2\right)}{2} - n\frac{\left|x_1 + x_2\right|}{\sqrt{2}} \le x \le \frac{\left(x_1 + x_2\right)}{2} + n\frac{\left|x_1 + x_2\right|}{\sqrt{2}}\right) \\ & \quad = \frac{1}{2}\operatorname{erf}\left(\frac{\left(\left(x_1 + x_2\right)/2\right) + n\left(\left|x_1 - x_2\right|/\sqrt{2}\right) - \mu}{\sigma\sqrt{2}}\right) \\ & \qquad + \frac{1}{2}\operatorname{erf}\left(\frac{n\left(\left|x_1 - x_2\right|/\sqrt{2}\right) + \mu - \left(\left(x_1 + x_2\right)/2\right)}{\sigma\sqrt{2}}\right). \end{aligned} \tag{9}$$

This equation can be simplified. Let $y = x_1 + x_2$ and $z = x_1 - x_2$. Then, (9) becomes

$$\begin{aligned} & P\left(\frac{y}{2} - n\frac{|z|}{\sqrt{2}} \le x \le \frac{y}{2} + n\frac{|z|}{\sqrt{2}}\right) \\ & \quad = \frac{1}{2}\operatorname{erf}\left(\frac{\left(y/2\right) + n\left(|z|/\sqrt{2}\right) - \mu}{\sigma\sqrt{2}}\right) \\ & \qquad + \frac{1}{2}\operatorname{erf}\left(\frac{n\left(|z|/\sqrt{2}\right) + \mu - \left(y/2\right)}{\sigma\sqrt{2}}\right) \\ & \quad = \frac{1}{2}\operatorname{erf}\left(\frac{y}{2\sigma\sqrt{2}} + \frac{n|z|}{2\sigma} - \frac{\mu}{\sigma\sqrt{2}}\right) \\ & \qquad + \frac{1}{2}\operatorname{erf}\left(\frac{n|z|}{2\sigma} + \frac{\mu}{\sigma\sqrt{2}} - \frac{y}{2\sigma\sqrt{2}}\right) \\ & \quad = \frac{1}{2}\operatorname{erf}\left(\frac{\left(y - 2\mu\right)}{2\sigma\sqrt{2}} + \frac{n|z|}{2\sigma}\right) \\ & \qquad + \frac{1}{2}\operatorname{erf}\left(\frac{n|z|}{2\sigma} - \frac{\left(y - 2\mu\right)}{2\sigma\sqrt{2}}\right). \end{aligned} \tag{10}$$

3. The Probability Averaged over the Estimate of the Mean

For simplicity, we have $P \equiv P((y/2) - n(|z|/\sqrt{2}) \le x \le (y/2) + n(|z|/\sqrt{2}))$. The expectation with respect to y can be written as

$$\begin{aligned} E_y\left(P\right) = {} & \frac{1}{2\sigma\sqrt{\pi}} \\ & \times \int_{-\infty}^{\infty}\left[\frac{1}{2}\operatorname{erf}\left(\frac{\left(y - 2\mu\right)}{2\sigma\sqrt{2}} + \frac{n|z|}{2\sigma}\right)\right. \\ & \qquad \left. + \frac{1}{2}\operatorname{erf}\left(\frac{n|z|}{2\sigma} - \frac{\left(y - 2\mu\right)}{2\sigma\sqrt{2}}\right)\right] \\ & \times e^{-(y-2\mu)/4\sigma^2}dy, \end{aligned} \tag{11}$$

where we have used the fact that the standard deviation of y is $\sqrt{2}\sigma$ because

$$\begin{aligned} \operatorname{var}(X_1 + X_2) &= \operatorname{var}(X_1) + \operatorname{var}(X_2), \\ \sigma_X &= \sqrt{\operatorname{var}(X)}; \end{aligned} \tag{12}$$

see [6].

Let $u = (y - 2\mu)/2\sigma$, so that $du = dy/2\sigma$. Equation (11) becomes

$$\begin{aligned} &E_y(P) \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left[\frac{1}{2} \operatorname{erf}\left(\frac{u}{\sqrt{2}} + \frac{n|z|}{2\sigma} \right) + \frac{1}{2} \operatorname{erf}\left(\frac{n|z|}{2\sigma} - \frac{u}{\sqrt{2}} \right) \right] \\ &\quad \times e^{-u^2} du \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \operatorname{erf}\left(\frac{u}{\sqrt{2}} + \frac{n|z|}{2\sigma} \right) e^{-u^2} du. \end{aligned} \tag{13}$$

At this point, consider the following integral:

$$g = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \operatorname{erf}\left(\frac{a}{\sqrt{m}} + b \right) e^{-a^2} da. \tag{14}$$

We will find three conditions on g that determine its structure. First, the following limit is clear from (14).

Condition 1. Consider

$$\lim_{m \to \infty} g = \operatorname{erf}(b). \tag{15}$$

Now, let us determine the following integral:

$$\begin{aligned} s &= \int_{-\infty}^{\infty} \operatorname{erf}\left(\frac{a}{\sqrt{m}} + b \right) da, \\ s &= \sqrt{m} \int_{-\infty}^{\infty} \operatorname{erf}\left(\frac{a}{\sqrt{m}} + b \right) \frac{da}{\sqrt{m}}, \\ s &= \sqrt{m} \int_{-\infty}^{\infty} \operatorname{erf}(v)\, dv, \end{aligned} \tag{16}$$

where $v = (a/\sqrt{m}) + b$ and $dv = da/\sqrt{m}$. Using the known integral of the error function, we find that

$$\begin{aligned} s &= \sqrt{m} \left[v \operatorname{erf}(v) + \frac{1}{\sqrt{\pi}} e^{-v^2} \right]_{-\infty}^{\infty}, \\ s &= \sqrt{m} \left[\left(\frac{a}{\sqrt{m}} + b \right) \operatorname{erf}\left(\frac{a}{\sqrt{m}} + b \right) + \frac{1}{\sqrt{\pi}} e^{-(a/\sqrt{m}+b)^2} \right]_{-\infty}^{\infty}. \end{aligned} \tag{17}$$

Condition 2. Consider

$$s = 2b\sqrt{m}. \tag{18}$$

Therefore, g must vary as b and $\sqrt{m}$.

Now, because the error function is odd, we can write

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \operatorname{erf}\left(\frac{a}{\sqrt{m}} \right) e^{-a^2} da = 0. \tag{19}$$

Therefore, (19) can be added to g to obtain

$$g = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left[\operatorname{erf}\left(\frac{a}{\sqrt{m}} + b \right) - \operatorname{erf}\left(\frac{a}{\sqrt{m}} \right) \right] e^{-a^2} da. \tag{20}$$

Since

$$\lim_{m \to 0} \operatorname{erf}\left(\frac{a}{\sqrt{m}} \right) = \lim_{m \to 0} \operatorname{erf}\left(\frac{a}{\sqrt{m}} + b \right) = \operatorname{sgn}(a), \tag{21}$$

we have the following.

Condition 3. Consider

$$\lim_{m \to 0} g = 0. \tag{22}$$

Based on Conditions 1, 2, and 3, g must have the following form:

$$g = \operatorname{erf}\left(b \frac{\sqrt{m}}{\sqrt{m + c}} \right), \tag{23}$$

where $c > 0$. Equations (14) and (23) suggest that we can write

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \operatorname{erf}(a + 1)\, e^{-a^2} da = \operatorname{erf}\left(\frac{1}{\sqrt{1 + c}} \right). \tag{24}$$

Based on (24), we can numerically determine c as

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \operatorname{erf}(a + 1)\, e^{-a^2} da \approx \frac{1}{\sqrt{\pi}} \sum_{i=-N}^{N} \operatorname{erf}(a + 1)\, e^{-a^2} \Delta a. \tag{25}$$

For $N = 200$ and $\Delta a = 0.1$, we find

$$\frac{1}{\sqrt{\pi}} \sum_{i=-N}^{N} \operatorname{erf}(a + 1)\, e^{-a^2} \Delta a = 0.6827. \tag{26}$$

Consequently,

$$\operatorname{erf}\left(\frac{1}{\sqrt{1 + c}} \right) = 0.6827 \tag{27}$$

or

$$\frac{1}{\sqrt{1 + c}} = 0.7071 \tag{28}$$

and, thus, $c = 1$, so that

$$g = \operatorname{erf}\left(b \frac{\sqrt{m}}{\sqrt{m + 1}} \right) \tag{29}$$

or

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \operatorname{erf}\left(\frac{a}{\sqrt{m}} + b \right) e^{-a^2} da = \operatorname{erf}\left(b \frac{\sqrt{m}}{\sqrt{m + 1}} \right). \tag{30}$$

Equation (30) implies that

$$\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \operatorname{erf}\left(\frac{u}{\sqrt{2}} + \frac{n|z|}{2\sigma} \right) e^{-u^2} du = \operatorname{erf}\left(\frac{n|z|\sqrt{2}}{2\sigma\sqrt{3}} \right) \tag{31}$$

or

$$E_y(P) = \operatorname{erf}\left(\frac{n|z|\sqrt{2}}{2\sigma\sqrt{3}} \right). \tag{32}$$

4. The Probability Averaged over the Sample Standard Deviation

The expectation with respect to z can be written as

$$E_z\left(E_y\left(P\right)\right) = \frac{1}{2\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} \operatorname{erf}\left(\frac{n\left|z\right|\sqrt{2}}{2\sigma\sqrt{3}}\right) e^{-z^2/4\sigma^2} dz, \quad (33)$$

where we have used the fact that the standard deviation of z is $\sqrt{2}\sigma$ based on previous information and the fact that $\operatorname{var}(aX) = a^2\operatorname{var}(X)$ [6]. Equation (33) can be expressed on a semi-infinite interval as

$$E_z\left(E_y\left(P\right)\right) = \frac{1}{\sigma\sqrt{\pi}} \int_{0}^{\infty} \operatorname{erf}\left(\frac{nz\sqrt{2}}{2\sigma\sqrt{3}}\right) e^{-z^2/4\sigma^2} dz. \quad (34)$$

Let $q = z/2\sigma$. So, $dq = dz/2\sigma$, and (34) becomes

$$E_z\left(E_y\left(P\right)\right) = \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} \operatorname{erf}\left(nq\sqrt{\frac{2}{3}}\right) e^{-q^2} dq, \quad (35)$$

or we can write

$$\left\langle P\left(\overline{x}_{N=2} - ns_{N=2} \le x \le \overline{x}_{N=2} + ns_{N=2}\right)\right\rangle = \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} \operatorname{erf}\left(nq\sqrt{\frac{2}{3}}\right) e^{-q^2} dq. \quad (36)$$

5. Computational Simulation

We can estimate $\langle P(\overline{x}_{N=2} - ns_{N=2} \le x \le \overline{x}_{N=2} + ns_{N=2})\rangle$ computationally. Simulate the normal, independent random variables X and $\{X_i\}$, for which

$$x \in X, \qquad x_i \in X_i. \quad (37)$$

Let the condition β be

$$\beta : \left(\overline{x}_{N=2} - ns_{N=2} \le x \le \overline{x}_{N=2} + ns_{N=2}\right). \quad (38)$$

If $M(\beta)$ is the number of trials in which the condition β is met and M is the total number of trials, then an estimate of $\langle P(\overline{x}_{N=2} - ns_{N=2} \le x \le \overline{x}_{N=2} + ns_{N=2})\rangle$ is given as

$$\left\langle P\left(\overline{x}_{N=2} - ns_{N=2} \le x \le \overline{x}_{N=2} + ns_{N=2}\right)\right\rangle \approx \frac{M\left(\beta\right)}{M}. \quad (39)$$

Figure 1 shows a plot of $(2/\sqrt{\pi}) \int_0^{\infty} \operatorname{erf}(nq\sqrt{2/3}) e^{-q^2} dq$ versus $M(\beta)/M$ for $M = 2000$.

6. Conclusions

We have computed the probability of a confidence interval based on minimal estimates of the mean and the standard deviation. Specifically, the estimates are based on two measurements. The effort addresses the problem of constructing a confidence interval based on minimal knowledge. The result is in the form of an integral of the error function that we have evaluated numerically and compared to a computer simulation that estimates the probability based on frequency. The comparison shows a high level of agreement that supports the validity of the derivation.

This work is part of a research effort interested in constructing confidence intervals based on various levels of knowledge. Example applications exist in geophysical exploration and cancer treatment.

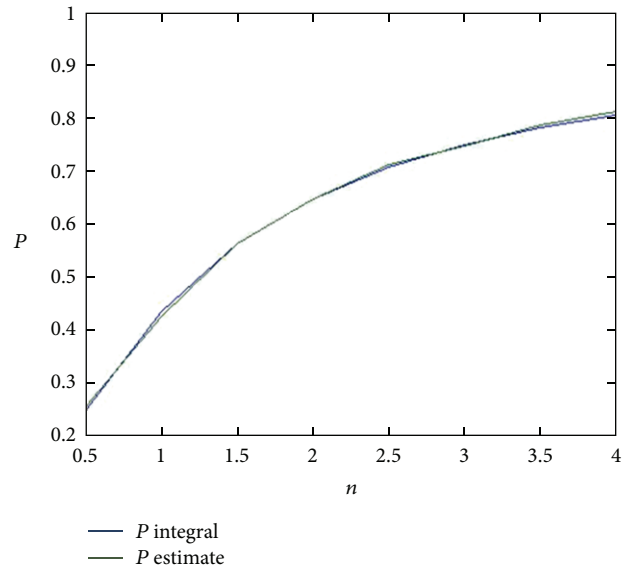

Figure 1: A plot of $(2/\sqrt{\pi}) \int_0^{\infty} \operatorname{erf}(nq\sqrt{2/3}) e^{-q^2} dq$ versus the estimate $M(\beta)/M$ for $M = 2000$.

Acknowledgment

Discussions with Gwendolyn Houston are appreciated.

References

[1] L. M. Houston, "The probability that a measurement falls within a range of n standard deviations from an estimate of the mean," *ISRN Applied Mathematics*, vol. 2012, Article ID 710806, 8 pages, 2012.

[2] L. M. Houston and G. L. Kinsland, "Minimal-effort time-lapse seismic monitoring: exploiting the relationship between acquisition and imaging in time-lapse data," *The Leading Edge*, vol. 17, no. 10, pp. 1440–1443, 1998.

[3] C. A. Perez, M. Bauer, S. Edelstein et al., "Impact of tumor control on survival in carcinoma of the lung treated with irradiation," *International Journal of Radiation Oncology, Biology and Physics*, vol. 12, no. 4, pp. 539–547, 1986.

[4] U. Balasooriva, J. Li, and C. K. Low, "On interpreting and extracting information from the cumulative distribution function curve: a new perspective with applications," *Australian Senior Mathematics Journal*, vol. 26, no. 1, pp. 19–28, 2012.

[5] L. M. Houston, G. A. Glass, and A. D. Dymnikov, "Sign-bit amplitude recovery in Gaussian noise," *Journal of Seismic Exploration*, vol. 19, no. 3, pp. 249–262, 2010.

[6] Y. A. Rozanov, *Probability Theory: A Concise Course*, Dover Publications, New York, NY, USA, 1977.