Research Article

The Probability That a Measurement Falls within a Range of n Standard Deviations from an Estimate of the Mean

Louis M. Houston

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

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

Received 5 November 2012; Accepted 25 November 2012

Academic Editors: M.-H. Hsu and X. Liu

Copyright © 2012 Louis M. Houston. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We derive a general equation for the probability that a measurement falls within a range of n standard deviations from an estimate of the mean. So, we provide a format that is compatible with a confidence interval centered about the mean that is naturally independent of the sample size. The equation is derived by interpolating theoretical results for extreme sample sizes. The intermediate value of the equation is confirmed with a computational test.

1. Introduction

A confidence interval is an interval in which a measurement or trial falls corresponding to a given probability [1, 2]. In statistics, confidence intervals centered about an estimate of the mean target the mean [1]. Because these confidence intervals rely on the standard error of the estimate, they increase as sample size decreases and they decrease as sample size increases. Consequently, the confidence interval, in statistics, is often a margin of error for the estimate of the mean [3].

We are interested in finding the probability of a confidence interval centered about an estimate of the mean that targets an arbitrary measurement and is independent of the sample size. In this way, we provide a format that is compatible with a confidence interval centered about the mean that is naturally independent of sample size, but has a width of n standard deviations. We approach this problem by considering the known case of a confidence interval centered about the mean and we derive the associated probability. We correlate this result to the case of infinite sample size. The next step is to calculate the probability

associated with the minimum sample size of one. This requires that we average the probability over different possible sample values. Finally, we propose an equation for probability that naturally interpolates our results and we show that the equation is consistent with intermediate probability values by comparing the equation to estimates of probability produced by computer simulation.

2. Expected Probability

In order to make the necessary confidence interval calculations, we will have to determine an expected or average probability. This can be examined by considering the cumulative distribution function [4, 5],

$$F(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-(t-\mu)^2/2\sigma^2} dt, \quad (2.1)$$

where x is a measurement, μ is the mean, and σ is the standard deviation.

Let $v = (t - \mu) / \sqrt{2}\sigma$. So, $dv = dt / \sqrt{2}\sigma$. Therefore, (2.1) becomes

$$\begin{aligned} F(x; \mu, \sigma^2) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{(x-\mu)/\sqrt{2}\sigma} e^{-v^2} dv \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^0 e^{-v^2} dv + \frac{1}{\sqrt{\pi}} \int_0^{(x-\mu)/\sqrt{2}\sigma} e^{-v^2} dv \\ &= \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right), \end{aligned} \quad (2.2)$$

where $\operatorname{erf}(x)$ is the so-called erf function [6, 7].

Now, we know that

$$F(x; \mu, \sigma^2) = P(X \leq x). \quad (2.3)$$

Rather than specifying a value of x , we would like to compute the probability averaged over all possible values of x . So we have

$$E(F(x; \mu, \sigma^2)) = E(P(X \leq x)). \quad (2.4)$$

This yields

$$E(F(x; \mu, \sigma^2)) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \frac{1}{2} e^{-(x-\mu)^2/2\sigma^2} dx + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \frac{1}{2} \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) e^{-(x-\mu)^2/2\sigma^2} dx \quad (2.5)$$

$$= \frac{1}{2} + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \frac{1}{2} \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) e^{-(x-\mu)^2/2\sigma^2} dx. \quad (2.6)$$

Let $z = (x - \mu) / \sqrt{2}\sigma$. So, $dz = dx / \sqrt{2}\sigma$. Therefore, (2.6) becomes

$$E\left(F(x; \mu, \sigma^2)\right) = \frac{1}{2} + \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{\infty} \operatorname{erf}(z) e^{-z^2} dz. \quad (2.7)$$

Since $\operatorname{erf}(z)$ is an odd function, we have

$$E\left(F(x; \mu, \sigma^2)\right) = \frac{1}{2}. \quad (2.8)$$

This can be easily confirmed by an alternative calculation. The probability density $f(x)$ is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}. \quad (2.9)$$

Therefore, we can write [8]

$$f = \frac{dF}{dx}. \quad (2.10)$$

The expected value [9] of the cumulative distribution function is thus

$$E(F) = \int_{-\infty}^{\infty} F f dx. \quad (2.11)$$

Note that when $x \rightarrow \infty$, $F = 1$ and when $x \rightarrow -\infty$, $F = 0$. Therefore, we can write

$$\begin{aligned} E(F) &= \int_0^1 F dF \\ &= \frac{1}{2}. \end{aligned} \quad (2.12)$$

Consequently, as per (2.4), we can write the expected probability as

$$E(P(X \leq x)) = \frac{1}{2}. \quad (2.13)$$

3. The Probability of a Confidence Interval

We want to determine the probability given as

$$P(\bar{x} - n\sigma \leq x \leq \bar{x} + n\sigma), \quad (3.1)$$

where

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (3.2)$$

is the estimate of the mean, μ . At the maximal value of N , we find that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N x_i \approx \mu. \quad (3.3)$$

For a normal distribution, the probability that a measurement falls within n standard deviations ($n\sigma$) of the mean μ (i.e., within the interval $[\mu - n\sigma, \mu + n\sigma]$) is given by

$$\begin{aligned} P(\mu - n\sigma \leq x \leq \mu + n\sigma) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu-n\sigma}^{\mu+n\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx \\ &= \frac{2}{\sigma\sqrt{2\pi}} \int_{\mu}^{\mu+n\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx. \end{aligned} \quad (3.4)$$

Now, let $u = (x - \mu)/\sqrt{2}\sigma$, so $du = dx/\sqrt{2}\sigma$. Then

$$P(\mu - n\sigma \leq x \leq \mu + n\sigma) = \frac{2}{\sigma\sqrt{2\pi}} \sqrt{2}\sigma \int_0^{n/\sqrt{2}} e^{-u^2} du \quad (3.5)$$

$$= \frac{2}{\sqrt{\pi}} \int_0^{n/\sqrt{2}} e^{-u^2} du \quad (3.6)$$

$$= \operatorname{erf}\left(\frac{n}{\sqrt{2}}\right). \quad (3.7)$$

Observe that we can express (3.7) in terms of the cumulative probability distribution function F . Thus, we can write

$$P(\mu - n\sigma \leq x \leq \mu + n\sigma) = F(\mu + n\sigma; \mu, \sigma^2) - F(\mu - n\sigma; \mu, \sigma^2). \quad (3.8)$$

The minimal value of N is one, so that $\bar{x} = x_1$.

In this case, we have

$$P(x_1 - n\sigma \leq x \leq x_1 + n\sigma) = F(x_1 + n\sigma; \mu, \sigma^2) - F(x_1 - n\sigma; \mu, \sigma^2). \quad (3.9)$$

We find that (3.9) reduces into

$$P(x_1 - n\sigma \leq x \leq x_1 + n\sigma) = \frac{1}{2} \operatorname{erf}\left(\frac{x_1 + n\sigma - \mu}{\sigma\sqrt{2}}\right) + \frac{1}{2} \operatorname{erf}\left(\frac{n\sigma + \mu - x_1}{\sigma\sqrt{2}}\right). \quad (3.10)$$

Table 1: Comparisons of numerical integration of $(1/\sqrt{\pi}) \int_{-\infty}^{\infty} \text{erf}(z + (n/\sqrt{2}))e^{-z^2} dz$ and $\text{erf}(n/2)$.

n	$(1/\sqrt{\pi}) \int_{-\infty}^{\infty} \text{erf}(z + n/\sqrt{2})e^{-z^2} dz$	$\text{erf}(n/2)$
0.5	0.2763	0.2763
1.5	0.7112	0.7112
2.5	0.9229	0.9229
3.5	0.9867	0.9867
4.5	0.9985	0.9985

The probability in (3.10) is written in terms of the value x_1 . Consequently, we can derive an expected probability by computing

$$E_{x_1}(P) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[\frac{1}{2} \text{erf}\left(\frac{x_1 + n\sigma - \mu}{\sigma\sqrt{2}}\right) + \frac{1}{2} \text{erf}\left(\frac{n\sigma + \mu - x_1}{\sigma\sqrt{2}}\right) \right] e^{-(x_1 - \mu)^2/2\sigma^2} dx_1 \quad (3.11)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[\frac{1}{2} \text{erf}\left(\frac{(x_1 - \mu)}{\sigma\sqrt{2}} + \frac{n}{\sqrt{2}}\right) + \frac{1}{2} \text{erf}\left(\frac{n}{\sqrt{2}} - \frac{(x_1 - \mu)}{\sigma\sqrt{2}}\right) \right] e^{-(x_1 - \mu)^2/2\sigma^2} dx_1. \quad (3.12)$$

Let $z = (x_1 - \mu)/\sigma\sqrt{2}$ which implies that $dz = dx_1/\sigma\sqrt{2}$. This reduces (3.12) into

$$E_{x_1}(P) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left[\frac{1}{2} \text{erf}\left(z + \frac{n}{\sqrt{2}}\right) + \frac{1}{2} \text{erf}\left(\frac{n}{\sqrt{2}} - z\right) \right] e^{-z^2} dz \quad (3.13)$$

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \text{erf}\left(z + \frac{n}{\sqrt{2}}\right) e^{-z^2} dz. \quad (3.14)$$

Equation (3.14) can be evaluated numerically to yield the following:

$$E_{x_1}(P(x_1 - n\sigma \leq x \leq x_1 + n\sigma)) = \text{erf}\left(\frac{n}{2}\right). \quad (3.15)$$

A portion of the numerical results is presented in Table 1.

Based on (3.7) and (3.15), we can propose the equation

$$P(\bar{x} - n\sigma \leq x \leq \bar{x} + n\sigma) = \text{erf}\left(\frac{n}{\sqrt{2}\sqrt{2^{1/N}}}\right), \quad (3.16)$$

in which we are by default referring to the expected probability.

Equation (3.16) clearly converges for the extreme estimates of the mean, since

$$\begin{aligned} \lim_{N \rightarrow \infty} \text{erf}\left(\frac{n}{\sqrt{2}\sqrt{2^{1/N}}}\right) &= \text{erf}\left(\frac{n}{\sqrt{2}}\right), \\ \lim_{N \rightarrow 1} \text{erf}\left(\frac{n}{\sqrt{2}\sqrt{2^{1/N}}}\right) &= \text{erf}\left(\frac{n}{2}\right). \end{aligned} \quad (3.17)$$

Figure 1: A plot of the equation $P(\bar{x} - n\sigma \leq x \leq \bar{x} + n\sigma) = \text{erf}(n/\sqrt{2}\sqrt{2}^{1/N})$ for four different values of N .

A plot of (3.16) for four values of N is shown in Figure 1. Observe that the curve for $N = 2$ is intermediate to the other curves. Consequently, we can have confidence in (3.16) if we can show that it is valid for $N = 2$.

4. Computational Test

We can estimate $P(\bar{x} - n\sigma \leq x \leq \bar{x} + n\sigma)$ computationally. Simulate the normal, independent random variables X and $\{X_i\}$, for which

$$x \in X, \quad x_i \in X_i. \quad (4.1)$$

Let the condition β be

$$\beta : (\bar{x} - n\sigma \leq x \leq \bar{x} + n\sigma). \quad (4.2)$$

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 P is given as

$$\bar{P} = \frac{M(\beta)}{M}. \quad (4.3)$$

Figure 2 shows a plot of $\bar{P}$ versus P for $N = 2$ and $M = 2000$.

Figure 2: A plot of $\bar{P}$ versus P for $N = 2$ and $M = 2000$.

5. Conclusions

We derive a general equation for the probability that a measurement falls within a range of n standard deviations from an estimate of the mean. So, we provide a format that is compatible with a confidence interval centered about the mean that is naturally independent of the sample size. It is consistent with our equation that probability reduces with sample size. However, for samples greater than ten, the value of probability begins to converge. The equation for probability is derived by considering the minimal and maximal sample sizes and producing an equation which naturally interpolates the results. Computer simulation is used to estimate probability for the sample size $N = 2$ that produces intermediate results that are in strong agreement with the general equation.

Acknowledgment

Discussions with Maxwell Lueckenhoff are appreciated.

References

- [1] J. F. Kenney and E. S. Keeping, "Confidence limits for the binomial parameter" and "Confidence interval charts" §11.4 and 11.5," in *Mathematics of Statistics*, part 1, pp. 167–169, D. Van Nostrand, Princeton, NJ, USA, 3rd edition, 1962.
- [2] A. Tennant and E. M. Badley, "A confidence interval approach to investigating non-response bias and monitoring response to postal questionnaires," *Journal of Epidemiology and Community Health*, vol. 45, no. 1, pp. 81–85, 1991.
- [3] D. G. Rees, *Essential Statistics*, Chapman & Hall/CRC, 4th edition, 2001.
- [4] D. Zwillinger and S. Kokoska, *CRC standard probability and statistics tables and formulae*, Chapman & Hall/CRC, Boca Raton, FL, 2000.

- [5] 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, 2012.
- [6] J. Spanier and K. B. Oldham, "The error function and its complement," in *An Atlas of Functions*, chapter 40, pp. 385–393, Hemisphere, Washington, DC, USA, 1987.
- [7] 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.
- [8] N. G. Ushakov, "Density of a probability distribution," in *Encyclopedia of Mathematics*, M. Hazewinkel, Ed., Springer, 2001.
- [9] L. M. Houston, G. A. Glass, and A. D. Dymnikov, "Sign data derivative recovery," *ISRN Applied Mathematics*, vol. 2012, Article ID 630702, 7 pages, 2012.