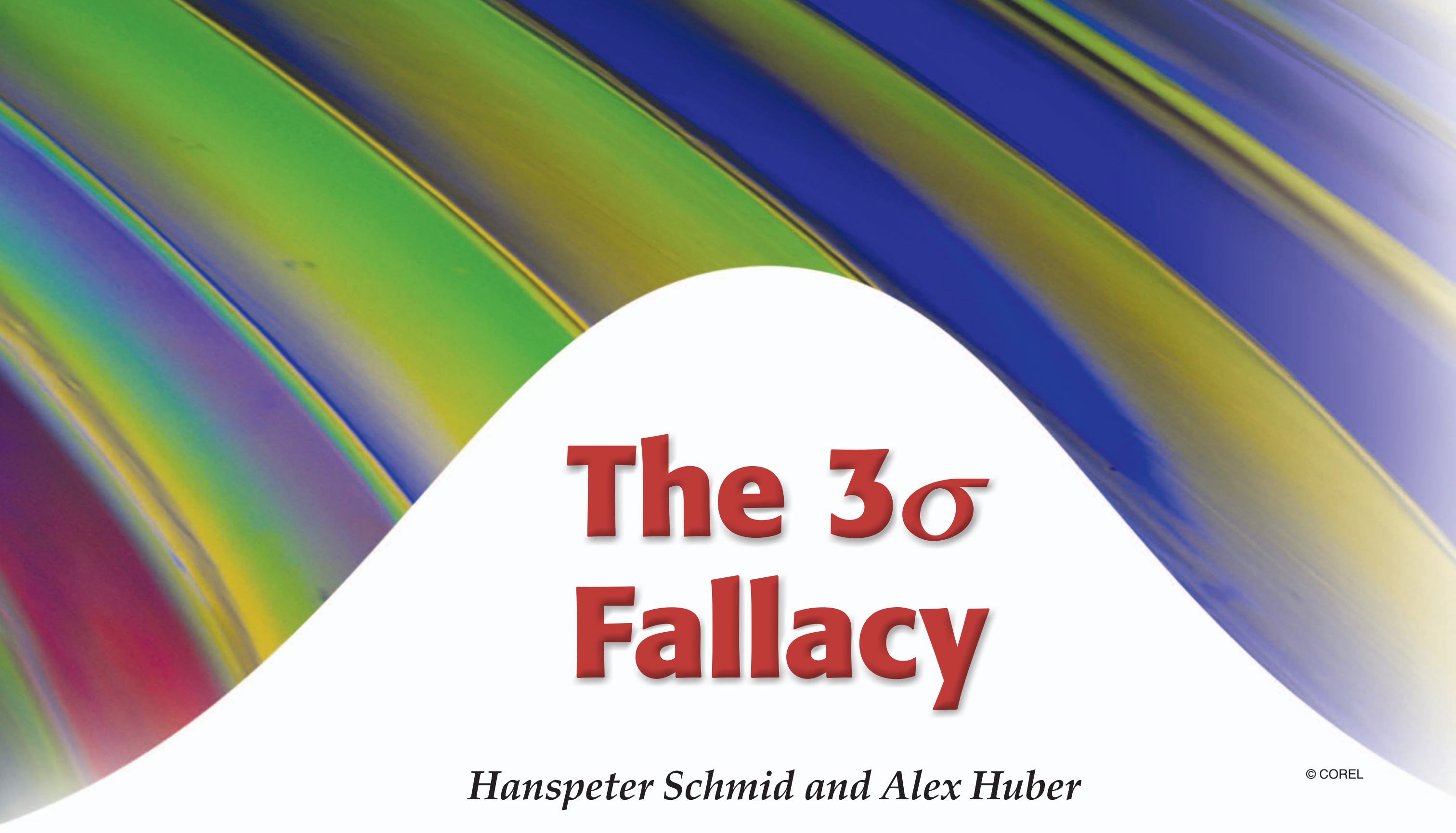

© COREL

The 3σ Fallacy

Hanspeter Schmid and Alex Huber

This article first appeared in
IEEE Solid-State Circuits Magazine,
vol. 6, no. 2, pp. 52–58, 2014.

Many solid-state circuits papers today report the mean μ and the standard deviation σ of measurement results obtained from a small number of test chips and then compare them with numbers other authors obtained. Almost none of them discuss confidence intervals, ranges of values for that standard deviation within which the true value lies with a certain probability. Many implicitly assume that the $\mu \pm 3\sigma$ range would contain all but 0.27% of chip samples to be expected in volume production. This is incorrect even if it is certain that the measured quantity is exactly normal distributed.

In this article, we shed some light on confidence and error intervals and show how the naive approach to interpreting $\mu \pm 3\sigma$ can lead to a misjudgement of error probabilities by orders of magnitude. We show that using standard deviations only works for normal distributions, and then we propose a better, distribution-independent way to report measurements in the future.

Along the way we show how many integrated circuits (ICs) you actually need to measure to obtain a range that contains, with a probability as small as 75%, with all but 0.27% of the ICs coming from the same batch as the measured ICs. This number is 1,027.

Introduction

We have all been in this situation: a small number of ICs—some ten or 20—come back from a multiproject-wafer (MPW) run, and then we are expected to make measurements, derive some statistical data from it, and draw conclusions from the derived data. The simplest way to do statistics is to assume that we are looking at a small number N of samples of a larger population. For example, we have $N = 24$ test chips from an MPW run. We assume they behave as if they were 24 ICs randomly taken from a huge batch of ICs.

Hanspeter Schmid (hanspeter.schmid@fhnw.ch) and Alex Huber (alex.huber@fhnw.ch) are with the University of Applied Sciences Norhtwestern Switzerland, Institute of Microelectronics, Steinackerstrasse 1, 5210 Windisch, Switzerland.

Digital Object Identifier 10.1109/MMM.2014.2355954
Date of publication: 12 November 2014

1527-3342/14©2014IEEE

We then measure a quantity x, getting N measurement values x_i. For the huge batch, these values have the mean μ_x and the standard deviation σ_x, which we would like to estimate. Without additional prior knowledge, the best estimates are [1]

$$\text{for } \mu_x\text{: } m_x = \frac{1}{N}(x_1 + \cdots + x_N); \tag{1}$$

$$\text{for } \sigma_x\text{: } s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_x)^2}. \tag{2}$$

The denominator $N-1$ makes the estimate unbiased for low sample sizes N. So far, so good.

In the area of sensor electronics, it has recently become fashionable to draw the $\pm 3\sigma$ limits in the published graphs, irrespective of the number of measured values available [2].

Showing the $\pm 3\sigma$ limits for small sample sizes is highly problematic. To start with, it is not $\pm 3\sigma$ that can be plotted, but $\pm 3s$. The implicit assumption (when mentioning 3σ) that the true standard deviation σ and its estimate s are the same is the first 3σ fallacy.

An even bigger problem is what it implies when authors plot $\pm 3\sigma$ limits. Most readers will assume that the $\pm 3\sigma$ limits drawn are bounds outside which only 0.27% of the huge batch's samples will be found. This is an even greater fallacy, because it is not even correct if we are certain that the values we look at are samples from a normal distribution.

In this article, we will show what is so wrong about using $\pm 3\sigma$: first, we show what $3s$ and 3σ would really mean if we were certain that we are looking at samples of a normally distributed batch. We are never certain of this, however, and if we measure trimmed ICs, we are even certain that it cannot be a normal distribution. Therefore, we discuss a method of doing statistics that also works if the underlying distribution is not known but has any halfway reasonable shape (being continuous is already more than sufficient).

The main question then remains: How many samples will be outside the limits? We propose a new standard method to define limits that can be used for benchmarking in future publications such that measurements become comparable even if sample sizes are very small, different, and coming from differently shaped distributions. You, the reader, can test this with your own data sets using our Web application [3].

A First Look at the Data and Percentiles

We will demonstrate everything using data from a real-world example: $N = 24$ measurement values taken from [2], temperature errors of an integrated, trimmed temperature sensor. The values sorted in ascending order are $x_i = -12.237, -9.712, -9.218, -7.235, -6.455, -4.869, -4.842, -4.407, -3.460, -2.527, -1.764, -1.711, -0.613, 0.252, 0.363, 1.193, 1.720, 2.185, 3.379, 5.496, 6.511, 8.722, 10.292, 19.126$ mK.

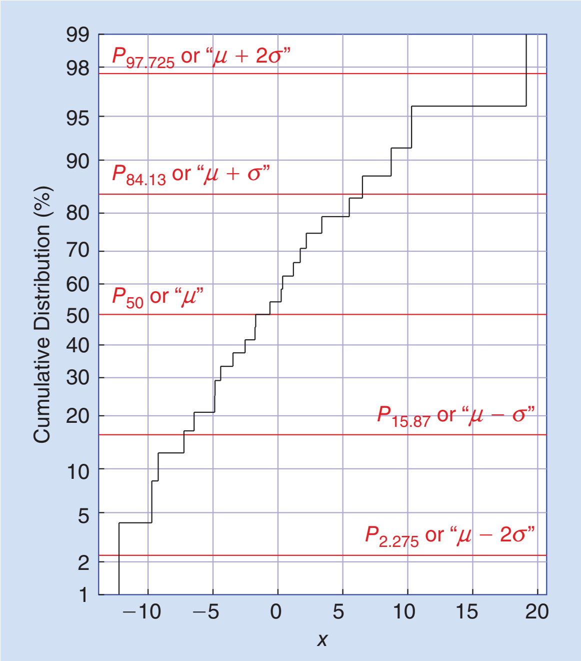

Figure 1. *The data series x_i plotted to look like a cumulative distribution. The vertical axis is scaled such that a perfect normal distribution would be represented as a straight line [1].*

To compare different sample sizes, we also look at a second set of data of sample size $N = 8$. It is just the first eight ICs that were measured: $y_i = -7.235, -1.711, 0.363, 3.379, 6.511, 8.722, 10.292, 19.126$ mK. We now have two differently sized samples of which we are certain have the same underlying statistics.

We have plotted the data series x_i as a cumulative distribution in Figure 1: at every value x_i, the curve steps by $100\%/N$. Like this, you can read off the graph, for every x, what percentage of the measured points lies below that x. We call the point below which $p\%$ of all points lie P_p, the pth percentile of the distribution:

$$\mathrm{P}\{x_i \le P_p\} = p\,.$$

The special scaling of the vertical axis would let a normal distribution appear as a straight line [1]. This approximately seems to be the case in between the lines $P_{15.87}$ and $P_{84.13}$—the 15.87- and 84.13-percentiles, which are marked "$\mu + \sigma$" and "$\mu - \sigma$" in [1] and many other textbooks because these are the percentiles where $\mu \pm \sigma$ are for a normal distribution. For the normal distribution only, samples lie below or above the $\pm n\sigma$ bounds with a probability

$$p = \frac{1}{2}\left[1 - \operatorname{erf}\left(\frac{n}{\sqrt{2}}\right)\right], \tag{3}$$

each, where erf is the error function. While giving $\pm\sigma$, $\pm 2\sigma$ and $\pm 3\sigma$ bounds is valid for normal distributions, the values that should be presented for unknown and potentially asymmetric distributions are: instead of $\mu \pm \sigma$, the two percentiles $P_{15.87}$ and $P_{84.13}$; instead of $\mu \pm 2\sigma$, $P_{2.275}$ and $P_{97.725}$; instead of $\mu \pm 3\sigma$, $P_{0.135}$ and $P_{99.865}$. Also, we

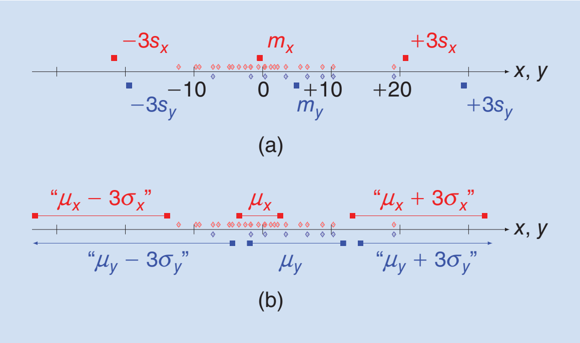

Figure 2. *(a) Standard way of drawing "$\mu \pm 3\sigma$" (actually $m \pm 3s$) in papers. (b) The intervals labeled "$\mu \pm 3\sigma$" are the intervals where the values $\mu \pm 3\sigma$ can lie for all possible μ and σ in their respective 95% confidence intervals. The intervals labeled μ_x and μ_y are simply the respective 95% confidence intervals.*

would then not give the mean value μ, but the median $M = P_{50.0}$, below which 50% of the samples are found.

This is annotated in Figure 1. As mentioned above, inside the $P_{15.87}$ and $P_{84.13}$ lines, our data set looks normally distributed, but outside these lines, we cannot tell much from such a small number of samples. At this point, the only way to get further with statistics basing on mean and standard deviation is to assume that the underlying distribution is a normal distribution.

Assuming a Normal Distribution

Let us now look at the two data sets x_i and y_i. Equations (1) and (2) give $m_x = -0.4088$, $s_x = 7.0758$, $m_y = 4.9308$, $s_y = 8.1285$. Not surprisingly, the results from the data sets x_i and y_i look quite different. A graphical representation of $m_x \pm 3s_x$ and $m_y \pm 3s_y$ is shown in Figure 2(a).

What should alarm readers and authors alike is that all $\pm 3s$ points now lie outside the range of which we have measurement data: by drawing these points, we have implied information about a possible value range of x and y for which we have no empirical evidence. Also, the points drawn are just estimates of the real values of μ and $\pm 3\sigma$, and nothing is said about confidence yet.

If and only if we are certain that the underlying distribution is a normal distribution, then we can actually calculate the ranges in which the true μ and σ lie for any confidence level. Since $\pm 3s$ implies an error probability of 0.27%, as mentioned above, it might be a good idea to ask for an interval within which the true values lie with a probability of 95%: $c = 0.95$ and $\alpha = 1 - c = 0.05$. It is well known that the confidence interval for μ is then [4]

$$\mathrm{P}\left\{m - t_{(1-\alpha/2,N-1)} \cdot \frac{s}{\sqrt{N}} \le \mu \le m + t_{(1-\alpha/2,N-1)} \cdot \frac{s}{\sqrt{N}}\right\} = c, \tag{4}$$

where $t_{(1-\alpha/2,N-1)}$ is the inverse cumulative student-t distribution, the solution of the integral equation

$$\int_{-\infty}^{t_{(1-\alpha/2,N-1)}} \frac{\Gamma\left(\frac{N}{2}\right)}{\sqrt{(N-1)\pi \cdot \Gamma\left(\frac{N-1}{2}\right)}} \times \left(1 + \frac{t^2}{N-1}\right)^{-\frac{N}{2}} dt = 1 - \alpha/2 \tag{5}$$

for the integration boundary $t_{(1-\alpha/2,N-1)}$. Similarly for σ [5],

$$\mathrm{P}\left\{\sqrt{\frac{(N-1)s^2}{\chi^2_{(\alpha/2,N-1)}}} \le \sigma \le \sqrt{\frac{(N-1)s^2}{\chi^2_{(1-\alpha/2,N-1)}}}\right\} = c, \tag{6}$$

where $\chi^2_{(\alpha/2,N-1)}$ is the inverse cumulative Chi-square distribution, the solution of the integral equation

$$\int_{-\infty}^{\chi^2_{(\alpha/2,N-1)}} \frac{x^{\frac{N-1}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{\frac{N-1}{2}} \cdot \Gamma\left(\frac{N-1}{2}\right)} dx = \frac{\alpha}{2}. \tag{7}$$

Solving these integrals numerically (in Scientific Python, e.g., implemented as `stdtrit` and `chdtri`) or using a table ([4], [5]) for the data set x, $N = 24$, gives estimated ranges for the true μ and σ:

$$\mu_x = -3.3966 \cdots 2.5791, \quad \sigma_x = 5.4994 \cdots 9.9256,$$

and for the data set y, $N = 8$,

$$\mu_y = 1.8648 \cdots 11.7264, \; \sigma_y = 5.3744 \cdots 16.5438.$$

Therefore, what we should plot are ranges rather than points, as shown in Figure 2(b). These ranges are quite large, because both μ and σ are uncertain, so, e.g., the confidence range for the location of the $\mu + 3\sigma$ point on the x axis extends from $\min(\mu) + 3\min(\sigma)$ to $\max(\mu) + 3\max(\sigma)$. The figure already makes it clear how little we actually know with $N = 24$ and eight samples, respectively.

Now imagine that someone assumes that the estimated $s_x = \sigma_x$. That person would then believe that the probability that a sample lies outside the range $m_x \pm 3s_x$ is only 0.27%. If the true μ and σ are the maxima of their respective confidence intervals, then the probability that a sample lies outside the range $m_x \pm 3s_x$ is actually 6.05%. This means that the error probability was underestimated by a factor of 22.4.

All this must be shocking enough for authors, reviewers, and readers alike who have simply plotted and requested 3σ bounds up to date. It gets even worse: Remember that even this is valid only if we have prior knowledge that the data we look at is exactly normally distributed. And this is a knowledge that we never have at all. Our example data comes from a trimmed temperature sensor, in which case we are certain that the distribution is not normal. In that case, the estimation errors we make can be arbitrarily much higher.

The sobering conclusion of this section is therefore not "apply this theory correctly instead of incorrectly."

The conclusion is "do not even use this theory." We should never just assume a normal distribution without having a valid reason to do so.

What if the Underlying Distribution Is Not Known?

So what can we do if we have no knowledge about the shape of the underlying distribution? The interesting answer is that there is a distribution-independent method to obtain confidence intervals for percentiles that is even simpler mathematically than the standard method described in the previous section.

The percentiles of a distribution have a very nice property than can be explained with a simple thought experiment: What do we know about the percentile P_p or the median $M = P_{50.0}$ if we have just one single measurement? The question sounds absurd, but being what they are, we know that having one single measurement value x_1, the percentile P_p lies below that value with a probability of $1-p$ and above it with a probability p. So it is for every measurement value, independent of all the others. Therefore, for a sorted list of samples x_i, $i = 1 \ldots N$, the probability that a percentile lies somewhere between two measured values follows the binomial distribution [6]

$$\mathrm{P}\{x_k \le P_p \le x_{k+1}\} = \binom{N}{k} p^k (1-p)^{N-k}, \quad k = 0 \cdots N, \tag{8}$$

where it is assumed that $x_0 \to -\infty$ and $x_{N+1} \to \infty$. This is valid independent of the underlying probability distribution, as long as that distribution is suffciently well behaved (its being continuous is already more than suffcient).

Figure 3 shows this both for the median and for the 15.78% percentile. Adding the probabilities on the intervals, we see that the probability that the true median is within the range of measured values is 87.5%, but the probability that $P_{15.87}$ is within is only 49.9%. This means that four measurements are only suffcient to estimate the median of the true distribution with a confidence of 87.5%.

This can now easily be generalized using (8) for any number N of measurements, any percentile p, and any integer $1 \le m < N/2$:

$$\mathrm{P}\{x_m \le M \le x_{N-m+1}\} = 1 - 2\sum_{k=0}^{m-1} \binom{N}{k} \frac{1}{2^N}, \tag{9}$$

$$\begin{aligned} \mathrm{P}\{x_m \le P_p\} &= P\{x_{N-m+1} \ge P_{1-p}\} \\ &= 1 - \sum_{k=0}^{m=1} \binom{N}{k} p^k (1-p)^{N-k}, \end{aligned} \tag{10}$$

where $p = 1/2$ is inserted into (8) to obtain (9). This lets us, as described in [7] for the median, decide which values x_i we should use as bounds for different percentiles and different confidence levels, as shown in Table 1.

For example, if you have ten samples and need a 75% confidence interval for M, Table 1 says 3, meaning that the interval $x_3 \cdots x_8$ is a 75% (or better) confidence interval. The most extreme data, $x_{1,2,9,10}$, are simply dropped. So we have a statistical method where ignoring outliers is not an ad-hoc strategy but a proven part of the procedure. The more measurements we have, the more outliers we can ignore, as Table 1 shows.

If research groups would publish their figures as we propose here, then a better figure (e.g., narrower temperature error range) obtained would be good enough an indicator to discuss whether a new circuit merits being published.

We will show examples in the following section, but let us stress, right here, a very important point: if we choose actually measured values as interval bounds, then it is impossible that any interval derived with this method ever exceeds the range of measured values, it is impossible that we imply information about ranges of x for which we have no evidence, and the interval limits will also automatically have the correct number of significant digits.

Note that this method estimates the median $M = P_{50.0}$ rather than the mean value μ. For a symmetric distribution, the median and the mean are the same, but in general, they differ. Calculating the median minimizes the mean *absolute* distance to all samples, but the mean value minimizes the mean *squared* distance, so the median may be the more informative measure because measurement outliers have much more influence on the mean than on the median.

A Suggestion for Statistical Benchmarking

Our suggestion for dealing with small data sets in future papers is to agree on a common confidence level c for comparisons and then, for comparability with the old $\pm 1,2,3\sigma$ thinking, publish ranges for the median M and for $P_{15.87}$ and $P_{84.13}$, $P_{2.275}$ and $P_{97.725}$, and $P_{0.135}$ and $P_{99.865}$.

A brief survey of recent papers published at the IEEE European Solid-State Circuits Conference (ESSCIRC)

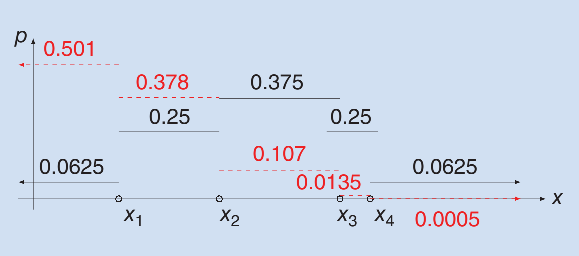

Figure 3. *The probability that the true median ($P_{50\%}$, black) and the percentile $P_{15.87\%}$ (red, dashed) of a process generating four measurements x_i, $i = 1 \ldots 4$, lie in the intervals defined by the measured data x_i.*

TABLE 1. This table shows which values to use as bounds. (Note: The values marked * give only 74.75% or 94.95% confidence.) For example, if $N = 42$ measurement results are available, and they are sorted like $x_1 \leq x_2 \leq \cdot \leq x_N$, then the 75% confidence interval for M is $x_{17} \cdots x_{26}$ $(42 - 17 + 1 = 26)$, the 95% confidence interval for M is $x_{15} \cdots x_{28}$, the 75% confidence interval for $P_{15.87}$ and $P_{84.13}$ is $x_5 \cdots x_{38}$, and the 95% confidence interval for $P_{15.87}$ and $P_{84.13}$ is $x_3 \cdots x_{40}$.

	M		$P_{15.87}$			M		$P_{15.87}$	
N	**75%**	**95%**	**75%**	**95%**	**N**	**75%**	**95%**	**75%**	**95%**
1	–	–	–	–	35	14	12	4	2
2	–	–	–	–	36	15	12	4	2
3	1	–	–	–	37	15	13	4	2
4	1	–	–	–	38	15	13	4	3
5	1	–	–	–	39	16	13	5	3
6	2	1	–	–	40	16	14	5	3
7	2	1	–	–	41	17	14	5	3
8	2	1	*1	–	42	17	15	5	3
9	3	2	1	–	43	18	15	5	3
10	3	2	1	–	44	18	16	5	3
11	4	2	1	–	45	19	16	5	3
12	4	3	1	–	46	19	16	6	3
13	4	3	1	–	47	20	17	6	4
14	5	3	1	–	48	20	17	6	4
15	5	4	1	–	49	20	18	6	4
16	6	4	1	–	50	21	18	6	4
17	6	5	2	–	51	21	19	6	4
18	7	5	2	1	52	22	19	6	4
19	7	5	2	1	53	22	19	7	4
20	7	6	2	1	54	23	20	7	4
21	8	6	2	1	55	23	20	7	*5
22	8	6	2	1	56	24	21	7	5
23	9	7	2	1	57	24	21	7	5
24	9	7	3	1	58	25	22	7	5
25	10	8	3	1	59	25	22	7	5
26	10	8	3	1	60	26	22	8	5
27	11	8	3	1	61	26	23	8	5
28	11	9	3	2	62	26	23	8	5
29	11	9	3	2	63	27	24	8	5
30	12	10	3	2	64	27	24	8	6
31	12	10	3	2	65	28	25	8	6
32	13	*11	4	2	66	28	25	8	6
33	13	11	4	2	67	29	26	9	6
34	14	11	4	2	68	29	26	9	6

(continued)

TABLE 1. *(continued)* This table shows which values to use as bounds. (Note: The values marked * give only 74.75% or 94.95% confidence.) For example, if $N = 42$ measurement results are available, and they are sorted like $x_1 \le x_2 \le \cdot \le x_N$, then the 75% confidence interval for M is $x_{17} \cdots x_{26}$ (42 − 17 + 1 = 26), the 95% confidence interval for M is $x_{15} \cdots x_{28}$, the 75% confidence interval for $P_{15.87}$ and $P_{84.13}$ is $x_5 \cdots x_{38}$, and the 95% confidence interval for $P_{15.87}$ and $P_{84.13}$ is $x_3 \cdots x_{40}$.

	M		$P_{15.87}$			M		$P_{15.87}$	
N	**75%**	**95%**	**75%**	**95%**	**N**	**75%**	**95%**	**75%**	**95%**
69	30	26	9	6	85	37	33	11	8
70	30	27	9	6	86	38	34	11	8
71	31	27	9	6	87	38	34	11	8
72	31	28	9	7	88	39	35	12	9
73	32	28	9	7	89	39	35	12	9
74	32	29	10	7	90	40	36	12	9
75	33	29	10	7	91	40	36	12	9
76	33	29	10	7	92	40	37	12	9
77	33	30	10	7	93	41	37	12	9
78	34	30	10	7	94	41	38	12	9
79	34	31	10	7	95	42	38	13	9
80	35	31	10	8	96	42	38	13	10
81	35	32	11	8	97	43	39	13	10
82	36	32	11	8	98	43	39	13	10
83	36	33	11	8	99	44	40	13	10
84	37	33	11	8	100	44	40	13	10

shows that having only $N = 8$ samples is quite common for academic and industry papers, so we set the confidence level c such that $P_{15.87}$ and $P_{84.13}$ are determined by the data extrema in the case $N = 8$. This means we calculate $c = P\{x_1 \le P_{15.87}\} = 1 - 0.251 = 0.749$. Therefore, allowing papers with $N = 8$ samples to participate in numerical benchmarks already decides that our common confidence level shall be 75%; authors with $N = 8$ measurements can then state that their measured data extrema are the limits of a 75% confidence interval on $P_{15.87}$ and $P_{84.13}$.

We can now ask the following question: What is the minimum N such that we can make statements about $P_{2.28}$ and $P_{0.135}$, which correspond to $\mu - 2\sigma$ and $\mu - 3\sigma$ in normal distributions? Solving (9) using $m = 1$ for N, which simply is

$$P\{x_1 \le P_p\} = 1 - (1-p)^N \ge 0.75 \implies N \ge \frac{\log(1-0.75)}{\log(1-p)},$$

gives minimum sample sizes of $N = 61$ and $N = 1{,}027$, respectively. This means that, for normal MPW sample sizes (the most we ever got back was 50), it is never possible to talk about $P_{2.28}$ and $P_{0.135}$ with even a confidence level as low as 75%. With respect to the infamous 6σ, observe that there the value is $N \approx 1.4 \cdot 10^9$, which more or less tells us: "forget about 6σ."

Having more measurements does, however, give an advantage. Evaluating (9) shows that, at a confidence level of 75%, $P_{15.87}$ and $P_{84.13}$ are in the range $x_1 \ldots x_N$ for $N = 8$ upwards. The confidence increases for higher N, and at $N = 17$, we get to the situation that $P_{15.87}$ and $P_{84.13}$ are in the range $x_2 \ldots x_{N-1}$ with 75% confidence, i.e., we can omit the two data extrema. From $N = 24$ upward, it is $x_3 \ldots x_{N-2}$, so we can omit the two lowest and highest values, and so on. This means that having more measurements makes it possible to ignore more outliers in the data set. The same thinking can be applied to the median and to 95% confidence, as shown in Table 1 and as performed by the companion Web application to this article [3].

This is where we can come back to the numerical examples right at the beginning: with $N = 8$ and $c = 0.75$, the $P_{15.87}$ and $P_{84.13}$ are in the range $x_1 \ldots x_8$, and the median is in the range $x_2 \ldots x_7$. For $N = 24$ and $c = 0.75$, the $P_{15.87}$ and $P_{84.13}$ are in the range $x_3 \ldots x_{22}$, and the median is in the range $x_9 \ldots x_{16}$. So an interval plot should be shown as in Figure 4(a). To compare,

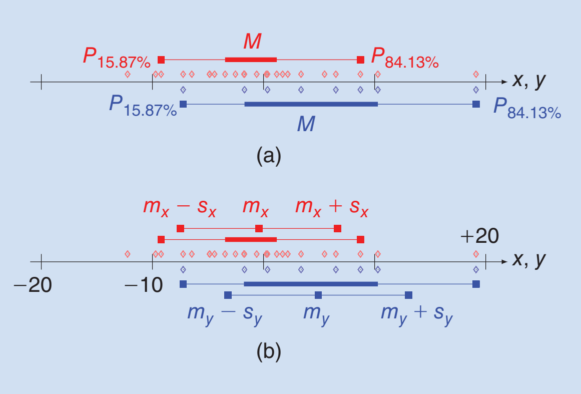

Figure 4. *(a) Data sets with median and percentiles estimated according to our method. (b) The same without M and P labels but drawn together with conventional $m \pm s$ limits.*

the same plot is replicated in Figure 4(b) together with the conventional "$\mu \pm \sigma$" limits. The range is narrower for $N = 24$ than for $N = 8$, so having more results lets us give a smaller range with the same confidence. Figure 4(b) also shows that the naively obtained bounds given by $m \pm s$ are so tight that they do not even contain the 75% confidence intervals of the respective percentiles.

In our opinion, if research groups would publish their figures as we propose here, then a better figure (e.g., narrower temperature error range) obtained would be good enough an indicator to discuss whether a new circuit merits being published. However, 75% confidence may not be enough to make decisions on future research.

To base design decisions on statistical evaluation, it would be better to use 95% confidence limits all the time and look at the range of the $P_{15.87}$ and $P_{84.13}$ of, e.g., temperature errors. If the new range is narrower, there is a good chance that the design change has brought an improvement, given that it can be expected to have a distribution of a similar shape. This means, however, that $N = 18$ devices or more need to be measured and (for $18 \leq N \leq 27$) that the data extrema are used as the 95% confidence interval for $P_{15.87}$ and $P_{84.13}$.

Finally, what if we really need information on failure probability at a level of 0.27%, corresponding to the usual $\pm 3\sigma$ thinking? Then we either have to measure 1,027 samples or more to obtain even a 75% confidence interval on that range, or we have to derive the shape of the underlying probability distribution from physical principles and then use very complicated statistics. There is, to our knowledge, no way that is more convenient but still correct.

Conclusions

The sobering conclusion of this article is that, giving the mean μ and the standard deviation σ calculated from a small number of samples, from samples that are not normal distributed, or even both, is quite meaningless. Giving any statistically obtained data without setting a confidence level is also meaningless. The only meaningful way to disclose data are conclusions of the form "We conclude that the parameter x is within the range $x_a \cdots x_b$ with probability p" or similar.

We have shown how to do this even if we do not know how the physical quantity we measure is actually distributed, which is quite useful, since that is a knowledge we very seldom have. Especially in trimmed ICs, we often have a very good model of how the majority of the values are distributed but almost no clue about the shape of the distributions for outliers, which unfortunately is precisely the region of interest when we want to estimate production yield. In a nutshell, unscientifically speaking: the normal distribution is just good enough for describing normality but not for describing extremes.

It is true that there are much more elaborated methods to do statistics: Bayesian statistics can be done if the shape of the underlying distribution is known [8]; Kernel density estimation can be used to estimate the shape of an unknown distribution from samples [9]; and confidence intervals for many very different quantities can be estimated with Bootstrapping [10]. However, all of these methods require a lot of knowledge about statistics and need to be adapted individually to each measurement situation and are, therefore, in our opinion, not useful for numerical benchmarking in scientific literature.

We are not sure whether it is a good idea at all to do statistical benchmarking on MPW-sized sample sets, but if we want to do it, then let us at least do it right. One scientifically justifiable way to go is then the method described in the section "A Suggestion for Statistical Benchmarking" of this tutorial paper.

Acknowledgments

We thank the anonymous reviewers for the valuable comments.

References

[1] H. J. C. Berendsen, *A Student's Guide to Data and Error Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[2] M. A. P. Pertijs, K. A. A. Makinwa, and J. H. Huijsing, "A CMOS smart temperature sensor with a 3σ inaccuracy of 0.1 °C from –55 °C to 125 °C," *IEEE J. Solid-State Circuits*, vol. 40, no. 12, pp. 2805–2815, Dec. 2005.

[3] H. Schmid. (2014, Apr. 19). Companion Web app to this paper. [Online]. Available: http://public.ime.fhnw.ch/threesigma/

[4] Wikipedia. (2014, Apr. 19). Student's *t*-distribution. [Online]. Available: http://en.wikipedia.org/wiki/Student_t

[5] Wikipedia. (2014, Apr. 19). Chi-squared distribution. [Online]. Available: http://en.wikipedia.org/wiki/Chi_square

[6] W. R. Thompson, "On confidence ranges for the median and other expectation distributions for populations of unknown distribution form," *Ann. Math. Statist.*, vol. 7, no. 3, pp. 122–128, 1936.

[7] J. L. van der Parren, "Tables for distribution-free confidence limits for the median," *Biometrika*, vol. 57, no. 3, pp. 613–617, 1970.

[8] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[9] Wikipedia. (2014, Apr. 19). Kernel density estimation. [Online]. Available: http://en.wikipedia.org/wiki/Kernel_density_estimation

[10] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall and CRC Press, 1998.