Probability and Statistics Notes

Note 12

Some Observations on the Usefulness of Estimates of the Mean and Standard Deviation of an Unknown Distribution

by

Paul Ryl

24 October 1984

R & D Associates
2305 Renard, SE / Suite 201
Albuquerque, New Mexico
87106

Abstract

Suppose we are interested in some random variable X such as the threshold power of a particular type of semiconductor. Suppose we are given that the population mean value of X is exactly 100 watts, and that the standard deviation is exactly 50 watts. Suppose that is all the information we are given about X . Is this the kind of information which would be of most use to us in deciding what fraction of the thresholds lay between, say, 60 and 140 watts? or lay outside that interval? or lay below that interval, i.e., below 60 watts? This note considers such questions as these. (The random variable X of interest could just as well instead be the magnitude of a current coupled into a system, or the margin of safety of a system in dB.)

Contents

Some Observations on the Usefulness of Estimates of the Mean and Standard Deviation of an Unknown Distribution.

Introduction.

Suppose we *know* the exact values of the mean and standard deviation of the distribution of a particular random variable (say of the threshold of electrical overstress permanent damage of a certain type of semiconductor, to make the example specific). Suppose, moreover, that *we also know* in closed form the pdf (probability density function) of that distribution. Can we then calculate the fraction of the population (of thresholds, in the example above) which will lie within, say, k standard deviations of that mean?

Answer: Not necessarily.

The majority of this note is devoted to providing an example which proves that this answer is correct. But before we get into the details, notice two things:

If we are given the exact values of the mean and standard deviation of a distribution, and are guaranteed that they are correct, but we are *not* given the distribution (e.g., the pdf in closed form) together with a guarantee that it is correct, then we are *even less* able to calculate such values as the fraction of the population which lies within k standard deviations of the mean.

And if we are given, not exact values of the mean and standard deviation of the distribution, but rather *estimates* of the population mean and standard deviation, such as values of the mean and standard deviation of a *sample*, still without any guarantee of what the distribution (family) is ... then we are *even less able yet* to calculate how large such fractions of the population as the above might be,

The Role of Chebyshev's Inequality.

Given that we know μ and σ, then we can always write Chebyshev's Inequality:

$$P(|X-\mu| \geq \tilde{k}) \leq \frac{\sigma^2}{\tilde{k}^2}$$

If for $\tilde{k}$ we substitute $k\sigma > 0$ then we can rewrite Chebyshev's Inequality as

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} ,$$

i.e.,

$$P(\mu-k\sigma < X < \mu+k\sigma) \geq 1 - \frac{1}{k^2} \qquad (1)$$

where X denotes the random variable of interest. Thus the fraction of the population of semiconductor thresholds which lies in the interval

$$(\mu - k\sigma , \mu + k\sigma) \qquad (2)$$

can be no less than $\frac{k^2-1}{k^2}$.

Is this helpful? Not necessarily. For two reasons. First, this (lower) bound on the probability is usually regarded as rather loose compared to bounds which can sometimes be discovered when the distribution is known. Consideration of the case in which the distribution is known to be Bernoulli (with $p = \frac{1}{2}$) and $k = 1.01$ will illustrate this point. In that case inequality (1) tells us that the fraction of the population lying in interval (2) is no less than $1-[1/(1.01^2)] \doteq 2\%$ (we might have guessed that already, without inequality (1)), whereas the truth of the matter is that the fraction of the population which lies in interval (2) in that case is actually *100* percent. The second reason is, whenever $k \leq 1$ we observe that inequality (1) becomes vacuous: it then merely states that the probability of an event is at least some non-positive number; but we already know that from Kolmogorov, who tells us that *all* probabilities are non-negative.

For these reasons Chebyshev's Inequality, inequality (1), although always true if μ and $\sigma > 0$ exist and are known (too good to be true already) and k is positive, may nonetheless not be too useful.

It should be pointed out that the difficulty in making use of the mean and standard deviation rests principally in the fact that these numbers, by themselves, carry *very little information about the population distribution;* that their values are never really known perfectly in real world situations, but can only be estimated approximately from incomplete, noisy sample data, although this does exacerbate the situation it is still secondary. For if

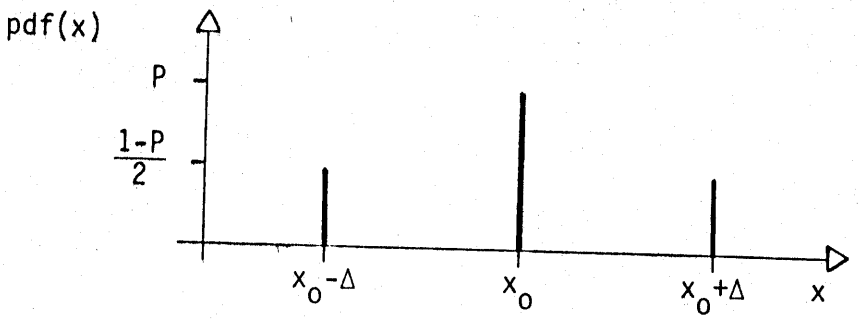
restriction to sample statistics $\bar{x}$ and s were the primary difficulty then we could solve our problem by making estimates using the Saw-Yang-Mo Inequality (cf. "The American Statistician", May 1984, p. 130), which is a version of inequality (1) but with $\bar{x}$ replacing μ , s replacing σ , and a more complicated right hand side. But no, the difficulty persists even if the values of μ and σ are granted known perfectly, so that inequality (1) applies directly. The rest of this note is devoted to providing an example to illustrate concretely that knowledge, even perfect knowledge, of the mean μ and the standard deviation σ of a distribution can be completely inadequate for making some kinds of estimates of where the population values lie ... in some cases even when the distribution (family) is known.

Content of the center standard deviation of the symmetric trinomial distribution.

The symmetric trinomial distribution has pdf

$$f(x) = \begin{cases} P & \text{if } x = x_0 \\ \frac{1-P}{2} & \text{if } x = x_0-\Delta \text{ or if } x = x_0+\Delta \\ 0 & \text{otherwise} \end{cases}$$

where $0 < \Delta$ and $P \in (0,1)$. A graph of this pdf is:



How might this distribution arise? It could, for example, be the threshold distribution for a type of semiconductor manufactured by three vendors, each using a different but very tightly controlled horizontal-vertical layout design (i.e., horizontal geometry vs vertical diffusion profile). (Observe that throughout this note we simplify the examples by assuming that the random variable of interest, say threshold, is univariate.)

The mean of this distribution is

$$\mu = \sum_{i=-1}^{1} x_i \, f(x_i) =$$

$$= (x_o - \Delta)\,\frac{1-P}{2} + x_o P + (x_o + \Delta)\,\frac{1-P}{2} =$$

$$= x_o \left(\frac{1-P}{2} + P + \frac{1-P}{2} \right) + \Delta \left(-\frac{1-P}{2} + \frac{1-P}{2} \right) = 0 =$$

$$= x_o \left(\frac{1}{2} - \underbrace{\frac{P}{2} + P - \frac{P}{2}}_{0} + \frac{1}{2} \right) =$$

$$= x_o \left(\frac{1}{2} + \frac{1}{2} \right) = x_o$$

In what follows we may assume wolog (without loss of generality) that

$$\mu = x_o = 0 \qquad (3) .$$

The variance of this distribution is

$$\sigma^2 = \sum_{i=-1}^{1} (x_i - \mu)^2 \, f(x_i)$$

$$\overset{(3)}{=\!=} \sum_{i=-1}^{1} x_i^2 \, f(x_i) =$$

$$\overset{(3)}{=\!=} (-\Delta)^2 \frac{1-P}{2} + (0 \cdot P) + \Delta^2 \frac{1-P}{2} =$$

$$= \Delta^2 \frac{1-P}{2} + \Delta^2 \frac{1-P}{2} =$$

$$= 2\,\Delta^2 \frac{1-P}{2} =$$

$$= \Delta^2 (1-P)$$

Therefore the standard deviation of this distribution is

$$\sigma = \Delta \sqrt{1-P} \qquad (4) .$$

Therefore

$$P \in (0,1) \iff 1-P \in (0,1) \iff$$

$$\iff \sqrt{1-P} \in (0,1) \overset{0<\Delta}{\iff}$$

$$\overset{0<\Delta}{\iff} \Delta\sqrt{1-P} \in (0,\Delta) \overset{(4)}{\iff}$$

$$\overset{(4)}{\iff} \sigma \in (0,\Delta) \iff$$

$$\iff \frac{\sigma}{2} \in (0, \frac{\Delta}{2}) \iff$$

$$\iff 0 < \frac{\sigma}{2} < \frac{\Delta}{2} \qquad (5) .$$

So now we can consider the center standard deviation I , i.e., the interval of width exactly one full standard deviation which is had by going out plus or minus one half a standard deviation from the mean, i.e., the interval

$$I = [\, \mu - \tfrac{1}{2}\sigma \,,\, \mu + \tfrac{1}{2}\sigma \,] \qquad (6) .$$

From inequality (5), above, we know that

$$I = [\, \mu - \frac{\sigma}{2} \,,\, \mu + \frac{\sigma}{2} \,] \subset [\, \mu - \frac{\Delta}{2} \,,\, \mu + \frac{\Delta}{2} \,] \overset{0<\Delta}{\subset}$$

$$\overset{0<\Delta}{\subset} [\mu-\Delta,\mu+\Delta] .$$

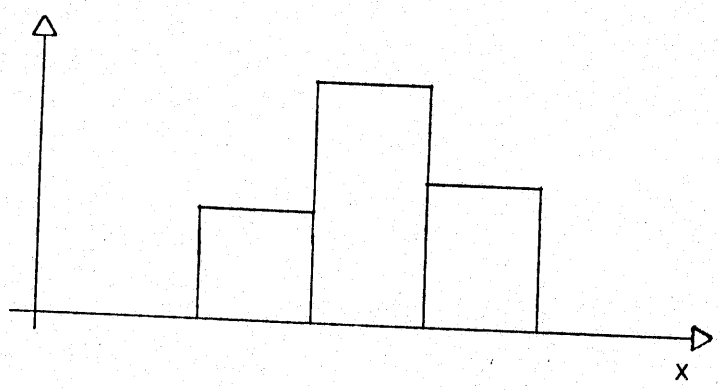
Hence we conclude that the center standard deviation I of the symmetric trinomial distribution contains the center lobe in the figure near the bottom of p. 5, above, but neither of the two side lobes.

As a corollary of this conclusion we know that the center standard deviation I, given by equation (6), contains exactly a fraction P of the population.

Therefore the interval (2), *the value of* $k \in [0,1)$ *being completely fixed,* and the distribution (family) being known perfectly, can contain any fraction $P \in (0,1)$ of a population.

That is, this example shows that even if we *know* the population's distribution (family) (e.g., we know that it is symmetric trinomial, but do not know P or Δ) , then it is still possible to have no idea at all what fraction of the population is contained within some fixed number $k \in [0,1)$ of standard deviations of the population mean. (Throughout the development of the example we have considered k to be fixed at $\frac{1}{2}$, to illustrate the point.) Finally, this is true *even if it be granted that in addition to the distribution family we also know the mean* μ *and standard deviation* σ *of the distribution perfectly*.

In closing, it can be illuminating to reflect upon what the histogram (i.e., empirical pdf) of data taken from a population with a distribution such as the trinomial can be expected to look like:



Possibly quite normalish, no? (In fact, observe that with three "bins" data from a symmetric trinomial distribution would almost certainly "pass" many standard quantitative GOF (goodness of fit) tests for normality.)

Conclusion.

It is possible that, even if μ and σ are known perfectly, and we are given a fixed value of $k \in [0,1)$, we may still be completely unable to say what fraction of the population lies in (or outside, or on either side of) the interval $[\ \mu - k\sigma\ ,\ \mu + k\sigma\ ]$ ... in some cases *even if the distribution (family) is known perfectly also* ... and this can be so even if the sample histogram is completely innocuous, the sample passes a normal GOF test, etc.

Recommendations.

Given sample data on a random variable (such as semiconductor thresholds), we know how to use order statistics from that sample to compute nonparametric confidence intervals for quantiles, either of the Bayes kind (cf., for example, earlier PSNs, Probability and Statistics Notes, of this series, e.g., PSN 2) or of the Neyman-Pearson kind (cf., for example, "Introduction to Mathematical Statistics", by Hogg and Craig, 1978, pp. 305, 306). Therefore if data has been taken directly on the random variable of interest (as opposed to being taken on "components" of that variable), and if what is needed is information about the quantiles of the distribution, and if it has been decided that only a few summary statistics from the sample data are all that will be reported rather than publishing the entire sample data set, then under some circumstances the sample size and some judiciously chosen order statistics may be considerably more useful to the reader than the sample mean and variance or standard deviation.

For example, if in the example given in the Abstract we were able to determine that 60 watts was the .2 quantile and 140 watts was the .75 quantile, then we would be able to answer several of the questions posed in the Abstract. The fraction of the thresholds which lie betweeen 60 and 140 watts would then be $.75 - .2 = 55\%$, the fraction outside that interval would be $1 - .55 = 45\%$, and the fraction below the interval would be $.2 = 20\%$.

Therefore it may be recommended that strong consideration be given to reporting the sample size n and the order statistics often used in confidence calculations, e.g., the extreme value statistics $x_{(1)}$ and $x_{(n)}$; then, if publication room allows, also the median (which we might loosely call $x_{(\frac{n}{2})}$); then, if there is room for a couple more numbers, the first and third sample quartiles (which we might loosely call $x_{(\frac{n}{4})}$ and $x_{(\frac{3n}{4})}$). Only after the reader has been given these values would the sample mean and variance or standard deviation be given when the distribution is unknown.

Appendix

Some remarks about why the example in the note works.

The trinomial distribution by itself is a five parameter distribution. If we know of a univariate random variable X only that it is trinomial, then we must learn the values of five numbers before we will be able to specify completely the probability behavior of X . For example, we could determine the sizes of the three values of x_i where the lobes are located in the figure near the bottom of p. 5, above, plus the heights of any two of those lobes. (In general an N-nomial random variable has a 2N-1 parameter distribution.)

When we gave that we were going to consider a *symmetric* trinomial distribution, we were really granting two different kinds of symmetry at the same time. We were giving that the *distances* from the right lobe to the center lobe and from the center lobe to the left lobe were the same:

$$x_1 - x_0 = x_0 - x_{-1} \quad ;$$

and we were also giving that the *heights* of the two side lobes were the same:

$$f(x_{-1}) = f(x_1) \quad .$$

By giving both of these two facts about the distribution we reduced the number of parameters from five to three. These remaining three parameters could be chosen to be x_0 , Δ , and P .

Next we granted wolog that $x_0 = 0$, equation (3). This reduced the number of parameters which we still had to be concerned about from three to two. (Alternatively, in the example in the Abstract we gave that $x_0 = 100$ watts.)

Then we offered one more value, viz., the exact value of σ (or of σ^2). For example, in the Abstract example we allowed that $\sigma = 50$ watts .

However, this was only *one* value, and we still had *two* parameters left to determine. So, for example, we knew the LHS (left hand side) of equation (4); but that was insufficient to determine the values of *both* of the two parameters Δ and P in the RHS (right hand side) of that equation. So there

was still one parameter of the original trinomial distribution floating free. That one remaining loose parameter was sufficient to derail any efforts to determine the fraction of the population which lay in the interval (2).

The foregoing remark also shows why knowing only the mean and standard deviation of a population (let alone of only a sample from the population) are by themselves so inadequate for discovering population quantiles. These are just two numbers. But the number of parameters of a real world distribution, such as of the electrical power overstress permanent damage threshold of a type of semiconductor, can be vastly larger. Not only the "locations" x_i of the thresholds of each of the vendors, to determine where the lobes are, plus the fractions $f(x_i)$ of that transistor type which come from each of the vendors, to determine the heights of the various lobes, but also numerous influences within each vendor which affect the shape (e.g., the "fatness") of each individual lobe; all these can be regarded as parameters which have a bearing on the probability that a single randomly selected semiconductor of that type will have its threshold in some pre-specified interval (a,b) . Thus the number of parameters for an actual real world physical variable distribution can be immense. Trying to describe such a situation by providing only two numbers, say μ and σ (let alone just $\bar{x}$ and s) can therefore introduce severe under-modeling.

Another aspect of this topic to notice is that when we assume a distribution (family), we may be assuming *much* more information about the distribution of the random variable than if we just assumed the value of a single parameter of the unknown actual distribution. For example, if we just outright *assumed* that a particular quantile, say the .2 quantile, was 60 watts, then we might not be making nearly as strong an assumption (i.e., might not be wishing away nearly as many parameters) as if we assumed that the threshold distribution was lognormal.

Finally, we can observe that the example in the note shows that we need under-model by *only one* parameter to make completely invalid any estimates which we might subsequently make of the fraction of the population which lies in a pre-specified interval [a,b] such as interval (6).