Understanding the 68/95/99.7 rule

Peter Burton

May 8, 2018

In Section 1 we present a procedure for making predictions about the long-term behavior of random processes. This procedure can be seen as an exposition of the so-called 68/95/99.7 rule. In Section 2 we present an example of a practical implementation of this procedure.

1 A theoretical procedure for making predictions

Let X be a real-valued random variable. One can think of X as a machine which outputs a real number according to some random procedure. If x is a real number, we use the notation $P(X = x)$ to denote the probability that X takes on the value x. Similarly, if a and b are real numbers with $a < b$, we use the notation $P(a \leq X \leq b)$ to denote the probability that the value of X lies between a and b. In mathematics, probabilities are expressed as numbers between 0 and 1. Probabilities can be multiplied by 100 to obtain percentages.

We will consider two kind of random variables. The first kind are called discrete. These are specified by a probability mass function $p(x) = P(X = x)$ for real numbers x. In the discrete case we have $P(X = x) = 0$ for all but finitely many x. However,

$$\sum_{x:p(x)>0} p(x) = 1, \tag{1}$$

so that $p(x) > 0$ for at least one x. (1) can be interpreted as saying it is certain that at least one outcome occurs. Usually, but not always, in the discrete case every x such that $P(X = x) > 0$ is an integer. A basic example of a discrete random variable is rolling a fair die. In this case we have

$$p(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5 \text{ or } 6 \\ 0 & \text{if } x \text{ is not among the above six values} \end{cases}$$

The second kind of random variables are called continuous. A continuous random variable is determined by a probability density function f from the real numbers to the nonnegative real numbers. The function f determines the distribution of X by the formula

$$\int_a^b f(x)\,\mathrm{d}x = P(a \leq X \leq b)$$

where a and b are real numbers with $a < b$. When dealing with a continuous random variable X, it does not really make sense to ask for the probability that value of X is equal to a single specific number. Rather, one

should ask for the probability that the value of X lies within some interval of positive length. Analoglously to (1) we have

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

A basic example of a continuous random variable is as follows. Consider breaking a stick of length 1 inch into two possibly unequal pieces. Let X be the distance from the left end to the breaking point. Assuming that the break is equally likely to occur at any point in the stick, the probability density function of X would be given by

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{if } x \leq 0 \text{ or } x \geq 1 \end{cases}.$$

Thus the probability that the break occurs in between $1/4$ and $1/3$ is

$$\int_{\frac{1}{4}}^{\frac{1}{3}} f(x)\,\mathrm{d}x = \int_{\frac{1}{4}}^{\frac{1}{3}} 1\,\mathrm{d}x = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

In each case, the function described above is the fundamental data used to define the distribution of X.

Suppose a random process modeled by X is performed repeatedly, in such a way that distinct trials do not affect one another. This last hypothesis is called independence. Our goal is to determine the long term behavior of this process. If the process is performed n times, this means we want to understand the sum $X_1 + \cdots + X_n$ where the X_j are independent random variables, each of which has the same distribution as X. Write S_n for the random number $X_1 + \cdots + X_n$. We will use statistics calculated from the fundamental data of X to understand S_n. An essential principle is that our predictions improve as n gets larger. **The most important caveat to what we say here is that everything rests on the assumption of independence.**

1.1 Step 1

The expectation $E(X)$ of X is given by

$$E(X) = \sum_{x:p(x)>0} xp(x) \tag{2}$$

when X is discrete and by

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x \tag{3}$$

when X is continuous. Thinking of X as being fixed, we write $\mu = E(X)$. Sometimes μ is called the mean, the average value, or the first moment of X. The expectation μ is the most fundamental statistic associated with X, as it allows for the most basic prediction for the long term value of a random process described by X. We now state this prediction, which is a version of the law of large numbers.

Basic prediction. We have $S_n \approx n\mu$.

The notation $\approx$ should be interpreted as saying that the quantities on either side are approximately equal.

1.2 Step 2

We now concentrate on approximating the error in our basic prediction. For this we need to introduce the notion of a generalized expectation. For any function g from the real numbers to the real numbers we define

$$E(g(X)) = \sum_{x:p(x)>0} g(x)p(x) \tag{4}$$

when X is discrete and

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x. \tag{5}$$

We record an important application of (4) for discrete X:

$$E(X^2) = \sum_{x:p(x)>0} x^2 p(x) \tag{6}$$

We also record the corresponding application of (5) for continuous X:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x. \tag{7}$$

Using (2) and (4) in the discrete case and (3) and (7) in the continuous case, we can define the second fundamental statistic. This the variance $\mathrm{Var}(X)$ of X, which is given by

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = E((X-\mu)^2). \tag{8}$$

It is traditional to write $\sigma = \sqrt{\mathrm{Var}(X)}$ and call σ the standard deviation. Variance and standard deviation are essentially the same concept. σ is important because the error in the basic prediction is naturally quantified in multiples of $\sigma\sqrt{n}$. Specifically, we have the following approximation for the probability that the basic prediction is correct up to an error of size at most $k\sigma\sqrt{n}$. This approximation is a version of the central limit theorem.

Basic error approximation. We have

$$P(|S_n - n\mu| \leq k\sigma\sqrt{n}) \approx 100(1 - 2\Phi(-k))\%. \tag{9}$$

Here Φ is the cumulative distribution function of the standard normal given by

$$\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2}\,\mathrm{d}x.$$

The factor of 100 is present to convert the probability into a percentage. As k gets larger, the function $\Phi(-k)$ decreases to 0 very quickly, so the quantity in parenthesis on the right becomes closer to 1. Therefore the probability on the right becomes higher. Alternate expressions for the value $1 - 2\Phi(-k)$ are $\Phi(k) - \Phi(-k)$ and $2\Phi(k) - 1$. This number represents the area under the portion of the classical bell curve lying over the interval from $-k$ to k. The function Φ can be easily computed to great precision with computer mathematics software. Usually we want to take $k = 1, 2$ or 3. This leads to the well-known 68/95/99.7 rule since we have

$$100(1 - 2\Phi(-1))\% \approx 68\%$$
$$100(1 - 2\Phi(-2))\% \approx 95\%$$
$$100(1 - 2\Phi(-3))\% \approx 99.7\%$$

1.3 Step 3

In this step we calculate a third statistic which allow us to quantitatively control the uncertainty implied in the fact that we have used $\approx$ rather than $=$ or $\leq$ in (9). This step is mostly important when making predictions with very close to 100% certainty, for example with the $k = 3$ case of the 68/95/99.7 rule. For the $k = 1$ and $k = 2$ cases of the rule, the basic error estimate from Step 2 is usually good enough.

We begin by recording an application of (4) for discrete random variables.

$$E(|X - \mu|^3) = \sum_{x:p(x)>0} |x - \mu|^3 p(x) \tag{10}$$

We also record the corresponding application of (5) for continuous random variables.

$$E(|X - \mu|^3) = \int_{-\infty}^{\infty} |x - \mu|^3 f(x)\, \mathrm{d}x. \tag{11}$$

The presence of absolute value signs sometimes makes the integral in (11) difficult to calculate by hand, but it can usually be evaluated to great precision with computer mathematics software. We adopt the notation

$$\rho = \frac{1}{\sigma^3} E(|X - \mu|^3). \tag{12}$$

The statistic ρ is called the third absolute standardized moment. ρ is related to another statistic called the skewness, which is defined to be $\frac{1}{\sigma^3} E((X-\mu)^3)$. Sometimes the notation ρ is also used for the skewness. For our purposes we will need the absolute value sign in the definition. We have the following quantitative error estimate to confirm our approximations from Step 2. This estimate is a version of the Berry-Esseen theorem.

Quantitative error estimate. We have

$$P(|S_n - n\mu| \leq k\sigma\sqrt{n}) \geq 100 \left(1 - 2\Phi(-k) - \frac{\rho}{\sqrt{n}}\right) \%.$$

The idea in this estimate is that the two terms we subtract on the right are small positive numbers, and so the quantity in parenthesis is close to 1. Consequently, the probability on the right is close to 100%. Interpreting this in terms of the 68/95/99.7 rule, we can get a more certain version of the rule by subtracting any percentage greater than or equal to $\frac{100\rho}{\sqrt{n}}\%$ from each of the three entries in the rule. For example, suppose that we calculate $\frac{100\rho}{\sqrt{n}} \leq 2$. Then we can get a more certain version of the 68/95/99.7 rule by subtracting 2% from each entry in the rule. Thus modified, we have a mathematically rigorous 66/93/97.7 rule, with no approximation error other than what comes from rounding decimals. This 66/93/97.7 rule would need to be changed in a different statistical situation. Note that the qualitative difference between a 66% probability and a 68% probability is negligible, and the qualitative difference between a 93% probability and a 95% probability is also small. However, a 97.7% probability is qualitatively quite different from a 99.7% probability. Thus we see how Step 3 is most important for the $k = 3$ case of the analysis from Step 2.

2 Example of a practical implementation of the procedure

Suppose we are consulting for a company which undertakes a transaction 100 times each day. There is some random uncertainty involved, so that the transaction can result in one of three possible outcomes. With

probability 1/2 they gain \$2, with probability 1/6 they gain \$4, and with probability 1/3 they lose \$3. We can model the outcome of a single transaction with a discrete random variable X whose probability mass function is given by

$$p(x) = P(X = x) = \begin{cases} 1/2 & \text{if } x = 2 \\ 1/6 & \text{if } x = 4 \\ 1/3 & \text{if } x = -3 \\ 0 & \text{if } x \text{ is not equal to } 2, 4 \text{ or } -3 \end{cases} \tag{13}$$

If the outcome of the transaction is genuinely random, it is reasonable to assume that the probabilities for distinct instances of the transaction are independent.

A useful thing to know when incorporating this into a business plan is how much money they can expect to gain or lose from these transactions over one year. During one year they make 36500 transactions. Therefore we need to understand $X_1 + \cdots + X_{36500}$ where the X_j are i.i.d. copies of X. Since we have fixed $n = 36500$, we will simply write S for the random number $X_1 + \cdots + X_{36500}$. We begin by calculating the most fundamental statistic, which is the expectation $E(X) = \mu$. Using (2) and (13) we have

$$\mu = 2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{6} + (-3) \cdot \frac{1}{3} = \frac{2}{3}. \tag{14}$$

Now we can make our basic prediction.

Basic prediction. We have

$$S \approx n\mu = 36500 \cdot \frac{2}{3} \approx 24333.33.$$

In other words, they will make approximately \$24333.33 over the year.

Moving on to Step 2, we would like to conduct the basic error approximation. In order to do this we need to know σ. We apply (6) and (13) to see that

$$E(X^2) = 2^2 \cdot \frac{1}{2} + 4^2 \cdot \frac{1}{6} + (-3)^2 \cdot \frac{1}{3} = \frac{23}{3}.$$

Therefore

$$\sigma = \sqrt{E(X)^2 - \mu^2} = \sqrt{\frac{23}{3} - \left(\frac{2}{3}\right)^2} = \sqrt{\frac{65}{9}}, \tag{15}$$

and so we see that the fundamental unit for estimating the error is

$$\sigma\sqrt{n} = \sqrt{36500 \cdot \frac{65}{9}} \approx 513.43.$$

We calculate $2\sigma\sqrt{n} \approx 1026.86$ and $3\sigma\sqrt{n} \approx 1540.29$. Applying the basic error approximation we obtain the following probabilities.

$$P(|S - 24333.33| \leq 513.43) \approx 100(1 - 2\Phi(-1))\% \approx 68\% \tag{16}$$

$$P(|S - 24333.33| \leq 1026.86) \approx 100(1 - 2\Phi(-2))\% \approx 95\% \tag{17}$$

$$P(|S - 24333.33| \leq 1540.29) \approx 100(1 - 2\Phi(-3))\% \approx 99.7\%. \tag{18}$$

We can consider the quantity $\frac{100k\sigma\sqrt{n}}{n\mu}$ to turn the inner error estimates into percentages relative to the basic prediction. Doing this we see that, for example,

$$\frac{100 \cdot 513.43}{24333.33} \approx 2.10$$

so that the $k = 1$ error is about 2.10% of the basic prediction. Similarly we can see that an error of 1026.86 is about 4.21% of the basic prediction and an error of 1540.29 is about 6.33% of the basic prediction. Therefore we could summarize the approximation (17) as saying that with 95% probability, the basic prediction is off by no more than 4.21%, or maybe we could just say 5% to get a rounder number.

Suppose, however that we wanted to understand whether it's reasonable to go to the $k = 3$ case and make an almost certain claim about the error in the basic prediction. In this case it would be appropriate to go to Step 3 and make use of the quantitative error estimate. In order to do this, we need to calculate ρ. Using (10) and (13) we see that

$$E(|X - \mu|^3) = \left|2 - \frac{2}{3}\right|^3 \cdot \frac{1}{2} + \left|4 - \frac{2}{3}\right|^3 \cdot \frac{1}{6} + \left|-3 - \frac{2}{3}\right|^3 \cdot \frac{1}{3} = \frac{1927}{81}.$$

Using (12) and (15) we can calculate

$$\rho = \frac{1927}{81} \cdot \left(\frac{65}{9}\right)^{-3/2}$$

and so

$$\frac{100\rho}{\sqrt{n}} = 100 \cdot \frac{1927}{81} \cdot \left(\frac{65}{9}\right)^{-3/2} \cdot \frac{1}{\sqrt{36500}} \approx 0.641\%$$

By the quantitative error estimate, we obtain a rigorous modification of the 68/95/99.7 rule by subtracting any percentage greater than 0.641%. For example we could choose to subtract 1% and obtain quantitative estimates

$$P(|S - 24333.33| \leq 512.44) \geq 67\% \tag{19}$$

$$P(|S - 24333.33| \leq 1026.44) \geq 94\% \tag{20}$$

$$P(|S - 24333.33| \leq 1540.29) \geq 98.7\%. \tag{21}$$

As usual, the qualitative difference between the approximations (16)-(18) and the precise statements (19) - (21) is only significant for the $k = 3$ regime of a very high probability error estimate.

Department of Mathematics
The University of Texas, Austin
pjburton@math.utexas.edu