Brit. J. Phil. Sci. **65** (2014), 621–649

Why are Normal Distributions Normal?

Aidan Lyon

ABSTRACT

It is usually supposed that the central limit theorem explains why various quantities we find in nature are approximately normally distributed—people's heights, examination grades, snowflake sizes, and so on. This sort of explanation is found in many textbooks across the sciences, particularly in biology, economics, and sociology. Contrary to this received wisdom, I argue that in many cases we are not justified in claiming that the central limit theorem explains why a particular quantity is normally distributed, and that in some cases, we are actually wrong.

doi:10.1093/bjps/axs046

Advance Access published on September 10, 2013

> Everyone believes in it: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact. (Henri Poincaré)[1]

1 Introduction

We seem to be surrounded by bell curves—curves more formally known as normal distributions or Gaussian distributions. All manner of things appear to be distributed normally: people's heights, IQ scores, examination grades, sizes of snowflakes, errors in measurements, lifetimes of lightbulbs, weights of loaves of bread, milk production of cows, chest sizes of Scottish soldiers, and so on. That we are surrounded by normal distributions seems to be a contingent fact. It seems that things could have been otherwise; uniform distributions, for example, could have been the norm—or, there could have even been no norm at all. Why, then, are normal distributions normal?

A very common answer to this question, found throughout the sciences, involves the central limit theorem (CLT) from probability theory. Roughly speaking, this theorem says that if a random variable, X, is the sum of a large number of small and independent random variables, then almost no matter how the small variables are distributed, X will be approximately normally distributed. Quantities, such as examination grades, snowflake sizes, and so on, seem to be determined by large numbers of such factors, and so by the CLT, these quantities are approximately normally distributed (so the explanation goes).

The following is a representative example of this sort of explanation:

> *Why is the Normal Curve Normal?*
>
> The primary significance of the normal distribution is that many chance phenomena are at least approximately described by a member of the family of normal probability density functions. If you were to collect a thousand snowflakes and weigh each one, you would find that the distribution of their weights was accurately described by a normal curve. If you measured the strength of bones in wildebeests, again you are likely to find that they are normally distributed. Why should this be so? [. . .] It turns out that if we add together many random variables, all having the same probability distribution, the sum (a new random variable) has a distribution that is approximately normal. [. . .] This result is formally called the *Central Limit Theorem*, and it provides the theoretical basis for why so many variables that we see in nature appear to have a probability density function that approximates a bell-shaped curve. If we think about random biological or physical processes, they can often be viewed as

[1] Quote attributed to Henri Poincaré by de Finetti ([1990], p. 63) with respect to the view that all, or almost all, distributions in nature are normal. (However, (Cramer ([1946], p. 232) attributes the remark to Lippman and quoted by Poincaré.)

> being affected by a large number of random processes with individually small effects. The sum of all these random components creates a random variable that converges on a normal distribution regardless of the underlying distribution of processes causing the small effects. (Denny and Gaines [2000], pp. 82–3)

One finds very similar statements throughout the sciences. From electrical engineering:

> The central limit theorem explains why the Gaussian random variable appears in so many diverse applications. In nature, many macroscopic phenomena result from the addition of numerous independent, microscopic processes: this gives rise to the Gaussian random variable. (Leon-Garcia [2008], p. 369)

To insurance and finance:

> The central limit theorem explains the wide applicability of the normal law to approximate the result of a stochastic experiment influenced by a large number of random factors. (Bening and Korolev [2002], pp. 36–7)

Sometimes the theorem is employed to explain generally why quantities in nature tend to be distributed normally, and sometimes it is employed for particular cases—for example, why people's heights are normally distributed.

I will argue (Section 3) that the general explanation for why normal distributions are normal is false and that, very often, so are the explanations in particular cases. I'll also argue that, often, we have no (or very little) epistemic grounds for giving such explanations. Moreover, I'll raise some doubts concerning the general explananda that normal distributions are 'normal'. What does that mean, exactly? And is it even true? (As we'll see in Section 3.4, many distributions in nature that are apparently normal may actually be log-normal or some other distribution, and in Section 3.6, we'll see that whenever there is a normal distribution in nature, there are many 'nearby' distributions that are not normally distributed.)

First, though, I'll give a brief overview of some basic technicalities and terminology that I'll use throughout the article.

2 Normal Distributions and the Central Limit Theorem

2.1 Normal distributions

The standard formulation of a normal distribution is given by specifying its variance, σ^2, and mean, μ, and is often written in short form as $N(\mu, \sigma^2)$.[2] Figure 1 (left) shows one particular distribution. The mean, μ, controls the location of the peak of the distribution, and σ^2 controls how 'fat' the distribution is: a larger value for σ^2 results in a 'fatter' bell curve.

[2] The probability density function for the normal distribution is: $p(x) = e^{-(x-\mu)^2/2\sigma^2}/(\sigma\sqrt{2\pi})$.

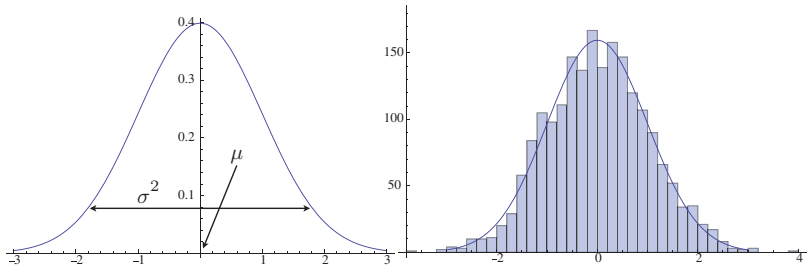

Figure 1. (Left) A normal distribution with mean 0 and variance 1. (Right) A bell curve approximating a data set (not normalized).

Many actual relative frequency distributions in nature can be well-approximated by a member of the family of normal distributions, with the appropriate mean and variance. Often, such relative frequency distributions are obtained by suitably normalizing frequency distributions (i.e. histograms). Figure 1 (right) shows a frequency data set, which hasn't been normalized, that is well-approximated by a bell curve.

2.2 The Central Limit Theorem

The name 'the CLT' is actually used to refer to many different theorems, so there is no single theorem that is the CLT. However, the theorem that is perhaps most commonly referred to by this name is the following:

> **The Central Limit Theorem:** Let $x_1, x_2, \ldots, x_n$ be a sequence of random variables that are identically and independently distributed, with mean $\mu = 0$ and variance σ^2. Let $S_n = 1/\sqrt{n}(x_1 + \cdots + x_n)$. Then the distribution of the normalized sum, S_n, approaches the normal distribution, $N(0, \sigma^2)$, as $n \to \infty$.

I'll introduce other versions of the theorem as they become relevant to the discussion later in this article. The above theorem, though, is a convenient starting point as it is one of the simplest and most often cited versions of the theorem.

The CLT is in some ways very counterintuitive. This is because the distribution of the x_i can be any distribution with mean of 0 and variance σ^2, and it can be hard to imagine how, by simply summing such random variables, one obtains a random variable (in the limit) with a normal distribution. Moreover, it can be the case that for finite sums with a small number of terms, the distribution of the resulting random variable is well-approximated by a normal distribution. Figure 2 shows that four independent random variables that are uniformly distributed over the $[-1, 1]$ interval are already well-approximated by a normal distribution when they are summed together.

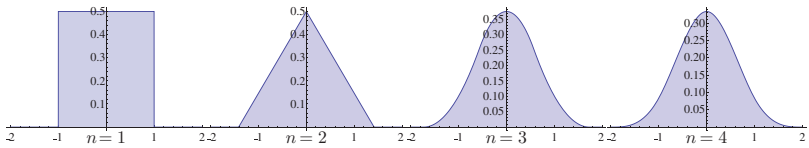

Figure 2. A random variable with uniform distribution over [−1, 1] added to itself repeatedly. After only four summations, the resulting distribution is very close to being a normal distribution.

2.3 Terminology

Strictly speaking, there are no quantities in nature that are distributed normally—or at least, there are very few. However, there are many quantities in nature with frequency distributions that are very well-approximated by normal distributions. Unfortunately, it can be tedious to keep writing (and reading) 'a quantity whose values are well-approximated by a normal distribution, when suitably normalized'. So, although it is not strictly speaking correct, it will nevertheless often be convenient to simply to refer to such quantities as normally distributed quantities. When the difference matters, I'll be explicit about it.

Random variables are mathematical objects, and quantities are physical things, such as numbers of people, strengths of forces, weights of objects, and so on. Nevertheless, it will often be convenient to refer to a quantity as a random variable.

A quantity is often determined by other quantities. For example, the amount of liquid in my coffee cup (a quantity) is the sum of (a way of being determined by) the amount of water and the amount of milk in my coffee cup (two other quantities). I will refer to the quantities that determine another quantity as the latter's factors. When convenient, I will also refer to factors as random variables. So, for example, the amount of liquid in my coffee cup is a random variable that is the sum of two other random variables: the amount of water and the amount of milk in my coffee cup.

The notions of a factor and determination are used in this article in a very metaphysically thin way. For example, it may be convenient to say that the amount of milk in my coffee cup is determined by two factors: the total amount of liquid minus the amount of water. It is true that we seem to have a notion of determination according to which only the amounts of water and milk determine the total amount of liquid, but I won't be using that notion in this article.

Finally, the concept of probability will often be used in this article. Our concept of probability is notoriously difficult to analyze. Fortunately, the way I need to use the notion in this article and the way it is used in the surrounding

literature allows us to understand it either as an actual relative frequency or as a purely mathematical notion, i.e. a function that satisfies Kolmogorov's axioms.

3 Explaining Normality

It's very common to find statements of the CLT explaining why a particular random variable is normally distributed. For example:

> [S]uppose you bake 100 loaves of bread, each time following a recipe that is meant to produce a loaf weighing 1,000 grams. By chance you will sometimes add a bit more or a bit less flour or milk, or a bit more or less moisture may escape in the oven. If in the end each of a myriad of possible causes adds or subtracts a few grams, the [CLT] says that the weight of your loaves will vary according to the normal distribution. (Mlodinow [2008], p. 144)

Virtually any textbook on applied statistics, or a field that relies heavily on statistics (for example, population ecology), that mentions the CLT includes a similar statement. When read literally, some of these texts are claiming that a mathematical theorem explains an empirical fact:

> The [CLT] explains why the normal distribution arises so commonly and why it is generally an excellent approximation for the mean of a collection of data (often with as few as 10 variables). (Gregersen [2010], p. 295)

Or, better:

> [I]t is undeniable that, in a large number of important applications, we meet distributions which are at least approximately normal. Such is the case, for example, with the distributions of errors of physical and astronomical measurements, a great number of demographical and biological distributions, etc.
>
> The central limit theorem affords a theoretical explanation of these empirical facts [. . .] (Cramer [1946], p. 231)

On the face of it, these statements seem quite odd, perhaps even false. It surely can't be the theorem by itself that is doing the explanatory work. At the very least, there must be additional premises that connect the theorem to their explananda.[3] Without additional premises, the theorem is explanatorily idle,

[3] Those who deny that there are mathematical explanations of empirical facts will also want to replace the theorem with something non-mathematical. Although I believe that this strategy in general would result in an explanatory impoverishment of the sciences (Lyon [2012]), my purpose here is not to argue for that claim. My goal is to understand what scientists mean when they say things like: 'The central limit theorem affords a theoretical explanation of these empirical facts', and so it will be necessary to be at least open-minded to the possibility of mathematical explanations of empirical facts. I won't be assuming that the mathematics in such explanations is indispensable or provides a basis for mathematical realism (see, for example, Baker [2005], [2009], [2012] for further discussion).

disconnected from any empirical phenomena. It's therefore worthwhile to tease out the premises that connect the theorem to quantities, such as weights of bread, and give the explanations in full detail. This turns out to be surprisingly difficult—so difficult that it seems that the CLT does not explain why quantities are normally distributed as often as the textbooks suggest.

3.1 Loaves of bread

As a sort of 'warm up' to the problems that follow, and to simplify matters, let's focus on a particular case: the distribution of the one hundred loaves of bread. According to the above quote, the CLT says that their weights will be normally distributed if each of a myriad of possible causes adds or subtracts a few grams. Let's suppose that the weights are normally distributed with a mean of 1000 g and with some variance. How might the CLT explain this? We know that the weight of a given loaf of bread is the sum of a number of factors, such as the weight of flour used; the weight of water used; the weight of yeast, sugar, and salt used; the weight of water lost during the baking process, and so on. Let's associate the random variable W with the weight of a loaf of bread, and the random variables x_i with the factors that sum together to determine the weight of the loaf of bread. We know that:

$$W = x_1 + x_2 + \cdots + x_n$$

for some fixed n. If n is large, then it seems we might be able to use the CLT to show that W is approximately normally distributed—as the above quote states. To do that, we need to show that the x_i satisfy the conditions of the CLT, i.e. that their distributions are identically and independently distributed.

However, this is straightforwardly not the case. A simple recipe for a 1,000 g loaf of bread calls for about 625 g of flour and 375 g of water (this is the standard 5:3 ratio used by bakers). Assuming a basic proficiency in baking of our baker, the mean weight of flour used would be about 625 g and the mean weight of water used would be 375 g. So at least two of the x_i are not identically distributed: as their distributions have different means.

To proceed with the explanation, we need to first transform our variables so that they all have a common mean.[4] A natural way to do this is to subtract the means away from their respective variables. So, for example, instead of using the weight of flour, x_1, we need to use the discrepancy between the weight of flour and the mean of x_1: x_1, namely, 625. Our new random variables, however, don't sum to the total weight of the bread, W. Instead, they sum to the

[4] Assuming we don't use another version of the CLT, which would also be a way of dealing with the difficulty. We will eventually have to do this anyway, and I discuss how one might use other versions of the CLT later in the article.

discrepancy between W and the mean of W, which is, say, 1000. So the sum that we need to examine is:

$$(W - 1,000) = (x_1 - 625) + (x_2 - 375) + \cdots + (x_n - c_n)$$

where c_i is the actual mean of the random variable x_i. If we can show that $W - 1000$ is normally distributed, then we can conclude that W is normally distributed because normal distributions are invariant under additions of a scalar.

3.2 Varying variances and probability densities

It remains to be shown that the $x_i - c_i$ are identically and independently distributed. For convenience, I'll call the new random variables x_i'. Let's first start with the identity of the distributions, which I'll call p_i. By definition, their means exist and are all 0, but there is more to a distribution than its mean. For example, there is its variance. Are the variances of each of the x_i' all identical? It seems plausible that in fact they are not. For example, the variance associated with the amount of salt is plausibly smaller than the variance associated with the amount of flour. If the variances are not all identical, then the CLT doesn't apply, and the explanation doesn't go through. Moreover, for all we know, it's possible that some of the p_i have different functional forms. For example, the weight of flour used (minus 625 g) may be a normal distribution, but the amount of water lost during the baking process (minus its actual mean) might be a symmetric bimodal distribution, with probability mass heaped over ± 1 and almost entirely absent over 0.

Fortunately, there is a more complicated version of the CLT that relaxes the condition that the distributions be of the same form (or have the same variance), so long as they satisfy what is called the Lindeberg condition.

> **Central Limit Theorem (Lindeberg–Feller):** Let x_i be mutually independent random variables with distributions p_i such that $E(x_i) = 0$ and $\mathrm{Var}(x_i) = \sigma_i^2$. Define $s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$. Then, if
>
> $$\text{(Lindeberg condition) For every } t > 0,\ \sqrt{s_n} \sum_{k=1}^{n} \int_{|y| \geq t s_n} y^2 p_i\{\mathrm{d}y\} \to 0$$
>
> then the distribution of $S_n = (x_1 + \cdots + x_n)/s_n$ tends to the normal distribution with mean of 0 and variance of 1 (Feller [1971], p. 262).

The Lindeberg condition is complicated, but it entails a simpler condition that is easier to understand and is still informative: it guarantees that the individual variances are small compared with their sum (Feller [1971], pp. 262–3).

Are the variances small with respect to their sum? It seems that they need not be. For example, the variance associated with the weight of flour might be quite large compared with the sum of the variances. One way for this to

happen is if the baker has very precise instruments for measuring water, sugar, yeast, salt, and so on, but a fairly imprecise instrument for measuring flour (for example, the baker 'eye balls' it). The resulting weights of the breads that the baker produces can be normally distributed in such a scenario, and yet this latest version of the CLT won't apply.

One option in response to this is to factor out the random variables with large variances.[5] For example, instead of trying to use the CLT to show that $(W - 1000)$ is normally distributed, one might try use the CLT to show that:

$$(W\prime - x_1' - x_2') = x_3' + \cdots + x_n'$$

is normally distributed—where x_2 corresponds to the error in water and is assumed to also have a reasonably large variance. Of course, it would then need to be the case that the variances of x_i' for $3 \leq i \leq n$ are small compared with their sum. If not, one would need to also factor out those x_i' with large variances (relative to the new standard of 'large'). However, there is the worry that once this process is completed, we are left with only a few random variables, or even with no random variables. We started with only six specified factors—the weights of flour, water, sugar, salt, yeast, and water lost during the baking process—and are plausibly down to four.[6] Of course, there will always be some more factors that are not specified (for example, water lost during the proofing process), but they will be very small compared with those already listed, and their distributions will have little effect on the total distribution. The CLT is about what happens in the limit, as n approaches infinity. In actual cases, where n is finite but large, we might reasonably expect the CLT to show why the distribution of interest is approximately normal. But if n is small, it's hard to see how the CLT could apply even approximately.

Moreover, as we have to apply a number of transformations after using the CLT, the approximation may get even worse. One now needs to include the additional premise into the explanation that, for example, deconvoluting the distributions of x_1' and x_2' from an approximate normal distribution results in an approximate normal distribution.

3.3 Tensile strengths and problems with summation

Consider the following:

> The [CLT] explains why many physical phenomena can be described, approximately, by a normal distribution. For example, the tensile strength of a component made of a steel alloy can be considered to be

[5] This method of factoring out problematic factors is very similar to Galton's ([1875], p. 45) solution to a similar problem involving the effect of aspect on fruit size.

[6] One response to this is that factors, such as the weight of flour, break down into many sub-factors with small variances. I discuss this sort of response in Sections 3.5–3.8.

> influenced by the percentages of the alloying elements such as manganese, chromium, nickel and silicon, the heat treatment it received, and the machining process used during its production. If each of these effects tends to combine with the others in determining the value of the tensile strength, then the tensile strength can be approximated by a normal distribution according to the [CLT]. (Roush and Webb [2000], p. 166)

Here, the quantity that is normally distributed (or supposed to be) is the tensile strength of a steel alloy component (for some machine, perhaps). The factors that determine the tensile strength of such a component are (among others): the percentages of manganese, chromium, nickel, and silicon that make up the alloy; the heat treatment; and the machining process used during production. Again, for the CLT to apply, we need to find a sequence of random variables that are identically and independently distributed, and that sum to the tensile strength of the component. Presumably, these would be random variables naturally associated with the factors just mentioned.

Forget for the moment the issue of whether the factors are identically and independently distributed (or satisfy the Lindeberg condition), and focus on just finding a set of factors that sum to the tensile strength. Before, with the bread example, the corresponding task was easy. The factors that determined the weight of the bread were all weights themselves, and measured in the same units (g). And so the sum of the factors equalled the weight of the bread. Here, however, this is not the case. The factors listed above are not tensile strengths; they are a diverse range of quantities, some of which do not seem to even have a standard numerical representation (for example, what are the units for 'machining process'?). So the factors cited could not possibly sum to the tensile strength of the component. They may combine in other ways to determine the tensile strength, but they do not do this by summation. The CLT, nevertheless, is about the *sum* of a sequence of random variables, and so it seems it can't be used (with the cited factors) to explain why tensile strength is normally distributed (at least not as straightforwardly as the quote suggests).

3.4 Products of factors and log-normal distributions

Although the CLT involves a sum, it isn't essential that the quantity of interest break down into a sum. As the logarithm of a product of factors is the sum of the logarithms of the factors, it may suffice for the quantity to be a product of factors. If X is an infinite product of random variables and the logarithms of those variables are appropriately distributed (for example, satisfy the Lindeberg condition), then $\log(X)$ is normally distributed, and X is log-normally distributed. If the variance of a log-normal distribution is small compared with its mean, the log-normal distribution is very similar to a

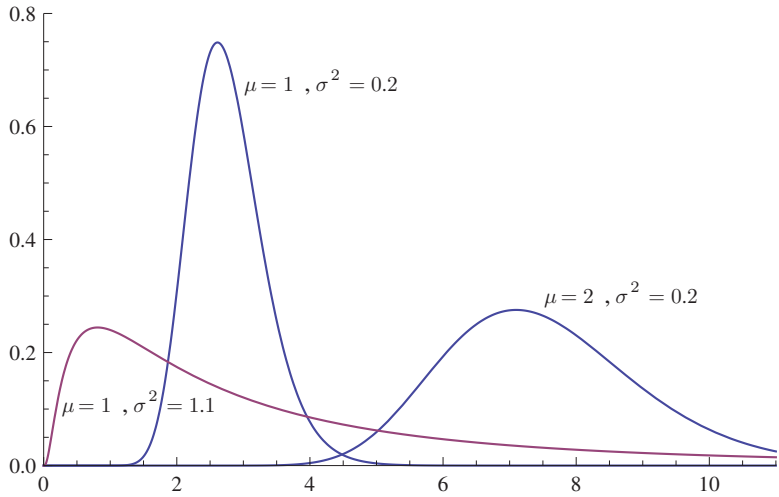

Figure 3. Three log-normal distributions. When the variance is small with respect to the mean, the log-normal distribution looks very similar to the normal distribution. However, when the variance is not small with respect to the mean (for example, $\mu = 1, \sigma^2 = 1.1$), the log-normal can look very different.

normal distribution (Figure 3). So if one can show that a quantity is approximately log-normally distributed, then that may suffice to show that it is approximately normally distributed.

Interestingly, this suggests that the general explanandum—that normal distributions are normal—is false. Perhaps log-normal distributions with small variances are normal and we confuse them for normal distributions. Or, perhaps both normal and log-normal distributions with all variances are normal. (Log-normal distributions with relatively high variances are common in nature—for example, survival times of species, concentrations of minerals in the Earth, times to first symptoms of infectious diseases, numbers of words per sentence, and personal incomes.[7]) Or, perhaps it is only log-normal distributions, of all variances, that are normal.

Limpert *et al.* ([2001])—a study of the use of both distributions in science—found no examples of original measurements that fit a normal distribution and not a log-normal one (it's trivial to find cases of the opposite). The only examples of a normal distribution fitting data better than a log-normal distribution were cases where the original measurements had been manipulated in some way ([2001], p. 350). Limpert *et al.* make a strong case that normal distributions are not as common as is typically assumed, and that log-normal

[7] For more examples, see Limpert *et al.* ([2001]).

distributions may in fact be more common. They argue that the multiplication operation is more common in nature than addition:

> Clearly, chemistry and physics are fundamental in life, and the prevailing operation in the laws of these disciplines is multiplication. In chemistry, for instance, the velocity of a simple reaction depends on the product of the concentrations of the molecules involved. Equilibrium conditions likewise are governed by factors that act in a multiplicative way. From this, a major contrast becomes obvious: The reasons governing frequency distributions in nature usually favor the log-normal, whereas people are in favor of the normal. (Limpert *et al.* [2001], p. 351)

(By referring to people's preference for the normal distribution, Limpert *et al.* are alluding to its mathematical and conceptual convenience/simplicity.) They also argue that quantities that can't take negative values (for example, people's heights) can't be normally distributed, as any normal distribution will assign positive probability to negative values. However, they can be log-normally distributed, as log-normal distributions are bounded below by zero ([2001], pp. 341–2). If they are correct, then the explananda of the CLT explanations that one finds in textbooks is false and all the examples mentioned so far—bread weights, human heights, bone strengths of wildebeests, tensile strengths, and so on—are all wrong. All these quantities would be log-normally distributed and so, presumably, their factors multiply instead of sum together.

Moreover, there have been cases where data that were once thought to be normal actually turned out to be better accounted for by some other distribution. For example, Peirce ([1873]) analysed twenty-four sets of five hundred recordings of times taken for an individual to respond to the production of a sharp sound, and concluded that the data in each set were all normally distributed. However, Wilson and Hilferty ([1929]) conducted a reanalysis of the data and found that each of these data sets were not normally distributed, for various reasons (see also Koenker [2009] for discussion). So the normal distribution may not be as normal as people previously thought. Indeed, prior to 1900 it seems that it was commonly accepted that normal distribution was practically universal and was labelled as 'the law of errors', thus prompting the remark by Poincaré quoted at the beginning of this article. Nevertheless, for the sake of argument, I will assume that the normal distribution is at least common in nature (see, for example, Frank [2009]) and, unless otherwise specified, I will run normal distributions and log-normal distributions together. My objections to CLT explanations apply to both cases equally well.

Returning to the example of the tensile strength of the steel alloy, that tensile strength may be a product of appropriate factors doesn't really help. Percentages of manganese, chromium, and so on, do not multiply into a tensile strength—and 'machining process' doesn't multiply with anything into

anything, let alone a tensile strength. Strictly speaking, we need not be limited to pure sums and pure products; combinations of the two may work as well. For example, if Z breaks up into a product and sum:

$$Z = X + Y = (x_1 x_2 \cdots x_n) + (y_1 + y_2 + \cdots + y_m),$$

then to show that Z is normally distributed, it would suffice to show that X and Y are normally distributed separately and that their sum is also normal. Explanations in terms of other, more complicated combinations of sums and products may work in a similar manner. But the same problems arise: no combination of sums and products of percentages, 'heat treatment', and 'machining process' will be a tensile strength.

3.5 Transforming factors and sub-factors

At the very least, some initial transformation of the variables has to be made. The percentage of an element has to become something such as total mass of that element, which may then be rescaled by some physical constant. And 'heat treatment' and 'machining process' need to be specified in some precise way so that they have numerical representations and physical units. Perhaps, then, there is some combination of sums and products of the factors that result in the tensile strength of the alloy component. But even that may not be enough; more complicated transformations may need to be applied to the factors—exponentials, sinusoidals, hyperbolics, and so on.

Suppose that there is some set of transformations of the factors x_i such that X is some combination of sums and products of them—call the transformed factors x_i'. It needs to be the case that the x_i' are appropriately distributed. However, we rarely know how the factors are distributed. Indeed, it is very rare that the factors are even specified and shown that they sum and/or multiply to X. (In Section 3.7, I examine a case where the factors are specified and I argue that they don't satisfy the conditions of the CLT.) It seems that it is often simply assumed that there is some set of appropriate factors that sum and/or multiply to the quantity of interest:

> [T]he [CLT] explains the common appearance of the 'Bell Curve' in density estimates applied to real world data. In cases like electronic noise, examination grades, and so on, *we can often regard* a single measured value as the weighted average of a large number of small effects. Using generalisations of the [CLT], we can then see that this would often (though not always) produce a final distribution that is approximately normal. (Mandal [2009], p. 31, emphasis added)

However, without explicitly knowing what the factors are, what their distributions are, and how they combine to determine the quantity of interest, it

seems we can't know whether the CLT actually applies, and therefore whether a CLT explanation for why the quantity of interest is normally distributed can be given.

Behind the purported CLT explanations, there seems to be the idea that for a given quantity of interest, there are so many factors, so many ways of breaking factors up into sub-factors, so many ways of combining factors, and so many ways of transforming factors, that there must be some set of appropriate factors—possibly transformed—that sum and/or multiply to the quantity. Once we have that set of factors, the CLT kicks in, and we can explain why the quantity of interest is normally distributed (so this line of thought goes).

3.6 Transformations of quantities

However, this line of reasoning has the potential to over-generate. There are many quantities in nature that are not normally distributed, and yet they have many factors, and there are many ways of transforming those factors, and so on. It's always possible to transform a quantity, X, that we know to be normally distributed into another quantity, Z, that is not normally distributed. For example, let H be the height of the males at a university, and $G = T(H)$ where:

$$\begin{aligned} T(H) &= H + 40, \quad \text{if } H < 175 \\ &= H - 40, \quad \text{if } H \geq 175. \end{aligned}$$

If H is normally distributed with mean 175 and variance 10, then G isn't normally distributed (for example, Figure 4).

If there are a large number of factors that determine H values, then this is also true for G. Do these factors sum to H in a way that satisfies the CLT? The line of thought from the previous section was that there are so many sets of factors that determine H, and so many ways of transforming them, and so on, that there must be some set of factors that combine in some combination of sums and products to equal H. However, this reasoning seems to apply equally well for G. How, then, can we justify applying the line of argument to H and not to G?

We know that if H is normally distributed, then G isn't, since G is $T(H)$ and this transformation neither preserves nor even approximates normality. This transformation is the final combination of sums and products of factors that determine G, and so it seems that it's because the transformation is part of what determines G that G is not normally distributed. Perhaps this is what separates H and G. To complete the argument, we need to show that there are no similar transformations in the determination of

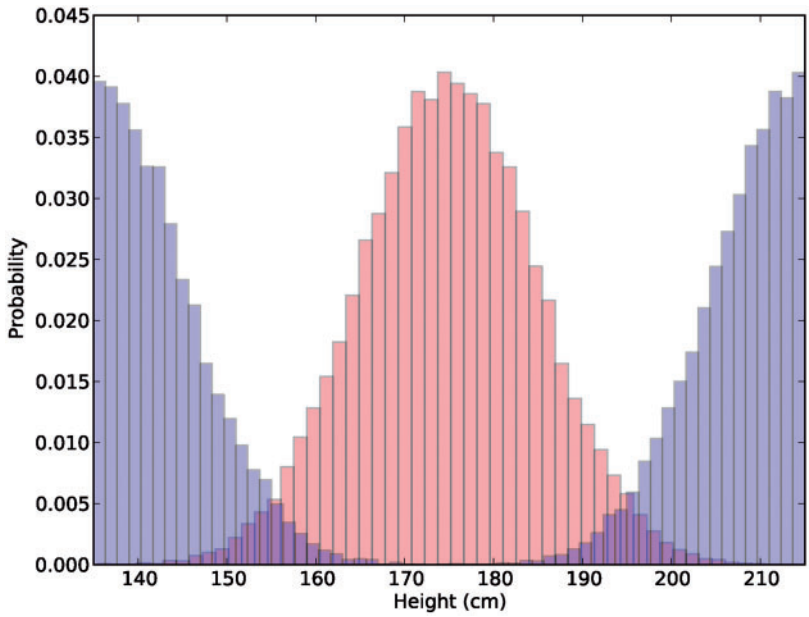

Figure 4. The middle (bell-shaped) histogram is the approximate normal distribution of H, and the outer (inverted) histogram is the distribution of G obtained by transforming the distribution of H.

H. The problem, though, is that H can be defined in terms of G as $H = T^{-1}(G)$, where:

$$T^{-1}(G) = G + 40, \quad \text{if } G < 175$$
$$= G - 40, \quad \text{if } G \geq 175.$$

T^{-1} doesn't preserve normality and it is part of the determination of H, so now the situation for H and G is reversed. Therefore, there doesn't seem to be any difference between H and G (that we know about) that would allow us to justify assuming the conditions of the CLT are true for H and not true for G. (At this point, one may be tempted to make an appeal to the principle of inference to the best explanation. If so, see Section 3.8.)

One difference between H and G is that H is in some sense 'natural' whereas G isn't.[8] However, the CLT makes no mention of naturalness, so it can't be the naturalness of H and unnaturalness of G alone that distinguishes the two quantities. The line of argument would have to be that at least one of the conditions of the CLT is more plausibly true for a natural variable than for an

[8] G is a grue-like quantity (Goodman [1955]).

unnatural one. I don't see how such an argument would go, however (similarly for other naturalness-like notions).[9]

Quantities in nature can always be transformed in these ways. Therefore, there is a sense in which the very question we are trying to answer—why are normal distributions normal?—is wrong. All distributions are normal. For every quantity that is normally distributed, there is another quantity that is a transformation of the first quantity and is distributed in some other way. The question, therefore, probably ought to be formulated as something like: why do the quantities in nature that we tend to focus on tend to be normally distributed? The answer to this question might plausibly involve the 'naturalness' of the quantities that we tend to focus on, and this might allow us to justifiably treat H and G differently. However, it's still not clear how the 'naturalness' of a random variable could be used to bolster a CLT explanation.

3.7 Quantitative genetics

In Sections 3.2 and 3.3, I discussed examples of CLT explanations that make only a cursory reference to the factors that are meant to sum to the quantity of interest and satisfy the conditions of the CLT. I argued that if we use the factors mentioned in the explanations, then those explanations simply don't work. In Sections 3.4 and 3.5, I considered two ways in which we might be able to save—or at least help—the explanations (by allowing combinations of sums and products, and finding alternative factors through transformations or breaking factors down into sub-factors). In Section 3.6, I discussed a problem with the second strategy for saving the explanations (the problem of potential over-generation). These arguments, I believe, show that we have little epistemic grounds for giving such explanations and that we are probably wrong in giving a CLT explanation for the normality of a given distribution.

The situation is worse for purported CLT explanations that make explicit reference to the factors that determine the quantity of interest. In these cases, very specific and salient factors are mentioned and play a crucial and central role in their CLT explanations. The clearest examples of such explanations are the CLT explanations for the normal distributions of various phenotypic traits studied in quantitative genetics.

[9] Even if such an argument could be made, there could still be a problem. Sometimes H and G will be equally natural variables:

> [T]he weight of an object depends on the product of its three linear dimensions with its density. Necessarily, if the linear dimension is precisely normally distributed, the triple product cannot be normally distributed and in fact the resultant distribution approaches log normality. (Koch [1969], p. 254)

Restricting our attention to natural variables, therefore, doesn't guarantee there won't be problematic transformations. (Thanks to Neil Thomason for pointing this out to me.

A common explanation for why people's heights (for example) are normally distributed is that a person's height is largely determined by their genes, and that numerous genes contribute additively to height, so by the CLT, heights are normally distributed:

> We will now consider the modern explanation of why certain traits, such as heights, are approximately normally distributed [...]
>
> We assume that there are many genes that affect the height of an individual. These genes may differ in the amount of their effects. Thus, we can represent each gene pair by a random variable x_i, where the value of the random variable is the allele pair's effect on the height of the individual. Thus, for example, if each parent has two different alleles in the gene pair under consideration, then the offspring has one of four possible pairs of alleles at this gene location. Now the height of the offspring is a random variable, which can be expressed as
>
> $$H = x_1 + x_2 + \cdots + x_n + W$$
>
> if there are n genes that affect height. (Here, as before, the random variable W denotes non-genetic effects.) Although n is fixed, if it is fairly large, then Theorem 9.5 [a slightly weaker version of Lindeberg–Feller version of the CLT] implies that the sum $x_1 + x_2 + \cdots + x_n$ is approximately normally distributed. Now, if we assume that the x_i's have a significantly larger cumulative effect than W does, then H is approximately normally distributed. (Grinstead and Snell [1997], pp. 347–8)

Gillespie also sketches a similar explanation:

> There is one very important property of the two-allele model that does change when more loci are added: The distribution of the genetic effects approaches a normal distribution [...] There is no a priori reason why the phenotypic [i.e. height] distribution should be so, well, normal. The Central Limit Theorem from probability theory does provide a partial explanation. This theorem states that the distribution of the sum of independent random variables, suitably scaled, approaches a normal distribution as the number of elements in the sum increases [...]
>
> Of course, the phenotypic distribution also has an environmental component that must itself be approximately normally distributed if the phenotypic distribution is to be normally distributed. In fact, this appears to be generally true as judged from an examination of the phenotypic distribution of individuals that are genetically identical, as occurs, for example, in inbred lines. Perhaps the environmental component is also the sum of many small random effects that add to produce their effects on the phenotype. (Gillespie [1998], pp. 129–30)

It's a little difficult to read Gillespie here as it seems he is not completely convinced of this argument. However, earlier he makes a reference to this two-allele model and writes:

> In the last section of this chapter, we will show how the one-locus model may be replaced with a more realistic multilocus model, where each locus

> may have only a couple of alleles. *The normality then comes*, when the number of loci is large, *from the Central Limit Theorem*. (Gillespie [1998], p. 106, my emphasis)

There are several reasons to be skeptical of such explanations.[10] (I'll focus on Grinstead and Snell's version since it's more precise and uses a more general version of the CLT.)

First, there is a problem with the construction of random variables whose values are each 'allele pair's effect on the height of the individual'. Consider an analogy with the lengths of Wikipedia articles. For some authors, it is possible to isolate their contribution to the article (for example, five lines of text). However, the contributions of others cannot be isolated in this way. One author may write one line with completely false information. This causes another author to delete that one line, and replace it with three more. If the first author hadn't written the false information, the second wouldn't have added the three lines. Because of this, we can't associate random variables with each author's contribution that all have the same units and sum to the total length of the article. Genes can work in similar ways to determine a phenotype such as height, and so for the same reason, we can't associate random variables with each allele pair's effect on height that are all measured in the same way and sum to total height.[11]

The second problem is closely related to the first, but still distinct. Genes can regulate each other's expression through gene regulatory networks. They also interact through the developmental mechanisms that convert gene products into the components of a trait (for example, muscle tissue). This means that a gene's contribution to height can be strongly dependent on the expression of other genes. However, independence of factors is required for the CLT to apply. Indeed, it's striking that there is no mention of the condition of independence in the above passage (or its surrounding text), even though the version of the CLT that the authors cite (Theorem 9.5) requires it. There are generalizations of the CLT that allow for some dependencies between factors, so long as other constraints are met (for example, Hoeffding and Robbins [1948]). However, it is by no means clear that genetic factors satisfy these constraints.

[10] They originate from, or at least are inspired by, Francis Galton's work in this area. See Stigler ([1986], Sections 2–3) for a historical account of the development of such explanations.

[11] Interestingly, the size distribution of featured Wikipedia articles (in bytes) is very well-approximated by a log-normal distribution (http://en.wikipedia.org/wiki/User:Dr_pda/Featured_article_statistics). A standard CLT explanation of this distribution might be that there are many factors that determine the size of a featured article—say the contributions of different authors—and these factors are independent and multiply together to form the total size of each article. Of course, this is not true; at best, they might sum together. To run a CLT-style explanation for the distribution of featured article size one would have to find some set of factors that are appropriately distributed and multiply to the total byte size of the article (or something that could be transformed to total byte size without destroying approximate log-normality). It's not at all clear what those factors would be.

Third, it's estimated that about 80% of human height is due to genetics (Visscher *et al.* [2006]). The remaining 20% or so is due to non-genetic factors, such as interactions with the environment and perhaps epigenetic effects. This means that W's contribution to H is by no means small. Of course, W breaks down into sub-factors. However, the same worries apply to those as well—for example, there might be interactions between environmental factors and any epigenetic ones.

Fourth, as mentioned earlier, there is some reason to believe that the distribution of heights is better accounted for by a log-normal distribution. Limpert *et al.* ([2001]) point out that quantities that can't take non-negative values can't be normally distributed, and so are more likely to be log-normally distributed. Heights—and many other traits studied in quantitative genetics—obviously don't take negative values, so they may be better understood as being log-normally distributed. In which case, it might be more reasonable to suppose that at least some of the effects of the gene pairs are multiplicative, rather than additive.

Some authors have noted that the conditions of the CLT do not generally apply for phenotypic traits:

> If our random variable is the size of some specified organ that we are observing, the actual size of this organ in a particular individual may often be regarded as the joint effect of a large number of mutually independent causes, acting in an ordered sequence during the time of growth of the individual. If these causes simply add their effects, which are assumed to be random variables, we infer by the central limit theorem that the sum is asymptotically normally distributed.
>
> In general it does not, however, seem plausible that the causes co-operate by simple addition. It seems more natural to suppose that each cause gives an impulse, the effect of which depends both on the strength of the impulse and on the size of the organ already attained at the instant when the impulse is working. (Cramer [1946], p. 219)[12]

By making some other assumptions, Cramer goes on to argue that the CLT can apply to these impulses and explain the observed distribution of organ sizes—in particular, Cramer shows how the impulses may generate a log-normal distribution. However, several of the assumptions that Cramer makes seem implausible. For example, Cramer supposes that the impulses are independent of each other, without telling us what they are. (This is also strange because Cramer introduces the argument as an example of how the CLT can be extended to cases in which the factors are not independent (Cramer [1946], p. 219).)

[12] However, Cramer later writes that 'it often seems reasonable to regard a random variable observed, for example, in some biological investigation as being the total effect of a large number of independence causes, which sum up their effects' (Cramer [1946], p. 232).

Hartl and Clark are also aware that the conditions of the CLT are not always satisfied:

> Many measurable quantities in the real world are determined by such sums of independent causes. For example, the multiple genetic and environmental factors that determine quantitative traits may be approximately additive in their effects, so that the central limit theorem is expected to hold. For many characters, the factors appear to multiply in their effects, and in these cases a logarithmic transformation gives a better approximation to the normal distribution [...]
>
> The key factor in arriving at a normal distribution is the independence of the component normal factors. *Interdependence of causal factors does occur in quantitative genetics, and this can result in departure from the normal distribution.* (Hartl and Clark [1989], pp. 434–5, my emphasis)

They are correct to note that interdependence of causal factors does occur in quantitative genetics. However, it doesn't automatically follow that this results in a departure from the normal distribution.[13] Interdependent causal factors don't necessarily destroy normality: a quantity that is comprised of factors that are dependent on each other can quite easily be normally distributed. Interdependence does, however, have the potential to destroy the applicability of the CLT.

It's worth noting that Hartl and Clark are perhaps a little optimistic in their claim that many measurable quantities are determined by sums of independent causes. As an example, they cite genetic and environmental factors that determined quantitative traits, but we know that such factors are often not independent of each other. Moreover, they seem to be overly optimistic when they write that such factors 'may be' approximately additive so the CLT can be 'expected to hold'. Of course, the factors may be approximately additive, but they also may not be. From the fact that they may be approximately additive, it doesn't follow that the CLT is expected to hold.

One way to potentially fix things so that the condition of independence is satisfied is to group dependent terms together and consider them as single terms themselves.[14] For example, if x_1 and x_2 are dependent on each other and so are x_4, x_5, and x_6, then we might proceed as follows:

$$\begin{aligned} H &= (x_1 + x_2) + x_3 + (x_4 + x_5 + x_6) + x_7 + \cdots + x_n + W \\ &= x_{1,2} + x_3 + x_{4,5,6} + x_7 + \cdots + x_n + W. \end{aligned}$$

Then, by definition, the terms in the sum would be independent of each other, thus satisfying the independence condition. This is the reverse of the proposal

[13] I don't believe that Hartl and Clark think that this automatically follows. I'm only emphasizing a point that can easily be missed.

[14] Thanks to Lekki Wood for pointing this out to me.

in Section 3.5, which was to break factors up into smaller sub-factors. Here, we are combining factors into sub-factors.

However, if there are large groups of factors that are dependent on each other, then some of the sub-factors will be large, thus violating the Lindeberg condition. Moreover, there may be a 'six degrees of separation' effect that results in very large sub-factors.[15] Even if x_1 and x_2 are independent, they may nevertheless need to be grouped together. If x_3 is dependent on both x_1 and x_2, then it needs to be grouped with x_1 and x_2, which entails that x_1 and x_2 need to be grouped together (by transitivity of grouping). In general, if the regulatory network is sufficiently connected in this way, all of the terms, or at least large groups of them, would have to be grouped together.

One might reply that it may suffice if the independence condition is satisfied only approximately. If there are only weak dependencies between the genetic factors, then it seems reasonable to expect that the CLT applies approximately and that this would result in an approximately normal distribution.[16] However, not all of the dependencies between genetic factors are weak; there are clearly some strong dependencies and the structure of genetic regulator networks is incredibly complex (see, for example, Zhao *et al*. [2008]). So one could resort to the sub-factors argument, but it would also have to be established that the strong dependencies are few enough and structured in the right way so that the sub-factors are not so large as to violate the conditions of the CLT. This may in fact be the case; I'm only arguing that this is what one would have to establish to maintain an approximate application of the CLT. (Note that one still also has to show that the other conditions are satisfied, for example, the additivity of the effects.)

3.8 Inference to the best explanation

One may respond to the problems that I've raised by an appeal to the principle of inference to the best explanation. If we observe a normally distributed quantity that we know to be determined by a large set of factors, then by inference to the best explanation, we should infer that those factors satisfy the conditions of the CLT. In fact, this sort of reasoning can be crucial to understanding the processes that generate the variety of patterns we observe in nature:

> In general, inference in biology depends critically on understanding the nature of limiting distributions. If a pattern can only be generated by a very particular hypothesized process, then observing the pattern strongly suggests that the pattern was created by the hypothesized generative process [. . .] (Frank [2009], p. 1564)

[15] Also known as the Kevin Bacon Game, where in six movies or fewer, one links Bacon to another actor by the co-star relation.

[16] Thanks to Georges Rey and an anonymous reviewer for bringing this reply to my attention.

Surely, then, it is reasonable to maintain that if we observe that a particular quantity is normally distributed, and we have no reason to suppose otherwise, we ought to assume (at least as an initial hypothesis) that the quantity is determined by a set of factors that satisfy the conditions of the CLT. However, Frank continues:

> By contrast, if the same pattern arises as a limiting distribution from a variety of underlying processes, then a match between theory and pattern only restricts the underlying generative processes to the broad set that attracts to the limiting pattern. Inference must always be discussed in relation to the breadth of processes attracted to a particular pattern. (Frank [2009], p. 1564)

And this is important because:

> We do not know all of the particular generative processes that converge to the Gaussian. Each particular statement of the central limit theorem provides one specification of the domain of attraction—a subset of the generative models that do in the limit take on the Gaussian shape. (Frank [2009], p. 1574)

We, therefore, can't infer that the conditions of any given version of the CLT are satisfied simply on the basis of observing a normal distribution. We also can't infer that the conditions of some version of the CLT are satisfied, because all the versions of CLT do not exhaust the space of generative models that lead to the normal (Gaussian) distribution.

4 Maximum Entropy Explanations

If the previous sections are correct, then the CLT doesn't explain why normal distributions are normal. Indeed, it is questionable whether normal distributions really are all that normal: for every normal distribution, there is a cluster of other distributions that are not normal distributions (Section 3.4), and in many cases throughout the sciences, data that were once thought to be normal turns out to actually be better accounted for by some other distribution, such as the log-normal (Section 3.6). In addition to this, I've argued that in many cases, the CLT fails to explain why particular distributions are normal—for example, if heights are normally distributed, then the standard CLT explanations do not work, and it's not clear how they could.

Although normal distributions may not be as normal as we once thought they were, they do appear to be quite common—especially when we restrict our focus to quantities that tend to interest us. The CLT doesn't seem to be able to explain this weaker claim, but could there be some other explanation? Is it just a coincidence, or simply due to our predilection for simple and symmetric distributions? Similarly, could there be an interesting explanation for why a given distribution in nature is normal?

One intriguing alternative explanation involves an important mathematical property that the normal distribution has:

> A further fact, which serves to 'explain' why it is that this 'order generated out of chaos' often has the appearance of a normal distribution, is that out of all distributions having the same variance the normal has maximum entropy (i.e. the minimum amount of information). (de Finetti [1990], p. 62)

Put more formally: out of all distributions with mean μ, variance σ^2, and support over all of R, the normal distribution, $N(\mu, \sigma^2)$, maximizes entropy. And the class of normal distributions $N(-, \sigma^2)$ all maximize entropy subject to the constraints of a fixed variance σ^2 and support over R.[17]

Entropy is a notoriously slippery notion. Its main claims to fame are its roles in information theory, thermodynamics, and statistical mechanics—although, it's not clear that it is the one concept appearing in all these theories (for example, Popper [1982], Section 6). Even as a purely mathematical property of a probability function, entropy is difficult to define. Perhaps the standard definition is: the entropy of a continuous distribution, p, is $H(p(x)) = -\int_{\infty}^{\infty} p(x) \log p(x) \mathrm{d}x$ (for example, Frank [2009]). However, this is a special case of a more general definition, which includes an (arguably) arbitrary measure function, $m(x)$: $H_m(p(x)) = -\int_{\infty}^{\infty} p(x) \log p(x)/m(x) \mathrm{d}x$ (see, for example, Jaynes [1968] for a discussion of this definition). For present purposes, I'll assume that the entropy of $p(x)$ is $H(p(x))$. In rough and intuitive terms, the entropy of a distribution is a measure of how flat and smooth that distribution is; the flatter and smoother a distribution, the higher its entropy. The flattest and smoothest distribution over R is the uniform distribution, and the flattest and smoothest distribution over R with a given finite mean and variance is the normal distribution with that mean and variance.

This maximum entropy property of the normal distribution makes it an important kind of an attractor: if we start with some arbitrary distribution and repeatedly perform operations on it so that it increases in entropy but maintains its mean and variance, then the distribution will approach the normal distribution. The evolution of the distribution of S_n in the CLT is exactly like this. It starts off as simply the distribution of x_1, but then after the addition of x_2 it becomes the distribution of $S_2 = (x_1 + x_2)/\sqrt{2}$, then $S_3 = (x_1 + x_2 + x_3)/\sqrt{3}$, and so on. Each time another (independent and identically distributed) x_i is added, the entropy is increased, and as there is always the normalizing factor $1/\sqrt{i}$, the variance is maintained, and so the distribution of S_n approaches the normal distribution. In short, the CLT specifies one particular way in which a sequence of distributions increase in entropy and

[17] For more on the history of the normal distribution and its maximum entropy property, see Jaynes ([2003]), Chapter 7 and Stigler ([1986]).

remain constant in mean and variance, and thus get attracted to a normal distribution (Jaynes [2003], p. 221). In fact, as we saw in the previous section, the different versions of the CLT specify different ways in which distributions can approach the normal distribution (Frank [2009], p. 1574). Each version of the CLT identifies a set of operations (a 'generative model') that increase entropy and preserve variance (preserving the mean is not necessary as this only controls the location of the peak of the resulting normal distribution).

As an example of how an ME explanation might work, consider the tensile strengths from Section 3.3 and suppose that they are, in fact, normally distributed. There are many factors that determine the tensile strength of a component, and they do so in many different and complicated ways. All these factors and their interactions are determined by some machining process, which is designed by some engineers for the purpose of building the components. However, the engineers are not just interested in building the components and they are also interested in quality control. No machining process is perfect, so there will always be some error about the desired tensile strength. When building the machine that produces the components, the engineers will make sure that these errors are within some acceptable range—they don't need perfection, just something close to it. This amounts to fixing the mean (the desired the tensile strength), and the variance (the acceptable range of errors) of the distribution of tensile strengths. Apart from that, the engineers don't care what the machine does, and it will naturally tend to a state of maximal disorder—i.e. a state of maximum entropy—subject to its engineered constraints. (This is an appeal to something like the second law of thermodynamics, and I discuss this in more detail below.) The distribution of tensile strengths that maximizes entropy subject to those constraints is a normal distribution, and so that is why the tensile strengths are normally distributed.

It's worth considering another example of a completely different kind. Let's assume that the heights of a particular human population are normally distributed. There are many factors that determine a person's height, and they do so in many different and complicated ways. All these factors and their interactions are determined by some evolutionary selection process (natural, sexual, and so on), which is determined by the population's environment. The overall selection pressure determines an ideal height;[18] though the selection pressure is not perfect, some variability about the ideal won't matter very much. In fact, there may even be a selection pressure to maintain some variability to hedge against fluctuating circumstances in the environment. This amounts to fixing the mean (the ideal height), and an upper bound on the variance (the variation in the population) of the distribution of heights. Apart from that, there is no other relevant selection pressure, and the population will

[18] In fact, a range of ideal heights would be enough for this example to work.

naturally tend to a state of maximal disorder—i.e. a state of maximum entropy—subject to its selection constraints. (This is another appeal to something like the second law of thermodynamics.) The distribution of heights that maximizes entropy subject to those constraints is a normal distribution, and so that is why the heights in the population are normally distributed.

If such maximum entropy (ME) explanations could be made to work, they would enjoy at least two virtues over CLT explanations. First, ME explanations would be more modally robust than a CLT explanation, in that the ME explanations could support a wider range of counterfactuals. This is because ME explanations are less committal to the precise details of the aggregation process that converts the factors into the quantity of interest. Second, ME explanations appear to be appropriately sensitive to one's choice of variables. This is because the entropy of a distribution depends on one's choice of variables—at least according to the standard formulation, $H(p(x))$. This causes problems for the principle of maximum entropy that some have used to determine objective Bayesian priors (see, for example, van Fraassen [1989], p. 303). However, in this context, such dependence on one's choice of variables may be a virtue, as the explanandum itself appears to be dependent on one's choice of variables—i.e. normal distributions are common only if we carve up the world in the right way.

The are important obstacles that need to be overcome before we can conclude that ME explanations save the day, however. First, the examples I gave above assumed that the quantities in question are normally distributed. But what if they are log-normal, or something else? It seems that ME explanations ought to explain other common distributions in nature. Frank ([2009]) has made a lot of progress in this direction, demonstrating that several common distributions in nature can be modelled in terms of processes maximizing entropy. However, Frank doesn't cover an important class of distributions: the log-normals. One might hope that these could be accounted for by transforming the log-normal distribution in question into a normal and arguing that entropy is maximized subject to the appropriate constraints on the transformed variable. It's not clear how such a story would go, however, for a quantity such as human height, H—one would have to argue that the variance of $\ln(H)$ is fixed, that the distribution of $\ln(H)$ is expected to maximize entropy, and that there are no other constraints on $\ln(H)$.

Second, it is by no means clear what entropy is. There are number of interpretations of entropy, and it is not clear how they can play the appropriate explanatory role. Perhaps the most popular interpretation is that the entropy of a probability distribution is a measure of the information that an agent with that distribution has (Frank [2009] appears to have this interpretation in mind). This renders entropy as a property of a subjective probability distribution. But what does how much information an agent has got to do with

distributions of actual frequencies in nature? Something is conceptually amiss. As our goal is to explain an actual frequency distribution—for example, the approximately normal distribution of human male heights—it's not clear that this way of thinking about entropy and probability is satisfactory. Popper ([1982], p. 109) makes a similar point about the explanatory role of entropy in statistical mechanics.) Whatever the appropriate understanding of entropy is, it seems that it has to be a physical property and that it can be a property of a diverse range of physical things—from tensile strengths, to human heights, to bread weights. Moreover, the interpretation of entropy needs to make it the sort of thing that tends to be maximized. In the examples I gave earlier, I made an appeal to the tendency of things in nature to increase in disorder, i.e. the second law of thermodynamics. However, it is not at all clear that the entropy of thermodynamics is the same entropy that normal distributions maximize.[19] One reason to think that it might be, however, is that the Maxwell–Bolztmann velocity distribution law is the law that the velocities of particles of an ideal gas in a maximum entropy state are normally distributed. This is by no means a knockdown argument, but it is suggestive.

These issues require much more space than is available here. ME explanations may be a way forward, but much more work needs to be done. My goal is not to argue that ME explanations are the correct explanations for why particular distributions are normal, log-normal, and so on, or why normal distributions are common in nature. I only intend to point out that developing ME explanations in detail may be a promising alternative explanatory strategy.

5 Conclusion

I began this article with a famous remark regarding the different attitudes experimentalists and mathematicians have (or at least had) towards the normality of normal distributions. At the end of his chapter on the normal distribution, Cramer writes of this remark:

> It seems appropriate to comment that both parties are perfectly right, provided that their belief is not too absolute: mathematical proof tells us that, *under certain qualifying conditions*, we are justified in expecting a normal distribution, while statistical experience shows that, in fact, distributions are often *approximately normal*. (Cramer [1946], p. 232, emphasis in original)

I have argued that the 'certain qualifying conditions' are not satisfied in important cases—for example, in quantitative genetics. For other examples (e.g. tensile strengths), I've argued that we often have little reason for

[19] Thanks to Craig Callender for pointing this out to me.

assuming that the qualifying conditions are satisfied. I've also argued that distributions are 'often approximately normal' only if we carve up the world the right way with our variables. Any explanation for the normality of normal distributions should take this into account, and CLT explanations do not. Moreover, it appears that normal distributions are not as normal as we once thought they were. There are many documented cases of data initially thought to be normally distributed turning out to be log-normally distributed (Limpert *et al*. [2001]), or distributed in some other way (for example, Wilson and Hilferty [1929]).

All of this is not to say that CLT explanations are never true. To the contrary, the CLT probably explains why, for example, sample averages tend to be normally distributed. However, note that sample averages are clearly sums of factors, and are often designed so that those factors are as close to being independent and identically distributed as possible. My arguments are only intended to show that often our CLT explanations fail to be veridical—typically when the quantities in question are more naturally occurring (for example, human heights) and have a composition much more complicated than that of simple sample averages.[20]

Finally, I've suggested that a promising alternative way to explain why a particular quantity is normally distributed is to appeal to the maximum entropy property that normal distributions have. This, in turn, may generalize into an explanation for why normal distributions, and many other distributions, are common in nature.

Acknowledgements

Thanks to Marshall Abrams, Alan Baker, Frederic Brouchard, Rachael Brown, Brett Calcott, Craig Callender, Lindley Darden, Kenny Easwaran, Branden Fitelson, Alan Hájek, Philippe Huneman, Hanna Kokko, Roberta Millstein, Georges Rey, Neil Thomason, Lekki Wood, and two anonymous reviewers for this journal. Special thanks to Branden Fitelson for very detailed and insightful comments on earlier versions of this paper.

Department of Philosophy,
University of Maryland,
College Park, MD, USA,
alyon@umd.edu

[20] Incidentally, that some normal distributions are not explained by the CLT should not be news. For example, the position of a free particle governed by the Schrödinger equation is a normally distributed variable. Yet there are no factors that determine the particle's position (at least, according to 'no hidden variables' interpretations of quantum mechanics), let alone factors that are appropriately distributed, and so on

References

Baker, A. [2005]: 'Are There Genuine Mathematical Explanations of Physical Phenomena?', *Mind*, **114**, p. 223.

Baker, A. [2009]: 'Mathematical Explanation in Science', *The British Journal for the Philosophy of Science*, **60**, pp. 611–33.

Baker, A. [2012]: 'Science-Driven Mathematical Explanation', *Mind*, **121**, pp. 243–67.

Bening, V. and Korolev, V. [2002]: *Generalized Poisson Models and Their Applications in Insurance and Finance*, Volume 7, Zeist, Netherlands: VSP.

Cramer, H. [1946]: *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

de Finetti, B. [1990]: *Theory of Probability*, Volume 2, Chichester: John Wiley.

Denny, M. and Gaines, S. [2000]: *Chance in Biology: Using Probability to Explore Nature*, Princeton: Princeton University Press.

Feller, W. [1971]: *An Introduction to Probability Theory and its Applications*, Volume 2., New York: John Wiley.

Frank, S. A. [2009]: 'The Common Patterns of Nature', *Journal of Evolutionary Biology*, **22**, pp. 1563–85.

Galton, F. [1875]: 'Statistics by Intercomparison, with Remarks on the Law of Frequency of Error', *Philosophical Magazine*, **4**, pp. 33–46.

Gillespie, J. H. [1998]: *Population Genetics: A Concise Guide*, Baltimore, Maryland: John Hopkins University Press.

Goodman, N. [1955]: *Fact, Fiction, and Forecast*, Cambridge, MA: Harvard University Press.

Gregersen, E. (*ed.*) [2010]: *The Britannica Guide to Statistics and Probability*, New York: Rosen Education Service.

Grinstead, C. and Snell, J. [1997]: *Introduction to Probability*, Providence, Rhode Island: American Mathematical Society.

Hartl, D. L. and Clark, A. G. (*eds*) [1989]: *Principles of Population Genetics*, Sunderland: Sinauer Associates.

Hoeffding, W. and Robbins, H. [1948]: 'The Central Limit Theorem for Dependent Random Variables', *Duke Mathematical Journal*, **15**, pp. 773–80.

Jaynes, E. [1968]: 'Prior Probabilities', *IEEE Transactions on Systems Science and Cybernetics*, **4**, pp. 227–41.

Jaynes, E. [2003]: *Probability Theory: The Logic of Science*, Cambridge: Cambridge University Press.

Koch, A. [1969]: 'The Logarithm in Biology, II: Distributions Simulating the Log-Normal', *Journal of Theoretical Biology*, **23**, pp. 251–68.

Koenker, R. [2009]: 'The Median is the Message: Wilson and Hilfertys Experiments on the Law of Errors', *The American Statistician*, **63**, pp. 20–5.

Leon-Garcia, A. [2008]: *Probability, Statistics, and Random Processes for Electrical Engineering*, Upper Saddle River, New Jersey: Pearson Education Inc.

Limpert, E., Stahel, W. A. and Abbt, M. [2001]: 'Log-Normal Distributions across the Sciences: Keys and Clues', *BioScience*, **51**, pp. 341–52.

Lyon, A. [2012]: 'Mathematical Explanations of Empirical Facts, and Mathematical Realism', *Australasian Journal of Philosophy*, **90**, pp. 559–78.

Mandal, B. [2009]: *Global Encyclopaedia of Welfare Economics*, New Delhi: Global Vision Publishing House.

Mlodinow, L. [2008]: *The Drunkard's Walk*, New York: Pantheon Books.

Peirce, C. S. [1873]: *On the Theory of Errors of Observation*, Report of the Superintendent of the U.S. Coast Survey, pp. 200–24.

Popper, K. [1982]: *Quantum Theory and the Schism in Physics*, NJ: Rowman and Littlefield.

Roush, M. and Webb, W. [2000]: *Applied Reliability Engineering*, University of Maryland, College Park: The Centre for Reliability Engineering.

Stigler, S. [1986]: *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard: Harvard University Press.

van Fraassen, B. [1989]: *Laws and Symmetry*, Oxford: Oxford University Press.

Visscher, P., Medland, S., Ferreira, M., Morley, K., Zhu, G., Cornes, B., Montgomery, G. and Martin, N. [2006]: 'Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings', *PLoS Genetics*, **2**, p. e41.

Wilson, E. and Hilferty, M. [1929]: 'Note on C.S. Peirce's Experimental Discussion of the Law of Errors', *Proceedings of the National Academy of Sciences of the United States of America*, **15**, p. 120.

Zhao, W., Serpedin, E. and Dougherty, E. [2008]: 'Inferring Connectivity of Genetic Regulatory Networks using Information-Theoretic Criteria', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **5**, pp. 262–74.