On Moments

Neil J. Hatfield
School of Mathematical Sciences,
University of Northern Colorado

Revised May 2019

This reading is meant to provide you with additional information regarding the role that moments play in the idea of distribution. This reading will have a brief refreshing (or intro) to a few ideas from Calculus that I'll make use of later on in the readings.

0 A Quick (P)Review of Calculus (You may skip this section)

Before we dig into the idea of moments in Statistics, we do need to recap[1] a couple of key ideas from Calculus.

First and foremost, Calculus is the Mathematics of Change. This is to say, that Calculus is the branch of mathematics that deals exclusively with studying how quantities' values vary and what those changes mean. Essentially, all problems in Calculus can be reduced to one of the following two fundamental problems:

1. How can we find out (describe) how fast a quantity is changing given we know how much of the quantity we have at any moment in time? (Answer: rates-of-change and derivatives)

2. How can we find out (describe) how much of a quantity we have given we know something about how the quantity is changing? (Answer: accumulation and integrals)

While some individuals might say that there's more to Calculus than these two questions, I would agree. The additional pieces to Calculus are all centered on interpreting the answers/consequences to those two questions.

For our present purposes, we will restrict our attention to the idea of accumulation, since distributions are about indefinite accumulation. Consider the following mathematical expression

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

[1] Or preview for those readers who have yet to take Calculus

As is true for every mathematical expression, there is a lot of meaning wrapped up in each part of the expression. To make sense of the expression, you have to keep in mind what each component means and the relationships between the components. Let's examine what each part of this expression means:

- F: under Euler's Function Notation conventions, this is the name of a special kind of function–an accumulation function. Accumulation functions tell us the total change in some quantity's value, but not the actual value of the quantity. (If we know an initial amount, say C, we can add that to the rule of F and transform the accumulation function into an amount function.)
  - In our course, we don't directly use accumulation functions, rather we use an amount function.
  - In particular we use the amount functions of the form $C_D(x|\theta) = F(x|\theta)+0$.
  - Yes, I realize that adding zero does not change the numeric value, but from a conceptual standpoint that "+0" moves us from talking about a change that could occur anywhere on the number line, to a specific change in location on the number line. Think of the "+0" as allowing us to change our conversation from a distance we're traveling to an actual location.
- x: This is a realization of the [previously defined in the problem] stochastic variable X. Ultimately, this will be a value that we identify in the problem to act as the upper limit of accumulation–a bound on our data event.
- $F(x)$ is the accumulation of changes up to the value x.
- $\int$: The integral symbol is really just a stylized "S" meaning "sum". The intent behind this symbol is for the reader to imagine infinitely many products that are then combined together. ("Integrate" literally means "to put together".)
- $-\infty$ and x: These are the "limits of integration"; the value below the integral symbol indicates where you will begin accumulating from while the value above indicates where you'll stop the accumulation.
- $f(t)dt$: This is the general form of the products that we're putting together. Each instance of the product represents a change in the dependent quantity in the co-variational relationship we're studying. In Statistics, this product is a change in the long-run relative frequency of accumulated outcomes.
- $f(t)$: This is the instantaneous-rate-of-change at the value t. In general, this value tells us how many times the change in the dependent quantity is as large as the corresponding change in the independent quantity. In Statistics, we will interpret $f(t)$ as telling us how many times the change in probability is as large as the corresponding change in the value of the stochastic variable from the value t (where we're thinking of any small change from t).
- dt: This represents an extremely small change in the value of the stochastic variable; infinitesimally small.

- t: While our stochastic variable is X, our realization (x) is used in the upper limit of accumulation. As is customary in Calculus, to help keep our thinking straight we alter the symbol used inside the product. Often we use t; t still represents all of the values of the stochastic variable...just the ones between the lower and upper limits of accumulation.

1 Specific Long-Run Behaviors

Distributions are the long-run behavior of a stochastic variable. We often think of distributions as being "complete" in that they contain every nuance and tiniest bit of detail. However, this can be overwhelming for us. There are times when instead of wanting to see the whole forest at once, we just want to look at particular tree. If the distribution is forest, then each individual tree is a particular behavior. These particular behaviors allow us to better understand the distribution.

In class we mentioned the particular behaviors referred to as the Distribution Median, Distribution Minimum, Distribution Maximum, and Distribution Mode. All four of these represent particular values of the stochastic variable that tell us key information about the stochastic variable's distribution.

- The Distribution Median tells us the value of the stochastic variable for which 50% of the infinitely many outcomes will be less than or equal to.
- The Distribution Minimum tells us the value of the stochastic variable that no other value may be smaller than. (Also called the Absolute Minimum.)
- The Distribution Maximum tells us the value of the stochastic variable that no other value may be greater than. (Also called the Absolute Maximum.)
- The Distribution Mode tells us the value of the stochastic variable that has the largest instantaneous-rate-of-change; the value of the stochastic variable that coincides with the largest increase in long-run relative frequency–potentially the most frequently occurring value.

The Distribution Minimum and Maximum stem from the domain of the stochastic variable and act as lower and upper bounds on the values the underlying stochastic process can produce. The Distribution Median comes from thinking backwards through the distribution function (i.e., start with an output of 0.5 from the distribution function and find the input value that gives 0.5 as the output). The Distribution Mode comes directly from the distribution function (and an application of calculus), but is not guaranteed to exist.

All these four specific behaviors are useful and they all have a similarity; they either deal with the domain of the stochastic variable or they deal with the distribution function, but not both. By bringing the two (domain and distribution function) together, allows us to get a better picture of the long-run behavior of the stochastic variable. This is where the concept of a "moment" in Statistics comes into play.

2 Moments

In Calculus and in Everyday Language, we often use "moment" to refer to a small interval of time. However, this is not always how this word is used in other fields. To better understand a "moment" in Statistics, we'll first look at the role of "moment" in Physics.

2.1 Examples from Physics

In Physics, a moment is a combination of a physical quantity (e.g., force, charge) and a distance. For example, torque (the first moment of force) is combination of force and rotational distance; thus torque measures the rate-of-change of an object's angular momentum. You can replace force with other quantities such as charge and mass to construct other moments. For example, the zeroth moment of mass is the total mass of an object while the first moment of mass tells you the center of mass for that object.

There are two formulas that are used to calculate a moment in Physics:

- The nth Moment at Single Point: $\mu_n = r^n Q$, where r is the distance from the reference point to the point under study and Q is the measure of the physical quantity (i.e., force, charge, mass, etc.).
- The nth Moment over Space: $\mu_n = \int r^n \rho(r) dr$, where r is the distance you're currently at in space from the reference point (a location), $\rho(r)$ is the measure of the physical quantity at that location, and dr is a small change in your distance.

The first is great if you want to talk about something like a molecule. The second works well if you want to talk about the entirety of an object. As an example, suppose you have a steel rod that has a tiny nick on one end's edge. If you only wanted to talk about a moment for the deepest part of the nick, the first formula would potentially work. However, if you wanted to talk about the steel rod as a whole, then the second formula would be what you would want to use.

2.2 Statistical Moments

In Statistics, we use a similar idea of combining two measures together to form a moment. Since we're concerned with distributions (i.e., long-run behavior) we do not use the "moment at a single point" idea. Rather, we want to bring the entirety of the stochastic variable (that is, the whole domain) into play. Thus, when we talk about a moment in Statistics, we will make use integral-based formula.

In Physics, we combined a physical quantity (Q or $\rho(r)$) with a distance (r); to adapt this for our purposes we will replace distance with the value of the stochastic variable and for the physical quantity, we'll use the rate-of-change of probability (LRRF) at the value of the stochastic variable. This gives rise to two new formulas:

- The nth "Raw" Moment: $\mu'_n = \int x^n f(x) dx$, where x is a realization of X, $f(x)$ is the rate-of-change of probability at x, and dx is the small change in the value of the stochastic variable.

- The nth "Central Moment: $\mu_n = \int (x - \mathrm{E}[X])^n f(x)dx$, where $\mathrm{E}[X]$ is the first (raw) moment, $x - \mathrm{E}[X]$ is the "centered" value of the stochastic variable, $f(x)$ is the rate-of-change of probability at x, and dx is the small change in the value of the stochastic variable.

While we can calculate many different moments using either formula, we typically only use the first formula (i.e., "raw" moments) to find the Zeroth and First Moments. We then prefer to use the "central" moment formula for the others.

2.3 The Use of the Moments

The Moments are specific behaviors that let us glean additional insight into the long-run behavior of our stochastic variable (i.e., the distribution). Unlike the specific behaviors previously mentioned (i.e., the Distribution Median, Distribution Minimum, Distribution Maximum, and Distribution Mode), moments blend together the stochastic variable's domain and the distribution function together to allow us this better insight. In other words, moments allow us to identify particular features of how the indefinite accumulation is happening.

One of the real powers of moments is that they allow us to identify a stochastic variable's distribution. When we speak of "the normal distribution", we are being a bit misleading with our phrasing. There is no such thing as "*the* normal distribution", rather there is a family (a collection) of infinitely many distributions that have common features but differ slightly. Think of the name of a "named distribution" as being the last name (surname, family name). Now, imagine yourself showing up to a family reunion and calling out the last name. Which single family member do you want to talk to? The moments give us a way (through their role as being parameter values) to find the right member of the family.

3 Five Essential Moments

While we can find as many moments as we would like, there are five that we use most often. These are the Zeroth Moment, the First Moment, the Second Central Moment, the Third Central Moment, and the Fourth Central Moment.

While I have included the various formulas (based in calculus) for these five moments, you **do not need to memorize** these formulas. If you are in an Introductory Statistics course, you also do not need stress about trying to perfectly understand these formulas.

3.1 The Zeroth Moment: Total Long-Run Relative Frequency, Total Probability

The Zeroth Moment is the one moment that is guaranteed to exist for all distributions. The Zeroth Moment measures the behavior of a stochastic process to actually accumulate *everything*. That is to say, that if we imagine running the stochastic process forever, we will see all possible values of the stochastic variable occur at one point in time or another. Mathematically, the Zeroth Moment (μ'_0) is found through the expression

$$\mu'_0 = \int x^0 f(x)dx = \int f(x)dx = 1$$

If we accumulate all of the changes in long-run relative frequency (i.e., we accumulate all possible outcomes), then we will have a total accumulation of 1 or 100%. The fact that the Zeroth Moment measures the total accumulation of probability is what ensures that all distributions have this moment. In order to be a true distribution, we must be able to see all possible outcomes when we repeat the stochastic process forever.

3.2 The First Moment: Expected Value, $\mathbf{E}[X]$

The First Moment (a.k.a., the first "raw" moment or the first moment centered around zero) generally goes by the name "Expected Value [of the stochastic variable]" (and sometimes "*the* mean"). The First Moment measures the specific behavior of what value the stochastic process appears to tend towards in the long-run. That is to say, the first moment measures the value of the stochastic variable that we *expect* to see on any trial of the stochastic process. Since we're keeping in mind the entirety of the long-run behavior, we will only have one Expected Value (if one exists) for the stochastic variable.

The Expected Value is generally thought about as being a value of the stochastic variable, but does not necessarily have to be part of the variable's domain. This is particularly true in the case of discrete stochastic variables. For example, if we let Y represent the number of pips showing on the upper face of a standard, fair, six-sided die, then $\mathrm{E}[Y] = 3.5$. The domain of Y is the set $\{1, 2, 3, 4, 5, 6\}$ which does not include 3.5. If we are asked to predict[2] the next outcome of the stochastic process (i.e., a short-run), we should use the Expected Value. For the die roll, even though 3.5 is not part of the domain, we should still use 3.5 as our prediction[3].

We can calculate the Expected Value through the mathematical expression

$$\mathrm{E}[X] = \mu'_1 = \int x^1 f(x)dx$$

where x is a realization of the stochastic variable X, $f(x)$ is the instantaneous-rate-of-change of probability at x, and dx is a small change in the value of the stochastic variable. The Expected Value can be thought of through an analogy to the idea of the mass of an object. Imagine a steel rod that has a total mass of 100 grams but the rod isn't necessary the same thickness from end to end. If we wanted to find the place along the rod to set a fulcrum and achieve perfect balance of the rod, we would seek to calculate the first moment of mass (a.k.a. the "center of mass") by looking at the distance along the rod from one end and the density of the steel at each distance. Density is a rate-of-change (of mass with respect to volume) and we

[2] Recall that prediction entails the anticipation that we're wrong.

[3] In practice, we usually "un-split the difference" and guess 3 or 4.

can think the distance along the rod as the ordered values of the stochastic variable's domain. This makes the Expected Value analogous to the center of mass and is one of the central ways to think about the First Moment as a "Measure of Center"[4].

There are some additional things to be aware of with the Expected Value. First and foremost, the Expected Value is not necessarily the most frequently occurring value in the long-run. There are some cases where the Expected Value is the more frequently occurring value, but there are other reasons for this to be happening. For example, if the Distribution Mode is the same value as the Expected Value, then, yes, that value will be the most frequently occurring; but if the Distribution Mode is a different value, then the Expected Value will not be the most frequently occurring value. Second, only when the distribution is symmetric (i.e., the graph of the rate-of-change of probability is symmetric) will the Expected Value mark out the axis of symmetry (i.e., when the Distribution Median and the Expected Value are the same value). Third, the Expected Value does not exist for all distributions. For example, the Cauchy distribution is a distribution that does not have an Expected Value. If you were to carry out the calculus to find the Expected Value for a Cauchy distribution, rather than getting a number, you end up with an expression that has no limit (i.e., increases without bound, ∞, in the long-run). In this case, there is no Expected Value. There are other cases where what you're studying results in division by zero in the calculation for the Expected Value; is results in the absence of the Expected Value.

The Expected Value is often used as a Location Parameter that helps us identify which member of the family of distributions applies to the current situation.

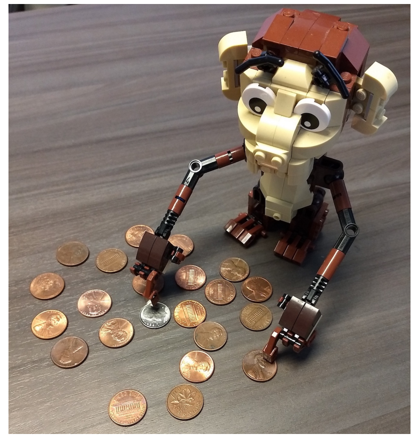
3.3 The Second Central Moment: Variance, Var$[X]$

The Second Central Moment is the first moment where we give ourselves a particular reference point when we calculate our "distances". In this case we use the Expected Value as our reference point. The Second Central Moment measures the dispersion of outcomes from the Expected Value. Thinking back to our three types of variation, this would be an attempt to measure the Variation Between Individuals/Variation Within a Collection. Our collection is the infinitely many outcomes of the stochastic process and rather than looking between each and every individual, we'll look back and forth between the Expected Value and each outcome. To help you think about this, we'll turn to Lego® Monkey and coins.

In Figure 1, the dime will play the role of the Expected Value and all of the pennies play the role of outcomes from our stochastic process. In Figure 1(a), all of the pennies are in the same place as the dime (don't worry about vertical displacement or that the tower isn't perfectly straight). The monkey essentially doesn't have to shift his attention at all from the dime to the pennies. This would be analogous to all of the outcomes being exactly the same as the Expected Value. In Figure 1(b), each of the pennies is fairly close to the dime; that is, the outcomes of the stochastic process are close to the Expected Value. The monkey has to shift his attention from

[4]For those individuals who can unpack the phrase "Measure of Center", this is analogy that they make.

(a) No Variance



(b) Low Variance

(c) High Variance

Figure 1: Lego® Monkey explores Variance

the dime (Expected Value) to each penny (outcome) but he does not have to shift far. In Figure 1(c), some of the pennies are much further from the dime; this would be like an outcome being further away from the Expected Value. The monkey has to "move"/"shift" his attention much more from the dime to different pennies and back. When we look at the Second Central Moment, or the Variance, we are getting a measure of the total amount of shifting our attention from the Expected Value to any outcome and back these will be for the infinitely many outcomes. Mathematically, this gives us the formula

$$\mathrm{Var}[X] = \sigma^2 = \mu_2 = \int (x - \mathrm{E}[X])^2 f(x)dx$$

where $x - \mathrm{E}[X]$ represents the displacement from the Expected Value–the distance from the Expected Value that each outcome is, $f(x)$ is the instantaneous-rate-of-change of probability at x, and dx is a small change in the value of the stochastic variable.

Variance is on a scale that has an absolute zero: there is no admissible value less than zero (i.e., no negatives). When $\mathrm{Var}[X] = 0$, we know that there is absolute no variation between the outcomes and the Expected Value (i.e., no dispersion) and thus, every value is the same. There is no upper limit to the scale that Variance is measured by; you can have Variance values of 0.5, 1, 15, 2543, or even 1.8 trillion. You can even end up with $\mathrm{Var}[Y] = \infty$ (that is, increasing without bound). The larger the value of Variance, the further you have to shift your attention between the Expected Value and each of the infinitely many outcomes. However, try not to get caught in the trap that just because you have a large Variance means you can't have outcomes close to the Expected Value. In Figure 1(c), there are a couple of pennies that are still close to the dime.

The Second Central Moment is often used as a Scale Parameter; again, helping us to pick which exact member of the distribution family works for our current situation. Depending on which named distribution we're looking at, the Variance can provide us with additional information such as the location of inflection points for the graph of the rate-of-change of probability function[5].

3.4 The Third Central Moment: Skewness, Skew$[X]$

The Third Central Moment (more formally, the Third Central-Standardized Moment) goes by the more common name of Skewness. Skewness is measure of how lopsided or asymmetrical a distribution is. There is an important piece to keep in mind when you think about Skewness: symmetry is often a visual aspect so you must be sure that you're looking at the right visualization when checking for Skewness. When we talk about a distribution's Skewness, we often examine at least one of the following:

- the graph of the rate-of-change of probability function (a.k.a. probability density function–PDF) for the distribution
- Data Visualizations using the sequence of accumulated values
  - Histograms or Bar charts
  - Box plots (Box and Whisker Plots)

We do not often look at the graph of the distribution function, as these functions do not have graphs that highlight the symmetry/asymmetry that Skewness measures.

The concept of Skewness can transcend the canvas of data visualizations. We escape the world of pictures by focusing on what the underlying stochastic process is doing. Skewness is providing a measure of the direction that the process generates outcomes that lie away from the main bulk of the outcomes. In this sense, Skewness can be thought of a measure of which direction we should look for potential outliers.

The mathematical formula for getting at Skewness is

$$\text{Skew}[X] = \gamma_1 = \frac{\mu_3}{\left(\sqrt[2]{\sigma^2}\right)^3} = \int \left(\frac{x - \text{E}[X]}{\sqrt[2]{\text{Var}[X]}}\right)^3 f(x)dx$$

where $\frac{x-\text{E}[X]}{\sqrt[2]{\text{Var[X]}}}$ is the standardized distance from the Expected Value for each value of the stochastic variable, $f(x)$ still represents the instantaneous-rate-of-change of probability at x, and dx is still a small change in the value of the stochastic variable.

[5]The normal (Gaussian) distribution is one such family where this applies.

Skewness is measured on the Real line, with us having particular interpretations of negative, zero, and positive values. Generally speaking, the further a value is away from zero, the more asymmetrical the graph of the PDF function, the histogram/bar chart, or box plot will look.

- **Negative Skewness** When $\text{Skew}[X] < 0$
  - **Non-Image Based:** The stochastic process generates some values that are smaller in magnitude than the main bulk of the outcomes. While process can generate values both smaller and larger than the Expected Value, those that are larger lie closer to the Expected Value, making a modal clump. There are values that are smaller than the Expected Value and which are further away from the Expected Value; there might not be very many, but there are at least a few. We might have potential outlier individuals who have a value much smaller than the Expected Value.
  - **Image Based:** The "tail" (i.e., the side of the PDF plot, histogram, box plot) that is associated with values that are less than the Expected Value is much longer than the other "tail". This tells us that there are values of the stochastic variable that are much smaller than the Expected Value versus values larger. If you imagine a number line where the Expected Value is at zero, the negative side of the number line will go on much longer than the positive side of the number will.
- **Zero Skewness** When $\text{Skew}[X] = 0$
  - **Non-Image Based:** The stochastic process generates value that are smaller than the main bulk outcomes, larger than the main bulk, and in the main bulk of outcomes. In essence, the process can produce values that are just as far away from the Expected Value in both directions rather than producing further away values in only one direction. We might have potential outlier individuals who have a value much smaller than the Expected Value AND we may have potential outlier individuals who have a value much larger than the Expected Value.
  - *Most Common Image Case*: the two "tails" are of the same length and the visualization is symmetric along the Expected Value. This would tell us that there are values of the stochastic variable that are just as far below the Expected Value as there are values above. Returning to the number line idea, the negative and positive sides of the number line go on the same amount of distance.
  - *Extremely Rare/Special Cases*: while not symmetric, the two "tails" are in such a way that they end up canceling each other out. (This is a result of the mathematical formula applied to some very interesting situations.)
- **Positive Skewness** When $\text{Skew}[X] > 0$
  - **Non-Image Based:** The stochastic process generates some values that are larger in magnitude than the main bulk of the outcomes. While process

can generate values both smaller and larger than the Expected Value, those that are smaller lie closer to the Expected Value, making a modal clump. There are values that are larger than the Expected Value and which are further away from the Expected Value; there might not be very many, but there are at least a few. We might have potential outlier individuals who have a value much larger than the Expected Value.

- **Image Based:** The "tail" (i.e., the side of the PDF plot, histogram, box plot) that is associated with values that are greater than the Expected Value is much longer than the other "tail". This tells us that there are values of the stochastic variable that are much larger than the Expected Value versus values smaller. If you imagine a number line where the Expected Value is at zero, the positive side of the number line will go on much longer than the negative side of the number will.

Some individuals refer to "left-" or "right-" skewness. These terms make sense when you focus on visualizations where the stochastic variable is on the horizontal axis, with values to the left being smaller than values to the right (i.e., $a < b < c < \ldots$). In such a situation "left" will mean "negative" and "right" will mean "positive". However, if you alter any part of this arrangement (e.g., use a vertical orientation or flip the ordering so that larger values are the left (i.e., $a > b > c > \ldots$), then "left" and "right" as typically used no longer convey any useful meaning.

The Third Central Moment, Skewness, can assist in identifying which member of the distribution family we want to reference through the role of a Shape Parameter.

3.5 The Fourth Central Moment: Kurtosis, Kurt$[X]$

The Fourth Central Moment (more formally, the Fourth Central-Standardized Moment) goes by Kurtosis and focuses on measuring the propensity (tendency) of the distribution to have extreme values; that is, for the stochastic process to produce objects/living beings that are outliers. While Kurtosis is measured on the Real line, we do not have a nice interpretation for the values. For example, we don't have an "absolute zero"; that is to say, that there is no value for Kurtosis which means "there are no outliers generated". To account for this, what Statisticians have done is find a value with which to peg the rest of the scale to, much like we do with the Celsius and Fahrenheit temperature scales. For any normal distribution, we know that the value of Kurtosis is 3. Thus, we end up judging the generation of outliers by what we would expect to see from a normal (Gaussian) distribution. Due to this comparison, many computer programs will report "Excess Kurtosis" rather than Kurtosis. Excess Kurtosis is simply $\text{Kurt}[X] - 3$. The mathematical formula that we use to find Kurtosis is

$$\text{Kurt}[X] = \frac{\mu_4}{\left(\sqrt[2]{\text{Var}[X]}\right)^4} = \int \left(\frac{x - \text{E}[X]}{\sqrt[2]{\text{Var}[X]}}\right)^4 f(x)dx$$

where $\frac{x-\text{E}[X]}{\sqrt[2]{\text{Var[X]}}}$ is the standardized distance from the Expected Value for each value of the stochastic variable, $f(x)$ still represents the instantaneous-rate-of-change of probability at x, and dx is still a small change in the value of the stochastic variable.

We interpret values of Kurtosis in the following way:

- **Platykurtic:** When Kurtosis is less than three ($\text{Kurt}[X] < 3$), then we have a distribution that has fewer extreme values (i.e., there are fewer outliers generated by the stochastic process) than what we would expect to see for a normal distribution. For the outliers that do exist, their values are typically not as far from the rest of the values as we might see in the normal distribution (i.e., the extremes aren't as "extreme" as they could be).

- **Mesokurtic:** When Kurtosis is three ($\text{Kurt}[X] = 3$), we have a distribution that generates extreme values at about the same rate as we would expect to happen when the long-run behavior matches that of the normal distribution.

- **Leptokurtic:** When Kurtosis is greater than three ($\text{Kurt}[X] > 3$), then we have a distribution that has more extreme values (i.e., there more outliers generated by the stochastic process) than what we would expected to see in situations where the long-run behavior matches that of a normal (Gaussian) distribution. The extreme values that we observe tend to be even further away from the rest of the values than for a normal distribution (i.e., the extremes are even more "extreme").

Kurtosis can assist in identifying which member of the distribution family we want to reference through the role of a Shape Parameter.