

## Why Symmetry?

For a few pages I would like to step aside from the direct business of analyzing data to address the question, "Why symmetry?" Well behaved variables are the key to data analysis that is likely to pay off, as compared to data analysis that churns the numbers a bit without much hope of getting beyond the obvious of means and variation. More sophisticated statistical techniques often condition their results on a premise that the data were well behaved to begin with. Picking the first criterion, what look for symmetry?

Here I will offer two answers to that question. It is abundantly clear that almost all data will show variation. But why? Two of the possible reasons for variation are error and complexity. I'm going to show you two analogies, one for error, one for complexity. And you will note that each of them leads me to expect symmetry. So beginning with error: Why symmetry?

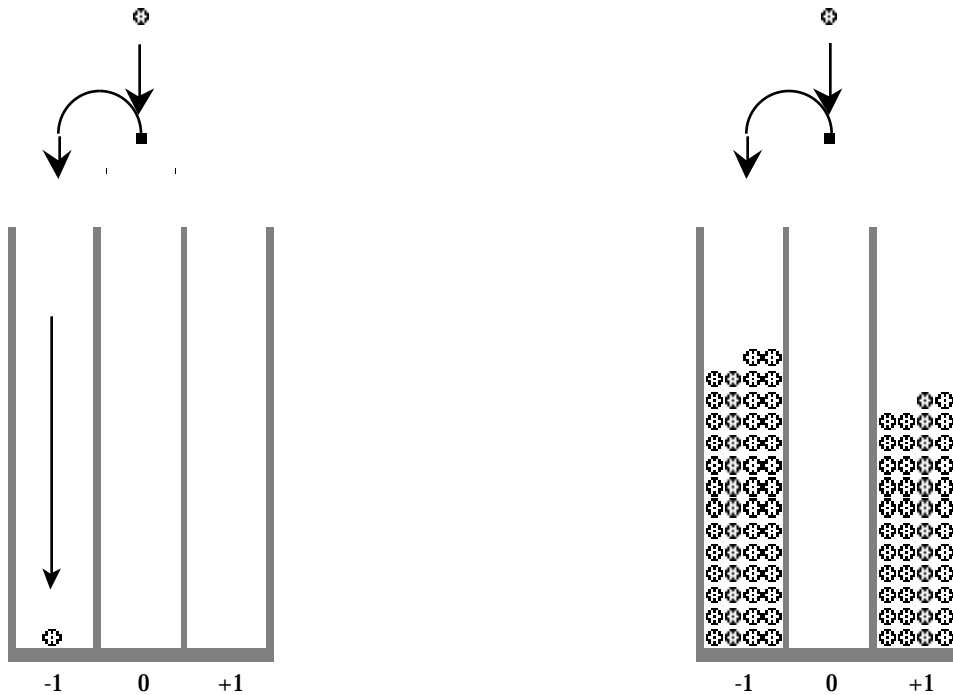
### **Error (The GaltonBoard)**

I want you to consider a mechanical model for measurement error: Even the simple process of determining the weight of an object requires a process. And that leads to error. Objects have to be weighed. If the scale is a balance beam then the balance beam has counter weights, and the counter weights themselves have to be measured. A balance beam has a pivot, and the pivot has to be perfectly shaped, which it never

is. And the pivot has to be perfectly placed, which it never is. We have to be careful that there is no dirt in the scale, no dust in the pans, and so forth — scales make errors. If a scale has springs in it, then springs have slight irregularities. They are influenced by temperature, and corrosion, not to speak of the fact that we have to know a few laws or a few practical tricks for translating the length of the springs in the scale into numbers that describe the weight of objects.

So measurement introduces errors. The interesting question is not whether or not measurements will include error — the question is, what will the distribution of estimates look like, *including* the error? To answer that question I need a hypothesis. The usual hypothesis is to suppose that error, however it is introduced, is unbiased: The error has a 50/50 chance of increasing the apparent value by some amount and an equal chance of decreasing it by the same amount.

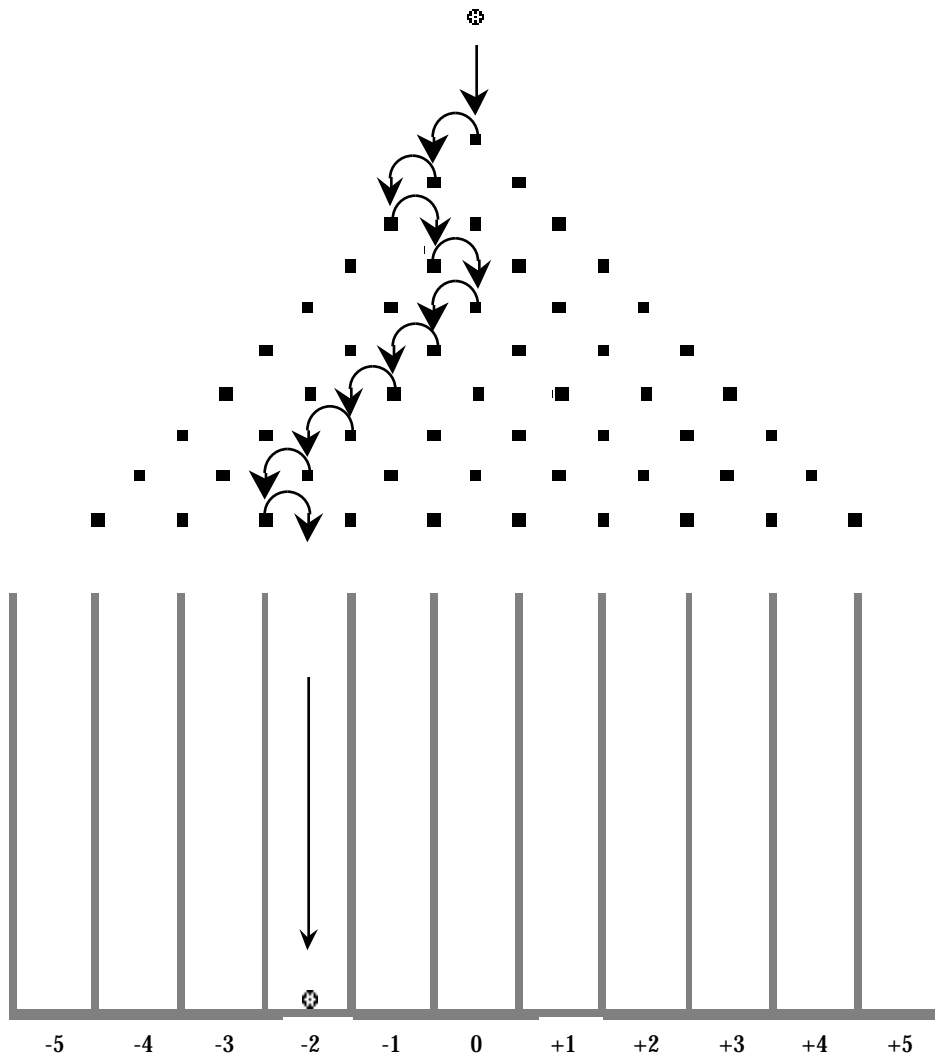
We can duplicate this hypothesis mechanically by imaging that we drop a ball in the direction of a slot that represents the right answer. But, before the ball can fall into the slot it encounters an obstacle that gives the ball a bounce, displacing the ball to the left or to the right of the direction that represents the right answer.



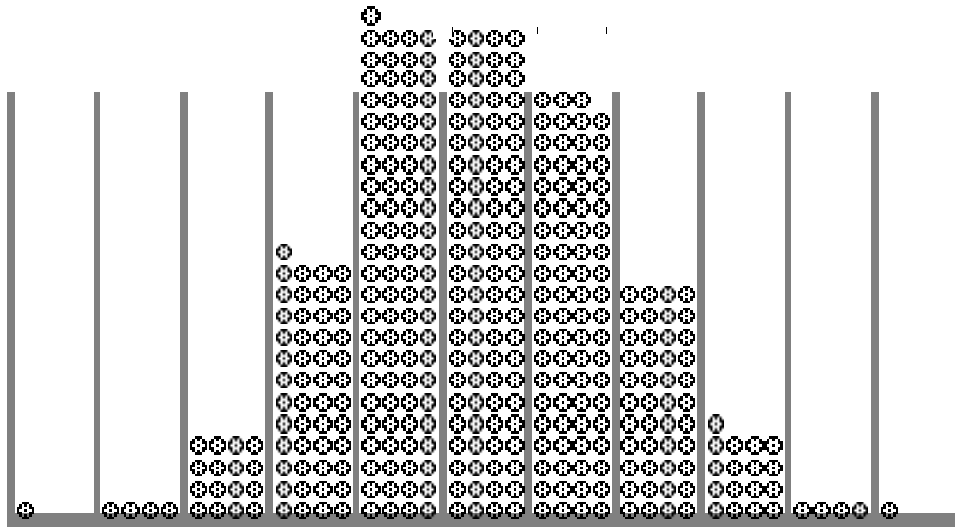
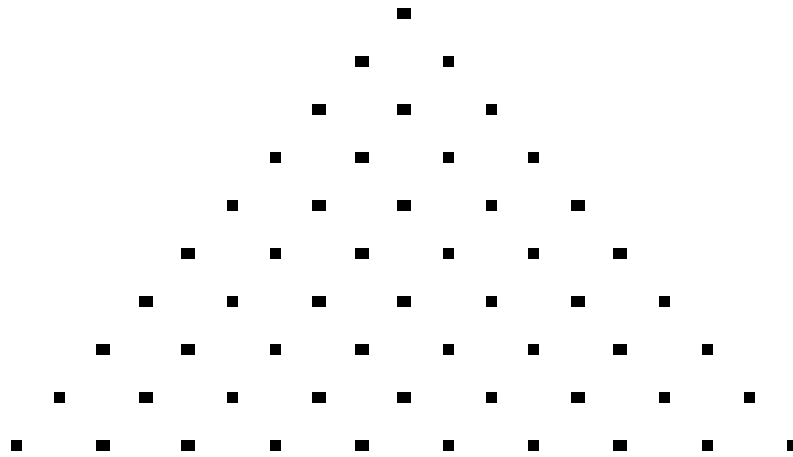
If we make several estimates of the quantity, then the process will repeat itself, making errors, some of the estimates will be less than the correct value, some will be larger than the correct value, and the result will be a distribution of estimates that is “more or less” symmetrical with respect to the correct value — “more or less” because the estimates will tend toward a fifty/fifty split but may not come out exactly fifty/fifty.

The hypothesis further states that a measurement process may include not one but several small errors, each of which has the same effect on the estimate — deflecting the estimate in a direction that makes it high, or deflecting the estimate in a direction which would make it low by the same amount. I can duplicate this process

mechanically by imagining a sequence of obstacles standing in the way of my bouncing ball, each one displacing the estimate, each one making the value a little smaller or a little larger.



And if we make several estimates of the quantity, the process will repeat its self, some of the estimates will be small, some will be large, and the result will approximate a symmetrical bell shaped distribution (called the “binomial distribution”).



	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
Count Equal:	1	4	16	49	93	92	79	44	17	4	1
	$n = 400$										

This is the standard hypothesis for errors introduced in the process of measurement. The exact specification of the model might be carried further — how large are the “bounces”, what is the width of the slots, how many layers of obstacles are there, and so forth. But for most purposes we are more interested in the moral of this story. The moral of the story is that if the variation of values is due to unbiased measurement error, then the distribution of values should be symmetrical and bell shaped.

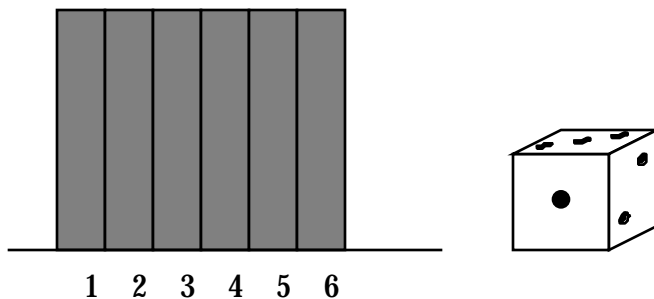
As is frequent in data analysis, the application of this principle is requires that we use it backward: When your data is *not* symmetrical and bell shaped, then you can *not* explain the variation as noise. When the data is not symmetrical and bell shaped, you’ve got some work to do to explain why not.

rewrite to match 25b

## Shape & Number II: Complex Processes

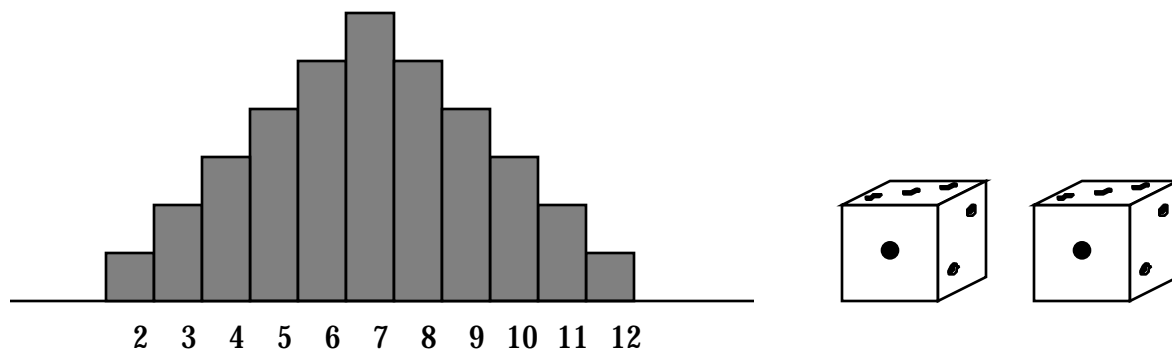
Starting again with the observation that there is something very important about the shape of a distribution, let me introduce another very powerful hypothesis that explains why you *should* expect bell shaped distributions, and why, therefore, it is interesting when that is not what you get. When it is presented mathematically, the principle at work here is known as the Central Limit Theorem, a real mathematical theorem here in the heart of data analysis. What it says, when it is interpreted is that complicated processes will tend to produce bell shaped results.

The central limit theorem is built on the difference between complex events and simple events. Suppose our simple event is analogous to the “process” of throwing a single six-sided die many times. If I threw a single die 6 times or, better, 600 times, I would expect to get something close to an even result, something close to



**Theoretical Shape of distribution: One Die, Six Possible Outcomes, 1 through 6, Equal Probabilities**

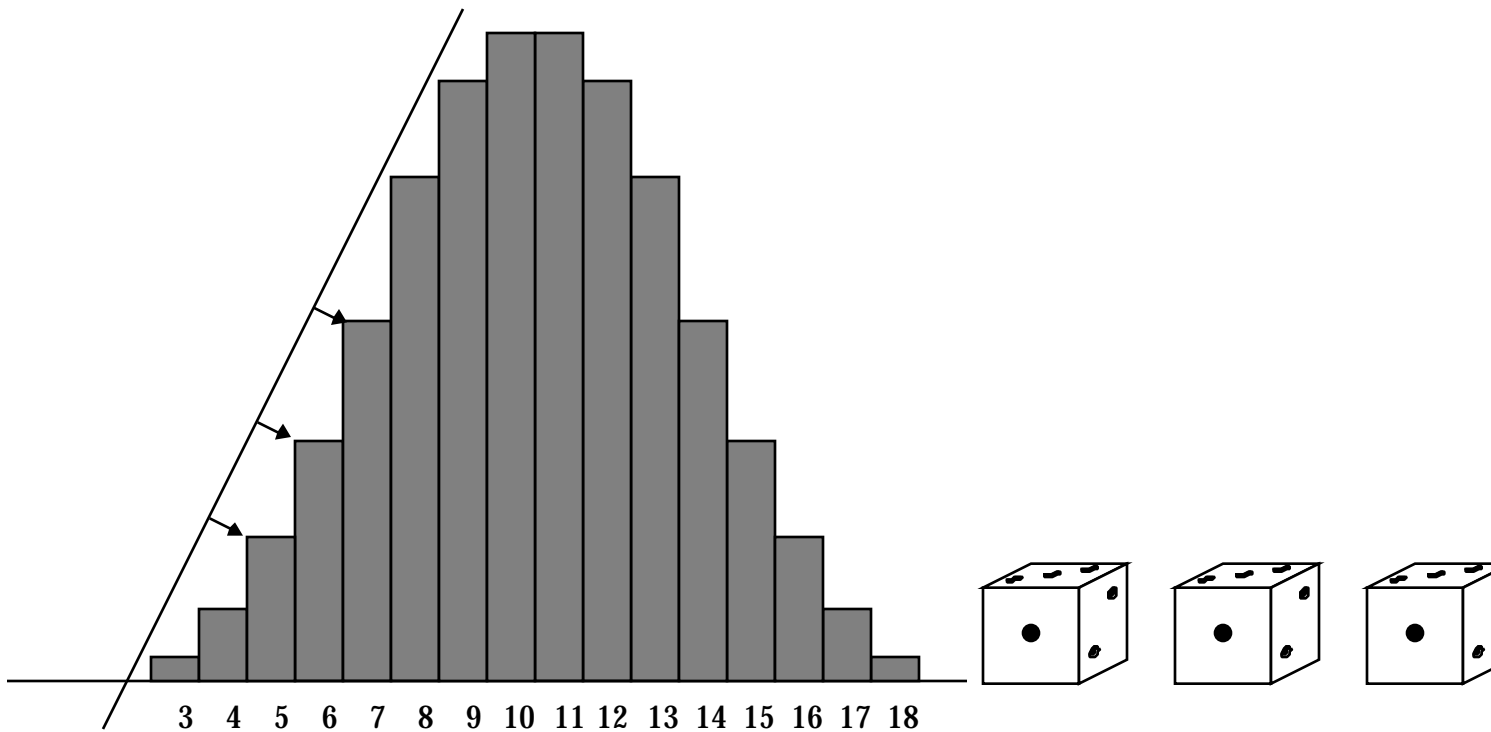
That's my simple event, and there is the distribution of its results, symmetrical to be sure, but not bell shaped. Now suppose I look at the result of a complicated event, the result of throwing two dice and recording their sum. As is well known, with two dice the results will be uneven: some numbers will be common, others will be rare, specifically, sevens will be relatively common while twos and twelves will be relatively rare. The reason for this is straightforward: There are thirty-six ways that the pair of dice can land and among these six of the thirty-six possibilities will add up to seven (while only one of the thirty six possibilities adds up to two and only one of the thirty six possibilities adds up to twelve). There is 1 way to get a 2, there are 2 ways to get a 3, 3 to get a 4, 4 to get a 5, 5 to get a 6, 6 to get a 7, 5 to get an 8, 4 to get a 9, 3 to get a 10, 2 to get an 11, and 1 to get a 12. So, in 360 throws of the dice or 3,600 throws of the dice I expect a distribution of results something like



**Theoretical Shape of distribution: Sum of Two Dice, Eleven Possible Outcomes, 2 through 12, "Triangular" Probabilities**

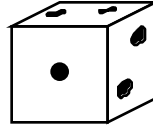
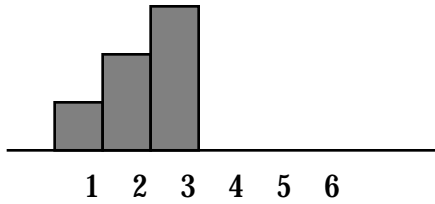
Because most people are familiar with dice there is nothing startling about this behavior of the dice. But it is actually quite remarkable in one way: Specifically, the shape for the composite is not the shape of the things of which it is composed: One die has a flat distribution. But the two dice, together, have a "triangular" distribution.

And with three dice you'll see a little flaring out, changing shape again toward what is called a bell-shaped or "normal" or "Gaussian" distribution.



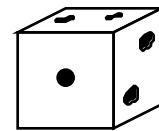
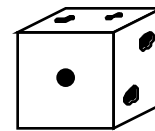
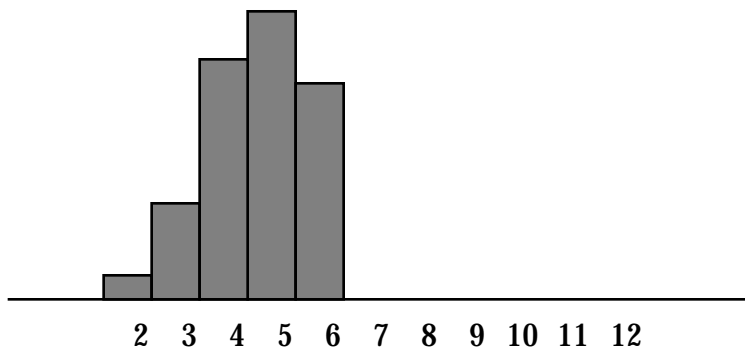
**Theoretical Shape of distribution: Sum of Three Dice, Eleven Possible Outcomes, 3 through 18, "Bell-Shaped" Probabilities (Rounded Concave Sides, Rounded Convex Peak)**

Dice are so familiar that this little demonstration may fail to convince, so let me try something more extreme. I'm going to change the simple event by taking a marker pen to my set of dice and writing my own numbers on their faces. I'm going to use one 1, two 2's, and three 3's. So for 6 or six hundred throws of this simple altered die I expect to get something that is neither symmetrical, nor bell shaped. For one die I should get



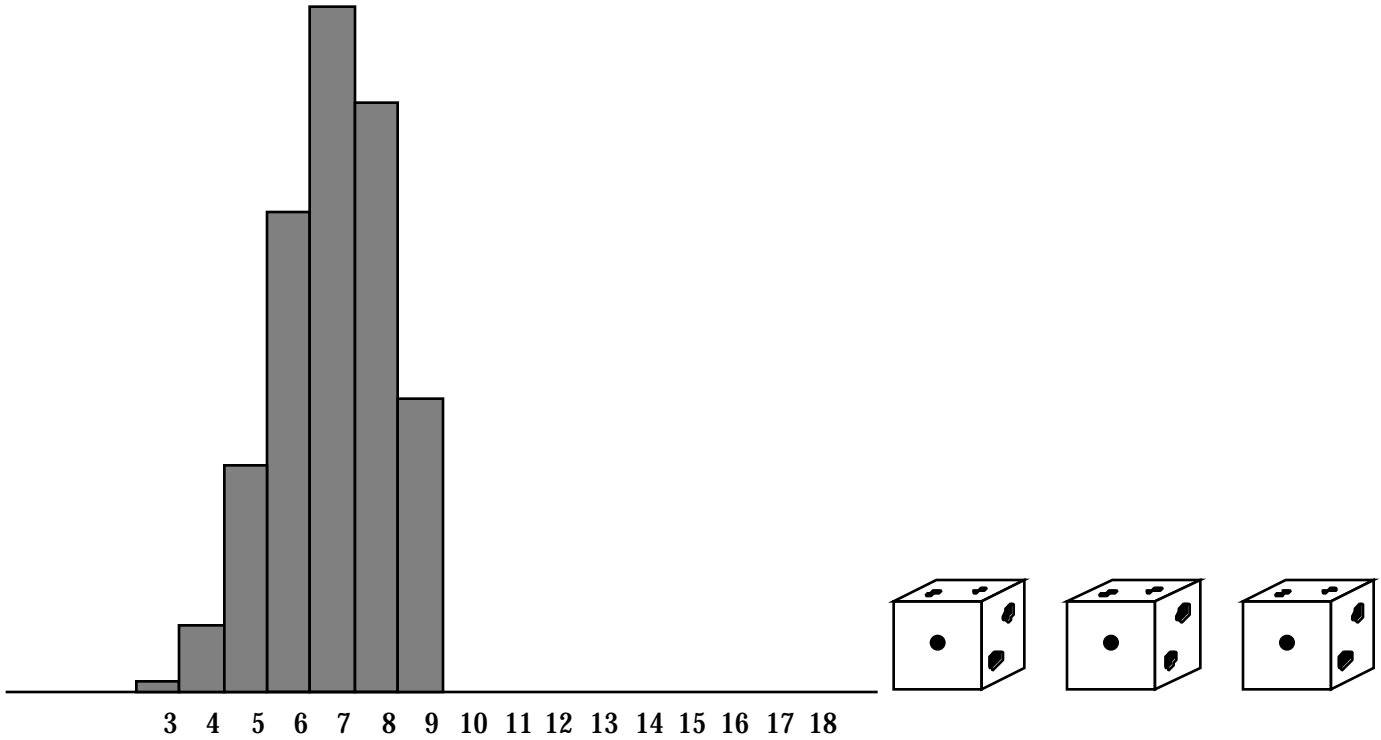
**Theoretical Shape of Distribution: One Altered Die with One 1, Two 2's, and Three 3's (Three possible outcomes, 1, 2, or 3 — peak to the right)**

There is my simple event, neither bell shaped nor symmetrical. What the Central Limit Theorem says is that the shape of the simple event doesn't matter: The compound event, adding the results of many simple events will always tend to be bell shaped and symmetrical. No matter how wierd the simple distribution: combine such events and the result will acquire properties of the Gaussian. Let me put it into action. Throwing two such dice with their individually triangular distributions, what do I get?



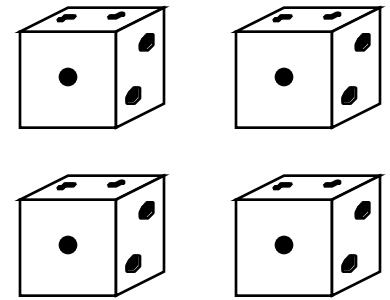
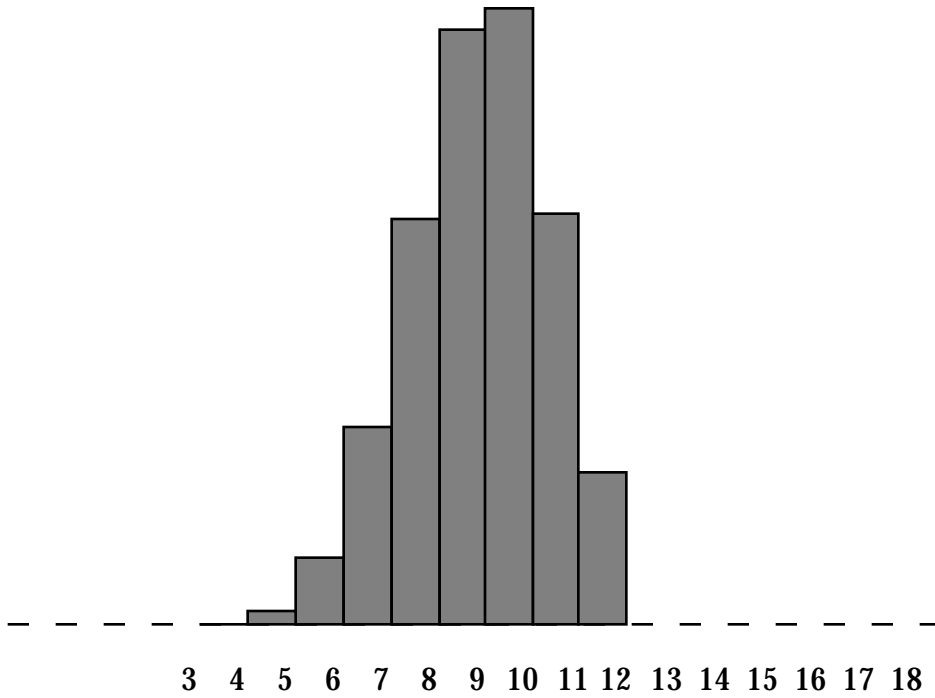
**Theoretical Shape of Distribution: Sum of Two Altered Dice, Five Possible Outcomes, 2 through 6 — Peak displaced Toward Center.**

Again, the composite shape is not the same as the simple shape. Why? For the same reason that ordinary dice tend to get sevens: With these two dice there are only nine ways of getting the extreme value on the right, while there are twelve ways of getting the peak that is displaced toward the center. Throwing three wierd dice, I get.

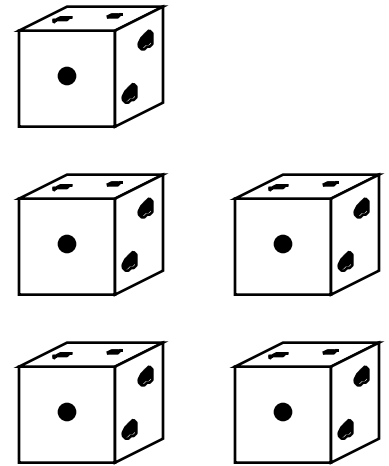
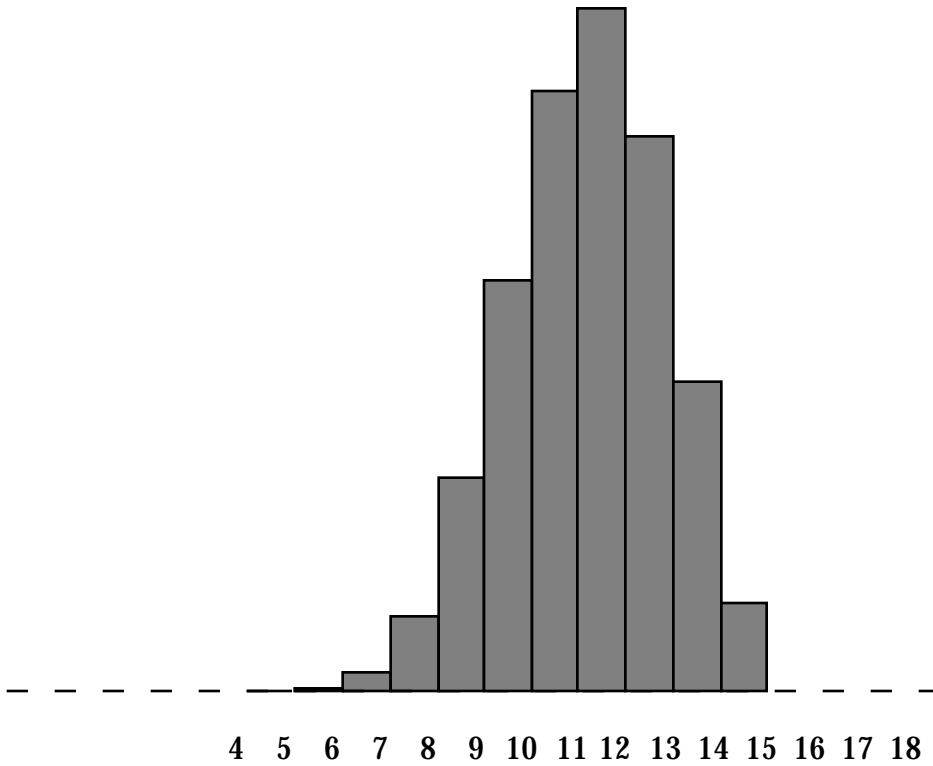


Theoretical Shape ofr Distribution: Sum of Three Altered Dice, Seven Possible Outcomes, 3 through 9, More Symmetrical (More Bell Shaped: Symmetrical, Rounded Concave Sides, Rounded Convex Peak.

And for four dice and five dice I get



Theoretical Shape of distribution: Sum of Four Altered Dice, Nine Possible Outcomes, 4 through 12, More Symmetrical (More Bell Shaped: Symmetrical, Rounded Concave Sides, Rounded Convex Peak)



Theoretical Shape of distribution: Sum of Five Altered Dice, Eleven Possible Outcomes, 5 through 15, More Symmetrical (More Bell Shaped: Symmetrical, Rounded Concave Sides, Rounded Convex Peak)

The shape is still not bell shaped and symmetrical, but it bears little resemblance to the shape shown by the simple event and it is definitely changing. With more dice, compounding more simple events, the Central Limit Proves that the result will get ever closer to being bell shaped and symmetrical. The Central Limit Theorem as mathematics is more precise than that, defining exactly which properties of these distributions become like the properties of the Gaussian distribution. But the point is that even under extreme circumstances there is good

reason to expect symmetrical bell-shaped distributions and to find it interesting when they *don't* happen.