Chapter 4.1 - The Normal Distribution

> "Everybody believes in the exponential law of errors [i.e., the Normal distribution]: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by observation." - Whittaker, E. T. and Robinson, G.

When statisticians discuss models and distributions we always talk about the "data generating process". The idea we're referring to is that all of the data we observe and record arises from a real world system that can be represented by a mathematical function.

It may seem like we're grasping at straws to make math have real life applications or maybe we lack all respect for nature and want to pretend we understand it. The truth is that math has this bizarre habit of making sense when applied to things that *aren't* math. Mathematicians never intend to interact with reality but reality keeps knocking on their office door asking for more.

Capturing the data generating process involves two steps:

- Understanding the constant or *deterministic* features of the process.
- Assuming the nature of the chaotic or *probabilistic* features of the process.

In this chapter we discuss the probabilistic components— and hopefully find some order among complete chaos.

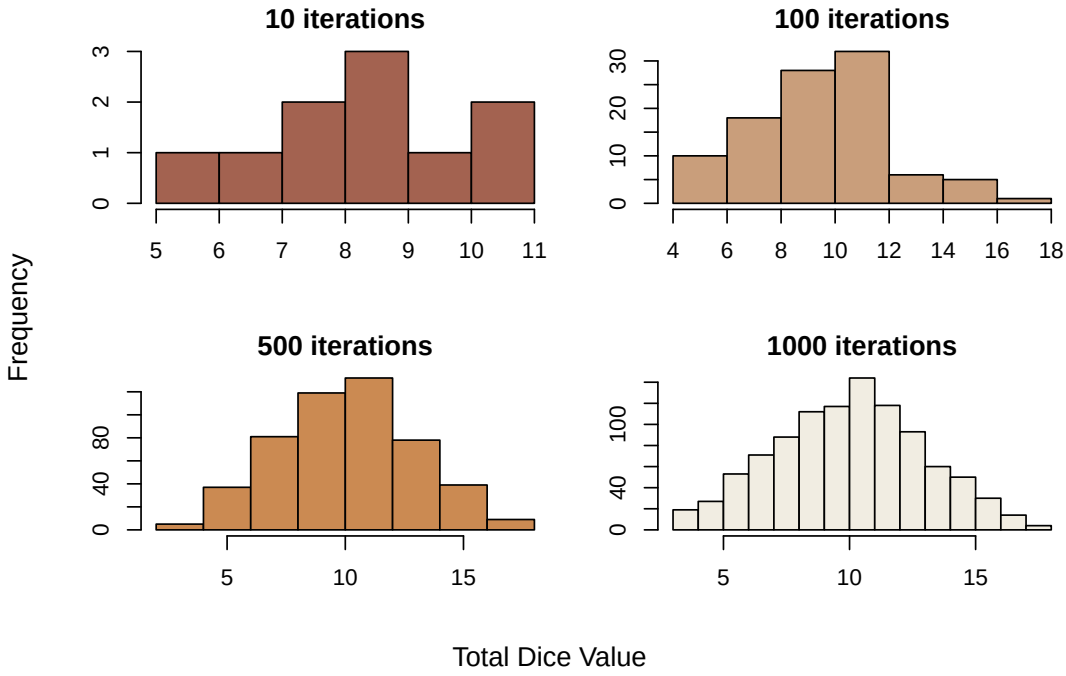
Consider a simple experiment of rolling 3 fair, six-sided dice, then adding up the numbers they land on.

```
set.seed(73) # reproducibility seed
dice=sample(1:6,3,T) # sample from 6 fair, six-sided dice
cat("Dice rolls:", dice, "\n", "Sum of the rolls:", sum(dice), "\n")
```

```
## Dice rolls: 5 1 1
##  Sum of the rolls: 7
```

We're going to repeat this action 1000 times and take a look at the results for first 10, 100, and 500 iterations, as well as the total iterations.

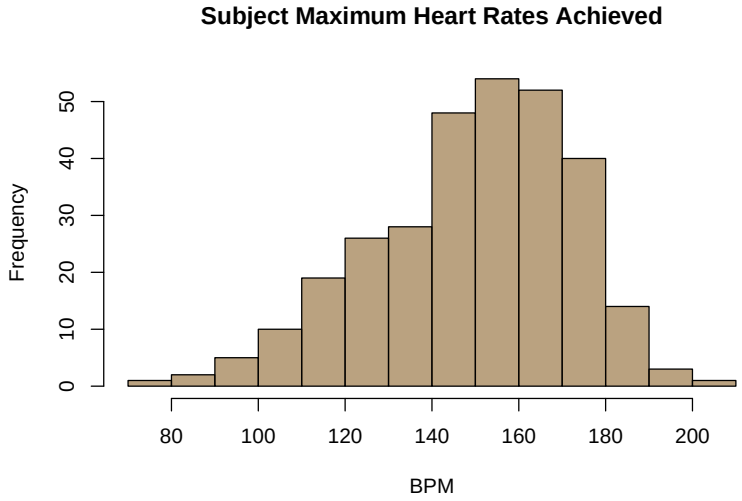
```
set.seed(73)
results=replicate(1000, sum(sample(1:6,3,T))) # 1000 iterations
```



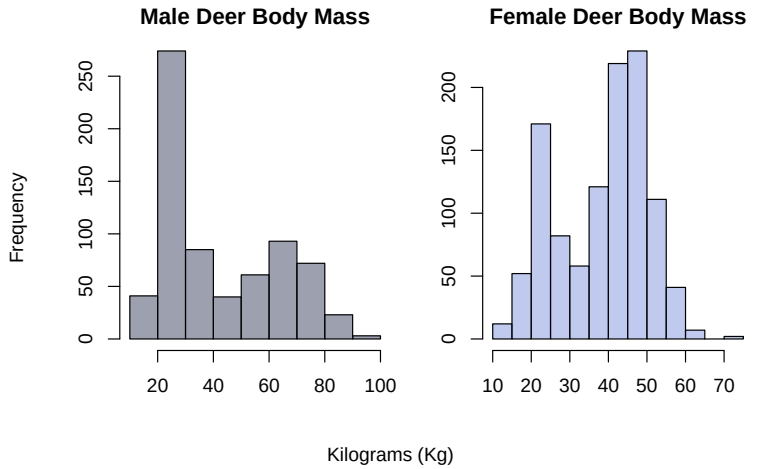
The pattern is clear; as we increase the number of iterations the results become a symmetric histogram with a central peak. We saw this shape before when we discussed the Empirical Rule but never addressed why it's so special that a rule was developed to describe it. As it turns out this shape is just special enough to have **it's own name**. We call this the **Gaussian** or **Normal distribution** and it's the most important function to statistics (and perhaps all of science).

This shape doesn't just pop up when adding up the total of dice rolls, we can find it *everywhere*.

Our cholesterol example comes from a much larger data set with a variety of biometrics on the subjects. If we look at the maximum heart rates the subjects achieved during a cardiovascular test we'll see it looks *vaguely* normal:

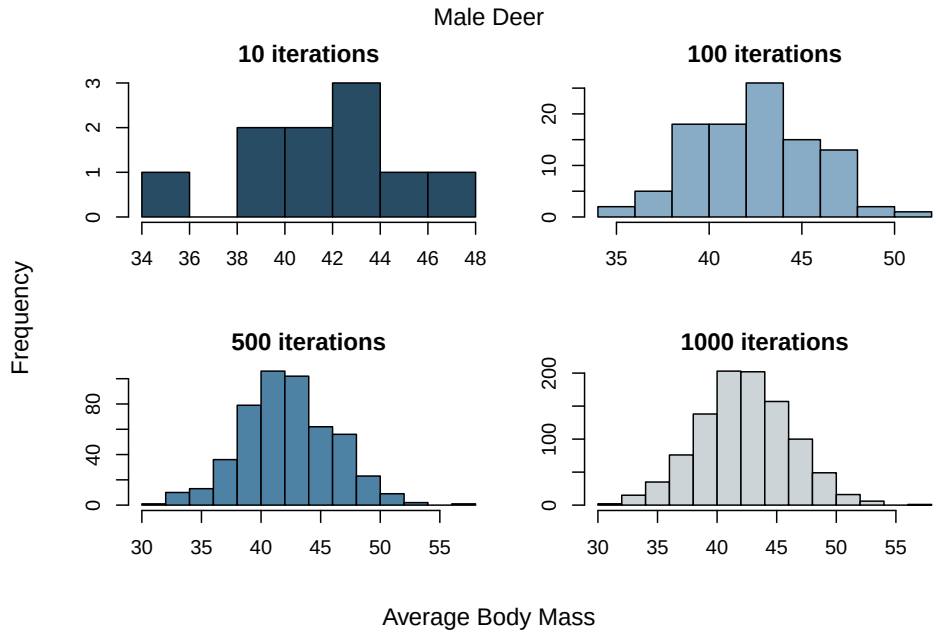


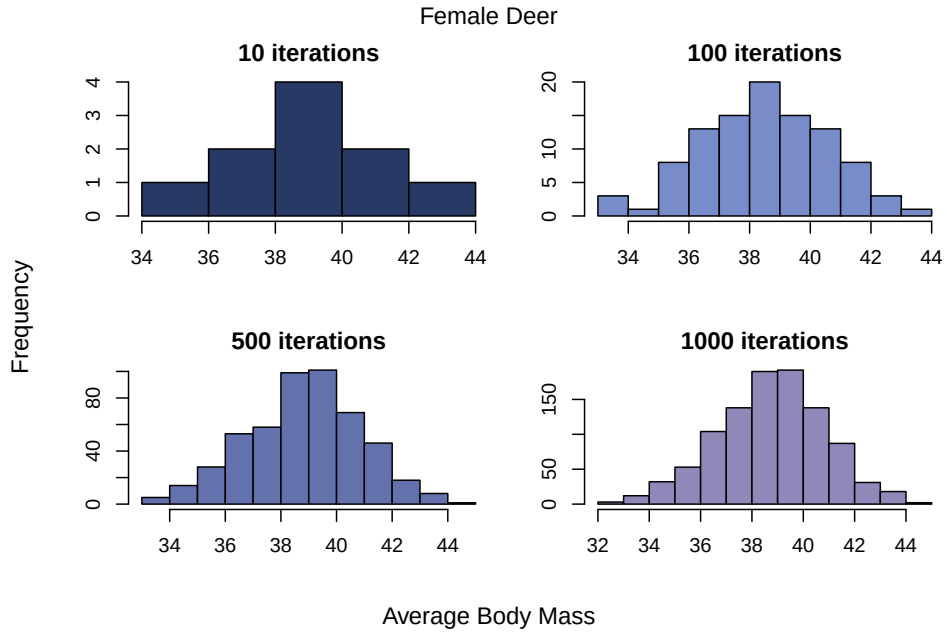
But why isn't it *truly* following a normal distribution? Well that's because real life processes don't match well with the normal distribution. The normal distribution is unique in that it models the *summations and averages* of data over the long run. For instance, if we look at the data for male and deer body mass it should look non-normal.



But we can use this data for an interesting exercise. Let's treat the weight of males and females as if they're two separate *populations* that we can pull samples from. We'll take a random sample of 30 deer from each group and then calculate the average of their body weights. We can then replicate this experiment the same way we did with the dice example, 10, 100, 500, and 1000 times.

```
set.seed(73)
males=replicate(1000,mean(sample(subset(deer$Body.mass.in.kg,deer$Sex=="Male"),
                                 30,replace=T)))
females=replicate(1000,mean(sample(subset(deer$Body.mass.in.kg,deer$Sex=="Female"),
                                   30,replace=T)))
```

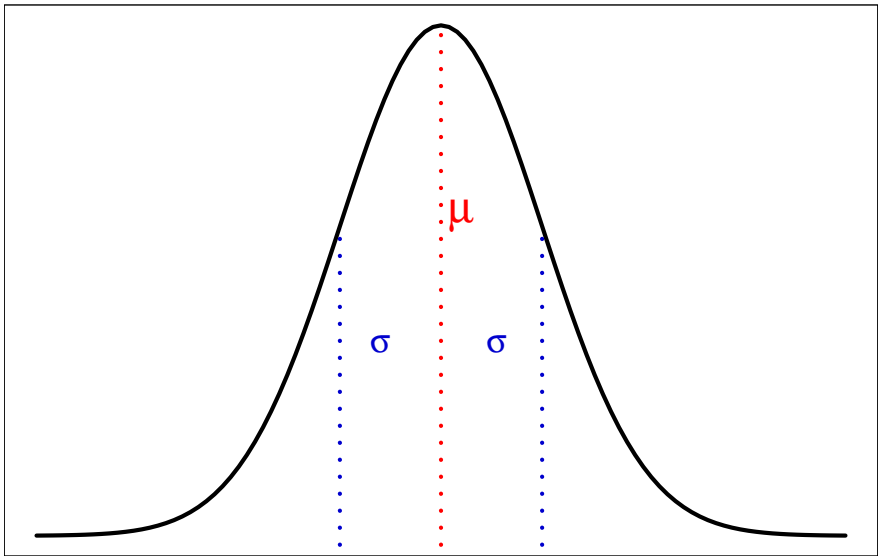




Again, this might seem like a cheeky attempt at making statistics work with the real world. But as we progress we'll see that the nature of the normal distribution makes it the ultimate model for population level dynamics. Since statistics is a field built around describing populations, this is a *very* attractive feature.

Probability Density Function

The **Normal Distribution** is characterized by its iconic symmetric, bell shape curve centered around its peak. When describing distributions we discuss the parameters of their distribution function. The parameters of the normal distribution are its **mean** (μ) and **standard deviation** (σ).



The probability density function for the normal distribution may seem a little clunky and unintuitive, but it has a lot of beautiful properties that explain its structure.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

The function is comprised of two "pieces": the normalizing constant and the kernel.

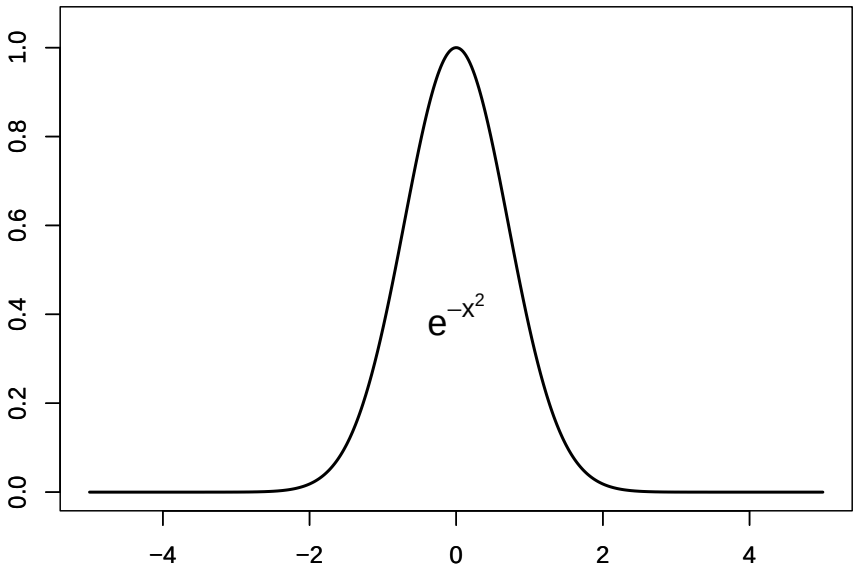
$$\underbrace{\frac{1}{\sigma\sqrt{2\pi}}}_{\text{Normalizing Constant}}$$

Scales the total area to 1

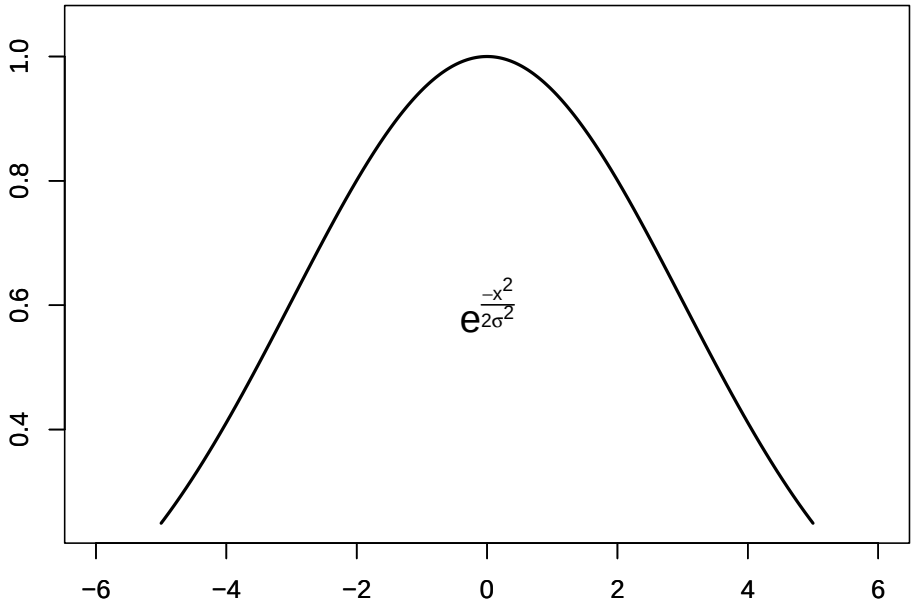
$$\underbrace{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}_{\text{Kernel}}$$

Defines the shape of the curve

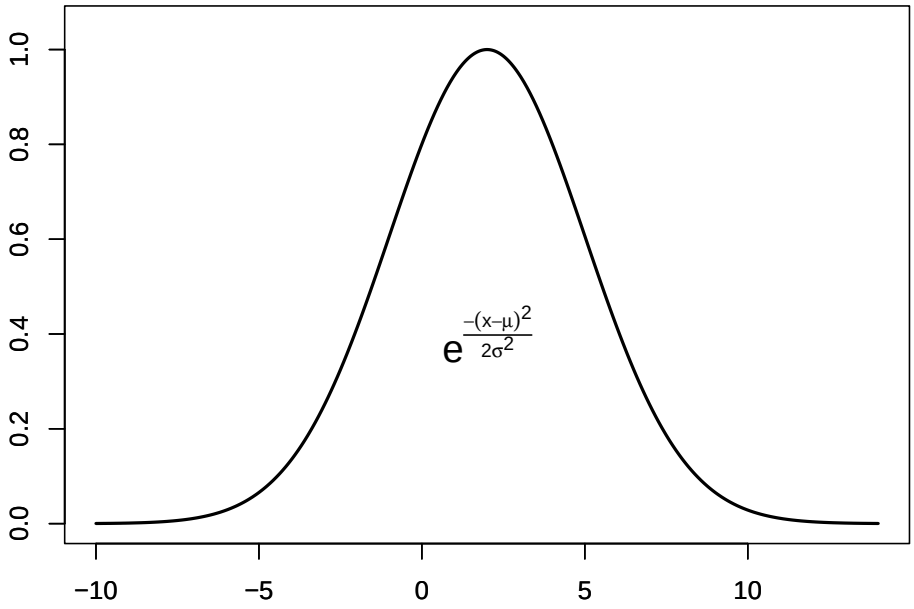
The use of an exponential in the kernel is what gives the normal distribution its "curves", which is made clear when we plot the function e^{-x^2}:



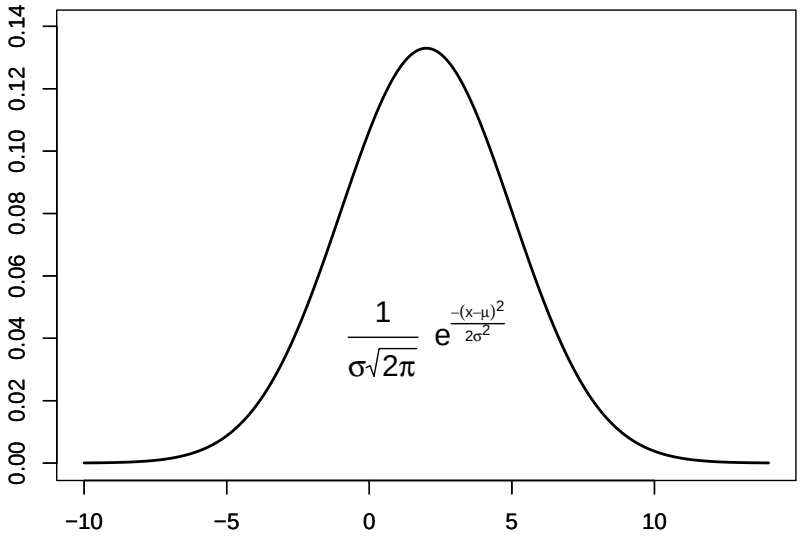
The standard deviation of a normal distribution controls how *spread* the curve is. When we add that piece of the kernel in we'll see that it's tails push outwards. Let's say that we're adding in a standard deviation of 3:



The mean controls the center of the distribution. If we include that component and set the mean equal to 2 we'll see the return of that "dip" in the tails as well as a shift in the center towards 2:



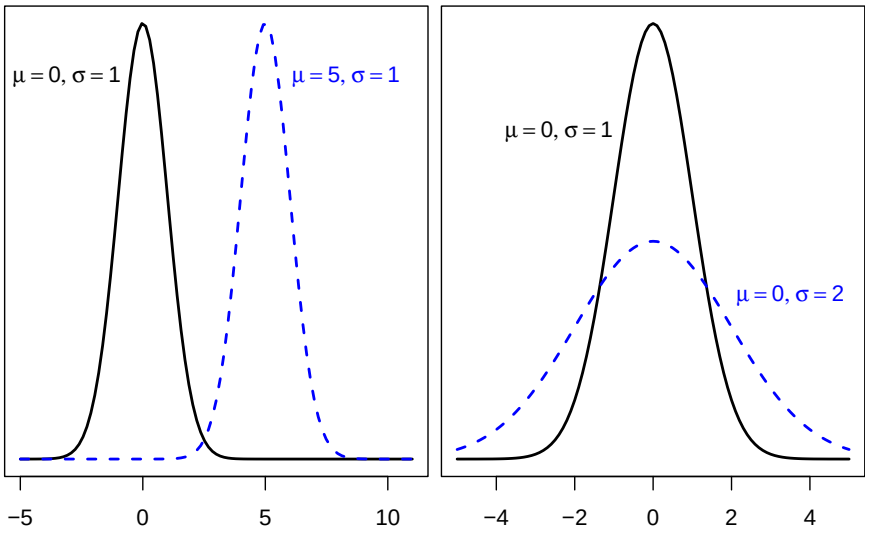
Right now our curve has a far greater area underneath it than 1.00. This is where the normalizing constant comes into play:



While it's not necessary to study the PDF in-depth we should always take the time to understand why the PDF is constructed the way it is. It's hard to make use of a tool when you don't even know what it looks like— it's easier to use a tool if you know *roughly* how it works.

The Standard Normal Distribution

The parameters of the normal distribution are incredibly helpful since they allow us to shift and squish the curve itself:

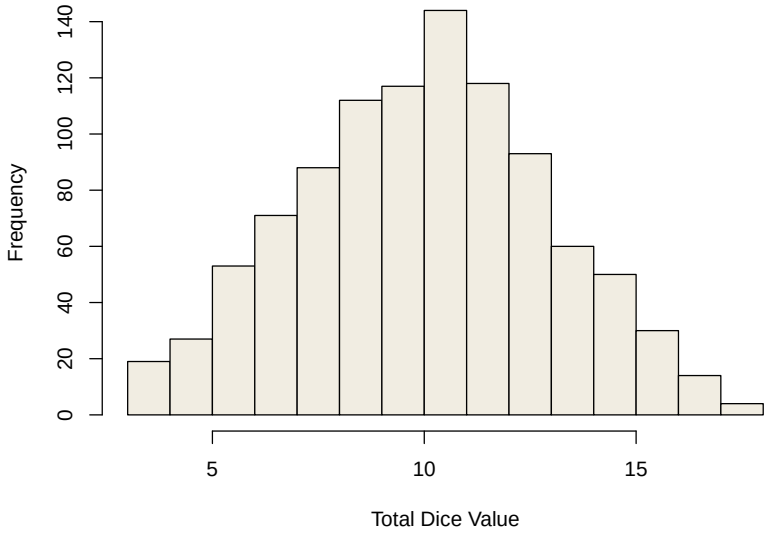


But they create some problems when it comes time to work with the PDF itself. The normal PDF is one of those impossible functions to integrate for reasons that are happily left as an exercise for the reader. Our main focus now is figuring out how to compute probabilities from the normal distribution despite this challenge.

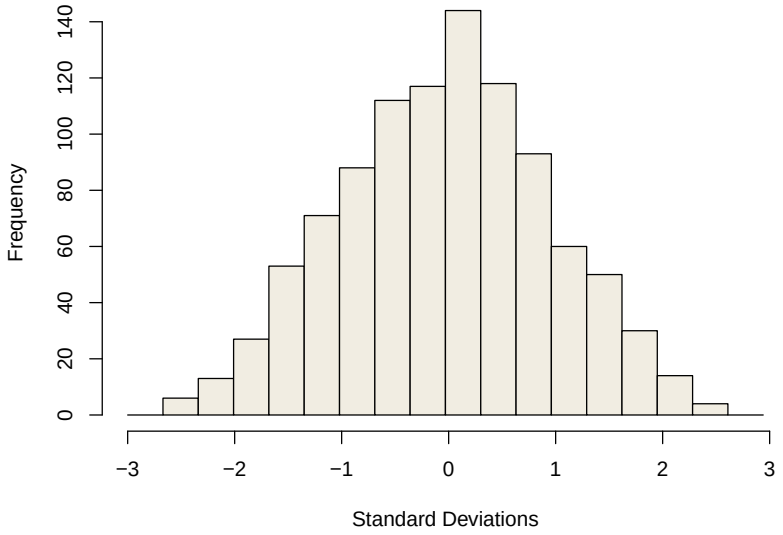
The reality is that since the integral isn't analytically tractable we have to use somewhat exotic methods to *approximate* the probability associated with any given interval beneath the normal curve. Since there are legitimately *infinite* possible combinations of means and standard deviations that we can plug into the normal PDF this is a very inconvenient thing to have to do. But what if we *could* somehow calculate all of the possible combinations and place them into a massive reference book?

Statisticians did that in the 1800s, kind of.

Let's look at the data from our dice experiment again. The histogram from 1000 iterations had that bell shape we're looking for the classify something as normal so we'll just focus on that:



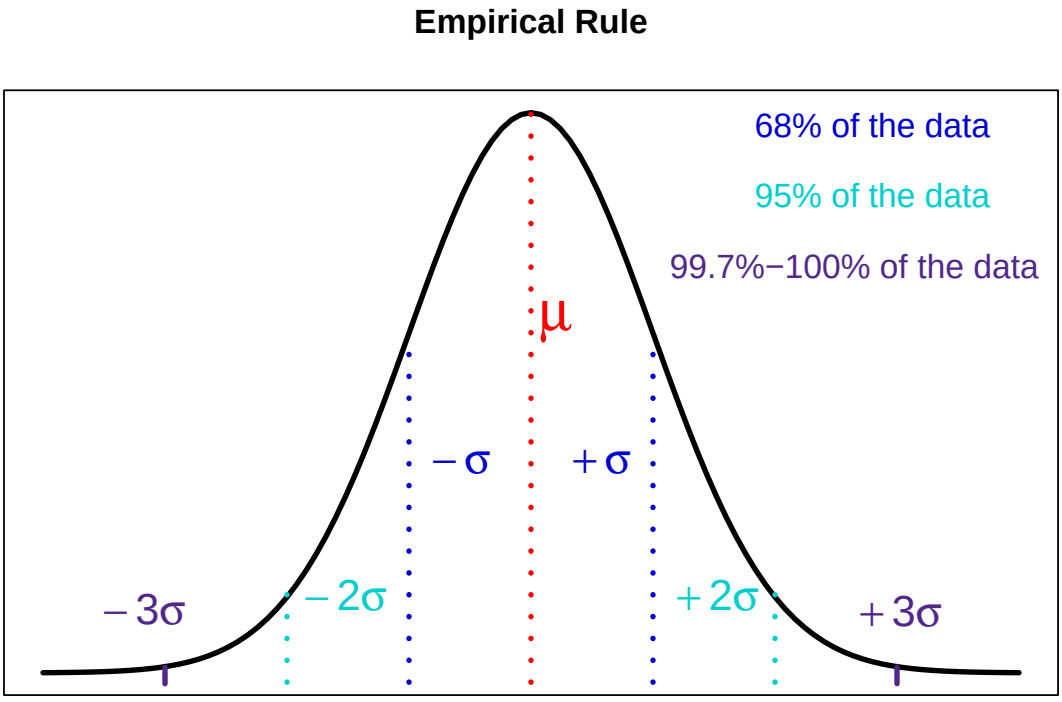
We know from chapter 1.5 that we can standardize data by subtracting it from the mean and dividing by the standard deviation. This converts everything to z-scores that measure every data points distance from the mean in standard deviations.



While the histogram looks a little different due to some issues with the plotting software, nothing else has changed besides this standardization.

The data is now normally distributed with $\mu = 0$ and $\sigma = 1$. What this means is that any calculations made for probabilities from this data are *equivalent* to the original data. So the probability of the 3 dice adding up to a value between 7 and 13 is the **exact same** as the probability that the *standardized* values are between -1 and 1 standard deviations of the mean.

We already know what that probability is thanks to the Empirical rule!



The beauty of this is that instead of making a reference book for every interval on every possible combination of means and standard deviations we can make a single table for this version of the normal distribution and **standardize** every time we want to make a probability calculation.

This special case of the normal distribution is called the **standard normal distribution** and it's purpose is exactly as we've described: create one unified table for calculating the probability density of normal random variables.

Standard Normal Distribution: A special case of the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

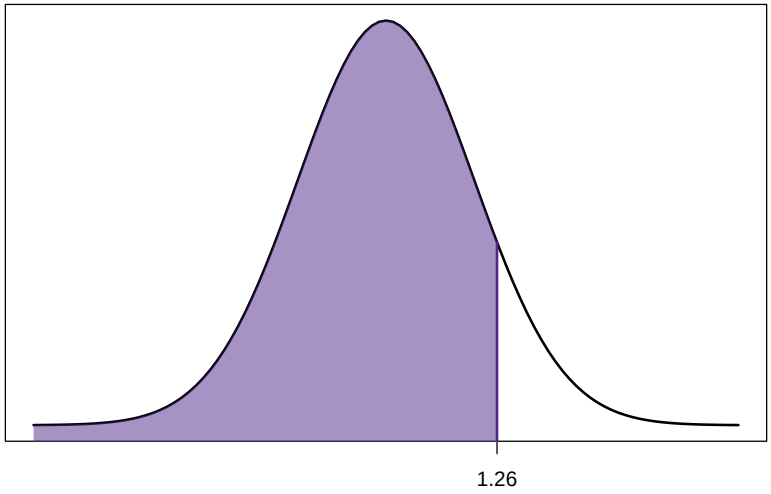
We use the letter Z to represent a standard normal random variable (referring to z-scores). The probability that a **standard normal random variable** Z is between a and b ($P(a < Z < b)$) is equal to the **area under the standard normal curve** over the interval $[a, b]$.

We calculate these probabilities using the **z-table**, a reference table with approximated values for the **area under the curve to the left** of each value between -4 and 4 indexed by 0.01.

Z-Table from -0.5 to 0.5

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224

We can use this table to solve any standard normal problem we might encounter. For instance, if we wanted to find the probability that a standard normal random variable is **less than** 1.26:



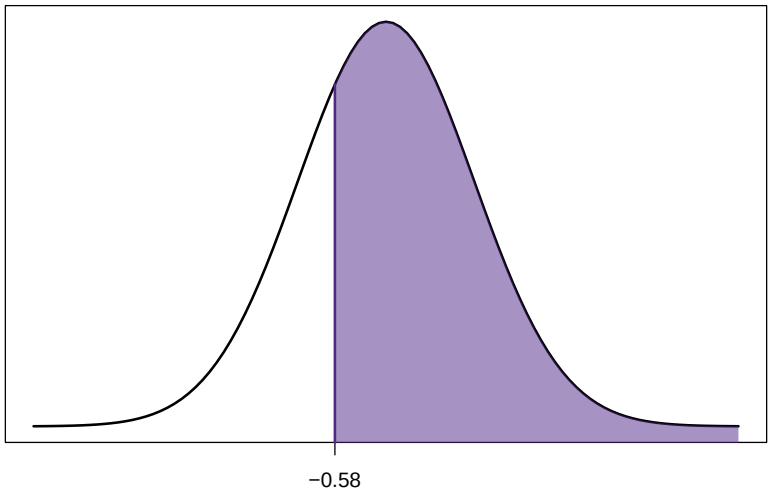
We would look into our z-table, go down to "1.2" on the left, look over to "0.06" on the top, and find their intersection.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	**0.8962**	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

Since the z-table reads to the **left** that means that all the values listed represent the probability that a standard normal random variable, Z, is **less than** the given value. So the answer is fairly direct:

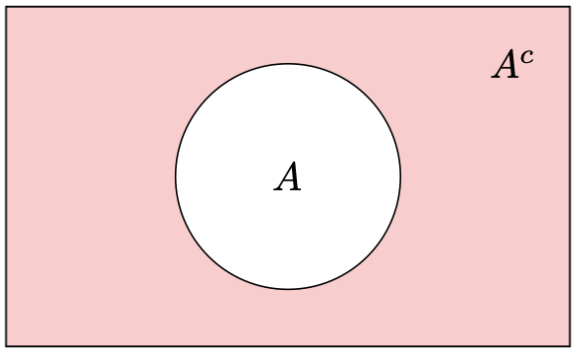
$$P(Z < 1.26) = 0.8962$$

What about the probability that Z is **greater** than a specific value?



Even though the z-table only reads to the left we can easily solve this problem by turning to our dear friend set theory!

For a probability distribution to be considered legitimate is has to have a total probability equal to 1.00 which means that the area underneath the curve is also 1.00 ... this should feel familiar.



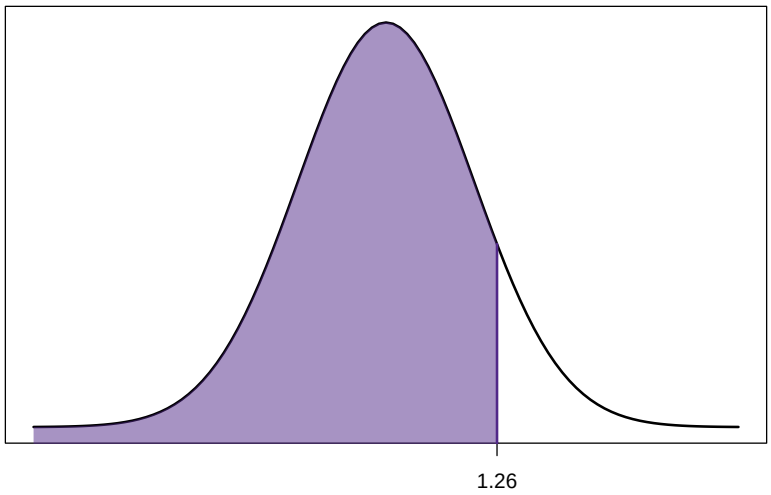
All we have to do is calculate the probability that Z is **less than** -0.58 and then subtract that probability from 1.00 to find the *complement*. This complement is the probability that Z **isn't** less than -0.58, which is the definition of being greater than -0.58.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	**0.2810**	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483

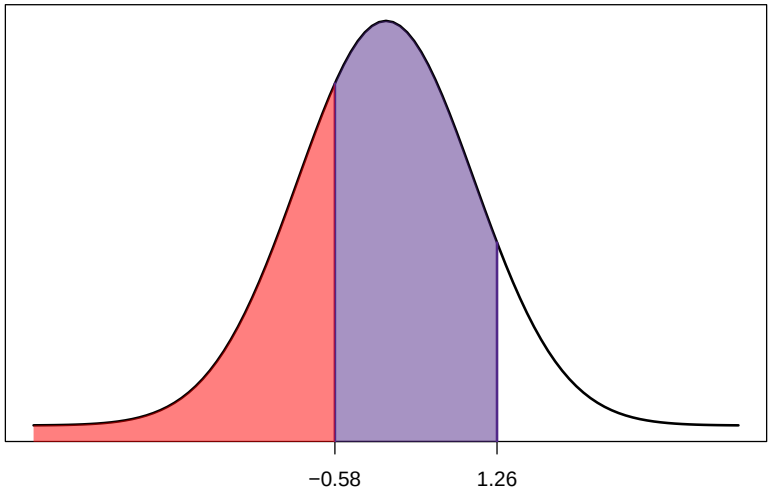
$$P(Z > -0.58) = 1 - 0.2810 = 0.7190$$

What about the area between these two values? When it comes to calculating intervals there's a few different tricks we can use.

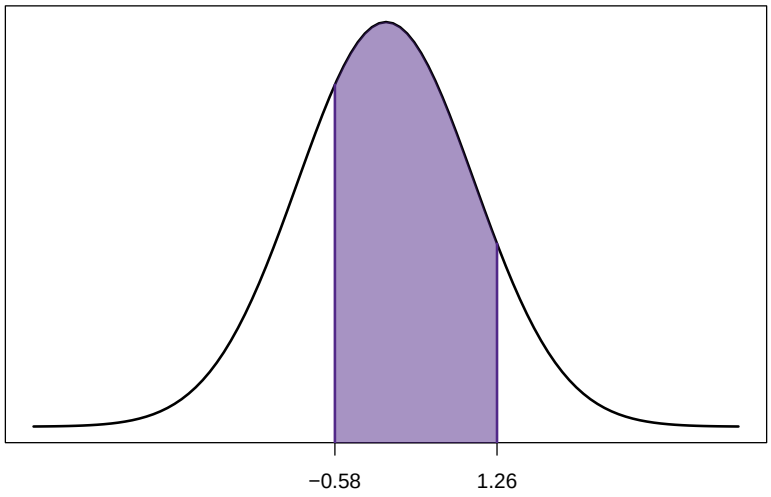
Since we've already calculated the area **to the left** of each value in this interval we can **subtract** them from one another. This is better explained visually. We can start with the area to the left of 1.26:



And then remove the area to the left of -0.58:

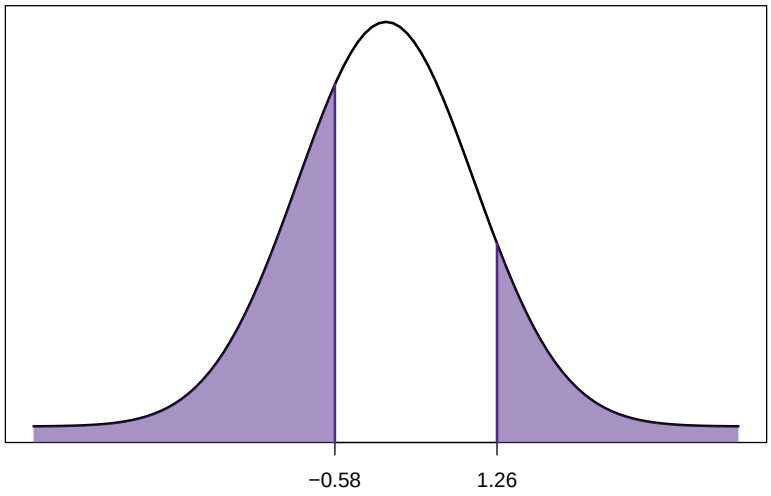


This leaves us with only the interval between the two.



$$P(-0.58 < Z < 1.26) = 0.8962 - 0.2810 = 0.6152$$

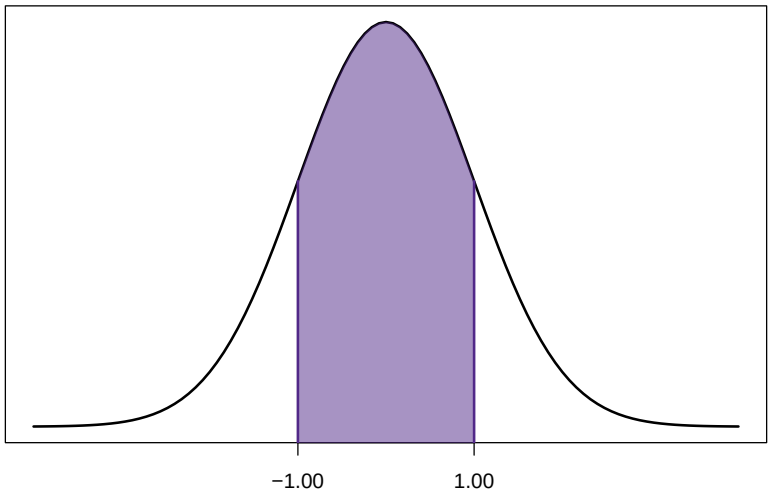
We can also apply the complement trick here by calculating just the area of the tails, so the area to the left of -0.58 and the right of 1.26:



The complement of this is that interval between them:

$$P(-0.58 < Z < 1.26) = 1 - (0.2810 + 0.1038) = 0.6152$$

Any time we're dealing with a symmetric interval we can get away with calculating the area the the left of the lower value or the area to the right of the upper value, **doubling** it, and finding the complement. This is a nice by-product of the normal distribution being **symmetric**.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	**0.8413**	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	**0.1587**	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611

$$P(Z > 1.00) = 1 - P(Z < 1.00) = 1 - 0.8413 = 0.1587 = P(Z < -1.00)$$

$$P(-1.00 < Z < 1.00) = 0.8413 - 0.1587 = 0.6826$$

$$P(-1.00 < Z < 1.00) = 1 - (2 \times 0.1587) = 1 - 0.3174 = 0.6826$$

The only other common standard normal problem we'll encounter is calculating the *bounds* of an interval when all we have is the *probability*.

$$P(-z_0 < 0.95 < z_0) = ?$$

This is an instance where we can leverage the symmetric property of the normal distribution. We know that the distribution has a total probability of 1.00 and the area to the left of one number should be equal to the area to the right of the opposite signed number.

If we find the complement of the desired *interval*, divide it in half, and locate the number with that probability then the area between it and it's mirror should be equal to the interval's probability.

$$\frac{1-0.95}{2} = 0.025$$

This involves searching the table for a specific probability, but this isn't a particularly challenging task (albeit slightly inconvenient):

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	**0.0250**	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294

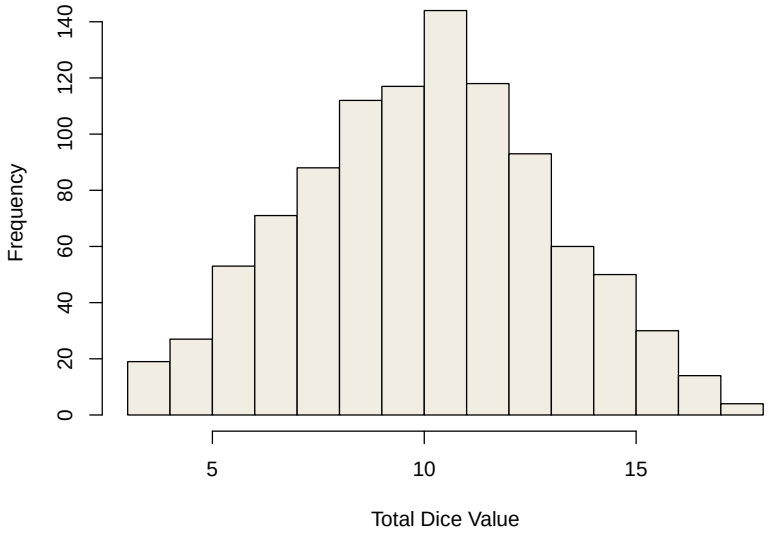
Now if we look to the opposite end of the table and find the area to the left of 1.96 we can subtract these areas from one another to check if they match the interval probability:

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	**0.9750**	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

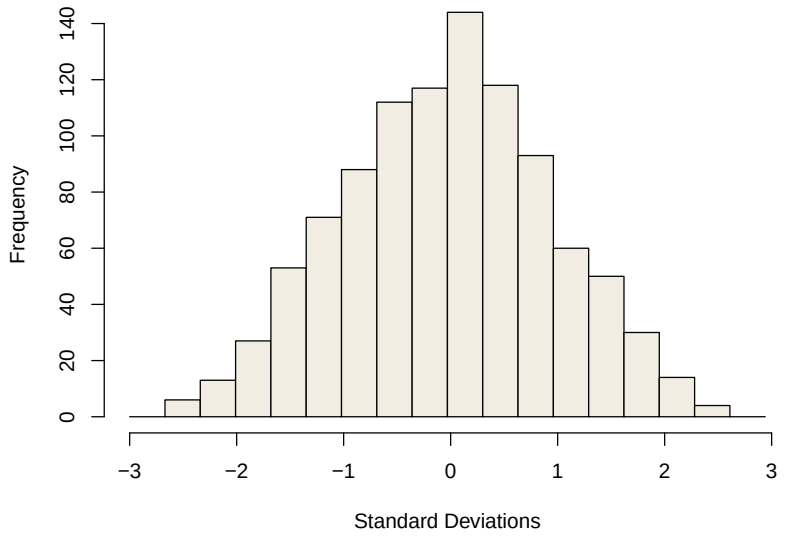
$$0.975 - 0.250 = 0.95 \qquad z_0 = 1.96$$

The General Normal Distribution

The whole point of developing the standard normal distribution and going through all of that z-table nonsense was so that we could work with *real data*; how do we do that?



The trick we used to introduce the standard normal distribution is the answer to that question. When we have *non-standard* normal data we *standardize* it by coverting the data into z-scores, use the z-table to answer any questions we have, and then *un-standardize* that information to make it relevant to the data again.



It's un-reasonable to assume anyone is going to calculate 1000 z-scores by hand. Remember: we need this to be useful with *and* without a computer. This is where we have to make *assumptions* to simplify the process.

Let's start with the assumption that the dice data is *actually* normally distributed. If this is true then we can fully describe it with just mean and variance.

```
# dice experiment
mean(results) # mean
var(results) # variance
sqrt(var(results)) # standard deviation
```

```
##
##  Mean: 10.491
##  Variance: 8.84076
##  Standard Deviation 2.973342
```

When representing the distribution of a random variable we use the notation:

$$X \sim f(\cdot)$$

Where X is the random variable, $\sim$ means "is distributed", and $f(\cdot)$ is the distribution with its parameters included. In the case of Z, the standard normal random variable, we can write it out as such:

$$Z \sim N(0, 1)$$

This is saying "Z is distributed normal, mean 0, variance 1". It should be noted that we use *variance* for the normal distribution when describing it with this notation, but since $1^2 = 1$ the variance and standard deviation are the same value.

When the mean and variance of a normal random variable are **not** 0 and 1 respectively we refer to its distribution as the **General Normal Distribution** (generally just called the normal distribution). To describe the general normal random variable, X, we write:

$$X \sim N(\mu, \sigma^2)$$

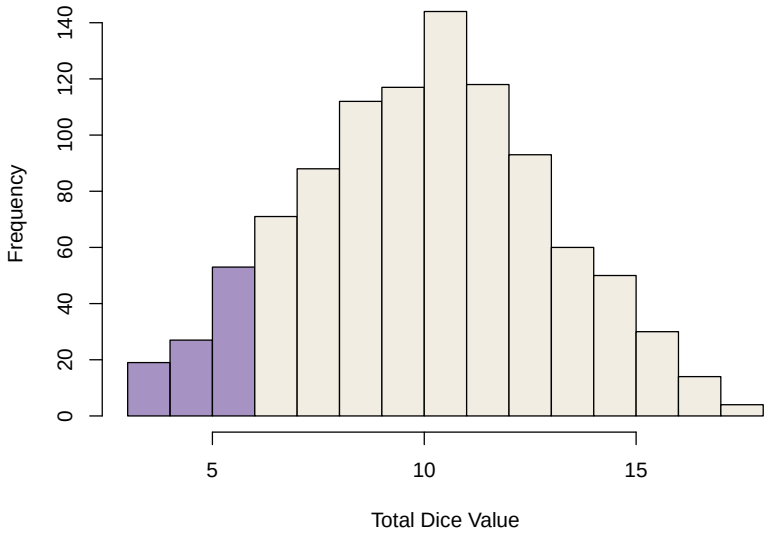
"X is distributed normal, mean μ, variance σ^2". Now let's consider the dice data to arise from a normal random variable, Y.

$$Y \sim N(10.5, 8.7)$$

If we were to *standardize* Y such that it becomes Z we just need to apply the z-score formula:

$$P(Y < 6) = P\left(Z < \frac{Y - 10.5}{\sqrt{8.7}}\right)$$

We can now use the z-table to calculate the probability of Y realizing to any interval within ± 4 standard deviations from it's mean. So let's do that, what would the probability be that we roll our 3 dice and their sum is less than 6?



$$P\left(Z < \frac{6 - 10.5}{\sqrt{8.7}}\right) = -1.515 \approx --1.52$$

We're rounding here so that we can actually *use* the z-table.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	**0.0643**	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681

Since this is an experiment we can check our work against the true results:

```
# proportion of values less than or equal to 6
sum(results <= 6)/1000
```

```
## [1] 0.099
```

These two values are close, but not exact, which is fine because we never expect exact in statistics. In this case there's an element of randomness to the program we're using to create the data and some estimation error due to attempting to use continuous methods on discrete values. Still, we can see that our z approximation gave enough information about the dice rolls that we could make realistic decisions without ever performing the experiment.

There is the problem of reversing our steps. We could ask how high of a roll we need to land on to be considered 90^{th} percentile, which has a simple enough solution for standard normal:

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	**0.8997**	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177

But 1.28 doesn't answer our question. However since a z-score is just the number of standard deviations away from the mean our particular value is we should be able to multiple the z-score by the *standard deviation* and add it to the *mean* to answer the question:

$$10.5 + (1.28)\sqrt{8.7} = 14.275$$

Considering we're dealing with a discrete question we can answer the problem fully by rounding. It's typically better to over-estimate values when dealing with probabilities and proportions but this is also very case-by-case. In this case we can safely assume 15 is within the $90th$ percentile.

Non-standard normal problems are fairly simple so long as we have a firm grasp on standard normal concepts. The two formulas we've seen here cover the full extent of introductory questions surrounding general normal distribution probabilities:

$$P(X < x) = P\left(Z < \frac{x - \mu}{\sigma}\right)$$

$$X = \mu + z \times \sigma$$

The normal distribution is an incredible useful and prevalent concept throughout statistics, but it's not the full extent of distribution theory. To consider ourselves "familiar" with distribution theory we still have to address two problems:

1. We currently have no clear way to deal with data that's not normally distributed or even discrete.
2. Our methods so far have used very large sample sizes. We have no methods to handle *realistic* sample sizes.