# Chapter 3.3 - Random Variables

> "He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may be cast." - Leonardo da Vinci

Analogies are convenient because they help us take very complex or obscure concepts and boil them down to simple examples. The analogy we typically consider for mathematical and "theoretical" statistics (used in a loose sense since statisticians consider performative exercises of math skills and an aversion to computer programing 'theory') is building the foundation of a house.

It's an excellent analogy because awful statisticians are born from a poor understanding of mathematical statistics and linear models, like a house with a rotting foundation. But there's a component that's rarely addressed in this analogy because statisticians don't build houses we sit in front of computer screens staring at terribly designed spreadsheets and screaming at error messages. Foundations are built from concrete and concrete *takes time to dry.*

If you're struggling with the materials we're building a foundation from you should consider that perhaps this is **intended**. Most students don't understand what the hell they learned in college until 3 to 4 years after they graduate. Don't let confusion prevent you from progressing, but most importantly *be kind to yourself.*

---

Let's flip a coin three times. What's the probability that we land on heads twice? Right now we only know how to do this by looking at all of the possible outcomes— so let's do that.

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

| **Flip 1** | **Flip 2** | **Flip 3** |
|---|---|---|
| H | H | H |
| H | H | T |
| H | T | H |
| T | H | H |
| H | T | T |
| T | H | T |
| T | T | H |
| T | T | T |

We can denote an event, landing on heads twice, and call it $A$. The probability of $A$, $P(A)$, is the summation of all outcomes where heads appears twice divided by the total number of possible outcomes.

$$P(A) = \frac{3}{8} = 0.375$$

If we instead wanted to look at the probability that *at least* 2 coin flips result in heads we would need a new event. This is annoying for two reasons: (a) it's not practical and (b) we know that we're only adding one outcome so this feels pointless. It would be more convenient to have a *different kind of event* that denotes the **number** of heads we observe. Then we could just use an inequality to do the rest of the work for us!

It should be obvious at this point in the textbook that I'm always alluding to something that already exists. We refer to this new kind of "event" as a **random variable**.

**Random variable** (shorthand: r.v.): A rule for assigning a numerical value to each outcome of a random experiment

It's general convention to use a **capital letter toward the end of the alphabet** to notate random variables (i.e., $X$, $Y$, $Z$).

Let $X$ be a random variable denoting the number of heads we observe in our three coin flips.

$$X = \{\text{the number of heads observed}\}$$

Random variables have their own version of a sample space, referred to as their **support**.

**Support:** The set of possible values a random variable can be.

The support of $X$ is $S_X = \{0, 1, 2, 3\}$. We generally use $S$ to denote a sample space, while $S_X$, where $X$ is some random variable, refers to **the support of** $X$.

As soon as we flip the coin three times and record the number of heads we can't really justify that the number is random. Rather, it has become a *realization* of its original random variable. Because of this we have to notate the *result* differently from the *proposal*.

$$X = \text{the value before the experiement has been performed (still random)}$$

$$x = \text{the value after the experiement has been performed (not random)}$$

We have to make this distinction because the language of statistics relies on the ability to check the probability that a random variable realizes to a given value. In this case we want to make the following expressions **make sense**:

$$P(X = x) \quad \text{means the probability that r.v. } X \text{ is equal to possible value } x$$

$$P(X > x) \quad \text{means the probability that r.v. } X \text{ is greater than possible value } x$$

---

## Discrete Random Variables

Random variables are separated into two major types: Discrete or Continuous. The difference between them is an entire schism between mathematical fields so it's good to tackle them one at a time.

No matter how many times we flip our coin we'll never be able to record a partial value. There will always be a 1 or a 0 marked down for the number of heads or tails. This property of our random variable is what defines it as a **discrete random variable**.

**Discrete**: The number of **possible values** in the **support** is **finite** or **countably** infinite.

It should be noted that partial values aren't a *requirement* so much as a good rule of thumb. Your random variable could take on partial values and still be discrete so long as the partial values can be **coded as discrete** (i.e., recording quantities as half steps like $0, 0.5, 1, 1.5, ...$ could be coded as just $0, 1, 2, 3, ...$ as long as we change the definition of the half steps).

$$\text{Let } \ Y = \{\text{The number of fish in a pond}\}$$

$$S_Y = \{0, 1, 2, ...\}$$

Can we count have half a fish? Technically yes. But half a fish isn't biologically important to count. If partial counts **feel arbitrary** then you're more than likely working with a **discrete** variable. While we can hypothetically have **infinite** fish, but the **realized value** will always be a **whole** number.

---

## Discrete Probability Distributions

We can take the table of possible outcomes for our coin tosses and convert it into a "probability model" for head flips.

| **Number of Heads** | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **Probability** | 0.125 | 0.375 | 0.375 | 0.125 |

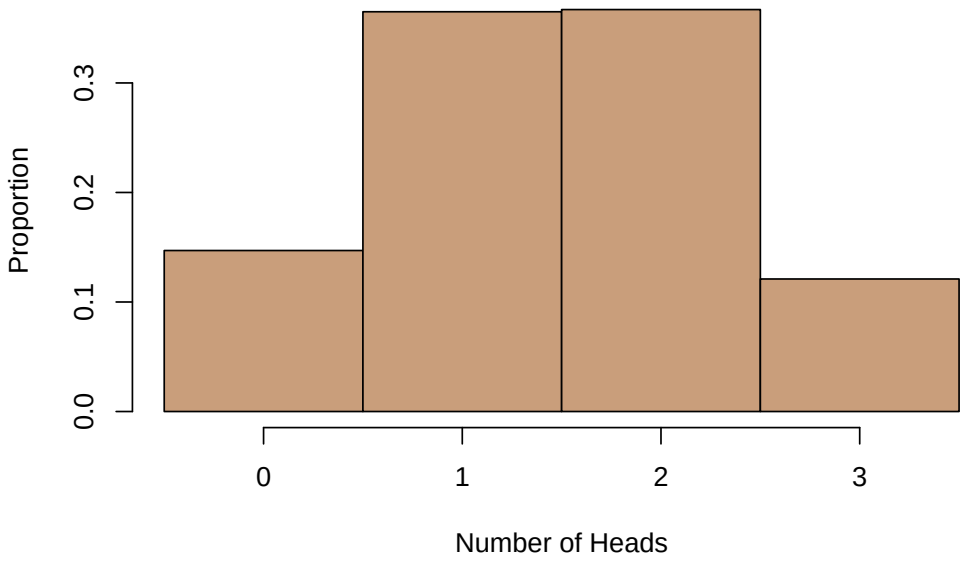
We can run an experiment to see if the probability distribution matches. Applying what we learned about the law of large numbers, let's flip three coins 1000 times.

```
set.seed(73) # reproducibility seed
x=rbinom(1000,3,0.5) # 3 fair coin flips repeated 1000 times
as.data.frame(table(x)) # table and report results
```

```
##   x Freq
## 1 0  147
## 2 1  365
## 3 2  367
## 4 3  121
```

These can be converted into proportions by dividing each set of outcomes by 1000, and the same can be visualized with a histogram:

| **Number of Heads** | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **Proportion** | 0.147 | 0.365 | 0.367 | 0.121 |




We can see that the results from each run of this experiment are spread out, or *distributed*, in a certain way. This is no accident, we refer to this spread of results as the **probability distribution** for this random variable.

The form of a random variable's **probability distribution** depends on whether it is **continuous** or **discrete**. For a **discrete random variable** the probability distribution is often a *list* of all possible values the random variable can take and their corresponding *probabilities of occurrence*

**Discrete probability distributions** satisfy the following two properties:

$$i. \quad 0 \leq P(X = x) \leq 1$$

$$ii. \quad \sum_{x} P(X = x) = 1$$

Remember that we're working inside of a sample space. Even though the sample space may contain multiple random variables with their own probability distributions, each distribution has to satisfy these properties to be considered legitimate.

---

## Expectation of Discrete Random Variables

If you flipped a coin 3 times, how many times would you land on heads **on average**? Why ponder the problem when we can just do it! We'll flip three coins 1000 times and take the mean of our results:

```
set.seed(73)
cat("Mean:", mean(rbinom(1000,3,0.5))) # average 1000 iterations of 3 coin tosses
```

```
## Mean: 1.462
```

This is a troubling result since we've already established that partial values are *impossible* for this random variable. It'd be nice to at least confirm whether this is the correct calculation. Let's refer back to the probability distribution for the number of heads in three coin flips.

| **Number of Heads** | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **Probability** | 0.125 | 0.375 | 0.375 | 0.125 |

When we ran the initial 1000 coin flip experiment and calculated its probability distribution, what did we do to compute those probabilities? We divided the number of times each value occurred by the total number of flips. If we reverse our steps we'll get back to the original results of the experiment:

| **Possible Values** | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **Result** | 147 | 365 | 367 | 121 |

The program that calculated the average from these results did so by using the mean formula we've previously covered. That's not very convenient since typing 1000 values into a calculator would take a little while. A trick we can use to avoid this is by multiplying the number of each outcome by the value it should have taken in the original data. This is called *weighting* the results.

| **Weighted Value** | 0 | 365 | 734 | 363 |
|---|---|---|---|---|

Now we're just adding 4 values instead of 1000, but we still need to divide the summation by 1000 since that's the true sample size the data arose from:

$$\frac{0 + 365 + 734 + 363}{1000} = \frac{1462}{1000} = 1.462$$

Interestingly enough, since those proportions are the result of multiplying by 1000 we can just add up the weighted proportions to get the **exact** same answer.

| **Weighted Proportion** | 0.000 | 0.365 | 0.734 | 0.363 |
|---|---|---|---|---|

$$0.00 + 0.365 + 0.734 + 0.363 = 1.462$$

We can use this technique with our *original* probability distribution that we calculated from theoretical results of flipping three coins:

| **Number of Heads** | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **Probability** | 0.125 | 0.375 | 0.375 | 0.125 |
| **Weighted Probability** | 0.000 | 0.375 | 0.750 | 0.375 |

$$0.000 + 0.375 + 0.750 + 0.375 = 1.5$$

This is *very* close to the average from our experiment, just as the probability distribution was close. However this value of 1.5 has a unique interpretation from the one from the experiment. If we were to keep repeating our three coin tosses an *infinite* number of times we would **expect** to see 1.5 of every 3 coin tosses land on heads across all of those infinite repetitions. This is known as the **expectation** of a random variable.

The **expectation** of a discrete random variable is the sum of all possible values of the random variable, weighted by their individual probabilities.

$$\mathbb{E}X = \sum_x x \ P(X = x)$$

This differs from a sample or population mean because it uses the idea of an *infinite sample size*. That said we can denote the expectation of a random variable as $\mu$, but because of it's alternate use cases we often stick to some form of $\mathbb{E}X$.

We can use this formula for *any* discrete random variable so long as we know the full probability distribution. Consider the discrete random variable, $Y$:

$$Y = \{\text{number of rabies positive bats in a cave}\}$$

| $y$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $P(Y = y)$ | 0.4 | 0.2 | 0.15 | 0.1 | 0.1 | 0.05 |

$$\mathbb{E}Y = \sum_y yP(Y = y)$$

$$0(0.4) + 1(0.2) + 2(0.15) + 3(0.1) + 4(0.1) + 5(0.05) = 1.45$$

Expectations can be thought of as "long-run averages" meaning that as we approach an infinite sample size we would see the average values of the experiment converge to the expectation.

If we were describing this result we would say: "over time we **expect** to have 1.45 rabies positive bats in the cave".

## Variance of a Discrete Probability Distribution

If we can measure the mean or center of a probability distribution we *should* be able to measure the **variance** or **spread** of the distribution. We'll stick to our coin flips for now— to keep things simple.

```
set.seed(73)
cat("Variance:", var(rbinom(1000,3,0.5))) # variance for 1000 iterations of 3 coin tosses
```

```
## Variance: 0.7853413
```

Let's build off of what we've already learned. We know that theoretical probability distribution is assuming infinite sample size, so we'll abuse that knowledge. The general formula for sample variance is:

$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$

If the sample size is infinite then we can throw out the $n-1$ in the denominator and replace it with the probabilities that are *derived* from this infinite sample.

$$\sum_i^n (x_i - \bar{x})^2 P(X = x)$$

The analog for sample mean in this case is the expectation and the data is just the set of values the random variable can take. We'll notate the expectation, $\mathbb{E}X$, as $\mu$ in this case.

$$\sum_x (x - \mu)^2 P(X = x)$$

It's left as an exercise for the reader to calculate the variance using the experimental data. Let's compare the output of our 1000 coin flips with the theoretical variance:

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(X = x)$ | 0.125 | 0.375 | 0.375 | 0.125 |
| $(x - \mu)^2$ | 2.25 | 0.25 | 0.25 | 2.25 |

$$2.25(0.125) + 0.25(0.375) + 0.25(0.375) + 2.25(0.125) = 0.75$$

Again, quite close to our approximation. The law of large numbers allows us to get *much closer* by blowing up the sample size:

```
set.seed(73)
cat("Variance:", var(rbinom(10^6,3,0.5))) # variance for 10^6 iterations of 3 coin tosses
```

```
## Variance: 0.7506564
```

We denote the **variance** of a random variable as $\mathbb{V}X$ or $\sigma_X^2$. Just like with the expectation, variance tends to be written out as $\mathbb{V}X$ to avoid confusion. The $\sigma_X^2$ notation does come in handy for explaining the next measurement: standard deviation.

$$\sigma_X = \sqrt{\sigma_X^2}$$

Together we have two general formulas for the spread of discrete probability distributions:

$$\mathbb{V}X = \sigma_X^2 = \sum_x (x - \mu)^2 P(X = x)$$

$$\sigma_X = \sqrt{\sigma_X^2}$$

In practice:

$$Y = \{\text{number of rabies positive bats in a cave}\}$$

| $y$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $P(Y = y)$ | 0.4 | 0.2 | 0.15 | 0.1 | 0.1 | 0.05 |

$$\sigma_Y^2 = \sum_y (y - \mu)^2 P(Y = y)$$

$$(0 - 1.45)^2(0.4) + (1 - 1.45)^2(0.2) + (2 - 1.45)^2(0.15) + (3 - 1.45)^2(0.1) + (4 - 1.45)^2(0.1) + (5 - 1.45)^2(0.05)$$

$$= 2.4475$$

$$\sigma_Y = \sqrt{2.4475} \approx 1.564$$

---

## Continuous Random Variables

Continuous random variables are problematic for introductory courses in statistics. In order to properly work with them students either need a solid grasp of calculus or elementary programming skills. Neither are really the focus of this textbook but we'll be seeing some trivial examples of each to help with the concepts.

**Continuous random variable**: The **support** consists of all numbers in an **interval** of the **real number line** and are *uncountable infinite*.

It's easier to define continuous random variables because the "rule of thumb" is just an outright rule: if we can continue to subdivide the partial values of a variable *endlessly*, then it's a continuous random variable.

$$\text{Let } \ W = \{\text{the weight of a cow}\}$$

$$S_W = (0, \infty)$$

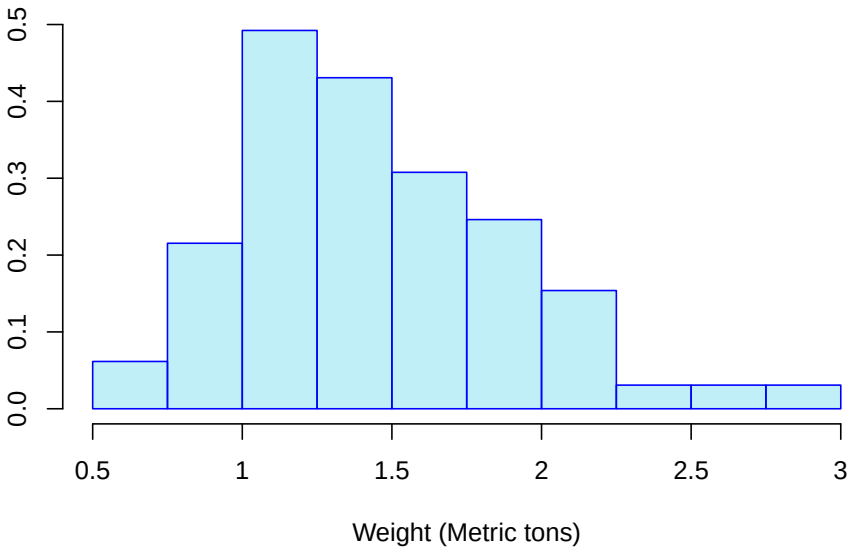
We should take a moment to discuss how we describe the support of continuous random variables. If we wanted to use the discrete method we would quickly realize how ridiculous continuous supports are.

$$S_W = \{0.0000001, 0.0000002, 0.0000003, ...\}$$

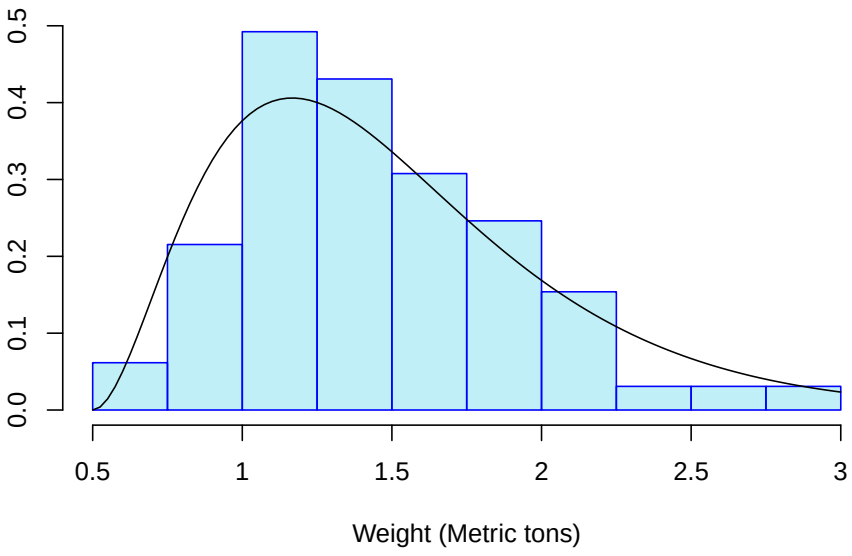
This isn't very informative *and* it's tedious. Instead we can use brackets, [ ], to state that the support *includes* that value and parenthesis, ( ), to state that the support *approaches* that value. In the case of weighing cows it would be ridiculous to have a cow with near-zero weight, but not impossible. However a cow with a zero weight isn't a cow. We also can't claim that any cows would actually **be infinite weight** as that's both biologically and mathematically impossible.

It should also create some confusion to measure cows with an unspecified unit of weight. This is intentional since *any unit* we use to measure these cows is hypothetically continuous since their weight is continuous. However we could "discretize" the weight measurements by placing them into the *classes* or *bins* of a histogram.

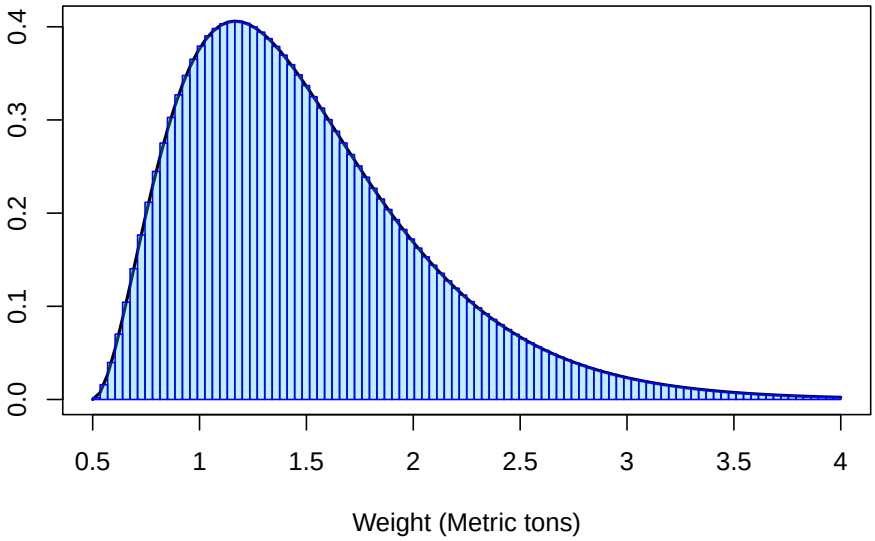
Say we're measuring the cows by metric tons. We place all of our measurements into a histogram where the classes are a quarter of a ton each.



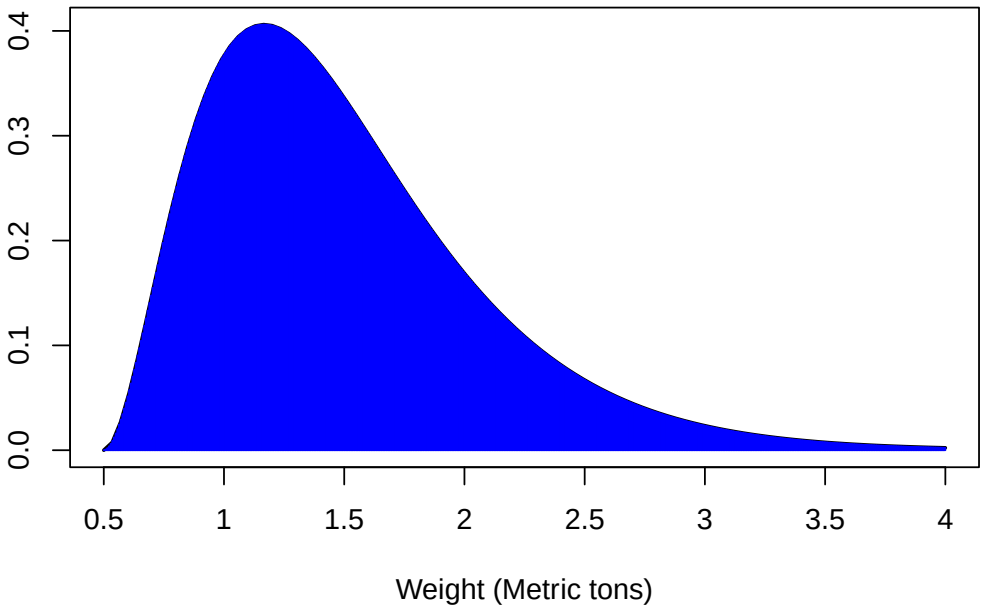
We can see how it's possible to draw a curve across our histogram to represent the **general trend** of the data:




If we increase the data we collect, this histogram could get more detailed, the bars could narrow. The more data we collect, the closer we would approach to the true **distribution** of the probabilities for our outcomes.



Inevitably we converge to a complete, smoothed curve:




This curve is used to describe the distribution of a continuous random variable. We refer to it as a **probability density curve** and it tells us what **proportion** of the population falls within any given interval.

By definition, the **area under the curve** between any two values $a$ and $b$ represents the probability that *random variable* $X$ takes a value **between** $a$ and $b$.

Those with basic knowledge of calculus can probably see what's happened here. The process of drawing rectangles below a curve to estimate it's area is known as a *Reimann sum* and the concept of slimming down those rectangles to an *infinitesimal* length is the basis behind an *integral*.

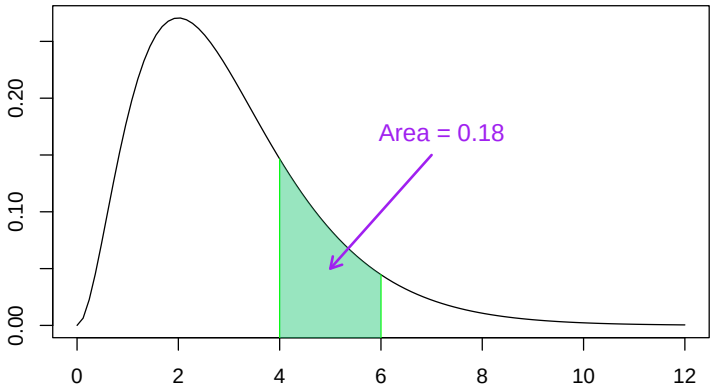
The methods in this book focus on circumventing integration wherever possible for two reasons:

- Integral calculus is far outside the level of mathematics background this book assumes.
- Applied statistics is riddled with "non-analytically tractable" integrals, meaning they cannot be solved by hand.

It's more beneficial for students to learn the alternative methods, especially since calculus is best taught and practiced by mathematicians.

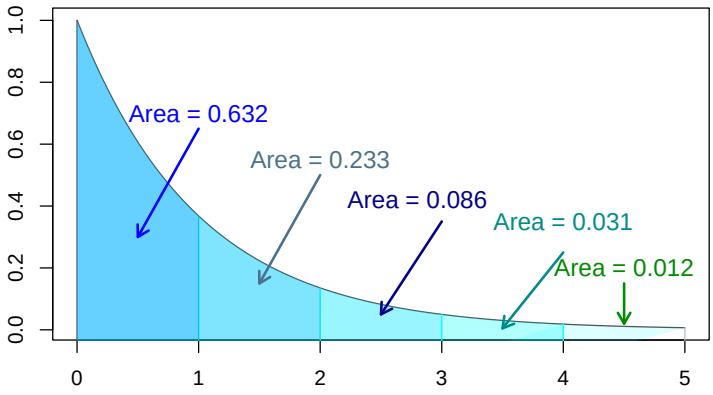
The easiest way we can simplify the process of computing continuous probabilities is for me to just do the work for you.

Probability density curves focus on *area* so when thinking about the probability of a random variable being realized between two values we can *visualize* that area with its associated probability:



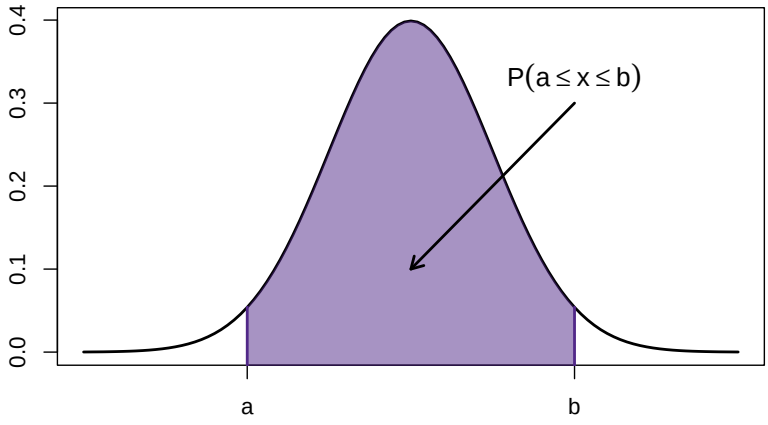

In this case the probability that the random variable, (shown by the density curve above), takes on a value between 4 and 6 is 0.18. It's key to recognize the lack of differentiation here between area, proportions, and probability. These are all the same thing in the scheme of probability density curves.

The method I've used to fill in this curve is a form of numerical integration called "Simpson's rule". All numerical methods are *approximations* meaning that they'll never achieve perfect accuracy. The curve below is estimated the same way but you'll find it's total area isn't 1.00:




This doesn't mean that the density curve represents an illegitimate distribution, it just means that we made some rounding errors while trying to estimate the areas in the first place. A good gut check on approximations is to ask yourself how important the difference in estimation is. Here the difference is 0.006, in some cases this could be significant, but in this situation we can round up to 1.00 without losing any sleep.

## Continuous Probability Distributions




For continuous random variables, probability is now "area under the curve". So only **intervals** will have *non-zero* probability; any single value will have a probability of **zero**.

$$P(X = a) = 0, \text{ for any single number } a$$

$$P(X = b) = 0, \text{ for any single number } b$$

Conceptually this may not register well at first. We'll walk through a few explanations as to why a continuous random variable has no chance of realizing to a specific value.

At every carnival there's a person on a small stage, dressed in a red and white pinstripe suit, claiming to be able to guess anyone's age. The game is simple: You pay a small entry fee then the carny sizes you up and makes their guess. If they're right within a small interval then you get nothing and they move on to the next person. If they're wrong then you get a prize based off of **how wrong** they are.

The problem is that they're right **a lot**. To be fair the range a lot of them set for guessing is huge, $3 - 5$ years give or take. But even if they tighten that range to 0 they would still guess right a surprising amount of times. That's because the way they're guessing is **discrete** so given any person's physical appearance they only have a small range of possible outcomes. The game would be almost **impossible** if they had to guess your age in *days*. It would be completely impossible if they had to guess your age in *minutes*. This is because:

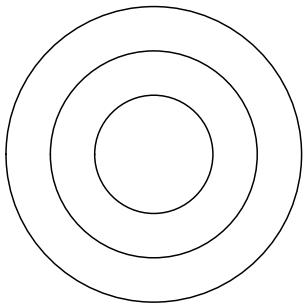
a. You would age at least a minute while sitting there waiting for them to make the guess.

b. There's far too many possible outcomes in measuring age with minutes.

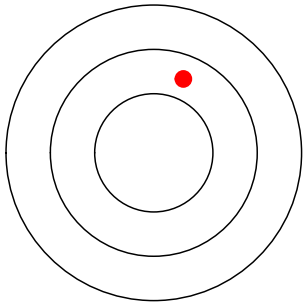
c. Nobody knows their age in minutes off the top of their head.

This is the general idea behind continuous variables having zero probability for specific values. There are far too many possibilities and there's always an increasing level of accuracy that the value can be taken to. No one person is exactly 6 million minutes old for a perceptible duration of time, we could keep measuring their age down to Planck time ($10^{-43}$ seconds) and then we could keep adding zeros.

Another example that's often used in mathematical statistics courses is the dart board.

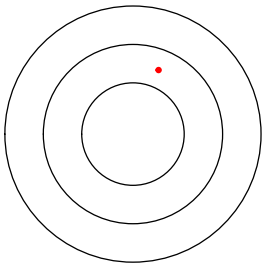


You throw a dart at the dart board with no particular target in mind.

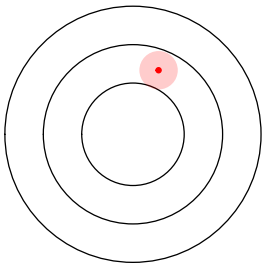


You're now given the challenge to hit the *exact* point on that dart board you did previously.

This is, of course, impossibly difficult— especially when we consider that the "point" on this board is wildly out of scale to where it would actually be.



If the challenge were to hit an *interval* around the original dart it would still be quite difficult, but it would be *possible*.



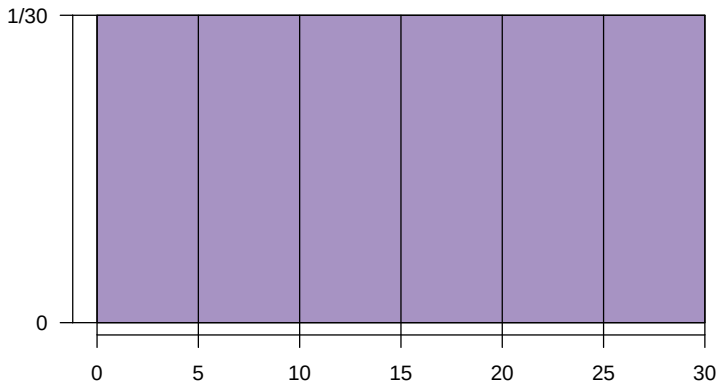
This feature of continuous distributions doesn't bar us from attempting to yield accurate estimates since we can always tighten the interval of interest. But we won't concern ourselves with exact outcomes, which is comforting in a sense.

As a result of this is that continuous distributions also don't distinguish between $\leq$ and $<$. That's to say:

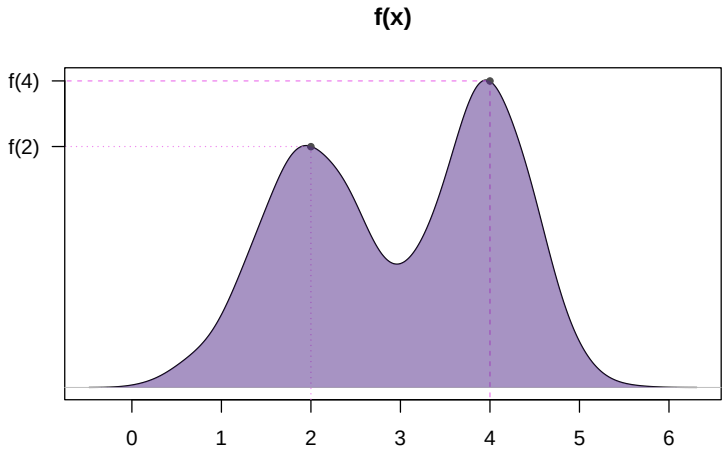
$$P(X \leq 1) = P(X < 1)$$

Since $\leq$ assumes that $X$ can be equal to some value, and $P(X = x) = 0$, it doesn't serve any purpose.

*Probability density curve* just refers to the graphical description of a **continuous** distribution. This doesn't inherently mean they're going to be curves. If every possible value of $X$ is equally likely then it takes on a **uniform distribution**. Hilariously, the "curve" for this distribution is a horizontal bar:




The curve used to describe the probability distribution of a continuous random variable is called a probability density curve, which is dictated by a function, $f(x)$, called the **probability density function** (PDF).




Probability density functions are key to understanding the behavior of continuous random variables because they act as models, like a model airplane, for the real phenomena that the random variable describes. Discrete random variables have their own version of this called the **probability mass function** (PMF) but this is just a difference in vocabulary— these functions accomplish the same task.

For a PDF to be considered legitimate it must integrate to 1, which is a fancy way of saying that the total probability has to be 1.00. Since integrals accomplish the same general task as sums we can think of this the same way we did with discrete random variables: all of the probabilities need to add up to 1.00.

The mean, variance, and standard deviation have the same interpretation for continuous variables as they do for discrete. The reason why this is a footnote rather than a detailed discussion of their calculation is because their formulas require calculus:

$$\mathbb{E}X = \int_{-\infty}^{\infty} x \ f(x)dx$$

$$\mathbb{E}X^2 = \int_{-\infty}^{\infty} x^2 \ f(x)dx$$

$$\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2$$

$$\sigma_X = \sqrt{\mathbb{V}X}$$

These calculations aren't inherently difficult, although there are plenty impossible ones, but they are *often* a lengthy exercise.

From a programming perspective they're quite easy to work with. Numerical integration methods, like Reimann sums or quadrature, can calculate these values very quickly. That said there's not much reason to do this. Computers capable of solving even simple mathematical problems have only been in use for a couple decades. Statistics has existed as a field of study much longer than this.

Since early statisticians couldn't use a MacBook Pro they developed their own methods for working around this mess of calculus. These methods evolved into their own field of study, one that's become so fundamental to the science of statistics that it's often hard for students to realize that it's not the **entirety of statistics**.

This field is known as "Distribution Theory" and it is (in this author's opinion) the second most impactful contribution that statisticians have made to the world.