3 Expectation and Variance

If we have a random variable X on some sample space Ω, we might need to be aware of some properties of X in order to set up an experiment or investigate a problem. Two such properties we might need to measure are the "average" of X (that is, if we repeat X many times and we average the values, what will we obtain?) and the "centrality" of X (that is, how far away from the "average" can values of X be found?). In this section, we will discover methods of measuring both of these properties.

3.1 Expectation

The "average" of X is more commonly known as the **expectation** or **expected value**, and it is denoted by the symbol "$\mathbb{E}$". The expectation of a random variable is the weighted average of the values produced by the random variable on each outcome, where each weight is determined by the probability of that outcome.

Definition 21 (Expectation of a random variable)**.** The expectation of a random variable $X : \Omega \to A$ on a sample space Ω is

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} \mathbb{P}[\omega]\, X[\omega].$$

Why do we require a weighted average instead of a plain average? Consider the following example.

Example 22. Assume we have two dice: die d_1 is a fair six-sided die, and die d_2 is a biased six-sided die where $\mathbb{P}[\{\boxdot\}] = 0.75$ and $\mathbb{P}[\{\boxdot\}] = \cdots = \mathbb{P}[\{\boxdot\}] = 0.05$ (that is, rolling a value between 2 and 6 occurs with probability 0.05 for each value). Let X_1 and X_2 denote the random variables associated with rolling die d_1 and die d_2, respectively. The average value we would expect to roll with die d_1 is

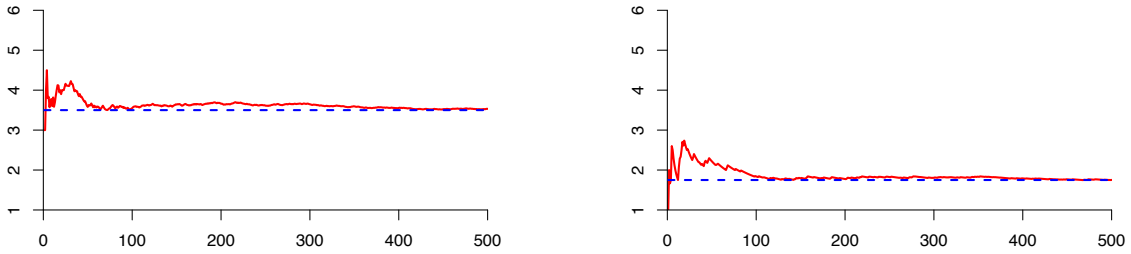
$$\mathbb{E}[X_1] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \cdots + \frac{1}{6} \cdot 6 = \frac{21}{6} = \frac{7}{2} = 3.5,$$

since each of the sides has an equal probability of appearing. With die d_2, however, we cannot say that the average value we would expect to roll is also 3.5. Even though die d_2 has the same number and value of sides as die d_1, we expect to roll a 1 with die d_2 far more often than any other number. Therefore, we must weight the outcome of rolling a 1 accordingly. Doing so gives us an average value of

$$\mathbb{E}[X_2] = \frac{75}{100} \cdot 1 + \frac{5}{100} \cdot 2 + \cdots + \frac{5}{100} \cdot 6 = \frac{7}{4} = 1.75,$$

which is a much more reasonable average value when we consider that the number of 1s we roll with die d_2 will skew the average closer to 1.

As an illustration of expectation, consider the following figures. The plot on the left shows the cumulative average of 500 randomized rolls of the fair die d_1, and the plot on the right shows the cumulative average of 500 randomized rolls of the biased die d_2. In both plots, the expectation of the experiment is denoted by a dashed line, and we can see that each of the cumulative averages converge to the expectation as we perform more trials.



By Definition 21, the expectation of a random variable is just a sum over each outcome in the sample space Ω. This definition works well for small sample spaces, like Ω_C for coin flips or Ω_D for die rolls. For larger sample spaces, however, it can be inconvenient to deal with a huge sum consisting of many terms. Luckily,

there is an alternative formula to calculate expectation that is written in terms of the values produced by the random variable X rather than the individual outcomes in Ω.

Theorem 23. *Let $X : \Omega \to A$ be a random variable. Then*

$$\mathbb{E}[X] = \sum_{a \in X[\Omega]} \mathbb{P}[X = a]\, a.$$

Proof. We know from Definition 21 that the expectation of X is $\mathbb{E}[X] = \sum_{\omega \in \Omega} \mathbb{P}[\omega]\, X[\omega]$. Rewriting this expression to collect all terms resulting in $X[\omega] = a$ for some value a, we get that

$$\mathbb{E}[X] = \sum_{a \in X[\Omega]} \left(\sum_{\substack{\omega \in \Omega \\ X[\omega] = a}} \mathbb{P}[\omega]\, X[\omega] \right).$$

Since $X[\omega] = a$ for all outcomes ω in the inner sum, we can replace $X[\omega]$ by a directly within the inner sum. By the distributivity property of multiplication, we can then move a outside of the inner sum to obtain

$$\mathbb{E}[X] = \sum_{a \in X[\Omega]} \left(a \sum_{\substack{\omega \in \Omega \\ X[\omega] = a}} \mathbb{P}[\omega] \right).$$

The new inner sum, $\sum_{\substack{\omega \in \Omega \\ X[\omega] = a}} \mathbb{P}[\omega]$, is equal to $\mathbb{P}[X = a]$. Substituting this into our formula, we get

$$\mathbb{E}[X] = \sum_{a \in X[\Omega]} a\, \mathbb{P}[X = a].$$ □

3.2 Properties of Expectations

Just like with events, we can define what it means for random variables to be independent. Given two random variables X and Y on some sample space Ω, we say that X and Y are independent if, for all values a_1 and a_2, the probability that $X[\omega] = a_1$ does not alter the probability that $Y[\omega] = a_2$ and vice versa. In other terms, X and Y are independent if

$$\mathbb{P}[X = a_1 \text{ and } Y = a_2] = \mathbb{P}[X = a_1] \times \mathbb{P}[Y = a_2].$$

Observe that the definition of independence for random variables is nearly identical to the definition of independence for events. Using the notion of independent random variables, we can prove our first property of expectations. Given two random variables X and Y, what can we say about the expectation of their product, $X \times Y$? If we know that X and Y are independent, then the following property gives us the answer.

Theorem 24. *Let X and Y be independent random variables on some sample space Ω. Then*

$$\mathbb{E}[X \times Y] = \mathbb{E}[X] \times \mathbb{E}[Y].$$

Proof. We have that

$$\mathbb{E}[X \times Y] = \sum_{a_1, a_2 \in (X \times Y)[\Omega]} \mathbb{P}[X = a_1 \text{ and } Y = a_2]\, a_1 a_2$$

by our alternate formulation of expectation given in Theorem 23. Since X and Y are independent, we can rewrite this expression as

$$\mathbb{E}[X \times Y] = \sum_{a_1, a_2 \in (X \times Y)[\Omega]} \left(\mathbb{P}[X = a_1] \times \mathbb{P}[Y = a_2]\right) a_1 a_2.$$

By splitting the sum and rearranging terms, we get

$$\begin{aligned}\mathbb{E}[X \times Y] &= \sum_{a_1 \in X[\Omega]} \left(\sum_{a_2 \in Y[\Omega]} \mathbb{P}[X = a_1]\, a_1 \times \mathbb{P}[Y = a_2]\, a_2 \right) \\ &= \sum_{a_1 \in X[\Omega]} \mathbb{P}[X = a_1]\, a_1 \sum_{a_2 \in Y[\Omega]} \mathbb{P}[Y = a_2]\, a_2 \\ &= \mathbb{E}[X] \times \mathbb{E}[Y]. \end{aligned}$$

□

Note that Theorem 24 does not hold if X and Y are dependent. To find the expectation of the product of dependent random variables X and Y, we require a slightly different formula which we will not discuss here.

Remark. Be careful: the converse of Theorem 24 does not necessarily hold. Even if we know that $\mathbb{E}[X \times Y] = \mathbb{E}[X] \times \mathbb{E}[Y]$, we cannot conclude that X and Y are independent.

Next, let's consider multiplication not by a random variable, but by a constant. Recall that the expectation of a random variable is essentially just a sum. If we introduce a multiplicative constant into the sum, we can imagine this process either as scaling each value taken by the random variable by some constant factor or, by the following property, as scaling the expectation itself by the same constant factor.

Theorem 25. *Let X be a random variable on some sample space Ω, and let c be a real number. Then*

$$\mathbb{E}[cX] = c\,\mathbb{E}[X].$$

Proof. We have that

$$\begin{aligned}\mathbb{E}[cX] &= \sum_{\omega \in \Omega} \mathbb{P}[\omega]\, (cX)[\omega] \\ &= \sum_{\omega \in \Omega} \mathbb{P}[\omega]\, c(X[\omega]) \\ &= c \sum_{\omega \in \Omega} \mathbb{P}[\omega]\, X[\omega] \\ &= c\,\mathbb{E}[X]. \end{aligned}$$

□

We can alternatively frame Theorem 25 as a special case of Theorem 24, since the expectation of a constant value c is the value itself ($\mathbb{E}[c] = c$) and since $\mathbb{E}[c]$ and $\mathbb{E}[X]$ are independent.

Our final property deals with addition instead of multiplication. Given two random variables X and Y, what can we say about the expectation of their sum, $X + Y$? Here, unlike in Theorem 24, we do not require X and Y to be independent.

Theorem 26 (Linearity of expectations)**.** *Let X and Y be random variables on some sample space Ω. Then*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Proof. We have that

$$\begin{aligned}\mathbb{E}[X + Y] &= \sum_{\omega \in \Omega} \mathbb{P}[\omega]\, (X + Y)[\omega] \\ &= \sum_{\omega \in \Omega} \mathbb{P}[\omega]\ (X[\omega] + Y[\omega]) \\ &= \sum_{\omega \in \Omega} (\mathbb{P}[\omega]\, X[\omega] + \mathbb{P}[\omega]\, Y[\omega]) \\ &= \sum_{\omega \in \Omega} \mathbb{P}[\omega]\, X[\omega] + \sum_{\omega \in \Omega} \mathbb{P}[\omega]\, Y[\omega] \\ &= \mathbb{E}[X] + \mathbb{E}[Y]. \end{aligned}$$

□

More generally, by combining the previous two properties and considering m random variables instead of two random variables, we have the following result.

Corollary 27. *Let $X_1, X_2, \ldots, X_m$ be random variables on some sample space Ω, and let $c_1, c_2, \ldots, c_m$ be real numbers. Then*

$$\mathbb{E}[c_1X_1 + c_2X_2 + \cdots + c_mX_m] = c_1\,\mathbb{E}[X_1] + c_2\,\mathbb{E}[X_2] + \cdots + c_m\,\mathbb{E}[X_m].$$

Proof. Follows from Theorems 25 and 26. □

3.3 Variance

The "centrality" of X is a measure of how "spread out" values of X are relative to the expectation of X. What does it mean for values to be "spread out"? To be sure, we need to formalize the notion of "spreading out", but we will begin our discussion by considering an example.

Let's consider an experiment where we roll pairs of fair dice and take the sum of the values on each face. There are 36 outcomes in the sample space $\Omega = (\Omega_D \times \Omega_D)$; one outcome for each combination of two dice. However, there are only 11 possible sums ($1 + 1 = 2$ through $6 + 6 = 12$), so by the pigeonhole principle, at least one sum must correspond to more than one outcome. Indeed, we have the following sums and corresponding outcomes:

2: ⚀⚀ (1/36)

3: ⚀⚁ ⚁⚀ (2/36)

4: ⚀⚂ ⚁⚁ ⚂⚀ (3/36)

5: ⚀⚃ ⚁⚂ ⚂⚁ ⚃⚀ (4/36)

6: ⚀⚄ ⚁⚃ ⚂⚂ ⚃⚁ ⚄⚀ (5/36)

7: ⚀⚅ ⚁⚄ ⚂⚃ ⚃⚂ ⚄⚁ ⚅⚀ (6/36)

8: ⚁⚅ ⚂⚄ ⚃⚃ ⚄⚂ ⚅⚁ (5/36)

9: ⚂⚅ ⚃⚄ ⚄⚃ ⚅⚂ (4/36)

10: ⚃⚅ ⚄⚄ ⚅⚃ (3/36)

11: ⚄⚅ ⚅⚄ (2/36)

12: ⚅⚅ (1/36)

As we can see, the probability of rolling a 7 (which is, not coincidentally, the expectation of the sum of two die rolls) is higher than the probability of rolling a value closer to either extreme. Because of the probability difference making it much less likely to roll an extreme value like 2 or 12, the outcomes are more concentrated around the expectation.

To define "spreading out"—or **variance**, as it is formally known—in a mathematical sense, we must figure out a way to model the distance of an outcome from the expectation. Let $\mu = \mathbb{E}[X]$, where X is the random variable corresponding to our experiment. What we want to do is determine how far each value $X[\omega]$ is from μ, for every outcome ω. Moreover, we want to weight each outcome according to its probability, in order to account for extreme values appearing less frequently (as we saw in our example).

It might seem that the correct way to measure this distance is by calculating the probability-weighted difference between each $X[\omega]$ and μ, and then summing all of the differences. However, if we follow this approach, we quickly run into a problem, for

$$\sum_{\omega\in\Omega} \mathbb{P}[\omega]\,(X[\omega] - \mu) = \left(\sum_{\omega\in\Omega} \mathbb{P}[\omega]\,X[\omega]\right) - \left(\sum_{\omega\in\Omega} \mathbb{P}[\omega]\,\mu\right),$$

making the first sum equal to $\mathbb{E}[X]$ and the second sum equal to μ. Since $\mu = \mathbb{E}[X]$ by definition, the difference is always zero. Not very helpful!

How can we overcome this cancellation issue while still maintaining the spirit of the equation by taking differences? Instead of taking the difference by itself, we could take the squared difference. This maintains the spirit of the equation while also avoiding the cancellation issue; squaring each difference means that we only add positive terms to our sum.

Remark. If we care about positive terms, then why do we square the differences instead of, say, taking absolute values? Although taking absolute values would seem to give the same result, squaring the differences preserves a number of properties that are quite useful in other areas of probability theory; for instance, squaring the differences preserves differentiability, while taking the absolute value does not. We won't get into any situations in this lecture where squaring the difference matters, but it's good to know why we would want to do so in the first place.

Taking the squared difference between $X[\omega]$ and μ leads to our formal definition of variance.

Definition 28 (Variance of a random variable)**.** The variance of a random variable $X : \Omega \to A$ on a sample space Ω is

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X-\mu)^2] \\ &= \sum_{\omega \in \Omega} \mathbb{P}[\omega]\,(X[\omega]-\mu)^2,\end{aligned}$$

where $\mu = \mathbb{E}[X]$.

Immediately from our definition, we get a handy formulation of variance exclusively in terms of expectation.

Theorem 29. *Let X be a random variable on some sample space Ω. Then*

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Proof. Let $\mu = \mathbb{E}[X]$. By Definition 28, we have that

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X-\mu)^2] \\ &= \mathbb{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + E[\mu^2] \\ &= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

□

Returning to our opening example, let's calculate the variance of our experiment of summing two dice.

Example 30. Assume we are conducting an experiment where we roll two dice and take the sum of the values on each face. Let X be the random variable mapping pairs of dice rolls to the set $[2, 12]$.

What is $\mathbb{V}[X]$? Since our two die rolls are independent of one another, we can break this problem down into two parts: let X_1 be the random variable corresponding to the first die roll, and let X_2 be the random variable corresponding to the second die roll. Just like with events and with expectations, we can calculate the variance of two independent random variables by simply summing the individual variances; that is,

$$\mathbb{V}[X] = \mathbb{V}[X_1] + \mathbb{V}[X_2].$$

Since the variance of a single die roll is

$$\begin{aligned}\mathbb{V}[X_1] &= \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 \\ &= \frac{1^2 + 2^2 + \cdots + 6^2}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12},\end{aligned}$$

and since $\mathbb{V}[X_1] = \mathbb{V}[X_2]$, we have that the variance of two die rolls is $\mathbb{V}[X] = (35/12) + (35/12) = 35/6$.

4 Probability Distributions

Thinking back to the first section of these notes, recall that we observed it is possible to use probability measures to define **probability distributions**. A probability distribution is an assignment of probabilities to possible values that a random variable can take. Formally speaking, a (discrete) probability distribution $P : X[\Omega] \to [0, 1]$ on a random variable X on some sample space Ω is a function mapping elements of the **value space** $X[\Omega]$ to a real number, where for any value $a \in X[\Omega]$,

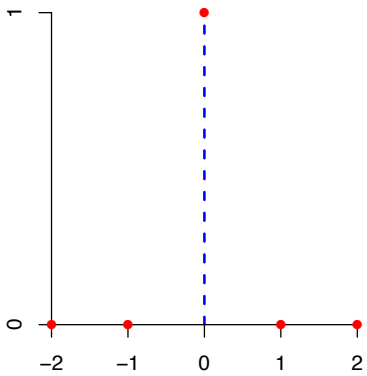
$$P(a) = \mathbb{P}[X = a] = \sum_{\omega \in \Omega} \{\mathbb{P}[\omega] \mid X[\omega] = a\}.$$

Before we continue, we require one small bit of terminology. We say that the **support** of a probability distribution P consists of all values $a \in X[\Omega]$ that have a nonzero probability. In other terms, for a random variable X and a probability distribution P, the support of P on X is the subset of values from the value space that X can take.

With these notions defined, we can consider a few examples of common (discrete) probability distributions.

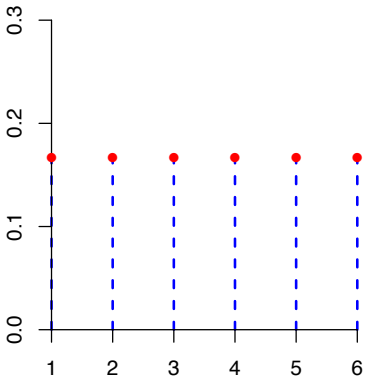
The simplest probability distribution that exists is the **degenerate** probability distribution, also known as the trivial probability distribution. In this distribution, a random variable X always takes a single value a with probability 1. Even though this doesn't seem like random behaviour—indeed, it is entirely deterministic—it still meets the definition of a random variable.

At right, we have an illustration of the degenerate probability distribution with support $\{0\}$. Here, we can see that $\mathbb{P}[X = 0] = 1$ and $\mathbb{P}[X \neq 0] = 0$.



We have already seen the **uniform** probability distribution earlier in these notes, when we discussed flipping fair coins and rolling fair dice. In this probability distribution, the support comes from a finite set of values $[a, b]$, and a random variable X has equal probability of taking each value.

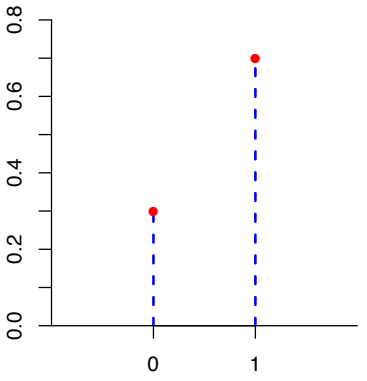
At right, we have an illustration of the uniform probability distribution; specifically, the distribution for our familiar example of rolling a fair die. Each of the six values that can be taken by the "die-roll" random variable X occur with an equal probability of $1/6$.



The **Bernoulli** probability distribution is our first example of a non-degenerate probability distribution where different values appear with different probabilities. In this distribution, a random variable X takes value 1 with some probability p and value 0 with the complementary probability $q = 1 - p$.

We can model this probability distribution by flipping an unfair coin. We can also think of this probability distribution as modelling the odds of "success" versus "failure".

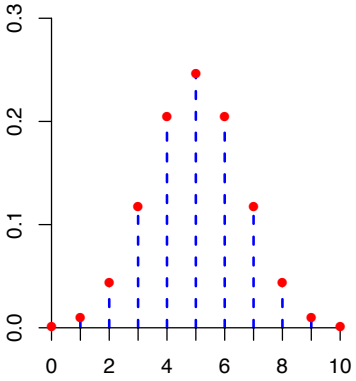
At right, we have an illustration of the Bernoulli probability distribution where $p = 0.7$ and $q = (1 - 0.7) = 0.3$.



If we take the sum of n independent Bernoulli distributions, each with the same probability p, we get the **binomial** probability distribution. This distribution gives the probability of obtaining a total of k successes out of n trials, where each trial succeeds with probability p.

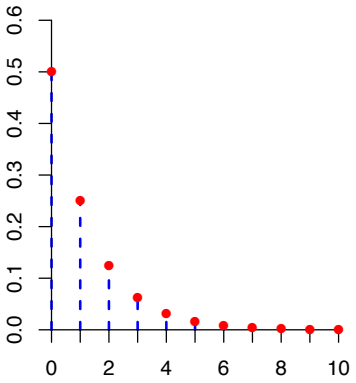
When $n = 1$, the binomial probability distribution becomes the Bernoulli probability distribution.

At right, we have an illustration of the binomial probability distribution where $n = 10$ and $p = 0.5$.



The **geometric** probability distribution is one of two probability distributions we will see that has infinite support (namely, $\mathbb{N}$). This distribution gives the probability that k independent Bernoulli trials will fail before we obtain the first success, where each trial succeeds with probability p. This distribution has infinite support because we could theoretically measure an infinite number of trials.

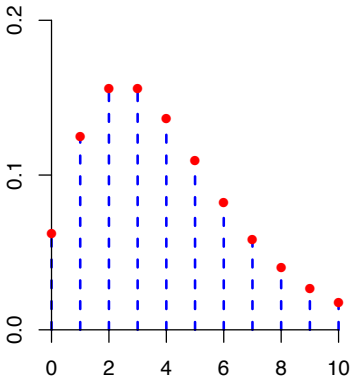
At right, we have an illustration of the geometric probability distribution where $p = 0.5$. Only the first ten trials are shown.



If we generalize the geometric probability distribution, we get the **negative binomial** probability distribution. This distribution gives the probability that k independent Bernoulli trials will fail before we obtain the nth success, where each trial succeeds with probability p. Again, this distribution has infinite support (namely, $\mathbb{N}$).

When $n = 1$, the negative binomial probability distribution becomes the geometric probability distribution.

At right, we have an illustration of the negative binomial probability distribution where $n = 4$ and $p = 0.5$. Only the first ten trials are shown.



There are, of course, many other probability distributions available to use; the six distributions we have seen here were chosen because they are among the most commonly used. The fine details of each of the probability distributions discussed in this section are summarized in the following table. As an exercise, try deriving each of the values of $\mathbb{P}[X = \omega]$, $\mathbb{E}[X]$, and $\mathbb{V}[X]$ given in the table from the definitions of each probability distribution.

Distribution	**Given**	**Support**	$\mathbb{P}[\boldsymbol{X = \omega}]$	$\mathbb{E}[\boldsymbol{X}]$	$\mathbb{V}[\boldsymbol{X}]$
Degenerate	—	$\{a\}$	1 if $\omega = a$	a	0
Uniform	—	$[a, b]$	$1 / (b - a)$ if $\omega \in [a, b]$	$(a + b) / 2$	$(b - a)^2 / 12$
Bernoulli	$0 \leq p \leq 1$	$\{0, 1\}$	p if $\omega = 1$, $(1 - p)$ if $\omega = 0$	p	$p(1 - p)$
Binomial	$0 \leq p \leq 1$, $n \geq 0$	$[0, n]$	$\binom{n}{k} p^k (1 - p)^{n-k}$	np	$np(1 - p)$
Geometric	$0 < p \leq 1$	$\mathbb{N}$	$(1 - p)^k p$	$(1 - p) / p$	$(1 - p) / p^2$
Neg. Binomial	$0 < p \leq 1$, $n \geq 0$	$\mathbb{N}$	$\binom{n+k-1}{n} (1 - p)^k p^n$	$n(1 - p) / p$	$n(1 - p) / p^2$