# *Basic Probability*

## *Introduction*

The world is an uncertain place. Making predictions about something as seemingly mundane as tomorrow's weather, for example, is actually quite a difficult task. Even with the most advanced computers and models of the modern era, weather forecasters still cannot say with absolute certainty whether it will rain tomorrow. The best they can do is to report their best estimate of the *chance* that it will rain tomorrow. For example, if the forecasters are fairly confident that it will rain tomorrow, they might say that there is a 90% chance of rain. You have probably heard statements like this your entire life, but have you ever asked yourself what exactly it means to say that there is a 90% chance of rain?

Let us consider an even more basic example: tossing a coin. If the coin is fair, then it is just as likely to come up heads as it is to come up tails. In other words, if we were to repeatedly toss the coin many times, we would expect about about half of the tosses to be heads and and half to be tails. In this case, we say that the **probability** of getting a head is 1/2 or 0.5.

Note that when we say the probability of a head is 1/2, we are *not* claiming that any sequence of coin tosses will consist of exactly 50% heads. If we toss a fair coin ten times, it would not be surprising to observe 6 heads and 4 tails, or even 3 heads and 7 tails. But as we continue to toss the coin over and over again, we expect the long-run frequency of heads to get ever closer to 50%. In general, it is important in statistics to understand the distinction between *theoretical* and *empirical* quantities. Here, the true (theoretical) probability of a head was 1/2, but any realized (empirical) sequence of coin tosses may have more or less than exactly 50% heads. (See Figures 1 – 3.)

Now suppose instead that we were to toss an unusual coin with heads on both of its faces. Then every time we flip this coin we will observe a head — we say that the probability of a head is 1. The probability of a tail, on the other hand, is 0. Note that there is no way

Figure 1: The true probability of a head is 1/2 for a fair coin.

Figure 2: A sequence of 10 flips happened to contain 3 head. The empirical frequency of heads is thus 3/10, which is quite different from 1/2.

Figure 3: A sequence of 100 flips happened to contain 45 heads. The empirical frequency of heads is 45/100, which is much closer to 1/2.

we can further modify the coin to make flipping a head even more likely. Thus, *a probability is always a number between 0 and 1 inclusive.*

## *First Concepts*

### *Terminology*

When we later discuss examples that are more complicated than flipping a coin, it will be useful to have an established vocabulary for working with probabilities. A probabilistic **experiment** (such as tossing a coin or rolling a die) has several components. The **sample space** is the set of all possible **outcomes** in the experiment. We usually denote the sample space by $\Omega$, the Greek capital letter "Omega." So in a coin toss experiment, the sample space is

$$\Omega = \{\mathsf{H}, \mathsf{T}\},$$

since there are only two possible outcomes: heads ($\mathsf{H}$) or tails ($\mathsf{T}$). Different experiments have different sample spaces. So if we instead consider an experiment in which we roll a standard six-sided die, the sample space is

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Collections of outcomes in the sample space $\Omega$ are called **events**, and we often use capital Roman letters to denote these collections. We might be interested in the event that we roll an even number, for example. If we call this event $E$, then

$$E = \{2, 4, 6\}.$$

Any subset of $\Omega$ is a valid event. In particular, one-element subsets are allowed, so we can speak of the event $F$ of rolling a 4, $F = \{4\}$.

### *Assigning probabilities to dice rolls and coin flips*

In a random experiment, every event gets assigned a probability. Notationally, if $A$ is some event of interest, then $P(A)$ is the probability that $A$ occurs. The probabilities in an experiment are not arbitrary; they must satisfy a set of rules or **axioms**. We first require that *all probabilities be nonnegative.* In other words, in an experiment with sample space $\Omega$, it must be the case that

$$P(A) \geq 0 \tag{1}$$

for any event $A \subseteq \Omega$. This should make sense given that we've already said that a probability of 0 is assigned to an impossible event,

and there is no way for something to be less likely than something that is impossible!

The next axiom is that *the sum of the probabilities of all the outcomes in $\Omega$ must be 1.* We can restate this requirement by the equation

$$\sum_{\omega \in \Omega} P(\omega) = 1. \tag{2}$$

This rule can sometimes be used to deduce the probability of an outcome in certain experiments. Consider an experiment in which we roll a fair die, for example. Then each outcome (i.e. each face of the die) is equally likely. That is,

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = a,$$

for some number $a$. Equation (2) now allows us to conclude

$$1 = \sum_{k=1}^{6} P(k) = \sum_{k=1}^{6} a = 6a,$$

so $a = 1/6$. In this example, we were able to use the symmetry of the experiment along with one of the probability axioms to determine the probability of rolling any number.

Once we know the probabilties of the outcomes in an experiment, we can compute the probability of any event. This is because *the probability of an event is the sum of the probabilities of the outcomes it comprises.* In other words, for an event $A \subseteq \Omega$, the probability of $A$ is

$$P(A) = \sum_{\omega \in A} P(\omega). \tag{3}$$

To illustrate this equation, let us find the probability of rolling an even number, an event which we will denote by $E$. Since $E = \{2, 4, 6\}$, we simply add the probabilities of these three outcomes to obtain

$$\begin{aligned} P(E) &= \sum_{\omega \in E} P(\omega) \\ &= P(2) + P(4) + P(6) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{2}. \end{aligned}$$

What is the probability that we get at least one `H`?

*Solution.* One way to solve this problem is to add up the probabilities

of all outcomes that have at least one H. We would get

$$\begin{aligned} P(\text{flip at least one H}) &= P(\text{HH}) + P(\text{HT}) + P(\text{TH}) \\ &= p^2 + p \cdot (1-p) + (1-p) \cdot p \\ &= p^2 + 2 \cdot (p - p^2) \\ &= 2p - p^2 \\ &= p \cdot (2-p). \end{aligned}$$

Another way to do this is to find the probability that we **don't** flip at least one H, and subtract that probability from 1. This would give us the probability that we **do** flip at least one H.

The only outcome in which we don't flip at least one H is if we flip T both times. We would then compute

$$P(\text{don't flip at least one H}) = P(\text{TT}) = (1-p)^2$$

Then to get the **complement** of this event, i.e. the event where we **do** flip at least one H, we subtract the above probability from 1. This gives us

$$\begin{aligned} P(\text{flip at least one H}) &= 1 - P(\text{don't flip at least one H}) \\ &= 1 - (1-p)^2 \\ &= 1 - (1 - 2p + p^2) \\ &= 2p - p^2 \\ &= p \cdot (2-p). \end{aligned}$$

Wowee! Both methods for solving this problem gave the same answer. Notice that in the second calculation, we had to sum up fewer probabilities to get the answer. It can often be the case that computing the probability of the complement of an event and subtracting that from 1 to find the probability of the original event requires less work. □

*Independence*

If two events $A$ and $B$ don't influence or give any information about the other, we say $A$ and $B$ are independent. Remember that this is not the same as saying $A$ and $B$ are disjoint. If $A$ and $B$ were disjoint, then given information that $A$ happened, we would know with certainty that $B$ did *not* happen. Hence if $A$ and $B$ are disjoint they could never be independent. The mathematical statement of independent events is given below.

**Definition 0.0.1.** *Let $A$ and $B$ both be subsets of our sample space $\Omega$. Then we say $A$ and $B$ are independent if*

$$P(A \cap B) = P(A)P(B)$$

*In other words, if the probability of the intersection factors into the product of the probabilities of the individual events, they are independent.*

We haven't defined set intersection in this section, but it is defined in the set theory chapter. The $\cap$ symbol represents *A and B* happening, i.e. the intersection of the events.

**Example 0.0.1.** *Returning to our double coin flip example, our sample space was*

$$\Omega = \{HH, HT, TH, TT\}$$

*Define the events*

$$A \doteq \{\textit{first flip heads}\} = \{HH, HT\}$$
$$B \doteq \{\textit{second flip heads}\} = \{HT, TT\}$$

***Notation:*** *We write the sign $\doteq$ to represent that we are defining something. In the above expression, we are defining the arbitrary symbols A and B to represent events.*

*Intuitively, we suspect that A and B are independent events, since the first flip has no effect on the outcome of the second flip. This intuition aligns with the definition given above, as*

$$P(A \cap B) = P(\{HT\}) = \frac{1}{4}$$

*and*

$$P(A) = P(B) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

*We can verify that*

$$P(A \cap B) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A)P(B)$$

*Hence A and B are independent. This may have seemed like a silly exercise, but in later chapters, we will encounter pairs of sets where it is not intuitively clear whether or not they are independent. In these cases, we can simply verify this mathematical definition to conclude independence.*

## *Expectation*

Consider the outcome of a single die roll, and call it $X$. A reasonable question one might ask is "What is the average value of $X$?". We define this notion of "average" as a weighted sum of outcomes.

Since $X$ can take on 6 values, each with probability $\frac{1}{6}$, the weighted average of these outcomes should be

$$\begin{aligned}\text{Weighted Average} &= \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 2 + \frac{1}{6}\cdot 3 + \frac{1}{6}\cdot 4 + \frac{1}{6}\cdot 5 + \frac{1}{6}\cdot 6 \\ &= \frac{1}{6}\cdot(1+2+3+4+5+6) \\ &= \frac{21}{6} \\ &= 3.5\end{aligned}$$

This may seem dubious to some. How can the average roll be a non-integer value? The confusion lies in the interpretation of the phrase *average roll*. A more correct interpretation would be the long term average of the die rolls. Suppose we rolled the die many times, and recorded each roll. Then we took the average of all those rolls. This average would be the fraction of 1's, times 1, plus the fraction of 2's, times 2, plus the fraction of 3's, times 3, and so on. But this is exactly the computation we have done above! In the long run, the fraction of each of these outcomes is nothing but their probability, in this case, $\frac{1}{6}$ for each of the 6 outcomes.

From this very specific die rolling example, we can abstract the notion of the *average value* of a random quantity. The concept of average value is an important one in statistics, so much so that it even gets a special bold faced name. Below is the mathematical definition for the **expectation**, or average value, of a random quantity $X$.

**Definition 0.0.2.** *The **expected value**, or **expectation** of $X$, denoted by $E(X)$, is defined to be*

$$E(X) = \sum_{x \in X(\Omega)} xP(X = x)$$

This expression may look intimidating, but it is actually conveying a very simple set of instructions, the same ones we followed to compute the average value of $X$.

The $\sum$ sign means to sum over, and the indices of the items we are summing are denoted below the $\sum$ sign. The $\in$ symbol is shorthand for "contained in", so the expression below the $\sum$ is telling us to sum over all items *contained in* our sample space $\Omega$. We can think of the expression to the right of the $\sum$ sign as the actual items we are summing, in this case, the weighted contribution of each item in our sample space.

The notation $X(\Omega)$ is used to deal with the fact that $\Omega$ may not be a set of numbers, so a weighted sum of elements in $\Omega$ isn't even well defined. For instance, in the case of a coin flip, how can we compute $\mathtt{H} \cdot \frac{1}{2} + \mathtt{T} \cdot \frac{1}{2}$? We would first need to assign *numerical values* to $\mathtt{H}$ and $\mathtt{T}$ in order to compute a meaningful expected value. For a coin flip we typically make the following assignments,

$$\mathtt{T} \mapsto 0$$
$$\mathtt{H} \mapsto 1$$

So when computing an expectation, the indices that we would sum over are contained in the set

$$X(\Omega) = \{0, 1\}$$

Let's use this set of instructions to compute the expected value for a coin flip.

### *Expectation of a Coin Flip*

Now let $X$ denote the value of a coin flip with bias $p$. That is, with probability $p$ we flip $\mathtt{H}$, and in this case we say $X = 1$. Similarly, with probability $1 - p$ we flip $\mathtt{T}$, and in this case we say $X = 0$. The expected value of the random quantity $X$ is then

$$\begin{aligned} E(X) &= \sum_{x \in X(\Omega)} xP(X = x) \\ &= \sum_{x \in \{0,1\}} xP(X = x) \\ &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= 0 \cdot P(\mathtt{T}) + 1 \cdot P(\mathtt{H}) \\ &= 0 \cdot (1 - p) + 1 \cdot p \\ &= p \end{aligned}$$

So the expected value of this experiment is $p$. If we were flipping a fair coin, then $p = \frac{1}{2}$, so the average value of $X$ would be $\frac{1}{2}$.

Again, we can never get an outcome that would yield $X = \frac{1}{2}$, but this is not the interpretation of the expectation of $X$. Remember, the correct interpretation is to consider what would happen if we flipped the coin many times, obtained a sequence of 0's and 1's, and took the average of those values. We would expect around half of the flips to give 0 and the other half to give 1, giving an average value of $\frac{1}{2}$.

**Exercise 0.0.1.** *Show the following properties of expectation.*

*(a) If $X$ and $Y$ are two random variables, then*

$$E(X + Y) = E(X) + E(Y)$$

*(b) If $X$ is a random variable and $c$ is a constant, then*

$$E(cX) = cE(X)$$

*(c) If $X$ and $Y$ are independent random variables, then*

$$E[XY] = E[X]E[Y]$$

*Proof.* For now, we will take (a) and (c) as a fact, since we don't know enough to prove them yet (and we haven't even defined independence of random variables!). (b) follows directly from the definition of expectation given above. □

*Variance*

The variance of a random variable $X$ is a nonnegative number that summarizes on average how much $X$ differs from its mean, or expectation. The first expression that comes to mind is

$$X - E(X)$$

i.e. the difference between $X$ and its mean. This itself is a random variable, since even though $EX$ is just a number, $X$ is still random. Hence we would need to take an expectation to turn this expression into the average amount by which $X$ differs from its expected value. This leads us to

$$E(X - EX)$$

This is almost the definition for variance. We require that the variance always be nonnegative, so the expression inside the expectation should always be $\geq 0$. Instead of taking the expectation of the difference, we take the expectation of the squared difference.

**Definition 0.0.3.** *The **variance** of $X$, denoted by $Var(X)$ is defined*

$$Var(X) = E[(X - EX)^2]$$

Below we give and prove some useful properties of the variance.

**Proposition 0.0.1.** *If $X$ is a random variable with mean $EX$ and $c \in$ is a real number,*

*(a) $Var(X) \geq 0$.*

*(b) $Var(cX) = c^2 Var(X)$.*

*(c) $Var(X) = E(X^2) - E(X)$.*

*(d) If $X$ and $Y$ are independent random variables, then*

$$Var(X+Y) = Var(X) + Var(Y)$$

*Proof.*

(a) Since $(X - EX)^2 \geq 0$, its average is also $\geq 0$. Hence $E[(X - EX)^2] \geq 0$.

(b) Going by the definition, we have

$$\begin{aligned}\text{Var}(cX) &= E[(cX - E[cX])^2] \\ &= E[(cX - cEX)^2] \\ &= E[c^2(X - EX)^2] \\ &= c^2 E[(X - EX)^2] \\ &= c^2 \text{Var}(X)\end{aligned}$$

(c) Expanding out the square in the definition of variance gives

$$\begin{aligned}\text{Var}(X) &= E[(X - EX)^2] \\ &= E[X^2 - 2XEX + (EX)^2] \\ &= E[X^2] - E(2XEX) + E((EX)^2) \\ &= E[X^2] - 2EXEX + (EX)^2 \\ &= E[X^2] - (EX)^2\end{aligned}$$

where the third equality comes from linearity of $E$ (Exercise 2.3 (a)) and the fourth equality comes from Exercise 2.3 (b) and the fact that since $EX$ and $(EX)^2$ are constants, their expectations are just $EX$ and $(EX)^2$ respectively.

(d) By the definition of variance,

$$\begin{aligned}\text{Var}(X+Y) &= E[(X+Y)^2] - (E[X+Y])^2 \\ &= E[X^2 + 2XY + Y^2] - \left((E[X])^2 + 2E[X]E[Y] + (E[Y])^2\right) \\ &= E[X^2] - (E[X])^2 + E[Y^2] - (E[Y]^2) + 2E[XY] - 2E[X]E[Y] \\ &= E[X^2] - (E[X])^2 + E[Y^2] - (E[Y]^2) \\ &= \text{Var}(X) + \text{Var}(Y)\end{aligned}$$

where the fourth equality comes from the fact that if $X$ and $Y$ are independent, then $E[XY] = E[X]E[Y]$. Independence of random variables will be discussed in the "Random Variables" section, so don't worry if this proof doesn't make any sense to you yet.

□

**Exercise 0.0.2.** *Compute the variance of a die roll, i.e. a uniform random variable over the sample space* $\Omega = \{1, 2, 3, 4, 5, 6\}$.

*Solution.* Let $X$ denote the outcome of the die roll. By definition, the variance is

$$\begin{aligned}
\text{Var}(X) &= E[(X - EX)]^2 \\
&= E(X^2) - (EX)^2 && \text{(Proposition 2.11 (c))} \\
&= \Big( \sum_{k=1}^{6} k^2 \cdot \frac{1}{6} \Big) - (3.5)^2 && \text{(Definition of Expectation)} \\
&= \frac{1}{6} \cdot (1 + 4 + 9 + 16 + 25 + 36) - 3.5^2 \\
&= \frac{1}{6} \cdot 91 - 3.5^2 \\
&\approx 2.92
\end{aligned}$$

□

**Remark 0.0.1.** *The square root of the variance is called the* ***standard deviation****.*

### *Markov's Inequality*

Here we introduce an inequality that will be useful to us in the next section. Feel free to skip this section and return to it when you read "Chebyschev's inequality" and don't know what's going on.

Markov's inequality is a bound on the probability that a nonnegative random variable $X$ exceeds some number $a$.

**Theorem 0.0.1** (Markov's inequality)**.** *Suppose $X$ is a nonnegative random variable and $a \in$ is a positive constant. Then*

$$P(X \geq a) \leq \frac{EX}{a}$$

*Proof.* By definition of expectation, we have

$$\begin{aligned}
EX &= \sum_{k \in X(\Omega)} kP(X = k) \\
&= \sum_{k \in X(\Omega) \text{ s.t. } k \geq a} kP(X = k) + \sum_{k \in X(\Omega) \text{ s.t. } k < a} kP(X = k) \\
&\geq \sum_{k \in X(\Omega) \text{ s.t. } k \geq a} kP(X = k) \\
&\geq \sum_{k \in X(\Omega) \text{ s.t. } k \geq a} aP(X = k) \\
&= a \sum_{k \in X(\Omega) \text{ s.t. } k \geq a} P(X = k) \\
&= aP(X \geq a)
\end{aligned}$$

where the first inequality follows from the fact that $X$ is nonnegative and probabilities are nonnegative, and the second inequality follows from the fact that $k \geq a$ over the set $\{k \in X(\Omega) \text{ s.t. } k \geq a\}$.

**Notation:** "s.t." stands for "such that".

Dividing both sides by $a$, we recover

$$P(X \geq a) \leq \frac{EX}{a}$$

□

**Corollary 0.0.1** (Chebyschev's inequality)**.** *Let $X$ be a random variable. Then*

$$P(|X - EX| > \epsilon) \leq \frac{Var(X)}{\epsilon^2}$$

*Proof.* This is marked as a corollary because we simply apply Markov's inequality to the nonnegative random variable $(X - EX)^2$. We then have

$$\begin{aligned} P(|X - EX| > \epsilon) &= P((X - EX)^2 > \epsilon^2) && \text{(statements are equivalent)} \\ &\leq \frac{E[(X - EX)^2]}{\epsilon^2} && \text{(Markov's inequality)} \\ &= \frac{\text{Var}(X)}{\epsilon^2} && \text{(definition of variance)} \end{aligned}$$

□

## *Estimation*

One of the main reasons we do statistics is to make inferences about a population given data from a subset of that population. For example, suppose there are two candidates running for office. We could be interested in finding out the true proportion of the population that supports a particular political candidate. Instead of asking every single person in the country their preferred candidate, we could randomly select a couple thousand people from across the country and record their preference. We could then estimate the true proportion of the population that supports the candidate using this sample proportion. Since each person can only prefer one of two candidates, we can model this person's preference as a coin flip with bias $p =$ the true proportion that favors candidate 1.

### *Estimating the Bias of a Coin*

Suppose now that we are again flipping a coin, this time with bias $p$. In other words, our coin can be thought of as a random quantity $X$ defined

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

where 1 represents `H` and 0 represents `T`. If we were just handed this coin, and told that it has some bias $0 \leq p \leq 1$, how would we estimate $p$? One way would be to flip the coin $n$ times, count the number of heads we flipped, and divide that number by $n$. Letting $X_i$ be the outcome of the $i^{th}$ flip, our estimate, denoted $\hat{p}$, would be

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

As the number of samples $n$ gets bigger, we would expect $\hat{p}$ to get closer and closer to the true value of $p$.

### *Estimating $\pi$*

In the website's visualization, we are throwing darts uniformly at a square, and inside that square is a circle. If the side length of the square that inscribes the circle is $L$, then the radius of the circle is $R = \frac{L}{2}$, and its area is $A = \pi(\frac{L}{2})^2$. At the $i^{th}$ dart throw, we can define

$$X_i = \begin{cases} 1 & \text{if the dart lands in the circle} \\ 0 & \text{otherwise} \end{cases}$$

The event “dart lands in the circle" has probability

$$p = \frac{\text{Area of Circle}}{\text{Area of Square}} = \frac{\pi\left(\frac{L}{2}\right)^2}{L^2} = \frac{\pi}{4}$$

So with probability $p = \frac{\pi}{4}$, a dart lands in the circle, and with probability $1 - \frac{\pi}{4}$, it doesn't.

By the previous section, we can estimate $p$ using $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ so that for large enough $n$, we have

$$\hat{p} \approx p = \frac{\pi}{4}$$

so that rearranging for $\pi$ yields

$$\pi \approx 4\hat{p}$$

Hence our approximation gets closer and closer to $\pi$ as the number of sample $n \to \infty$ causing $\hat{p} \to p$.

### *Consistency of Estimators*

What exactly do we mean by “closer and closer"? In this section, we describe the concept of **consistency** in order to make precise this notion of convergence. Our estimator in the last section, $4\hat{p}$ is itself random, since it depends on the $n$ sample points we used to compute it. If we were to take a different set of $n$ sample points, we would likely get a different estimate. Despite this randomness, intuitively we believe that as the number of samples $n$ tends to infinity, the estimator $4\hat{p}$ will converge in some probabilistic sense, to $\pi$.

Another way to formulate this is to say, no matter how small a number we pick, say 0.001, we should always be able to conclude that the probability that our estimate differs from $\pi$ by more than 0.001, goes to 0 as the number of samples goes to infinity. We chose 0.001 in this example, but this notion of probabilistic convergence should hold for any positive number, no matter how small. This leads us to the following definition.

**Definition 0.0.4.** *We say an estimator $\hat{p}$ is a **consistent** estimator of $p$ if for any $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\hat{p} - p| > \epsilon) = 0.$$

Let's show that $4\hat{p}$ is a *consistent* estimator of $\pi$.

*Proof.* Choose any $\epsilon > 0$. By Chebyshev's inequality (Corollary 2.13),

$$
\begin{aligned}
P(|4\hat{p} - \pi| > \epsilon) &\leq \frac{\text{Var}(4\hat{p})}{\epsilon^2} \\
&= \frac{\text{Var}\left(4 \cdot \frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\epsilon^2} && \text{(Definition of } \hat{p}) \\
&= \frac{\frac{16}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right)}{\epsilon^2} && (\text{Var}(cY) = c^2\text{Var}(Y)) \\
&= \frac{\frac{16}{n^2}\sum_{i=1}^{n}\text{Var}(X_i)}{\epsilon^2} && (X_i\text{'s are independent}) \\
&= \frac{\frac{16}{n^2} \cdot n \cdot \text{Var}(X_1)}{\epsilon^2} && (X_i\text{'s are identically distributed}) \\
&= \frac{\frac{16}{n} \cdot p(1-p)}{\epsilon^2} && (\text{Var}(X_i) = p(1-p)) \\
&= \frac{16 \cdot \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right)}{n\epsilon^2} && \left(p = \frac{\pi}{4}\right) \\
&\to 0
\end{aligned}
$$

as $n \to \infty$. Hence we have shown that $4\hat{p}$ is a consistent estimator of $\pi$. □