This is a draft; please send me corrections and/or suggestions.

Convergence of the sample mean

Suppose that $\langle X_i \rangle_{i=1}^N$ are a sequence of i.i.d. random variables[1] representing the outcomes from some experiment repeated N times. We have discussed properties of the sample mean,

$$\overline{X} = \frac{1}{N} \sum_{i=1}^N X_i.$$

In particular, $\mathbb{E}[\overline{X}] = \mathbb{E}[X]$, and $\mathrm{Var}(\overline{X}) = \frac{1}{N}\mathrm{Var}(X)$. A related result — the weak law of large numbers (WLOLN) — says that as N grows large, $\mathrm{P}(|\overline{X} - \mathbb{E}[X]| > \varepsilon) \to 0$ for all $\varepsilon > 0$. Taken together, these results say that the more observations we make, the closer the sample mean should be to the true mean. We will extend this logic slightly for an example.

Corollary: *let $\overline{X}_N$ and $\overline{X}_{N+1}$ be the sample means of $\langle X_i \rangle_{i=1}^N$ and $\langle X_i \rangle_{i=1}^{N+1}$, respectively. Then as $N \to +\infty$,*

$$\mathit{Var}(\overline{X}_{N+1} - \overline{X}_N) \to 0.$$

Proof: this corollary says that not only are sample means coming arbitrarily close to the population mean, they are also tending arbitrarily close to one another (or rather, arbitrarily close to the one following). This is fairly intuitive: if things are converging to a particular point, then the distance between two consecutive observations should be shrinking. A formal proof is a matter of applying what we know about independent random variables.

$$\begin{aligned}
\mathrm{Var}(\overline{X}_{N+1} - \overline{X}_N) &= \mathrm{Var}\left(\frac{1}{N+1}\sum_{i=1}^{N+1} X_i - \frac{1}{N}\sum_{i=1}^N X_i\right) \\
&= \mathrm{Var}\left(\frac{N}{N(N+1)}\sum_{i=1}^N X_i - \frac{N+1}{N(N+1)}\sum_{i=1}^N X_i + \frac{1}{N+1}X_{N+1}\right) \\
&= \mathrm{Var}\left(\frac{1}{N+1}X_{N+1} - \frac{1}{N(N+1)}\sum_{i=1}^N X_i\right) \\
&= \mathrm{Var}\left(\frac{1}{N+1}X_{N+1}\right) + \sum_{i=1}^N \mathrm{Var}\left(-\frac{1}{N(N+1)}X_i\right) \\
&= \left(\frac{1}{N+1}\right)^2 \mathrm{Var}(X_{N+1}) + \sum_{i=1}^N \left(\frac{1}{N(N+1)}\right)^2 \mathrm{Var}(X_i) \\
&= \frac{\mathrm{Var}(X)}{(N+1)^2} + \frac{N\mathrm{Var}(X)}{N^2(N+1)^2} \\
&= \left(\frac{1}{N(N+1)}\right)\mathrm{Var}(X).
\end{aligned}$$

Since $\mathrm{Var}(X)$ is fixed, as $N \to +\infty$ the right-hand multiple will go to 0, so the variance of the difference will go to 0.

□

[1] When the sequence $\langle X_i \rangle_{i=1}^N$ is i.i.d., we generally drop the subscript and think of each being distributed identically to some other random variable X. This allows us to avoid having to justify "which X_i we are talking about."

An interesting consequence here is that not only does the variance of the difference go to 0, but it does so *quickly*, on order N^2. That means that with 10 observations, the variance of the difference of sample means will be only 1% of the variance of the underlying random variable! This is because not only is the variance of an individual sample mean decreasing, but also because an extra observation will have so little effect on the sample mean, since it is downweighted by $\frac{1}{N+1}$.

In section, I sampled the class to determine how long people sleep on a normal school night[2]. We collected the following data:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_i	7	7	4	6	5	6	8	7	5	5
$\overline{X}_N$	7	7	6	6	$\frac{29}{5}$	$\frac{35}{6}$	$\frac{43}{7}$	$\frac{50}{8}$	$\frac{55}{9}$	6
	7	7	6	6	5.8	5.833	6.143	6.25	6.111	6
$\overline{X}_N - \overline{X}_{N-1}$	–	0	−1	0	−0.2	0.033	0.310	0.107	−0.139	−0.111

We can read what we want into this data — this speaks to the question of when, precisely, is N large? — but we can see that with a low number of observations the difference in sample means is swinging wildly about, while with a larger number the difference in sample means is coming closer and closer to 0.

If people send me the data points from the 1:00pm section, I will fill in a section on Chebyshev's inequality here.

Estimation potpourri

You have two [possibly-biased] coins. The first lands heads-up with probability p_1, while the second lands heads-up with probability p_2. You run the following experiment: flip the first coin; if it lands heads-up, flip the second coin. Let H_1 and H_2 be two random variables; $H_1 = 1$ when the first coin lands heads-up and $H_1 = 0$ otherwise; $H_2 = 1$ when the second coin lands heads-up and $H_2 = 0$ otherwise (including if the second coin is not flipped).

(a) *What is the joint PMF of H_1 and H_2?*

Solution: consider first the outcome space of the experiment: both coins may land heads-up, the first coin may be heads-up while the second is tails-up, or the first coin may be tails-up (and the second is not flipped).

The probability of the first outcome is the probability that the first coin lands heads-up (p_1) multiplied by the probability that the second coin lands heads-up (p_2). The probability of the second outcome is the probability that the first coin lands heads-up (p_1) multiplied by the probability that the second coin lands tails-up ($1 - p_2$). The probability of the third outcome is the probability that the first coin lands tails-up ($1 - p_1$).

The PMF is then given by

$$\begin{aligned} f_{H_1 H_2}(1,1) &= p_1 p_2, \\ f_{H_1 H_2}(1,0) &= p_1(1 - p_2), \\ f_{H_1 H_2}(0,0) &= 1 - p_1. \end{aligned}$$

(b) *What are the marginal PMFs of H_1 and H_2?*

[2] Several people cheated and tried to give ranges; tough cookies.

Solution: the clearest method of computing the marginal distributions is by expressing the probabilities as a table. We have

	1	0	H_2
1	p_1p_2	$p_1(1-p_2)$	p_1
0	0	$1-p_1$	$1-p_1$
H_1	p_1p_2	$1-p_1p_2$	

. The marginal of H_1 is obtained by summing across the columns, while the marginal of H_2 is obtained by summing down the rows. We then have

$$\begin{aligned} f_{H_1}(1) &= p_1, & f_{H_2}(1) &= p_1p_2 \\ f_{H_1}(0) &= 1-p_1, & f_{H_2}(0)&1-p_1p_2. \end{aligned}$$

We also can compute these values directly from the definition of the marginal distributions. That is,

$$\begin{aligned} f_{H_1}(1) &= \sum_{h_2 \in H_2} f_{H_1H_2}(1,h_2) \\ &= f_{H_1H_2}(1,0) + f_{H_1H_2}(1,1) \\ &= p_1(1-p_2) + p_1p_2 \\ f_{H_1}(1) &= p_1p_2. \end{aligned}$$

$$\begin{aligned} f_{H_1}(0) &= \sum_{h_2 \in H_2} f_{H_1H_2}(0,h_2) \\ &= f_{H_1H_2}(0,0) + f_{H_1H_2}(0,1) \\ f_{H_1}(0) &= 1-p_1. \end{aligned}$$

$$\begin{aligned} f_{H_2}(1) &= \sum_{h_1 \in H_1} f_{H_1H_2}(h_1,1) \\ &= f_{H_1H_2}(0,1) + f_{H_1H_2}(1,1) \\ f_{H_2}(1) &= p_1p_2. \end{aligned}$$

$$\begin{aligned} f_{H_2}(0) &= \sum_{h_1 \in H_1} f_{H_1H_2}(h_1,0) \\ &= f_{H_1H_2}(0,0) + f_{H_1H_2}(1,0) \\ &= (1-p_1) + p_1(1-p_2) \\ f_{H_2}(0) &= 1-p_1p_2. \end{aligned}$$

(c) *What is the correlation between H_1 and H_2, $\rho_{H_1H_2}$?*

Solution: by definition,

$$\rho_{H_1H_2} = \frac{\mathrm{Cov}(H_1,H_2)}{\sqrt{\mathrm{Var}(H_1)\mathrm{Var}(H_2)}}.$$

We can shortcut the variance computations; since the marginal of each of H_1, H_2 is a Bernoulli distribution (two possible outcomes), the formula for the variance of a Bernoulli random variable may be applied. In this case,

$$\mathrm{Var}(H_1) = p_1(1-p_1), \qquad \mathrm{Var}(H_2) = p_1p_2(1-p_1p_2).$$

We know that covariance is

$$\mathrm{Cov}(H_1,H_2) = \mathbb{E}[H_1H_2] - \mathbb{E}[H_1]\mathbb{E}[H_2].$$

Again, we can take the shortcut of using the formula for the expectation of a Bernoulli random variable,

$$\mathbb{E}[H_1] = p_1, \quad \mathbb{E}[H_2] = p_1 p_2.$$

The leading expectation we must compute directly,

$$\mathbb{E}[H_1 H_2] = \sum_{(h_1,h_2)} h_1 h_2 f_{H_1 H_2}(h_1, h_2) = (0)(0)(1-p_1) + (1)(0)p_1(1-p_2) + (1)(1)p_1 p_2 = p_1 p_2.$$

The correlation is then given by

$$\rho_{H_1 H_2} = \frac{p_1 p_2 - p_1(p_1 p_2)}{\sqrt{p_1(1-p_1)p_1 p_2(1-p_1 p_2)}} = \sqrt{\frac{p_2 - p_1 p_2}{1 - p_1 p_2}}.$$

(d) *Let X represent the number of heads obtained in the experiment. What is the PMF of X?*

Solution: from the definition of H_1 and H_2, we can see that $X = H_1 + H_2$. It follows that

$$\begin{aligned}
f_X(0) &= \mathrm{P}(X=0) = \mathrm{P}(H_1 + H_2 = 0) = \mathrm{P}(H_1 = 0, H_2 = 0) = f_{H_1 H_2}(0,0) = 1 - p_1, \\
f_X(1) &= \mathrm{P}(X=1) = \mathrm{P}(H_1 + H_2 = 1) = \mathrm{P}(H_1 = 1, H_2 = 0) = f_{H_1 H_2}(1,0) = p_1(1-p_2), \\
f_X(2) &= \mathrm{P}(X=2) = \mathrm{P}(H_1 + H_2 = 2) = \mathrm{P}(H_1 = 1, H_2 = 1) = f_{H_1 H_2}(1,1) = p_1 p_2.
\end{aligned}$$

Then the PMF of X is

$$f_X(x) = \begin{cases} 1 - p_1 & \text{if } x = 0, \\ p_1(1-p_2) & \text{if } x = 1, \\ p_1 p_2 & \text{if } x = 2. \end{cases}$$

(e) *You compute X_i over 10 experiments; you find*

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
2	0	0	1	0	1	2	1	0	1

What is the maximum-likelihood estimate for (p_1, p_2)?

Solution: to find the maximum-likelihood estimate for these two parameters, we must first determine the probability of the observed experimental outcome. Since the experimental outcomes are presumed to be independent, we have

$$\mathrm{P}(\langle X_i \rangle_{i=1}^{10}) = \prod_{i=1}^{10} \mathrm{P}(X_i) = \prod_{i=1}^{10} f_X(X_i).$$

Using the PMF we found in part (d), we can express this product as

$$\mathrm{P}(\langle X_i \rangle_{i=1}^{10}) = (1-p_1)^4 (p_1(1-p_2))^4 (p_1 p_2)^2 = p_1^6 (1-p_1)^4 p_2^2 (1-p_2)^4.$$

The maximum-likelihood estimate is the value of the parameters which maximizes the probability of the observed outcome — we know this outcome has happened, so the true nature of the world should intuitively reflect this; i.e., the known outcome of the world should be the most probable outcome, prior to observation. Determining these values is as simple as taking the first derivative of the above probability and setting it equal to zero. We find

$$\begin{aligned}
& & \frac{\partial}{\partial p_1}\mathrm{P}(\langle X_i\rangle_{i=1}^{10}) &= p_2^2(1-p_2)^4\left[6p_1^5(1-p_1)^4 - 4p_1^6(1-p_1)^3\right]\\
&\rightsquigarrow & 6p_1^5(1-p_1)^4 - 4p_1^6(1-p_1)^3 &= 0\\
&\rightsquigarrow & 6(1-p_1) - 4p_1 &= 0\\
&\iff & p_1 &= \frac{6}{10} = \frac{3}{5};\\
& & \frac{\partial}{\partial p_2}\mathrm{P}(\langle X_i\rangle_{i=1}^{10}) &= p_1^6(1-p_1)^4\left[2p_2(1-p_2)^4 - 4p_2^2(1-p_2)^3\right]\\
&\rightsquigarrow & 2p_2(1-p_2)^4 - 4p_2^2(1-p_2)^3 &= 0\\
&\rightsquigarrow & 2(1-p_2) - 4p_2 &= 0\\
&\iff & p_2 &= \frac{2}{6} = \frac{1}{3}.
\end{aligned}$$

Then the maximium-likelihood estimate for these parameters is $(p_1^*, p_2^*) = (\frac{3}{5}, \frac{1}{3})$.

That H_1 and H_2 represent Bernoulli trials is reflected in these outcomes: we have witnessed the first coin heads-up 6 times out of 10, so the intuitive estimate for p_1 is $\frac{6}{10} = \frac{3}{5}$; of the 6 times the second coin was flipped, we have witnessed it land heads-up 2 times, so the intuitive estimate for p_2 is $\frac{2}{6} = \frac{1}{3}$. This is a nice feature of the Bernoulli setup (and is useful for checking your answers) but does not necessarily generalize to other distributions.

(f) *Suppose $p_1 = \frac{3}{5}$ and $p_2 = \frac{1}{3}$. Use facts about linear combinations of random variables (H_1 and H_2) to compute $\mathbb{E}[X]$ and Var(X).*

Solution: we know that $X = H_1 + H_2$. Since expectation is a linear operator, we have

$$\mathbb{E}[X] = \mathbb{E}[H_1 + H_2] = \mathbb{E}[H_1] + \mathbb{E}[H_2] = p_1 + p_1p_2 = p_1(1+p_2) = \frac{4}{5}.$$

Variance is slightly more involved, but equally formulaic.

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}(H_1 + H_2)\\
&= \mathrm{Var}(H_1) + \mathrm{Var}(H_2) + 2\mathrm{Cov}(H_1, H_2)\\
&= p_1(1-p_1) + p_1p_2(1-p_1p_2) + 2(p_1p_2 - p_1^2p_2)\\
&= \frac{6}{25} + \frac{4}{25} + 2\left(\frac{1}{5} - \frac{3}{25}\right)\\
&= \frac{14}{25}.
\end{aligned}$$

(g) *You run this experiment 7 times. Use Chebyshev's inequality to place an upper bound on $P(\overline{X} \leq \frac{1}{5})$.*

Solution: we know $\mathbb{E}[\overline{X}] = \mathbb{E}[X] = \frac{4}{5}$, and $\mathrm{Var}(\overline{X}) = \frac{1}{7}\mathrm{Var}(X) = \frac{2}{25}$ (remember, there are 7 experiments). Then we see

$$\mathrm{P}\left(\overline{X} \leq \frac{1}{5}\right) = \mathrm{P}\left(\overline{X} - \frac{4}{5} \leq -\frac{3}{5}\right) \leq \mathrm{P}\left(\left|\overline{X} - \frac{4}{5}\right| \geq \frac{3}{5}\right).$$

It may help to graph on a number line why the ≤ turns into a ≥ when we apply the absolute value.

To apply Chebyshev's inequality, we need to find $k\sigma_{\overline{X}} = \frac{3}{5}$. This is

$$\begin{aligned} && k\sigma_{\overline{X}} &= \frac{3}{5} \\ \iff && k\sqrt{\frac{2}{25}} &= \frac{3}{5} \\ \iff && k &= \frac{3}{\sqrt{2}}. \end{aligned}$$

Chebyshev's inequality tells us then that

$$\mathrm{P}\left(\left|\overline{X} - \frac{4}{5}\right| \geq \frac{3}{\sqrt{2}}\sqrt{2}25\right) \leq \frac{1}{\left(\frac{3}{\sqrt{2}}\right)^2} = \frac{2}{9}.$$

Then an extremely weak upper bound on the desired probability is

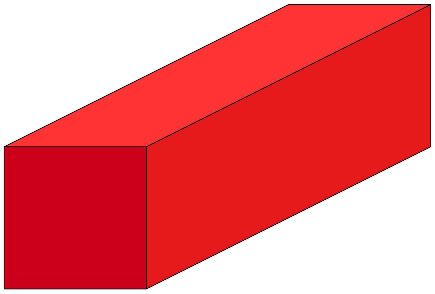
$$\mathrm{P}\left(\overline{X} \leq \frac{1}{5}\right) \leq \frac{2}{9}.$$

Continuous random variables

Up until this point we have been dealing with discrete random variables; these are good for addressing whether or not something happens, what kind of thing happens, how many times it happens, etc. This concept will only capture so many features of the world. Consider, for example, the number of gallons of water you use daily, as an Angeleno: it could be 100 gallons or it could be 110 gallons. It could also be 105.1 gallons 105.11 gallons, 105.111 gallons, etc. The crux is this: the amount of water you use could be any among a continuous set of values; a discrete random variable cannot capture this[3]!

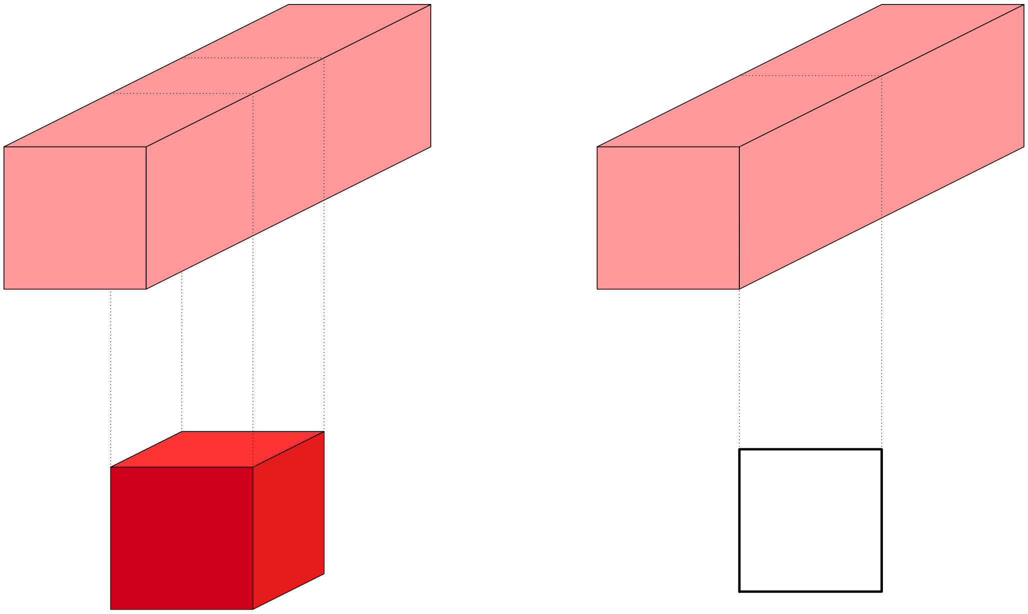
There is a slight hiccup, though. With discrete random variables, any outcome may happen with a strictly positive probability (hence, *probability mass*). But with the infinite outcomes a continuous random variable may take[4], if each occurred with positive probability we would certainly have an overall probability greater than 1; an [uncountably] infinite quantity of positive numbers must sum to infinity. So instead of speaking of probability mass, we speak of *probability density.*

The analogy that I like to use is this: imagine a brick. This brick has a definite mass. If we take a small portion of this brick, it also has mass. However, once we take an *infinitesimally small* portion of the brick, it has no mass. That is, since it has no size, it contains nothing and therefore has no mass. However, it still has a density! In fact, to determine the overall mass of the brick, we can integrate over the densities of each infinitesimal slice. That is, since density is mass per unit volume, if we integrate over the entire volume we will get the mass back.



[3] Although it could say you use between 108 and 109 gallons of water, between 109 and 110 gallons of water, etc.

[4] The math is slightly more nuanced than this, but this is good enough for government work.



A continuous random variable functions in a similar way: each individual outcome (analogous to the infinitesimal slice of the brick) has zero probability mass, but it still has a probability density. When we integrate over the probability density, we get back the probability mass that we are used to dealing with. This motivates the fact that rather than probability mass functions (PMFs), we now deal with *probability density functions*, or PDFs.

If the PDF of a random variable X is f_X, we have the following:

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f_X(x) dx.$$

PDFs have properties that more or less align with the properties we learned about PMFs:

- $f_X(x) \geq 0$;
- $\int_{\text{Support}(X)} f_X(x) dx = 1$.

In particular, many of the things we did before with PMFs may be done with PDFs, substituting the summation ($\sum$) with integration ($\int$).

Question: *the PDF of X is $f_X(x) = ax$, and X has support $[0, 1]$. What is a?*

Solution: we know that $\int_0^1 f_X(x)dx = 1$. So we have

$$\begin{aligned} & & \int_0^1 f_X(x)dx &= 1 \\ \iff & & \int_0^1 axdx &= 1 \\ \iff & & \frac{1}{2}ax^2\Big|_{x=0}^1 &= 1 \\ \iff & & \frac{a}{2} &= 1 \\ \iff & & a &= 2. \end{aligned}$$

Question: *the PDF of X is $f_X(x) = a(k+x)(k-x)$, ($a > 0$) and X has support $[-B, B]$.*

(a) *What values can B take?*

Solution: we know that $f_X(x) \geq 0$ for all x in the support of X. This means

$$\begin{aligned} & & a(k+x)(k-x) &\geq 0 \\ \iff & & k^2 - x^2 &\geq 0 \\ \iff & & k^2 &\geq x^2 \\ \iff & & |x| &\leq |k|. \end{aligned}$$

Since we know that $X \in [-B, B]$, the only way to have $|x| \leq |k|$ for all x in the support of X is if $B \in (0, k]$.

(b) *What is a as a function of B and k?*

Solution: to find a, we appeal to the same method as in the previous question.

$$\begin{aligned} & & \int_{-B}^B f_X(x)dx &= 1 \\ \iff & & \int_{-B}^B a(k+x)(k-x)dx &= 1 \\ \iff & & \int_{-B}^B k^2 - x^2 dx &= \frac{1}{a} \\ \iff & & \left(k^2x - \frac{1}{3}x^3\right)\Big|_{x=-B}^B &= \frac{1}{a} \\ \iff & & 2B\left(k^2 - \frac{1}{3}B^2\right) &= \frac{1}{a} \\ \iff & & a &= \left[2B\left(k^2 - \frac{1}{3}B^2\right)\right]^{-1}. \end{aligned}$$

Henceforth, assume $B = k$.

(c) *What is a?*

Solution: we substitute in,

$$a = \left[2B\left(k^2 - \frac{1}{3}B^2\right)\right]^{-1} = \left[2k\left(\frac{2k^2}{3}\right)\right]^{-1} = \frac{3}{4k^3}.$$

(d) *What is* $P(X \leq \frac{k}{2})$*?*

Solution: we can use the definition of the cumulative distribution function (CDF) to determine this value. That is,

$$\mathrm{P}\left(X \leq t\right) = F_X\left(t\right) = \int_{-k}^{t} f_X(x)dx.$$

Then we have

$$\begin{aligned} F_X(t) &= \int_{-k}^{t} \frac{3}{4k^3}(k^2 - x^2)dx \\ &= \frac{3}{4k^3}\left(k^2x - \frac{1}{3}x^3\right)\Big|_{x=-k}^{t} \\ &= \frac{3}{4k^3}\left(k^2t - \frac{1}{3}t^3\right) + \frac{3}{4k^3}\left(\frac{2k^3}{3}\right) \\ &= \frac{3k^2t - t^3}{4k^3} + \frac{1}{2}. \end{aligned}$$

It follows that

$$\mathrm{P}\left(X \leq \frac{k}{2}\right) = F_X\left(\frac{k}{2}\right) = \frac{12k^3 - k^3}{32k^3} + \frac{1}{2} = \frac{27}{32}.$$

(e) *What is* $P(-\frac{k}{2} \leq X \leq \frac{k}{2})$*?*

Solution: for a continuous random variable, the definition of this probability is

$$\mathrm{P}\left(-\frac{k}{2} \leq X \leq \frac{k}{2}\right) = \int_{-\frac{k}{2}}^{\frac{k}{2}} f_X(x)dx.$$

Appealing to calculus, we can rewrite this as

$$\int_{-\frac{k}{2}}^{\frac{k}{2}} f_X(x)dx = \int_{-k}^{\frac{k}{2}} f_X(x)dx - \int_{-k}^{-\frac{k}{2}} f_X(x)dx = F_X\left(\frac{k}{2}\right) - F_X\left(-\frac{k}{2}\right).$$

This general property is quite useful; we can also see it in the following way: let $E_1 = \{X \leq -\frac{k}{2}\}$, $E_2 = \{X \leq \frac{k}{2}\}$, and $E_3 = \{-\frac{k}{2} \leq X \leq \frac{k}{2}\}$. Then we have $E_1 \cap E_3 = \emptyset$, but $E_1 \cup E_3 = E_2$. Since the probability of the union of mutually-exclusive events is the sum of their individual probabilities, we then have

$$\mathrm{P}(E_1) + \mathrm{P}(E_3) = \mathrm{P}(E_2) \quad \Longrightarrow \quad \mathrm{P}(E_3) = \mathrm{P}(E_2) - \mathrm{P}(E_1) = F_X\left(\frac{k}{2}\right) - F_X\left(-\frac{k}{2}\right).$$

As we have already computed $F_X(\frac{k}{2})$, we only need $F_X(-\frac{k}{2})$. This is

$$F_X\left(-\frac{k}{2}\right) = -\frac{12k^3 - k^3}{32k^3} + \frac{1}{2} = \frac{5}{32}.$$

It follows that

$$\mathrm{P}\left(-\frac{k}{2} \leq X \leq \frac{k}{2}\right) = \frac{27}{32} - \frac{5}{32} = \frac{11}{16}.$$