Introduction to Random Variables

Nathaniel E. Helwig

University of Minnesota

1 What is a Random Variable?

The concept of "randomness" is fundamental to the field of statistics. As mentioned in the probability theory notes, the science of statistics is concerned with assessing the uncertainty of inferences drawn from random samples of data. Now that we've defined some basic concepts related to set operations and probability theory, we can more formally discuss what it means for things to be random.

Definition. According to the Merriam-Webster online dictionary[1], the word random means "lacking a definite plan, purpose or pattern" or "relating to, having, or being elements or events with definite probability of occurrence". In probability and statistics, we use the second definition, such that a random process is any action that has a probability distribution.

The concept of uncertainty is inherent to the definition of randomness. For any random process, we cannot be certain about the action's outcome because there is some element of chance involved. Note that the opposite of a random process is a "deterministic process", which is some action that always results in the same outcome.

***Example* 1.** Flipping a two-sided (heads-tails) coin is a random process because we do not know if we will observe a heads or tails. Flipping a one-sided (heads-heads) coin is a deterministic process because we know that we will always observe a heads.

***Example* 2.** Drawing the first card from a shuffled deck of 52 cards is a random process because we do not know which card we will select. Drawing the first card from a sorted deck of 52 cards is a deterministic process because we will select the same card each time.

[1] https://www.merriam-webster.com/dictionary/random

Definition. In probability and statistics, a random variable is an abstraction of the idea of an outcome from a randomized experiment. More formally, a random variable is a function that maps the outcome of a (random) simple experiment to a real number. Random variables are typically denoted by capital italicized Roman letters such as X.

A random variable is an abstract way to talk about experimental outcomes, which makes it possible to flexibly apply probability theory. Note that you cannot observe a random variable X itself, i.e., you cannot observe the function that maps experimental outcomes to numbers. The experimenter *defines* the random variable (i.e., function) of interest, and then *observes* the result of applying the defined function to an experimental outcome.

Definition. The realization of a random variable is the result of applying the random variable (i.e., function) to an observed outcome of a random experiment. This is what the experimenter actually observes. The realization of a random variable is typically denoted using lowercase italicized Roman letters, e.g., x is a realization of X.

Definition. The domain of a random variable is the sample space S, i.e., the set of possible realizations that the random variable can take.

***Example* 3.** Suppose we flip a fair (two-sided) coin $n \geq 2$ times, and assume that the n flips are independent of one another. Define X as the number of coin flips that are heads. Note that X is a random variable given that it is a function (i.e., counting the number of heads) that is applied to a random process (i.e., independently flipping a fair coin n times). Possible realizations of the random variable X include any $x \in \{0, 1, \ldots, n\}$, i.e., we could observe any number of heads between 0 and n.

***Example* 4.** Suppose that we draw the first card from a randomly shuffled deck of 52 cards, and define X as the suit of the drawn card. Note that X is a random variable given that it is a function (i.e., suit of the card) that is applied to a random process (i.e., drawing the first card from a shuffled deck). Possible realizations of the random variable X include any $x \in \{1, 2, 3, 4\}$, where 1 = Clubs, 2 = Diamonds, 3 = Hearts, and 4 = Spades. Note that it is not necessary to code the suits using numeric values, i.e., we could write that $x \in \{\text{Clubs}, \text{Diamonds}, \text{Hearts}, \text{Spades}\}$. However, mapping the suits onto numbers is (i) notationally more convenient given that we can more compactly denote the possibilities, and (ii) technically more correct given our definition of a random variable.

2 Discrete versus Continuous Random Variables

A random variable has a probability distribution that associates probabilities to realizations of the variable. Before explicitly defining what such a distribution looks like, it is important to make the distinction between the two types of random variables that we could observe.

Definition. A random variable is discrete if its domain consists of a finite (or countably infinite) set of values. A random variable is continuous if its domain is uncountably infinite.

***Example* 5.** Suppose we flip a fair (two-sided) coin $n \geq 2$ times, and assume that the n flips are independent of one another. Define X as the number of coin flips that are heads. The random variable X is a discrete random variable given that the domain $S = \{0, \ldots, n\}$ is a finite set (assuming a fixed number of flips n). Thus, we could associate a specific probability to each $x \in S$. See Example 13 from the "Introduction to Probability Theory" notes for an example of the probability distribution with $n = 3$.

***Example* 6.** Consider the face of a clock, and suppose that we randomly spin the second hand around the clock face. Define X as the position where the second hand stops spinning (see Figure 1). The random variable X is a continuous random variable given that the domain $S = \{x \mid x \text{ is a point on a circle}\}$ is an uncountably infinite set. Thus, we cannot associate a specific probability with any given $x \in S$, i.e., $P(X = x) = 0$ for any $x \in S$, but we can calculate the probability that X is in a particular range, e.g., $P(3 < X < 6) = 1/4$.

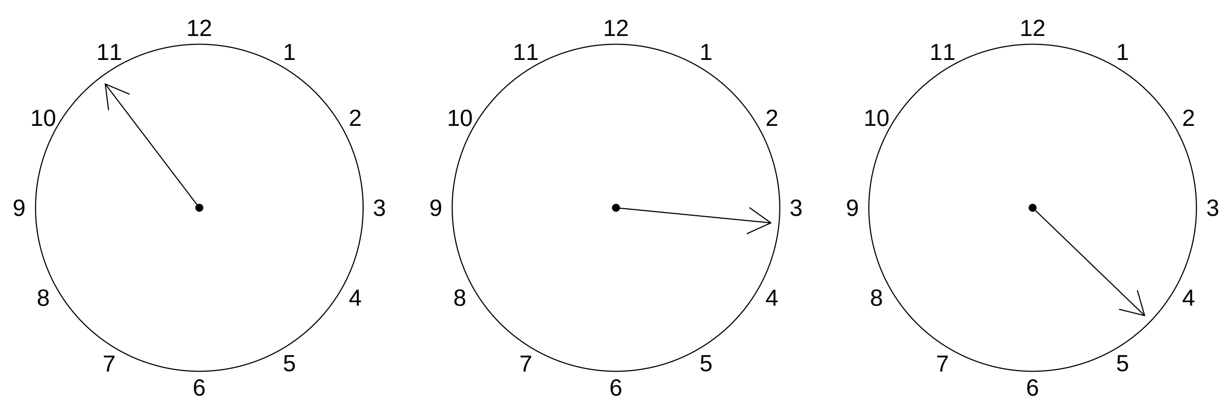

Figure 1: Clock face with three random positions of the second hand.

3 Probability Mass and Density Functions

For discrete random variables, we can enumerate all of the possible realizations of the random variable, and associate a specific probability with each possible realization. In contrast, for continuous random variables, we cannot enumerate all of the possible realizations of the random variable, so it is impossible to associate a specific probability with each possible realization. As a result, we must define the probabilities of discrete and continuous random variable occurrences using distinct (but related) concepts.

Definition. The probability mass function (PMF) of a discrete random variable X is the function $f(\cdot)$ that associates a probability with each $x \in S$. In other words, the PMF of X is the function that returns $P(X = x)$ for each x in the domain of X.

Any PMF must define a valid probability distribution, with the properties:

- $f(x) = P(X = x) \geq 0$ for any $x \in S$
- $\sum_{x \in S} f(x) = 1$

***Example* 7.** See Figure 2 for the PMF corresponding to the coin flipping example with $n = 5$ and $n = 10$ independent flips. As we might expect, the most likely realizations of X are those in which we observe approximately $n/2$ heads. Note that it is still possible to observe $x = 0$ or $x = n$ heads, but these extreme results are less likely to occur.

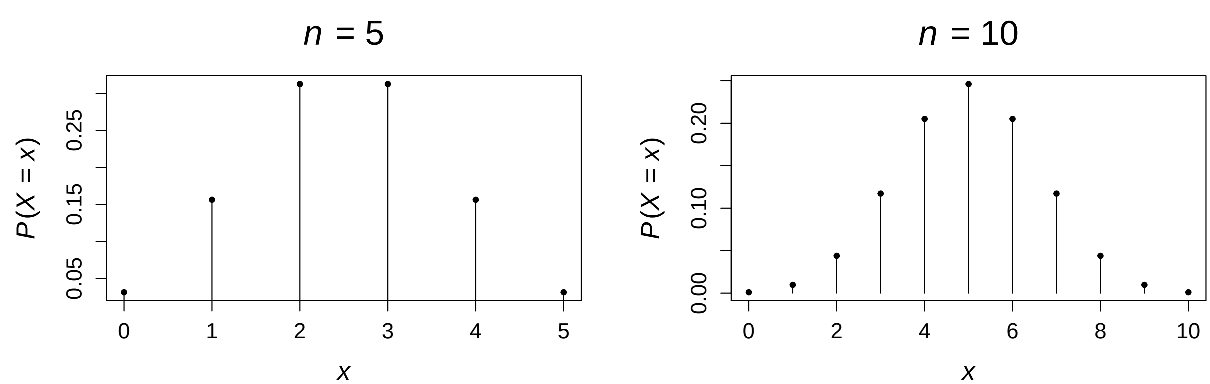

Figure 2: Probability mass function for coin flipping example with $n = 5$ and $n = 10$.

Definition. The probability density function (PDF) of a continuous random variable X is the function $f(\cdot)$ that associates a probability with each range of realizations of X. The area under the PDF between a and b returns $P(a < X < b)$ for any $a, b \in S$ satisfying $a < b$.

Any PDF must define a valid probability distribution, with the properties:

- $f(x) \geq 0$ for any $x \in S$
- $\int_a^b f(x)dx = P(a < X < b) \geq 0$ for any $a, b \in S$ satisfying $a < b$
- $\int_{x \in S} f(x)dx = 1$

***Example* 8.** Suppose that we randomly spin the second hand around a clock face n independent times. Define Z_i as the position where the second hand stops spinning on the i-th replication, and define $X = \frac{1}{n}\sum_{i=1}^n Z_i$ as the average of the n spin results. Note that the realizations of X are any values $x \in [0, 12]$, which is the same domain as Z_i for $i = 1, \ldots, n$. See Figure 3 for the PDF with $n = 1$ and $n = 5$ independent spins. Note that with $n = 1$ spin, the PDF is simply a flat line between 0 and 12, which implies that $P(x < X < x + 1)$ is equal for any $x \in \{0, \ldots, 11\}$. This makes sense given that, for any given spin, the second hand could land anywhere on the clock face with equal probability. With $n = 5$ spins, the PDF has a bell shape, where values around the midpoint of $x = 6$ have the largest density. With $n = 5$ spins, we have that $\int_4^8 f(x)dx = P(4 < X < 8) \approx 0.80$ (i.e., about 80% of the realizations of X are between 4 and 8).

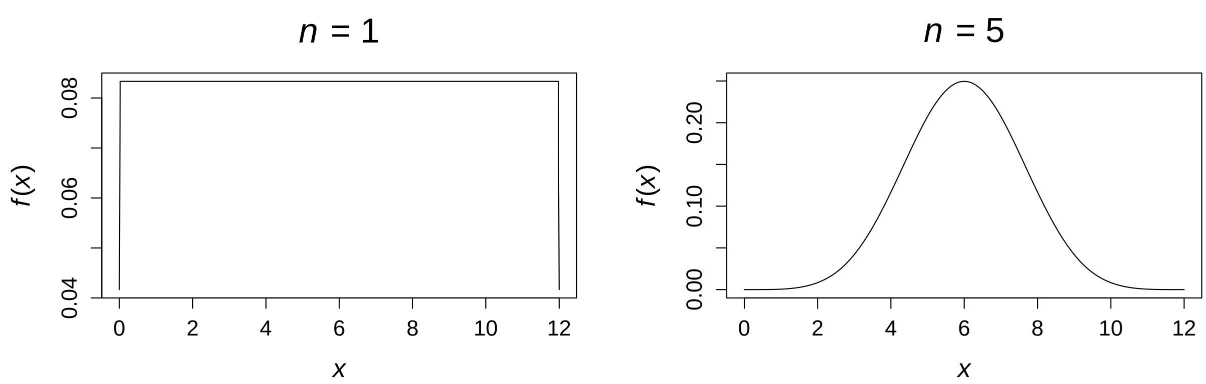

Figure 3: Probability density function for clock spinning example with $n = 1$ and $n = 5$.

4 Cumulative Distribution Function

Definition. The cumulative distribution function (CDF) of a random variable X is the function $F(\cdot)$ that returns the probability $P(X \leq x)$ for any $x \in S$. Note that the CDF is the same as the probability distribution that was defined in Section 4 of the "Introduction to Probability" notes, such that the CDF is a function from S to $[0, 1]$, i.e., $F : S \to [0, 1]$.

Any CDF must define a valid probability distribution, i.e., $0 \leq F(x) \leq 1$ for any $x \in S$, which comes from the fact that $F(x) = P(X \leq x)$ is a probability calculation. Note that a CDF can be defined for both discrete and continuous random variables:

- $F(x) = \sum_{z \in S, z \leq x} f(z)$ for discrete random variables
- $F(x) = \int_{-\infty}^{x} f(z) dz$ for continuous random variables

Furthermore, note that probabilities can be written in terms of the CDF, such as

$$P(a < X \leq b) = F(b) - F(a)$$

given that the CDF is related to the PMF (or PDF), such as

- $f(x) = F(x) - \lim_{a \to x^-} F(a)$ for discrete random variables
- $f(x) = \frac{dF(x)}{dx}$ for continuous random variables

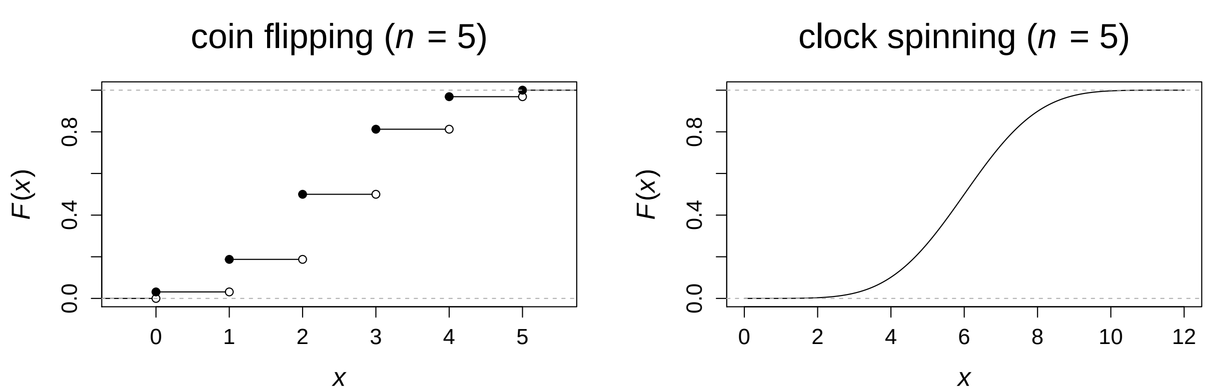

Figure 4: Cumulative distribution function for the coin and clock examples with $n = 5$.

5 Quantile Function

Definition. The quantile function of a random variable X is the function $Q(\cdot)$ that returns the realization x such that $P(X \leq x) = p$ for any $p \in [0, 1]$. Note that the quantile function is the inverse of the CDF, such that $Q(\cdot)$ is a function from $[0, 1]$ to S, i.e., $Q : [0, 1] \rightarrow S$.

More formally, quantile function can be defined as $Q(p) = \min_{x \in S} F(x) \geq p$, where min denotes the minimum. Thus, for any input probability $p \in [0, 1]$, the quantile function $Q(p)$ returns the smallest $x \in S$ that satisfies the inequality $F(x) \geq p$. For continuous random variables, we have the relationship $Q = F^{-1}$. There are a handful of quantiles that are quite popular. The "quartiles" are defined as

- First Quartile: $p = 1/4$ returns x that cuts off the lower 25%
- Second Quartile (Median): $p = 1/2$ returns x that cuts the distribution in half
- Third Quartile: $p = 3/4$ returns x that cuts off the upper 25%

which are some of the most commonly used percentiles.

Definition. The $100p$th percentile of a distribution is the quantile x such that $100p\%$ of the distribution is below x for any $p \in (0, 1)$. For example, the 20th percentile is the quantile corresponding to $p = 1/5$, such that 20% of the distribution is below $Q(1/5)$.

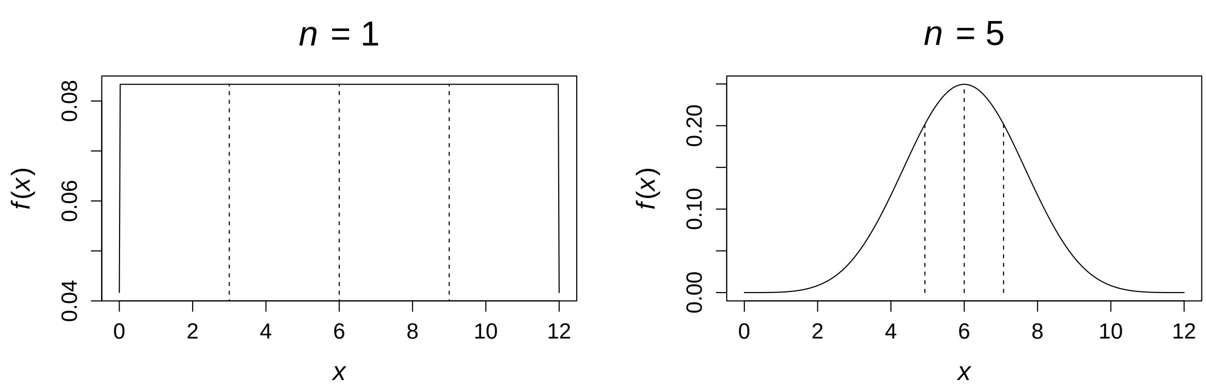

Figure 5: Density function and quartiles for clock spinning example with $n = 1$ and $n = 5$.

6 Expected Value and Expectation Operator

We have already defined one way to talk about the "center" of a distribution: the median. As a reminder, the median of a probability distribution is the quantile value x that cuts the distribution in half, such that 50% of the distribution is below the median and 50% of the distribution is above the median, i.e., $P(X < x) = P(X > x) = 1/2$ if x is the median. Here we will discuss another way to measure the "center" of a distribution, which is useful for understanding what to expect of a random variable.

Definition. The expected value of a random variable X is a weighed average where the weights are defined according to the PMF or PDF. The expected value of X is defined as $\mu = E(X)$ where $E(\cdot)$ is the expectation operator, which is defined as $E(X) = \sum_{x \in S} x f(x)$ for discrete random variables or $E(X) = \int_{x \in S} x f(x) dx$ for continuous random variables.

To gain some insight into the expectation operator $E(\cdot)$, suppose that we have sampled n independent realizations of some random variable X. Let $x_1, \ldots, x_n$ denote the n independent realizations of X, and define the arithmetic mean as $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$. As the sample size n gets infinitely large, the arithmetic mean converges to the expected value μ, i.e.,

$$\mu = E(X) = \lim_{n \to \infty} \bar{x}_n$$

which is due to the *weak law of large numbers* (also known as Bernoulli's theorem).

Expectation Rules

1. If $a \in \mathbb{R}$ is fixed constant, then $E(a) = a$.

2. If X is a random variable with mean $\mu = E(X)$, and if $a, b \in \mathbb{R}$ are fixed constants with $b \neq 0$, then $E(a + bX) = E(a) + bE(X) = a + b\mu$.

3. $E(X_1 + \cdots + X_p) = E(X_1) + \cdots + E(X_p)$, which reveals that *expectations of summations are summations of expectations* (ESSE).

4. $E(b_1 X_1 + \cdots + b_p X_p) = b_1 E(X_1) + \cdots + b_p E(X_p)$, which reveals that *expectations of weighted summations are summations of weighted expectations* (EWSSWE).

Note that rules #3 and #4 are true regardless of whether $X_1, \ldots, X_p$ are independent.

***Example* 9.** For the coin flipping example, note that $X = \sum_{i=1}^n Z_i$ where Z_i is the i-th coin flip. Applying the ESSE rule, we have that the expected value of X can be written as $E(X) = \sum_{i=1}^n E(Z_i)$. Since the coin is assumed to be fair, the expected value of Z_i has the form $E(Z_i) = \sum_{x=0}^1 xf(x) = 0(1/2) + 1(1/2) = 1/2$ for any given $i \in \{1, \ldots, n\}$. Thus, the expected value of X can be written as $E(X) = \sum_{i=1}^n (1/2) = n/2$. In other words, we would expect to observe $n/2$ heads given n independent flips of a fair coin, which makes intuitive sense (about half of the flips should be expected to be heads if the coin is fair).

***Example* 10.** For the clock spinning example, note that $X = \sum_{i=1}^n a_i Z_i$ where Z_i is the i-th clock spin and $a_i = 1/n$ for all $i = 1, \ldots, n$. Applying the EWSSWE rule, we have that the expected value of X can be written as $E(X) = (1/n) \sum_{i=1}^n E(Z_i)$. And note that $E(Z_i) = E(Z_1)$ for all $i = 1, \ldots, n$ because the n spins are independent and identically distributed (iid). Thus, we just need to determine the expected value of a single spin, which has the form $E(Z) = \int_0^{12} zf(z)dz$. The density function for a single clock spin is simply a rectangle that gives equal density to each observable value between 0 and 12 (see Figure 3), which implies that $f(z) = \frac{1}{12}$ for $z \in [0, 12]$ and $f(z) = 0$ otherwise. This implies that the expected value of a single spin has the form $E(Z) = \frac{1}{12} \int_0^{12} zdz = \frac{1}{12} [\frac{1}{2} z^2]_{z=0}^{z=12} = \frac{1}{24}(144 - 0) = 6$, so the expected value of X has the form $E(X) = \frac{1}{n} \sum_{i=1}^n 6 = \frac{1}{n}(6n) = 6$.

7 Variance and Standard Deviation

As we saw in the previous section, the expectation operator can be used to quantify the center of a random variable's distribution, which is useful for understanding what to expect (i.e., the average value) of the variable. In this section, we will see that the expectation operator can also be used to help use quantify the "spread" of a distribution.

Definition. The <u>variance</u> of a random variable X is a weighted average of the squared deviation between a random variable and its expectation, where the weights are defined according to the PMF or PDF. The variance of X is defined as $\sigma^2 = E[(X - \mu)^2]$ where $E[\cdot]$ is the expectation operator: $E[(X - \mu)^2] = \sum_{x \in S} (x - \mu)^2 f(x)$ for discrete random variables or $E[(X - \mu)^2] = \int_{x \in S} (x - \mu)^2 f(x) dx$ for continuous random variables. The variance of X can be written $E[(X - \mu)^2] = E(X^2) - \mu^2$ where $\mu = E(X)$. In other words, the variance of X is the expected value of the squared X minus the square of the expected value of X.

To gain some insight into the variance, suppose that we have sampled n independent realizations of some random variable X. Let $x_1, \ldots, x_n$ denote the n independent realizations of X, and define the arithmetic mean of the squared deviations from the average value, i.e., $\tilde{s}_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$ where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$. As the sample size n gets infinitely large, the arithmetic mean of the squared deviations converges to the variance σ^2, i.e.,

$$\sigma^2 = E[(X - \mu)^2] = \lim_{n\to\infty} \tilde{s}_n^2$$

which is due to the *weak law of large numbers* (also known as Bernoulli's theorem).

Variance Rules

1. If $a \in \mathbb{R}$ is fixed constant, then $\mathrm{Var}(a) = 0$.
2. If X is a random variable with variance $\sigma^2 = \mathrm{Var}(X)$, and if $a, b \in \mathbb{R}$ are fixed constants with $b \neq 0$, then $\mathrm{Var}(a + bX) = \mathrm{Var}(a) + b^2\mathrm{Var}(X) = b^2\sigma^2$.
3. For independent variables, $\mathrm{Var}(X_1 + \cdots + X_p) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_p)$, which reveals that *variances of summations are summations of variances* (VSSV) when the summed variables are mutually independent.
4. For independent variables, $\mathrm{Var}(b_1X_1 + \cdots + b_pX_p) = b_1^2\mathrm{Var}(X_1) + \cdots + b_p^2\mathrm{Var}(X_p)$, which reveals that *variances of weighted summations are summations of weighted variances* (VWSSWV) when the summed variables are mutually independent.
5. More generally, $\mathrm{Var}(b_1X_1 + b_2X_2) = b_1^2\mathrm{Var}(X_1) + b_2^2\mathrm{Var}(X_2) + 2b_1b_2\mathrm{Cov}(X_1, X_2)$, where $\mathrm{Cov}(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$ is the covariance between X_1 and X_2.

Note that rules #3 and #4 are only true if $X_1, \ldots, X_p$ are mutually independent.

***Example* 11.** For the coin flipping example, remember that $X = \sum_{i=1}^n Z_i$ where Z_i is the i-th coin flip. Applying the VSSV rule (which is valid because the Z_i are independent), we have that the variance of X can be written as $\mathrm{Var}(X) = \sum_{i=1}^n \mathrm{Var}(Z_i)$. Since the coin is assumed to be fair, $\mathrm{Var}(Z_i) = \sum_{x=0}^1 (x - 1/2)^2 f(x) = (1/4)(1/2) + (1/4)(1/2) = 1/4$ for any given $i \in \{1, \ldots, n\}$, which uses the fact that $E(Z_i) = 1/2$. Thus, the variance of X can be written as $\mathrm{Var}(X) = \sum_{i=1}^n (1/4) = n/4$.

***Example* 12.** For the clock spinning example, remember that $X = \sum_{i=1}^n a_i Z_i$ where Z_i is the i-th clock spin and $a_i = 1/n$ for all $i = 1, \ldots, n$. Applying the VWSSWV rule, we have that the variance of X can be written as $\mathrm{Var}(X) = \frac{1}{n^2} \sum_{i=1}^n \mathrm{Var}(Z_i)$. And note that $\mathrm{Var}(Z_i) = \mathrm{Var}(Z_1)$ for all $i = 1, \ldots, n$ because the n spins are independent and identically distributed (iid). Thus, we just need to determine the variance of a single spin, which has the form $\mathrm{Var}(Z) = E(Z^2) - E(Z)^2$. Remembering that $f(z) = \frac{1}{12}$ and $E(Z) = 6$, we just need to calculate $E(Z^2) = \frac{1}{12} \int_0^{12} z^2 dz = \frac{1}{12} [\frac{1}{3} z^3]_{z=0}^{z=12} = \frac{1}{36}(1728 - 0) = 48$. Thus, the variance of a single spin Z has the form $\mathrm{Var}(Z) = 48 - 36 = 12$, which implies that the variance of X is $\mathrm{Var}(X) = \frac{1}{n^2} \sum_{i=1}^n 12 = 12/n$.

Definition. The standard deviation of a random variable X is the square root of the variance of X, i.e., the standard deviation of X is defined as $\sigma = \sqrt{E[(X-\mu)^2]}$.

If X is a random variable with mean μ and variance σ^2, then the transformed variable

$$Z = \frac{X - \mu}{\sigma}$$

has mean $E(Z) = 0$ and variance $\mathrm{Var}(Z) = E(Z^2) = 1$. To prove this result, we can use the previous expectation and variance rules. First, note that $Z = a + bX$ where $a = -\frac{\mu}{\sigma}$ and $b = \frac{1}{\sigma}$. Applying Expectation Rule #2 gives $E(Z) = a + bE(X) = -\frac{\mu}{\sigma} + \frac{\mu}{\sigma} = 0$, and applying Variance Rule #2 gives $\mathrm{Var}(Z) = b^2 \mathrm{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1$.

Definition. A standardized variable has mean $E(Z) = 0$ and variance $E(Z^2) = 1$. Such a variable is typically denoted by Z (instead of X) and may be referred to as "z-score".

8 Moments of a Distribution

Definition. The k-th moment of a random variable X is the expected value of X^k, i.e., $\mu_k' = E(X^k)$. The k-th central moment of a random variable X is the expected value of $(X-\mu)^k$, i.e., $\mu_k = E[(X-\mu)^k]$, where $\mu = E(X)$ is the expected value of X. The k-th standardized moment of a random variable X is the expected value of $(X-\mu)^k/\sigma^k$, i.e., $\tilde{\mu}_k = E[(X-\mu)^k]/\sigma^k$, where $\sigma = \sqrt{E[(X-\mu)^2]}$ is the standard deviation of X.

Note: The mean μ is the first moment and the variance σ^2 is the second central moment.

To gain some insight into the moments of a distribution, suppose that we have sampled n independent realizations of some random variable X. Letting $x_1, \ldots, x_n$ denote the n independent realizations of X, we have that

$$\mu'_k = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i^k$$
$$\mu_k = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_n)^k$$
$$\tilde{\mu}_k = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}_n}{\tilde{s}_n}\right)^k$$

where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\tilde{s}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$. Note that these are all results of the law of large numbers, which states that averages of iid data converge to expectations.

Definition. The third standardized moment is referred to as skewness, which is a measure of the lack of symmetry of a probability distribution. Positive (or right-skewed) distributions are have heavy right tails. Negative (or left-skewed) distributions have heavy left tails.

Definition. The fourth standardized moment is referred to as kurtosis, which is a measure of the peakedness of a probability distribution. Kurtosis above 3 is leptokurtic (more peaked than normal) and kurtosis below 3 is platykurtic (less peaked than normal).

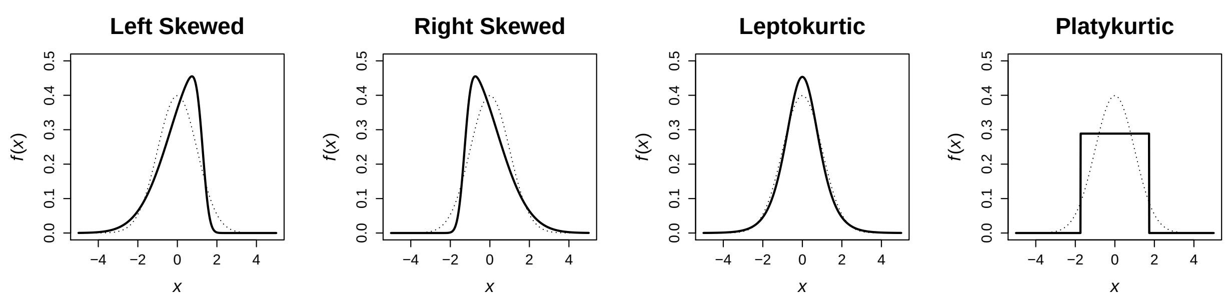

Figure 6: Density functions for distributions with different values of skewness and kurtosis. The dotted line in each subplot denotes the density of the standard normal distribution.