201-SN1 - Probability & Statistics

B. Cordy

Fall 2025

Contents

Chapter 1

Descriptive Statistics

1.1 Sampling

Definition 1.1 *Given a research question to be addressed by a study or experiment, the group that the question concerns is called a **population**. Our goal is to try to answer the question using data about some smaller subset of the population, known as a **sample**.*

Example 1.1 *We are interested in whether a new treatment for pollen allergies is more effective than an existing treatment. To try to answer this question, we can take a large group of people with pollen allergies, give half the new treatment, give the other half the existing treatment, and have them report their results so we can compare.*

We're interested in studying individuals with pollen allergies, in general, that's the population in this scenario. Only those who took one of the two treatments and reported their results back are members of the sample.

Key Point 🔑
Someone who took one of the treatments, but did not report their results, is not in our sample. The sample is the group we have data about.

Perhaps the simplest way to illustrate the distinction is to think of the population as the group we *want to know about*, and the sample as the group we *do know about*.

Example 1.2 *A semiconductor fabrication plant wants to estimate the proportion of defective chips among their total production. To do this, they test every twentieth chip that leaves the production line.*

Since they're interested in what proportion of all chips they produce are defective, the population here consists of all chips produced at their plant. The sample includes only those that are tested.

Statistics & Parameters

Definition 1.2 *A numerical property of a population is called a **parameter**, while a numerical property of sample is called a **statistic**.*

Example 1.3 *In a forest where 41.3% of trees are cedar, a biologist (who is unaware of this fact) estimates the proportion of cedar trees by selecting twenty different trees in the forest at random. She notes that eleven of the twenty trees in her sample are cedar, so estimates that $\frac{11}{20} = 55\%$ of trees in the forest are cedar. In this context the value of 55% she computes is a statistic, and the true value of 41.3% is a parameter.*

In some contexts, the distinction is more subtle. The point is to consider how a quantity relates to the question that it's being used to address. Does the value give a definitive answer based on complete information, with no uncertainty, or is it based on partial information or subject to some degree of uncertainty?

Example 1.4 *LeBron James' free throw percentage in 2024 was 56.6%. If this figure is being used to answer the question 'what was LeBron James' free throw percentage in 2024', then it's a parameter, but if it's being used to answer the question 'how likely is LeBron James to make his next free throw' then it's a statistic.*

Remark ☛

Some quantities colloquially called statistics are not, in fact, statistics. If a college concludes that 23% of its current students are enrolled in the science program by examining their complete records of student enrollment, they haven't computed a statistic, they've computed a parameter.

The discipline of statistics is about making decisions using sample data. There are so many contexts where sample data is the only kind of information or the highest quality kind of information we have access to, which is why the applications of statistics are so varied.

Given a research question, and relevant sample data, the first question to ask is: how was the sample obtained? Our ability to do statistics well depends on one factor more than any other: do we have an accurate model of the sampling process? In other words, do we understand the process that generated the data we're working with?

Three Kinds of Samples

Definition 1.3 *If we take as our sample those members of the population that are easiest to access, we obtain a **convenience sample**.*
*If we order our population in some way and periodically select members at regular intervals, we obtain a **systematic sample**.*
*If we select members of the population at random, with all members having an equal chance of appearing on every selection, we obtain a **simple random sample**.*

Remark ☝

Sampling can be done **with replacement**, which means the same individual may appear many times in the sample, or **without replacement**, which means that each individual can appear at most once in the sample.

Think about this process for generating a simple random sample: we put the names of every individual in our population (supposing these are unique) into a hat and draw names. When sampling with replacement, after drawing a name we write it down, put it back in the hat, and shake the hat. When sampling without replacement, the names that are drawn are not put back into the hat.

In the rest of the course, whenever we refer to a random sample, this will be shorthand for a simple random sample taken with replacement. This means the sampling process is equivalent to repeatedly rolling a die labelled with the names of the members of our population, with the outcome of any one roll having no effect on the outcome of any other. This will allow us, later on in the course, to provide precise justifications and error bounds for our estimates and predictions by applying the theory of probability.

Warning ⚠

In colloquial language, the term random sample could be interpreted as a sample derived from an unknown or not well-understood process. In statistics, a random sample (a simple random sample taken with replacement) results from a completely specified process, namely, drawing names from a hat in manner where each name is equally likely to appear on each selection, regardless of which names are drawn before or after, or repeatedly rolling a die where each side corresponds to a different individual in the population.

Example 1.5 *A semiconductor fabrication plant wants to estimate the proportion of defective chips among their total production. To do this, they test every twentieth chip that leaves the production line.*

In this context, the sample obtained will be a systematic sample, not a random sample. If they select a random number between zero and twenty and start testing every twentieth chips after that many chips have been produced, they will still not obtain a random sample, since a sample which contains two consecutively produced chips cannot occur.

Example 1.6 *In her statistics class, Anne is doing a project where she wants to estimate how many hours per week students in her age group spend commuting to and from school. She gathers data by asking ten of her friends about their commuting routine.*

In this case, Anne will obtain a convenience sample. Conclusions she draws about students in her age group are dubious, since she's only sampling students she happens to be close with.

Notice that in practice, it is often very difficult to obtain a random sample. It requires a complete list of all members of the population under study, and if these members are people, a way to guarantee they will volunteer the information the experimenters are interested in knowing. If you read any corporate software license agreement, you will almost certainly see a section which affirms all your user data is accessible for studies carried out by the company that developed the software. This ensures the company can obtain high quality samples from the population of its users.

An interesting sampling case study is the COVID-19 outbreak on the MV Diamond Princess, which was quarantined in February 2020. Under normal circumstances, there's no way to obtain a sample of people who are known to have been exposed to a disease (and may or may not have any symptoms), as we cannot forcibly expose people for ethical reasons, and subjects cannot self-select because those who are asymptomatic are often unaware they've been exposed at all. Because no one was allowed to leave the ship, researchers were able to derive accurate estimates of parameters like the proportion of COVID-19 cases which present as asymptomatic.

1.2   Variables & Levels of Measurement

Definition 1.4 *The members of the population are referred to as* **individuals**. *Any property of these individuals which can take different values for different individuals is called a* **variable**.

Model	Year	Colour	Mileage	Condition
Toyota Yaris	2015	Blue	85 000	B
Honda CRV	2018	Black	140 000	C
Ford F-150	2022	Red	46 000	A
Subaru Outback	2019	Blue	77 000	B

Typically, if we represent our data in a table, as above, each individual corresponds to a row of the table (in this case the individuals are used cars in some car seller's inventory), and each variable corresponds to a column. Variables can describe any aspect of the individuals in the data set. It's often useful to categorize them based on which operations we can meaningfully apply.

Remark ☛

In statistics, the letter n is reserved for the sample size, that is, the number of individuals in the dataset, and will appear frequently in formulas and explanations. In the dataset above, $n = 4$.

Definition 1.5 *As we answer the three questions below, each affirmative answer moves the variable one level up in the hierarchy of* **levels of measurement**.

$$\textbf{\textit{Categorical}} \; \rightarrow \; \textbf{\textit{Ordinal}} \; \rightarrow \; \textbf{\textit{Interval}} \; \rightarrow \; \textbf{\textit{Ratio}}$$

- *Can we naturally order the values? (The $<$ and $>$ operations are meaningful)*

- *Can we subtract and average the values? (The $+$ and $-$ operations are meaningful)*

- *Can we multiply and divide the values? (The $\cdot$ and $\div$ operations are meaningful)*

Ratio: Variables like height, distance, or anything else you can measure with a ruler, are at the ratio level. They are naturally ordered and we can perform arithmetic with them. In particular, ratios are meaningful. One distance can be twice as far as another, or two-thirds as far.

Interval: Variables like shoe size or temperature (in $°C$) are at the interval level. Ratios of values are not meaningful. The shoe size 8 is not twice as large as 4, and $10\,°C$ is not twice as hot as $5\,°C$. Values can be put a scale with equal increments, so we can form differences, say a difference of $5\,°C$ between two temperatures, and also compute averages.

Ordinal: Variables like letter grade or level of satisfaction (very unsatisfied, unsatisfied, neutral, satisfied, or very satisfied) are at the ordinal level. None of the four basic arithmetic operations $(+, -, \cdot, \div)$ are meaningful, but there's a clear natural ordering.

Categorical: Variables like name, location, or colour are at the categorical level. The values of the variable have no natural ordering, they are simply labels that can be used to group individuals into categories.
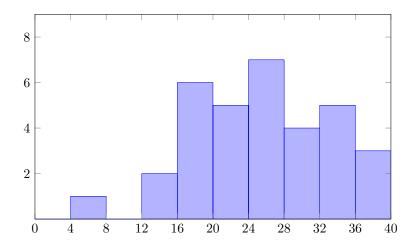
In the case of our used car data in the table above, Model and Colour are at the categorical level, Condition is at the ordinal level, Year is at the interval level, and Mileage is at the ratio level.

Remark ☛
This system of classification of variables into four groups is widespread, but not universal. There are alternative systems, but the one described above is widely known and over time has become part of standard statistical vocabulary.

1.3 Histograms

Any variable, regardless of its level of measurement, has some collection of values it may take, and each of these values occurs some number of times in the data. This information, taken all together, constitutes the ***distribution*** of the variable. The most intuitive way to understand the distribution of a variable is often through a ***histogram***.

The horizontal axis, which represents the variable's value, is divided into segments of equal width, known as ***bins*** or ***classes***, and the height of each bar is the number of individuals whose values fall into

that particular bin. Each bin has a **lower limit** and an **upper limit**. The distance between consecutive lower limits is called the **bin width** or **class width**.
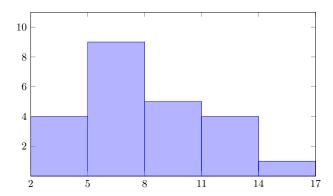
The above discussion only applies to variables at the interval or ratio levels of measurement. When we don't have a scale which can be divided into segments of equal width, we can instead create one bin for each possible value of the variable, and the result is usually called a **bar graph**.

Example 1.7 *Create a histogram with five bins, to visualize the distribution of ages in a room of kids whose ages are given, in order, below.*

$$2 \quad 3 \quad 3 \quad 4 \quad 5 \quad 5 \quad 6 \quad 7 \quad 7 \quad 7 \quad 7 \quad 7 \quad 7 \quad 8 \quad 8 \quad 9 \quad 9 \quad 10 \quad 11 \quad 11 \quad 11 \quad 12 \quad 14$$

We need to partition these 23 values into five bins of equal width. One way to do this is to compute the **range** *of the data (the difference between the smallest and largest value), then add one and divide that by the number of bins, in this case, five.*

This calculation yields $\frac{14-2+1}{5} = 2.6$, which we'll round up to 3, and take as the bin width. Adding one to the range and always rounding up will ensure the largest value doesn't exceed the upper limit of the last bin. Taking the smallest data value as the lower limit of the first bin, and repeatedly adding the bin width to obtain the lower limits of the other bins gives the table and histogram below.

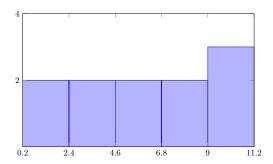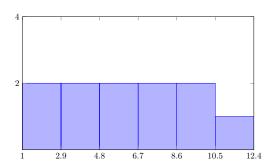| Bins	$2-4$	$5-7$	$8-10$	$11-13$	$14-16$
Counts	4	9	5	4	1

The precise details of the process that creates the bins is not important, as long as the horizontal axis is divided into consecutive segments of equal width in a way that gives a useful visualization of the distribution of the variable.

Example 1.8 *Consider the first eleven positive integers. The histogram below on the left was created from these values by using 2.2 for the bin width, and 0.2 for the lower limit of the first bin. The histogram below on the right was created from the same values by using 1.9 for the bin width, and 1 for the lower limit of the first bin.*

There is nothing incorrect about either of the two histograms, it is simply a quirk of the way the axis was divided into bins that results in the difference in appearance. This illustrates how it's sometimes possible to build intentional bias into a histogram.

As mentioned above, we won't dwell on the details of the process of creating bins, but suffice to say it is good practice to have some consistent, explicit process (usually the one built-in to whatever software package is being used to generate the histogram) and not tweak the details until the result is deemed satisfactory, which one might call 'bin hacking'.

1.4 Measures of Center

It's often useful to quantify the notion of the center of a distribution by producing a single 'typical' or 'central' value. The definition below gives three ways to do this.

Definition 1.6 *Let $x_1, x_2, \ldots, x_n$ be the values of a statistical variable that occur across the n individuals in the dataset.*

- *The **mean** value of the variable is given by $\frac{1}{n}\sum_i x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$.*

- *The **median** value of the variable is obtained by ordering $x_1, x_2, \ldots, x_n$ smallest to largest, then taking the value in the middle of the resulting list if n is odd, and the mean of the two values nearest the middle if n is even.*

- *The **mode** is the value of the variable which occurs most frequently. Note that there may be more than one mode, and in these cases we say the variable is multimodal.*

Example 1.9 *Consider the set of values $1, 1, 2, 3, 5$. We can compute the mean $\frac{1}{5}(1+1+2+3+5) = 2.4$, and since the values are already given in sorted order, we can see the median is 2 and the mode is 1.*

Note that these three measures of centre correspond to three levels or measurement. The mean is defined at the interval level and above, the median is defined at the the ordinal level and above, and the mode is always defined, even at the nominal level.

Warning ⚠

In colloquial language, the average almost always refers to the mean, but there are many situations where it doesn't. The average letter grade cannot refer to the mean, since letter grades are ordinal data, and income and wealth averages quoted in the media are often medians, not means.
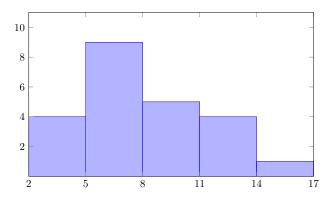
Here's a scenario that illustrates the relevance of each measure: In a guessing game, five numbered balls are put into an urn. The distribution of numbers matches our data in the last example (two balls labelled with 1, one with 2, one with 3, and one with 5). A ball is drawn at random from the urn, and the player's task is to guess the number the ball will be labelled with.

If there is a $10 reward for guessing correctly, and no payout for guessing incorrectly, clearly the best strategy is to guess the mode, 1. What's not so immediate is that if instead the player is rewarded $10 minus the error in their guess (so guessing 3 then drawing a 5 will yield a reward of $10 - 2 = 8$ dollars), the best strategy is to guess the median, 2, and if the player is rewarded $10 minus the squared error in their guess (so guessing 3 then drawing a 5 will yield a reward of $10 - 2^2 = 6$ dollars), then the best strategy is to guess the mean, 2.4.

This is not a quirk that happens because of the particular choice of labels on the balls. In general, regardless of the distribution of values, the median minimizes the average absolute error and the mean minimizes the average squared error.

Example 1.10 *In the preceding section, we constructed the histogram shown below, which summarizes the distribution of the variable whose values are:*

$$2 \quad 3 \quad 3 \quad 4 \quad 5 \quad 5 \quad 6 \quad 7 \quad 7 \quad 7 \quad 7 \quad 7 \quad 7 \quad 8 \quad 8 \quad 9 \quad 9 \quad 10 \quad 11 \quad 11 \quad 11 \quad 12 \quad 14.$$

There are 23 *values in the ordered list, so the median is the* 12th *value, which is a* 7, *while the mean is* $\frac{1}{23}(2 + 3 + 3 + ... + 14) = 7.57$. *Notice that the mean is a little bit larger than the median. When the distribution of a variable has a longer tail on one side, the mean is pulled towards the tail more than the median.*

The most extreme form of this effect occurs when a data value which is much larger than or much smaller than all the other values is present. Such values could occur by mistake, as typos, or could occur because of the nature of the population that the data was sampled from.

Key Point 🔑

Adding an extra value (or two, or three) to a dataset which is very different from the rest, that is, adding an **_outlier_**, will affect the mean more than the median. It's common to summarize this by saying the median is more *stable* than the mean.

For this reason, I encourage students to pay more attention to the median grades in their classes than the mean grades. If a few students miss a test and receive a zero, this can have a noticeable effect on the mean, and give students a false impression of where they stand relative to their classmates. Students are also often concerned about whether their grade ranks them among the top half of the class, which is a question the median answers, not the mean.

In general, if the data being studied contains outliers, the median is often preferred over the mean. This is why it's more common to measure median income than mean income. The distribution of income has such a long tail that citing the mean income can be deceptive when studying populations with high income inequality.

Example 1.11 *Mr. Brown's grade three class has fifteen students and the mean height is 117 cm. Mrs. Green's grade three class has twelve students and the mean height is 122 cm. What is the mean height in the combined group of twenty-seven students?*

We know that $\frac{S_B}{15} = 117$ and $\frac{S_G}{12} = 122$, where S_B and S_G are the sums of the heights of all students in Mr. Brown's and Mrs. Green's classes, respectively. Therefore, $S_B = 15 \cdot 117$ and $S_G = 12 \cdot 122$. We can then calculate the mean of the entire group as

$$\frac{S_B + S_G}{27} = \frac{15 \cdot 117 + 12 \cdot 122}{27} = 119.22.$$

The example above, and the one below, are typical of more involved problems about means, which generally require you to pass back and forth between means and sums. You should get used to doing this, as the same idea will reappear later in the course in problems involving the Central Limit Theorem REF!.

Example 1.12 *Mr. Brown's grade three class has fifteen students and the mean height is 117 cm. If a new student who is 114 cm tall joins the class, how will the mean change?*

As above, we know that $\frac{S_B}{15} = 117$, so $S_B = 15 \cdot 117$. We can then calculate the mean of the new group of sixteen students as below, and conclude the mean height will drop by 0.19 cm.
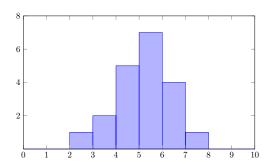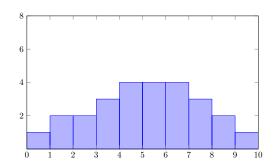
$$\frac{S_B + 114}{16} = 116.81$$

In later parts of the course, it will be extremely important to distinguish the mean of a population from the mean of a random sample taken from that population. The greek letter μ is used for the mean of a population (a parameter), while $\bar{x}$ is used for the mean of a sample (a statistic). Whenever possible, we'll try to use greek letters for parameters and latin letters for statistics.

Example 1.13 *A random number generator produces the values 1, 2, 3, 4, 5, and 6 with equal probability. Three random numbers are produced, and the results are 2, 2, and 5.*

The mean value generated by the random number generator is $\mu = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$, while the mean in the sample of three values is $\overline{x} = \frac{1}{3}(2 + 2 + 5) = 3$.

1.5   Measures of Dispersion

The two distributions pictured in the histograms above have a similar mean, somewhere between 5 and 6 in both cases, but differ noticeably in how the data is distributed around the mean. The variable whose distribution is pictured on the right is more likely to take values farther from it's mean. We say it has higher **dispersion** than the variable on the left. In this section we'll introduce a few ways to quantify this notion.

We could measure dispersion by simply taking the difference between the smallest and largest data values. This is known as the **range** of the variable. However, the range is extremely sensitive to outliers, and ignores almost all of the data (all but the largest and smallest values). We can do better.

Definition 1.7 *Let Q_1 denote the median of all values smaller than the median, and let Q_3 denote the median of all values larger than the median. These are known, respectively, as the first and third quartile. The **interquartile range** is given by $IQR = Q_3 - Q_1$.*

The idea here is to divide the sorted data into blocks: the smallest 25%, the middle 50%, and the largest 25%. The interquartile range is the width of the middle block. This is a quick way to quantify dispersion, which as with median, is not strongly influenced by the presence of outliers.

Example 1.14 *Calculate median and IQR of the variable that takes the values below.*

$$1 \ \ 1 \ \ 2 \ \ 4 \ \ 7$$

The values are already in sorted order, so we can see the median is 2, and the first and third quartiles are $Q_1 = \frac{1+1}{2} = 1$, and $Q_3 = \frac{4+7}{2} = 5.5$. Then $IQR = 5.5 - 1 = 4.5$.

Although the interquartile range is useful for summarizing data, by far the most important measure of dispersion in statistical theory and practice is the variance, together with its square root, the standard deviation.

Definition 1.8 *The average squared distance from the mean is called the **variance**, and denoted σ^2. If $x_1, x_2, \ldots, x_n$ are the values of a statistical variable that occur across the n individuals in the dataset, and μ is the mean of these values, then the variance is given by $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n}(x_i - \mu)^2$. The square root of the variance, σ, is known as the standard deviation.*

Example 1.15 *Find the variance and standard deviation of the variable that takes the values below.*

$$1 \quad 1 \quad 2 \quad 4 \quad 7$$

First we compute $\mu = \frac{1}{5}(1 + 1 + 2 + 4 + 7) = 3$. Then we can calculate the variance as

$$\sigma^2 = \tfrac{1}{5}[(1-3)^2 + (1-3)^2 + (2-3)^2 + (4-3)^2 + (7-3)^2] = \tfrac{1}{5}[4 + 4 + 1 + 1 + 16] = \tfrac{26}{5},$$

and the standard deviation is $\sigma = \sqrt{\frac{26}{5}} = 2.28$.

Warning ⚠

The standard deviation is not the average distance from the mean. We know $\sqrt{a^2 + b^2} \neq a + b$, and similarly, the average of a sequence of terms is not equal to the square root of the average of their squares.

The actual average distance from the mean is known as the mean absolute deviation, which we will not study in this course. It's a fine measure of dispersion, but does not play a central role in statistics, as the variance and standard deviation do.

Bessel's Correction

Suppose we want to determine the age of the oldest person in a population, and we do so by taking a random sample, and using the age of the oldest person in the sample as our estimate. It should be clear that, typically, the result will be an underestimate. Why? This process will certainly never produce an overestimate, and unless the random sample happens to contain the oldest person in the population, or someone else of the same age, the estimate will be too low. Thus, if we average over all possible samples, the value of our estimate will be lower than the actual age of the oldest person in the population.

This problem doesn't occur when using the sample mean to estimate the population mean. It turns out that the average value of the sample mean, across all possible samples, is always precisely equal to the population mean (we say the sample mean is an **unbiased estimator** of the population mean, while the sample maximum is a **biased estimator** of the population maximum). These notions are developed rigorously in the Probability & Random Variables option course, 201-PRV.

Unfortunately, if we calculate the variance in our sample using the formula introduced earlier, this results in a biased estimator of the population variance, which undershoots on average. The good news is that we can fix the problem easily, and obtain an unbiased estimator for the population variance, by making a small change in the way we calculate the variance of a sample, called Bessel's correction, after German astronomer and mathematician Friedrich Bessel.

Definition 1.9 *The (Bessel corrected) sample variance is $s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ and the (Bessel corrected) sample standard deviation, s, is its square root.*

There are two important changes here. Firstly, the mean is denoted $\bar{x}$, since we're working with sample data, and secondly, instead of averaging as usual, by dividing by the number of terms in the sum, we decrease the divisor by one. This is Bessel's correction.

Example 1.16 *Find the mean and sample variance of the sample of five values 2, 0, 7, 3, 3.*

$$\bar{x} = \tfrac{1}{n} \sum_i x_i = \tfrac{1}{5}(2 + 0 + 7 + 3 + 3) = 3$$

$$s^2 = \tfrac{1}{n-1} \sum_i (x_i - \bar{x})^2 = \tfrac{1}{4}[(-1)^2 + (-3)^2 + (4)^2 + (0)^2 + (0)^2] = \tfrac{13}{2}$$

In practice, the Bessel corrected variance and standard deviation are used almost universally. Try calculating the standard deviation of a small list of values using spreadsheet software on a computer. You'll find two commands to calculate the standard deviation, one where Bessel's correction is applied (STDEV.S in Excel or Google Sheets) and one where it isn't (STDEV.P in Excel or Google Sheets).

Remark 👆
Sample variance and sample standard deviation always refer to the Bessel corrected values. When referring to the original formula in Definition 4.6, we'll omit the prefix sample, or use the term population explicitly.

Example 1.17 *Consider a standard fair six-sided die. Find the mean and variance for the outcome of a single die roll. Find the mean and variance for a sample of four rolls whose results are 2, 3, 6, 1.*

Treating the values 1, 2, 3, 4, 5, 6 as a complete population, we have

$$\mu = \tfrac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

$$\sigma^2 = \tfrac{1}{6}[(-2.5)^2 + (-1.5)^2 + (-0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2] = \tfrac{35}{12}.$$

From the sample of values 2, 3, 6, 1, we have

$$\bar{x} = \tfrac{1}{4}(2 + 3 + 6 + 1) = 3$$

$$s^2 = \tfrac{1}{3}[(-1)^2 + (0)^2 + (3)^2 + (2)^2] = \tfrac{14}{3}.$$

Chapter 2

Sets & Counting

2.1 Sets and Set Notation

Sets formalize the concept of a collection, and the language of set theory is ubiquitous in many branches of mathematics, so being familiar with the language and notation of set theory will be useful not only in this course, but also in mathematics more generally.

Definition 2.1 *A **set** is an unordered collection of objects, called its **elements**, which could be anything. Typically we denote sets with capital letters, and elements of sets with lower case letters when we want to make a clear distinction.*

Example 2.1 *Let $A = \{\heartsuit, p, 2, 3, \diamondsuit\}$. This set has five elements. Note that if $B = \{3, 5, p, \heartsuit, \diamondsuit\}$, then $A = B$. Sets are equal when they have the same elements.*

Key Point 🔑
Sets have no notion of ordering or repetition. The sets $S = \{a, b, c\}$ and $T = \{c, b, b, a\}$ are equal. You'll never (from now on) see an element repeated in a set for this reason.

It's helpful to conceive of a set as a way of dividing the entire universe into two groups, the things that are in the set, and the things that are not. When asked whether any particular thing is in the set, there is always a definitive answer of yes or no. Two sets are equal when the answers to that question always agree.

Elements and Subsets

If a is an element of the set A, we write $a \in A$, and if not, we write $a \notin A$. If $A = \{\heartsuit, p, 2, 3, \diamondsuit\}$ then $\heartsuit \in A$ and $2 \in A$, but $1 \notin A$. The set S satisfying $p \in S$ and $q \in S$, but $x \notin S$ for all other x can be written as $S = \{p, q\}$.

Example 2.2 *Let $S = \{2, 3, 5, 7\}$ and $T = \{5, 10, 15\}$. Then $15 \notin S$, but $15 \in T$.*

If A and B are sets, and every element of A is also an element of B, we say that A is a **subset** of B, and write $A \subseteq B$. Notice that every set is a subset of itself.

Example 2.3 *Let $X = \{2, 3, 5, 7, 9\}$, $Y = \{2, 5, 6\}$, and $Z = \{2, 7, 9\}$. Then $Z \subseteq X$. However, $Y \not\subseteq X$ since 6 is an element of Y, but it's not an element of X.*

Example 2.4 *Let $P = \{a, b, c, d\}$. Then $c \in P$ and $\{c\} \subseteq P$. Note that $\{c\} \notin P$ (there are four elements of P and $\{c\}$ is not one of them) and $c \not\subseteq P$ (c is not a set, so cannot be a subset of anything).*

One set that comes up often enough that it's deserving of its own name is the set with no elements, known as the **empty set**, and denoted $\emptyset = \{\}$. When asked whether anything is in $\emptyset$, the answer is always no. Note that $\emptyset \subseteq A$ for any set A, since if not, there would have to be something in $\emptyset$ which is not an element of A, but $\emptyset$ has no elements, so the statement $\emptyset \subseteq A$ is vacuously true.

In some contexts, we also define a **universal set** U as the set of all objects in the universe we are considering. For example, if we're in a setting where the only objects being considered are real numbers, we can let $U = \mathbb{R}$.

Example 2.5 *Consider a single roll of a six-sided die. Let $U = \{1, 2, 3, 4, 5, 6\}$. If we define L as the set of outcomes which are less than five, then $L = \{1, 2, 3, 4\}$, and if we define E as the set of outcomes which are even, then $E = \{2, 4, 6\}$.*
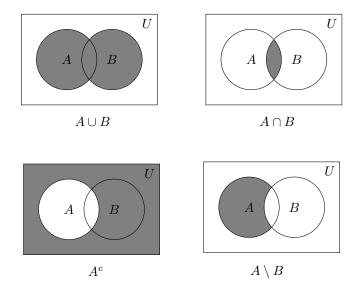
One useful way to describe a set is with **set comprehension**. This notation builds a new set out of a known one by selecting only those elements that satisfy a given condition. The condition can be specified in algebraic notation, in written language, or a mix of the two. The point is that the condition is clear and unambiguous to the reader.

Example 2.6 *Let $S = \{1, 2, 3, 4, 5\}$, and define $T = \{x \in S \mid x > 3\}$. Then $T = \{4, 5\}$.*

Example 2.7 *The set $\{\frac{p}{q} \in \mathbb{Q} \mid q = 2 \text{ and } p \text{ is positive}\}$ is the set of positive half-integers. These are the positive numbers which are either a whole number, or halfway between two whole numbers.*

2.2   Operations on Sets

Given two sets $A, B \subseteq U$, we'll define four fundamental operations, the **union** $A \cup B$, the **intersection** $A \cap B$, the **complement** A^c, and the **difference** $A \setminus B$.

$$A \cup B$$

$$A \cap B$$

$$A^c$$

$$A \setminus B$$

One can interpret $\cup$ as 'or', $\cap$ as 'and', and ^{c} as 'not', but remember that $\cup$ refers to the *inclusive or*, as in 'to participate you must be older than 16 or have permission of a parent' (*possibly both!*), not the *exclusive or*, as in 'the meal comes with a choice of soup or salad' (*not both!*).

Warning ⚠
In colloquial language, the word 'or' is ambiguous. The set $A \cup B$ is the set of all elements that are in A or in B or in both.

The set difference $A \setminus B$ is what remains when all elements of B are removed from the set A. Note that we can express this as $A \setminus B = A \cap B^c$ (the set of elements which are in A *and not* in B), so the difference operation does not add any more expressive power to the language, since it can be defined in terms of the others. However, it's often a convenient shorthand.

Example 2.8 *Let* $U = \{p, q, r, s, t, u, v\}$*, and take* $A = \{p, q, r, s\}$ *and* $B = \{p, r, t, v\}$*. Then we can compute* $A \cap B = \{p, r\}$*,* $A^c = \{t, u, v\}$*, and* $(A \cup B)^c = (\{p, q, r, s, t, v\})^c = \{u\}$*.*

Example 2.9 *Let* $U = \{\heartsuit, \diamondsuit, \clubsuit, \spadesuit\}$*, and take* $S = \{\heartsuit, \diamondsuit, \clubsuit\}$ *and* $T = \{\heartsuit\}$*. Then we can compute* $S \setminus T = \{\diamondsuit, \clubsuit\}$ *and* $T \setminus S = \emptyset$*.*

The Algebra of Sets

The set operations described above satisfy certain algebraic laws. For the most part, these laws should seem intuitively clear once you've had enough experience with the set operations, and for this reason, you should not attempt to memorize them. The exceptions, however, are the **distributive laws** and **DeMorgan's laws**, which take some getting used to, and are not (at least to me and most others) immediate from the meaning of the operations.

Identity	$A \cup \emptyset = A, \quad A \cap U = A$
Domination	$A \cup U = U, \quad A \cap \emptyset = \emptyset$
Idempotent	$A \cup A = A, \quad A \cap A = A$
Complement	$A \cup A^c = U, \quad A \cap A^c = \emptyset, \quad (A^c)^c = A$
Commutative	$A \cup B = B \cup A, \quad A \cap B = B \cap A$
Absorption	$A \cup (A \cap B) = A, \quad A \cap (A \cup B) = A$
Associative	$(A \cup B) \cup C = A \cup (B \cup C)$
	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributive	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
De Morgan's	$(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c$

Example 2.10 *Use laws of set algebra to simplify the expression* $(A \cup (B \cap A^c)) \setminus B$.

$$
\begin{aligned}
(A \cup (B \cap A^c)) \setminus B &= (A \cup (B \cap A^c)) \cap B^c \\
&= ((A \cup B) \cap (A \cup A^c)) \cap B^c \\
&= ((A \cup B) \cap U) \cap B^c \\
&= (A \cup B) \cap B^c \\
&= (A \cap B^c) \cup (B \cap B^c) \\
&= (A \cap B^c) \cup \emptyset \\
&= A \cap B^c
\end{aligned}
$$

Remark ☛

The example above might leave you with some questions. Why were the laws used in that particular order? What does it even mean for one set expression to be 'simpler' than another?

Questions like these are addressed in a branch of mathematics known as Boolean algebra, which we're just scratching the surface of. Take the example as an illustration of how the laws above can be used in a productive way, and an invitation to think about some of the basic questions in Boolean algebra, the answers to which are beyond the scope of this course.

2.3   Inclusion-Exclusion Principle

The number of elements in a set A is known as its ***cardinality***, denoted $N(A)$ or $|A|$. Note that sets can have infinitely many elements, and we will write $N(A) = \infty$ when this occurs.

Example 2.11 *Let* $A = \{\heartsuit, \spadesuit, \diamondsuit\}$, *and* $B = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, ...\}$. *Then* $N(A) = 3$ *and* $N(B) = \infty$.

There is only one set whose cardinality is zero, the empty set $\emptyset$. There are countless sets whose cardinality is infinite. The set B in the example above, the natural numbers $\mathbb{N} = \{0, 1, 2, 3, 4, ...\}$, the integers $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$, and the real numbers $\mathbb{R}$ are examples you've encountered before.
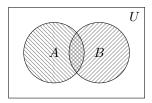
It's important, and may come as a surprise, to know that some of these infinite sets have more elements than others. The sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{E} = \{2, 4, 6, ...\}$ (the positive even numbers) all have the same number of elements, but the set $\mathbb{R}$ has more. The classic introduction to how this works is known as **Hilbert's Hotel**.

If a set has a larger cardinality than $\mathbb{N}$, we say it's **uncountable**. If it has a finite cardinality, or the same cardinality as $\mathbb{N}$, we say it's **countable**.

In the case where the cardinalities $N(A)$ and $N(B)$ are both finite, there is an important relationship between $N(A)$, $N(B)$, $N(A \cup B)$, and $N(A \cap B)$ which comes in very useful when calculating probabilities, which we'll get to in the next chapter.

Theorem 2.1 *(Inclusion-Exclusion Principle)* $N(A \cup B) = N(A) + N(B) - N(A \cap B)$.

Pf. Consider the sum $N(A) + N(B)$. Every element in exactly one of the two sets A and B is counted once in this sum, but an element which is in both sets (that is, an element of $A \cap B$) is counted twice.
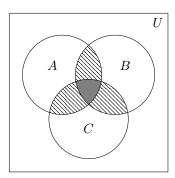

In $N(A \cup B)$, every element of $A \cup B$ (including those in $A \cap B$) is counted only once. To count the elements in $A \cap B$ twice, as above, we can sum $N(A \cup B) + N(A \cap B)$. Thus, $N(A \cup B) + N(A \cap B) = N(A) + N(B)$, and rearranging gives $N(A \cup B) = N(A) + N(B) - N(A \cap B)$.

$\square$

Example 2.12 *How many cards in a standard deck (52 cards, 13 of each suit) are hearts or face cards?*

If we let U be the set of all cards in the deck, H the set of all hearts, and F the set of all face cards (the J, Q, and K of each suit), then $N(H \cap F) = 3$ (the J, Q, and K of hearts), and hence

$$N(H \cup F) = N(H) + N(F) - N(H \cap F) = 13 + 12 - 3 = 22.$$

The inclusion-exclusion principle can be extended to unions of three sets or more. In the case of three, if we write $N(A \cup B \cup C) = N(A) + N(B) + N(C)$ then we have double counted the outcomes in the lined areas in the Venn diagram below, and triple counted the outcomes in $A \cap B \cap C$ (the shaded central area).

To correct for double counting the lined areas, we can subtract the cardinalities of $A \cap B$, $B \cap C$, and $C \cap A$, but each of these includes $A \cap B \cap C$. Elements of this triple intersection are counted three times and subtracted out three times. Thus, we need to add them back in to count them once. All together we obtain the formula below.

$$
\begin{aligned}
N(A \cup B \cup C) = {} & N(A) + N(B) + N(C) \\
& - N(A \cap B) - N(B \cap C) - N(C \cap A) \\
& + N(A \cap B \cap C)
\end{aligned}
$$

Example 2.13 *At a daycare, 24 children will eat Apples, 28 children will eat Bananas, and 30 children will eat Clementines. If 18 will eat Apples & Bananas, 22 will eat Bananas & Clementines, 19 will eat Apples & Clementines, and 14 will eat all three fruits, how many children will eat at least one of the three?*

$$
\begin{aligned}
N(A \cup B \cup C) = {} & N(A) + N(B) + N(C) \\
& - N(A \cap B) - N(B \cap C) - N(C \cap A) \\
& + N(A \cap B \cap C) \\
= {} & 24 + 28 + 30 - 18 - 22 - 19 + 14 \\
= {} & 37
\end{aligned}
$$

Notice that when we apply the inclusion-exclusion principle, we can solve for any of the terms. If we were given the cardinality of the union of all three sets, we could instead solve for the cardinality of the intersection of all three sets.
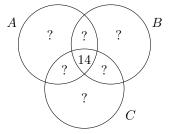
Example 2.14 *At a daycare, 24 children will eat Apples, 28 children will eat Bananas, and 30 children will eat Clementines. If 18 will eat Apples & Bananas, 22 will eat Bananas & Clementines, 19 will eat Apples & Clementines, and 37 will eat at least one of the three fruits, how many children will eat all three?*

$$
\begin{aligned}
N(A \cup B \cup C) = {} & N(A) + N(B) + N(C) \\
& - N(A \cap B) - N(B \cap C) - N(C \cap A) \\
& + N(A \cap B \cap C) \\
37 = {} & 24 + 28 + 30 - 18 - 22 - 19 + N(A \cap B \cap C) \\
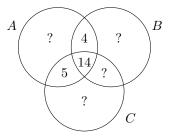14 = {} & N(A \cap B \cap C)
\end{aligned}
$$

These kind problems can also be done by filling in the regions of the Venn diagram carefully with the correct counts, while being very careful to never double count. In many cases, this approach is faster and more natural then writing out the algebraic form of the law.

Example 2.15 *At a daycare, 24 children will eat Apples, 28 children will eat Bananas, and 30 children will eat Clementines. If 18 will eat Apples & Bananas, 22 will eat Bananas & Clementines, 19 will eat Apples & Clementines, and 14 will eat all three fruits, how many children will eat only apples, and neither of the other two fruits?*
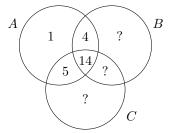
We'll start with the count in the centre, 14, and move from the inside out.

Since $N(A \cap B) = 18$, and the 14 children in the centre will all eat Apples and Bananas, the region directly above the centre has 4 children in it, that is, there are 4 children who will eat Apples and Bananas but not Clementines $(A \cap B \cap C^c)$. Similarly, since $N(A \cap C) = 19$ and $N(A \cap B \cap C) = 14$, there are 5 children who will eat Apples and Clementines but not Bananas $(A \cap B^c \cap C)$.
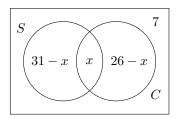

Finally, $N(A) = 24$, and three of the four regions that make up A have been filled out, and total $14 + 4 + 5 = 23$ children. Thus, only a single child eats Apples and none of the other fruits $(A \cap B^c \cap C^c)$.

Key Point 🔑
Always fill in the Venn diagram from the inside out. Even if the quantity you're solving for is in the center, call it x and start filling in from the center, moving out.

Example 2.16 *In a group of 45 people, 31 of them went swimming, 26 of them went canoeing, and 7 of them did neither activity. How many did both activities?*

    Let x be the number of people that did both activities. If we label the Venn diagram as above and sum the values in all four regions, then $(31 - x) + x + (26 - x) + 7 = 45$. Solving yields $x = 19$.

    Alternatively, we could use the algebraic form of the inclusion-exclusion principle, after noting that since 7 people did neither activity, $N((S \cup C)^c) = 7$, and hence $N(S \cup C) = 45 - 7 = 38$. Now applying the inclusion-exclusion principle,

$$N(S \cup C) = N(S) + N(C) - N(S \cap C)$$
$$34 = 31 + 26 - N(S \cap C)$$
$$N(S \cap C) = 31 + 26 - 38$$
$$N(S \cap C) = 19$$

2.4   Counting Principles & Factorials

Suppose you order fries and root beer at a fast food restaurant. There are two options for the size of the fries and three options for the size of the root beer.

    Since there are two options for the size of the fries, and three options for the size of the root beer, you have a total of six possible choices when you order both. These are enumerated below, the first letter in each ordered pair denoting the size of the fries (small or large), and the second letter the size of the root beer (small, medium, or large).

$$(S, s) \quad (S, m) \quad (S, l) \quad (L, s) \quad (L, m) \quad (L, l)$$

These are exactly the elements of the cartesian product $F \times R$, where $F = \{S, L\}$ and $R = \{s, m, l\}$.

If it turns out that, unfortunately, there's only enough change in your pocket to pay for a single item (which could be any size), then you only have five choices. Either you only order fries, and choose an element from F, or only order root beer, and choose an element from R. This is equivalent to choosing an element of $F \cup R = \{S, L, s, m, l\}$.

To summarize, the number of ways to choose an element from F *and* an element from R is $|F| \cdot |R|$, while the number of ways to choose an element from F *or* an element from R is $|F| + |R|$. This idea generalizes considerably, and forms the foundational principle that will allow us to solve many problems.

Proposition 2.1 *(Fundamental Counting Principle) In a sequence of k choices, if the first choice can be made in n_1 ways, the second choice can be made in n_2 ways, and so on, the number of ways of making all choices is the product $n_1 n_2 \cdots n_k$, while the number of ways of selecting one of the choices to make and making only that choice is the sum $n_1 + n_2 + \ldots + n_k$.*

Example 2.17 *How many ways are there to select three students to be class representatives in three different classes, one with twenty students and two with thirty?*

There are three choices to be made, choosing a representative for each of the three classes, and since we must make all three choices, the number of ways this can be done is $20 \cdot 30 \cdot 30 = 18\,000$.

Example 2.18 *How many four symbol sequences can be made with the symbols in $S = \{\heartsuit, \diamondsuit, \clubsuit, \spadesuit\}$?*

We can imagine filling in four blank spaces with symbols, and write in each blank space the number of choices we have for the symbol that gets placed there. If we allow the same letter to be used many times, then the number of sequences is $\underline{4} \cdot \underline{4} \cdot \underline{4} \cdot \underline{4} = 256$.

On the other hand, if each space must be filled with a distinct symbol, then initially there are four choices for the first symbol, but after making that choice, there are three remaining for the second symbol, and so on. In total then, the number of sequences with distinct symbols is $\underline{4} \cdot \underline{3} \cdot \underline{2} \cdot \underline{1} = 24$.

This second result is the number of ways of arranging four different symbols in a line, and counting arrangements is something we'll do frequently, so it's convenient to have a shorthand for this kind of descending product of consecutive integers.

Definition 2.2 *Given $n \in \mathbb{N}$, the **factorial** of n, denoted $n!$, is the number of ways of arranging n distinct objects in a line, and is defined by*

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 2 \cdot 1$$

Remark ☛
Though the usual interpretation of the factorial in terms of counting arrangements no longer applies, we define $0! = 1$ for convenience. This makes many formulas involving factorials (like the Taylor series formula) much easier to write.

Factorials have some very nice algebraic properties that make them easy to work with, in particular, ratios of factorials can be simplified easily.

$$\frac{8!}{6!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{8 \cdot 7 \cdot \not{6} \cdot \not{5} \cdot \not{4} \cdot \not{3} \cdot \not{2} \cdot \not{1}}{\not{6} \cdot \not{5} \cdot \not{4} \cdot \not{3} \cdot \not{2} \cdot \not{1}} = 56$$

2.5   Permutations & Combinations

Permutations

Example 2.19 *How many ways are there to choose three students from a class of twenty and arrange them in a line?*

We can use the same approach as in the last example. We're filling in three blanks, we have twenty options to choose from, and repeats are not permitted. Thus, there are $\underline{20} \cdot \underline{19} \cdot \underline{18} = 6\,840$ ways to do it.

The notation $^{20}P_3$ is used to refer to the number of ways of selecting 3 objects from a set of 20 distinct options and ordering them. Thus, we have $^{20}P_3 = 6\,840$.

Definition 2.3 *If an ordered list of r distinct objects is selected from a set of n objects, the result is called an r-**permutation**. The number of r-permutations of n objects, denoted by ^nP_r, is given by*

$$^nP_r = n \cdot (n-1) \cdot \cdots \cdot (n-r+1) = \frac{n!}{(n-r)!}$$

Remark ☛
We can write the formula for permutations conveniently as above, but it's usually more helpful in problems to think about the fundamental counting principle directly, and to interpret the formula for permutations as a truncated factorial (a factorial with the end cut off).

Example 2.20 *How many ways there there to create a sequence of five letters with no repetitions using any of the 26 lower case letters of the alphabet?*

Taking 5 letters from a set of 26 and arranging them, the number of options is

$$^{26}P_5 = \frac{26!}{(26-5)!} = \frac{26!}{21!} = 26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7\,893\,600$$

Permutations with Identical Objects

What if some objects are indistinguishable? If this is the case, there are fewer possibilities. For example, how many arrangements of the letters in the string $AAAB$ are there?

Notice that we just have to decide which of the four possible positions we'll place the B into. Once that's done, then A's go everywhere else, so there are only four distinct arrangements.

Now suppose that we number the three As. How many rearrangements of the string $A_1 A_2 A_3 B$ are there? Since the four letters are now distinguishable, there are $4! = 24$. Let's write these out, grouping together in columns all strings that will be identical after the labels on the A's are removed.

$$
\begin{array}{llll}
A_1 A_2 A_3 B & A_1 A_2 B A_3 & A_1 B A_2 A_3 & B A_1 A_2 A_3 \\
A_1 A_3 A_2 B & A_1 A_3 B A_2 & A_1 B A_3 A_2 & B A_1 A_3 A_2 \\
A_2 A_1 A_3 B & A_2 A_1 B A_3 & A_2 B A_1 A_3 & B A_2 A_1 A_3 \\
A_2 A_3 A_1 B & A_2 A_3 B A_1 & A_2 B A_3 A_1 & B A_2 A_3 A_1 \\
A_3 A_1 A_2 B & A_3 A_1 B A_2 & A_3 B A_1 A_2 & B A_3 A_1 A_2 \\
A_3 A_2 A_1 B & A_3 A_2 B_1 A & A_3 B A_2 A_1 & B A_3 A_2 A_1
\end{array}
$$

Notice that rearranging the labels will result in a string which is identical once the labels are removed. This means if we list all possible arrangements of the labeled strings, each possible unlabeled string will appear 3! times, since there are 3! ways of arranging the labels. Alternatively, think about how there are 3! ways to *apply the labels* to the three A's, and each of the 24 labeled strings can be obtained by first selecting an unlabeled string (i.e. a column) then applying the labels.

The upshot is that *we can count the number of unlabeled strings by dividing the number of labeled strings by the number of ways of rearranging the labels*, in this case, $4!/3! = 4$. The same principle applies if there are many different kinds of indistinguishable objects, and the general result can be summarized in the formula below.

Proposition 2.2 *(Mississippi Formula) The number of ways of rearranging n objects, with indistinguishable groups of sizes $k_1, k_2, \ldots, k_m$, is given by*

$$
\frac{n!}{k_1! \cdot k_2! \cdot \ \cdots \ \cdot k_m!}
$$

Example 2.21 *How many arrangements of the letters in MISSISSIPPI are there?*

There are eleven letters in total, with identical groups of size 4 (the I's), 4 (the S's), and 2 (the P's). Therefore, the number of arrangements is

$$
\frac{11!}{4! \cdot 4! \cdot 2!} = \frac{11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} = 11 \cdot 10 \cdot 9 \cdot 7 \cdot 5 = 34\,650.
$$

Example 2.22 *An urn contains four green and five red marbles, which are removed one at a time. How many different sequences of red and green marbles can result from this process?*

In this case, we are rearranging nine objects, four G's and five R's. Each arrangement, read left to right, corresponds to a sequence of marbles that could appear. There are identical groups of size 4 (the green marbles), and 5 (the red marbles). Therefore, the number of arrangements is

$$\frac{9!}{4! \cdot 5!} = \frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 9 \cdot 2 \cdot 7 = 126.$$

Combinations

Suppose now we're interested in the number of ways that a subset of r objects can be selected from a larger set of n distinct objects. We're only concerned with which objects are selected not with their order, so if the same objects appear, but in a different order, we consider this to be the same outcome.

Consider how many ways there are to selecta group three children from a family with five. To do this, we can imagine lining up the children any fixed order, let's say by age. We can now specify which children are chosen using a string such as $YYNYN$, which would indicate we chose the first two children in the line as well as the fourth. In fact, every distinct five letter string made up of Ys and Ns which has exactly three Ys will represent a different choice of three children, and every choice of three children determines such a string.

Now we can use the Mississippi formula to count the number of different choices by counting strings. Each string is a rearrangement of five objects with identical groups of sizes three and two, hence there are $\frac{5!}{3!2!} = 5 \cdot 2 = 10$ possibilities.

The general principle is that once the elements of a set have been lined up in some arbitrary order, each subset is given by an *ordered* list of Y's and N's, where Y denotes membership in the subset, and N denotes non-membership. We can count the number of these ordered lists of Y's and N's using the Mississippi formula.

Definition 2.4 *If a subset of r objects is selected from a set of n objects, the result is called an r-combination. The number of r-**combinations**, denoted ^nC_r, or more commonly $\binom{n}{r}$, is given by*

$$^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Warning ⚠

Note carefully the distinction between ^nP_r and ^nC_r. The former counts the number of ways of selecting r objects *and arranging them*, while the latter is the number of selections possible when *any rearrangement is regarded as the same outcome*.

Example 2.23 *In Lotto 6/49, if you purchase a ticket you choose 6 distinct numbers between 1 and 49. The ticket wins if its numbers match 6 distinct numbers between 1 and 49 which are randomly drawn, regardless of the order they are drawn in. How many different tickets would you need to buy to guarantee a win?*

Since the order of the numbers on the ticket is irrelevant, the number of different tickets available is $\binom{49}{6} = \frac{49!}{6!43!} = 13\,983\,816$, so this is the minimum number you would need to purchase to be guaranteed one will win.

Example 2.24 *How many ways are there to choose a president, a treasurer, and an organizing committee of three from a group of 25 people? Assume no person can hold more than one position.*

There are 25 choices for president, then 24 for treasurer, then $\binom{23}{3}$ for the organizing committee. Note that when choosing the organizing committee, the order in which members are chosen is irrelevant. The committee consisting of Alice, Bob, and Chelsea and the committee consisting of Chelsea, Alice, and Bob are the same. We must make all three choices, so by the fundamental counting principle the number of possibilities is

$$25 \cdot 24 \cdot \binom{23}{3} = 1\,062\,600.$$

Notice that we could just as well have proceeded in the opposite order, that is, first chosen the organizing committee, then the treasurer, then the president, and still obtained

$$\binom{25}{3} \cdot 22 \cdot 21 = 1\,062\,600.$$

Example 2.25 *How many ways are there to divide a group of eight people into a team of three and a team of five?*

To create a team of three and a team of five, we choose three individuals for the team of three, then the five remaining make up the team of five.

$$\binom{8}{3}\binom{5}{5} = 56$$

Example 2.26 *How many ways are there to divide a group of eight people into two teams of four?*

To create two teams of four, we choose four for the first team, and those that remain make up another team of four. We have to be very careful though. In the example above the two teams were distinguishable (one has more people than the other), but here they are not.

If we select A, B, C, and D for the first team and E, F, G, and H for the second, this results in the outcome as selecting E, F, G, and H for the first team and A, B, C, and D for the second. The teams are indistinguishable, so every outcome has been double counted! Thus, we need to divide out by the number of rearrangements of the teams.

$$\frac{\binom{8}{4}\binom{4}{4}}{2!} = 35$$

Notice that if the two teams were distinguishable, for example, if there were a red team and blue team with four people on each, there would be $\binom{8}{4}\binom{4}{4} = 70$ ways.

Remark ☝

As the last example shows, it's very easy to end up inadvertently double counting, even for those with lots of experience with these kinds of problems. Solving such a problem amounts to coming up with a sequence of choices which can be made to produce each possible outcome. You need to convince yourself that your sequence of choices can not only produce every possible outcome, but produce it in a unique way (making any of the choices differently results in a different outcome).

Chapter 3

Probability

3.1 Sample Spaces & Events

Suppose the outcome of an experiment is uncertain, but we can represent the collection of all possible outcomes by writing it as a set.

Definition 3.1 *A set which represents the collection of all possible outcomes of an experiment is known as a **sample space**, typically denoted Ω.*

Specifying a concrete representation of the set of all outcomes of an experiment will often clarify the context of a problem and eliminate many potential sources of confusion.

Example 3.1 *Consider a single roll of a single six-sided die. In this experiment, there are six possible outcomes, corresponding to the six distinct faces of the die. Thus, we can take $\Omega = \{1, 2, 3, 4, 5, 6\}$.*

Example 3.2 *Consider measuring the breaking strength (in Newtons) of a plank of wood. If we have no information about the size or shape of the plank, or about the precision with which the breaking strength will be measured, we should allow any positive real value. Thus, $\Omega = \{x \in \mathbb{R} \mid x \geq 0\}$.*

Example 3.3 *Suppose that a coin is flipped, and then a marble is drawn from an urn which contains black, white, and grey marbles. In this scenario, we could take $\Omega = \{HB, HW, HG, TB, TW, TG\}$, using HB to represent flipping a head and drawing a black marble, HW to represent flipping a head and drawing a white marble, and so on.*

Example 3.4 *Suppose you play a game where a coin is flipped repeatedly until the first tail appears. We could take $\Omega = \{T, HT, HHT, HHHT, HHHHT, \dots\}$, the set of all finite sequences of zero or more H's followed by a single T.*

As a sample space is a set, it can be countable of uncountable. We'll see later on in the course that the tools of calculus will come to our aid when dealing with probability problems that involve uncountable sample spaces.

The choice of sample space depends on the experiment and on what question the experimenters are trying to answer. If we roll a pair of dice, we might use the sample space $\Omega = \{2, 3, 4, \ldots, 11, 12\}$ to count the sum of the two resulting values, or $\Omega = \{(1,1), (1,2), (2,1), (2,2), \ldots, (5,6), (6,5), (6,6)\}$ to distinguish the dice and represent the value on each.

One big advantage of the larger sample space is that each of the outcomes is *equally likely*. Imagine one die is red and the other is blue (the colours of the dice certainly won't influence the probabilities in this context). If neither die has any bias, then when the pair is rolled, each die is equally likely to land with any of its faces up. The red die showing 2 and blue showing 5 is as likely as the red die showing 6 and blue showing 6, and so on. In other words, each ordered pair of outcomes (m, n) for $1 \leq m \leq 6$ and $1 \leq n \leq 6$ is as likely as any other.

If we write the number showing on the red die across the top row, the number showing on the blue die down the left column, and the sum of the two results in a table, we obtain the following.

$_B\backslash^R$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Note that outcomes in the sample space $\Omega = \{2, 3, 4, \ldots, 11, 12\}$ are not equally likely. There are six ordered pairs $(1,6)$, $(2,5)$, $(3,4)$, $(4,3)$, $(5,2)$, and $(6,1)$ which yield a sum of seven, while only the single pair $(1,1)$ gives a sum of two. This means rolling a sum of seven is six times more likely than rolling a sum of two.

Events in a Sample Space

Definition 3.2 *Events* *are subsets of a sample space in which we might be interested. We can describe an event using words or with more formal set notation.*

Example 3.5 *Suppose we roll a single die. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, let E be the event 'an even number is rolled' and let G be the event 'a number greater than four is rolled'. Then we can write $E = \{2, 4, 6\}$ and $G = \{5, 6\}$.*

Example 3.6 *Consider a point selected at random inside a square with corners at $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$. Let $\Omega = \{(x,y) \mid 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1\}$, and let T be the event 'the point is in the top half of the square'. Then using set notation we would write $T = \{(x,y) \mid 0 \leq x \leq 1 \text{ and } 0.5 \leq y \leq 1\}$.*

Since events $A, B \subseteq \Omega$ are sets, we can form the events $A \cup B$ (at least one of A and B occur), $A \cap B$ (both A and B occur), and A^c (A does not occur). Te set operations allow us to build up complex events out of simpler pieces, or vice-versa, break a complex event into simpler pieces that are easier to deal with.

Let's return for a moment to rolling a pair of dice. If one of the dice shows a five, is it possible the sum is seven? Yes, since the outcomes $(5,2)$ and $(2,5)$ are elements of both events $F = $ 'one of the dice shows five' and $S = $ 'the sum is seven', that is, $F \cap S = \{(5,2), (2,5)\}$.

If one of the dice shows a five, is it possible the sum is three? No. If one die shows a five, the lowest possible sum is six. There are no elements common to the events $F = $ 'one of the dice shows five' and $T = $ 'the sum is three', that is, $F \cap S = \emptyset$.

Definition 3.3 *Two events A and B in a sample space Ω are **mutually exclusive** if $A \cap B = \emptyset$, that is, if there are no outcomes in Ω where both events occur.*

Example 3.7 *Consider an experiment where a coin is flipped, and then a die is rolled, and let $\Omega = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$. The events $H = $ 'a head is flipped' and $E = $ 'an even number is rolled' are not mutually exclusive, since $H \cap E = \{H2, H4, H6\}$.*

Example 3.8 *Suppose you're waiting for a bus, and you measure the number of seconds it takes for the bus to arrive. Let $E = $ 'the bus arrives in under 3 minutes' and $F = $ 'the bus takes more than 5 minutes to arrive'. Then E and F are mutually exclusive. Formally, let $\Omega = \{x \in \mathbb{R} \mid x \geq 0\}$. In interval notation, we can write $E = [0, 180)$ and $F = (300, \infty)$, then clearly $E \cap F = \emptyset$.*

3.2 Probability Models & Probability Laws

Consider a single roll of a fair six-sided die, and let $\Omega = \{1, 2, 3, 4, 5, 6\}$. Since the die is fair, we should assign each of the six outcomes an equal probability of one-sixth.

If we want to find the probability any event $E \subseteq \Omega$ occurs, we can simply sum up the probabilities of each of the outcomes. If $E = \{2, 4, 6\}$ (an even number is rolled), then we write $P(E) = \frac{3}{6}$. To assign a probability to any event in this sample space, we can simply count up the number of outcomes in the event, and divide that number by six to obtain its probability.

Definition 3.4 *A sample space Ω, together with a way of assigning probabilities to all events $E \subset \Omega$, is known as a **probability model**.*

Example 3.9 *Describe a probability model on the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a six-sided die where each outcome is equally likely except for six, which is twice as likely as all the other outcomes.*

Let $P(\{1\}) = q$. Then $P(\{6\}) = 2q$, and since it's certain that one of the six outcomes will occur,

$$P(\{1, 2, 3, 4, 5, 6\}) = 1$$
$$P(\{1\}) + P(\{2\}) + \ldots + P(\{5\}) + P(\{6\}) = 1$$
$$q + q + \ldots + q + 2q = 1$$
$$7q = 1$$
$$q = \tfrac{1}{7}$$

Therefore, the desired probability model assigns the outcomes one through five the probability $\frac{1}{7}$, and the outcome six the probability $\frac{2}{7}$.

Key Point 🔑

For a *countable* sample space, a probability model is an assignment of probabilities to the elements of Ω. Once this is done, the probability of any event can be computed by summing the probabilities of the outcomes it contains. For an uncountable sample space, the situation is more subtle, as you'll see in the following example.

Example 3.10 *Let $\Omega = \{(x, y) \mid 0 \le x \le 1 \text{ and } 0 \le y \le 1\}$, a unit square in $\mathbb{R}^2$. How can we describe a probability model in which a point in this unit square is selected at random?*

Consider the smaller square $E = \{(x, y) \mid 0 \le x \le \frac{1}{2} \text{ and } 0 \le y \le \frac{1}{2}\}$ shaded below. Since this smaller square occupies one-quarter of the area, we should certainly have $P(E) = \frac{1}{4}$.

This probability shouldn't change if we move the smaller square around inside the larger one, or even if we cut it into pieces and move them around, as long as we don't add or remove any area. This leads us a to a simple realization: in this model, the probability of any event is its area. Note that the area of Ω is one since it's a unit square. If not, we would take $P(E)$ to be the proportion of the area of Ω that E occupies.

So the probability model has a very simple description. Each event $E \subseteq \Omega$ is assigned a probability equal to its area.

Remark ✊

There are some very interesting subtleties of probability models on uncountable sample spaces that the previous example gives us an opportunity to discuss.

Firstly, each element of Ω is a point, with no area, and hence an event such as $E = \{(\frac{1}{3}, \frac{2}{3})\}$ is assigned the probability zero. Every event which contains only a single point has probability zero, but one of them will occur. The implication is that events with probability zero can sometimes occur. This is indeed the way things work in the world of uncountable sample spaces.

Secondly, it's not clear whether we've actually described the probability model completely. Consider an event like Q = 'the point selected has rational coordinates'. This is an event which could occur, and a well-defined subset of Ω, but what is the area of Q? This is a deep question which ultimately leads to a branch of mathematics called measure theory, discussed in more advanced courses.

Basic Laws of Probability

A probability model assigns of probabilities to events, but not every kind of assignment is permitted. If would be absurd, for example, to create a probability model on the sample space $\Omega = \{H, T\}$ where $P(\{H\}) = 0.6$ and $P(\{T\}) = 0.9$.

Proposition 3.1 *Any probability model on a sample space Ω must satisfy the following.*

- $P(\Omega) = 1$, *and* $0 \leq P(E) \leq 1$ *for any event* $E \subseteq \Omega$.

- $P(E^c) = 1 - P(E)$ *for any event* $E \subseteq \Omega$.

- *If A and B are mutually exclusive events, then* $P(A \cup B) = P(A) + P(B)$.

It's reasonable to think about these requirements as either the defining characteristics of a probability model, or as the basic laws of probability. Note that the third law generalizes to sequences of mutually exclusive events. Formally, as long as $A_i \cap A_j = \emptyset$ for all i and j between 1 and n,

$$P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n).$$

Informally, whenever any finite number of events in Ω do not overlap (in the sense of a Venn diagram), the probability at least one occurs is the sum of their individual probabilities.

Remark ✊

The second condition above actually follows from the first and third. It's a good exercise to prove this for yourself (*Hint*: Read the next example, and note that E and E^c are always mutually exclusive events).

Example 3.11 *Use the basic probability laws to show that $P(\emptyset) = 0$ in any probability model.*

Note that since $\Omega \cap \emptyset = \emptyset$, the events Ω and $\emptyset$ are mutually exclusive. Therefore

$$P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset),$$

but since $P(\Omega) = 1$, the equation above says $1 = 1 + P(\emptyset)$, and hence $P(\emptyset) = 0$.

Example 3.12 *What is the probability that a single card drawn from a standard shuffled deck is a king or not a face card?*

Let $K = $ 'the card drawn is a king' and $F = $ 'the card drawn is a face card'. Since the events K and F^c are mutually exclusive (a card cannot be both a king and not a face card), we have

$$P(K \cup F^c) = P(K) + P(F^c) = P(K) + 1 - P(F) = \tfrac{4}{52} + 1 - \tfrac{12}{52} = \tfrac{44}{52}.$$

Inclusion-Exclusion

What if we want to know $P(A \cup B)$ for events A and B which are not mutually exclusive? Fortunately, the inclusion-exclusion principle in Section 2.3 is also a valid law of probability. For any two events A and B in any probability model,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Example 3.13 *What is the probability that a single card drawn from a standard shuffled deck is red or not a face card?*

Let $R = $ 'the card drawn is red' and $F = $ 'the card drawn is a face card'. The events R and F^c are not mutually exclusive (the seven of hearts, for example, is red and not a face card). In fact, there are twenty cards that are red and not face cards, ace through ten of hearts and ace through ten of diamonds. Thus,

$$P(R \cup F^c) = P(R) + P(F^c) - P(R \cap F^c) = \tfrac{26}{52} + 1 - \tfrac{12}{52} - \tfrac{20}{52} = \tfrac{46}{52}.$$

Alternatively, notice that $(R \cup F^c)^c = R^c \cap F$ by DeMorgan's law. Now $R^c \cap F = $ 'the card drawn is a black face card', and there are six of these, so we have

$$P(R \cup F^c) = 1 - P((R \cup F^c)^c) = 1 - P(R^c \cap F) = 1 - \tfrac{6}{52} = \tfrac{46}{52}.$$

Example 3.14 *What is the probability that a single card drawn from a standard shuffled deck is either a diamond, a queen, or has an even number on it?*

Applying the three-event version of the inclusion-exclusion principle to the events $D = $ 'the card drawn is a diamond', $Q = $ 'the card drawn is a queen', and $E = $ 'the card drawn has an even number on it',

$$P(D \cup Q \cup E) = P(D) + P(Q) + P(E) - P(D \cap Q) - P(Q \cap E) - P(E \cap D) + P(D \cap Q \cap E)$$
$$= \tfrac{13}{52} + \tfrac{4}{52} + \tfrac{20}{52} - \tfrac{1}{52} - \tfrac{0}{52} - \tfrac{5}{52} + \tfrac{0}{52} = \tfrac{31}{52}.$$

3.3   Conditional Probability

Suppose a single die is rolled. The result is unknown, but we're told it was an even number. What is the probability the result was six?

Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, $S =$ 'the outcome was a six', and $E =$ 'the outcome was an even number'. Initially, each outcome in Ω is equally likely, and $P(S) = \frac{1}{6}$. However, given the knowledge that an even number was rolled, the sample space has been reduced from all the outcomes in Ω to only those in E. Now that E has occurred, each of the three even outcomes has probability $\frac{1}{3}$, while each of the odd outcomes has probability zero.

We write $P(S \,|\, E) = \frac{1}{3}$ to indicate that the probability of S under the assumption that E has occurred is $\frac{1}{3}$, and we read $P(S \,|\, E)$ as the **conditional probability** of S given E.

Example 3.15 *In a certain school, 35% of students receive an A grade in reading, 22% of students receive an A grade in mathematics, and 16% of students receive an A grade in both. Of those students who received an A grade in reading, what proportion also received an A grade in mathematics?*

Let the number of students at the school be denoted by n. Then the number that received an A in reading is $0.35n$, the number that received an A in mathematics is $0.22n$, and the number that received an A in both is $0.16n$. If we are considering only students who received an A in reading, we are dealing with a pool of $0.35n$ students. The only students in this pool who received an A in math must have received an A in both subjects, so there are total of $0.16n$ such students.

Thus, the proportion of A students in reading who also received an A in math is $\frac{0.16n}{0.35n} = \frac{0.16}{0.35} \simeq 0.46 = 46\%$. Note that if we select a student at random and define the events $R =$ 'the student received an A in reading' and $M =$ 'the student received an A in math', then we can rewrite this calculation as

$$P(M \,|\, R) = \frac{P(M \cap R)}{P(R)} = \frac{0.16}{0.35} \simeq 0.46.$$

We'll use this idea to formally define conditional probability, but it's often more productive to view conditional probability as a reduction of the sample space. In the example above, if you gathered all students who received an A in reading in a room together and asked those who received an A in math to raise their hands, 46% of them would raise their hands (if they're honest).

Definition 3.5 *Given two events E and F in some sample space Ω, the conditional probability of E given F is denoted $P(E \,|\, F)$, and defined by*

$$P(E \,|\, F) = \frac{P(E \cap F)}{P(F)}.$$

Note that division by zero could be problematic here, so we'll only define the conditional probability of E given F when $P(F) \neq 0$.

The Intersection Law

This definition can be used either to compute $P(E \mid F)$ when $P(E \cap F)$ and $P(F)$ are known, or to compute $P(E \cap F)$ when $P(E \mid F)$ and $P(F)$ are known. In fact, we'll often use it in this second way, so it's worth renaming some variables and moving terms around to make it easier to use in that direction.

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$
$$P(B \mid A)P(A) = P(B \cap A)$$
$$P(B \mid A)P(A) = P(A \cap B)$$
$$P(A \cap B) = P(A)P(B \mid A)$$

Informally, if A and B both occur, then A must occur and B must occur in this new setting where the event A has already happened.

Example 3.16 *Suppose we draw two cards from a shuffled deck without replacement. What is the probability that both are hearts?*
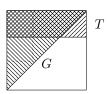
Let H_1 = 'the first card drawn is a heart' and H_2 = 'the second card drawn is a heart'. Then both cards are hearts when $H_1 \cap H_2$ occurs.

$$P(H_1 \cap H_2) = P(H_1)P(H_2 \mid H_1) = \frac{13}{52} \cdot \frac{12}{51} = \frac{156}{2652} \simeq 0.059$$

The key point is that if H_1 has occurred, a heart has been removed from the deck. There are then 51 cards remaining, 12 of which are hearts, so $P(H_2 \mid H_1) = \frac{12}{51}$.

Example 3.17 *Consider a point selected at random in a unit square, as in Example 3.10. As before, we take $\Omega = \{(x,y) \mid 0 \le x \le 1 \text{ and } 0 \le y \le 1\}$. If the point has $x < y$, what is the probability it lies in the top third of the square?*

We're looking for $P(T \mid G)$ where $T = \{(x,y) \mid 0 \le x \le 1 \text{ and } \frac{2}{3} \le y \le 1\}$ and $G = \{(x,y) \mid 0 \le x < y \le 1\}$. To compute this probability, we can draw a picture.

Recall from Example 3.10 that in this probability model, the probability of any event is the area it occupies. The small triangle in the top-right (the event $T \setminus G$) is half of a $\frac{1}{3}$ by $\frac{1}{3}$ square, hence it has area $\frac{1}{18}$, so

$$P(T \cap G) = \tfrac{1}{3} - \tfrac{1}{18} = \tfrac{5}{18} \quad \text{and} \quad P(G) = \tfrac{1}{2}$$

Therefore, $P(T \mid G) = \dfrac{P(T \cap G)}{P(G)} = \dfrac{\frac{5}{18}}{\frac{1}{2}} = \frac{5}{9}.$

Note that the intersection law generalizes to sequences of events. Formally,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2 \,|\, A_1)P(A_3 \,|\, A_1 \cap A_2) \cdots P(A_n \,|\, A_1 \cap A_2 \cap \ldots \cap A_{n-1})$$

where the ith factor on the right is the conditional probability of A_i under the assumption that all prior events in the sequence have occurred.

Example 3.18 *What is the probability that when we draw four cards from a deck without replacement, all four cards are different suits?*

Let $D_i =$ 'the ith card is a different suit from all those cards the came before', then $P(D_1) = 1$, since no cards were drawn before the first, and we calculate

$$P(D_1 \cap D_2 \cap D_3 \cap D_4) = P(D_1)P(D_2 \,|\, D_1)P(D_3 \,|\, D_1 \cap D_2)P(D_4 \,|\, D_1 \cap D_2 \cap D_3)$$
$$= 1 \cdot \frac{39}{51} \cdot \frac{26}{50} \cdot \frac{13}{49} \approx 0.1055.$$

Example 3.19 *Suppose that seven lightbulbs are in a box, but only two work. Bob will randomly select lightbulbs one at a time and test them. What is the probability he'll find both working lightbulbs after exactly three tests?*

Let $W_i =$ 'the ith lightbulb tested works'. Then we need W_3 to occur, and exactly one of W_1 and W_2, since if W_1 and W_2 both occur then Bob will be finished after only two tests.

$$P((W_1 \cap W_2{}^c \cap W_3) \cup (W_1{}^c \cap W_2 \cap W_3))$$
$$= P(W_1 \cap W_2{}^c \cap W_3) + P(W_1{}^c \cap W_2 \cap W_3)$$
$$= P(W_1)P(W_2{}^c \,|\, W_1)P(W_3 \,|\, W_1 \cap W_2{}^c) + P(W_1{}^c)P(W_2 \,|\, W_1{}^c)P(W_3 \,|\, W_1{}^c \cap W_2)$$
$$= \frac{2}{7} \cdot \frac{5}{6} \cdot \frac{1}{5} + \frac{5}{7} \cdot \frac{2}{6} \cdot \frac{1}{5} = \frac{10}{210} + \frac{10}{210} \simeq 0.095$$

Note that $W_1 \cap W_2{}^c \cap W_3$ and $W_1{}^c \cap W_2 \cap W_3$ are mutually exclusive, so we can pass from the first to the second line without using the inclusion-exclusion principle.

3.4 The Law of Total Probability

If two cards are drawn from a shuffled deck, what is the probability the second card drawn is a heart? It may be tempting to reply that the question is not well-posed, and in order to answer, we would require more information, namely, whether the first card drawn was a heart or not.
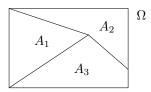
However, the question above is indeed well-posed. If we repeat the experiment a very large number of times, then in some proportion of trials, the second card will be heart. In order to compute this value analytically, we'll need the result below.

Theorem 3.1 (*Law of Total Probability*) *If* $\{A_i\}_{i=1}^n$ *is any mutually exclusive sequence of events with* $\bigcup_{i=1}^n A_i = \Omega$*, then for any event* B,

$$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_n)P(A_n).$$

Note that we require $P(A_i) > 0$ *for each* A_i*, so that the conditional probabilities on the right are defined.*

When a sequence $\{A_i\}_{i=1}^n$ is mutually exclusive and $\bigcup_{i=1}^n A_i = \Omega$, we say that the sequence *partitions the sample space*. This means that every possible outcome is in exactly one of the A_i. The law of total probability says that for any partition of Ω, we can compute the probability of an event by taking a weighted average of conditional probabilities across the partition, where the weights are the probabilities that the corresponding piece of the partition occurs.

We can obtain a special case of the law of total probability which occurs frequently in problems by considering the sequence A, A^c, for some event A. Note that regardless of what event A happens to be, this sequence always partitions Ω.

Corollary 3.1 *If* A *is an event with* $0 < P(A) < 1$*, then for any event* B,

$$P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c).$$

Example 3.20 *If two cards are drawn from a shuffled deck, what is the probability the second card drawn is a heart?*

Let $H_1 = $ 'the first card is a heart' and $H_2 = $ 'the second card is a heart'. We'll apply the corollary above to compute the probability of H_2.

$$P(H_2) = P(H_2 \mid H_1)P(H_1) + P(H_2 \mid H_1{}^c)P(H_1{}^c)$$
$$= \frac{12}{51} \cdot \frac{1}{4} + \frac{13}{51} \cdot \frac{3}{4} = \frac{51}{204} = \frac{1}{4}$$

8 *This result makes intuitive sense. Because we're computing* $P(H_2)$ *in the absence of any information about the first card, we may as well have put that first card on the bottom of the deck after it's drawn without ever looking at it, so we're still drawing a card from a complete shuffled deck.*

Pf. (of Theorem 3.1) If the sequence $\{A_i\}_{i=1}^n$ partitions Ω, and $B \subseteq \Omega$,

$$\Omega = A_1 \cup A_2 \cup \ldots \cup A_n$$
$$B \cap \Omega = B \cap (A_1 \cup A_2 \cup \ldots \cup A_n)$$
$$B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n).$$

Now the sequence of events $B \cap A_1$, $B \cap A_2$, ..., $B \cap A_n$ is mutually exclusive because $\{A_i\}_{i=1}^n$ was mutually exclusive, and this new sequence is obtained by reducing the number of outcomes in each event. The events had empty intersections initially, so removing outcomes will keep the intersections empty.

$$\begin{aligned}
P(B) &= P((B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)) \\
&= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n) \\
&= P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_n \cap B) \\
&= P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \cdots + P(A_n)P(B \mid A_n) \\
&= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_n)P(A_n)
\end{aligned}$$

$\square$
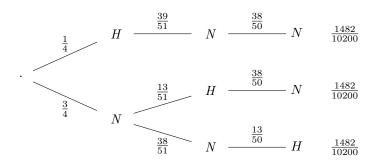
Tree Diagrams

Tree diagrams are a nice graphical tool for organizing calculations, and have the law of total probability built in to their structure. Each column represents a step in an experiment. If we draw three balls from an urn, for example, each column will represent a draw. If we roll a die and then flip a coin, the first column will represent the roll and the second will represent the flip.

Example 3.21 *What is the probability that, when three cards are drawn from a shuffled deck without replacement, exactly one heart appears?*

In the tree diagram below, the three columns represent three draws from a deck of cards. When a draw results in a heart, we denote that with an H, and when a draw does not result in a heart, we denote that with an N.

Each edge is labelled with a conditional probability. For example, the top-center edge which runs from H to N is labelled with $P(N_2 \mid H_1)$, the probability a non-heart is drawn on the second draw, after drawing a heart on the first. Note that we only depict the branches where the desired event ('exactly one heart appears') occurs. To find the probability of this event, we simply multiply along each branch, and add the results. If $E =$ 'exactly one heart is drawn', then

$$P(E) = \frac{1482}{10200} + \frac{1482}{10200} + \frac{1482}{10200} = \frac{4446}{10200} \simeq 43.6\%.$$
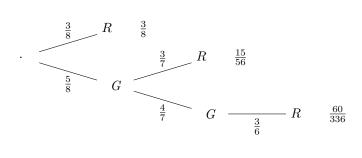
This works because the branches represent mutually exclusive events (at each fork, a given trial of the experiment can proceed along only one path), so we can add the probabilities of the branches without having to appeal to the inclusion-exclusion principle. Furthermore, the intersection law states that we can compute the probability of all events along a branch occurring by multiplying their conditional probabilities. Note that if we applied the law of total probability in the above example, partitioning the sample space with $\{H_1 \cap H_2, H_1 \cap N_2, N_1 \cap H_2, N_1 \cap N_2\}$, we would obtain

$$
\begin{aligned}
P(E) &= P(E \mid H_1 \cap H_2)P(H_1 \cap H_2) + P(E \mid H_1 \cap N_2)P(H_1 \cap N_2) \\
&\quad + P(E \mid N_1 \cap H_2)P(N_1 \cap H_2) + P(E \mid N_1 \cap N_2)P(N_1 \cap N_2) \\
&= 0 + P(E \mid H_1 \cap N_2)P(N_2 \mid H_1)P(H_1) + P(E \mid N_1 \cap H_2)P(H_2 \mid N_1)P(N_1) \\
&\quad + P(E \mid N_1 \cap N_2)P(N_2 \mid N_1)P(N_1) \\
&= P(N_3 \mid H_1 \cap N_2)P(N_2 \mid H_1)P(H_1) + P(N_3 \mid N_1 \cap H_2)P(H_2 \mid N_1)P(N_1) \\
&\quad + P(H_3 \mid N_1 \cap N_2)P(N_2 \mid N_1)P(N_1)
\end{aligned}
$$

and this is exactly how we calculated the result above. Each path from left to right corresponds to a term in the sum above. The event E never occurs if two hearts are drawn, so $P(E \mid H_1 \cap H_2) = 0$ and that term (which corresponds to a branch not depicted in the diagram) vanishes. Although working through the notation is good practice, I expect you'll agree that in cases like this, the tree diagram is more intuitive and simplifies the bookkeeping.

If at any point in the tree diagram, the event we're interested in is sure to occur, there's no need to continue along that branch any further.

Example 3.22 *Suppose that an urn contains three red and five green balls. If three balls are drawn from the urn without replacement, what is the probability that at least one red ball will be drawn?*
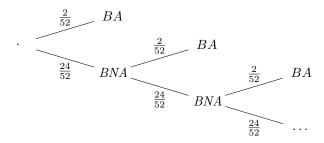
As soon as a red ball is drawn, there's no need to continue the branch. Regardless of what happens from that point on, a red ball will have been drawn. If $A =$ 'at least one red ball is drawn', then

$$
P(A) = \frac{3}{8} + \frac{15}{56} + \frac{60}{336} = \frac{276}{336} \simeq 82.1\%.
$$

Example 3.23 *Suppose we draw cards with replacement from a shuffled deck until the first ace appears. What is the probability that all cards drawn are black?*

Here BA denotes 'black ace' and BNA denotes 'black non-ace'. There's no limit to the number of draws that could occur before the first black ace appears, so the tree diagram is infinite. However, we can still sum the products along each branch. Let $OB =$ 'only black cards are drawn', then

$$P(OB) = \frac{2}{52} + \frac{24}{52} \cdot \frac{2}{52} + \frac{24}{52} \cdot \frac{24}{52} \cdot \frac{2}{52} + \frac{24}{52} \cdot \frac{24}{52} \cdot \frac{24}{52} \cdot \frac{2}{52} + \cdots$$

$$= \frac{2}{52} \left(1 + \frac{24}{52} + \left(\frac{24}{52} \right)^2 + \left(\frac{24}{52} \right)^3 + \cdots \right)$$

$$= \frac{2}{52} \left(\sum_{i=0}^{\infty} \left(\frac{24}{52} \right)^n \right)$$

$$= \frac{2}{52} \cdot \frac{1}{1 - \left(\frac{24}{52} \right)} = \frac{1}{14} \simeq 7.1\%$$

Note that the sum of the infinite series above was evaluated with the geometric series sum formula, $a + ar + ar^2 + \cdots = \frac{a}{1-r}$ when $-1 < r < 1$.

3.5 Independence

Suppose we flip a coin and then roll a die. Let $H =$ 'a head is flipped', $E =$ 'an even number is rolled' and $L =$ 'a number larger than three was rolled'. Note that

$$P(E) = \frac{1}{2} \quad \text{and} \quad P(E \,|\, H) = \frac{1}{2}$$

since knowing that a head was flipped doesn't change the probability an even number is rolled. On the other hand, we can calculate

$$P(E) = \frac{1}{2} \quad \text{and} \quad P(E \,|\, L) = \frac{2}{3}.$$

In this case, knowledge that the number rolled was a four, five, or six increases the probability that the result of the roll was even. The events E and H are said to be *independent*, while the events E and L are *not independent*.

Definition 3.6 *Events A and B are called **independent** if $P(B) = P(B \,|\, A)$.*

Example 3.24 *Suppose we flip a coin two times. Note that $\Omega = \{HH, HT, TH, TT\}$ is a sample space with equally likely outcomes for the experiment.*

If we consider the events $H_1 =$ 'the first flip is a head' and $H_2 =$ 'the second flip is a head', then $P(H_2) = \frac{1}{2}$ and $P(H_2 \mid H_1) = \frac{1}{2}$, so these events are independent.

If we instead let $H_1 =$ 'the first flip is a head' and $T_1 =$ 'the first flip is a tail', then $P(T_1) = \frac{1}{2}$ and $P(T_1 \mid H_1) = 0$, so these two events are not independent.

Example 3.25 *Consider two cards drawn from a shuffled deck. Let $S_1 =$ 'a spade is drawn on the first draw' and $S_2 =$ 'a spade is drawn on the second draw'. Then S_1 and S_2 are independent provided the draws are done with replacement, but not independent if the draws are done without replacement. In this second case, $P(S_2 \mid S_1) = \frac{12}{51}$, but using the law of total probability, $P(S_2) = \frac{1}{4}$.*

Using the definition of the conditional probability $P(A \mid B)$, we can state what it means for two events to be independent in a different way.

$$P(A) = P(A \mid B) \;\leftrightarrow\; P(A) = \frac{P(A \cap B)}{P(B)} \;\leftrightarrow\; P(A)P(B) = P(A \cap B)$$

Thus, when A and B are independent events, $P(A \cap B) = P(A)P(B)$. This relation is often used as the definition of independence. As we just saw, it's equivalent to the definition we gave earlier but also applies when $P(A) = 0$ or $P(B) = 0$ (in these cases $P(A \mid B)$ or $P(B \mid A)$ would be undefined).

Key Point 🔑

Independence is a symmetric relation, which you can interpret intuitively as 'the probability of one event occurring is not influenced by the occurrence of the other' without worrying about which event comes first. The alternate form of the definition $P(A \cap B) = P(A)P(B)$ confirms this, as swapping the roles of A and B yields the same equation.

Theorem 3.2 *If A and B are independent, then A and B^c are independent.*

Pf. Note that $A = (A \cap B) \cup (A \cap B^c)$ and hence $P(A) = P(A \cap B) + P(A \cap B^c)$ since the two events on the right are mutually exclusive. Isolating $P(A \cap B^c)$,

$$\begin{aligned}
P(A \cap B^c) &= P(A) - P(A \cap B) \\
&= P(A) - P(A)P(B) \\
&= P(A)(1 - P(B)) \\
&= P(A)P(B^c).
\end{aligned}$$

$\square$

Corollary 3.2 *If A and B are independent, then A^c and B^c are independent.*

Pf. If A and B are independent, then A and B^c are independent by Theorem 3.2, but independence is symmetric, so applying Theorem 3.2 again, this time to B^c and A, we conclude B^c and A^c are independent. $\square$

Remark ☝

The upshot of the previous theorem and its corollary is that the independence relation is unaffected by complements. This can be useful, since calculating $P(A^c \,|\, B)$, or $P(A \,|\, B^c)$, or $P(A^c \,|\, B^c)$ may be easier than calculating $P(A \,|\, B)$.

Given a sequence of events $\{A_i\}_{i=1}^n$, we say this sequence is independent if the occurrence of any collection of these events does not influence the chance that any other event in the sequence occurs. Formally, it's easiest to state this by extending the relation $P(A \cap B) = P(A)P(B)$ to all subsequences.

Definition 3.7 *The sequence of events $\{A_i\}_{i=1}^n$ is independent if for any subsequence $A_{i_1}, A_{i_2}, \dots, A_{i_k}$ of events taken from $\{A_i\}_{i=1}^n$,*

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$$

Example 3.26 *Suppose we toss a fair coin twice. Let $H_1 =$ 'The first toss is heads', $H_2 =$ 'The second toss is heads', and $E =$ 'Exactly one toss results in heads'. As we already observed, H_1 and H_2 are independent. Moreover, $P(E) = \frac{1}{2}$ since exactly one head is flipped in two of the four equally likely outcomes, and $P(E \,|\, H_1) = \frac{1}{2}$ since assuming H_1 occurs, E occurs when the second flip is a tail. Thus, H_1 and E are independent, and essentially the same argument shows H_2 and E are independent.*

Since each pair of events in the sequence $\{H_1, H_2, E\}$ is independent, we say this sequence is pairwise independent. However, note that if we assume both H_1 and H_2 occur, the probability E occurs becomes zero, that is, $P(H_1 \cap H_2 \cap E) = 0$, but we can calculate $P(H_1)P(H_2)P(E) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$.

The example above illustrates that even if every pair of events in a sequence is independent, there could still be some event in the sequence whose probability is influenced by the joint occurrence of some larger collection of other events. In other words, for sequences of events, pairwise independence does not imply independence.

Example 3.27 *Suppose that a 64 bit sequence (a sequence of 64 zeros and ones) is sent over an unreliable channel. The probability that each bit gets flipped is 0.005 independent of what happens to the other bits. What is the probability that the entire sequence is received without errors?*

Let $F_i =$ 'the ith bit gets flipped', then $P(F_i) = 0.005$, and the probability that the ith bit is received correctly is $P(F_i^c) = 1 - 0.005 = 0.995$. If $E =$ 'the entire string is received without any flipped bits' then

$$
\begin{aligned}
P(E) &= P(F_1^c \cap F_2^c \cap \dots \cap F_{64}^c) \\
&= P(F_1^c)P(F_2^c) \cdots P(F_{64}^c) \\
&= 0.995 \cdot 0.995 \cdot \dots \cdot 0.995 \\
&\simeq 0.72556 \simeq 72.6\%
\end{aligned}
$$

Chapter 4

Random Variables

4.1 Discrete Random Variables

Definition 4.1 *A **random variable** is a function from a sample space Ω to $\mathbb{R}$. If the range of this function is countable, the random variable is called discrete. In this section, we'll deal only with discrete random variables.*

Since a sample space is used to represent the possible outcomes of an experiment, a random variable associates a real number with each possible outcome. In other words, a random variable is a variable whose value is determined by the outcome of an experiment involving some element of chance.

Random variables are typically denoted using capital letters at the end of the alphabet, such as X, Y, or Z, and the probability the random variable X takes the real number value x is denoted $P(X = x)$.

Example 4.1 *Let X denote the outcome of single roll of a fair die. Then X is a random variable defined on the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$.*

Not surprisingly, $P(X = 1) = \frac{1}{6}$, $P(X = 2) = \frac{1}{6}$, and so on. In general, $P(X = x) = \frac{1}{6}$ for all $x \in \{1, 2, 3, 4, 5, 6\}$ since the die is fair. We could also ask for $P(X > 4)$, and again you should not be surprised that $P(X > 4) = P(X = 5) + P(X = 6) = \frac{2}{6}$.

Remark ☛
It may take some time to get used to this new notation. If X is a random variable, an equation like $X = 2$ or an inequality like $X > 2$ is an event, that may or may not occur depending on the value X takes, so the expression $P(X = 2)$ denotes the probability of an event.

Example 4.2 *Let Y denote the number of heads in three flips of a fair coin. Then Y is a random variable defined on the sample space $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.*

The range of Y is $\{0, 1, 2, 3\}$. We can calculate the probability Y takes any of these values with the techniques of the previous chapter, and summarize the results in. a table.

$$P(Y = 0) = P(\{TTT\}) = \frac{1}{8}$$
$$P(Y = 1) = P(\{HTT, THT, TTH\}) = \frac{3}{8}$$
$$P(Y = 2) = P(\{HHT, HTH, THH\}) = \frac{3}{8}$$
$$P(Y = 3) = P(\{HHH\}) = \frac{1}{8}$$

y	$P(Y = y)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

Key Point 🔑

When we actually perform the experiment and flip the coin three times, the random variable will take a real number value, known as a ***realization***. In the table above, the leftmost column is labelled with every possible realization of Y. Upper case letters will denote random variables, and the corresponding lower case letters will be used for realizations.

Probability Mass Functions

Given a random variable X, the function $f_X : \mathbb{R} \to \mathbb{R}$ that assigns to each possible realization of X the probability it occurs (the function which is described by the table in our example above) is called the probability mass function of X. If x is any value outside the range of X, then we omit this value from the table.

Definition 4.2 *The **probability mass function** (abbreviated pmf) f_X for the random variable X is defined by $f_X(x) = P(X = x)$. We often represent this function with a table, and visualize it with a figure known as a lollipop plot.*
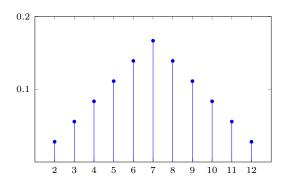
Example 4.3 *Let X denote the sum of the two numbers showing when a pair of fair six-sided dice is rolled. Graph the probability mass function f_X with a lollipop plot.*

Each die ie equally likely to show any value in the set $A = \{1, 2, 3, 4, 5, 6\}$, so for a pair of dice, all outcomes in $\Omega = A \times A$ occur with probability $\frac{1}{36}$.

$D_1 \backslash^{D_2}$	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Thus, $P(Y = 2) = \frac{1}{36}$, $P(Y = 3) = \frac{2}{36}$, $P(Y = 4) = \frac{3}{36}$, and so on. Drawing one lollipop at each realization of X whose height is its probability, we obtain the lollipop plot shown below.

Proposition 4.1 *Suppose that X is a discrete random variable with range $A = \{x_1, x_2, x_3, \ldots\}$. Then the probability mass function f_X satisfies the following properties.*

1. *$f_X(x) \geq 0$ for all $x \in \mathbb{R}$.*

2. *$f_X(x) = 0$ for all $x \notin A$.*

3. *$\sum_{x \in A} f_X(x) = 1$.*

Remark ☝
These properties completely characterize the class of probability mass functions for discrete random variables. In other words, if f is a function with these three properties, then it is the pmf of a discrete random variable.

Example 4.4 *Find the value of k so that the function h below is a pmf.*

$$h(x) = \begin{cases} \left(\frac{1}{3k}\right)^x & \text{if } x = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$
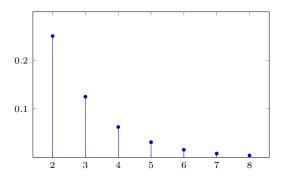
As long as k is positive, h will satisfy properties (i) and (ii) above. The last property allows us to determine k uniquely.

$$\sum_{x \in A} h(x) = \sum_{x=1}^{\infty} \left(\frac{1}{3k}\right)^x = \frac{1}{3k} + \left(\frac{1}{3k}\right)^2 + \left(\frac{1}{3k}\right)^3 + \cdots = \frac{\frac{1}{3k}}{1 - \frac{1}{3k}} = \frac{1}{3k - 1}$$

The geometric series sum formulas was used to evaluate the infinite sum. Setting the last expression equal to 1 gives $3k - 1 = 1$, and hence $k = \frac{2}{3}$. Using this value of k gives the function h below, which we

can graph using a lollipop plot. Only the first few values can be shown, since the graph continues off to the right indefinitely.

$$h(x) = \begin{cases} \left(\frac{1}{2}\right)^x & \text{if } x = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Cumulative Distribution Functions
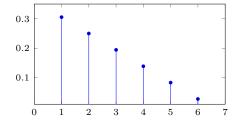
Definition 4.3 *The **cumulative distribution function** (abbreviated cdf) F_X for the random variable X is defined by $F_X(x) = P(X \leq x)$.*
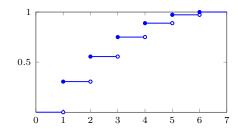
Notice that we use capital letters for cdfs and lower case letters for pmfs. This is conventional and we'll always do it (for a good reason, which we'll see later). It's also worth mentioning that the word 'cumulative' is frequently dropped, and the cdf is simply called the distribution function.

Example 4.5 *Suppose we roll a pair of dice, and let X be the minimum of the two values that appear on the dice. Find the pmf f_X and the cdf F_X.*

The random variable X takes integer values between 1 and 6 inclusive, and referring back to the table in Example 4.3, we can find the probability of each realization. We have $P(X = 1) = \frac{11}{36}$ *(outcomes in the first row or column)*, $P(X = 2) = \frac{9}{36}$ *(outcomes in the second row to the right of the 3, or in the second column below the 3)*, $P(X = 3) = \frac{7}{36}$ *(outcomes in the third row to the right of the 5, or in the third column below the 5)*, *and so on.*

The graph of the pmf f_X is shown on the left, and the cdf F_X is shown on the right.

Notice that $F_X(2.5) = P(X \leq 2.5) = P(X = 1) + P(X = 2) = \frac{11}{36} + \frac{9}{36} \simeq 0.56$. The cumulative distribution function adds up the probabilities that X takes any value smaller than or equal to its input.

Consider two real numbers a and b with $a \leq b$. Given any random variable X, $F_X(a) \leq F_X(b)$, because if X is smaller than a then it must also be smaller than b, in other words, the event $X \leq a$ is a subset of the event $X \leq b$, and hence it cannot be more likely. This means that a cdf is always non-decreasing. This and other key properties that all cdfs share are listed in the proposition below.

Proposition 4.2 *For any random variable X, the distribution function F_X satisfies the following properties.*

1. $F_X(x) \geq 0$ *for all $x \in \mathbb{R}$.*

2. F_X *is non-decreasing.*

3. $\lim\limits_{x \to -\infty} F_X(x) = 0$ *and* $\lim\limits_{x \to \infty} F_X(x) = 1$.

4. $F_X(x)$ *is right-continuous, meaning* $\lim\limits_{x \to a^+} F_X(x) = F_X(a)$.

> **Remark** ☛
> These properties completely characterize the class of distribution functions, that is, any function satisfying all four properties is the cdf of some random variable.
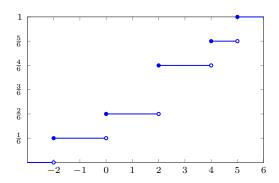
When dealing with cdfs, $F_X(a+)$ is used as an abbreviation of $\lim_{x \to a^+} F_X(x)$, and similarly, $F_X(a-)$ is used to abbreviate $\lim_{x \to a^-} F_X(x)$. Using this notation we can express the probabilities of many events succinctly in terms of the cdf.

Event	Probability
$X \leq a$	$F_X(a)$
$X < a$	$F_X(a-)$
$X = a$	$F_X(a) - F_X(a-)$
$X \geq a$	$1 - F_X(a-)$
$X > a$	$1 - F_X(a)$
$a < X < b$	$F_X(b-) - F_X(a)$
$a \leq X \leq b$	$F_X(b) - F_X(a-)$

Of particular interest is that fact that $P(X = a) = F_X(a) - F_X(a-)$. This is a formal statement of the observation that the possible values of a discrete random variable X are the values where F_X has a discontinuity, and the distance the graph jumps at a is $P(X = a)$.

Example 4.6 *Consider the function whose graph is given below. Check that this function is a cdf for some random variable. If we call the random variable with this cdf Z, find $P(1 < Z < 2)$, $P(Z = 4)$, and $P(Z > 1)$.*

It's easy to see that this function is non-negative, non-decreasing, and tends to 0 as $x \to -\infty$ and 1 as $x \to \infty$ (we're assuming the graph continues off to the left and right along the horizontal lines shown). Furthermore, it's right-continuous since at each jump the value of the function is equal to the limit from the right.

$$P(1 < Z < 2) = F_Z(2-) - F_Z(1) \qquad\qquad P(Z = 4) = F_Z(4) - F_Z(4-)$$
$$= \frac{2}{6} - \frac{2}{6} = 0 \qquad\qquad\qquad\qquad = \frac{5}{6} - \frac{4}{6} = \frac{1}{6}$$

$$P(Z > -1) = 1 - F_Z(-1)$$
$$= 1 - \frac{1}{6} = \frac{5}{6}$$

Independent Random Variables

We've seen in Section 3.5 that events A and B in Ω are independent if $P(A \mid B) = P(A)$, or equivalently, if $P(A \cap B) = P(A)P(B)$. This latter form of the definition extends to discrete random variables in a natural way.

Definition 4.4 *Two random variables X and Y defined on the same sample space Ω are **independent** if $P(X = x \cap Y = y) = P(X = x)P(Y = y)$ for all realizations x and y.*

Remark ☛
The condition $P(X = x \cap Y = y) = P(X = x)P(Y = y)$ for all realizations x and y implies that $P(X \in A \cap Y \in B) = P(X \in A)P(Y \in B)$ for any events $X \in A$ and $Y \in B$. The definition is stated in terms of realizations since it's easier to check that way.

Example 4.7 *Suppose a fair die is rolled twice. Let X be the result of the first roll and Y be the result of the second. Then X and Y are independent random variables, since for any x and y in $\{1, 2, 3, 4, 5, 6\}$, we have $P(X = x \cap Y = y) = \frac{1}{36}$, while $P(X = x) = \frac{1}{6}$ and $P(Y = y) = \frac{1}{6}$.*

Example 4.8 *Let X be a number chosen at random from the set $\{1, 2, 3, 4, 5, 6\}$, and let Y be the remainder when X is divided by 3. Then X and Y are not independent random variables, since, for instance, $P(X = 3 \cap Y = 1) = 0$, because if $X = 3$ then we must have $Y = 0$, but we can calculate $P(X = 3)P(Y = 1) = \frac{1}{6} \cdot \frac{2}{6} \neq 0$.*

As it did in Section 3.5, the notion of independence extends to sequences. If the distribution of each random variable in the sequence is unaffected by knowing the realizations any number of other random variables in the sequence take, then the sequence is independent.

Example 4.9 *Four fair six-sided dice are rolled. What is the probability at least one shows a six?*

Let X_i be the result that shows on the ith die. Then we have

$$
\begin{aligned}
P(X_1 = 6 \cup ... \cup X_4 = 6) &= P((X_1 \neq 6 \cap ... \cap X_4 \neq 6)^c) \\
&= 1 - P(X_1 \neq 6 \cap ... \cap X_4 \neq 6) \\
&= 1 - P(X_1 \neq 6) \cdot ... \cdot P(X_4 \neq 6) \\
&= 1 - \tfrac{5}{6} \cdot ... \cdot \tfrac{5}{6} \\
&= 1 - \left(\tfrac{5}{6}\right)^4.
\end{aligned}
$$

Independence of the random variables X_1, X_2, X_3, and X_4 is used to pass from the intersection on the second line to the product on the third. Note that this example was a pure probability problem, but now that we have access to the language of random variables, we can refer to the outcome on each die in a simple and precise way, without having to define the relevant events, which helps make the solution as clear as possible.

4.2   Expected Value

Suppose we flip a fair coin three times and count the number of heads that appear. If we repeat this experiment many times, what do we expect the average number of heads to be?

We saw in the last section that if X is the number of heads observed, then X is a discrete random variable whose distribution is given in the table below.

y	$P(Y = y)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

If we perform the experiment many times, then we should expect to observe no heads about 12.5% of the time, one head 37.5% of the time, and so on. If we weight each possible number of heads by the probability it occurs, we obtain

$$0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = 1.5.$$

Not surprisingly we should, on average, obtain 1.5 heads in every three flips of a fair coin. This notion of the average value of a random variable is formalized in the definition below.

Definition 4.5 *Given a discrete random variable X with range $A = \{x_1, x_2, x_3, \dots\}$, the **expected value** of X, also called the **mean** of X, is denoted $E(X)$ or μ_X, and is computed by multiplying each possible realization of X with its probability and summing the results.*

$$E(X) = \sum_{x \in A} x \cdot P(X = x) = \sum_{x \in A} x \cdot f_X(x)$$

Warning ⚠
The sum in the definition above could be infinite, in which case it may or not converge. If the sum is not absolutely convergent, then $E(X)$ is undefined.

Remark ☛
Given a statistical variable in a dataset with n individuals, if we define a random variable X to be the value that variable takes on a randomly selected individual in the dataset, then $E(X)$ is indeed the mean of that variable, as defined in Section 1.4, so the definition above agrees with, and generalizes, the definition of the mean given there.

Example 4.10 *Alice needs to decide whether to buy individual metro tickets at \$3.25 each, or an unlimited weekend pass at \$13.75. If she knows she'll take the metro either three, four, or five times this weekend with equal probability, what is the best decision?*

Let X represent the amount of money Alice will spend if she buys individual tickets. Then X is equally likely to take the three values 9.75, 13, and 16.25. Therefore,

$$E(X) = 9.75 \cdot \frac{1}{3} + 13 \cdot \frac{1}{3} + 16.25 \cdot \frac{1}{3} = 13.$$

On the other hand, if she purchases the unlimited pass, she'll pay \$13.75. Thus, if Alice is interested in saving money on average, she should opt for the individual tickets.

Example 4.11 *Consider the random variable Z whose cdf is given. Find $E(Z)$.*

$$F_Z(z) = \begin{cases} 0 & \text{if } z < -3 \\ \frac{1}{3} & \text{if } -3 \le z < 1 \\ \frac{1}{2} & \text{if } 1 \le z < 5 \\ \frac{5}{6} & \text{if } 5 \le z < 6 \\ 1 & \text{if } z \ge 6 \end{cases}$$

Since $P(Z = a) = F_Z(a) - F_Z(a-)$ we can recover the pmf f_Z by examining the discontinuities of F_Z. From the pmf we compute $E(Z)$.

$$f_Z(z) = \begin{cases} \frac{1}{3} & \text{if } z = -3 \\ \frac{1}{6} & \text{if } z = 1 \\ \frac{1}{3} & \text{if } z = 5 \\ \frac{1}{6} & \text{if } z = 6 \end{cases}$$

$$E(Z) = \sum_{z \in A} z \cdot f_Z(z) = -3 \cdot \frac{1}{3} + 1 \cdot \frac{1}{6} + 5 \cdot \frac{1}{3} + 6 \cdot \frac{1}{6} = \frac{11}{6}$$

One immediate application of expected value is to games of chance. Given such a game, if we let X represent the player's profit from playing the game once, then if $E(X) > 0$ it's a **winning bet**, if $E(X) < 0$ it's a **losing bet**, and if $E(X) = 0$, it's a **fair game**.

Example 4.12 *On a roulette wheel, there are 18 black numbers, 18 red numbers, and 2 green numbers. Any wager made on black will double if successful. Find the expected value of a $20 bet on black, assuming each of the numbers on the roulette wheel is equally likely to be selected.*

Let X represent the profit from a bet of $20 on black. Then X can take only values in $A = \{20, -20\}$ since the bet is either won or lost. There are a total of 38 possible outcomes, 18 of which result in a win, so $P(X = 20) = \frac{18}{38}$ and $P(X = -20) = \frac{20}{38}$.

$$E(X) = 20 \cdot \frac{18}{38} + (-20) \cdot \frac{20}{38} = -\frac{40}{38} = -1.05$$

Therefore, this is a losing bet. If you were to sit at a roulette table and repeatedly make $20 bets on black for a very long time, you should expect your long-term losses to total about $1 per game played. Conversely for the casino, they should expect to make a $1 profit for every $20 bet on black played.

Example 4.13 *(St-Petersburg lottery) A coin is flipped until the first tail appears. You are initially given $1 before the coin is flipped for the first time, and every time a head appears, the prize doubles. When the first tail appears, you receive the prize. If the sequence of flips that appears is $HHHT$, for example, you would receive $8. What is the expected value of the prize?*

If we define a random variable X on the sample space $\Omega = \{T, HT, HHT, \ldots\}$ whose value is the prize, then X can take any value in $A = \{1, 2, 4, 8, 16, \ldots\}$, i.e., any value of the form 2^k for $k \in \mathbb{N}$. Furthermore, $X = 2^k$ when the outcome is a sequence with k heads followed by a tail. Thus,

$P(X = 2^k) = \left(\frac{1}{2}\right)^{k+1}$.

$$E(X) = \sum_{x \in A} x P(X = x)$$
$$= \sum_{k=0}^{\infty} 2^k P(X = 2^k)$$
$$= 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + 8 \cdot \frac{1}{16} + \cdots$$
$$= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots \to \infty$$

The expected value is undefined, since the sum diverges by growing arbitrarily large.

This is a very strange result, since it implies that even if you're required to pay \$1 000 000 to play this game, it's still a winning bet. How can this be?

Notice that this is not a game one could actually play, as there's only a finite amount of money in the world. If this example is redone with an upper limit on the payout, the expected value will be defined. Using the current global GDP as the payout limit, the result is around \$45. It's also worth considering that someone with \$$2 \times 10^{100}$ has twice as much money as someone with \$$1 \times 10^{100}$ in a literal sense, but both could spend the rest of their lives throwing briefcases of money into a fire and neither would make a dent in their net worth. In practice, twice as much money doesn't have twice as much utility.
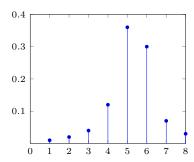
Regardless, if one were actually able to play this game as stated, and play it as many times as one desires, it would be rational do so even if every play cost \$1 000 000, since one would *eventually* win an amount so large that the sum of all prior losses would be rendered insignificant.
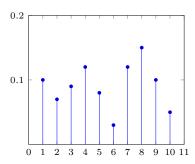
4.3   Variance

Expected value gives a natural measure of center for random variables. In fact, if the lollipops in the lollipop plot of the pmf f_X each represent literal masses whose weights are proportional to their heights, then the value of $E(X)$ is the point on the x-axis where the distribution balances, and that's certainly a reasonable definition for the center.

We can also use expected value to quantify the amount of dispersion present in the distribution of a random variable. Consider for example the two random variables X and Y whose pmfs are given on the left and right respectively.

The values of X tend to be clustered close to the center of the distribution, only very rarely straying under 4 or over 7. On the other hand, Y takes on values far from the center of its distribution quite often. In other words, the distance between the value of X and the center of its distribution is usually smaller than the distance between the value of Y and its center of its distribution.

In Section 1.5 we defined the variance of a statistical variable in a dataset as the average squared distance from the mean, and we'll define the variance of a random variance analagously, as the expected squared distance between a random variable and its expected value.

Definition 4.6 *The **variance** of a discrete random variable X is denoted $\mathrm{Var}(X)$ or σ_X^2, and given by*

$$\mathrm{Var}(X) = E[(X - \mu_X)^2] = \sum_{x \in A} (x - \mu_X)^2 f_X(x).$$

*The **standard deviation** of X is then $\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\mathrm{Var}(X)}$.*

Remark ☞
In this section, many expressions are made clearer by using the notation μ_X, or simply μ, in place of $E(X)$, so keep in mind that μ_X is the expected value of X, which is a constant value at the center of the distribution of X.

Example 4.14 *Determine the expected value, variance, and standard deviation of the random variable X whose pmf is given below.*

$$f_X(x) = \begin{cases} \frac{2}{8} & \text{if } x = 1 \\ \frac{1}{8} & \text{if } x = 3 \\ \frac{3}{8} & \text{if } x = 5 \\ \frac{2}{8} & \text{if } x = 6 \\ 0 & \text{otherwise} \end{cases}$$

First, $E(X) = 1 \cdot \frac{2}{8} + 3 \cdot \frac{1}{8} + 5 \cdot \frac{3}{8} + 6 \cdot \frac{2}{8} = 4$, that is, $\mu_X = 4$. Then we can calculate the variance as

$$\mathrm{Var}(X) = E[(X - \mu_x)^2]$$
$$= (1-4)^2 \cdot \frac{2}{8} + (3-4)^2 \cdot \frac{1}{8} + (5-4)^2 \cdot \frac{3}{8} + (6-4)^2 \cdot \frac{2}{8}$$
$$= 9 \cdot \frac{2}{8} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 4 \cdot \frac{2}{8} = \frac{30}{8}.$$

Thus, the standard deviation of X is $\sigma_X = \sqrt{\frac{30}{8}} \simeq 1.936$.

Example 4.15 *In Big Win Bob's game, each player pays \$1 to place a bet on a number between 1 and 100, a winning number is selected at random, and the prize for a correct bet is \$75. In Conservative Charles' game, each player pays \$1 to bet on a number between 1 and 4, a winning number is selected at random, and the prize for a correct bet is \$3. Which game would you rather play?*

Let X represent a player's net gain in a single play of Bob's game, and let Y represent a player's net gain in a single play of Charles' game.

$$E(X) = 74 \cdot \frac{1}{100} + (-1) \cdot \frac{99}{100} = -0.25$$
$$E(Y) = 2 \cdot \frac{1}{4} + (-1) \cdot \frac{3}{4} = -0.25$$

Thus, it does not matter which game you play in the long run. Over a very large number of plays the games are equivalent, either way the player loses 25¢ per game on average. However, the variances of X and Y are not equal.

$$\mathrm{Var}(X) = (74 - (-0.25))^2 \cdot \frac{1}{100} + (-1 - (-0.25))^2 \cdot \frac{99}{100} \simeq 55.69$$
$$\mathrm{Var}(Y) = (2 - (-0.25))^2 \cdot \frac{1}{4} + (-1 - (-0.25))^2 \cdot \frac{3}{4} \simeq 1.68$$

This large difference in the variances indicates Big Win Bob's game is much riskier. In Bob's game players win rarely but win big, while in Charles' game players win small amounts quite often.

Calculating the variance by hand is a tedious process, and in practice such calculations are essentially always done with computers. It's important to do a few examples by hand though, so that you understand the subtleties and develop some intuition. In the example above, for instance, you can see how the variance of Big Win Bob's game is dominated by the contribution of the rare outcome where the player wins \$100 because the difference between that outcome and the expected value is squared.

Example 4.16 *If X represents the outcome of a single roll of a fair six-sided die, find σ_X.*

First, we compute the mean $E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}$. Then we calculate the variance as

$$Var(X) = \sum_{x \in A} \left(x - \frac{7}{2} \right)^2 \cdot \frac{1}{6} \ , \ where \ A = \{1, 2, 3, 4, 5, 6\}$$
$$= \left(-\frac{5}{2} \right)^2 \cdot \frac{1}{6} + \left(-\frac{3}{2} \right)^2 \cdot \frac{1}{6} + \left(-\frac{1}{2} \right)^2 \cdot \frac{1}{6} + \left(\frac{1}{2} \right)^2 \cdot \frac{1}{6} + \left(\frac{3}{2} \right)^2 \cdot \frac{1}{6} + \left(\frac{5}{2} \right)^2 \cdot \frac{1}{6}$$
$$= \frac{25}{24} + \frac{9}{24} + \frac{1}{24} + \frac{1}{24} + \frac{9}{24} + \frac{25}{24} = \frac{35}{12}$$

The standard deviation is then $\sigma_X = \sqrt{\frac{35}{12}} \approx 1.707$.

Proposition 4.3 *If X is a discrete random variable with $\mathrm{Var}(X) = 0$, then there is some realization x of X such that $P(X = x) = 1$. That is, X takes a single real value with probability one.*

Pf. Suppose there were two different realizations x_1 and x_2 which both have positive probabilities of occurring. Then at least one of $(x_1 - \mu_X)^2$ and $(x_2 - \mu_X)^2$ is nonzero (since μ_X cannot be equal to both x_1 and x_2), but this means at least one of the terms in the sum that defines $E[(X - \mu_X)^2]$ is positive. There are no negative terms in this sum since both squares and probabilities are always non-negative, hence $Var(X) > 0$. $\qquad\square$

4.4 The Uniform Distribution

Definition 4.7 *If X is a discrete random variable whose range is an arithmetic progression, and X takes each value with equal probability, we say that X has a **uniform distribution** on A.*
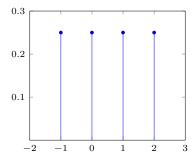
Example 4.17 *Suppose that X represents the result of a single fair die roll, then X is uniformly distributed on $A = \{1, 2, 3, 4, 5, 6\}$, since $P(X = x) = \frac{1}{6}$ for all $x \in A$.*

The notation $X \sim Uniform(n)$ indicates X is a discrete random variable which is uniformly distributed on $A = \{1, 2, 3, ..., n\}$. We could write $X \sim Uniform(6)$ to describe the distribution of the random variable in the example of a die roll above.

If X is uniformly distributed on the set A (some arithmetic progression), then its pmf has the form

$$f_X(x) = \begin{cases} \frac{1}{n} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases},$$

which yields a lollipop plot with a finite number lollipops of the equal height, spaced evenly along the horizontal axis. For example, the lollipop plot below shows the pmf of a random variable uniformly distributed on $A = \{-1, 0, 1, 2\}$.

The expected value of a uniformly distributed variable on the set A is simply the average of the values in A. Formally, if X is uniformly distributed on the set $A = \{a_1, a_2, \ldots, a_n\}$, then

$$E(X) = a_1 \cdot \frac{1}{n} + a_2 \cdot \frac{1}{n} + \cdots + a_n \cdot \frac{1}{n} = \frac{a_1 + a_2 + \cdots + a_n}{n}.$$

In particular, if $X \sim Uniform(n)$, then $E(X) = \frac{1+2+3+\cdots+n}{n} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n+1}{2}$.

Remark ☛

When an element is being chosen from a set, and each element in that set has the same probability of being selected, we say the choice is made *uniformly at random*.

One could 'randomly' select a number from the set $A = \{0, 1, 2, 3\}$ by flipping a coin three times and counting the number of heads, but not all outcomes would be equally likely. Typically when an author uses the words 'at random' in mathematics, what's meant is that each possible outcome is equally likely, not that the outcome is the result of some unspecified random process. To be as clear as possible, it's good practice to use the term *uniformly*.

4.5   The Bernoulli Distribution
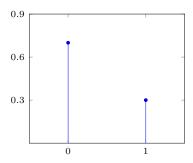
Suppose that we perform an experiment with only two possible outcomes, usually called success and failure, and the probability of success is known to be p. Such an experiment is called a Bernoulli trial, after Swiss mathematician Jacob Bernoulli. If X is a random variable that counts the number of successes in a single such trial (either zero or one), we say that X has a Bernoulli distribution with parameter p.

Definition 4.8 *The random variable X has a **Bernoulli distribution** with parameter p, written $X \sim$ Bernoulli(p), if*

$$f_X(x) = \begin{cases} p & if \ \ x = 1 \\ 1 - p & if \ \ x = 0 \\ 0 & otherwise \end{cases}.$$

In other words, X takes the value 1 with probability p, and 0 with probability $1 - p$.

Think of a Bernoulli random variable as counting the number of heads observed in a single flip of a coin with a given bias. If X represents the number of heads in a single flip of a fair coin, then $X \sim Bernoulli(0.5)$, and if Y represents the number of heads in a single flip of a biased coin which lands on heads only 30% of the time, then $Y \sim Bernoulli(0.3)$. The pmf of Y is graphed below.

Proposition 4.4 *If $X \sim Bernoulli(p)$, then $E(X) = p$ and $\mathrm{Var}(X) = p(1-p)$.*

Pf. These results both follow from direct calculations. From the probability mass function above we have $E(X) = 0 \cdot (1-p) + 1 \cdot p = p$. We can then calculate

$$\mathrm{Var}(X) = E[(X - \mu_X)^2] = E[(X - p)^2] = (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p$$
$$= p^2(1-p) + (1-p)^2 p = p(1-p)(p + (1-p)) = p(1-p).$$

$\square$

Bernoulli random variables are simple but fundamental in probability and statistics. The Binomial, Geometric, and Poisson distributions, all of which come up frequently in scientific practice, can be defined in terms of Bernoulli random variables. In this course, we'll cover only the Binomial distribution, the others are studied in the Probability & Random Variables option course, 201-PRV.
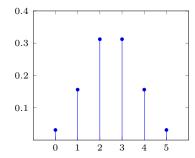
4.6 The Binomial Distribution
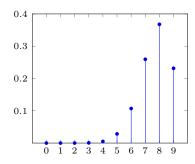
Suppose we perform a fixed number of Bernoulli trials (call this number n). Each trial has the same probability of success (call this value p), and the results of each trial form a sequence of independent Bernoulli random variables. If X counts the total number of successes after all the trials have been performed, then we say that X has a Binomial distribution.

Definition 4.9 *The random variable X has a **binomial distribution** with parameters n and p, written $X \sim Binomial(n, p)$, if*

$$f_X(x) = \begin{cases} \dbinom{n}{x}(p)^x(1-p)^{n-x} & \text{if } x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}.$$

Below on the left is the pmf of a binomial distribution with $n = 5$, $p = 0.5$, and on the right, the pmf of a binomial distribution with $n = 9$, $p = 0.85$.

To derive this probability mass function, imagine recording of the results of the Bernoulli trials as an n letter long string of S's and F's, representing success and failure respectively. By the Mississippi formula from Section 2.5, the number of such strings in which x successes occur is $\frac{n!}{x!(n-x)!} = \binom{n}{x}$.

Consider any such string, say $SFSSSFF$. In this case $n = 7$ and four successes occurred. The probability this sequence of outcomes occurs is not hard to calculate since we're performing independent Bernoulli trials, each with the same probability of success. Every S occurs with probability p, and every F with probability $1 - p$. Since there are four S's and three F's, and the trials are independent, this string occurs with probability $p^4(1 - p)^3$.

In total then, there are $\binom{n}{x}$ sequences of outcomes with x successes, and each sequence occurs with probability $p^x(1 - p)^{n-x}$, so $P(X = x) = \binom{n}{x}p^x(1 - p)^{n-x}$.

The number of heads in n flips of a coin is the example that should immediately come to mind when you think of a binomial distribution, but binomial distributions also appear in many practical contexts involving samples.

Example 4.18 *Suppose that in a certain country's election, 56% of the voting population will vote for a particular candidate, Bob. If you select seven members of the voting population uniformly at random, and ask them who they'll vote for (assume they'll answer honestly), what is the probability your poll will incorrectly conclude that Bob will not win a majority of the vote?*

Let X represent the number of people who respond by saying that they'll vote for Bob. If we regard such a response as a success, and each interview is done without knowledge of the results of any of the others, then $X \sim Binomial(7, 0.56)$.

The result we're seeking is the probability that there are fewer than four successes, which we calculate from the pmf as follows.

$$P(X \leq 3) = \sum_{i=0}^{3} P(X = i) = \sum_{i=0}^{3} \binom{7}{i}(0.56)^i(0.44)^{7-i}$$
$$\simeq 0.0032 + 0.0284 + 0.1086 + 0.2304$$
$$= 0.3706 \simeq 37\%$$

There is a 37% chance that the results of this poll will incorrectly predict that Bob will not win a majority of the vote.

Remark ☝

It would be easier to use the cdf to evaluate $P(X \leq 3)$, but unfortunately, the cdf for a binomial distribution does not have a simple closed form, so we're stuck summing the values of the pmf.

One of the reasons the binomial distribution appears in many contexts is that the process of evaluating a claim by doing repeated independent trials, which could be interviews, simulations, or literal laboratory experiments, is such a fundamental practice.

Theorem 4.1 *If $X \sim Binomial(n, p)$, then $E(X) = np$ and $\mathrm{Var}(X) = np(1 - p)$.*

These formulas follow from properties of expected value and sums of random variables which are not covered in this course, but appear in the Probability & Random Variables option course, 201-PRV.

Example 4.19 *Let X be the number of doubles in five rolls of a pair of dice, and let Y be the number of heads in three flips of a fair coin. Which quantity is larger on average? Which quantity is more variable?*

X is a binomial random variable with $n = 5$ and $p = \frac{6}{36} = \frac{1}{6}$, so $E(X) = np = \frac{5}{6}$ and $\mathrm{Var}(X) = np(1-p) = 5(\frac{1}{6})(\frac{5}{6}) = 25/36 \simeq 0.694$.

Y is a binomial random variable with $n = 3$ and $p = \frac{1}{2}$, so $E(Y) = np = \frac{3}{2}$ and $\mathrm{Var}(X) = np(1-p) = 3(\frac{1}{2})(\frac{1}{2}) = 3/4 \simeq 0.75$.

Therefore, Y is larger on average. The second question is somewhat ambiguous, but if we use the variance as our measure of variability, Y is a little more variable then X.

Example 4.20 *Edward lives in a country with an unreliable mail system, where 7% of parcels do not make it to their destination. He wants to send two books to his brother, each with a value of \$20. If he sends them in one parcel, the postage is \$5.20, and if he sends them separately, the postage is \$3.30 for each book. To minimize his expected losses, which method is preferable?*

Let $X \sim Bernoulli(0.07)$, then sending both books in one parcel, the \$5.20 postage charge always applies, and \$40 is lost if $X = 1$. If L represents Edward's loss, then $L = 5.20 + 40X$, and therefore $E(L) = 5.20 + 40E(X) = 5.20 + 40 \cdot 0.07 = 8$.

On the other hand, if the two books are sent separately, then $L = 6.60 + 20Y$, where $Y \sim Binomial(2, 0.07)$. Thus, $E(L) = 6.60 + 20E(Y) = 6.60 + 20 \cdot 2 \cdot 0.07 = 9.4$, so sending both books in one package is the better option.

4.7 Continuous Random Variables

In Section 4.1, we defined a random variable as a function from Ω to $\mathbb{R}$, but have so far only considered random variables that take values in some countable subset $A \subseteq \mathbb{R}$. We saw that we could specify the distribution of such a random variable using a probability mass function, or a cumulative distribution function.
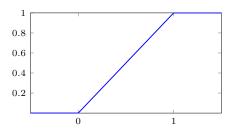
Our goal is now to understand what happens when a random variable can take any value in some interval $I \in \mathbb{R}$. Consider for example a random variable X which represents a number chosen uniformly at random from the interval $[0, 1]$. Unfortunately, it's not possible to express the distribution of X as a table, since $[0, 1]$ is not a countable set, so we can't list its values one-by-one.

Fortunately, the cumulative distribution function comes to the rescue. What is $P(X \leq \frac{2}{3})$? If we divide the interval into thirds, then $X \leq \frac{2}{3}$ when X falls into one of the left two thirds, which should happen two thirds of the time if X is equally likely to fall anywhere in $[0, 1]$.

Thus, $P(X \leq \frac{2}{3}) = \frac{2}{3}$. By a similar argument, it should be clear that $P(X \leq x) = x$ for any $x \in [0, 1]$ and therefore, the cumulative distribution function of X is the function given below.

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$
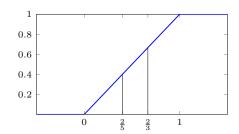

Using this cdf, we can determine the probability that X will be in various subsets of $[0, 1]$. Let's take the interval $(\frac{2}{5}, \frac{2}{3})$. Using the cdf above, we can compute

$$\begin{aligned} P(X \in (\tfrac{2}{5}, \tfrac{2}{3})) &= P(\tfrac{2}{5} < X < \tfrac{2}{3}) \\ &= P(X < \tfrac{2}{3}) - P(X \leq \tfrac{2}{5}) \\ &= F_X(\tfrac{2}{3}-) - F_X(\tfrac{2}{5}) \\ &= \tfrac{2}{3} - \tfrac{2}{5} = \tfrac{4}{15} \end{aligned}$$

Key Point 🔑

If the cdf of X is continuous, then $\lim_{x \to a^-} F_X(x) = \lim_{x \to a^+} F_X(x) = F_X(a)$, and hence $F_X(a-)$, $F_X(a+)$ and $F_X(a)$ are all equal. This implies a random variable X with a continuous cdf will take any fixed realization with probability zero, since
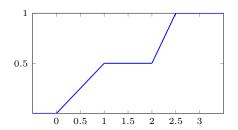
$$P(X = x) = P(X \leq x) - P(X < x) = F_X(x) - F_X(x-) = 0.$$

Furthermore, $P(X \in (a, b))$, $P(X \in [a, b))$, $P(X \in (a, b])$, and $P(X \in [a, b])$ are all equal to $F_X(b) - F_X(a)$. Including or excluding endpoints makes no difference.

Definition 4.10 *If X is a random variable whose cumulative distribution function $F_X(x) = P(X \leq x)$ is continuous, then we say X is a **continuous random variable**.*

Example 4.21 *Let X be a random variable with cdf given below. Find $P(X \in (\frac{1}{3}, 1) \cup [\frac{3}{2}, \frac{5}{2}])$.*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2}x & \text{if } 0 \leq x < 1 \\ \frac{1}{2} & \text{if } 1 \leq x < 2 \\ x - \frac{3}{2} & \text{if } 2 \leq x < \frac{5}{2} \\ 1 & \text{if } x \geq \frac{5}{2} \end{cases}$$
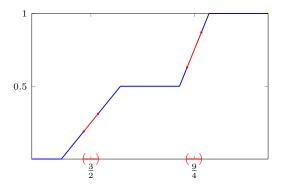
Since the events $X \in (\frac{1}{3}, 1)$ and $X \in [\frac{3}{2}, \frac{5}{2}]$ are mutually exclusive,

$$\begin{aligned} P(X \in (\tfrac{1}{3}, 1) \cup [\tfrac{3}{2}, \tfrac{5}{2}]) &= P(X \in (\tfrac{1}{3}, 1)) + P(X \in [\tfrac{3}{2}, \tfrac{5}{2}]) \\ &= F_X(1) - F_X(\tfrac{1}{3}) + F_X(\tfrac{5}{2}) - F_X(\tfrac{3}{2}) \\ &= \tfrac{1}{2} - \tfrac{1}{2} \cdot \tfrac{1}{3} + 1 - \tfrac{1}{2} \\ &= \tfrac{1}{2} - \tfrac{1}{6} + 1 - \tfrac{1}{2} = \tfrac{5}{6}. \end{aligned}$$

Probability Density Functions

Consider a continuous random variable X, and two real numbers r and s. We now know that $P(X = r)$ and $P(X = s)$ are both zero, but we can still ask whether X is more likely to end up close to r or close to s.

We can formalize this by asking for the probability X lies in each of the tiny intervals $(s - \epsilon, s + \epsilon)$ and $(r - \epsilon, r + \epsilon)$, for some very small positive number ϵ. Take for instance the random variable X in Example 4.21 above. Is it more likely this variable takes a value close to $\frac{1}{2}$, or a value close to $\frac{9}{4}$?
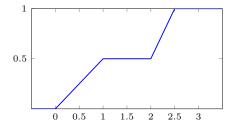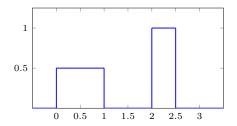
The probability X lies in the interval (a, b) is $F_X(b) - F_X(a)$, the difference in the value of the cdf at the endpoints. For the cdf shown above, the graph is twice as steep when it's near $\frac{9}{4}$ than it is when it's near $\frac{1}{2}$, and hence this difference is twice as large. Therefore, X is twice as likely to assume values near $\frac{9}{4}$ than it is to assume values near $\frac{1}{2}$.

In general then, the probability X lies close to a given value is proportional to the slope of the cdf at that value. But we know how to measure the slope of any function, with the derivative. Thus, given any $x \in \mathbb{R}$, if we evaluate the derivative of F_X at x, the larger the result, the more likely it is that X takes a value very close to x. Remember that the probability $X = x$ must be zero for all $x \in \mathbb{R}$, so when we evaluate the derivative of F_X at x, the result is not a probability, instead it's known as a probability density, and the derivative of F_X is called the probability density function of X.

Definition 4.11 *If X is a continuous random variable with cdf F_X, the **probability density function** of X is denoted f_X and is defined by $f_X(x) = \frac{d}{dx} F_X(x)$.*

The cdf F_X of the continuous random variable X defined in Example 4.21 above is shown once again on the left, and the corresponding probability density function f_X is on the right. It's standard practice to draw vertical lines on the graph of the probability density function as below. These lines are technically not part of the graph (they are jump discontinuities), but can make it easier to read.

We'll often use pdf to abbreviate probability density function. The rationale for the term 'probability density' is the following: when considering a rod of metal or other material, we can only ask for the mass of a contiguous piece of the material, since the mass at a point is always zero. We can, however, consider the density of the material to be well defined at a point, as the rate of change in mass at that point as we move along the rod. The situation with continuous random variables is analogous: a probability density measures the rate of change of the cdf as we move along it, in other words, the rate at which probability is accumulating as we move past that point.

Proposition 4.5 *If X is a continuous random variable, then the probability X lies in the interval (a, b) is the area under f_X on (a, b).*
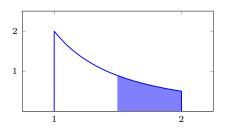
Pf. Calculating $P(X \in (a, b))$ using the cdf F_X as we have done in the examples above, and applying the fundamental theorem of calculus, we have

$$P(X \in (a, b)) = F_X(b) - F_X(a) = \int_a^b \frac{d}{dx} F_X(x)\, dx = \int_a^b f_X(x)\, dx.$$

This integral is the area under f_X on (a, b). Note that the result would be the same with any of the intervals $(a, b]$, $[a, b)$, or $[a, b]$.   □

Example 4.22 *Let X be the continuous random variable whose pdf is given below. Find $P(X > \frac{3}{2})$.*

$$f_X(x) = \begin{cases} \frac{2}{x^2} & \text{if } 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

$$P(X > \tfrac{3}{2}) = \int_{\frac{3}{2}}^{2} f_X(x)\, dx = \int_{\frac{3}{2}}^{2} \frac{2}{x^2}\, dx = -\frac{2}{x}\Big|_{\frac{3}{2}}^{2} = -\frac{2}{2} - \left(-\frac{2}{\frac{3}{2}}\right) = \frac{1}{3}$$

Key Point 🔑

The area above the interval I under the probability density function f_X is the probability X takes a value in I. This is why the probability density function is so useful. It gives a representation of the distribution of a continuous random variable which is as geometrically intuitive as a histogram, and allows us to apply the tools of calculus directly to compute probabilities.

Proposition 4.6 *If X is any continuous random variable with cdf F_X, the pdf f_X satisfies each of the properties below.*

1. *For all $x \in \mathbb{R}$, $f_X(x) \ge 0$.*

2. *$\int_{-\infty}^{\infty} f_X(x)\, dx = P(-\infty < X < \infty) = 1$.*

3. *$\int_{-\infty}^{t} f_X(x)\, dx = P(X \le t) = F_X(t)$.*

Informally, any non-negative function which encloses an area of one between its graph and the horizontal axis is the pdf of some continuous random variable whose cdf can be recovered through integration, as in the third property above.

Example 4.23 *Consider the function f given below. Find c so that this function is the pdf of some continuous random variable.*

$$f(x) = \begin{cases} cxe^{-x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Since $e^x > 0$ always, we have $xe^{-x} > 0$ when $x > 0$, and thus f is non-negative when $c > 0$. We must find a positive value for c such that the area under f is one, so we evaluate the area under f.

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{0}^{\infty} cxe^{-x}\, dx = c\int_{0}^{\infty} xe^{-x}\, dx = c(-x-1)e^{-x}\Big|_{0}^{\infty}$$
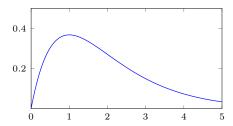
The last equality comes from using integration by parts. To complete the evaluation, we can use

L'Hôpital's rule.

$$c(-x-1)e^{-x}\Big|_0^\infty = \lim_{x\to\infty}\left(c(-x-1)e^{-x}\right) - c(-0-1)e^0$$

$$= c\lim_{x\to\infty}\left(\frac{-x-1}{e^x}\right) + c$$

$$\overset{H}{=} c\lim_{x\to\infty}\left(\frac{-1}{e^x}\right) + c$$

$$= 0 + c = c$$

The area under f is simply c, and therefore if we take $c = 1$, f is a pdf. The graph of f with $c = 1$ is given below.

As you can see, the material that was covered in Calculus I & II will be very relevant in the world of continuous random variables. If you're a bit rusty, don't get too worried. The integrals that appear in problems involving continuous random variables are often either very basic, or not possible to evaluate using the methods of elementary calculus (we'll see an example in the next section). The example above was created specifically to remind you about integration by parts and L'Hôpital's rule.

Expectation for Continuous Random Variables

The expected value of a continuous random variable X is defined in essentially the same way the expected value of a discrete random variable, but in place of a probability mass function we have a probability density function, and in place of a sum we have the continuous analogue, an integral.

Definition 4.12 *If X is a continuous random variable, the expected value of X is denoted $E(X)$ or μ_X, and defined by*

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Note that this improper integral may not converge, in which case the expected value is undefined.

Example 4.24 *Find the expected value of the random variable X whose pdf is given by $f_X(x) = 2/x^2$ on $[1, 2]$.*

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_1^2 x\,\frac{2}{x^2}\, dx = 2\int_1^2 \frac{1}{x}\, dx = 2\ln(x)\Big|_1^2 = 2\ln(2)$$

Example 4.25 *Consider the random variable X whose pdf is given below. Show $E(X)$ is undefined.*

$$f_X(x) = \frac{1}{\pi(1 + x^2)} \quad on \quad (-\infty, \infty)$$

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{\infty} x \frac{1}{\pi(1 + x^2)}\, dx = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x}{1 + x^2}\, dx$$

  To evaluate a 'doubly improper' integral like this, we must split it into two improper integrals to see if both of those converge. In the two resulting integrals, we make the substitution $u = 1 + x^2$, so $du = 2x\, dx$ and hence $dx = \frac{1}{2x} du$.

$$
\begin{aligned}
E(X) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x}{1 + x^2}\, dx \\
&= \frac{1}{\pi} \int_{-\infty}^{0} \frac{x}{1 + x^2}\, dx + \frac{1}{\pi} \int_{0}^{\infty} \frac{x}{1 + x^2}\, dx \\
&= \frac{1}{\pi} \int_{\infty}^{1} \frac{x}{u} \frac{1}{2x}\, du + \frac{1}{\pi} \int_{1}^{\infty} \frac{x}{u} \frac{1}{2x}\, du \\
&= \frac{1}{2\pi} \int_{\infty}^{1} \frac{1}{u}\, du + \frac{1}{2\pi} \int_{1}^{\infty} \frac{1}{u}\, du \\
&= \frac{1}{2\pi} \ln(u) \Big|_{\infty}^{1} + \frac{1}{2\pi} \ln(u) \Big|_{1}^{\infty} \\
&= \frac{1}{2\pi} \left(\ln(1) - \lim_{u \to \infty} ln(u) \right) + \frac{1}{2\pi} \left(\lim_{u \to \infty} ln(u) - \ln(1) \right)
\end{aligned}
$$

  Neither of the two improper integrals converge since $\lim_{u \to \infty} \ln(u) = \infty$, and hence $E(X)$ is undefined.

Remark ☝
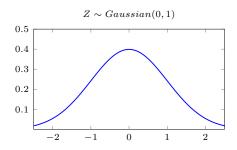
This result may seem odd, since you would intuitively place the center of this distribution at zero. It turns out that there exist distributions (this one is known as a Cauchy distribution) that have so much area in their tails that they have an undefined mean. These fat-tailed distributions are very interesting, as many of the usual estimation methods we'll see later on in the course don't work well, or even at all, when the characteristic of a population under study follows a fat-tailed distribution.
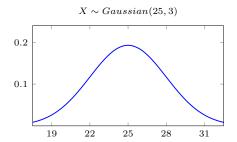
  The variance of a continuous random variable is also defined in essentially the same way as in the discrete case, as the average squared distance between a value of the random variable and its mean. As with the expected value, the computation amounts to evaluating an integral. The details appear in the Probability & Random Variables option course, 201-PRV. For our purposes, it will suffice to understand that in all contexts, variance measures the average squared distance between a quantity and its mean, and standard deviation is always the square root of the variance.

4.8   The Gaussian Distribution

Definition 4.13 *The random variable X has a **Gaussian distribution** with parameters μ (the mean) and $\sigma > 0$ (the standard deviation) if its pdf is given by the function below on the interval $(-\infty, \infty)$.*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The Gaussian distribution gets its name from mathematician Carl Friedrich Gauss, and is also known as the Normal distribution. It will play a central role in the statistical procedures we'll develop in the next chapter.

Unfortunately, the pdf f_X for a Gaussian distribution is not integrable in elementary terms. In other words, its antiderivative can't be expressed using only the functions and operations you've encountered in Calculus II, which means we can't find areas under the curve (probabilities) using those methods.

The traditional way to get around this problem is to use a large table of values as a reference. Unless otherwise stated, the letter Z will be reserved for the random variable $Z \sim Gaussian(0,1)$ from now on, and a Z-table is simply a large table of values of the cdf $F_Z(z)$, that is, a list of $P(Z \leq z)$ for a huge collection of realizations $z \in \mathbb{R}$.

What if X is a Gaussian random variable with a mean which is not zero and a standard deviation which is not one? As you might expect given the two graphs above, the distributions of any two Gaussian random variables differ by shifting and scaling, in other words, one is a linear function of the other. The tools required to prove this proposition are developed in the Probability & Random Variables option course, 201-PRV.

Proposition 4.7 *If $X \sim Gaussian(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma} \sim Gaussian(0,1)$.*

The process of translating values of any normally distributed random variable into corresponding values of $Z \sim Gaussian(0,1)$ is known as **standardization**, and $Z \sim Gaussian(0,1)$ is known as the **standard normal distribution**. To standardize a realization x of $X \sim Gaussian(\mu, \sigma)$ we simply compute $z = \frac{x-\mu}{\sigma}$, and we can treat the resulting z as a value of $Z \sim Gaussian(0,1)$.

Example 4.26 *The height of a randomly selected Canadian adult female, in inches, is approximated by the random variable $X \sim Gaussian(64, 2.5)$. What proportion of Canadian adult females are taller than 5'8"? What proportion have heights between 5'3" and 5'7"?*
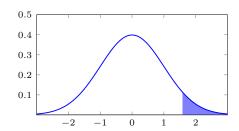
To standardize $5'8" = 68"$, we calculate $z = \frac{68-64}{2.5} = 1.6$. Then we have

$$
\begin{aligned}
P(X > 68) &= P(Z > 1.6) \\
&= 1 - P(Z \leq 1.6) \\
&= 1 - 0.9452 \\
&= 0.0548 \simeq 5.5\%.
\end{aligned}
$$
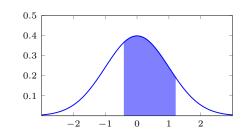
Thus, about 5.5% of Canadian adult females are taller than 5'8". For the second part we standardize $5'3" = 63"$ and $5'7" = 67"$ to obtain $z_1 = \frac{63-64}{2.5} = -0.4$ and $z_2 = \frac{67-64}{2.5} = 1.2$, then calculate

$$
\begin{aligned}
P(63 < X < 67) &= P(-0.4 < Z < 1.2) \\
&= P(Z \leq 1.2) - P(Z \leq -0.4) \\
&= 0.8849 - 0.3446 \\
&= 0.5404 \simeq 54\%.
\end{aligned}
$$
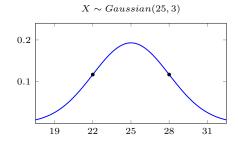
Therefore, around 54% of Canadian adult females are between 5'3" and 5'7".

Since there is now almost universal access to computers which can calculate areas under the pdf of $Z \sim Gaussian(0, 1)$ efficiently using numerical integration, the Z-table is something of a mathematical artifact whose uses are limited to classrooms and standardized tests.
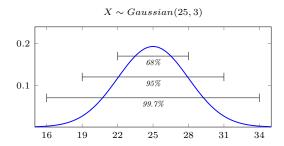
Gaussian Distribution Shortcuts

Any Gaussian distribution has two inflection points, and these points occur exactly one standard deviation from the mean. This fact is useful when drawing a quick sketch of a Gaussian distribution.

For approximate reasoning with Gaussian distributions without the aid of a Z-table or computer, the proposition below presents a well-known rule of thumb.

Proposition 4.8 *(68-95-99.7 Rule) For any Gaussian distribution, 68% of the area under its pdf occurs within one standard deviation of the mean, 95% occurs within two standard deviations, and 99.7% occurs within three standard deviations.*

Primarily a useful rule of thumb for quick approximate calculations, the rule also illustrates how quickly the distribution decays when it moves away from its mean. To contrast this with the strange behaviour we saw in Example 4.25 at the end of the last section, we can say that Gaussian distributions have thin tails.

Example 4.27 *If $X \sim Gaussian(64, 2.5)$ is used to model the height of a randomly selected Canadian female, provide a range of heights which will include 95% of all Canadian adult females.*

Using the 68-95-99.7 Rule, 95% of Canadian adult females will have heights within two standard deviations of the mean height, 64". So we calculate $64 - 2(2.5) = 59$ and $64 + 2(2.5) = 69$, then conclude that 95% of all Canadian adult females are taller than $59" = 4'11"$ and shorter than $69" = 5'9"$.

Index