Probability

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
np.random.seed(999)
```

> The term **probability** refers to the study of randomness and uncertainty. In any situation in which one of a number of possible outcomes may occur, the discipline of probability provides methods for quantifying the chances, or likelihoods, associated with the various outcomes — Jay L. Devore (2012).

Sample Spaces and Events

An **experiment** is any activity or process whose outcome is subject to uncertainty. It is a procedure that can be infinitely repeated and *has a well-defined set of possible outcomes*.

The Sample Space of an Experiment

> We called the *set of all possible outcomes* of an experiment the **sample space** of the experiment, denoted by S.

For examples:

- One of the simplest experiment is *tossing a coin*. The possible outcomes of this experiment is "*the coin comes up heads*" and "*the coin comes up tails*". We may write the sample space of tossing a coin as

  $$S = \{H, T\}$$

  where H represents "*heads*" and T represents "*tails*".

- Consider the experiment of *tossing a six-faced die*. If we are interested in the number that shows on the top face, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. If we interested only in whether the number on the top face is odd or even, then the sample space is $S = \{\text{odd}, \text{even}\}$.

- Consider an experiment consists of *examining a single fuse to see whether it is defective*. The sample space for this experiment can be abbreviated as $S = \{N, D\}$ where N

represents "*not defective*" and D represents "*defective*". If we instead *examine three fuses in sequence* and note the result of each examination, then an outcome for the entire experiment is any sequence of N's and D's of length 3, so

$$S = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\}.$$

- An experiment consists of flipping a coin and then flipping it a second time if a head occurs on the first flip. If a tail occurs on the first flip, then a die is tossed once. The sample space of this experiment if we interested in the side of the coin that comes up and the top face of the die is

$$S = \{HH, HT, T1, T2, T3, T4, T5, T6\}.$$

We called each outcome in a sample space a **sample point**. For example, when tossing a coin, the sample space is $S = \{H, T\}$ so the outcomes H and T are samples points.

We will commonly interested in size of sample space—*the number of sample points in a sample space*, abbreviated $n(S)$. From the examples shown above, we can simply just count the sample point.

- $S = \{H, T\} \longrightarrow n(S) = 2.$
- $S = \{1, 2, 3, 4, 5, 6\} \longrightarrow n(S) = 6.$
- S being a sample space of examining three fuses in sequence $\longrightarrow n(S) = 8.$

Events

> An **event** is *any collection of outcomes contained in the sample space (any subset of sample space)*. An event is **simple** if it consists of exactly one outcome and **compound** if it consists of more than one outcome.
>
> $$\text{Event} \subseteq S.$$

When an experiment is performed, a particular event E is said to occur if the resulting outcome is contained in E. In general, exactly one simple event will occur, but many compound events will occur simultaneously.

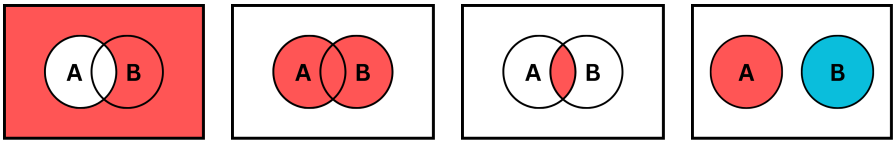
For example, consider the experiment of tossing three coins in sequence. So the sample space of this experiment is $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$. Thus there are eight simple events, each consists of only a single sample point of the sample space, among which are $E_1 = \{HHH\}$ and $E_2 = \{HTT\}$. Some compound events include:

- $A = \{HHT, HTH, THH\} =$ the event where exactly one coin comes up tails.
- $B = \{HHH, TTT\} =$ the event that all three coins come up the same side.
- $C = \{HHT, HTH, THH, HTT, THT, TTH, TTT\} =$ the event that at least one coin comes up heads.

Some Relations from Set Theory

An event is just a set, so any operations that can be done on sets can also be done on events to create another event:

1. The **complement** of an event E, denoted by E' or $\bar{E}$ or $\neg E$, is the event that E *does not occur*.
2. The **union** of two events A and B, denoted by $A \cup B$ and read "*A or B*", is the event where *either A or B or both occur*.
3. The **intersection** of two events A and B, denoted by $A \cap B$ and read "*A and B*", is the event where *both A and B occur*.
4. If $A \cap B = \emptyset$ where $\emptyset$ is the *null event* (the event consisting of no outcomes), then A and B are said to be **mutually exclusive** or **disjoint** events. That is they can not occur simultaneously.



Complement (A')	Union	Intersection	Mutually Exclusive

Probability Interpretations

We will revisit the concept of **relative frequency** mentioned in Chapter 0.1.

$$\text{relative frequency of a value } x = \frac{\text{number of times } x \text{ occurs}}{\text{total number of observation in the data set}}$$

Now we define the relative frequency in n repeated random experiments of an event A as

$$f_A = \frac{n_A}{n}$$

where n_A is the number of occurrences of A.

For n repeated random experiments, if n tends to infinity (or get larger and larger), then f_A will converge to some constant and we will call that constant the **probability** of A, denoted by $P(A)$:

$$P(A) = \lim_{n \to \infty} f_A = \lim_{n \to \infty} \frac{n_A}{n}.$$

For example, consider a (six-faced) die tossing experiment and let A be the event that the die comes up six. The result from tossing the die 10 times are as follows:

In [2]:
```
tossing_result = np.random.randint(1,7,size=10)
tossing_result_1000 = np.concat((tossing_result, np.random.randint(1,7,size=990)))
```

In [3]:
```
d = {
    'Toss #': np.arange(1,11),
    'Toss sesult': tossing_result,
    'Relative frequency of A': (tossing_result == 6).cumsum() / np.arange(1,11)
}
toss_10 = pd.DataFrame(data=d).set_index('Toss #')
toss_10
```

Out[3]:

Toss #	Toss sesult	Relative frequency of A
1	1	0.000000
2	5	0.000000
3	6	0.333333
4	2	0.250000
5	1	0.200000
6	2	0.166667
7	4	0.142857
8	2	0.125000
9	4	0.111111
10	1	0.100000

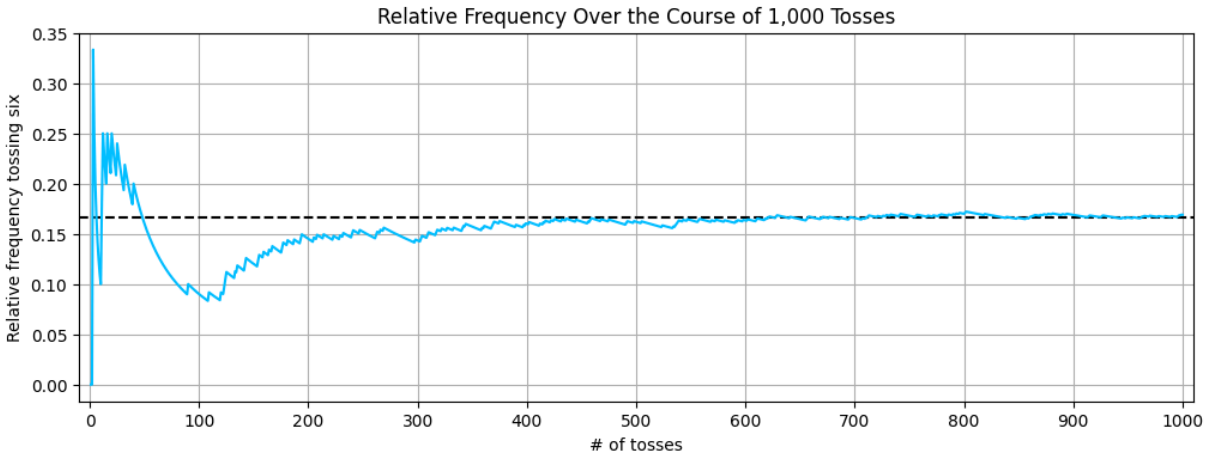
We can see that the relative frequency fluctuates over the course of few first tosses. But as the number of tosses increase, the figure shown below show that the relative frequency

stabilizes and converges to a constant of about $1/6$ which is indeed the probability of tossing a six.

In [4]:
```
fig, ax = plt.subplots(figsize=(12,4))

ax.plot([-10,1010], [1/6, 1/6], linestyle='--', color='black')
ax.plot(np.arange(1,1001), (tossing_result_1000 == 6).cumsum() / np.arange(1,1001),

ax.set_axisbelow(True); ax.grid(True)
ax.set_xticks(range(0, 1100, 100))
ax.set_xlim(-10,1010)
ax.set_title("Relative Frequency Over the Course of 1,000 Tosses")
ax.set_xlabel("# of tosses")
ax.set_ylabel("Relative frequency tossing six");
```



Probability Axioms

For any event A, *the objective of probability is to assign to each event A a number $P(A)$, called the probability of the event A, which will give a precise measure of the chance that A will occur.* To ensure that the probability assignments will be consistent with our intuitive notions of probability, all assignments should satisfy the following axioms of probability:

First Axiom: Non-negativity

The probability of an any event E is a non-negative real number:

$$P(E) \geq 0.$$

Second Axiom: Normalization

The assumption of unit measure: The probability of the sample space is 1. That is, the probability that at least one of the simple events (singleton subsets) of the sample space will occur is 1:

$$P(S) = 1.$$

Third Axiom: Additivity

If $A_1, A_2, A_3, \ldots$ is an infinite collection of disjoint (mutually exclusive) events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Notice that these axioms were inspired by and reflect the behavior of relative frequencies (surely, since we defined probability from relative frequency). One property of probability we can derived from these axioms is that for any event E, $P(E) \leq 1$ since the second axiom states that $P(S) = 1$ so for any particular event E that is a subset of the sample space S, the probability of that event has to be less than that of the sample space.

Rules and Properties of Probability

In this section, we will derived some of the rules of probability from the three axioms of probability.

Probability of Null Event & Additive Rule

Let $\emptyset$ denoted the **null event** which is a empty set containing no outcomes. Since null events are disjoint ($\because \emptyset \cap \emptyset = \emptyset$), so the third axiom gives

$$P\left(\bigcup_{i=1}^{\infty} \emptyset\right) = P(\emptyset) = \sum_{i=1}^{\infty} P(\emptyset).$$

This can happen only if $P(\emptyset) = 0$.

The fact that the probability of null event is zero further implies that the third axiom is also valid for a finite collection of disjoint events: Consider a collecting of disjoint events $A_1, A_2, \ldots, A_k, A_{k+1}, A_{k+2}, \ldots$ where $A_{k+1}, A_{k+2}, \ldots$ are all null events, so

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{k} A_i\right) + P\left(\bigcup_{i=k+1}^{\infty} A_i\right)$$

$$= P\left(\bigcup_{i=1}^{k} A_i\right) + P(\emptyset) = P\left(\bigcup_{i=1}^{k} A_i\right)$$

$$= \sum_{i=1}^{k} P(A_i).$$

Therefore, for $A_1, A_2, \ldots, A_k$ being a collection of **disjoint (mutually exclusive)** events, then

$$P(A_1 \cup A_2 \cup \cdots \cup A_k) = P(A_1) + P(A_2) \cdots + P(A_k)$$

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i).$$

This property is called the **additive rule** for finite collection of disjoint events.

For a collection of events that are not mutually exclusive, we can instead use the *principle of inclusion-exclusion* from set theory to determined the probability of unions, for instances:

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

for any events A and B.

Complement Rule

Consider two events A and A'. By definition

1. $A \cap A' = \emptyset$, so they are mutually exclusive
2. $A \cup A' = S$

So we can apply the addition rule for disjoint event and get

$$P(A \cup A') = P(A) + P(A') = P(S).$$

And now we can apply the second axiom to get the **complement rule**: For any event A,

$$P(A) + P(A') = 1 \quad \leftrightarrow \quad P(A) = 1 - P(A').$$

Example Consider a system of five electrical components connected in series such that if one (or more) component fails, the entire system disconnects and ultimately fails. Let A represent the event that the system fails, i.e., "at least one component fails." Denote a

component that fails by F and one that doesn't fail by S (for success). The event A can be explicitly written as:

$$A = \{FSSSS, SFSSS, SSFSS, \ldots, FFFFF\}.$$

There are, in fact, 31 outcomes in A, so calculating the probability of A directly would be tedious. However, the complement event A', where the system does not fail, contains only one outcome: $\{SSSSS\}$. We can use the complement rule to simplify the calculation.

Suppose that each component has a 10% chance of failure (and a 90% chance of success), and failures are independent across components. Then:

$$\begin{aligned} P(A) &= 1 - P(A') = 1 - P(\{SSSSS\}) \\ &= 1 - (0.9)^5 \\ &= 1 - 0.59049 \\ &\approx 0.41 \text{ or } 41\%. \end{aligned}$$

We will explore later why $P(\{SSSSS\}) = 0.9^5$ when we discuss the *multiplication rule*.

Equally Likely Outcomes

Consider an experiment with $n(S) = N$ different outcomes with all outcomes have **equal chance to occur**. These include many simple experiment such as tossing a fair coin or a fair die, or drawing a card from a standard deck. Let $S = \{e_1, e_2, \ldots, e_N\}$ be the sample space of experiment with equally likely outcomes and E_i be simple event where $E_i = \{e_i\}$ for $i = 1, 2, \ldots, N$, from the axioms, we can calculate:

$$P(S) = 1 = P(E_1) + P(E_2) + \cdots + P(E_N).$$

since all outcomes are equally likely to happen, that is $P(E_1) = P(E_2) = \cdots = P(E_N)$, so

$$P(E_1) + P(E_2) + \cdots + P(E_N) = N \cdot P(E_i) = 1$$

$$P(E_i) = \frac{1}{N} = \frac{1}{n(S)}.$$

That is the probability that each outcome will occur is $1/n(S)$ where $n(S)$ is the size of the sample space.

Now consider an event E of equally likely experiment, with $n(E)$ denoting the number of sample points contained in E. Then

$$P(E) = \sum_{E_i \in E} P(E_i) = \sum_{E_i \in E} \frac{1}{n(S)} = \frac{n(E)}{n(S)}.$$

Example Suppose you are rolling a 12-sided fair die and want to determine the probability of rolling an even number greater than 6. Let A represent the event of rolling an even number above 6. In this case:

$$A = \{8, 10, 12\}.$$

Since the die is fair, all outcomes are equally likely. The total number of outcomes in the sample space is $n(S) = 12$ and the number of favorable outcomes is $n(A) = 3$. The probability of A is:

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{12}.$$

Thus, there is a 25% chance of rolling an even number above 6.

Example Suppose that somebody secretly rolls two fair six-sided dice, and we wish to compute the probability that the sum of the face-up values $D1 + D2$ is equal to some integer s. The sample space of 36 combinations of rolled values of the two dice is shown in the table below:

Two Dice Sums		D_1					
		1	2	3	4	5	6
D_2	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

Let E_s be the event that the sum of two dice is equal to s. From the table we can count $n(E_s)$ for each integer s in the range 2 to 12 and calculate the probability given that each of the 36 outcomes is equally likely to occur using the formula $P(E_s) = n(E_s)/36$:

s	2	3	4	5	6	7	8	9	10	11	12
$n(E_s)$	1	2	3	4	5	6	5	4	3	2	1
$P(E_s)$	2.78%	5.56%	8.33%	11.11%	13.89%	16.67%	13.89%	11.11%	8.33%	5.56%	2.78%

Independence & Multiplication Rule

Independence: The events A and B are called **independent** if and only if *knowing that one occurs doesn't change the probability that the other occurs*.

For example, consider tossing a fair coin multiple times. Knowing that the first toss results in heads does not affect the probability that the second toss will be heads or tails. Thus, "first toss comes up heads" is independent of "second toss comes up heads" and "second toss comes up tails." However, "second toss comes up heads" and "second toss comes up tails" are not independent of each other, since knowing that the second toss already comes up heads gives information that it's impossible for the second toss to come up tails, and vice versa.

Multiplication rule for independent events: If A and B are independent, then the probability that both A and B occur is equal to the product of their probabilities:

$$P(A \text{ and } B) \;=\; P(A \cap B) \;=\; P(A)P(B).$$

Example Consider the same problem of a system of five electrical components connected in series as in the previous example from the complement rule section. If we want to calculate the probability that the system does not fail, $P(A') = P(\{SSSSS\})$, given that $P(S) = 0.9$. Since each component fails or succeeds independently of the others, we can use the multiplication rule for independent events to calculate:

$$\begin{aligned} P(A') &= P(S_1 \cap S_2 \cap S_3 \cap S_4 \cap S_5) \\ &= P(S_1) \cdot P(S_2) \cdot P(S_3) \cdot P(S_4) \cdot P(S_5) \\ &= (0.9)^5 = 0.59049. \end{aligned}$$

Conditional Probability

Sometimes the probabilities of various events **depend on what is known about the experiment at the time we are interested in**. Knowing that something has already happened might affect the probability of other events occurring. As we just discussed in the previous section, we call a set of events *independent* if there is no connection between them, and we can use the multiplication rule for independent events to calculate the probabilities related to them. But when the events are *not independent*, we need to use the concept of **conditional probability**.

Conditional probability is a measure of the probability of an event occurring, **given that another event is already known to have occurred**. We write "the conditional probability of A *given* B" as

$$P(A \mid B).$$

For example, suppose we do an analysis on spam emails and found that spam emails have 8% chance to contain the word "*money*" which is higher that the chance of the same word occurring in ham emails (emails that are not spams) of 1%. We may write:

$$P(\text{``money''} \mid \text{spam}) = 8\% \quad \text{and} \quad P(\text{``money''} \mid \text{ham}) = 1\%.$$

The conditional probability $P(A \mid B)$ can be understood as the fraction of the probability of B that intersects with A. You can think of this as *the sample space narrowing down to* B, and we are only interested in the part of A that lies within the new sample space B (i.e., $A \cap B$):

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \; P(B) > 0.$$

General Multiplication Rule

The formula of conditional probability often get rearranged as:

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

This formula is called the **general multiplication rule**, which extents the multiplication rule for two events that are not independent. Notice that in the special case where A and B are independent, knowing that B happened does not affect the probability of A, so $P(A \mid B) = P(A)$ which implies $P(A \cap B) = P(B)P(A)$ which is the multiplication rule mentioned earlier.

Moreover, we can define independent events using conditional probability:

> Events A and B are independent if and only if $P(A \mid B) = P(A)$, so $P(A \cap B) = P(A)P(B)$.

To extend the multiplication rule to more than two events, consider a set of events $A_1, A_2, A_3, \ldots, A_k$. We may write the intersection of every events in this set as

$$\begin{aligned} A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_k &= (A_1 \cap A_2) \cap A_3 \cap \cdots \cap A_k \\ &= \big((A_1 \cap A_2) \cap A_3\big) \cap \cdots \cap A_k \\ &\;\;\vdots \\ &= \Big(\big((A_1 \cap A_2) \cap A_3\big) \cap \cdots\Big) \cap A_k\Big). \end{aligned}$$

Hence,

$$\begin{aligned} P(A_1 \cap A_2 \cap \cdots \cap A_k) &= P(A_1)P(A_2 \mid A_1) \cdot P(A_3 \cap A_4 \cap \cdots \cap A_k) \\ &= P(A_1)P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot P(A_4 \cap \cdots \cap A_k) \\ &\;\;\vdots \\ &= P(A_1)P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdots P(A_k \mid A_1 \cap A_2 \cap \cdots \cap A_{k-1}) \end{aligned}$$

Example Suppose that somebody rolls two fair six-sided dice (again), this time wish to compute the probability that the face-up value of the first one D_1 is 2, given the information that their sum $D_1 + D_2$ is no greater than 5.

Let A be the event "the first die is 2 ($D_1 = 2$)" and B be the event "the sum of the dice is less than or equal to 5 ($D_1 + D_2 \leq 5$)". We can highlight the sample points contained in these two events in the sample space table as follows:

A

Two Dice Sums		D_1					
		1	2	3	4	5	6
D_2	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

B

Two Dice Sums		D_1					
		1	2	3	4	5	6
D_2	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

Think of this situation as the sample space narrowing from all 36 possible outcomes to just 10 outcomes contained in B. We then only interest in the outcomes in A that is within this new sample space, that is the event $A \cap B$:

Two Dice Sums		D_1					
		1	2	3	4	5	6
D_2	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

From the information, we calculate

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{3/36}{10/36} = \frac{3}{10} = 30\%.$$

Example A lot contains 12 items of which 4 are defective. Three items are drawn randomly from the lot one after the other. What is the probability that all three picked items are non-defective?

First, let

- A be the event "the first item is non-defective".
- B be the event "the second item is non-defective".
- C be the event "the third item is non-defective".

Therefore, the probability that all three items are non-defective is

$$P(A \cap B \cap C) = P(A)P(B \mid A)P(C \mid A \cap B) = \frac{8}{12}\frac{7}{11}\frac{6}{10} = \frac{14}{55} \approx 25.45\%.$$

Explanation:

- The probability $P(A)$ is obviously the proportion of non-defective items to all items $8/12$.
- $P(B|A)$ is the probability that non-defective item is picked after one non-defective has already been picked. So there are 11 items left with 4 defectives and 7 non-defectives.
- $P(C|A \cap B)$ now two non-defective has been picked. So there are 10 items left with 6 non-defectives.

The Law of Total Probability

Recall that a set of events $A_1, A_2, \ldots, A_k$ are **mutually exclusive** or **disjoint** if they are not intersected, that is no pair of two contain any common outcomes.

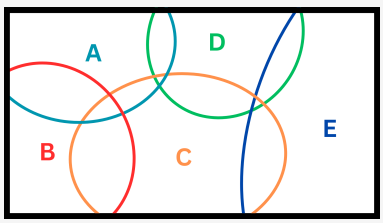
If $A_1, A_2, \ldots, A_k$ are events of sample space S and

$$A_1 \cup A_2 \cup \cdots \cup A_k = \bigcup_{i=2}^{k} A_i = S$$

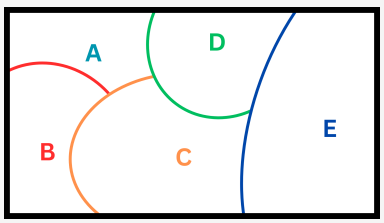
that is they covered the whole sample space, then $A_1, A_2, \ldots, A_k$ are called **collectively exhaustive (or just "exhaustive") events**.

And if a set of events are *both mutually exclusive and exhaustive*, then they are called a **partition** of the sample space S

Collectively Exhaustive

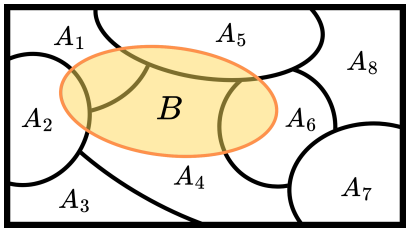


Both Collectively Exhaustive and Mutually Exclusive (Partition of Sample Space)



Now consider any event B of the sample space S. If $A_1, A_2, \ldots A_k$ is a partition of S, then we can split B into many parts intersected with events in the partition:

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_k).$$



Using the multiplication rule and addition rule, we can derived the *law of total probability* :

> **The law of total probability:** Let $A_1. A_2, \ldots, A_k$ be mutually exclusive and collectively exhaustive (a partition of S). Then for any event B,
>
> $$\begin{aligned} P(B) &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots P(B \mid A_k)P(A_k) \\ &= \sum_{i=1}^{k} P(B \mid A_i)P(A_i). \end{aligned}$$

Example Consider 3 different email accounts containing 100 emails in total. Account #1 contain 70 emails, account #2 contain 20 emails, and the remaining 10 emails is in account #3. Of the emails into account #1, only 1% are spam, whereas the corresponding percentages for accounts #2 and #3 are 2% and 5%, respectively. What is the probability that a randomly selected emails from any account is a spam?

To answer this question. Let first define the notation:

- E_i be the event "the email comes from account #i" for $i = 1, 2, 3$.
- S be the event "the email is spam".

The given information imply that:

- $P(A_1) = 0.7, P(A_2) = 0.2, P(A_3) = 0.1$.
- $P(S \mid A_1) = 0.01, P(S \mid A_2) = 0.02, P(S \mid A_3) = 0.05$.

Now we can simply substitute the values into the equation for the law of total probability:

$$\begin{aligned} P(S) &= P(S \mid A_1)P(S) + P(S \mid A_2)P(S) + P(S \mid A_3)P(S) \\ &= (0.01)(0.7) + (0.02)(0.2) + (0.05)(0.1) \\ &= 0.016 = 1.6\% \end{aligned}$$

So, in the long run, 1.6% of the emails will be spam.

Bayes' Theorem

Bayes' theorem (alternatively **Bayes' law** or **Bayes' rule**) gives a rule for *inverting conditional probabilities*, allowing one to find the probability of a cause given its effect $P(B \mid A)$ from the probability of a effect given its cause $P(A \mid B)$.

The Statement of Bayes' Theorem

Suppose that we know the value of $P(B \mid A)$ but we want to calculate $P(A \mid B)$, then what can we do? Let's start by rearranging the conditional probability formula:

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

Since the order of "and" (intersect) doesn't matter, we can write the general multiplication rule as:

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

We can then plug in this into the conditional probability formula:

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

And that is what we wanted: we derived the formula that express $P(A \mid B)$ in terms of $P(B \mid A)$. This formula is called **Bayes' theorem**.

In real-applications (most of the time), the value of $P(B)$ is not directly given. However, we can the law of total probability to calculate it using A and A' as a partition of sample space (since A and A' are disjoint and exhaustive). Therefore, Bayes' theorem is often written in a expanded form as

> $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A').} \quad \text{where} \quad P(B) \neq 0$$

Generalization of Bayes' Theorem

In the last section we expanded the formula for Bayes' theorem using the law of total probability by expanding the term $P(A)$ using A and A' as a partition of sample space. More generally, we can consider A as a part of any partition of the sample space and expand it accordingly. This leads to a generalized expression of Bayes' theorem:

> Let \set$A_1, A_2, \ldots, A_k$ be a set of mutually exclusive and exhaustive events (i.e., a partition of the sample space). Then for any event B where $P(B) \neq 0$. Then
>
> $$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{P(B)} = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{k} P(B \mid A_i)P(A_i)}$$
>
> for any $1 \leq j \leq k$.

Bayesian Analysis & Famous Examples

As we saw in the last section. Bayesian probability interprets probabilities based on a *degree of belief* in an event since this interpretation is defined by conditional probability, which

determines the probability of an event given the probability (belief) of another event.

In Bayesian analysis, we begin with some **prior** knowledge (or prior belief) about the event. Then, as we examine new data, we update the prior probability using Bayes' theorem to arrive at the **posterior** knowledge (or posterior belief).

From the formula for Bayes' theorem

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{P(B)} = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{k} P(B \mid A_i)P(A_i)}.$$

$P(A_j)$ is the **prior probability** before the event B occurs and $P(A_j \mid B)$ is the **posterior probability**, the probability under the condition that B occurs.

A great example would be a spam filter. We might have a dataset of emails in which 20% of the data are spam emails, so the prior probability that an email is a spam is 20%. After that, we examime the emails for certain keywords, calculate the probabilities that the keywords appear in spam or non-spam emails, and use these probabilities to update the prior probability to get a better estimate of whether a new email is spam or not.

Example: False Positive

Suppose that 1 in 1000 of the population has a certain rare disease for which an diagnostic test has been developed. If an infected person is tested, a positive result will occur 99% of the time. Whereas a person with no disease will get an erroneous positive result ('*false positive*') 2% of the time.

The technical term **"false positive"** refers to an error in which a test result incorrectly indicates the *presence* of a condition (e.g. a disease when the disease is not present), while a **"false negative"** is the opposite error, where the test result incorrectly indicates the *absence* of a condition when it is actually present.

To determine whether the test is effective, we ask: " If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?"

We will solve this problem using Bayes' theorem. First let

- D be the event "the person has the disease".
- $D^{'}$ be the event "the person does not have the disease".
- t be the event "positive test result".
- $t^{'}$ be the event "non-positive (negative) test result".

From the information state in the description of the problem, we can imply:

- $P(D) = 1/1000,\ P(D^{'}) = 999/1000.$
- $P(t \mid D) = 0.99,\ P(t^{'} \mid D) = 0.01.$

- $P(t \mid D') = 0.02,\ P(t' \mid D') = 0.98.$

We can calculate the probability that a person has the disease given that he/she tested positive $P(D \mid t)$ using the formula

$$P(D \mid t) = \frac{P(t \mid D)P(D)}{P(t)} = \frac{P(t \mid D)P(D)}{P(t \mid D)P(D) + P(t \mid D')P(D')}.$$

Substitute in the values, we get

$$P(D \mid t) = \frac{(0.99)(1/1000)}{(0.99)(1/1000) + (0.02)(999/1000)} \approx 0.047 = 4.7\%.$$

This result seems counterintuitive; we can see that the accuracy of the test is relatively low, despite a very small false positive rate. This is because *the proportion of the population with the disease is very small compared to the proportion without the disease*. As a result, even a small error rate, when applied to the much larger group of non-infected individuals, can lead to a significant number of false positives.

Example: The Monty Hall Problem

The Monty Hall problem is a classic probability puzzle that can be solved using Bayes' theorem. The result is often counterintuitive for most people. It is based, nominally, on the American television game show *Let's Make a Deal* and is named after its original host, Monty Hall.

Suppose you are on a game show, and you are given the choice of three doors. Behind one door is a car, and behind the other two doors are goats. You pick a door, say Door No. 1. The host, who knows what is behind each door, opens another door, say Door No. 3, which has a goat. He then asks, "Do you want to switch your choice to Door No. 2?"

Now the question is: **Should you switch your choice?**

To begin with, let's run a simulation of the game and look at the wining rate both for when you choose to switch the choices and not switch the choices:

In [5]:
```
def monty_hall_doors():
    doors = np.array(["goat", "goat", "car"])
    np.random.shuffle(doors)
    return doors

win_change, win_not_change = np.array([]), np.array([])

# run 1,000 games
for _ in range(1000):
    car = np.random.randint(3)
    first_pick = np.random.randint(3)
    opened_door = [x for x in range(3) if x != car and x != first_pick][0]
    change_pick = [x for x in range(3) if x != opened_door and x != first_pick][0]
```

```
    if first_pick == car:
        win_not_change = np.append(win_not_change, 1)
        win_change = np.append(win_change, 0)
    elif change_pick == car:
        win_not_change = np.append(win_not_change, 0)
        win_change = np.append(win_change, 1)
```

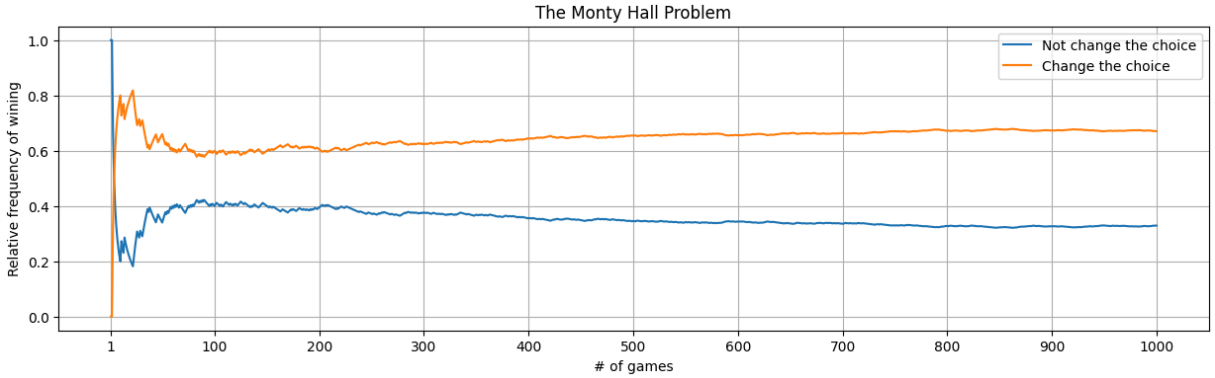
In [6]:
```
fig, ax = plt.subplots(figsize=(15,4))

x = np.arange(1,1001)

ax.plot(
    x, win_not_change.cumsum() / x,
    label="Not change the choice"
)

ax.plot(
    x, win_change.cumsum() / x,
    label="Change the choice"
)

ax.set_axisbelow(True); ax.grid(True)
ax.legend()
ticks = np.arange(0,1001,100)
ticks[0] = 1
ax.set_xticks(ticks)
ax.set_title("The Monty Hall Problem")
ax.set_xlabel("# of games")
ax.set_ylabel("Relative frequency of wining");
```



Most people might initially think that it doesn't matter whether you switch your choice or not, assuming there's a 50-50 chance of each remaining door hiding the car. However, simulation results consistently show that you have a higher chance of winning the car if you choose to switch your choice.

Now let's solve the problem analytically using Bayes' theorem. Let's first define some notation:

- Let C_i be the event "the car is behind door No.i" for $i = 1, 2, 3$.
- Let O_i be the event "the host opened door No.i, that is there is a goat behind door No.i ".

The prior probability $P(C_i) = 1/3$ for all $i = 1, 2, 3$ since initially, we only know that one of the three doors hides the car. Suppose you choose door No.1, and the host opens door No.2. The posterior probability is given by:

$$P(C_1 \mid O_2) = \frac{P(O_2 \mid C_1)P(C_1)}{P(O_2 \mid C_1)P(C_1) + P(O_2 \mid C_2)P(C_2) + P(O_2 \mid C_3)P(C_3)}$$

Given that the car is behind door No.1, the host can either open door No.2 or door No.3. If this is the case, we assume the host will randomly choose one of the two doors with a 50% chance for each. Thus, $P(O_2 \mid C_1) = P(O_3 \mid C_1) = 1/2$.

The host will never open the door hiding the car, so $P(O_2 \mid C_2) = 0$. If the car is behind door No.3, the host can only open door No.2 (since you initially chose door No.1, and the host cannot open it). Therefore, $P(O_2 \mid C_3) = 1$.

Combining all of this, we calculate:

$$P(C_1 \mid O_2) = \frac{(1/2)(1/3)}{(1/2)(1/3) + (0)(1/3) + (1)(1/3)} = \frac{1}{3}.$$

Similarly, for the case where the car is behind door No.3:

$$P(C_3 \mid O_2) = \frac{(1)(1/3)}{(1/2)(1/3) + (0)(1/3) + (1)(1/3)} = \frac{2}{3}.$$

This result implies that there is a 2/3 chance of winning the car if you switch to door No.3, which is greater than the 1/3 chance of winning if you stick with door No.1. A similar calculation applies in other cases (e.g., picking a different door initially). Thus, we conclude that switching your choice increases your chances of winning the game.

Example: Warner's Randomized Response Model

Suppose we want to know, "*What percentage of students have cheated during an exam in college?*" This question can be difficult to answer because students may not respond truthfully in a direct survey due to the sensitive nature of the question.

A research method called **randomized response**, proposed by S. L. Warner, allows respondents to address sensitive issues while maintaining confidentiality through randomization. In this method, **unknown to the interviewer**, chance determines whether the respondent answers truthfully or simply replies "yes," regardless of the truth.

For example, in a survey, students may first be instructed to toss a coin twice and then respond to one of two questions based on the outcome of the first toss:

- If the first toss is tails, answer Question 1: Have you ever cheated on an exam in college?
- If the first toss is heads, answer Question 2: Did you get tails on the second toss?

Since the interviewer does not know whether a "yes" response comes from Question 1 (cheating) or Question 2 (a coin toss result), this method allows respondents to answer truthfully while maintaining their anonymity.

Although an individual "yes" response cannot be interpreted directly, the proportion of cheaters can be estimated from the overall responses. We can split the event of a student answering "yes" into two parts:

$$\begin{aligned} P(``\text{yes}") &= P(``\text{yes}" \cap \text{Q1}) + P(``\text{yes}" \cap \text{Q2}) \\ &= P(\text{Q1})P(``\text{yes}" \mid \text{Q1}) + P(\text{Q2})P(``\text{yes}" \mid \text{Q2}). \end{aligned}$$

The proportion of cheaters, $P(``\text{yes}" \mid \text{Q1})$, can be expressed as:

$$P(``\text{yes}" \mid \text{Q1}) = \frac{P(``\text{yes}") - P(\text{Q2})P(``\text{yes}" \mid \text{Q2})}{P(\text{Q1})}.$$

All the terms can be determine:

- $P(\text{Q1})$ and $P(\text{Q2})$: The chance that the student is answer to question 1 or 2, which is the chance that the student gets "tail" or "head" on the first toss. That is 50%.
- $P(``\text{yes}" \mid \text{Q2})$: The chance that the student answer "yes" to question 2, which is equal to the chance that the student gets "tail" on the second toss. Which is also 50%.
- $P(``\text{yes}")$: The proportion of students answered "yes", which can be determined from the survey results.

Suppose that 27 students answered "yes" and 30 students answered "no". So we estimate

$$P(``\text{yes}" \mid \text{Q1}) = \frac{(27/57) - (0.5)(0.5)}{0.5} = 0.4473 = 44.73\%.$$

And that is the estimation of fraction of cheaters.