Chapter 3

Probability

We now turn our attention towards making inference in sample from the frequency of an occurrence in the population. Informally, we will define the probability of an event or outcome as the long run frequency of this event or outcome. This frequency may be over time or repeated trials. For example, the number six will appear one million times if we roll a fair die six million times. The long run frequency or probability of a six appearing will be one out of six or $\frac{1}{6}$. Formally, we will take the possible events or outcomes as data collect them in the form of a *probability space.*

3.1 Probability Space

Definition A probability space is a triple $(X, \mathcal{B}, P)$ consisting of

1. A set X called the *sample space*. We can think of this set as all possible states of some system or all possible outcomes of some experiment.

2. A collection $\mathcal{B} \subseteq 2^X$ (where 2^X denotes all subsets of X). We can think of $\mathcal{B}$ as a collection of observable events or outcomes.

3. A function $P : \mathcal{B} \to [0, 1]$. For each event $E \in \mathcal{B}$, $P(E)$ gives the probability of E occurring by mapping the event to the interval $[0, 1]$.

The set X is arbitrary, but the set $\mathcal{B}$ may not consist of all subsets of X. We take X as our universe, so it is possible that X is not a finite set. However, $\mathcal{B}$ will be finite by our construction, because we want to map a finite set of outcomes to a closed interval. Typically, X will be a finite set in most applications we consider, so $\mathcal{B}$ will consists of all possible subsets of X (i.e. $\mathcal{B} = 2^X$). Right now, P is also an arbitrary function, so we need to impose some structure on P to calculate probabilities.

Definition The function P and set $\mathcal{B}$ must satisfy the following properties

1. $P(X) = 1$ or $P(\emptyset) = 0$

2. $X \in \mathcal{B}$ and $\mathcal{B}$ is closed under the set operations union, intersection and complementation (i.e. if $E, F \in \mathcal{B}$, then $E \cup F \in \mathcal{B}$ and $E \cap F \in \mathcal{B}$ and $E^c = X \backslash E \in \mathcal{B}$). This property implies $P(E^c) = 1 - P(E)$.

3. If $E, F \in \mathcal{B}$ are disjoint (i.e. $E \cap F = \emptyset$), then $P(E \cup F) = P(E) + P(F)$. This property is known as additive countability of P. We can use mathematical induction to show that if $E_1, E_2, ..., E_N$ are all pairwise disjoint (i.e. $E_i \cap E_j = \emptyset$ if $i \neq j$), then $P(\bigcup_{i=1}^N E_i) = \sum\limits_{i=1}^{N} P(E_i)$. This property also holds if N is ∞.

Note that pairwise disjoint means that the two events or outcomes are *mutually exclusive.* We call events or outcomes *collectively exhaustive* when they form a partition of X.

Definition A collection of sets $E_1, E_2, ...$ is called a *partition* of the space X if

1. $A_i \cap A_j = \emptyset \; \forall i \neq j$
2. $X = E_1 \cup E_2 \cup ...$

Example Suppose we roll a pair of fair dice, what is $P(\text{sum of dice is even})$?

Listing all the possible outcomes. We see that an even number appears half the time and any one of these events is possible and equally likely, so $P(\text{sum of dice is even}) = \frac{1}{2}$.

Die 1	Die 2	Sum
1	1	2
1	2	3
1	3	4
.	.	.
.	.	.
.	.	.
6	6	12

In this example, we see

$$X = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

and

$$\mathcal{B} = \{\emptyset, X, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}, \{10\}, \{11\}, \{12\}, \{1, 2\}, \{2, 3\}...\}.$$

Technically, $\mathcal{B}$ includes X, the empty set ($\emptyset$), and all subsets of X, but we will often omit irrelevant ones when we write $\mathcal{B}$ for practical purposes. Hence, we only write out the outcomes of interest. For example, $P(2 \cap 3) = 0$ and $\{2, 3\}$ is a subset of X, but it is not a relevant outcome. So we use a shorthand and only list relevant outcomes as

$$\mathcal{B} = \{\{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}, \{10\}, \{11\}, \{12\}\}$$

and take as given that P is defined as zero over every other subset in $\mathcal{B}$ and $P(X) = 1$. In this sense, we will write $\mathcal{B}$ the same way as X for practical purposes with the implicit understanding that $\mathcal{B} \subseteq 2^X$ and $X \subset \{\emptyset, X\} \subseteq \mathcal{B}$.

3.2 Constructive Theory

Given our definitions, we can now turn to their implications. We can actually construct quite a bit of probability theory from just these definitions.

Proposition 3.2.1 *For any E and F, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.*

Proof *We can verify that*

$$E \cup F = E \cup (F \cap E^c) \Rightarrow P(E \cup F) = P(E) + P(F \cap E^c)$$

by additive countability. Similarly,

$$F = (F \cap E) \cup (F \cap E^c) \Rightarrow P(F) = P(F \cap E) + P(F \cap E^c).$$

Therefore,

$$P(E \cup F) = P(E) + P(F \cap E^c) = P(E) + P(F) - P(E \cap F) \qquad \blacksquare$$

Proposition 3.2.2 *If $E \subset F$, then $P(F) = P(E) + P(F \cap E^c) \geq P(E)$.*

Proof

$$\begin{aligned} F &= \underbrace{(E\cap F)}_{=E}\cup(F\cap E^c)\\ \Leftrightarrow F &= E\cup(F\cap E^c)\\ \Rightarrow P(F) &= P(E)+P(F\cap E^c) \textit{ by additive countability}\\ \Rightarrow P(F) &= P(E)+P(F\cap E^c)\geq P(E) \textit{ because } P(F\cap E^c)\in[0,1]. \quad \blacksquare \end{aligned}$$

Proposition 3.2.3 *If $E\cap F=\emptyset$ then $P(E)\leq P(F^c)$.*

Proof

$$E\cap F=\emptyset\Leftrightarrow E\subset F^c.$$

From the previous proposition, $P(E)\leq P(F^c)$ follows immediately. $\blacksquare$

Proposition 3.2.4 $P(E)=P(F)=P(E\cap F)\Rightarrow P[(E\cap F^c)\cup(E^c\cap F)]=0$

Proof

$$E=(E\cap F)\cup(E\cap F^c)\Rightarrow P(E)=P(E\cap F)+P(E\cap F^c)$$

by additive countability. Since $P(E)=P(E\cap F)$, $P(E\cap F^c)=0$. By the same logic, $P(E^c\cap F)=0$. Now

$$P[(E\cap F^c)\cup(E^c\cap F)=P(E\cap F^c)+P(E^c\cap F)=0$$

because $(E\cap F^c)\cap(E^c\cap F)=\emptyset$.

Proposition 3.2.5 $P(E)=P(F)=P(E\cap F)=1\Rightarrow P(E\cap F)=1$

Proof *Since $P(E)=P(F)=1$, we know $P(E^c)=P(F^c)=0$.*

$$\begin{aligned} P(E\cap F) &= P((E^c\cup F^c)^c) \textit{ by DeMorgan's Law}\\ &= 1-P(E^c\cup F^c)\\ &= 1-[P(E^c)+P(F^c)+P(E^c\cap F^c)]\\ \Leftrightarrow P(E\cap F) &= 1+P(E^c\cap F^c) \end{aligned}$$

$E^c\cap F^c\subset E^c\Rightarrow P(E^c\cap F^c)\leq P(E^c)=0\Rightarrow P(E^c\cap F^c)=0$. Hence, $P(E\cap F)=1$. $\blacksquare$

And finally...

Proposition 3.2.6 $P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(B\cap C)-P(A\cap C)+P(A\cap B\cap C)$.

Proof

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

We can verify that $(A\cup B)\cap C=(A\cap C)\cup(B\cap C)$, hence

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

We can also verify that $(A\cap C)\cup(B\cap C)=A\cap B\cap C$. Hence, after some substitution

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

3.3 Conditional Probability

So far, our probability structure measures the chances of an event occurring or outcome. However, we have information which helps us predict an outcome or event in many situations. We want to *condition* on this information to obtain an updated probability. Formally, we define a new probability space.

Definition Let $(X, \mathcal{B}, P)$ be a probability space and let $Y \in \mathcal{B}$ such that $P(Y) > 0$. We define the probability space $(Y, \mathcal{B}_Y, P_Y)$ as

1. $Y \subset X$ is the set of states such that $P(Y)$
2. The collection of observables events is defined as $\mathcal{B}_Y = \{E \cap Y : E \in \mathcal{B}\}$
3. The probability measure $P_Y : \mathcal{B}_Y \to [0,1]$ where

$$P(A|Y) = P_Y(A) = \frac{P(A)}{P(Y)} \quad \forall\ A \in \mathcal{B}_Y$$

   Y is the event that we know occurred and A is the event yet to occur.

Example We roll a fair die. What is $P(\text{even }\#|\# > 3)$?
$Y = \{4, 5, 6\}$ and two numbers are even out of the equally likely outcomes, hence $P(\text{even }\#|\# > 3) = \frac{2}{3}$.

Definition For $Y \subset X$ with $P(Y) > 0$ and any $E \subset X$, we define the *conditional probability* of E given Y as

$$P(E|Y) = \frac{P(E \cap Y)}{P(Y)} \equiv P_Y(E) \Rightarrow P(E \cap Y) = P(Y) \cdot P(E|Y).$$

Example Suppose we select three committee members from two men and three women. We select each person with equal probability. Given that the committee consists of at least one man and one woman (Y), what is $P(\#men > \#women|Y)$?

$$P(\#men > \#women|Y) = \frac{P[(\#men > \#women \cap Y]}{P(Y)}$$

$$P(Y) = P(\text{at least 1M and 1W}) = \frac{{}_3C_1}{{}_5C_3} = \frac{3}{10}$$

$$P(\#men > \#women \text{ and at least 1M and 1W}) = \frac{1}{{}_5C_3} = \frac{1}{10}$$

$$\Rightarrow P(\#men > \#women|Y) = \frac{1/10}{3/10} = \frac{1}{3}$$

In some situations the probability of an event does not depend on another event occurring. We call these events *independent.*

Definition The events E and F in the probability space $(X, \mathcal{B}, P)$ are *independent* if

$$P(E \cap F) = P(E) \cdot P(F)$$

Remark From the definition of conditional probability, we see that E and F are independent if and only if $P(E|F) = P(E)$. In other words, knowing F occurred does not affect the probability that E occurs.

Example We flip a fair coin twice. What is $P(\text{H on } 2^{nd} \text{ flip}|\text{H on } 1^{st} \text{ flip})$?
X is the set of all outcomes, so $X = \{HH, HT, TH, TT\}$.
Y is the set of outcomes where H occurred on the 1^{st} flip, so $Y = \{HH, HT\}$.
A is the set of outcomes where H occurs on the 2^{nd} flip.
$\Rightarrow A \cap Y = \{HH\}$. Now

$$P(A|Y) = \frac{P(A \cap Y)}{P(Y)} = \frac{1/4}{1/2} = \frac{1}{2} = P(A).$$

Hence, A and Y are independent.

Proposition 3.3.1 *Consider three events $X, Y, Z \in \mathcal{B}$ such that $P(X \cap Y) \neq 0$. Then $P(X \cap Y \cap Z) = P(X) \cdot P(Y|X) \cdot P(Z|X \cap Y)$. This proof is left as an exercise.*

Proposition 3.3.2 *Suppose E and F are independent. Then E and F^c are independent. This proof is left as an exercise.*

Remark Pairwise independence does not necessarily imply full independence. For example,

$$\begin{aligned}
P(A) = P(B) = P(C) &= \frac{1}{2} \\
P(A \cap B) = P(A) \cdot P(B) &= \frac{1}{4} \\
P(B \cap C) = P(B) \cdot P(C) &= \frac{1}{4} \\
P(A \cap C) = P(A) \cdot P(C) &= \frac{1}{4} \\
P(A \cap B \cap C) &= \frac{1}{4}
\end{aligned}$$

But $P(A) \cdot P(B) \cdot P(C) = \frac{1}{8}$.

3.3.1 Bayes' Rule

The most useful applications of conditional probability often use a different characterization of the definition. We call this characterization Bayes' Rule or Bayes' Theorem. This characterization has a nice interpretation. We consider two events, I and C. We think of I as the set of *information* we possess and C as the *cause* under consideration. We have some initial belief about some cause and update our beliefs about the underlying cause based on the information we observe.

Theorem 3.3.3 *Let $(X, \mathcal{B}, P)$ be a probability space and $C, I \in \mathcal{B}$ with $P(I) > 0$. Then*

$$P(C|I) = \frac{P(C)P(I|C)}{P(I)}.$$

The proof is left as an exercise.

We call $P(C)$ the *prior probability*, because it indicates our prior beliefs of the cause. We call $P(C|I)$ the *posterior probability*, because it is the probability of the cause after updating based on information I. We often refer to $P(I|C)$ as the *likelihood*, because it indicates the chances of observing the information we observe from the cause under consideration. $P(I)$ is the unconditional (or marginal) probability of observing the information we see.

Example Let C be the event that a patient has a disease. We know $P(C) = 0.01$ from the incidence of this disease in the general population being 1 in 100. I is the event that a patient tests positive for this disease. The test is 99% accurate, meaning $P(I|C^C) < 0.01$ and $P(I^C|C) < 0.01$. In other words, the test reports a negative result when the patient has the disease less than 1% of the time. The test also reports positive when the patient does not have the disease less than 1% of the time. Thus, we know

$$\begin{aligned}
&P(I|C) \approx 1 \\
&P(I) = P(I|C)P(C) + P(I|C^c)P(C^c) \approx 0.01 + (0.01)(0.99) \approx 0.02
\end{aligned}$$

Applying Bayes' Rule,

$$P(C|I) = \frac{P(C)P(I|C)}{P(I)} \approx \frac{0.01 \cdot 1}{0.02} = 0.5$$

This probability means that someone who tests positive for the disease only has a 50% of having the disease despite that the test is 99% accurate!

In this application, we calculated $P(I)$ by partitioning the space X into C and C^c. We can make finer partitions of the sample space in a way governed by what we sometimes call the "Rule of Total Probability."

Proposition 3.3.4 *Suppose we form a partition of the space $C_1 \cup C_2 \cup C_3 \cup ... = X$. Recall that a partition consists of pairwise disjoint sets. Then*

$$P(I) = \sum_i P(I|C_i) \cdot P(C_i)$$

as long as the set of events $\{C_i\}$ is countably infinite or finite.

3.4 Bernoulli Trials

Flipping a coin or rolling a die serve an examples of a repeated trial experiment. We can represent these d outcomes for a trial with a simple probability space. This experiment is known as a Bernoulli trial. Typically, we only deal with Bernoulli trials with two outcomes $\{0, 1\}$, but we define it generally.

Definition Let $(D, \mathcal{B}, P)$ be a probability space where

1. $D = \{0, 1, ..., d-1\}$ is the sample space or set of states (e.g. $\{Heads\} = \{1\}$)
2. $\mathcal{B}$ is a collection of subset of D
3. $P(i) = p_i$ for $i = 0, 1, ..., d-1$

Let $D^{(n)}$ denote the cartesian product of D with itself n times. Hence, $D^{(n)}$ consists of all ordered n-tuples $(x_1, x_2, ..., x_n)$ where $x_i \in D$ and $i = 1, 2, ..., n$. Now we construct the probability space $(D^{(n)}, \mathcal{B}^{(n)}, P^{(n)})$.

Proposition 3.4.1 *Let $(D^{(n)}, \mathcal{B}^{(n)}, P^{(n)})$ be the probability space defined above. Then for each $k = 0, 1, ..., n$*

$$P^{(n)}\{x_1, ..., x_n \in D^{(n)} : x_i = 0 \textit{ for } k \textit{ choices of } i = 1, ..., n\} = {}_nC_k \cdot p^k \cdot (1-p)^{n-k}.$$

In other words, the set of possible realizations of these n trials has a binomial distribution with the probability mass function described above.