# STAT 830

# Probability Theory

In this section I want to define the basic objects. I am going to give full precise definitions and make lists of various properties – even prove some things rigorously – but then I am going to give examples. In different versions of this course I require more or less understanding of the objects being studied.

**Definition**: A **Probability Space** (or **Sample Space**) is an ordered triple $(\Omega, \mathcal{F}, P)$ with the following properties:

- $\Omega$ is a set (it is the set of all possible outcomes of some experiment); elements of $\Omega$ are denoted by the letter $\omega$. They are called elementary outcomes.
- $\mathcal{F}$ is a family of subsets (we call these subsets **events**) of $\Omega$ with the property that $\mathcal{F}$ is a $\sigma$-field (or Borel field or $\sigma$-algebra) – that is $\mathcal{F}$ has the following **closure** properties:
  1. The empty set denoted $\emptyset$ and $\Omega$ are members of $\mathcal{F}$.
  2. $A \in \mathcal{F}$ implies $A^c = \{\omega \in \Omega : \omega \notin A\} \in \mathcal{F}$.
  3. $A_1, A_2, \cdots$ in $\mathcal{F}$ implies $A = \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.
- $P$ is a function whose domain is $\mathcal{F}$ and whose range is a subset of $[0, 1]$. The function $P$ must satisfy:
  1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
  2. **Countable additivity**: $A_1, A_2, \cdots$ **pairwise disjoint** ($j \neq k$ $A_j \cap A_k = \emptyset$)
$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

These axioms guarantee that we can compute probabilities by the usual rules, including approximation. Here are some consequences of the axioms:

$$A_i \in \mathcal{F}; i = 1, 2, \cdots \text{ implies } \cap_i A_i \in \mathcal{F}$$

$$A_1 \subseteq A_2 \subseteq \cdots \text{ implies } P(\cup A_i) = \lim_{n\to\infty} P(A_n)$$
$$A_1 \supset A_2 \supset \cdots \text{ implies } P(\cap A_i) = \lim_{n\to\infty} P(A_n)$$

The last two of these three assertions are sometimes described by saying that $P$ is *continuous*. I don't like this jargon because it does not agree very well with the standard meaning of a continuous function. There is (in what I have presented so far) no well defined *topology* or *metric* or other way to make precise the notion of a sequence of sets converging to a limit.

### 0.0.1 Examples

It seems wise to list a few examples of these triples which arise in various more or less sophisticated probability problems.

**Example 1: Three Cards Problem**

I imagine I have three cards – stiff pieces of paper. One card is green on both sides. One is red on both sides. The third card is green on one side and red on the other. I shuffle up the three cards in some container and pick one out, sliding it out of its container and onto the table in such a way that you can see only the colour on the side of the card which is up on the table. Later, when I talk about conditional probability, I will be interested in probabilities connected with the side which is face down on the table but here I just want to list the elements of $\Omega$ and describe $\mathcal{F}$ and $P$.

I want you to imagine that the sides of the card are labelled (in your mind, not visibly on the cards) in such a way that you can see that there are six sides of the card which could end up being the one which is showing. One card, the `RR` card has red on both sides and $\omega_1 = RR1$ means the first of these two sides is showing which $\omega_2 = RR2$ denotes the outcome that the second of these two sides is showing. I use $\omega_3 = RG1$ to denote the outcome where the Red / Green card is selected and the red side is up and $\omega_4 = RG2$ to denote the outcome where the same card is drawn but the green side is up. The remaining two elementary outcomes are $\omega_5 = GG1$ and $\omega_6 = GG2$ in what I hope is quite obvious notation.

So now $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$ is the sample space with six elements. There are many other possible notations for the elements of this sample space of course. I now turn to describing $\mathcal{F}$ and $P$.

In problems where $\Omega$ is finite or countably infinite we almost always take $\mathcal{F}$ to be the family of all possible subsets of $\Omega$. So in this case $\mathcal{F}$ is the collection of all subsets of $\Omega$. To make a subset of $\Omega$ we must decide for each of the six elements of $\Omega$ whether or not to put that element in the set. This makes 2 possible choices for $\omega_1$, then for each of these 2 choices for $\omega_2$ and so on. So there are $2^6 = 64$ subsets of $\Omega$; all 64 are in $\mathcal{F}$. In order to be definite I will try to list the pattern:

$$\mathcal{F} = \{\emptyset, \{\omega_1\}, \ldots, \{\omega_6\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \ldots, \{\omega_5, \omega_6\}, \ldots, \Omega\}$$

My list includes 1 set with 0 elements, 6 sets with 1 element, 6 choose 2 sets with 2 elements (total of 15), 6 choose 3 with 3 elements (20 such), 6 choose 4 (=15) with 4 elements, 6 with 5 elements and $\Omega$.

Finally I am supposed to describe $P$. The usual way, when $\Omega$ is finite, to assign probabilities is to give some probability, say $p_i$ to the $i$th elementary outcome $\omega_i$. In our case it is reasonable to assume that all 6 sides of the cards have the same chance of ending up visible so all

$$p_i = P(\{\omega_i\}) = \frac{1}{6}.$$

Then the probability of any subset of $\Omega$ is found by adding up the probabilities of the elementary outcomes in that set. So, for instance

$$P(\{\omega_1, \omega_3, \omega_4\}) = \frac{3}{6} = \frac{1}{2}.$$

The event "the side showing is red" is a subset of $\Omega$, namely,

$$\{\omega_1, \omega_2, \omega_3\}.$$

The event "the side face down is red" is also subset of $\Omega$, namely,

$$\{\omega_1, \omega_2, \omega_4\}.$$

The event "the side face down is green" is

$$\{\omega_3, \omega_5, \omega_6\}.$$

**Example 2: Coin Tossing till First Head Problem**

Now imagine tossing a coin until you get "heads" which I denote H. To simplify the problem I will assume that you quit tossing either when you get H OR when you have tossed the coin three times without getting H. Letting T denote tails the elements of $\Omega$ are, in obvious notation:

$$\{\omega_1, \omega_2, \omega_3, \omega_4\} \equiv \{H, TH, TTH, TTT\}$$

Again $\mathcal{F}$ is the collection of all $2^4 = 16$ subsets of $\Omega$ and we specify $P$ by assigning probabilities to elementary outcomes. The most natural probabilities to assign are $p_1 = 1/2$, $p_2 = 1/4$ and $p_3 = p_4 = 1/8$. I will return to this assumption when I discuss independence.

**Example 3: Coin Tossing till First Head Problem, infinite case**

Now imagine tossing the coin until you get "heads" no matter how many tosses are required. Let $\omega_k$ be a string of $k$ tails T followed by H. Then

$$\Omega = \{\omega_0, \omega_1, \omega_2, \cdots\}$$

which has infinitely many elements. Again $\mathcal{F}$ is the collection of all subsets of $\Omega$; the number of such subsets is uncountably infinite so I won't make a list! We specify $P$ by assigning probabilities to elementary outcomes. In order to add a bit to the example I will consider a biased coin. The most natural probabilities to assign are then

$$p_i = P(\{\omega_i\}) = p(1-p)^i.$$

This list of numbers adds up to 1, as it must, to ensure $P(\Omega) = 1$; you should recognize the sum of a geometric series.

**Example 4: Coin Tossing forever**

In order to discuss such things as the law of large numbers and many other probability problems it is useful to imagine the conceptual experiment of tossing the coin forever. In this case a single "elementary outcome", $\omega$ is actually an infinite sequence of Hs and Ts. One $\omega$ might be

$$HTHTHTHTHTHT\cdots$$

where the heads and tails alternate for ever. It would be typical to say

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots);\ \text{such that each } \omega_i \in \{H, T\}\}.$$

You can think about how many elements there are in $\Omega$ by taking a typical $\omega$ and replacing each H with a 1, then each T with a 0. Then put "0." in front and think of the result as a binary number between 0 and 1. So for instance the sequence above of alternating 0s and 1s is

$$\omega = 0.1010101010\cdots = \frac{1}{2}\left(1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \cdots\right)$$

which is just 2/3 by summing a geometric series.

The summary is that there are as many elements in $\Omega$ as there are numbers between 0 and 1 – an uncountably infinite number. It turns out that this is the situation where we just can't cope, logically, with having $\mathcal{F}$ be the collection of *all* subsets of $\Omega$. If you want to know which subsets go into $\mathcal{F}$ you need to find out about *Borel* sets.

In fact we take $\mathcal{F}$ to be "the smallest $\sigma$-field" which contains all sets of the form

$$B_i \equiv \{\omega \in \Omega : \omega_i = H\}$$

which is the subset of $\Omega$ obtained by keeping only outcomes whose $i$th toss is H. There is a bit of mathematical effort to prove the existence of any such "smallest" $\sigma$-field; it is the intersection of all $\sigma$-fields which contain the given special sets. Much greater effort is needed to understand the structure of this $\sigma$-field but I want to emphasize that if you can give a truly clear and explicit description of a subset of $\Omega$ that subset will be a Borel set – a member of $\mathcal{F}$.

Finally we have to say something about how to compute probabilities. Let's start with an intuitive presentation using the idea that we might be talking about independent tosses of a fair coin; I will define independence precisely later but for now I just want you to use what you already know about independent events. Let

$$C = B_1 \cap B_2^c \cap B_3 \cap B_4^c \cap B_5 \cap B_6^c \cdots .$$

The only point in $C$ is the sequence of alternating heads and tails I wrote down up above. So what is the probability of $C$. Certainly

$$P(C) \geq P(B_1 \cap B_2^c \cap B_3 \cap B_4^c \cap B_5 \cap B_6^c \cdots B_{2n}^c)$$

for any $n$. For independent tosses of a fair coin we compute the probability of this intersection by just multiplying 1/2 by itself $2n$ times to get $2^{-n}$. But

if $P(C) \leq 2^{-n}$ for all $n$ then $P(C) = 0$. In the same way we can check that $P(\{\omega\}) = 0$ for every elementary outcome $\omega$!

This just means we *cannot* compute probabilities of an event by adding up probabilities of elementary outcomes in the event – that always gives 0. Instead we use the idea of independence and the *assumption* that the various $B_i$ are independent and have probability $1/2$ to compute any probability we want; sometimes this is *hard*.

## 0.1 Random Variables

:

**Definition**: A **Vector valued random variable** is a function function $X : \Omega \mapsto R^p$ such that, writing $X = (X_1, \ldots, X_p)$,

$$P(X_1 \leq x_1, \ldots, X_p \leq x_p)$$

is defined for any constants $(x_1, \ldots, x_p)$. Formally the notation

$$X_1 \leq x_1, \ldots, X_p \leq x_p$$

describes a subset of $\Omega$ or **event**:

$$\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_p(\omega) \leq x_p\} \ .$$

Remember $X$ is a function on $\Omega$ so $X_1$ is also a function on $\Omega$; that is why we can stick in the argument $\omega$ of the function.

ASIDE: In almost all of probability and statistics the dependence of a random variable on a point in the probability space is hidden! You almost always see $X$ not $X(\omega)$.

There is a subtle mathematical point being made here. Not every function from $\Omega$ to $R^p$ is a random variable or random vector. The problem is that the set

$$\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_p(\omega) \leq x_p\}$$

might not be in $\mathcal{F}$! For our fourth example this is a potential mathematical (but not practical) problem.

### 0.1.1 Borel sets

In this subsection I give a small presentation of the notion of Borel sets in $R^p$. The material is not really part of this course.

**Definition**: The **Borel** $\sigma$-field in $R^p$ is the smallest $\sigma$-field in $R^p$ containing every open ball.

**Definition**: For clarity the open ball of radius $r > 0$ centred at $x \in R^p$ is

$$\{y \in R^p : ||y - x|| < r\}$$

where

$$||u|| = \sqrt{\sum_1^p u_i^2}$$

for a vector $u \in R^p$. The quantity $||u||$ is called the Euclidean norm of $u$; it is also the usual notion of length of a vector.

Every common set is a Borel set, that is, in the Borel $\sigma$-field.

**Definition**: An $R^p$ valued **random variable** is a map $X : \Omega \mapsto R^p$ such that when $A$ is Borel then $\{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$. This is equivalent to

$$\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_p(\omega) \leq x_p\} \in \mathcal{F}$$

for all $(x_1, \ldots, x_p) \in R^p$.
**Jargon and notation**: we write $P(X \in A)$ for $P(\{\omega \in \Omega : X(\omega) \in A\})$ and define the **distribution** of $X$ to be the map

$$A \mapsto P(X \in A)$$

which is a probability on the set $R^p$ with the Borel $\sigma$-field rather than the original $\Omega$ and $\mathcal{F}$. We also write

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

and call this set the *inverse image* of $A$ under $X$. So the distribution of $X$ is

$$P_X(A) = P(X^{-1}(A))$$

which is defined for all Borel sets $A \in R^p$.
**Remark**: The definition of a random variable depends only on the functions and the $\sigma$-fields involved and NOT on the probability $P$.

**Definition**: The **Cumulative Distribution Function** (cdf) of $X$ is the function $F_X$ on $R^p$ defined by

$$F_X(x_1, \ldots, x_p) = P(X_1 \leq x_1, \ldots, X_p \leq x_p)\,.$$

I will not always use the subscript $X$ to indicate which random vector is being discussed. When there is no real possibility of confusion I will just write $F$.

Here are some properties of $F$ for $p = 1$:

1. $0 \leq F(x) \leq 1$.
2. $x > y \Rightarrow F(x) \geq F(y)$ (monotone non-decreasing).
3. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.
4. $\lim_{x \searrow y} F(x) = F(y)$ (right continuous).
5. $\lim_{x \nearrow y} F(x) \equiv F(y-)$ exists.
6. $F(x) - F(x-) = P(X = x)$.
7. $F_X(t) = F_Y(t)$ for all $t$ implies that $X$ and $Y$ have the same distribution, that is, $P(X \in A) = P(Y \in A)$ for any (Borel) set $A$.

**Proof**: The values of $F$ are probabilities so they are between 0 and 1. If $F$ is the cdf of $X$ and $y < x$ then

$$\{X \leq y\} \subseteq \{X \leq x\}$$

so

$$F(y) = P(X \leq y) \leq P(X \leq x) = F(x).$$

Since $F$ is monotone the assertions about limits may be checked by considering a sequence $x_n$. For instance, to prove the first half of the third assertion we take $x_n$ to be any sequence decreasing to $-\infty$ – such as $x_n = -n$, say. If

$$A_n = \{X \leq x_n\}$$

then

$$A_1 \supseteq A_2 \supseteq \cdots$$

and

$$\cap_{n=1}^{\infty} A_n = \emptyset$$

so by the "continuity" of $P$

$$0 = P(\emptyset) = \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} F(x_n).$$

The argument at $\infty$ uses unions in place of intersections and a sequence $x_n$ increasing to $\infty$.

Assertion 4 considers a sequence $x_n$ decreasing to $y$ and then with the $A_i$ as above we find

$$\cap_{n=1}^{\infty} A_n = \{X \leq y\}$$

so that right continuity of $F$ comes from the continuity of $P$. Assertion 5 does the parallel thing with unions and shows $F(y-) = P(X < y)$.

Assertion 6 comes from the fact that

$$\{X < x\} \cup \{X = x\} = \{X \leq x\}.$$

The union is disjoint so

$$F(y-) + P(X = x) = F(y).$$

The final point, property 7, is much more sophisticated – much harder to prove. If you want to read about it you can look at the appendix on Monotone Class arguments if I ever get it done. •

For $p = 1$ any function $F$ with properties 1, 2, 3 and 4 is the cumulative distribution function of some random variable $X$. For $p > 1$ the situation is a bit more complicated. Consider the case $p = 2$ and two points $(u_1, u_2)$ and $(v_1, v_2)$. If $v_1 \geq u_1$ and $v_2 \geq u_2$ then the event $X_1 \leq u_1, X_2 \leq u_2$ is a subset of the event $X_1 \leq v_1, X_2 \leq v_2$. This means that

$$F(u_1, u_2) = P(X_1 \leq u_1, X_2 \leq u_2) \leq P(X_1 \leq v_1, X_2 \leq v_2) = F(v_1, v_2).$$

In this sense $F$ is monotone non-decreasing. But even if $F$ is continuous, monotone non-decreasing and satisfies properties 1 and 3 above we cannot be sure it is a cdf. Think about the rectangle

$$R \equiv \{(x_1, x_2) : u_1 < x_1 \leq v_1, u_2 < x_2 \leq v_2\}$$

The probability that $X$ lands in this rectangle must be at least 0 but in terms of $F$ you should be able to check that

$$\begin{aligned} P(X \in R) &= P(u_1 < X_1 \leq v_1, u_2 < X_2 \leq v_2) \\ &= F(v_1, v_2) - F(u_1, v_2) - F(v_1, u_2) + F(u_1, u_2). \end{aligned}$$

So this combination of values of $F$ at the four corners of the rectangle must be non-negative. For a thorough discussion of the properties of multivariate cumulative distributions see some reference which **I must add**.

## 0.2 Discrete versus Continuous Distributions

**Definition**: The distribution of a random variable $X$ is called **discrete** (we also say $X$ is discrete) if there is a countable set $x_1, x_2, \cdots$ such that

$$P(X \in \{x_1, x_2 \cdots\}) = 1 = \sum_i P(X = x_i)\,.$$

In this case the **discrete density** or **probability mass function** of $X$ is

$$f_X(x) = P(X = x)\,.$$

**Definition**: The distribution of a random variable $X$ is called **absolutely continuous** (again we also say $X$ is absolutely continuous) if there is a function $f$ such that

$$P(X \in A) = \int_A f(x)dx \tag{1}$$

for any (Borel) set $A$. This is a $p$ dimensional integral in general. Equivalently

$$F(x) = \int_{-\infty}^{x} f(y)\,dy\,.$$

**Definition**: Any $f$ satisfying (1) is a **density** of $X$.

There are a few important warnings and observations here:

- Many statisticians use the word *continuous* instead of the phrase *absolutely continuous* for this property.

- Others use the word *continuous* to mean only that $F$ is a continuous function.
- If $X$ is absolutely continuous then for most (*almost all*) $x$ the function $F$ is differentiable at $x$ and

$$F'(x) = f(x)\,.$$

- Absolute continuity is the property which is needed for a function to be equal to the integral of its derivative. If the function is continuously differentiable, for instance, then it is continuous. If $F$ is continuously differentiable except at a finite number of points where it is continuous then $F$ is absolutely continuous.

**Example**: The Uniform[0,1] distribution. We say that $X$ is Uniform[0,1] if

$$F(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 < x < 1 \\ 1 & x \geq 1\,. \end{cases}$$

which is equivalent to

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ \text{undefined} & x \in \{0, 1\} \\ 0 & \text{otherwise}\,. \end{cases}$$

**Example**: The standard exponential distribution. We say that $X$ is exponential with mean 1 (sometimes written Exp(1)) if

$$F(x) = \begin{cases} 1 - e^{-x} & x > 0 \\ 0 & x \leq 0\,. \end{cases}$$

or equivalently

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ \text{undefined} & x = 0 \\ 0 & x < 0\,. \end{cases}$$

**Remark**: I am not going to create notes on all the well known distributions. I expect you will know something about all the famous distributions (including the uniform and exponential distributions I just mentioned).