Why probability theory? Results of quantum measurements are random!

Outcomes of experiments (classical results) in quantum theory are random variables

In order to define a probability on a set we need a few basic elements,

- **Sample space** Ω: The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.
- **Set of events** (or **event space**) $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ (called **events**) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment).[1].

  → focus, $\mathcal{F}$ = all subsets of Ω
- **Probability measure**: A function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties,
  - $P(A) \geq 0$, for all $A \in \mathcal{F}$
  - $P(\Omega) = 1$
  - If $A_1, A_2, \dots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

↑ disjoint union of sets

These three properties are called the **Axioms of Probability**.

Example: Consider the event of tossing a six-sided die.

- $\Omega = \{\text{all possible outcomes of an experiment}\}$

  $= \{1, 2, 3, 4, 5, 6\}$

- $\mathcal{F} = \{\text{events}\} = \{\text{subsets of } \Omega\}$

  $= \{\emptyset, \{1\}, \{2\}, \dots, \{6\}, \{1,2\}, \dots, \{1,4,5\}, \dots$

  ↑ empty subset

  $\dots \{1,2,3,4,5,6\}\}$

event: "outcome is an odd number"

$A = \{1, 3, 5\} \in \mathcal{F}$

- probability measure $\mathbb{P} : \mathcal{F} \to \mathbb{R}$

$$\boxed{\mathbb{P}(A) = \frac{|A|}{|\Omega|}}$$

$|X|$ = cardinality of the set X
= # of elements in X

↳ $|\Omega| = 6$ = the number of all possible outcomes

↳ $\mathbb{P}(\emptyset) = 0$

$\mathbb{P}(\Omega) = \mathbb{P}(\{1,2,3,4,5,6\}) = 1$

$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \dots = \mathbb{P}(\{6\}) = \frac{1}{6}$

$\mathbb{P}(\{\text{outcome is odd}\}) =$
$\mathbb{P}(\{1,3,5\}) = \frac{3}{6} = \frac{1}{2}$

↳ check $\mathbb{P}\left(\bigsqcup_i A_i\right) = \sum \mathbb{P}(A_i)$:

$\frac{1}{2} = \mathbb{P}(\{\text{odd}\}) = \mathbb{P}(\{1,3,5\}) =$
$= \mathbb{P}(\{1\}) + \mathbb{P}(\{3\}) + \mathbb{P}(\{5\})$
$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$

Example: 2 coins

→ the result of one coin toss can be "heads" → H

"tails" → T

first → ← second

- $\Omega = \{HH, HT, TH, TT\}$
- $\mathcal{F}$ = all subsets of Ω
- $\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{4}$ "the coins are fair"

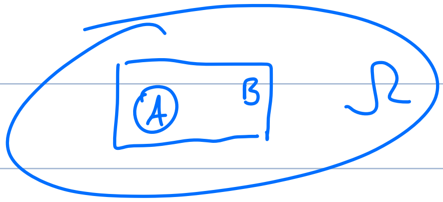
$$\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}) = \frac{1}{4}$$

Properties:

- If $A \subseteq B \implies P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$. $A \cap B$ = event that both A and B happen
- (Union Bound) $P(A \cup B) \leq P(A) + P(B)$.
- $P(\Omega \setminus A) = 1 - P(A)$. $\Omega \setminus A = A^c$ = the event that A does not happen
- (Law of Total Probability) If $A_1, \ldots, A_k$ are a set of disjoint events such that $\cup_{i=1}^k A_i = \Omega$, then $\sum_{i=1}^k P(A_k) = 1$.

Proof ideas

- $A \subseteq B$



$$B = A \sqcup (B \setminus A) \Rightarrow \mathbb{P}(B) = \mathbb{P}(A) + \underbrace{\mathbb{P}(B \setminus A)}_{\geq 0}$$

$$\Rightarrow \mathbb{P}(B) \geq \mathbb{P}(A)$$

- $\mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A)$: $\Omega = A \sqcup (\Omega \setminus A)$

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(\Omega \setminus A)$$

Conditional probability and independence

Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as,

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)} \leq 1$$

$A \cap B \subseteq B \Rightarrow$

$\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$

In other words, $P(A|B)$ is the probability measure of the event A after observing the occurrence of event B. Two events are called independent if and only if $P(A \cap B) = P(A)P(B)$ (or equivalently, $P(A|B) = P(A)$). Therefore, independence is equivalent to saying that observing B does not have any effect on the probability of A.

Example 1: 2 coins, probability that first toss is head, given that there is at least one head

$A = \{HH, HT\}$

$B = \{HH, HT, HH\}$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{P(B)} = \frac{2/4}{3/4} = \frac{2}{3}$$

$\mathbb{P}(A) = \frac{1}{2} \rightsquigarrow \mathbb{P}(A|B) \neq P(A) \Rightarrow A, B$ not indep.

Example 2: 2 coins, probability that the second toss is tails, given that the first one is head

$A = \{HT, TT\}$

$B = \{HH, HT\}$

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{HT\})}{\mathbb{P}(B)} = \frac{1/4}{2/4} = \frac{1}{2}$$

$\mathbb{P}(A) = \frac{1}{2} \Rightarrow$ A and B are independent

2 Random variables

Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads. Here, the elements of the sample space Ω are 10-length sequences of heads and tails. For example, we might have $w_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$. However, in practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails. Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails. These functions, under some technical conditions, are known as **random variables**.

More formally, a random variable X is a function $X : \Omega \longrightarrow \mathbb{R}$.[2] Typically, we will denote random variables using upper case letters $X(\omega)$ or more simply X (where the dependence on the random outcome ω is implied). We will denote the value that a random variable may take on using lower case letters x.

Example: In our experiment above, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses ω. Given that only 10 coins are tossed, $X(\omega)$ can take only a finite number of values, so it is known as a **discrete random variable**. Here, the probability of the set associated with a random variable X taking on some specific value k is

$$P(X = k) := P(\{\omega : X(\omega) = k\}).$$

Example: coin tosses - total number of Heads

$\Omega = \{ \omega : \omega =$ length 10 string with letters H or T $\}$

$|\Omega| = 2^{10} = 1024.$

$\mathcal{F} =$ all subsets of Ω $\quad \mathbb{P}(A) = \frac{|A|}{|\Omega|}$

$X : \Omega \to \mathbb{R} \quad X(\omega = (x_1, x_2, \dots, x_{10})) = |\{ i : x_i = H \}|$

- $\mathbb{P}(X = 0) = \mathbb{P}(\{\omega : X(\omega) = 0\}) = \mathbb{P}(\{$no heads$\})$

$= \mathbb{P}(\{TT \cdots T\}) = 1/2^{10}$

$\mathbb{P}(X = 1) = \mathbb{P}(\{$exactly 1 heads$\}) = \frac{10}{2^{10}}$, etc

2.2 Probability mass functions

When a random variable X takes on a finite set of possible values (i.e., X is a discrete random variable), a simpler way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume. In particular, a *probability mass function (PMF)* is a function $p_X : \Omega \to \mathbb{R}$ such that

$$p_X(x) \triangleq P(X = x).$$

In the case of discrete random variable, we use the notation $Val(X)$ for the set of possible values that the random variable X may assume. For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of coin, then $Val(X) = \{0, 1, 2, \dots, 10\}$.

Properties:

- $0 \leq p_X(x) \leq 1$.
- $\sum_{x \in Val(X)} p_X(x) = 1$.
- $\sum_{x \in A} p_X(x) = P(X \in A)$.

Example: 10 coin tosses - A = between 2 and 5 Heads

$\mathbb{P}(2 \leq X \leq 5) = \mathbb{P}(X \in \{2, 3, 4, 5\}) =$

$= \sum_{x \in \{2,3,4,5\}} p_X(x) = p_X(2) + p_X(3) + p_X(4) + p_X(5)$

2.4 Expectation (also known as the mean or the average)

Suppose that X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$ is an arbitrary function. In this case, $g(X)$ can be considered a random variable, and we define the **expectation** or **expected value** of $g(X)$ as

$$E[g(X)] \triangleq \sum_{x \in Val(X)} g(x) p_X(x).$$

Example: 2 coins, expectation of the number of heads

$\mathbb{P}(\{HH\})$ $\mathbb{P}(\{HT,TH\})$

$X(\{HH\}) = 2$

$X(\{HT\}) = X(\{TH\}) = 1$

$X(\{TT\}) = 0$

$$\mathbb{E}X = 2 \cdot \overset{\mathbb{P}(\{HH\})}{\frac{1}{4}} + 1 \cdot \overset{\mathbb{P}(\{HT,TH\})}{\frac{2}{4}} + 0 \cdot \frac{1}{4}$$

$$= 1$$

on average, we get "Heads" once when tossing two coins

Properties:

- $E[a] = a$ for any constant $a \in \mathbb{R}$.
- $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.
- (Linearity of Expectation) $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.
- For a discrete random variable X, $E[1\{X = k\}] = P(X = k)$.

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A]$$

Example: 10 coins, expectation of the number of heads

$$\mathbb{E}X = \sum_{x=0}^{10} x \cdot \mathbb{P}(X = x) = \sum_{x=0}^{10} x \cdot \frac{\binom{10}{x}}{2^{10}}$$

$$= \frac{10}{2} = 5$$

on average, we get 5 times "H" if we toss a coin 10 times.

2.5 Variance

The **variance** of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean. Formally, the variance of a random variable X is defined as

$$Var[X] \triangleq E[(X - E(X))^2] \geq 0$$

Using the properties in the previous section, we can derive an alternate expression for the variance:

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2, \end{aligned}$$

where the second equality follows from linearity of expectations and the fact that $E[X]$ is actually a constant with respect to the outer expectation.

Properties:

- $Var[a] = 0$ for any constant $a \in \mathbb{R}$.
- $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$.

- $Var(X) = 0 \iff \mathbb{E}\left[\underbrace{(X - \mathbb{E}(x))^2}_{\geq 0}\right] = 0$

$\iff X = \mathbb{E}(X)$ with prob 1

$\implies X$ is constant

- $var[aX] = \mathbb{E}[(ax)^2] - (\mathbb{E}[ax])^2$

$= a^2(\mathbb{E}X^2 - (\mathbb{E}X)^2) = a^2 var(X)$

2.6 Some common random variables

- $X \sim Bernoulli(p)$ (where $0 \leq p \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise.

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

$\text{val}(X) = \{0, 1\}$

$\mathbb{P}(X=1) = p \qquad \mathbb{P}(X=0) = 1-p$

- $X \sim Binomial(n, p)$ (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

X : sum of n independent Bernoulli(p)

Example : n cointosses, $X = \#$ heads

$$X = X_1 + X_2 + \cdots + X_n$$

$$X_i = \begin{cases} 1 & \text{if} \quad i\text{th toss is "H"} \\ 0 & \text{if} \quad \underline{\hspace{3cm}} \text{ "T"} \end{cases}$$

if coin is fair : X_i are indep Bernoulli$(\frac{1}{2})$

$\Rightarrow X$ = Binomial r.v. of param $(n, 1/2)$

- $X \sim Geometric(p)$ (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

$$p(x) = p(1-p)^{x-1}$$

$$\mathbb{P}(X=1) = \mathbb{P}(\{\text{you get heads on the } 1^{st} \text{ toss}\})$$

$$= p$$

$$\mathbb{P}(X=2) = \mathbb{P}(\{1^{st} \text{ time you get "H" is on toss } 2\})$$

$$= \mathbb{P}(\{TH\}) = (1-p)\,p$$

$$\mathbb{P}(X=n+1) = p(1-p)^{n}$$

$$\mathbb{E}[X] = \sum_{n\geq 1} n\,p(1-p)^{n-1} = p\sum_{n\geq 1} n(1-p)^{n-1} = p\frac{\partial}{\partial p}\sum_{n\geq 1}(1-p)^{n}$$

$$= p\frac{\partial}{\partial p}\,\frac{1-p}{p} = p^{-1} = 1/p$$

3 Two random variables

Thus far, we have considered single random variables. In many situations, however, there may be more than one quantity that we are interested in knowing during a random experiment. For instance, in an experiment where we flip a coin ten times, we may care about both $X(\omega) =$ the number of heads that come up as well as $Y(\omega) =$ the length of the longest run of consecutive heads. In this section, we consider the setting of two random variables.

$Y(\{THTHHHTHH\}) = 3$

3.2 Joint and marginal probability mass functions

If X and Y are discrete random variables, then the **joint probability mass function** $p_{XY} : \mathbb{R}\times\mathbb{R} \to [0,1]$ is defined by

$$p_{XY}(x,y) = P(X = x, Y = y).$$

Here, $0 \leq P_{XY}(x,y) \leq 1$ for all x, y, and $\sum_{x\in Val(X)} \sum_{y\in Val(Y)} P_{XY}(x,y) = 1$.

How does the joint PMF over two variables relate to the probability mass function for each variable separately? It turns out that

$$p_X(x) = \sum_y p_{XY}(x,y).$$

$P_Y(y) = \sum_x P_{XY}(x,y)$

and similarly for $p_Y(y)$. In this case, we refer to $p_X(x)$ as the **marginal probability mass function** of X. In statistics, the process of forming the marginal distribution with respect to one variable by summing out the other variable is often known as "marginalization."

Example: 2 coins

$X = \#\text{heads} \in \{0,1,2\}$

$\Omega = \{HH, HT, TH, TT\}$

$Y = \mathbb{1}_{2^{nd}\text{ toss is heads}} = \{0,1\}$

Y \ X	0	1	2
0	1/4		
1	0	1/4	

...

$\mathbb{P}(0,0) = \mathbb{P}(X = Y = 0) = \mathbb{P}(X = 0 \text{ and } "\cdot T")$

$= \mathbb{P}(TT) = 1/4$

$\mathbb{P}(1,1) = \mathbb{P}(\{\cdot H\}) = \mathbb{P}(\{TH\}) = 1/4$

3.6 Independence

Two random variables X and Y are **independent** if $F_{XY}(x, y) = F_X(x)F_Y(y)$ for all values of x and y. Equivalently,

- For discrete random variables, $p_{XY}(x, y) = p_X(x)p_Y(y)$ for all $x \in Val(X)$, $y \in Val(Y)$.
- For discrete random variables, $p_{Y|X}(y|x) = p_Y(y)$ whenever $p_X(x) \neq 0$ for all $y \in Val(Y)$.

Example: 2 coins

Recall Events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

$X = \mathbb{1}_{1st}$ coin is heads

$Y = \mathbb{1}_{2^{nd}}$ coin is tails

$\mathbb{P}(X) = \frac{1}{2} = \mathbb{P}(Y)$

$\mathbb{P}(X \cap Y) = \mathbb{P}(\{HT\}) = \frac{1}{4}$

$\Rightarrow \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} \Rightarrow$ OK.

Lemma 3.1. *If X and Y are independent then for any subsets $A, B \subseteq \mathbb{R}$, we have,*

$$P(\underbrace{X \in A}_{E}, \underbrace{Y \in B}_{F}) = P(\underbrace{X \in A}_{E})P(\underbrace{Y \in B}_{F})$$

By using the above lemma one can prove that if X is independent of Y then any function of X is independent of any function of Y.

X, Y indep. r.v $\Rightarrow$ $\mathbb{P}(E \cap F) = \mathbb{P}(E) \cdot \mathbb{P}(F)$

(Cond for indep. of events)

For $A, B \subseteq \mathbb{R}$ $\quad \mathbb{P}(X \in A \text{ and } Y \in B) = \mathbb{P}(X \in A) \cdot$
$\cdot\, \mathbb{P}(Y \in B)$

$\Rightarrow$ the r.v. X and Y are independent.

(reciprocal statement)