1 Probability Theory

1.1 Axiomatic Definition (Kolmogoroff)

Basic Building Blocks

- Ω: the *sample space*, a set of *basic outcomes*.
- $S \subseteq 2^{\Omega}$: the *event space*, a *σ-field* on Ω:
  - non-empty: $S \neq \emptyset$
  - closed under complement: $\forall x \in S : \overline{x} \in S$
  - closed under union: $\forall x, y \in S : x \cup y \in S$

A *probability distribution* for an event space S is a (total) function: $P : S \to \mathbb{R}$ which fulfills Axioms (1) through (3):

1. For arbitrary events A, $A \in S \Rightarrow P(A) \geq 0$
2. $P(\Omega) = 1$
3. For arbitrary sequences of pairwise disjoint events $A_1, A_2, \ldots \in S$:

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

Theorem 1 (Empty Event) $P(\emptyset) = 0$

Theorem 2 (Finite Sums) For arbitrary finite sequences of pairwise disjoint events $A_1, A_2, \ldots, A_n$:

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

Theorem 3 (Subtraction Rule) For arbitrary $A \in S$:

$$P(\overline{A}) = 1 - P(A)$$

Theorem 4 (Probability Range) For arbitrary $A \in S$:

$$0 \leq P(A) \leq 1$$

Theorem 5 (Subevent Probability) For $A, B \in S$:

$$A \subseteq B \Rightarrow P(A) \leq P(B)$$

Theorem 6 (Addition Rule) For arbitrary $A, B \in S$:

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

Law of Large Numbers: Given the axioms (and theorems) above, we can approach the "statistical" definition of probability by noting that since the relative frequency $\frac{n(A)}{n}$ can vary between experimental runs, we can speak of the probability that it lies in a given interval. If the actual probability of the event is $P(A)$, then we can say for an arbitrary $\varepsilon \in \mathbb{R}$ that:

$$\lim_{n\to\infty} P\left(\left|\frac{n(A)}{n} - P(A)\right| < \varepsilon\right) = 1$$

This assertion is known as the **Law of Large Numbers**.

1.2 Stochastic (In)Dependence

Definition 1 (Stochastic Independence) Two events A and B are said to be *independent* iff:

$$P(A \cap B) = P(A) \times P(B)$$

Theorem 7 If two events A and B are independent, then A and $\overline{B}$, $\overline{A}$ and B, as well as $\overline{A}$ and $\overline{B}$ are also independent.

1.2.1 Conditional Probability

Definition 2 Given $P(B) > 0$, the *conditional probability of A given B* $P(A|B)$ is defined as:

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

Theorem 8 Assuming $P(B) > 0$, two events A and B are independent iff

$$P(A|B) = P(A)$$

Theorem 9 (Multiplication Rule) For all $A, B \in S$:

$$P(B)P(A|B) = P(A \cap B) = P(A)P(B|A)$$

Theorem 10 (Chain Rule) For $n \in \mathbb{N}$, $A_1, \ldots, A_n \in S$:

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

... somewhat prettier:

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \prod_{i=2}^{n} P\left(A_i \,\middle|\, \bigcap_{j=1}^{i-1} A_j\right)$$

The "Chain Rule", together with assumptions of independence between events, provides a powerful tool for reduction of the computational complexity of many stochastic models (read: "get used to it").

Theorem 11 (Bayes' Theorem)

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

1.3 Random Variables

- Axioms Ω can be any (non-empty) set.
- It is often easier (sometimes even more natural) to restrict our observations to a set of numbers, such as $\mathbb{R}$.
- *Random variables* are used to map any sample space Ω (partially) to $\mathbb{R}$.

Definition 3 (Random Variable) Let S be an event space over a set Ω of basic outcomes, and let P be a probability distribution over S. A *random variable (over Ω)* is a function:

$$X : \Omega \to \mathbb{R}$$

such that for all $x \in \mathbb{R} : \{\omega \in \Omega \mid X(\omega) = x\} \in S$

Example 1 (Coin Toss: Random Variable) Let $\Omega = \{heads, tails\}$. Then, $X = \{(heads, 0), (tails, 1)\}$ is a random variable over Ω:

$$\begin{array}{rcl} X(heads) & = & 0 \\ X(tails) & = & 1 \end{array}$$

Example 2 (Two Dice: Random Variable) For the throw of two dice, let $\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$. Then, $X = \bigcup_{i=1}^{6} \bigcup_{j=1}^{6} \{((i,j), i+j)\}$ is a random variable over Ω:

$$\begin{array}{rcccl} X(1,1) & = & 1+1 & = & 2 \\ X(1,2) & = & 1+2 & = & 3 \\ & & \vdots & & \\ X(6,6) & = & 6+6 & = & 12 \end{array}$$

1.3.1 Discrete *vs.* Continuous Random Variables

A *discrete random variable* is one whose image under Ω ($X(\Omega)$) is finite or countable. A *continuous random variables* is one whose image under Ω is uncountably infinite and continuous (for instance, $X(\Omega) = [0,1]$), and whose cumulative distribution function is differentiable with a continuous derivative except in at most a finite number of points.

1.4 Probability Mass Functions

A *probability mass function* (sometimes simply called a *probability function*) is determined by a random variable X and a probability distribution P over $\Omega = \text{dom}(X)$.

Definition 4 (Probability Mass Function) Let X be a random variable over a set Ω of basic outcomes, and let P be a probability distribution over Ω. Then, the *probability mass function (pmf)* associated with X is given by:

$$p_X(x) = p(X = x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

It can be shown that p_X is always a probability distribution over $\mathbb{R}$. Often, the subscript X is omitted from the probability mass function p_X when the random variable concerned is clear from the context.

Example 3 (Coin Toss: Probability Function) Assuming the coin is fair, $P(heads) = P(tails) = \frac{1}{2}$. Therefore, $p_X(0) = p_X(1) = 0.5$, and $\forall x \in \mathbb{R} - \{0, 1\} : p_X(x) = 0$.

Example 4 (Two Dice: Probability Function) Probability mass is distributed according to the following table:

x	$X^{-1}(x)$	$p_X(x)$
2	$\{(1,1)\}$	1/36
3	$\{(1,2),(2,1)\}$	2/36
4	$\{(1,3),(2,2),(3,1)\}$	3/36
5	$\{(1,4),(2,3),(3,2),(4,1)\}$	4/36
6	$\{(1,5),(2,4),(3,3),(4,2),(5,1)\}$	5/36
7	$\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$	6/36
8	$\{(2,6),(3,5),(4,4),(5,3),(6,2)\}$	5/36
9	$\{(3,6),(4,5),(5,4),(6,3)\}$	4/36
10	$\{(4,6),(5,5),(6,4)\}$	3/36
11	$\{(5,6),(6,5)\}$	2/36
12	$\{(6,6)\}$	1/36

1.5 Some Properties of Random Variables

1.5.1 Sample Space

Definition 5 (Sample Space of a Random Variable) The *sample space of a random variable* X is just the image of X under Ω, and is written Ω_X

$$\Omega_X := X(\Omega) = \bigcup_{\omega \in \Omega} \{X(\omega)\}$$

1.5.2 Expectation Value

Definition 6 (Expectation Value) The *expectation value* $E(X)$ of a random variable X is simply the *mean* or *average value* of that variable, computed as a weighted sum of the variable's sample space:

$$E(X) = \sum_{x \in \Omega_X} p_X(x) \cdot x$$

One common convention uses the Greek letter μ to denote the expectation value of a random variable, when the particular variable is clear from the surrounding context: $\mu = E(X)$.

Example 5 (Coin Toss: Expectation Value)

$$\begin{array}{lccccccc} E(X) & = & X(heads) & \cdot & p_X(X(heads)) & + & X(tails) & \cdot & p_X(X(tails)) \\ & = & 0 & \cdot & 0.5 & + & 1 & \cdot & 0.5 \\ & = & 0.5 \end{array}$$

Example 6 (Two Dice: Expectation Value)

$$\begin{aligned} E(X) &= \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \frac{3}{36} \cdot 4 + \frac{4}{36} \cdot 5 + \frac{5}{36} \cdot 6 + \frac{6}{36} \cdot 7 + \\ &\quad \frac{5}{36} \cdot 8 + \frac{4}{36} \cdot 9 + \frac{3}{36} \cdot 10 + \frac{2}{36} \cdot 11 + \frac{1}{36} \cdot 12 \\ &= 7 \end{aligned}$$

Every function $g : \mathbb{R} \to \mathbb{R}$ can be used to map a random variable X to a new random variable $Y = g(X)$. The expectation value of such a functionally composed random variable is given by:

$$E(Y) = E(g(X)) = \sum_{x \in \Omega_X} p(x) \cdot g(x)$$

In particular, it is interesting (and useful) to note that ...

Theorem 12 (Expectation Value of Linear Functions) If $g : \mathbb{R} \to \mathbb{R}$ is a linear function – that is, if $g(x) = ax + b$ for some constants $a, b \in \mathbb{R}$ and for all $x \in \mathbb{R}$, then $E(g(X))$ can be computed as a function of $E(X)$:

$$E(g(X)) = E(aX + b) = aE(X) + b$$

Theorem 13 (Sum of Expectation Values) The sum of linear combinations of arbitrary random variables can be computed in terms of the expectation values of the variables themselves:

$$E(aX + bY) = aE(X) + bE(Y)$$

Theorem 14 (Product of Independent Expectation Values) If X and Y are independent random variables, then:

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

1.5.3 Variance

Definition 7 (Variance) The *variance* of a random variable X is a measure of how widely that variable's values are distributed, computed as the average square difference between the variable's values its mean:

$$\begin{array}{rcl} Var(X) & := & E((X - E(X))^2) \\ & = & E(X^2) - E(X)^2 \end{array}$$

It is common to write σ^2 to refer to the variance of a random variable, when the random variable in question is clear from the surrounding context: $\sigma^2 = Var(X)$. This is largely due to the fact that the *standard deviation* – commonly written σ – is defined as the square root of the variance: $\sigma = \sqrt{Var(X)}$.

Example 7 (Coin Toss: Variance)

$$\begin{array}{rcl} Var(X) & = & E((X - 0.5)^2) \\ & = & \sum_{x \in \Omega_X} p_X(x) \cdot (x - 0.5)^2 \\ & = & 0.5 \cdot (0 - 0.5)^2 + 0.5 \cdot (1 - 0.5)^2 \\ & = & 0.5 \cdot -0.5^2 + 0.5 \cdot .5^2 \\ & = & 0.5 \cdot 0.25 + 0.5 \cdot .25 \\ & = & 0.125 + 0.125 \\ & = & 0.25 \end{array}$$

Example 8 (Two Dice: Variance)

$$\begin{array}{rcl} Var(X) & = & E(X^2) - E(X)^2 \\ & = & 2^2 \cdot \frac{1}{36} + 3^2 \cdot \frac{2}{36} + 4^2 \cdot \frac{3}{36} + 5^2 \cdot \frac{4}{36} + 6^2 \cdot \frac{5}{36} + 7^2 \cdot \frac{6}{36} + \\ & & 8^2 \cdot \frac{5}{36} + 9^2 \cdot \frac{4}{36} + 10^2 \cdot \frac{3}{36} + 11^2 \cdot \frac{2}{36} + 12^2 \cdot \frac{1}{36} \\ & & -7^2 \\ & = & 4 \cdot \frac{1}{36} + 9 \cdot \frac{2}{36} + 16 \cdot \frac{3}{36} + 25 \cdot \frac{4}{36} + 46 \cdot \frac{5}{36} + 49 \cdot \frac{6}{36} + \\ & & 64 \cdot \frac{5}{36} + 81 \cdot \frac{4}{36} + 100 \cdot \frac{3}{36} + 121 \cdot \frac{2}{36} + 144 \cdot \frac{1}{36} \\ & & -49 \\ & = & 35/6 \\ & \approx & 5.83 \end{array}$$

1.6 Combinatorics

Combinatorics is the study of *combinations* of elements – in particular, the number(s) of ways of building lists (sets) of length (size) k out of a set of n possible elements. We have four cases to consider, depending on:

1. whether or not we care in which order items are selected, and
2. whether or not we can re-use previously selected items.

	−Reuse	**+Reuse**
+Order	$(n)_k$	n^k
−Order	$\binom{n}{k}$	$\binom{n+k-1}{k}$

Where:

$$\begin{aligned} n^k &= \overbrace{n \cdot n \cdot \ldots \cdot n}^{k \text{ factors}} \\ n! &= n \cdot (n-1) \cdot \ldots \cdot 1 \\ (n)_k &= n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!} \\ \binom{n}{k} &= \frac{(n)_k}{k!} = \frac{n!}{(n-k)!k!} \end{aligned}$$

For proofs, see Krenn and Samuelsson (1997, Section 1.2.7).

Example 9 (Lotto) For a lottery game in which 6 balls are selected from a pool of balls uniquely numbered from 1 to 49, where the balls are not replaced after having been drawn and where order does not matter, there are $\binom{49}{6} = 13,983,816$ possible results.

1.7 Some Common Probability Distributions

- In practical applications, we often have no prior knowledge of the probability distribution(s) for the events we wish to model.
- Although the *relative frequency* $\frac{n(A)}{N}$ can serve as an estimate for $P(A)$, such estimates become quite cumbersome when a large sample space (i.e. natural language data) is involved.
- Instead, we can often use prior knowledge of the domain we wish to model to assign our phenomena to one of a number of common *families of distributions*, and formulating our model in terms of the general properties ("parameters") of the distribution family.
- *Caveat:* such a technique can go horribly, horribly wrong if we choose the wrong underlying distribution, but it's pretty darned handy nonetheless.

1.7.1 Bernoulli Distribution

A stochastic experiment with exactly two possible outcomes is called a *Bernoulli trial*, and is described by a parametric distribution known as a *Bernoulli distribution*, which is completely specified by the probability p of one of the two possible outcomes. Random variables for Bernoulli trials are typically mapped to the values 0 ("failed trial") and 1 ("successful trial").

For a random variable X and a probability p, we write $X \sim b(1, p)$ to indicate that X has a Bernoulli distribution with $p_X(1) = p$. For such a random variable, the following properties hold:

$$\begin{array}{rrcrcl}
\textbf{Success Probability:} & b(1;1,p) & = & p_X(1) & = & p \\
\textbf{Failure Probability:} & b(0;1,p) & = & p_X(0) & = & 1-p \\
\textbf{Mean:} & E(b(1,p)) & = & E(X) & = & p \\
\textbf{Variance:} & Var(b(1,p)) & = & Var(X) & = & p \cdot (1-p)
\end{array}$$

Example 10 (Fair Coin) A random variable $X = \{heads \mapsto 1, tails \mapsto 0\}$ describing a single coin toss has a Bernoulli distribution:

$$X \sim b(1, P(heads))$$

Example 11 (Lotto) Let X be a random variable describing a lottery game as in Example 9, where for a 6-number ticket $t = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, let $X(t) = 1$ iff all of the numbers x_i are drawn, and let $X(t) = 0$ otherwise. If the lottery game is fair, then X has a Bernoulli distribution with probability $\frac{1}{13{,}983{,}816} \approx 7.15 \times 10^{-8}$:

$$X \sim b\left(1, 0.0000000715\right)$$

1.7.2 Binomial Distribution

A finite series of Bernoulli trials in which each trial is independent from all other trials is described by a *Binomial distribution*, and is completely specified by the number n of trials in the series and the probability p of success in any trial.

For a random variable X, a natural number n, and a probability p, we write $X \sim b(n, p)$ to indicate that X has a binomial distribution over n independent trials with $p_X(1) = p$ in any of those trials, and for a natural number $k \leq n$, we write $b(k; n, p)$ to indicate the probability of k successes in n such trials – but how do we find it?

- Given an $X \sim b(n, p)$ and a $k \leq n$, the probability of any specific (ordered) sequence containing k successes and $n - k$ failures is $p^k(1-p)^{n-k}$, since the individual trials are independent.

- We are interested only in the number k of successful trials, ignoring the order in which they occur, so we must count every sequence containing k successes as a single event.

- There are exactly $\binom{n}{k}$ such sequences – this is just the number of ways of picking k successful trial indices from a "bag" of n trial indices *without* regard to order and *without* re-using previously selected indices.

Therefore, for a random variable $X \sim b(n,p)$, the following properties hold:

$$\begin{array}{rrcl}
\textbf{k-Success:} & b(k;n,p) & = & \binom{n}{k} p^k (1-p)^{n-k} \\
\\
\textbf{Mean:} & E(X \sim b(n,p)) & = & \sum_{k=0}^{n} k \cdot b(k;n,p) \\
 & & = & np \\
\\
\textbf{Variance:} & Var(X \sim b(n,p)) & = & \sum_{k=0}^{n} (k-np)^2 \cdot b(k;n,p) \\
 & & = & np(1-p)
\end{array}$$

Example 12 (2 Fair Coins) Let X be a random variable modelling the toss of 2 fair coins which maps each experiment to the number of coins which come up "heads". Then, X has a binomial distribution:

$$\begin{array}{rcl}
X & \sim & b(2, 0.5) \\
\\
b(0;2,0.5) & = & 1 \cdot 0.5^0 \cdot (1-0.5)^{2-0} = 0.25 \\
b(1;2,0.5) & = & 1 \cdot 0.5^1 \cdot (1-0.5)^{2-1} = 0.5 \\
b(2;2,0.5) & = & 1 \cdot 0.5^2 \cdot (1-0.5)^{2-2} = 0.25 \\
\\
E(X) & = & 2 \cdot 0.5 = 1 \\
Var(X) & = & 2 \cdot 0.5 \cdot (1-0.5) = 0.5
\end{array}$$

Example 13 (Lotto) Let X be a random variable describing a sequence of fair lottery drawings as in Example 9 whose values are the number of full (6-position) matches ("winning tickets") in the sequence of lottery drawings. Suppose that Lotto drawings occur weekly and that X describes Lotto participation once a week since the birth of Socrates (ca. 470 BC), or roughly 128648 consecutive weeks. Then, X has a binomial distribution:

$$\begin{array}{rcl}
X & \sim & b(n = 128648, p \approx 7.15 \times 10^{-8}) \\
\\
b(1;n,p) = E(X) & \approx & 128648 \cdot 7.15 \times 10^{-8} \approx 0.0092
\end{array}$$

1.7.3 Normal (Gaussian) Distribution

The *Normal* or *Gaussian distribution* – courtesy of Johann Carl Friedrich Gauss – is a continuous distribution, and is often used as an approximation for the binomial distribution when n is large (since computation of the factorials involved in the binomial coefficient is highly problematic in these cases).

Definition 8 (Normal Distribution) Let X be a continuous random variable with mean μ and variance σ^2. X is said to be *normally distributed*, written $X \sim N(\mu, \sigma)$ iff it is characterized by the probability density function:

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

Roughly speaking, when dealing with continuous distributions, pointwise probabilities are replaced with interval probabilities, and summation in the formulae for discrete distributions are replaced with integrals.

1.8 Parameter Estimation

- Often, we can't observe probabilities $P(X)$ directly, but we can perform a series of experiments and tally the observed values $\vec{x}$ of a random variable X.
- If we can additionally sort our phenomena into a parameterized model, we can express the distribution function for X as conditioned on the (unknown) parameters $\theta \in \Xi$ of that model: $P(x) = P(x|\theta)$. For a binomial distribution $X \sim b(n, p)$, $\theta = p$.
- Given an observed set of values and an expression of their dependance on one or more model parameters, we can attempt to find the model parameters which best account for the observed data.
- Having thus estimated a model's parameters, we can effectively fix the (imperfect) model, using it to predict the outcomes of further experiments.

1.8.1 Maximum Likelihood Estimation

Definition 9 (Likelihood Function) A *likelihood function* for parameters $\theta \in \Xi$ and an observed sample $X = x$ of a random variable X is a function $L : \Xi \to \mathbb{R}$ which can be expressed as a (linear function of a) conditional probability function[1] with its first argument held fixed to the observation x:

$$L(\theta) = aP(X = x|\theta) + b \qquad a, b \in \mathbb{R}$$

[1] Despite the formulation of likelihood functions in terms of probabilities, likelihood functions are not in general probability functions themselves.

Example 14 (Likelihood: Coin Toss) A coin is tossed 10 times, resulting in 8 heads and 2 tails. By prior knowledge, we assume that the trials are independent, and thus a random variable X modelling the process has a binomial distribution: $X \sim b(10, p)$. The model's parameters θ are given by p, so $\Xi = [0, 1]$ and:

$$L(p) = \binom{10}{8} p^8 (1-p)^2$$

Definition 10 (Maximum Likelihood) Given a likelihood function $L : \Xi \to [0, 1]$ the parameters $\theta \in \Xi$ which maximize $L(\theta)$ are called the *maximum likelihood estimate* for θ:

$$MLE(\theta) = \arg\max_{\theta} L(\theta)$$

Example 15 (MLE: Coin Toss) Continuing Example 14, we find (unsurprisingly) that:

$$MLE(p) = \arg\max_{p} \binom{10}{8} p^8 (1-p)^2 = 0.8$$

1.8.2 Bayesian Statistics

Parameter estimation in Bayesian statistics takes into account *prior belief* in the acceptability of model parameters θ. Formally, this involves an additional probability distribution on Ξ and the application of Bayes' law in order to revise prior beliefs in the face of new evidence, thus computing a *posterior distribution*:

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)}$$

Although the prior probability $P(x)$ of x is unknown, we can ignore it if we just want to find the parameters which maximize the posterior distribution – that is, if we want to compute a new "best" value for θ given x:

$$MAP(\theta) = \arg\max_{\theta} P(\theta|x)$$

Given our (possibly new) parameter estimate θ, we can turn our attention to new data. Iteration of this process is known as *Bayesian updating*.

Example 16 (Bayesian Updating: Coin Toss) Consider the coin toss experiment from Example 14. Suppose our prior belief is modelled by by the function $G : \mathbb{R} \to \mathbb{R}$, where:

$$G(p) = \begin{cases} 6p(1-p) & \text{if } 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Then for the observed sequence of 8 heads and 2 tails, we have:

$$\begin{aligned} P(p|x) &= \frac{P(p)P(x|p)}{P(x)} \\ &= \frac{p^8(1-p)^2 \cdot 6p(1-p)}{P(x)} \\ &= \frac{6p^9(1-p)^3}{P(x)} \end{aligned}$$

so that:

$$\begin{aligned} MAP(p) &= \arg\max_p P(p|x) \\ &= \arg\max_p \frac{6p^9(1-p)^3}{P(x)} \\ &= 0.75 \end{aligned}$$

References

B. Krenn and C. Samuelsson. *The Linguist's Guide to Statistics.* URL `http://www-old.coli.uni-saarland.de/~thorsten/c-alg/stat_cl.ps.gz`. Unpublished manuscript, 1997.