Introduction to Probability Theory – Part I [1]

Erhard Hinrichs

Seminar für Sprachwissenschaft
Eberhard-Karls Universität Tübingen

[1] Largely based on material from Sharon Goldwater's tutorial *Basics of Probability Theory* (henceforth abbreviated as SGT), available at: `https://homepages.inf.ed.ac.uk/sgwater/math_tutorials.html`

On-line Resources

https://homepages.inf.ed.ac.uk/sgwater/math_tutorials.html

Events and Probabilities

- What is probability and why do we care?
- Sample spaces and events
- The probability of an event
- Probability distributions

What is Probability Theory?

Probability is common sense reduced to calculation. (Pierre-Simon Laplace)

Probability theory is a branch of mathematics that allows us to reason about events that are inherently random. SGT, p. 2

Probability theory is the mathematical framework that allows us to analyze chance events in a logically sound manner. The probability of an event is a number indicating how likely that event will occur. This number is always between 0 and 1, where 0 indicates impossibility and 1 indicates certainty. `https://seeing-theory.brown.edu/basic-probability/index.html`

What is Probability?

Intuitively, we understand the notion of probability as a measure of how likely it is that an event occurs or that a state of affairs holds.
However: it is far from trivial to arrive at a non-circular definition of probability.

- If we assign probabilities to a set of possible outcomes of an experiment, we often use assumptions about their relative likelihood; e.g. that the outcomes are equally likely.
- If we define likelihood in terms of probability, but then appeal to the notion of (equal) likelihood in our account of probability, then the account becomes circular.

Two Ways for Avoiding Circular Accounts of Probability

- **Alternative 1**: invoke the notion of the relative frequencies of the observed outcomes of a series of experiments.
  - resulting in Frequentist Theories of Probability; formalized by Egon Pearson, Jergy Newman, R.A. Fisher, and others.
- **Alternative 2**: ground the notion of probability in logic and in the notion of common sense reasoning and degrees of belief.
  - resulting in Bayesian theories, formalized by Cox and Jaynes

Plausible Reasoning

Suppose some dark night a policeman walks down a street, apparently deserted. Suddenly he hears a burglar alarm, looks across the street, and sees a jewelry store with a broken window. Then a gentleman wearing a black mask comes crawling out through the broken window, carrying a bag which turns out to be full of expensive jewelry. The policeman doesn't hesitate at all in deciding that this gentleman is dishonest. But by what reasoning process does he arrive at this conclusion?

E.T. Jaynes (2003).Probability Theory - The Logic Science. Cambridge University Press, p. 3.

An Alternative Explanation

It might be ... that this gentleman was the owner of the jewelry store and he was coming home from a masquerade party, and didn't have a key with him. However, just as he walked by his store, a passing truck threw a stone through the window, and he was only protecting his own property.

E.T. Jaynes (2003), p. 3

Deductive Reasoning (*Apodeixis*)

Classical Syllogisms (*modus ponens* and *modus tollens*)

(S1)

If *A* is true, then *B* is true.
A is true.

B is true. ∴

(S2)

If *A* is true, then *B* is true.
B is false.

A is false. ∴

Plausible Reasoning

Weak Syllogisms (*Epagogy*)

$$\text{(S3)} \qquad \frac{\begin{array}{l}\text{If } A \text{ is true, then } B \text{ is true.}\\ B \text{ is true.}\end{array}}{A \text{ becomes more plausible.} \qquad \therefore}$$

$$\text{(S4)} \qquad \frac{\begin{array}{l}\text{If } A \text{ is true, then } B \text{ is true.}\\ A \text{ is false.}\end{array}}{B \text{ becomes less plausible.} \qquad \therefore}$$

E.T. Jaynes (2003), p. 5

Revisiting the Policeman Scenario

... the brain, in doing pausible reasoning, not only decides whether something becomes more plausible or less plausible, but that it evaluated the degree of pausibility somehow. The plausibility for rain by 10 AM depends very much on the darkness of those clouds at 9:45. And the brain also makes use of old information as well as the specific data of the problem.

To illustrate that the policeman was also making use of the past experience of policemen in general, we have only to change that experience. Suppose that events like these happened several times every night to every policeman – and that in every case the gentleman turned out to be completely innocent. Very soon, policemen would learn to ignore such trivial things.

Revisiting the Policeman Scenario (continued)

Thus, in our reasoning we very much depends on our ***prior information*** *to help us in evaluating the degree of plausibility in a new problem. The reasoning process goes on unconsciously, almost instantaneously, and we conceal how complicated it really is by calling it* ***common sense****.*

E.T. Jaynes (2003), p. 6

Use of Probabilities in Computational Linguistics, SGT p. 3

- **Generation/prediction.** Reasoning from causes to effects: Given a known set of causes and knowledge about how they interact, what are likely/unlikely outcomes?
  - Example: Given an alphabet and knowledge about co-occurrences of characters and given a string of characters, what is/are the most likely next character(s)?
- **Inference.** Reasoning from effects to causes: Given knowledge about possible causes and how they interact, as well as some observed outcomes, which causes are likely/unlikely?
  - Example: Observe many headlines from a newspaper corpus, determine which words are domain-specific and which are domain-independent?

Basic Definitions: Sample Space, Event, Statistical Experiment; SGT, p.4

- A STATISTICAL EXPERIMENT is an action or occurrence that can multiple different outcomes, all of which can be specified in advance, but where the particular outcome that will occur cannot be specified because it depends on random chance.
- The SAMPLE SPACE of a statistical experiment is the set of all possible outcomes (also known as SAMPLE POINTS).
- An EVENT is a subset of the sample space.
- An IMPOSSIBLE EVENT is a an event which contains no outcomes.
- A CERTAIN EVENT is an event that contains all possible outcomes.

Example 2.2.1, SGT, p.4

Imagine I flip a coin, with two possible outcomes: heads (H) or tails (T). What is the sample space for this experiment? What about for three flips in a row? *Solution*: For the first experiment (flip a coin once), the sample space is just $\{$H;T$\}$. For the second experiment (flip a coin three times), the sample space is $\{$HHH;HHT;HTH;HTT;THH;THT;TTH;TTT$\}$

Example 2.2.3. Suppose I have two bowls, each containing 100 balls numbered 1 through 100. I pick a ball at random from each bowl and look at the numbers on them. How many elements are in the sample space for this experiment?

Solution: Using basic principles of counting (see the Sets and Counting tutorial), since the number of possible outcomes for the second experiment doesn't depend on the outcome of the first experiment, the total number of possible outcomes is 100^2, or 10,000.

Example 2.2.4. Which set of outcomes defines the event that the two balls add up to 200? Solution: There is only one outcome in this event, namely $\{(100,100)\}$

Simple Random Sampling (SRS)

- SRS is a technique for selecting a subset of entities from a given population.
- All entities in the population have an equal chance of being selected.
- SRS is a method of choice if the size of a population is unknown or prohibitively large to inspect in its entirety.
- Advantages of SRS:
    - A widely-used-best-practise method across scientific disciplines
    - Avoids data bias in selection
    - For an illustration see Kahn Academy: https://www.youtube.com/watch?v=acfjqWTwee0

Sampling With and Without Replacement

The Urn Model: Given a population of n elements $a_1, a_2, \ldots a_n$, any sequence $a_{j_1}, a_{j_2}, \ldots a_{j_n}$ is an *ordered sample of size r* drawn from the population. Let us assume that the elements of the ordered sample are chosen one-by-one. Then there are two scenarios:

- **Sampling with replacement:** Each selection is made from the entire population of n elements
- **Sampling without replacement**: Once an element is chosen from the population, it removed from the population.

How many distinct samples are there?

For a population of n elements and a sample size r, there are:

- n^r distinct samples with replacement.
- $n \times (n-1)\ldots(n-r+1)$ without replacement.

Transferring these concepts to NLP and CL

Example 1: Lexical Decision: Participants in an experiment are asked to decide by a button press whether a spoken or written word is an actual word for a given language. This is binary decision task with possible outcomes $\{$Y,N$\}$.

Example 2: Text Classification: Participants are asked to classify emails as spam or not spam. This is a binary decision task with possible outcomes $\{$S,N$\}$.

Example 3: Word Frequency: Counting the frequency of a word in different corpora of fixed or variable lengths.

Example 4: Text Generation: In a grossly simplified model, the production of a text can be viewed as a sequence of random samplings from a dictionary of finite size.

Example 5: Word Prediction: Participants are asked to predict the next word for a given sequence of words.

The Probability of an Event; SGT, p. 5

The probability of an event $P(E)$:

$$P(E) = \frac{|E|}{|S|} \quad \text{if all outcomes in the sample space } S \text{ are equally likely.} \tag{1}$$

The probability of the impossible event:

$$P(\emptyset) = 0. \tag{1.1}$$

The probability of the certain event:

$$P(S) = 1. \tag{1.2}$$

Example 2.3.2; SGT, p. 5

Suppose I have two bowls, each containing 100 balls numbered 1 through 100. I pick a ball uniformly at random from each bowl and look at the numbers on them. What is the probability that the numbers add up to 200?

Solution: There is exactly one outcome: (100,100), in a sample space of 100^2 outcomes altogether. Therefore, the probability of the outcome (100,100) is 1/10,000.

Example 2.3.3; SGT, p. 5

Let E be the event that the numbers on the balls in the previous example add up to exactly 51. What is the probability of E? Solution: We already know the size of the sample space, but we also need to determine the cardinality of E, i.e., the number of outcomes in this event.

There are altogether 50 distinct outcomes: (1,50), (2,49), . . . (49,2), (50,1). Therefore, the probability of E is 50/10,000, or 0.5%.

Disjoint Events and Partitions of the Sample Space; SGT, p. 6

- Two events E_1 and E_1 are MUTUALLY EXCLUSIVE (or DISJOINT) iff they have no outcomes in common, i.e. iff their intersection $E_1 \cap E_1 = \emptyset$
- A set of n mutually exclusive events E_1,..., E_n, whose union $E_1 \cup E_1 \cdots \cup E_n$ is equal to the sample space S is called a PARTITION of S.

Partition of the Sample Space

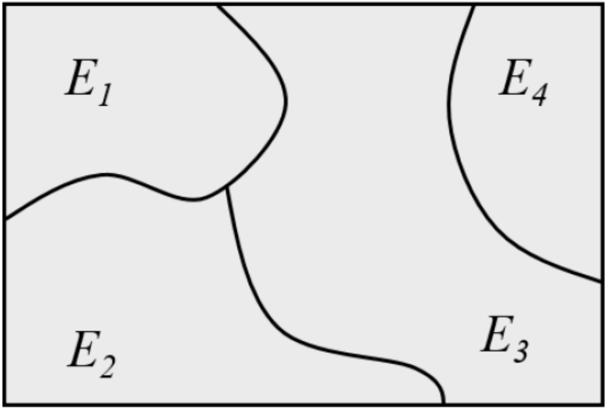


Figure: SGT tutorial p. 6

Probability and Probability Distribution; SGT, p. 6

Let $\{E_1 \ldots E_n\}$ be a partition of S. For each $E_i \in \{E_1 \ldots E_n\}$, $P(E_i)$ is a PROBABILITY and $\{P(E_1) \ldots P(E_n)\}$ is a PROBABILITY DISTRIBUTION iff the following properties hold:

$$\textbf{Property 1}: \quad 0 \leq P(E_i) \leq 1$$

$$\textbf{Property 2}: \quad \sum_{i=1}^{n} P(E_i) = 1$$

Probability Mass (Function); Uniform Distribution

- A probability is a function from an event to values between 0 and 1 (inclusive).
- A probability distribution is a function from partitions of events of a sample space S to sets of values between 0 and 1 (inclusive).
- A probability distribution assigns one unit of PROBABILITY MASS to the sample space S and distributes this probability mass in some fashion between all the E_i.
- A probability distribution is, therefore, sometimes also referred to as a PROBABILITY MASS FUNCTION.
- A UNIFORM distribution is a probability distribution where all events in a partition are equally likely, i.e. have the same amount of probability mass.

Example 2.4.2.: SGT p. 7

What is the distribution over the sum of two fair dice? Is it uniform or not?

Solution: We need to figure out the probability of each of the events. As noted above, there are 11 different events, corresponding to sums from 2 to 12. Also, note that we can define our sample space in terms of 36 equally likely outcomes, corresponding to the six possible outcomes on the first die multiplied by the six possible outcomes on the second die:

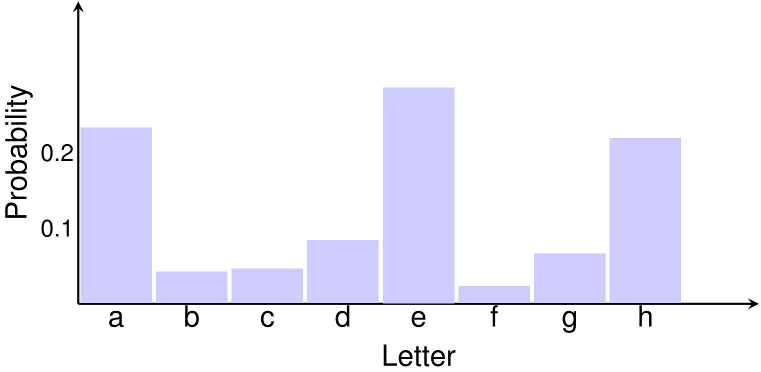
(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Example 2.4.2. continued

We can use Eq (1) to determine the probabilities of each of the events. For example, P(sum is 2)=1/36 because there is only one outcome in this event, whereas P(sum is 3)= 2/36 because there are two equally likely outcomes in this event.

A probability distribution over eight letters[2]

Lett.	a	b	c	d	e	f	g	h
Prob.	0.23	0.04	0.05	0.08	0.29	0.02	0.07	0.22



[2] slide due to Cagri Cöltekin

Event Complement; SGT, p. 10

$$P(\neg E) = 1 - P(E) \tag{2}$$

Example 3.1.1. Suppose I have a list of words, and I choose a word uniformly at random. If the probability of getting a word starting with **t** is 1/7, then what is the probability of getting a word that does not start with **t**?

Solution: Let E be the event that the word starts with **t.** Then $P(\neg E)$ is the probability we were asked for, and it is 1 - $P(E)$, or 6/7.

Event Union

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3}$$

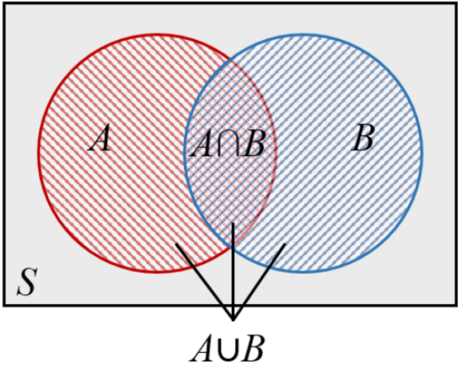


Figure: GW tutorial p. 11

Example 3.1.2; SGT p. 10

Name	Home country	Year
Andrew	UK	1
Sebastian	Germany	1
Wei	China	1
Fiona	UK	1
Lea	Germany	2
Ajitha	UK	1
Sarah	UK	2

What is the probability $P(E_1 \cup E_2)$, with E_1 = the student is female and E_2 = the student is from the UK?
Solution: The probability of $P(E_1)$ is $P(\frac{|E_1|}{|S|}) = \frac{4}{7}$. The probability of $P(E_2)$ is $P(\frac{|E_2|}{|S|}) = \frac{4}{7}$. However, $E_1 \cap E_2 = \{$Fiona, Ajitha, Sarah$\}$, and $P(E_1 \cap E_2) = \frac{3}{7}$. Hence:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \frac{4}{7} + \frac{4}{7} - \frac{3}{7} = \frac{5}{7}$$

Law of Total Probability (aka: Sum Rule)

Let $\{E_1 \ldots E_n\}$ be a partition of the sample space S and $B \subseteq S$.

$$P(B) = \sum_{i=1}^{n} P(B \cap E_i) \tag{4}$$

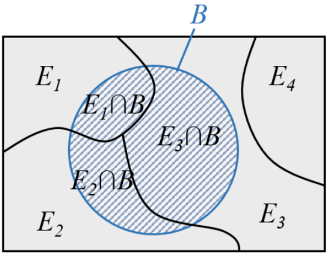


Figure: GW tutorial p. 12

Example 3.2.1; SGT, p. 11

Consider the scenario from Exercises 2.8 and 3.1.2. We partition the sample space according to the country that each student comes from, with E1 = "student is British", E2 = "student is Chinese", and E3 = "student is German". Also let B be the event that the student is female. Apply the law of total probability to compute P(B), and check that the result is the same as when computing P(B) directly.

Solution

$$\begin{aligned} P(B) &= P(B \cap E_1) + P(B \cap E_2) + P(B \cap E_3) \\ &= 3/7 + 0/7 + 1/7 = 4/7 \end{aligned}$$

This is the same result we get by computing P(B) directly (counting the number of female students and dividing by the total number of students).

Joint Probability Table (JPT)

It is often convenient to represent joint probabilities in a joint probability table. For example, consider a sample space of all students on a college campus. Let us assume that all students pay either in-state or out-of-state tuition and live either on-campus or off-campus. Then we can construct a JPT with each cell representing a joint probability.

	on-campus	off-campus
in-state	(in,on)	(in,off)
out-of-state	(out,on)	(out,off)

The rows and the columns refer to two different event classes that each form a partition of the entire event space. Each partition is called an *attribute* or a *dimension*.

JPT Example continued

Let us assume that probability of paying in-state and out-of state tuition is .55 and .45, respectively, and the probability of living on-campus and off-campus is .40 and .60, respectively. Then we can fill in the JPT as follows.

	on-campus	off-campus
P(in) = .55	P(in,on) = .22	P(in,off) = .33
P(out) = .45	P(out,on) = .18	P(out,off) = .27
	P(on) = .40	P(off) = .60

Then the probabilities to the left of each row and below each column are called marginal probabilities. Each marginal probability is the sum of all joint probabilities of the particular value for an attribute.