Statistical NLP: course notes

Çağrı Çöltekin — SfS / University of Tübingen

2020-05-11

These notes are prepared for the class *Statistical Natural Language Processing* taught in Seminar für Sprachwissenschaft, University of Tübingen.

This work is licensed under a Creative Commons "Attribution 3.0 Unported" license.

2 *Probability theory*

In this chapter, we will review some of the basic concepts from the probability theory. The concepts introduced in this chapter will be very important in many of the subjects we will cover during the course. Like the other background chapters, you may skip this chapter if you are familiar with the probability theory.

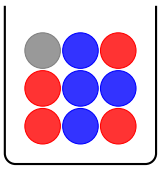

Figure 2.1: Topics in elementary probability are generally taught with examples from drawing one or more colorful balls from and urn. A simple experiment with the setup above is drawing a single ball from the urn. In this case, the possible outcomes are balls with one of three colors above. Probabilities are associated with set of outcomes. In the above setup, probability of drawing a red ball, $P(\bullet)$ is $4/9$ Similarly, $P(\bullet) = 4/9$ and $P(\bullet) = 1/9$. We can also assign probabilities to set of outcomes, for example, probability of drawing a red or blue ball, $P(\{\bullet, \bullet\}) = 8/9$, and probability of the sample space, $P(\{\bullet, \bullet, \bullet\}) = 1$. If our experiments involve drawing two balls with replacement (putting the ball we draw first back before drawing the other), then the outcomes are the all combinations of two ball colors. Examples probabilities for the outcomes of this experiment are $P(\{\bullet\bullet\}) = 16/81$, $P(\{\bullet\bullet\}) = 4/81$, $P(\{\bullet\bullet, \bullet\bullet\}) = 20/81$.

2.1 *Axioms of the probability*

Probability is a measure of (un)certainty of an event. In daily usage, we often associate some events with some probabilities. We talk about high- or low-probability events, or sometimes we express our notion of probability by percentages or odds.

Formally, we quantify the probability of an event with a number between 0 and 1. An event with probability of 0 is impossible, and an event with probability of 1 happens with certainty. Otherwise, any number in between, expresses the certainty we associate with the occurrence of the event. For example, an event with probability of 0.5 is as likely to happen as it may not.

The events we talk about are outcomes of trials (some sort of experiment or observation). In general, an event is a set of outcomes. The set of all possible outcomes of a trial is called its sample space, and conventionally indicated by the Greek letter omega (Ω). Figure 2.1 demonstrate these concepts with a common (probably familiar) example from typical lectures on elementary probability .

Formally, probabilities has to follow the following set of axioms.

1. $P(E) \in \mathbb{R}$, $P(E) \geqslant 0$. Probability of an event E has to be a positive number.

2. $P(\Omega) = 1$. The probability of all possible outcomes of a trial is one. Note that this also means that probability of no event can be larger than 1.

3. For two *disjoint* events E_1 and E_2, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$. In general, for N disjoint events. $P(E_1 \cup \ldots \cup E_N = \sum_{i=1}^{N} P(E_i)$.

The above three axioms forms the basis of all probabilistic statements. The other rules that we will discuss later on can all be derived from these three rules.

These axioms, hence the probability theory, defines *how* to use or manipulate the probabilities. It does not specify *what* a probabil-

ity is, and how to assign probabilities to events. There are different schools of thought that assign probabilities to events using (somewhat) incompatible ways. Since uncertainty, hence probability, plays an important role in any statistical study, we will encounter some of these differences during the course of this course.

2.2 *Random variables*

The concept of *random variable* is central to the probability theory. The value of a random variable is subject to uncertainties.[1] The value of a random variable depend on the outcome of a set of random events. For example, we are often interested in studying outcomes of experiments or observations that involve uncertainty such as

- height or weight of a person selected (randomly) form a population
- length of a randomly chosen document
- whether an email is spam or not
- the first word of a book, or the first word uttered by a baby

[1] Although, the value of a *random variable* is not necessarily unpredictable, as in the typical daily use of the term 'random'. Most random variables we study involve some uncertainties, but also some regularities or structure that we can exploit for making inferences. Somewhat surprisingly, even some aspects of a 'truly random' processes include these tendencies, which is very important for machine learning, and *statistical inference* in general.

In this context, we can think of a random variable as a (well-defined) function from the set of outcomes (Ω) of an experiment to real numbers. Clearly, the outcome of some of these trials are not trivial to express as a real number. Often, for mathematical convenience, an arbitrary mapping may be used, e.g., from a set of words to consecutive integers. However, some of the useful quantities (e.g., expected value or variance) does only make sense for random events for which a meaningful mapping to the (real) numbers exists. For example, we can define random variables whose values are words or syntactic representations of sentences (trees), but, in these cases quantities like mean will not be useful (What is the average word?).[2] Note that the values of random variables are not probabilities. Probabilities are associated with the outcomes of the random experiment.

[2] Technically, a random variable takes real numbers. A more correct term used for non-numeric random variables is *random element*. We will be using the term random variable for both cases.

Mapping some outcomes to numbers is easy, for example, if we are measuring the frequency of a sound signal, or counting the number of words in a document, the measurement already gives us a real value. The mapping may not be trivial in some other cases, such as whether the random variable representing a product review is positive or negative, or the part of speech of a word. In these cases, there are often conventional methods for mapping these outcomes to numbers. For example, the Boolean variables (such as binary outcome of a review) are typically mapped to 0 and 1 for negative and positive outcomes respectively.

For trials having outcomes with more than two categories (as in our part of speech example), a possible solution is to map them to integers arbitrarily, for example, as on the row labeled 'Integer' in Table 2.1. However, this mapping implies an ordering, even though the assignments are arbitrary. A more convenient method is to map

Part of speech	Noun	Verb	Adjective	Adverb	...
Integer	1	2	3	4	...
Binary vector	000001	000010	000100	001000	...

Table 2.1: Alternative numeric representations for categorical (nominal) variables.

each outcome to a vector of binary (0 or 1) values, where we set only a particular member of the vector to 1, and we set the rest to 0 for a particular outcome. This representation is called *one-of-k* or *one-hot* representation, and exemplified on the row labeled 'Binary vector' in Table 2.1. In statistical literature it is also called *dummy coding*.[3] This particular representation is used quite often in machine learning while encoding categorical features or outcomes. For example whether a particular word (out of all words in a dictionary) occurs in a document or not.[4]

In the discussion above, we implicitly made a distinction between two types of random variables. Some random variables are continuous, they can take any real number as values (such as height or frequency). Others, on the other hand, have discrete values. For example, number of words in a document or sentence can only take integer values, and whether a document is spam or not can have only two values (true or false). This distinction between the continuous and categorical random variables is important, in general, and in our discussion of some important properties and specific distributions of random variables below.

To make the discussion more concrete, we will use an example hypothetical random variable. Assume that predicting the length of utterances in a (particular type of) dialog is important for our purposes. So, our random variable, X, is a discrete variable, taking only integer values (length of utterances). For simplicity, we will also assume that the longest possible utterance is 11 words.

Table 2.2 lists each possible value of X, with its associated probability. We denote the random variable with uppercase X, while lowercase x stands for a possible value of the variable. The first row of Table 2.2 shows the $P(X = 1)$, whose probability turns out to be 0.155. For now, we do not ask where these probabilities come from.[5] However, it is important to note that the probabilities in the table sum to 1 (except for possible rounding error).

In our utterance-length example (in Table 2.2), there is a natural mapping between the utterance length and the value of the random variable. The value of the random variable is simply the integer that corresponds to the utterance length in words. This mapping is not always straightforward. To demonstrate that we give another hypothetical example in Table 2.3.

We assume that we are working on an ancient language, which is written only with eight letters. The random variable X is a mapping between the event (observing a particular the letter) and the value of the random variable (x). This mapping is more arbitrary than the example with sentence lengths above. Here, we use integers, but one-hot encoding would be more convenient in real-world applications.

[3] Note that the coding as we describe here involves some redundancy. We can be slightly more economical by mapping one of the outcomes to all zeros. Although this economy is not really useful for many purposes, it is sometimes used (especially in statistics) when one of the outcomes is a special base case (coded as all zeros). This makes comparison of other outcomes to the base case easier.

[4] If the outcomes/items (such as words) represented have some features in common, we can even use real values (rather than binary one-hot) in these vectors. We will later discuss such 'dense' representations.

Table 2.2: Probabilities of all possible values (x) of an example random variable X (utterance lengths).

x	$P(X = x)$
1	0.155
2	0.185
3	0.210
4	0.199
5	0.102
6	0.066
7	0.039
8	0.023
9	0.012
10	0.005
11	0.004

[5] They are *estimated* from a real corpus of spoken language, but we will delay the details for now.

Table 2.3: Probability distribution of variable X, associated with letter probabilities. Note that the mapping between the letters and the values of the random variable x is arbitrary.

letter	x	$P(X = x)$
a	1	0.233
b	2	0.042
c	3	0.046
d	4	0.084
e	5	0.286
f	6	0.026
g	7	0.063
h	8	0.219

In both examples, we specified probability of a particular value of a random variable like $P(X = x)$, for example $P(X = 1)$ in Table 2.3 is 0.23. However, when there is no ambiguity, we will skip the name of the random variable and write, for example, $P(1) = 0.23$, or even $P(a) = 0.23$.

2.2.1 *Probability distributions*

A *probability distribution* provides a mapping between values of a random variable (or the corresponding outcomes of a random trial) to probabilities. If the random experiment at hand can only have finite number of outcomes, a possible way to define the probability distribution is a vector or table of probabilities (as in Table 2.3, for example). In many cases, however, we use more compact function to specify the probability distribution. In the following we will revisit a few common ways of specifying probability distributions.

2.2.2 *Probability mass function*

Probability mass function (PMF) is a function that maps the values of a discrete random variable to their probabilities. The PMF of a random variable maps all possible values of the random variable to exact probability of the associated event. For example, probability of an utterance of length 3 is 0.21 in our example (Table 2.2). A probability mass function defines a discrete probability distribution. Table 2.2 defines such a PMF, which is shown graphically in Figure 2.2.

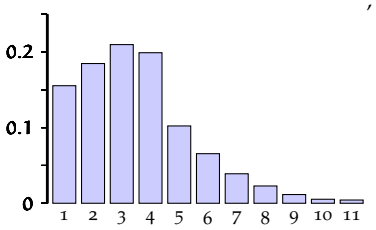

Figure 2.2: Graphical representation of probability mass function defined in Table 2.2.

2.2.3 *Probability density function*

Continuous random variables do not have probability mass functions, but analogously they can be defined through a *probability distribution function* (PDF). For continuous distributions, the probability of a single value of the random variable is zero. We can only talk about non-zero probabilities of intervals. This may be unintuitive at first sight. However, if you consider the fact that there are infinite number of real numbers between any arbitrary range $[a, b]$ ($a \neq b$), the probability of any single value is statistically equivalent to zero.

As an example of a continuous random variable, suppose we were measuring the durations of the utterances rather than the number of words as in our discrete random variable example, we cannot assign a non-zero probability to an utterance being 1.40 s. However, we can assign a probability to a range, say between 1.20 s and 1.60 s. As a result, the values of a PDF are not probabilities. Figure 2.3 shows a probability density function. For the sake of demonstration, we will pretend this to be the distribution of durations of utterances in a spoken language corpus.[6] Figure 2.3 clearly shows that the values of the PDF are not probabilities. For example, the value corresponding to $x = a$ is grater than 1. Another important property of the PDF, more difficult measure from the figure, is that the area under the curve is 1 unit. Although probabilities of individual real numbers

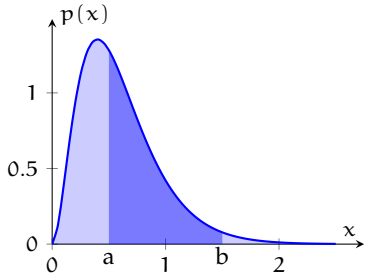

Figure 2.3: Example probability density function (PDF). Area under the whole curve, potentially stretching to infinity, is 1 units. The area under a particular interval is the probability that the random variable takes a value within the interval.

[6] The plot actually shows a well-known density function. Precisely, this is the probability density function of the gamma distribution with parameters $k = 3$, $\theta = 0.2$, but this is not important for this introduction.

are zero, we can calculate the probabilities of ranges, like the one highlighted in Figure 2.3. A PDF defines a continuous probability distribution similar to a PMF defines a discrete distribution. [7] The probability is simply the definite integral of the PDF in the interval we are interested in.

$$P(a \leqslant x \leqslant b) = \int_a^b p(x)dx$$

We will give brief descriptions of some of the important continuous probability distributions later in this chapter.

2.2.4 *Cumulative distribution function*

Cumulative distribution function (CDF), or *distribution function* of a random variable X, $F_X(x)$, yields the probability that X will take a value less than or equal to x.

$$F_X(x) = P(X \leqslant x)$$

In case of discrete distributions, it is simply the sum of probabilities of all values up to and including x. And, for continuous distributions, it is the area of under the probability density function in interval $[-\infty, x]$. As may you have already guessed, to evaluate the value of the CDF of a continuous random variable, we need to integrate the PDF in this range instead of summing. Output of a CDF is a probability regardless of whether the random variable is continuous or discrete. The cumulative distribution function plays an important role in statistics, particularly in hypothesis testing.

Table 2.4 repeats the probability mass function from Table 2.2 (in column 2), and also shows the values cumulative distribution function. Note that the CDF converges to 1.0 at the maximum value of the random variable X.

2.2.5 *Expected value*

Expected value of a random variable is its arithmetic mean (μ). Given a list of numbers (a sample), their arithmetic mean is simply their sum divided by the total number of numbers in the list. Assume that we have the following numbers, say as the length of utterances of interest in a (small) spoken language corpus:

$$1, 2, 3, 3, 3, 4, 4, 5, 7, 11$$

To find the arithmetic mean, we simply sum these numbers up, and divide the number of items in our list. Which gives us 4.3.

Rather than a fixed sample as in the example above, we often need to calculate the mean of a probability distribution, defined as a set of numbers with their probabilities, we calculate the average by weighting each value with its probability. The mean, or the expected value, of the probability distribution in Table 2.2 is

$$0.155 \times 1 + 0.185 \times 2 + 0.210 \times 3 + 0.199 \times 4 + 0.102 \times 5 + 0.066 \times 6 + 0.039 \times 7 + 0.023 \times 8 + 0.012 \times 9 + 0.005 \times 10 + 0.004 \times 11 = 3.516\,.$$

[7] A common convention for distinguishing, a PMF from a PDF is using uppercase P() for the former and lowercase p() for tha latter. $P(X = x)$ yields the probability of discrete random variable being x, but the values returned by p() are not probabilities, although higher values will indicate neighborhoods of high probability density.

Table 2.4: The values of *probability mass function* (column 2) and *cumulative distribution function* (column 3) of the earlier example random variable, X, presented in Table 2.2.

x	$P(X = x)$	$P(X \leqslant x)$
1	0.155	0.155
2	0.185	0.340
3	0.210	0.550
4	0.199	0.749
5	0.102	0.851
6	0.066	0.917
7	0.039	0.956
8	0.023	0.979
9	0.012	0.990
10	0.005	0.996
11	0.004	1.000

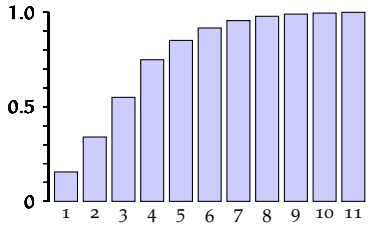

Figure 2.4: Graphical representation of cumulative distribution function defined in Table 2.4.

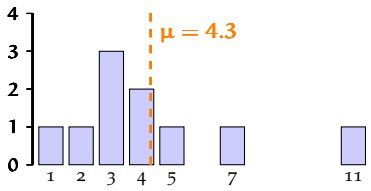

Figure 2.5: A graphical representation (histogram) of the sample (1, 2, 3, 3, 3, 4, 4, 5, 7, 11). Dashed orange line marks the mean.

(2.1)

Note that the expected value is not (necessarily) the most likely value. In our example, it may even be a value that the random variable cannot take.[8] Nevertheless, it is an important quantity indicating the central tendency of a probability distribution.

[8] The mean, however, is the value that minimizes the prediction error.

The expected value of a discrete random variable is

$$E[X] = \sum_x P(x)x. \tag{2.2}$$

where x ranges over all values of X, and $P(x)$ is a shorthand for $P(X = x)$, the probability of a random variable X taking the value x as defined by its probability mass function. In plain words, we multiply each value with its probability, and sum them up.

In general, the expected value of a function of a random variable $E[f(X)]$ can be calculated using,

$$E[f(X)] = \sum_x P(x)f(x). \tag{2.3}$$

In Equation 2.2 above, we simply used the identity function $f(x) = x$.

For continuous variables, we need an infinite sum, so, we integrate rather than sum:

$$E[f(X)] = \int_{-\infty}^{\infty} p(x)f(x)dx$$

Note that, here, $p(x)$ is a probability density function.

2.2.6 *Median and mode*

Like the expected value, *median* and *mode* are two other quantities (statistics) that are useful in characterizing the central tendency of probability distributions.

The *median* of a random variable is the value that splits the probability mass (or density) into two equal parts. More formally, for a random variable X, the median is defined as a number m that satisfies

$$P(X \leqslant m) \geqslant \frac{1}{2} \quad \text{and} \quad P(X \geqslant m) \geqslant \frac{1}{2}. \tag{2.4}$$

Going back to our finite sample of numbers, [1, 2, 3, 3, 3, 4, 4, 5, 7, 11], the easiest way to find the median is to find the number that splits the ordered list into two equal halfs. For odd number of items, this is the items in the middle. For even number of items, as in our case, median is conventionally defined as the mean of the two middle numbers, which is 3.5 for our example.[9]

[9] Technically any number that satisfies inequalities in (2.4) is a median, which is any number between 3 and 4 (not including). In general, m is not guaranteed to be unique, but for the sake of simplicity, we will continue talking about 'the median' of a distribution.

Given a probability distribution, the median is found simply by solving the inequalities above for m. For example, for a continuous probability distribution, that would mean solving

$$\int_{-\infty}^{m} p(x)f(x)dx \leqslant \frac{1}{2}$$

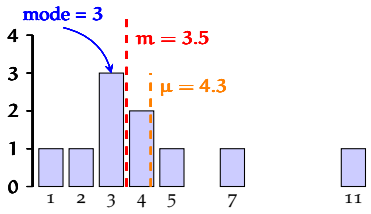

Figure 2.6: The histogram in Figure 2.5 repeated. This time we also mark median (m, red dashed line) and the mode, and the mode.

for m.

The *mode* of a finite sample is the value that occurs most often. In our example data above, the most frequent value is 3, so it is the mode (see Figure 2.6). The mode of a probability distribution is the value(s) that correspond to maxima of probability mass or density functions. The examples we had so far are *unimodal*, they have only one mode. However, some distributions can be *bimodal*, or in general, *multimodal*. A probability distributions is called multimodal, if there are multiple modes (peaks with possibly different heights), as in Figure 2.7.

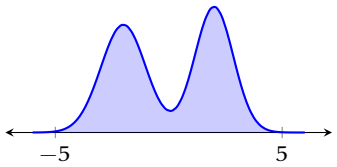

Figure 2.7: An example multimodal (bimodal) continuous probability distribution.

Multimodal distributions are interesting, as they often indicate a *confounding* variable. For example, distribution of heights or weights of university students are probably bimodal, since the differences in gender will result in having two peaks around the means of males and females.

The *mean*, *median* and *mode* are measures of central tendency. The mean is most commonly used measure of central tendency in many tasks in statistics and machine learning, since it has nice algebraic properties. However, as you can see in Figure 2.6, mean is affected from extreme values. Since the distribution in the figure is skewed, and there are extreme values (i.e., the effect of 11 in the figure is much higher than the other data points closer to the center), mean is moved from the center of the distribution towards these extreme values. The median is not affected by extreme values, but it does not have the same nice algebraic properties. The median is often used in statistics as a robust measure (a measure that is not affected by outliers) of central tendency. The mode is easy to interpret, but it also lacks the nice algebraic properties of mean. Furthermore, mode is only determined by the maxima, it is not affected by the other values in the distribution at all. For symmetric unimodal distributions (such as Gaussian distribution that we will briefly introduce below), the mean, median and mode are the same.

2.2.7 *Variance and standard deviation*

The mean, median and mode are measures of central tendency of a distribution. They all are very useful as single-number summaries. However, as any summary, they tell only part of the story. Another aspect of a probability distribution is its spread. Standard deviation and variance are the two (related) measures of spread.

Figure 2.8 presents two distributions with equal center, but different spread. Variance is one of the measures that quantify this difference. It is defined as

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2] \,.$$

In plain words, variance is the mean of the squared differences of the values of the random variable from its mean. It can be shown trivially that this formula is identical to $\mathrm{E}[X^2] - (\mathrm{E}[X])^2$.[10] Variance is often easier to manipulate algebraically. However for quick and

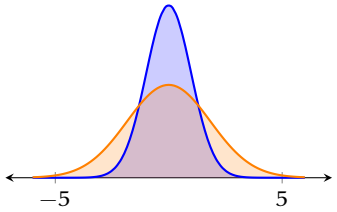

Figure 2.8: Two probability distributions with the same *mean*, *median* and *mode*, but different *variances*. More precisely, both distributions are Gaussian distributions with $\mu = 0$, and $\sigma = 0.7$ (narrow, blue) and $\sigma = 1.3$ (wide, orange).

[10] Both formulations result in loss of precision if implemented as is in a computer program. There are alternative formulations that are numerically (more) stable.

easy interpretation, its square root, *standard deviation*, is more useful, since it is in the same units as the random variable itself. Standard deviation is often denoted using Greek letter sigma (σ), then variance is naturally denoted by σ^2.

For a finite sample of size n, or a set of outcomes with equal probabilities, we can calculate the variance using the formula,

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

where μ is the arithmetic mean, or the expected value. Getting back to our example sample [1, 2, 3, 3, 3, 4, 4, 5, 7, 11], we calculate the variance by subtracting the mean (4.3 as calculated Section 2.2.5) from each number, squaring them, summing them up, and dividing the number of items in the sample, which is 7.41.[11] The standard deviation is then $\sqrt{7.41} = 2.72$.

[11] You are encouraged to perform this calculation.

To calculate the variance of a discrete distribution (rather than a sample) from its PMF, remember that expected value of any function of a random variable can be computed using Equation 2.3. Since the squared difference from mean is a function of the random variable, the expected value of it, the variance, is

$$\sum_{i=1}^{n} P(x_i)(x_i - \mu)^2 \tag{2.5}$$

For our utterance-length example, this leads to[12]

$$\begin{aligned}
&0.155 \times (1 - 3.516)^2 + 0.185 \times (2 - 3.516)^2 + 0.210 \times (3 - 3.516)^2 + \\
&0.199 \times (4 - 3.516)^2 + 0.102 \times (5 - 3.516)^2 + 0.066 \times (6 - 3.516)^2 + \\
&0.039 \times (7 - 3.516)^2 + 0.023 \times (8 - 3.516)^2 + 0.012 \times (9 - 3.516)^2 + \\
&\qquad 0.005 \times (10 - 3.516)^2 + 0.004 \times (11 - 3.516)^2 = 3.886
\end{aligned}$$

[12] If you perform the calculations as shown, you will find a different value (3.872) rather than the value displayed. This has to do with the fact that the numbers displayed here are rounded, calculating the variance using Equation 2.5 is *numerically unstable*. This is also a problem in real use since computers also represent real numbers with a limited precision. In practice other, numerically stable, formulations (algorithms) are used.

The standard deviation is, then, $\sqrt{3.886} = 1.971$. Note that, unlike the variance, the standard deviation is in the same units with the data. So, we can say that the standard deviation is 2.016 words.

For continuous distributions, as you should already be expecting, we replace the sum in Equation 2.5 with integral. The other properties of variance and its interpretation do not change.

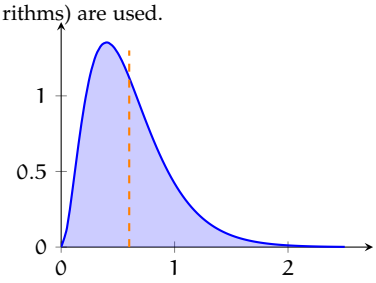

Figure 2.9: A (positively) skewed probability density function. This is the same distribution from Figure 2.3. In addition, we mark the location of the mean (vertical dashed line, 0.60). Note that larger part of the area under the curve falls to the left of the mean. The distribution has a longer right tail (hence, the expected value is 'pulled towards' the extreme values on the tail, compared to mode and median.).

2.2.8 *Symmetry and skewness*

Besides the spread (measured by variance or standard deviation), there are other important properties of probability distributions. Here, we will informally note another property, *skewness*, that we will sometimes make use of. We have already used some terms like 'symmetric' or 'asymmetric' distribution. A *symmetric probability distribution* has the same amount of probability mass on both sides of its mean. On the other hand, an *asymmetric probability distribution* is skewed, it has more probability mass on the right or left side of its mean. The ends of the probability distributions where there is little probability mass are called the *tails* of a distribution. A *positively skewed* distribution has a longer right tail, while a *negatively skewed* distribution has a longer left tail.

2.3 *Some well-known probability distributions*

Some natural processes generate quantities that follow certain well-known probability distributions. In this section we will introduce some of these well-known probability distributions. Probably, the most well-known probability distribution is the *Gaussian*, or the *normal*, distribution with the bell-shaped density function. These distributions can be specified by a set of parameters. For example, the normal distribution is generally parametrized by its mean μ, and the standard deviation σ (or equivalently its variance σ^2). A common notation to indicate that a random variable follows a known probability distribution is

$$X \sim \text{Normal}(\mu, \sigma^2) \quad \text{or} \quad X \sim \mathcal{N}(\mu, \sigma^2).$$

where X is the random variable, and 'Normal' or $\mathcal{N}$ is the conventional name or abbreviation for the distribution. Most distributions have alternative parametrizations that may be convenient in different applications. Usually, however, one of the parametrizations is considered as being canonical or standard. In case of the normal distribution, the mean (μ) and variance (σ^2) are the standard parameters. However, the normal distribution is sometimes parametrized by its mean and precision (inverse of variance, $\tau = 1/\sigma^2$).

2.3.1 *Uniform distribution*

There are both continuous and discrete flavors of the uniform distribution. The discrete uniform distribution assigns equal probabilities to all values in an interval $[a, b]$.[13] The canonical parameters of the uniform distribution are the end points of the range a and b. The mean and median are $\frac{a+b}{2}$, and the variance is $\sigma_2 = \frac{(b-a+1)^2-1}{12}$.

Continuous uniform distribution is similar to the discrete one. It is also parametrized by the extreme values in the interval, a and b. It's mean and median are $\frac{a+b}{2}$. The variance formula is slightly different ($\frac{(b-a)^2}{12}$).[14]

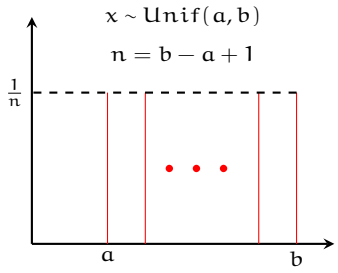

Figure 2.10: Probability mass function of the discrete uniform distribution.

[13] Remember that the interval $[a, b]$ (with square brackets) indicates an inclusive interval on both ends. Hence, the total number of integers in this interval is $b - a + 1$.

[14] For those who like a bit of math, it is a nice exercise to try to derive the variance expressions for both continuous and discrete uniform distributions. The important bit you need to remember is variance is $E[(X - E[X])^2]$. For discrete uniform distribution, you will also need to know how to calculate the sum of the squares of consecutive integers. For the continuous uniform distribution, you will need to integrate polynomials.

2.3.2 *Bernoulli distribution*

A *Bernoulli trial* is a simple random experiment with two outcomes. The most common text-book example of a Bernoulli trial is a coin flip. It can yield either heads (H) or tails (T). As a more practical example, we can also view spam detection as a Bernoulli trial: a given email is either spam or not. So, any email you receive is a Bernoulli trial. Or similarly, if we want to predict the gender of the author of a document (female or male). Yet another example: whether the output of a machine learning system for a single test instance is correct or incorrect.

The Bernoulli distribution characterizes the outcomes of Bernoulli trials. We map one of the outcomes to 1 and the other outcome to 0.[15] Then, the random variable X distributed according to Bernoulli

[15] The outcome mapped to 1 is sometimes called 'success', or the 'positive outcome'. The outcome mapped to 0 is sometimes called 'failure', or the 'negative' outcome. However, the assignment is generally arbitrary, or by conventions that does not always reflect the semantics of the words (e.g., a medical diagnostic indicating an illness may be said to have returned a 'positive result').

distribution has a single parameter p such that $P(X = 1) = p$ and $P(X = 0) = 1 - p$. For specifying the Bernoulli PMF, often an alternative notation is used for convenience:

$$P(X = k) = p^k(1-p)^{1-k}$$

where k is either 0 or 1. Note that when $k = 0$, the first term become 1 ($p^0 = 1$ for any p) yielding $1 - p$, and for $k = 1$ the second term is 1 and the result is p.

For example, in a coin toss trial with a fair coin, $p = 0.5$, both outcomes are equally likely. In my email account, announced publicly, the probability of a new email being spam is probably well above 0.5 (but fortunately spam filters seem to do a fine job, thanks to probability theory).

The expected value of a Bernoulli-distributed random variable is p, and the variance is $p(1-p)$. Note that the variance is highest when $p = 0.5$.

As you probably guessed from the examples above already, Bernoulli distribution has a wide range of applications despite its simplicity.

2.3.3 *Binomial distribution*

The *binomial distribution* is a generalization of the Bernoulli distribution to n trials. The value of the random variable is the number of 'successes' in the experiment. The binomial distribution has two parameters, p and n. Similar to the Bernoulli distribution the parameter p is the probability of the success in a single trial, and n is the number of trials. Given these parameters, [16]

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$
$$\mu = np$$
$$\sigma^2 = np(1-p)\,.$$

Note that the notation $\binom{n}{k}$ used in the definition of the probability mass function is the binomial coefficient. The rest of the PMF function is simply the probability of k successes in n independent Bernoulli trials with parameter p. Note also that like the Bernoulli distribution, the variance is highest at $p = 0.5$, and decreases as p gets closer to 0 or 1.

A typical example of the binomial distribution is n consecutive coin tosses (with a biased or fair coin). But we will, for example, use it for the number of correctly parsed sentences by a parser (out of all sentences in a test set).

[16] The binomial coefficient (also read as 'n choose k') is defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

It yields the number k successes (without ordering) in n trials. It has an important quantity for many combinatorial problems, among others. For example, in an experiment involving four independent coin tosses, there will be $2^4 = 16$ possible outcomes. If we are interested in all outcomes with 3 heads (arbitrarily labeled as 'success'),

$$\binom{4}{3} = \frac{4!}{3!1!} = \frac{1 \times 2 \times 3 \times 4}{1 \times 2 \times 3 \times 1} = 4\,.$$

You can verify this by exhaustively listing all outcomes of four coin tosses, where the only configurations with three heads are THHH, HTHH, HHTH, HHHT.

2.3.4 *Categorical distribution*

The categorical distribution is similar to Bernoulli distribution, but instead of binary outcomes, it characterizes experiments with k mutually exclusive outcomes. A typical example is the outcome of a dice roll. Categorical distribution is parametrized by the parameter k, the

number of mutually exclusive outcomes, and a vector $\mathbf{p} \in \mathbb{R}^k$ whose elements, $p_1, \ldots, p_k$, indicate the probability of the corresponding outcome.[17] Since the events we model are exhaustive and mutually exclusive, they form the sample space, and their probabilities sum to 1.

[17] Note that there are only $k - 1$ independent p parameters, since they have to sum to 1.

It should be already clear that the Bernoulli distribution is a special case of the categorical distribution, where $k = 2$. However, unlike for Bernoulli distribution, starting the index from 1 rather than 0 is more convenient. Note that the assignment of outcomes to the integer values is often arbitrary.

Given a random variable X is distributed with categorical distribution with parameters k and $p_1, \ldots, p_k$,

$$\begin{aligned} P(X = x_i) &= p_i \\ E[X = x_i] &= p_i \\ Var(X = x_i) &= p_i(1 - p_i) \end{aligned}$$

Alternatively, we can write the probability mass function as

$$P(X = x) = \prod_{i=1}^{k} p^{[i=x]_i}$$

where the notation $[i = x]$ is 1 if $i = x$ is true, 0 otherwise.

In machine learning, instead of integers in the interval $[1, k]$, a common practice is to use k-valued 'one-of-k' vectors as described on page 17. In this notation, we can write the PMF as

$$P(X = \mathbf{x}) = \prod_{i=1}^{k} p_i^{x_i}$$

where $\mathbf{x}$ is one-of-k vectors. Since only one of the x_i will be non-zero for a given $\mathbf{x}$, this is another convenient way to write the PMF of the categorical distribution. This notation also works well with the multinomial distribution we will discuss next.

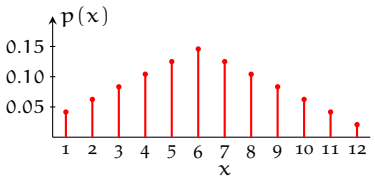

Figure 2.11: An example categorical PMF: probabilities of obtaining values 1 through 12 in a roll of two fair dice as either on the face of a single die, or as the sum of both. This distribution has to be (implicitly) internalized by any good backgammon player.

2.3.5 *Multinomial distribution*

The *multinomial distribution* arise when a k-way event is repeated n times. It is a generalization of the categorical distribution, where a categorical distribution is a multinomial distribution with $n = 1$. The relation between categorical distribution and the multinomial distribution is the same as the relationship between the binomial distribution and the Bernoulli distribution.[18] Sometimes the distinction between the categorical and multinomial distributions are blurred, but not paying attention the differences may lead to incorrect results in some cases. It is also a generalization of the binomial distribution with $k = 2$. Bernoulli distribution is also a special case, where $k = 2$ and $n = 1$.

[18] Based on this analogy, some authors call categorical distribution 'multinoulli' distribution.

The multinomial distribution is an important distribution for machine learning, and particularly for natural language processing. The outcome of a k-way multinomial event is conveniently expressed,

with a k-valued vector, where elements correspond to the counts of corresponding outcomes. For example, we may be interested in the distributions of part-of-speech tags in a document. For simplicity, say we are only interested in the distribution of nouns, verbs and adjectives. We can map these categories to consecutive integers (arbitrarily), 1, 2, 3, and the 'other' category to 4. On a 100-word document a possible outcome is $(33, 15, 9, 43)$, which means there were 33 nouns, 15 verbs, 9 adjectives and 43 other POS categories. This is a direct extension of the one-of-k representation we used for the outcome of a categorical random variable. We also carry over the parameter vector $\mathbf{p}$ which gives the probabilities of corresponding categories. The probability mass function then can be written as

$$P(X = \mathbf{x}) = \frac{n!}{x_1! \dots x_k!} \prod_{i=1}^{k} p_i^{x_i}.$$

Note that the first product in right hand side makes sure that the order of events does not matter.

To make things more concrete, let's return to our example with POS tags, where our sample was $\mathbf{x} = (33, 15, 9, 43)$ and assume that we are interested whether this document belongs to an author who is known to write documents with a distribution $\mathbf{p} = (0.3, 0.2, 0.1, 0.4)$. Now we can try to find the probability of such a document coming from this probability distribution by placing all into the PMF formula.[19]

[19] A warnings is (again) in order: while doing calculations with too large and too small numbers as in here, numerical instabilities (underflows or overflows) may occur.

$$P(X = (33, 15, 9, 43)) = \frac{100!}{33!15!9!43!} 0.3^{33} 0.2^{15} 0.1^{9} 0.4^{43} = 0.0005283718$$

We will delay the interpretation of this probability, but you should ask yourself what that probability is exactly, and how to interpret it. What have we found out? All the simplifications aside, does this document belong to the author? Is this probability large (a strong indication) or not?

The expected number of times a particular outcome, x_i, is observed in n trials is $E[x_i] = np_i$. The variance (of a particular outcome) is also similar to binomial distribution, $Var(x_i) = np_i(1 - p_i)$.

2.3.6 *Beta distribution*

The distributions we have discussed so far has been all discrete probability distributions. We will briefly introduce a few important continuous distributions as well, starting from a simple distribution.

The *Beta distribution* is a continuous distribution with a support on a bounded interval between 0 and 1. This means the probability distribution function is defined on this interval, which makes the Beta distribution ideal for distributions of probability values. This may sound too abstract at first, but think about having a machine that bends coins some random amount, resulting in bent coins with varying probability, p, of turning up heads in a coin toss trial. We first put our coins through the machine, and then perform the coin flip. Coin flip part is modeled properly by a Bernoulli distribution,

and what our coin bending machine does can be modeled nicely by the Beta distribution. This sort of modeling decisions often arise in *Bayesian* statistics, where the Beta distribution is said to be the *conjugate prior* of the Bernoulli and binomial distributions because of the convenience of using these two distributions together.

The Beta distribution is parametrized by two positive real numbers, α and β. The probability *density* function of the Beta distribution is

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}.$$

Fully understanding this equation is not essential for our purposes. However, you should note the similarity between the numerator of the right hand side and the probability mass function of the Bernoulli distribution. The gamma function ($\Gamma()$) in the denominator is a generalization of *factorial* function to real numbers (for positive integers $n! = \Gamma(n+1)$). Note that $p(x)$ is a probability density function. Unlike probability mass functions of the discrete probability distributions we were discussing so far, it does not return probabilities.

The mean and variance of the Beta distribution is,

$$E[X] = \frac{\alpha}{\alpha+\beta}$$
$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Again, the details are not essential for our purposes, but there are a few properties of the distribution to note. If α and β are equal, the mean will be 0.5, and the distribution will be symmetric. Figure 2.12 presents example PDFs with equal α and β. If α is larger than β, higher probability mass will be reserved for the values over 0.5. Otherwise, the probability mass will be shifted to the left. Figure 2.13 presents a few example Beta PDF functions with different α and β values.

Returning to our coin-bending machine example that produces bent coins according to various Beta distributions in Figure 2.12 and 2.13, we would have different expected p values for the coin tosses. For example, if α is larger than β, e.g., the blue distribution in Figure 2.13, the machine would produce coins that yield heads (arbitrary 'success' or 'positive' category) most of the time. Otherwise, e.g., the orange distribution, the machine would produce coins that produce tails most of the time. For the symmetric distributions, $\alpha = \beta > 1$ configuration produces coins with that are close to a fair coin, with increasing fairness with increasing α and β. When $\alpha = \beta < 1$, each coin is likely to be biased towards heads or tails, but with equal probability density shared by both types of unfair coins.

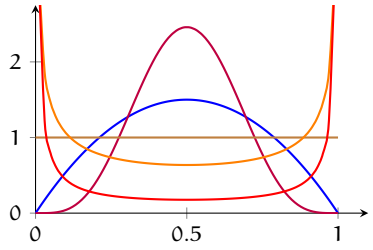

Figure 2.12: Beta distributions with equal α and β parameters. 2 (blue), 5 (purple), 1 (brown), 0.5 (orange), 0.10 (red). All distributions are symmetric with $\mu = 0.5$. The parameters $\alpha = \beta = 1$ leads to the continuous uniform distribution in $[0, 1]$. As α and β increase, the variance of the distribution decreases. You should also note that the distribution becomes similar to the Gaussian distribution for large values of α and β.

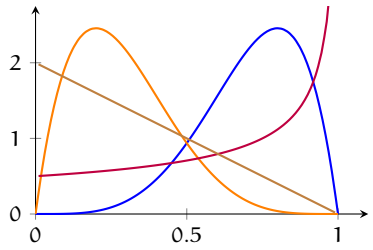

Figure 2.13: Example asymmetric beta PDFs: $\alpha = 5$, $\beta = 2$ (blue), $\alpha = 2$, $\beta = 5$ (orange), $\alpha = 1$, $\beta = 0.5$ (purple), $\alpha = 1$, $\beta = 2$ (brown).

2.3.7 *Dirichlet distribution*

Dirichlet distribution is a generalization of the Beta distribution. Like the Beta distribution, the support of Dirichlet distribution is over real

numbers in interval $[0, 1]$, but instead of parameters α and β, we have a k-dimensional vector $\boldsymbol{\alpha}$. As a result the Dirichlet distribution can express a distribution over k probabilities, making it ideal to serve as *priors* to categorical and multinomial distributions which are parameterized by a vector of probabilities. The PDF of the Dirichlet distribution is

$$p(x) = \frac{\prod_{i=1}^{k} x^{\alpha_i - 1}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}.$$

Again, it is not essential to fully grasp the mathematical definition. For our purposes, it suffices to know that the Dirichlet distribution assigns probabilities to a vector of probabilities.

A well-known application of the Dirichlet distribution in the natural language processing is the model known as *latent Dirichlet allocation* (LDA), which we will introduce later. Here we will give a brief informal description as a way to motivate the Dirichlet distribution. The LDA models a set of documents as having 'latent' (unobserved, unlabeled) dimensions corresponding to 'topics'. For example, a document may be about 'politics', but maybe also a bit of 'finance' is mixed too, but no or little 'sports'. Each document is modeled as a multinomial distribution over words. The probability parameters of the multinomial distributions, then, are determined by the topic distribution which is assumed to be a Dirichlet distribution.

2.3.8 *Gaussian distribution*

The *Gaussian (or normal) distribution* is probably the most important distribution for probability and statistics. It arises naturally for many continuous random variables. We will discuss its importance more when we discuss the *central limit theorem* below. For now, we will only define some properties of it. The normal distribution is the well-known distribution with the bell-shaped probability distribution function. It is parametrized by the mean μ and the variance σ^2. The formula for the normal PDF is

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

The normal distribution with $\mu = 0$ and $\sigma = 1$ is called the *standard normal distribution*. Changing μ changes the location of the distribution without affecting its shape, and changing σ changes the scale of the distribution (larger the sigma wider/shorter the PDF) without affecting its center. As a result the μ and σ (or σ^2) are sometimes called *location* and *scale* parameters respectively.

A useful fact (especially in statistics) about the normal distribution is that approximately 68 % of the probability mass (darkest area in Figure 2.14) falls between the interval $\mu \pm \sigma$, and 95 % of the probability mass (the second darkest shade in Figure 2.14) falls between the interval $\mu \pm 2\sigma$.

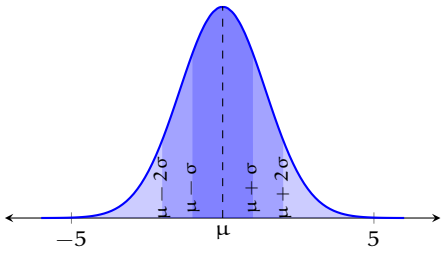

Figure 2.14: Normal PDF, with $\mu = 0$ and $\sigma = 1$.

2.3.9 *Student's t-distribution*

The *Student's t-distribution* (or simply t-distribution) is a probability distribution similar to the normal distribution. It has a central role in statistics, particularly in hypothesis testing. We will not go into details of description of the t-distribution here (although the name has a fun story that has to do with beer). The main use of the t-distribution is when one wants to estimate the parameters (μ) of a distribution associated by a population from a limited sample. It turns out, such an estimate is overconfident if one assumes that the sample means are normally distributed. The t-distribution corrects this since it has 'heavier tails' in comparison to the normal distribution. It has a single parameter *degrees of freedom*, ν (we compare it with a standard normal distribution with $\mu = 0$ and $\sigma = 1$). As the degrees of freedom increase, the distribution approaches to normal distribution. Figure 2.15 compares the t-distribution with the normal distribution.

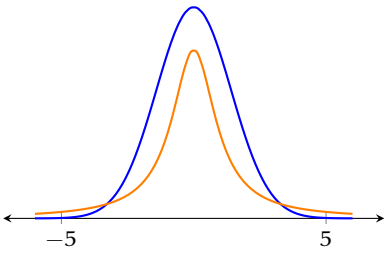

Figure 2.15: Standard normal distribution (blue) in comparison to the t-distribution with one degree of freedom ($\nu = 1$) (orange). The t-distribution has more probability mass at the tails. Note that this demonstration shows a particularly exaggerated difference since $\nu = 1$ is the case with the largest difference between the t-distribution and the standard normal distribution. The difference diminishes with increased degrees of freedom.

2.4 *Joint probability*

Our discussion of the random variables so far involved only a single value. Such random variables are called *univariate* random variables. For almost any practical application of probability, we have to deal with multiple (univariate) random variables. Multiple random variables that take real numbers define a *multivariate* distribution. Joint probability distribution for a set of discrete random variables may be specified by a (multi-dimensional) table listing probabilities of all possible combinations of the values each random variable take. In general, we note the joint probability mass function of two random variables X and Y as $P(X, Y)$. Analogously, a probability density function of a bivariate random variable (joint probability density of two continuous random variables) is noted as $p(X, Y)$. The notation naturally extends to distributions with more than two variables, e.g., $P(X, Y, Z)$ or $p(X, Y, Z)$.

For an example joint probability distribution, we will return to the letter probabilities from our hypothetical ancient language with eight letters (Table 2.3). It turns out, the ancient language we deal with had a few distinguishable dialects. The experts (of course hypothetical) distinguish northern, southern and eastern dialects, and they indicate that the official eastern dialect was more common (with probability 0.70) in the transcripts that survived. The others are less common, with probabilities 0.20 and 0.10 for north and south respectively. Note that 'dialect' of a document forms another categorical distribution (or multinomial distribution depending on what we are modeling).

We present joint probabilities of letters and dialects in Table 2.5. The table specifies the joint probability distribution $P(\text{letter}, \text{dialect})$, where 'letter' and 'dialect' are the random variables. Note that the probabilities in each cell in this table corresponds to probabilities of

Table 2.5: Joint probability table for letters and dialects in our example ancient language.

let.	east	north	south
a	0.198	0.007	0.028
b	0.029	0.012	0.001
c	0.025	0.009	0.012
d	0.062	0.015	0.007
e	0.172	0.097	0.017
f	0.017	0.003	0.007
g	0.050	0.013	0.000
h	0.146	0.044	0.029

joint events, observing a particular letter belonging to a particular dialect in the documents we have. For example, the first row and the second column indicates the probability of picking a letter from our corpus, which turns out to be the letter 'a' within a document written in the 'northern' dialect. We represent this probability using the notation $P(\text{letter=a}, \text{dialect=north})$. When the random variable we refer to is clear from the values, we will simplify this notation as $P(a, \text{north})$.

In Table 2.5, all values on the first column is naturally higher, since this column corresponds to the most common dialect. Similarly, the probabilities on fifth row is also rather high, since 'e' is an overall high-probability letter (see Table 2.3). An early note here is that the columns or rows are not just multiples of other columns or rows. This is an important property showing that the random variables are not *independent*. The (in)dependence of random variables is an important property, and we will discuss it further shortly.

2.5 Conditional probability

In many situations, we are interested in the probability of an event given that another event has happened. For example, we might be interested in the probabilities of the letters, given a particular dialect. The quantity that expresses this is the *conditional probability*. The notation used for conditional probability is $P(X = x \mid Y = y)$ or in our simplified notation $P(x \mid y)$, read as 'probability of x given y'. While the notation above defines the probability of an event *given* another event, $P(X \mid Y)$, where X and Y are random variables, defines a conditional distribution. In our example on letters and dialects, $P(\text{letter=a} \mid \text{dialect=north})$ is the probability of the letter 'a' given that the dialect is the northern dialect. While $P(\text{letter} \mid \text{dialect})$ is the conditional distribution of letters given the dialect.

Table 2.6: Conditional probabilities of $P(\text{letter} \mid \text{dialect})$.

let.	east	north	south
a	0.283	0.033	0.284
b	0.041	0.062	0.009
c	0.036	0.047	0.116
d	0.089	0.074	0.069
e	0.246	0.486	0.165
f	0.024	0.013	0.067
g	0.072	0.065	0.000
h	0.208	0.220	0.289

Table 2.6 shows the conditional probabilities, $P(\text{letter} \mid \text{dialect})$. Each cell on the table represents probability of observing a letter, given the dialect is the one marked at the head of the column. For example, the first row and second column indicates the probability of picking the letter 'a' among the documents that belong to 'northern' dialect (compare this with the interpretation of the same cell in Table 2.6). The distribution presented on Table 2.6 is a distribution over letters. It is important to note that unlike joint probability, the conditional probability is asymmetric.

We presented the conditional probabilities in Table 2.6 without telling how we did that. The joint probability distribution of two (or more) random variables specify the relevant conditional distribution (although calculating conditional distribution from the joint distribution may not always be trivial). The relation between joint and conditional distributions is

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)} \tag{2.6}$$

This calculation, together with the joint probability $P(X, Y)$, requires

$P(Y)$. Given the joint distribution $P(X, Y)$, $P(Y)$ can be calculated by summing over all possible values of X, which is called *marginalization*:[20]

$$P(Y = y) = \sum_x P(X = x, Y = y) \tag{2.7}$$

[20] Sometimes we talk about 'marginalizing out', since, in a sense, we are taking one of the variables out of the joint distribution by summing over all its values.

This means we are summing up all values on the joint distribution table by rows (or columns depending on the variable). The resulting probabilities are called *marginal probabilities*, since it is customarily written at the margins of the joint distribution table as in Table 2.7. The column margin is the same letter distribution in our original example from Table 2.3.

Table 2.7: Joint probability table for letters and dialects with marginal probabilities. $P(d)$ is the (marginal) probability of dialects, and $P(l)$ is the probability of letters in the corpus.

let.	east	north	south	$P(l)$
a	0.20	0.01	0.03	0.23
b	0.03	0.01	0.00	0.04
c	0.03	0.01	0.01	0.05
d	0.06	0.01	0.01	0.08
e	0.17	0.10	0.02	0.29
f	0.02	0.00	0.01	0.03
g	0.05	0.01	0.00	0.06
h	0.15	0.04	0.03	0.22
$P(d)$	0.70	0.20	0.10	1.00

If the conditional distribution $P(Y \mid X)$ and the $P(X)$ is known, the marginal probability of a random Y can also be calculated using the relation between the joint and conditional probabilities defined in Equation 2.6:

$$P(Y = y) = \sum_x P(Y = y \mid X = x)P(X = x)$$

From Equation 2.6, it is easy to show that the joint probability distribution of two variables X and Y can be calculated by

$$P(X, Y) = P(X|Y)P(Y) \quad \text{or} \quad P(Y, X) = P(Y|X)P(X).$$

Although both formulas lead to the same result, in practice, the calculations one needs to carry out are different. Preference towards one or the other may be more practical in solutions of different problems. The above rule generalizes to more than two variables as well. For three variables,

$$\begin{aligned} P(X, Y, Z) = &\quad P(X|Y, Z)P(Y|Z)P(Z) = \quad P(X|Y, Z)P(Z|Y)P(Y) \\ = &\quad P(Y|X, Z)P(X|Z)P(Z) = \quad P(Y|X, Z)P(Z|X)P(X) \\ = &\quad P(Z|X, Y)P(X|Y)P(Y) = \quad P(Z|X, Y)P(Y|X)P(X). \end{aligned} \tag{2.8}$$

The alternative ways of expanding the joint probability is called *factorization*s of it. We are free to choose the factorization that is most convenient for the particular application at hand.

In general, for any number of random variables, we can write

$$P(X_1, X_2, \ldots, X_n) = P(X_1|X_2, \ldots, X_n)P(X_2, \ldots, X_n). \tag{2.9}$$

Now, we can go on and expand the term $P(X_2, \ldots, X_n)$ recursively until we reach a single variable. This is called the *chain rule* of probabilities. As before, we have multiple ways to factorize the joint distribution since the ordering of the variables is not significant for the joint distribution.

For n variables, how many possible factorizations are there?

All the concepts discussed in this section generalizes to the continuous random variables as well. As usual, we use probability density function (instead of probability mass function) for specifying the continuous distributions, and sums become integrals.

2.6 *Bayes' formula*

Now we are ready to introduce one of the most important formulas in probability and statistics.

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

This identity can easily be derived from the relation between the joint and conditional probabilities (Equation 2.6). As a result, in itself, it is a simple statement of probability theory.[21] However, it is very important for machine learning and statistical inference.

[21] Although this formula plays a role in the controversy between Bayesian - frequentist approaches to statistics (see Section 2.12), there is nothing controversial about the formula itself.

A common example given for the use of Bayes' formula is medical diagnosis using a test of some sort. For the sake of the exercise, we will pretend to be a doctor. In a routine checkup, one of the medical tests was positive (indicated having a particular illness/condition) for one of our patients. The decision we are faced is the probability of the patient being ill, given that the test was positive. The test was performed using a device with the following specifications:

$P(\text{t+}|\text{ill}) = 0.99$, that is, the test returns positive 99 % of the cases if the patient is ill.

$P(\text{t+}|\text{healthy}) = 0.02$, that is, the test returns a false positive in 2 % of the cases.

Seeing these numbers, it is an easy mistake to think that our patient has the condition with a 99 % probability. Bayes' formula tells us how to calculate this properly:

$$P(\text{ill}|\text{t+}) = \frac{P(\text{t+}|\text{ill})P(\text{ill})}{P(\text{t+})}$$

The formula also tells us that we need additional information for this calculation. Fortunately (as in any made-up example), we know the unconditional probability of this particular condition, which turns out to be 0.000 10. So, one in 10 000 people is expected to have the condition. How do we calculate $P(\text{t+})$? The solution is marginalization. Since we know all possible values for the condition, we can simply marginalize the illness variable out.

$$P(\text{t+}) = P(\text{t+}, \text{ill}) + P(\text{t+}, \text{health}) = P(\text{t+} \mid \text{ill})P(\text{ill}) + P(\text{t+} \mid \text{healthy})P(\text{healthy})$$

Now we can put all in the Bayes' formula and calculate the expected probability of the patient being ill.

$$P(\text{t+}) = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.02 \times 0.9999}$$

If you do the math, you will find that the probability of the patient having the condition given the test result is less than 0.5 %, that is 5 in a thousand. This result comes as a surprise to most people. Besides showing the importance of tests with low false-positive rates, it also shows that human intuitions about probabilities are often not as accurate as one hopes for.

The Bayes' formula, this basic statement about probabilities, has a very important role in machine learning and statistics. We will return to it in later lectures with different use cases.

2.7 *Statistical independence*

The concept of statistical (in)dependence plays a very important role in machine learning and statistics. In any successful machine learning application, we are able to predict value of a random variable because it depends on another one. For example, financial institutions can make informed decision about granting loans based on applicants' financial status and past actions, because whether one will pay their loans back or not is not independent from these variables. Or, in NLP, we can guess the 'sentiment' in a product review from the words within the review, because the author's sentiment towards the product and his/her choices of words are not independent. If knowing one variable helps us (to some extent) guess the other variable, then we say that they are dependent, otherwise they are independent.

This leads to an important fact: if two random variables X *and* Y *are independent* then

$$P(X|Y) = P(X) \quad \text{and} \quad P(Y|X) = P(Y). \tag{2.10}$$

This simply says that knowing one of the variables does not change the probability of the other variable. From equations 2.10 and 2.6, we can easily derive that joint probability of *independent random variables* X and Y are simply the product of the individual probabilities:

$$P(X, Y) = P(X)P(Y)$$

This identity is very handy especially when we have many variables, as it simplifies the calculation of the joint probability greatly (compare with application of chain rule without independence in Equation in 2.8).

Independence assumptions as described above may not always be realistic. However, it happens often that two otherwise dependent variables become independent if we know the value of a third variable. This is called *conditional independence*. If the variables X and Y are conditionally independent given Z,

$$P(X, Y|Z) = P(X|Z)P(Y|Z).$$

Equivalently, this also means that

$$P(X|Y, Z) = P(X|Z).$$

The conditional independence comes handy in simplifying some problems. For example, in *naive Bayes* classifier the conditional independence (assumption) greatly simplifies the model, allows estimation of probabilities that are otherwise very difficult to estimate. A popular application of the naive Bayes classifier is spam filtering. To simplify let's assume we only track three words, and binary random variables W_1, W_2 and W_3 indicate whether each of the words occur in a given email or not. Our task, is then to estimate the probability that a given email is spam based on occurrences of these words,

$P(\text{spam} \mid W_1, W_2, W_3)$. This probability turns out to be difficult to estimate since for most choices of words, it is unlikely to observe enough number of spam and non-spam documents containing all of these words (especially when tracking many words rather than just three). So, we estimate $P(W_1, W_2, W_3 \mid \text{spam})$, and calculate the probability of email being spam given the words in (or not in) it using the Bayes' formula.

The probability $P(W_1, W_2, W_3 \mid \text{spam})$ is not easy to estimate either. Many combinations of (even these three) words will never be observed in any collections of emails. However, if we assume that the occurrence of words are independent of each other given the email is spam (or not), the calculations become much simpler. Although this is not necessarily a correct assumption (occurrence of words in a document is not independent of each other), it turns out the damage is not big: the resulting model does well in practice while simplifying the estimation considerably. Given the conditional independence assumption, we can simply write

$$P(W_1, W_2, W_2 \mid \text{spam}) = P(W_1|\text{spam})P(W_2|\text{spam})P(W_3|\text{spam}).$$

The probabilities on the right hand side above can simply be estimated from the number of times each word occurs in spam and non-spam documents.

2.8 *Expected value of a joint probability distribution*

We defined expected value of any function of a random variable in Equation 2.3, which generalizes to the case of joint probability distributions. Expected value of a function of two random variables is,

$$E[f(X, Y)] = \sum_x \sum_y P(x, y) f(x, y). \tag{2.11}$$

If we want to get the expected value of each random variable from the joint distribution, we simply set the function $f(x, y) = x$, which yields

$$\mu_X = E[X] = \sum_x \sum_y P(x, y) x. \tag{2.12}$$

Note that the inner sum is in fact marginalizing y out (Equation 2.7). In general, this formulation of expected value generalizes to continuous variables, and joint distributions of many random variables. Especially when we deal with joint distribution of a large number of variables, it is handy to use the vector notation. For example, we can represent any joint value (x, y) from the joint distribution of X and Y as a vector $z = (x, y)$. Then, Equation 2.11 can equivalently expressed as

$$E[f(X, Y)] = \sum_{z \in XY} P(z) f(z). \tag{2.13}$$

Here, the result is also a two-dimensional vector, and XY represents set of all combinations of the values both random variables take.

2.9 Covariance

In Section 2.2.7, we defined *variance* of a single random variable as squared expected difference from its mean. For joint distributions, we use the same idea to calculate the variances of component distributions. For a bivariate discrete distribution, we can calculate the variance of one of the variables with

$$\sigma_X = \sum_x \sum_y P(x, y)(x - \mu_x)^2.$$

When we have a joint distribution (more than one variable), however, another relevant and important quantity is the *covariance*. Covariance of two random variables is defined as

$$\text{cov}(X, Y) = \sigma_{XY} = E[(X - E[X])(Y - E[Y])]. \tag{2.14}$$

In words, it is the expected value of the product of the differences from means for each variable. Although this may not be easy to understand at first sight, it is worth having a look at the formula more carefully. If both variables are larger or smaller than their mean at the same time, the product in Equation 2.14 will be positive. Furthermore, higher the difference from the mean, the higher the product will be. On the other hand, if one variable is larger than its mean, and the other variable is smaller than its mean, the result will be negative, and similarly, the larger the absolute values, the larger the absolute value of the product in Equation 2.14. If one of these conditions is a general trend throughout all values the random variables take, the covariance will be a (large) positive or negative number respectively. If the positive and negative products occur by chance, then they will cancel each other, and covariance will be zero (or will have a small absolute value). In summary, if the variables co-vary we will get a non-zero covariance.

From the definitions of variance and covariance, you can see that both concepts are related. Variance is the covariance of a variable with itself.[22] And in many cases, it is convenient to define a *variance–covariance matrix*, or simply *covariance matrix*. The covariance matrix of the joint distribution of k random variables, $X_1 \ldots X_k$, looks like

[22] Hence, it is always positive.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_k} \\ \sigma_{X_2X_1} & \sigma_{X_2} & \cdots & \sigma_{X_2X_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_kX_1} & \sigma_{X_2X_k} & \cdots & \sigma_{X_k} \end{bmatrix}.$$

Since $\sigma_{XY} = \sigma_{YX}$, the covariance matrix is a symmetric matrix, and its diagonal contains the variances of the individual random variables.

2.10 Correlation

The value of covariance depends on the scale or unit of the variables. A normalized 'unitless' measure, *correlation*, defined as

$$\text{corr}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.15}$$

is easier to interpret. The *correlation coefficient* defined above ranges between -1 and 1. A correlation of 1 indicates a perfect increasing relationship between two variables, while -1 indicates a perfect inverse relationship. If the correlation coefficient (hence, the covariance) is 0, then the random variables do not have a linear relationship. This quantity is known commonly known as *Pearson's correlation coefficient*, since it was developed by the statistician Karl Pearson.[23]

[23] There are a few other correlation coefficients used in statistics. However, if the name is not specified, the term 'correlation coefficient' refers to Pearson's correlation coefficient.

You should have already realized a relation between correlation (or covariance) and independence. If two random variables are independent, their covariance (and correlation) is 0. However, the reverse is not correct. Covariance only measures linear relationships. As a result, covariance may be zero for strongly dependent variables, if the dependence is not linear. For example, the covariance between any random variable and its square ($\text{cov}(X, X^2)$) is 0,[24] despite the fact that they clearly are not independent.

[24] Try proving this.

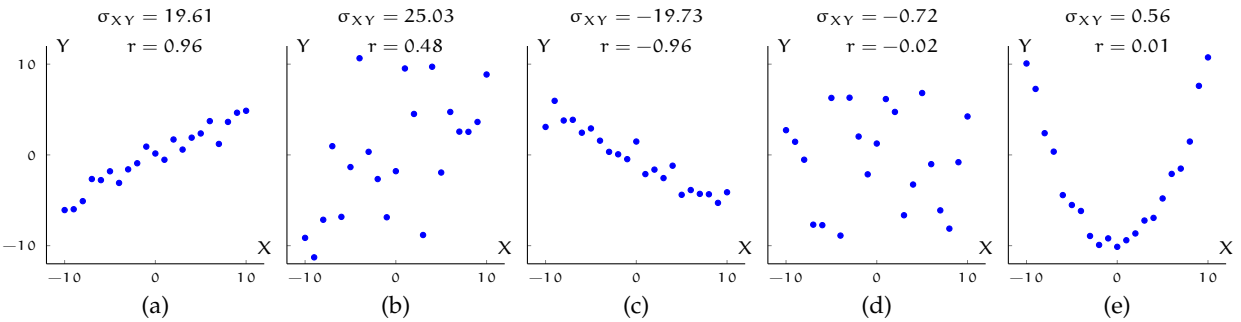

Figure 2.16: Scatter plots of samples from random variables with different dependence relations: (a) high positive correlation, (b) moderate positive correlation (c) high negative correlation, (d) uncorrelated variables, and (e) no correlation, but strong (quadratic) dependence.

Figure 2.16 demonstrates various levels of correlation and (in)dependence of two random variables. In Figure 2.16a and Figure 2.16c, the variables are highly correlated. For both pairs, the absolute values of correlation coefficient (and covariance) are large. The difference between them is the direction of the correlation, which is indicated by the sign of the correlation coefficient. The variables in Figure 2.16a are positively correlated, both increase and decrease at the same time. On the other hand, the ones in Figure 2.16a are negatively correlated, one decreases while the other increases. In both cases, the important thing to note is that the variables are highly dependent. Although we cannot predict one of the variables from the other with certainty, knowing one of the variables gives a lot of information about the other variable. Figure 2.16b presents a milder (positive) correlation. Here, the amount of information we get by knowing one of the variables is not as much as in (a) and (c). This is clearly indicated by the correlation coefficient. However, note that the covariance between the variables in (b) is higher than the covari-

ance in (a) - due to high variance of Y. It is difficult to interpret covariances directly, but correlation coefficient is clearly interpretable. The variables plotted in Figure 2.16d are not correlated. The graph does not indicate any reasonable relation between the two variables either. As a result both covariance and correlation scores are small. The variables plotted in Figure 2.16e are clearly not independent. Knowing one informs us about the other as much as in (a) and (c). However, since the relationship is not linear, the correlation coefficient is almost 0. This exemplifies why lack of correlation is not necessarily and indication of independence. We will introduce other, more general, measures of (in)dependence in Chapter 3.

2.11 *Correlation and Causation*

Informally, the term correlation is used for any type of dependence between variables, not just linear relationships. We noted earlier that lack of correlation does not indicate independence. Another common confusion about correlation (or general idea of dependence) is related to its relation with *causation*. Correlation does not indicate causation, although if one variable causes the other we expect to see correlation. If two variables are dependent (or correlated), they do not have to have a direct causal link. Sometimes, the dependence may be due a common dependence to another variable. For example, the fact that height of someone and their salaries are positively correlated is largely explained by the fact that in most of our present societies, women tend earn less, and tend to be shorter than men. Hence, at least part of the correlation we observe is not a causal relation between height and the salary, but a common causal ancestor, gender.

A more scientific example is the unmistakeable correlation shown in Figure 2.17 between per-capita chocolate consumption and number of Nobel laureates in a country. Although the correlation is sound, you should think twice before going into a chocolate diet with the hope of winning a Nobel prize.

Another issue is that statistical dependence (correlation) does not indicate the direction of the causation. For example, for a bystander observing a fire, it may seem like 'as more fire trucks arrive, the fire gets worse', but the direction of causation is probably the other way around. So although, we won't be doing any analysis of causation, you've been warned for this common error (see also the cartoon in Figure 2.18).

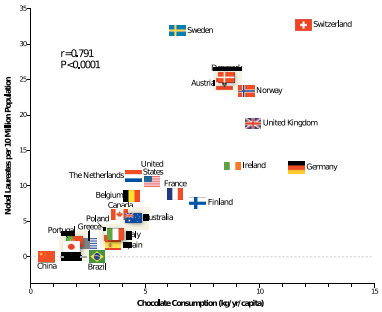

Figure 2.17: From (Messerli 2012). The figure shows a strong correlation between chocolate consumption in a country and number of Nobel prizes awarded to its citizens.

2.12 *Where do the probabilities come from?*

So far, we treated probabilities as numbers (between 0 and 1, inclusive) and did not say much about where do they come from. The question of what exactly a probability is a difficult one, and there have been (rival) alternative views on it. Two well-known views, we will encounter often is *frequentist* (classical) and *Bayesian* (probabilis-

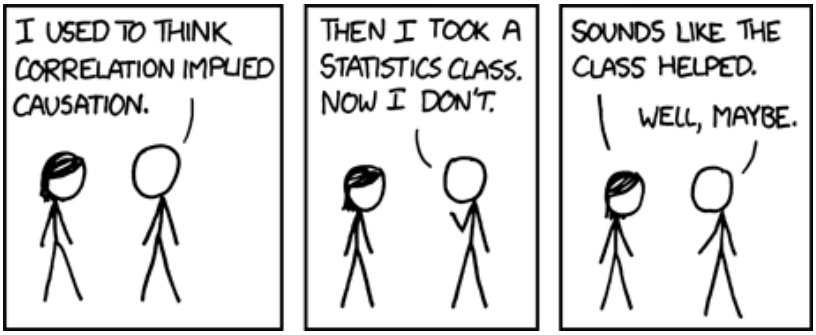

Figure 2.18: http://xkcd.com/552/

tic) approaches. Although both approaches agree on all the aspects of the probability theory we discussed above, the interpretations of probability in these views differ, and often lead to different methods of estimation or learning from data. This section is an informal, rather 'philosophical', note on this difference. We will occasionally return to it later in this course.

In frequentist view, the notion of probability is related to long-run relative frequency. That is, probability of an event is the relative frequency of its occurrence. For example, if we are interested in probability of a particular word, we can find its relative frequency on a large corpus. That is, the number of times the word appears in the corpus divided by the number of words in the corpus. Frequency-based probabilities leads to the estimation method called *maximum likelihood estimation* (MLE) we will discuss later. The MLE is prevalent in statistics and machine learning. The MLE results in an *unbiased* estimate, which means the estimate is guaranteed to converge to the actual value being estimated in the limit.[25]

[25] As we will discuss in many places, however, MLE *overfits* the data used for the estimation, the findings may not be general enough to be useful outside the data used for estimation. There are some modifications to MLE, that makes it more resistant to overfitting.

In frequentist view of probability, not every statement can be assigned to a probability value. Some quantities we are often interested are fixed, hence there is no notion of repeated experiments, and hence, no probability value. For example, we can assign a probability value to whether the next sentence in a conversation (or in a book) will be 10 words long. Because this is a repeatable experiment. However, we cannot assign a probability value to average number of words in a sentence in English (to simplify, say, in all written documents so far). Even though we do not know this quantity, there is a single number expressing it. Hence, in the frequentist view we cannot talk about the probability of average sentence length being 10. This seemingly 'philosophical' standpoint has a very big impact on how results are evaluated in experimental studies, and hence, on the current scientific enterprise.

Intuitively, frequency of an event and its probability is strongly related. However, our everyday notion of probability does not necessarily involve repetitive experiments. If you were given a fair coin, you would not need a large number of experiments to conclude that the probability of heads (or tails) is 0.5. On the practical side, some events do not occur frequently enough to be estimated reliably. As we will discuss late, most objects of interest in NLP, such as words,

are particularly bad at showing up in where we look for them (in corpora).

The Bayesian notion of probability is based on (subjective) 'degree of belief', matching more closely to our everyday notion of probability. In this view, probabilities are degrees of belief, and updated 'rationally', with the data at hand and based on the rules of probability theory. In particular, based on the Bayes' formula we discussed in Section 2.6. Now that we consider probabilities as degrees of beliefs, we can make probabilistic statements on things we are not allowed in case of frequentist tradition. For example, we can easily talk about the probability of average sentence length being 10,[26] and we can update this probability (in fact the whole probability distribution) with the data we observe. One particular benefit of the Bayesian estimation is that we do not need anything other than probability theory for estimation.

[26] Or, maybe between 9.50 and 10.50, since probabilities for single values is 0 for continuous random variables.

A common criticism for Bayesian estimation is is that one needs to choose a (subjective) prior probability (distribution) before starting the data-driven estimation. The word 'subjective' does not sound like a good idea for science. As a result, Bayesian methods are often criticized for not being 'objective'. However, the prior information does not necessarily involve 'personal' beliefs. For example, in our example with average number of words in a sentence, there is nothing wrong to start with the presupposition that the average number of sentences cannot be a negative number. Further, it is hardly a wrong assertion to assume that very large numbers, e.g., 1 000, are very unlikely (especially considering we are estimating *average* sentence length). There are many other forms of prior assumptions that are perfectly justified, based on the knowledge accumulated in the field–often based on data from earlier studies. Although making truly subjective decisions is not desirable while interpreting experimental results in science, if we shift our interest towards engineering rather than science, the use of prior information (subjective or otherwise) is not an issue at all. There is nothing wrong with using a subjective method, as long as it performs well for the task at hand.

A practical note on difference between the two estimation methods is about the computational power required. Bayesian estimation typically requires more computation power. This difference is becoming less important with the increasing power offered by developments in computer hardware and approximate estimation techniques being developed. However, it is still an important factor in many problems.

The debate is old and far from being settled yet. In this course, we do not take sides in this debate. We introduce methods that stem from both approaches for estimating probabilities. The aim of this section is to inform the reader about these different approaches since they have important consequences for most of the methods we discuss.

Summary

This refresher covers a large number of concepts briefly. The information provided above is (more than) necessary for following the topics from probability in this lecture. However, interested students are encouraged to consult other sources. There are many excellent books on probability theory, it is difficult to suggest a single one here. Below we list only a few.

MacKay (2003) covers most of the topics discussed in a way very relevant to machine learning. The complete book is available freely online (see the link in the reference below). Grinstead and Snell (2012) is a more conventional introduction to probability theory. This book is also freely available online. For an influential, but not quite conventional approach (from a Bayesian perspective) see Jaynes (2007).

Bibliography

Grinstead, Charles Miller and James Laurie Snell (2012). *Introduction to probability*. American Mathematical Society. ISBN: 9780821894149. URL: http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html.

Jaynes, Edwin T (2007). *Probability Theory: The Logic of Science*. Ed. by G. Larry Bretthorst. Cambridge University Press. ISBN: 978-05-2159-271-0.

MacKay, David J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. ISBN: 978-05-2164-298-9. URL: http://www.inference.phy.cam.ac.uk/itprnn/book.html.

Messerli, Franz H (2012). "Chocolate consumption, cognitive function, and Nobel laureates". In: *The New England journal of medicine* 367.16, pp. 1562–1564.