A review of Probability and Statistics

This document contains a summary of background material from probability and statistics. Important concepts are **highlighted**.

1 Probability

A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, consisting of a set Ω, a σ-algebra $\mathcal{F}$ of **measurable** subsets of Ω, called **events**, and a **probability measure** $\mathbb{P}$. That $\mathcal{F}$ is a σ-algebra means that it contains $\emptyset$ and Ω and that it is closed under countable unions and complements. The probability measure $\mathbb{P}$ is a non-negative function $\mathbb{P}\colon \mathcal{F} \to [0,1]$ such that $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$, and

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$$

for a collection $\{A_i\} \subset \mathcal{F}$ with $A_i \cap A_j = \emptyset$ for $i \neq j$. We interpret $\mathbb{P}(A \cup B)$ as the probability of A *or* B happening, and $\mathbb{P}(A \cap B)$ as the probability of A *and* B happening. Note that $(A \cup B)^c = A^c \cap B^c$, where A^c is the complement of A in Ω.

If the A_i are not necessarily disjoint, then we have the important **union bound**

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}(A_i).$$

This bound is sometimes also referred to as the *zero-th moment method*.

Example 1.1. Suppose we are rolling a die and A is the event that the result is an even number, while B is the event that the result is at least 4. Then

$$\mathbb{P}(A) = \frac{1}{2}, \quad \mathbb{P}(B) = \frac{1}{2}, \quad \mathbb{P}(A \cup B) = \frac{2}{3} < 1 = \mathbb{P}(A) + \mathbb{P}(B).$$

Clearly, the union bound gives a trivial bound here. In general, the union bound becomes less useful the more the sets involved overlap.

We say that an event A holds **almost surely** if $\mathbb{P}(A) = 1$ (note that this does not mean that the complement of A in Ω is empty).

Random variables

A measurable function between two measure spaces is a function for which the preimage of a measurable set is measurable. A **random variable** is a measurable map

$$X\colon \Omega \to \mathcal{X},$$

where $\mathcal{X}$ is typically $\mathbb{R}$, $\mathbb{R}^d$, $\mathbb{N}$, or a finite set $\{0, 1, \dots, k\}$, all considered as measure spaces with the Borel σ-algebra (the smallest σ-algebra containing the open sets, where for discrete sets we consider the discrete topology). For a measurable set $A \subset \mathcal{X}$, we write

$$\mathbb{P}(X \in A) := \mathbb{P}(\{\omega \in \Omega\colon X(\omega) \in A\}).$$

We will usually use upper-case letters X, Y, Z for random variables, and lower-case letters x, y, z for the values that these can take.

Example 1.2. A random variable specifies which events "we can see". For example, let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and define $X\colon \Omega \to \{0, 1\}$ by setting $X(\omega) = \mathbf{1}\{\omega \in \{4, 6\}\}$, where $\mathbf{1}$ denotes the indicator function. Then

$$\mathbb{P}(X = 1) = \frac{1}{3}, \quad \mathbb{P}(X = 0) = \frac{2}{3}.$$

If all the information we get about Ω is from X, then we can only determine whether the result of rolling a die gives an even number greater than 3 or not, but not the individual result.

The map $A \mapsto \mathbb{P}(X \in A)$ for subsets of $\mathcal{X}$ is called the **distribution** of the random variable. The distribution completely describes the random variable, and there will often be no need to refer to the domain Ω. If $F\colon \mathcal{X} \to \mathcal{Y}$ is another measurable map, then $F(X)$ is again a random variable. In particular, if $\mathcal{X}$ is a subset of $\mathbb{R}^d$, we can add and multiply random variables to obtain new random variables, and if X and Y are two distinct random variables, then (X, Y) is a random variable in the product space. In the latter case we also write $\mathbb{P}(X \in A, Y \in B)$ instead of $\mathbb{P}((X, Y) \in A \times B)$. Note that this also has an interpretation in terms of intersections of events: it is the probability that *both* $X \in A$ and $Y \in B$.

A **discrete** random variable takes countable many values, for example in a finite set $\{1, \dots, k\}$ or in $\mathbb{N}$. In such a case it makes sense to talk about the probability of individual outcomes, such as $\mathbb{P}(X = k)$ for some $k \in \mathcal{X}$. An **absolutely continuous** random variable takes values in $\mathbb{R}$ or $\mathbb{R}^d$ for $d > 1$, and is defined as having a **density** $\rho(x) = \rho_X(x)$, such that

$$\mathbb{P}(X \in A) = \int_A \rho(x)\, \mathrm{d}x.$$

In the case where $\mathcal{X} = \mathbb{R}$, we consider the **cumulative distribution function** (cdf) $\mathbb{P}(X \leq t)$ for $t \in \mathbb{R}$. The complement, $\mathbb{P}(X > t)$ (or $\mathbb{P}(X \geq t)$), is referred to

as the **tail**. Many applications are concerned with finding good bounds on the tail of a probability, as the the tail often models the probability of rare events. If X is absolutely continuous, then the probability of X taking a particular single value vanishes, $\mathbb{P}(X = a) = 0$. For a random variable $Z = (X, Y)$ taking values in $\mathcal{X} \times \mathcal{Y}$, we can consider the **joint density** $\rho_Z(x, y)$, but also the individual densities of X and Y, for which we have

$$\rho_X(x) = \int_{\mathcal{Y}} \rho_Z(x, y) \, \mathrm{d}y.$$

The ensuing distributions for X and Y are called the **marginal distributions**.

Example 1.3. Three of the most common distributions are:

- **Bernoulli distribution** $\mathrm{Ber}(p)$, taking values in $\{0, 1\}$ and defined by

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

  for some $p \in [0, 1]$. We can replace the range $\mathcal{X} = \{0, 1\}$ by any other two-element set, for example $\{-1, 1\}$, but then the relation to other distributions may not hold any more.

- **Binomial distribution** $\mathrm{Bin}(n, p)$, taking values in $\{0, \dots, n\}$ and defined by

  $$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{1.1}$$

  for $k \in \{0, 1, \dots, n\}$ and some $p \in [0, 1]$. We can also write a binomial random variable as a sum of Bernoulli random variables, $X = X_1 + \cdots + X_n$, since $X = k$ if and only if k of the summands have the value 1.

- **Normal distribution** $\mathcal{N}(\mu, \sigma^2)$, also referred to as Gaussian, with mean μ and variance σ^2, defined on $\mathbb{R}$ and with density

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

  This is the most important distribution in probability and statistics, as most other distributions can be approximated by it.

Moments

The **expectation** (or mean, or expected value) of a discrete random variable is defined as

$$\mathbb{E}[X] = \sum_{k \in \mathcal{X}} k \cdot \mathbb{P}(X = k).$$

For an absolutely continuous random variable with density $\rho(x)$, it is defined as

$$\mathbb{E}[X] = \int_{\mathcal{X}} \rho(x) \, \mathrm{d}x.$$

Note that the expectation does not always need to exist since the sum or integral need not converge. When we require it to exist, we often write this as $\mathbb{E}[X] < \infty$.

Example 1.4. The expectation of a Bernoulli random variable with parameter p is $\mathbb{E}[X] = p$. The expectation of a Binomial random variable with parameters n and p is $\mathbb{E}[X] = np$. For example, if one were to flip a biased coin that lands on heads with probability p, then this would correspond to the number of heads one would "expect" after n coin flips. The expectation of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ is μ. This is the location on which the "bell curve" is centred.

One of the most useful properties is **linearity of expectation**. If $X_1, \ldots, X_n$ are random variables taking values in a subset of $\mathbb{R}^d$ and $a_1, \ldots, a_n \in \mathbb{R}$, then

$$\mathbb{E}[a_1 X_1 + \cdots + a_n X_n] = a_1 \, \mathbb{E}[X_1] + \cdots + a_n \, \mathbb{E}[X_n].$$

Example 1.5. The expected value of a Bernoulli random variable with parameter p is

$$\mathbb{E}[X] = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = p.$$

The linearity of expectation then immediately gives the expectation of the Binomial distribution with parameters n and p. Since such a random variable can be written as $X = X_1 + \cdots + X_n$, with X_i Bernoulli, we get

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np.$$

This would be (slightly) harder to deduce from the direct definition (1.1), when one would have to use the binomial theorem.

If $F \colon \mathcal{X} \to \mathcal{Y}$ is a measurable function, then the expectation of the random variable $F(X)$ can be expressed as

$$\mathbb{E}[F(X)] = \int_{\mathcal{X}} F(x) \rho(x) \, \mathrm{d}x \tag{1.2}$$

in the case of an absolutely continuous random variable, and similarly in the discrete case. [1]

An important special case is the indicator function

$$F(X) = \mathbf{1}\{X \in A\} = \begin{cases} 1 & X \in A \\ 0 & X \notin A. \end{cases}$$

Then

$$\mathbb{E}[\mathbf{1}\{X \in A\}] = \mathbb{P}(X \in A), \tag{1.3}$$

as can be seen by applying (1.2) to the indicator function. The identity (1.3) is useful, as it allows to properties of the expectation, such as linearity, in the study of probabilities of events. The expectation also has the following monotonicity property: if $0 \leq X \leq Y$, where X, Y are real-valued random variables, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

[1] We will not always list the formulas for both the discrete and continuous, when the form of one of these cases can be easily guessed from the form of the other case. In any case, the sum in the discrete setting is also just an integral with respect to the discrete measure.

Another important identity for random variables is the following. Assume X is absolutely continuous, takes values in $\mathbb{R}$, and $X \geq 0$. Then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, \mathrm{d}t.$$

Using this identity, one can deduce bounds on the expectation from bounds on the tail of a probability distribution.

The **variance** of a random variable is the expectation of the square deviation from the mean:

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

The variance measures the "spread" of a distribution: random variables with a small variance are more likely to stay close to their expectation.

Example 1.6. The variance of the normal distribution is σ^2. The variance of the Bernoulli distribution is $p(1-p)$ (verify this!), while the variance of the Binomial distribution is $np(1-p)$.

The variance scales as $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$. In particular, it is translation invariant. The variance is in general not additive (but it is, if the random variables are independent).

Besides the expectation and the variance, the **higher moments** $\mathbb{E}[X^k]$ of a random variable are of interest. These are encoded in the **exponential moment generating function** (mgf)

$$\mathbb{E}\left[\mathrm{e}^{\lambda X}\right] = \sum_{k=0}^{\infty} \frac{\mathbb{E}[X^k]}{k!}.$$

The moments need not all be finite. Related to the moment generating function is the **characteristic function**

$$\varphi_X(t) = \mathbb{E}\left[\mathrm{e}^{itX}\right].$$

The characteristic function is multiplicative:

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t). \tag{1.4}$$

It allows one to express the density $\rho(x)$ of an absolutely continuous random variable X as

$$\rho(x) = \frac{1}{2\pi} \int_{t=-\infty}^{\infty} \mathrm{e}^{-itx} \varphi_X(t)\, \mathrm{d}t.$$

The characteristic function plays an important role in a common proof of the central limit theorem.

Example 1.7. Let X be a Bernoulli random variable with parameter p. Then the characteristic function is $\varphi_X(t) = 1 - p + p\mathrm{e}^{it}$. For a binomial random variable, $X \sim \mathrm{Bin}(n, p)$, the characteristic function is $(1 - p + p\mathrm{e}^{it})^n$. This follows from the multiplicative property (1.4).

Independence

A set of random variables $\{X_i\}$ taking values in the same range $\mathcal{X}$ is called **independent** if for any subset $\{X_{i_1}, \ldots, X_{i_k}\}$ and any subsets $A_j \subset \mathcal{X}$, $1 \leq j \leq k$, we have

$$\mathbb{P}(X_{i_1} \in A_1, \ldots, X_{i_k} \in A_k) = \mathbb{P}(X_{i_1} \in A_1) \cdots \mathbb{P}(X_{i_k} \in A_k).$$

In words, the probability of any of the events happening simultaneously is the product of the probabilities of the individual events. A set of random variables $\{X_i\}$ is said to be **pairwise independent** if every subset of two variables is independent. Note that pairwise independence does not imply independence.

Example 1.8. Assume you toss a fair coin two times. Let X be the indicator variable for heads on the first toss, Y the indicator variable for heads on the second toss, and Z the random variable that is 1 if $X = Y$ and 0 if $X \neq Y$. Taken individually, each of these random variables is a Bernoulli random variable with $p = 1/2$. They are also pairwise independent, as is easily verified, but not independent, since

$$\mathbb{P}(X = 1, Y = 1, Z = 1) = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(X = 1)\mathbb{P}(Y = 1)\mathbb{P}(Z = 1).$$

Intuitively, the information that $X = 1$ and $Y = 1$ already implies $Z = 1$, so adding this constraint does not alter the probability on the left-hand side.

We say that a set of random variables $\{X_i\}$ is **i.i.d.** if they are **independent and identically distributed**. This means that each X_i can be seen as a *copy* of X_1 that is independent of it, and in particular all the X_i have the same expectation and variance.

One of the most important results in probability (and, arguably, in nature) is the (strong) **law of large numbers**. Given random variables $\{X_i\}$, define the sequence of averages as

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n).$$

Since each random variable is, by definition, a function on a sample space Ω, we can consider the pointwise limit

$$\lim_{n\to\infty} \overline{X}_n,$$

which is the random variable that for each $\omega \in \Omega$ takes the limit $\lim_{n\to\infty} \overline{X}_n(\omega)$ as value.[2]

Theorem 1.9 (Law of Large Numbers)**.** *Let $\{X_i\}$ be a sequence of i.i.d. random variables with $\mathbb{E}[X_1] = \mu < \infty$. Then the sequence of averages $\overline{X}_n$ converges almost surely to μ:*

$$\mathbb{P}(\lim_{n\to\infty} \overline{X}_n = \mu) = 1.$$

[2]That this is indeed a random variable in the formal sense follows from measure theory, we will not be concerned with those details.

Example 1.10. Let each X_i be a Bernoulli random variable with parameter p. One could think this as flipping a coin that will show heads with probability p. Then $\overline{X}_n$ is the *average* number of heads when flipping the coin n times. The law of large numbers asserts that as n increases, this average approaches p almost surely. Intuitively, when flipping the coin a billion times, the number of heads we get divided by a billion will be indistinguishable from p: if we do not know p we can estimate it in this way.

Some useful inequalities

In applications it is often not possible to get precise expressions for a probability we are interested in, most often because we don't know the exact distribution we are dealing with and only have access to parameters such as the expectation or the variance. There are several useful inequalities that help us bound the tail or deviation probabilities. For the following, we assume $\mathcal{X} \subset \mathbb{R}$.

- **Jensen's Inequality** Let $f\colon \mathbb{R} \to \mathbb{R}$ be a convex function, that is, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for $\lambda \in [0,1]$. Then
$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$
- **Markov's Inequality** ("first moment method") For $X \geq 0$ and $\lambda > 0$,
$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}[X]}{\lambda}.$$
- **Chebyshev's Inequality** ("second moment method") For $\lambda > 0$,
$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \lambda) \leq \frac{\mathrm{Var}(X)}{\lambda^2}.$$
- **Exponential Moment Inequality** For any $s, \lambda \geq 0$,
$$\mathbb{P}(X \geq \lambda) \leq \mathrm{e}^{-s\lambda}\,\mathbb{E}[\mathrm{e}^{sX}].$$

Note that both the Chebyshev and the exponential moment inequality follow from the Markov inequality applied to certain transformations of X.

The normal distribution

The normal, or Gaussian distribution deserves special attention. A random variable X on $\mathbb{R}$ is **normally distributed** or **Gaussian** with mean μ and variance σ^2, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if it has density

$$\gamma_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

If $X \sim \mathrm{N}(0,1)$, we denote by

$$\Phi(z) := \mathbb{P}(X \leq z) = \int_{-\infty}^{z} \gamma_{0,1}(x)\,\mathrm{d}x$$

the cumulative distribution function. There is not closed-form expression for this integral, and traditionally the values of this function would be computed numerically and recorded in tables. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $(X-\mu)/\sigma \sim \mathrm{N}(0,1)$, so it suffices to have a table for Φ in the standard normal case (with $\mu = 0$ and $\sigma^2 = 1$). See Figure 1.1 for an illustration of the Gaussian distribution. We note that while there is no closed-form expression for the distribution function, the moments are well known and easy to compute. The moment generating function of a Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$ is

$$\mathbb{E}[\mathrm{e}^{\lambda X}] = \mathrm{e}^{\mu\lambda + \frac{\sigma^2\lambda^2}{2}}.$$

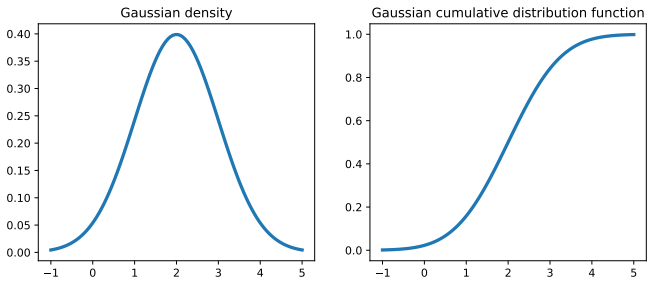

Figure 1.1: The density [left] and distribution function [right] of a Gaussian random variable.

The importance of the Gaussian distribution stems from the **central limit theorem**, which states that the average of independent random samples tends to the normal distribution.

Theorem 1.11. *(Central Limit Theorem) Let $X_1, \ldots, X_n$ be i.i.d. random variables with expected value μ and finite variance $\sigma^2 > 0$, and set $\overline{X}_n = (X_1 + \cdots + X_n)/n$. Then*

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}(\overline{X}_n - \mu) \leq z) = \Phi(z/\sigma).$$

A multivariate Gaussian is a random vector $X = (X_1, \ldots, X_n)$ with density

$$\gamma(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}.$$

The vector μ is the mean and the matrix $\boldsymbol{\Sigma}$ is the **covariance matrix** of X. The multivariate Gaussian has important invariance property: if $X \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ and $\boldsymbol{Q} \in O(n)$ is orthogonal, then $\boldsymbol{Q}X \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$.

Conditional probability and expectation

Given events $A, B \subset \Omega$ with $\mathbb{P}(B) \neq 0$, the **conditional probability** of A conditioned on B is defined as

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

One interprets this as the probability of A if we assume B. That is, if we observed B, then we replace the whole set Ω by B and consider B to be the new space of events, considering only the part of events A that lie in B. We can rearrange the expression for conditional probability to

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

from which we get the sometimes useful identity

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c), \tag{1.5}$$

where B^c denotes the complement of B in Ω.

Since by exchanging the role of A and B we get $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$, we arrive at the famous **Bayes rule** for conditional probability:

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

defined whenever both A and B have non-zero probability. These concepts clearly extend to random variables, where we can define, for example,

$$\mathbb{P}(X \in A|Y \in B) = \frac{\mathbb{P}(X \in A, Y \in B)}{\mathbb{P}(X \in B)}.$$

Note that if X and Y are independent, then the conditional probability is just the normal probability of X: knowing that $Y \in B$ does not give us any additional information about X! If we fix an event such as $\{Y \in B\}$, then we can define the **conditioning** of the random variable X to this event as the random variable X' with distribution

$$\mathbb{P}(X' \in A) = \mathbb{P}(X \in A|Y \in B).$$

In particular, $\mathbb{P}(X \in A|Y \in B) + \mathbb{P}(X \notin A|Y \in B) = 1$.

Example 1.12. Consider the case of testing for doping at a sports event. Let X be the indicator variable for the presence of a certain drug, and Y the indicator variable for whether the person tested has taken the drug. Assume that the test is 99% accurate when the drug is present and 99% accurate when the drug is not present. We would like to know the probability that a person who tested positive actually took the drug, namely $\mathbb{P}(Y = 1|X = 1)$. Translated into probabilistic language, we know that

$$\begin{aligned} &\mathbb{P}(X = 1|Y = 1) = 0.99, \quad \mathbb{P}(X = 0|Y = 1) = 0.01 \\ &\mathbb{P}(X = 0|Y = 0) = 0.99, \quad \mathbb{P}(X = 1|Y = 0) = 0.01. \end{aligned}$$

Assuming that only 1% of the participants have taken the drug, which translates to $\mathbb{P}(Y = 1) = 0.01$, we find that the overall probability of a positive test result is, using (1.5),

$$\begin{aligned}\mathbb{P}(X = 1) &= \mathbb{P}(X = 1|Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(X = 1|Y = 1)\mathbb{P}(Y = 1)\\ &= 0.01 \cdot 0.99 + 0.99 \cdot 0.01 = 0.0198.\end{aligned}$$

Hence, using Bayes' rule, we conclude that

$$\mathbb{P}(Y = 1|X = 1) = \frac{\mathbb{P}(X = 1|Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X = 1)} = \frac{0.99 \cdot 0.01}{0.0198} = 0.5.$$

That is, we get the surprising result that even though our test is very unlikely to give false positives and false negatives, the probability that a person tested positive has actually taken the drug is only 50%. The reason is that the event itself is highly unlikely.

We now come to the notion of **conditional expectation**. Let X, Y be random variables. If X is discrete, then the conditional expectation of X conditioned on an event $Y = y$ is defined as

$$\mathbb{E}[X|Y = y] = \sum_k k\mathbb{P}(X = k|Y = y). \tag{1.6}$$

This is simply the expectation of the random variable X' with distribution $\mathbb{P}(X' \in A) = \mathbb{P}(X \in A|Y = y)$. Intuitively, we assume that $Y = y$ is given/has been observed, and consider the expectation of X under this additional knowledge.

Example 1.13. Assume we are rolling dice, let X be the random variable giving the result, and let Y be the indicator variable for the event that the result is at most 4. Then $\mathbb{E}[X] = 3.5$ and $\mathbb{E}[X|Y = 1] = 2.5$ (verify this!). This is the expected value if we have the additional information that the result is at most 4.

In the absolutely continuous case we can define a conditional density

$$\rho_{X|Y=y}(x) = \frac{\rho_{X,Y}(x, y)}{\rho_Y(y)}, \tag{1.7}$$

where $\rho_{X,Y}$ is the joint density of (X, Y) and ρ_Y the density of Y. By a common abuse of notation, we will often write

$$\rho(x|y) := \rho_{X|Y=y}(x).$$

The conditional expectation is then defined

$$\mathbb{E}[X|Y = y] = \int_{\mathcal{X}} x\rho(x|y)\,\mathrm{d}x. \tag{1.8}$$

We replace the *density* ρ_X of X with an updated density $\rho_{X|Y=y}$ that takes into account that a value $Y = y$ has been observed when computing the expectation of X.

When looking at (1.6) and (1.8), we get a different number $\mathbb{E}[X|Y = y]$ for each $y \in \mathcal{Y}$, where we assume $\mathcal{Y}$ to be the space where Y takes values. Hence, we can *define* a random variable $\mathbb{E}[X|Y]$ on Y as follows:

$$\mathbb{E}[X|Y](y) = \mathbb{E}[X|Y = y].$$

If $X = f(Y)$ is completely determined by Y, then clearly

$$\mathbb{E}[X|Y](y) = \mathbb{E}[X|Y = y] = \mathbb{E}[f(Y)|Y = y] = \mathbb{E}[f(y)|Y = y] = f(y),$$

since the expected value of a constant is just that constant, and hence $\mathbb{E}[X|Y] = f(Y)$ as a random variable.

Using the definition of the conditional density (1.7), Fubini's Theorem and expression (1.8), we can write the expectation of X as

$$\begin{aligned}
\mathbb{E}[X] &= \int_{\mathcal{X}} x\rho_X(x)\,\mathrm{d}x \\
&= \int_{\mathcal{X}} x \int_{\mathcal{Y}} \rho_{(X,Y)}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} x\rho_{(X,Y)}(x,y)\,\mathrm{d}x \right) \mathrm{d}y \\
&= \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} x\rho_{(X|Y=y)}(x)\,\mathrm{d}x \right) \rho_Y(y)\,\mathrm{d}y = \int_{\mathcal{Y}} \mathbb{E}[X|Y = y]\rho_Y(y)\,\mathrm{d}y.
\end{aligned}$$

One can interpret this as saying that we get the expected value of X by integrating the expected values conditioned on $Y = y$ with respect to the density of Y. In the discrete case, the identity has the form

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X|Y = y]\mathbb{P}(Y = y).$$

The above identities can be written more compactly as

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

In the context of machine learning, we assume that we have a (hidden) map $f\colon \mathcal{X} \to \mathcal{Y}$ from an input space to an output space, and that, for a given input $x \in \mathcal{X}$, the *observed* output is $y = f(x) + \epsilon$, where ϵ is random noise with $\mathbb{E}[\epsilon] = 0$. If we consider the input as a random variable X, then the output is random variable

$$Y = f(X) + \epsilon,$$

with $\mathbb{E}[\epsilon|X] = 0$ ("the expected value of ϵ, when knowing X, is zero"). We are interested in the value $\mathbb{E}[Y|X]$. For this, we get

$$\mathbb{E}[Y|X] = \mathbb{E}[f(X)|X] + \mathbb{E}[\epsilon|X] = f(X),$$

since $f(X)$ is completely determined by X.

Graphical models

Rather than just considering the dependence between two random variables, in statistical applications one often deals with many random variables that depend on each other in more intricate ways. For example, given random variables X_1, X_2, X_3, X_4 with joint density ρ, we can determine the joint density

$$\rho(x_1, x_2, x_3, x_4) = \rho(x_1, x_2|x_3, x_4)\rho(x_3, x_4). \tag{1.9}$$

Formally, this means that if we consider the random vectors $X = (X_1, X_2)$ and $Y = (X_3, X_4)$, then (1.9) simply reduces to the informal way of writing (1.7). We can further exploit the dependency between X_3 and X_4 and between X_1 and X_2, which leads to the **factorization**

$$\rho(x_1, x_2, x_3, x_4) = \rho(x_1|x_2, x_3, x_4)\rho(x_2|x_3, x_4)\rho(x_3|x_4)\rho(x_4). \tag{1.10}$$

Not every variable may be relevant for explaining another variable. One way of studying a set of random variables and their dependencies is using the formalism of **graphical models**. In a graphical model, the random variables are mapped to the nodes in a **directed acyclic graph**. A directed graph $G = (V, E)$ consists of a set of nodes V and a relation between the nodes $E \subset V \times V$, with the elements of E called edges. The term 'directed' means that the order of the elements making up an edge $(v, w) \in E$ matters. A path in G from v to w consists of a sequence of edges $(v_1, w_1), \ldots, (v_k, w_k)$ such that $v_1 = v$, $w_i = v_{i+1}$ for $1 \leq i \leq k-1$, and $w_k = w$. A cycle is a path from v to itself, and a graph is **acyclic** if it does not contain any cycle. Given a DAG with nodes $V = \{1, \ldots, n\}$, each node i has a set of parents, or predecessors, $\pi(i)$. If we associated to each node a random variable X_i and denote the joint density by ρ, then this density factors as product of conditional densities as follows:

$$\rho(x_1, \ldots, x_n) = \prod_{i \in V} \rho(x_i|x_j, j \in \pi(i))$$

An example of a graphical model is shown in Figure 1.2.

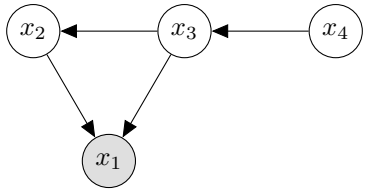

Figure 1.2: The graphical model $\rho(x_1|x_2, x_3)\rho(x_2|x_3)\rho(x_3|x_4)\rho(x_4)$.

Such a graph is also called a **Bayesian network**, due to the repeated use of Bayes' rule. A special case are **latent variable models** $\rho(x|z)$, where an observed random variable X depends on hidden, or latent, data Z. Graphical models are important in the study of **generative models** in machine learning.

2 Statistics

Statistics uses probability theory to analyse empirical data. Statistical inference aims to use data analysis to determine properties of some underlying probability distribution.

Estimation and the bias-variance trade-off

Suppose we observe samples $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, where $\mathcal{X}$ is a space of outcomes of an experiment. In statistics, the assumption is often made that these samples are realizations of a random variable X, distributed according to some probability distribution. We may be interested in determining the whole distribution that gave rise to the data, or, as is more often the case, just some properties of it. The setting for a statistical estimation problem thus consists of:

- A space of outcomes $\mathcal{X}$;
- A parametrized family of probability distributions $\{\mathbb{P}_\theta \colon \theta \in \Theta\}$;
- A map $g \colon \Theta \to \mathcal{Y}$, mapping a parameter to a quantity of interest to us.

For example, if $\mathcal{X} = \mathbb{R}$, $\theta = (\mu, \sigma^2)$ and $\mathbb{P}_\theta = \mathcal{N}(\mu, \sigma^2)$, then $g(\theta) = g(\mu, \sigma^2) = \mu$ if we are only interested in the mean. A **point estimator** is any map $T \colon \mathcal{X}^n \to \mathcal{Y}$ that maps any collection of n data points to a "guess" of $g(\theta)$.

If $(X_1, \ldots, X_n)$ are i.i.d. random variables on $\mathcal{X}$ distributed according to $\mathbb{P}_\theta$ and $\mathcal{Y}$ is a measure space, then $T = T(X_1, \ldots, X_n)$ is itself a random variable on $\mathcal{Y}$. We denote by $\mathbb{E}_\theta$ and Var_θ the expectation and variance with respect to the distribution on $\mathcal{Y}$ induced by the distribution $\mathbb{P}_\theta$ on $\mathcal{X}$. The **mean square error (MSE)** of an estimator T is defined as

$$\mathbb{E}_\theta[(T - g(\theta))^2].$$

The **bias** of T is defined as

$$b_T(\theta) = \mathbb{E}_\theta[T] - g(\theta).$$

The **standard error** of T is defined as

$$\sigma_T(\theta) = \sqrt{\mathrm{Var}_\theta(T)}.$$

The bias-variance trade-off can then be described as decomposing the mean square error as a sum of bias and variance:

$$\mathbb{E}_\theta[(T - g(\theta))^2] = b_T(\theta)^2 + \mathrm{Var}_\theta(T).$$

Ideally, we would like the bias and the variance to be both small, but typically this is not possible. An estimator is called **unbiased** if $b_T(\theta) = 0$ for all θ. Variants of the bias-variance trade-off are of fundamental importance in machine learning.

Example 2.1. Consider n coin flips, with outcomes $X = 1$ (head) or $X = 0$ (tail), where the probability of head is p. Assuming we are after p, the estimator is a map $T\colon \{0,1\}^n \to [0,1]$. The bias-variance trade-off is perhaps best illustrated by considering extreme cases. If we fixe some $p_0 \in [0,1]$ and take T to be the constant map, $T(x_1,\dots,x_n) = p_0$, then clearly $\mathrm{Var}_p(T) = 0$, but the bias is $b_T(p) = p_0 - p$. A more principled approach is to choose

$$T(X_1,\dots,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then $\mathbb{E}_p[T] = p$ and the method is therefore unbiased. However, the variance is

$$\mathrm{Var}_p(T) = \frac{p(1-p)}{n} > 0.$$

While the variance is positive, we see in this example that it convergence to 0 if we increase the number of samples n.

Example 2.2. If $X_1,\dots,X_n$ are i.i.d. $\mathcal{N}(\mu,\sigma^2)$ random variables, then

$$T(X_1,\dots,X_n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator of the mean μ, while

$$T(X_1,\dots,X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X - \overline{X}_n)^2$$

is an unbiased estimator of the variance σ^2. Note the unusual choice of normalization: intuitively, one would think of dividing by n instead of $n-1$, but the factor $1/(n-1)$ is essential to make this estimator unbiased.

If we want to evaluate the behaviour of an estimator as the sample size n increases, we need to consider families of estimators $\{T_n\}$. A family of estimators $\{T_n\}$ is called **consistent**, if for all $\epsilon > 0$ and all $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathbb{P}(|T_n - g(\theta)| > \epsilon) = 0.$$

The variance estimator in Example 2.2 is consistent.

Maximum likelihood

Arguably the most important method for constructing an estimator is **maximum likelihood**. Assume the distribution induced by $\mathbb{P}_\theta$ on $\mathcal{X}^n$ has density $\rho_\theta(x_1,\dots,x_n)$. For samples $(x_1,\dots,x_n) \in \mathcal{X}^n$, the **likelihood function** is defined as

$$L\colon \Theta \to \mathbb{R}, \quad L(\theta) = \rho_\theta(x_1,\dots,x_n).$$

The **maximum likelihood estimator** of a parameter θ is then defined as

$$T(x_1, \ldots, x_n) = \arg\max_{\theta} L(\theta).$$

In words, one selects the parameter θ for which the density $\rho_\theta(x_1, \ldots, x_n)$ is the largest, i.e., the density for which the observed data is "most likely". In practice, it is often more convenient to maximize the logarithm of the likelihood, also known as **log-likelihood**.

Example 2.3. Consider the example with n coin flips from Example 2.1. Here we use the probability function $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ instead of the density, which is given by

$$L(p) = p^k(1-p)^{n-k}$$

if k entries are 1 and $n-k$ entries are 0. Maximizing the logarithm of this function gives

$$\arg\max_p \left(k\log(p) + (n-k)\log(1-p)\right) = \frac{k}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

This recovers the consistent estimator from Example 2.1.

Example 2.4. Consider the problem of estimating the mean μ of a normal distribution with unit variance from n samples $x_1, \ldots, x_n$. The likelihood function is then the density of a multivariate normal distribution, and the logarithm of this density is given

$$\log L(\mu) = -\sum_{i=1}^{n}(x_i - \mu)^2 + \text{constant}.$$

Maximizing the log-likelihood is therefore equivalent of solving a linear least squares problem, consisting of minimizing the average squared deviation from the data. The solution to this problem is precisely the sample mean $(1/n)\sum_{i=1}^{n} x_i$.

While in some cases the maximum likelihood estimator can be computed explicitly, in other cases iterative algorithms such as **expectation-maximization (EM)** are used.

Maximum a posteriori estimation

In Bayesian statistics, probability is interpreted as a measure of uncertainty, rather than as a frequency. Moreover, the parameters that enter into a model are interpreted as random quantities in their own right. With this interpretation, densities parametrized by a set Θ are interpreted as conditional densities $\rho(x|\theta)$. Similarly, the **likelihood function** is interpreted as a conditional density

$$\rho(x_1, \ldots, x_n|\theta).$$

The assumed density $g(\theta)$ on Θ is called teh **prior density**. The method of **maximum a posteriori estimation (MAP)** aims to estimate θ my maximizing the **posterior**

density $\rho(\theta|x_1, \ldots, x_n)$, which can be expressed in terms of the likelihood and the prior distribution using Bayes' Theorem. Getting rid of constants that do not influence the resulting maximization problem, we can define the MAP estimator as

$$T_{\mathrm{MAP}}(x_1, \ldots, x_n) = \arg\max_{\theta} \rho(x_1, \ldots, x_n|\theta)g(\theta).$$

We see that the difference to maximum likelihood is seen through the presence of the prior density. As with maximum-likelihood, a variant of the expectation-maximization algorithm can be used to compute the MAP estimator.

Bibliography