Statistics Primer

ELEGANT AND powerful techniques are at the fingertips of statisticians. Although difficult at first, speaking their language can be quite powerful. Probability theory provides a set of tools that can be used to quantify uncertain events. In the context of robotics, probability theory allows us make decisions in the presence of uncertainty caused by phenomena such as noisy sensor data or interaction with unpredictable humans. This section introduces a few fundamental concepts including probability, random variables, distributions, and Gaussian random vectors.

When we talk about probability, we generally talk in terms of *experiments* and *outcomes*. When an experiment is conducted, a single outcome from the set of possible outcomes for that experiment results. For example, an experiment could be flipping a coin and the set of possible outcomes is {heads, tails}. If the experiment were to take a measurement in degrees Kelvin, then the set of possible outcomes would be the interval $[0, \infty)$. An *event* is defined to be a subset of the possible outcomes.

Let $\mathcal{S}$ denote the set of all possible outcomes for a given experiment, and let E be an event, i.e., $E \subset \mathcal{S}$. The *probability* of the event E occurring when the experiment is conducted is denoted $\Pr(E)$. Pr maps $\mathcal{S}$ to the interval $[0, 1]$. In the example of flipping a fair coin, $\Pr(\text{heads}) = 0.5$, $\Pr(\text{tails}) = 0.5$, and $\Pr(\text{heads} \cup \text{tails}) = 1$. In general, the probability must obey certain properties:

1. $0 \leq \Pr(E) \leq 1$ for all $E \subset \mathcal{S}$.

2. $\Pr(\mathcal{S}) = 1$.

3. $\sum_i \Pr(E_i) = Pr(E_1 \cup E_2 \cup \ldots)$ for any countable disjoint collection of sets $E_1, E_2, \ldots$. This property is known as *sigma additivity*. In particular, we have $\sum_{i=1}^{n} \Pr(E_i) = Pr(E_1 \cup E_2 \cup \ldots \cup E_n)$.

4. $\Pr(\emptyset) = 0$.

5. $\Pr(E^c) = 1 - \Pr(E)$, where E^c denotes the complement of E in $\mathcal{S}$.

6. $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$.

Technically, the first three axioms imply the last three.

Events may or may not depend upon each other. If the occurance of E_1 has no effect on E_2, then E_1 and E_2 are *independent;* otherwise they are dependent. We say E_1 and E_2 are independent if $\Pr(E_1 \wedge E_2) = \Pr(E_1) \cdot \Pr(E_2)$. One way to express the dependence of two events is through *conditional probability*. For events E_1 and E_2, $\Pr(E_1 \mid E_2)$ is the *conditional probability* that E_1 occurs given that E_2 occurs. If E_1 and E_2 are independent and $\Pr(E_2) > 0$, then $\Pr(E_1 \mid E_2) = \Pr(\mathrm{E}_1)$. For dependent events, Bayes' rule expresses the relationship between the conditional probabilities for two events, again assuming $\Pr(E_2) > 0$:

$$\Pr(E_1 \mid E_2) = \frac{\Pr(E_2 \mid E_1)\Pr(E_1)}{\Pr(E_2)}.$$

Bayes' rule is a useful formula; it is the foundation of the estimation methods presented in chapter 9.

I.1 Distributions and Densities

Within robotics, a somewhat simplified but nevertheless sufficient model of a *random variable* is a mapping from the set of events to the real line, usually denoted $X : \mathcal{S} \to \mathbb{R}$. A simple example of a random variable is to consider a single coin flip and define $X = 0$ when the outcome is heads and $X = 1$ when the outcome is tails. As another example, consider flipping a fair coin ten times. A random variable can be the number of heads that appeared or the number of times heads appeared sequentially, etc. Random variables are useful because they represent events as real numbers. With real numbers, we can perform calculations and analysis that are difficult or impossible to perform on the abstract events.

A *distribution* is an abstract concept that corresponds to all probability statements that can be made about a random variable. Before we discuss the various distributions used to describe random variables, we first distinguish between contiunous and discrete random variables. A random variable is said to be *discrete* if its range (the

values that it maps to) is a set of discrete points. We call a random variable *continuous* if its range forms a continuum on the real line and, using a term that is defined further below, it possesses a probability density function.

Discrete random variables are commonly described using one of two types of distributions. The first is the *cumulative distribution function* (CDF) which is denoted $F_X(a) = \Pr(X \leq a)$. The second is the *probability mass function* (PMF), which is defined to be $f_X(a) = \Pr(X = a)$.

Continuous random variables are described with two analogous distributions. The first is the cumulative distribution function (figure I.1), which is defined for continuous random variables exactly the same way that it is defined for discrete random variables. The second is the *probability density function* (PDF) (figure I.2), which is denoted f_X and is defined such that

$$\Pr(a \leq X \leq b) = \int_{x=a}^{b} f_X(x)\, dx.$$

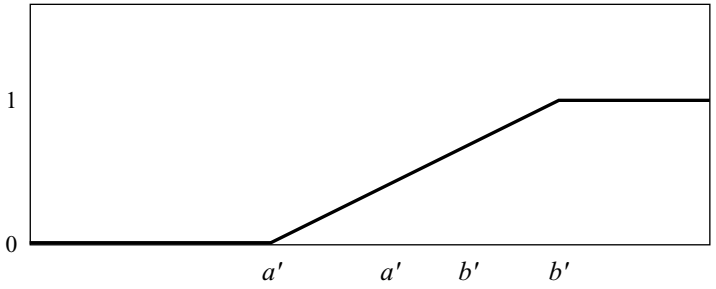

Figure I.1 Cumulative uniform distribution.

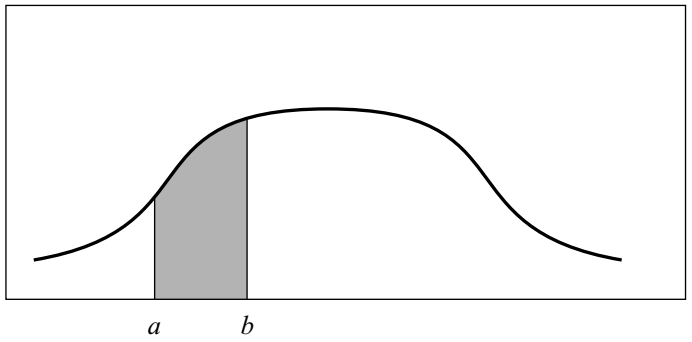

Figure I.2 Probability density function.

Note that for a continuous random variable, $Pr((X = a)) = \int_{x=a}^{a} f_X(x)\,dx = 0$. This can be disconcerting to the newcomer. Another way to view this is: consider the odds of landing exactly on a. Since the point a is a set of measure zero, it should have zero probability of occurring.

Some distributions are so common that they have their own name. For example, the uniform distribution is a family of continuous distributions over an interval. It can either be described by the CDF

$$U(x; a, b) = \begin{array}{ll} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{array}$$

or by the corresponding PDF

$$u(x; a, b) = \begin{array}{ll} 0 & x < a \\ \frac{1}{b-a} & a \leq x \leq b \\ 0 & x \geq b. \end{array}$$

One can calculate all sorts of probabilties using either the CDF or the PDF. Consider for $a', b' \in [a, b]$ with $a' < b'$. Using the CDF we can compute $Pr(a' \leq x \leq b') = U(b'; a, b) - U(a'; a, b)$. Alternatively we can use the PDF to compute $Pr(a' \leq x \leq b') = \int_{x=a'}^{b'} u(x; a, b)\,dx$.

I.2 Expected Values and Covariances

We previously defined a random variable to be a function that maps the event space to the real line. Similarly, we define a *random vector* to be a mapping from the event space to the space of real-valued vectors of some dimension. In other words, a random vector X is a map $X : \mathcal{S} \to \mathbb{R}^n$. Note that a random variable is just a special case of a random vector where $n = 1$.

The *expected value* (or *mean*) for a discrete random vector is defined to be

$$E(X) = \sum_i x_i f_X(x_i),$$

where x_i is the ith value that random variable X can take and f_X is the PMF associated with X. Note that $E(X)$ is a vector in $\mathbb{R}^n$, where n is the dimension of X. It is tempting to think that the expected value is the outcome most likely to occur, but this is not generally the case. The expected value of a single fair die roll is 3.5 which, of course, cannot occur.

The expected value (or mean) of a continuous random vector is defined to be

$$E(X) = \int_{x \in \mathbb{R}^n} x f_X(x)\, dx,$$

where f_X is the PDF associated with X. As in the case of discrete random vectors, $E(X)$ is a vector in $\mathbb{R}^n$. We also denote $E(X)$ with $\bar{X}$. Expectation is a linear operator, which means that $E(aX + bY) = aE(X) + bE(Y)$.

The *variance* of a (scalar) random variable x is $E((X - \bar{X})^2)$. For a scalar random variable the variance is denoted σ^2. For a random vector we can consider the variance of each element X_i of X individually. The variance of X_i is denoted σ_i^2.

Now we want to consider the effect of one variable on another. This is termed *covariance* between two random variables X_i and X_j. Let $\sigma_{ij} = E((X_i - \bar{X}_i)(X_j - \bar{X}_j))$. By this definition σ_{ii} is the same as σ_i^2, the variance of X_i. For $i \neq j$, if $\sigma_{ij} = 0$, then X_i and X_j are independent of each other. The *covariance matrix* of a random vector X is defined to be

$$P_X = E((X - \bar{X})(X - \bar{X})^T).$$

The $n \times n$ matrix P_X contains the variances and covariances within the random vector X. Specifically, the element in the ith row, jth column of P_X will be identical to the σ_{ij} defined above.

I.3 Multivariate Gaussian Distributions

A random vector X is said to have a multivariate Gaussian distribution if it is described by the PDF

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n |P_X|}} e^{-\frac{1}{2}(x-\bar{X})^T P_X^{-1}(x-\bar{X})}, \tag{I.1}$$

where $\bar{X} \in \mathbb{R}^n$ is the mean vector and $P_X \in \mathbb{R}^{n \times n}$ is the covariance matrix. It can be verified by direct substitution that $\bar{X}$ and P_X are in fact the mean and covariance matrix of X as defined in the section above.