Introduction

The first two sections discuss discrete random variables and continuous random variables respectively, mostly focusing on some of the important distributions for each. Section three explores some of the important properties of random variables, including, independence, conditional probability, and Bayes' theorem. This document is designed to be a probability reference, so it does not provide derivations or visual explanations.

1 Discrete Random Variables

Within this section X, Y, Z are discrete random variables and they represent discrete outcomes x, y, z respectively.

1.1 Distributions

Let $x \in S$, where S is a set. A probability distribution $P(X = x)$ must satisfy two requirements:

1. $P(X = x) \geq 0 \ \forall x$
2. $\sum_{x \in S} P(X = x) = 1$

$P(X = x)$ is also called the probability mass function (pmf) $p(x)$ for a discrete random variable X. The relationship is simple: $p(x) = P(X = x)$. Furthermore, the cumulative mass function (cmf) can be determined from the pmf:

$$F(a) = P(X \leq a)$$

1.1.1 Joint Distribution

A joint probability distribution, such as $P(X = x, Y = y)$, represents the probability of X and Y taking on their own outcomes simultaneously. In other words, $P(X = x, Y = y)$ can be used to find the probability of $X = x$ **and** $Y = y$.

1.1.2 Marginal Distribution

A marginal distribution only considers the probability distribution of one random variable X in the presence of other random variables. For two random variables, the marginal distribution for X can be found. Let $y \in T$, where T is a set.

$$P(X = x) = \sum_{y \in T} P(X = x, Y = y)$$

The marginal distribution exists for more than two variables. Let $z \in U$, where U is a set.

$$P(X = x) = \sum_{y \in T} \sum_{z \in U} P(X = x, Y = y, Z = z)$$

1.1.3 Conditional Distribution

The conditional distribution of a random variable gives the probability distribution of a random variable after a different random variable has a known outcome. The distribution is given by

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

1.1.4 Bernoulli Distribution

Define X such that $X = 1$ when an outcome is a success and $X = 0$ when an outcome is a failure. Define $p = P(X = 1)$. Then we say that $X \sim \text{Bernoulli}(p)$.

$$p(x) = \begin{cases} p & \text{if } X = 1 \\ 1 - p & \text{if } X = 0 \end{cases}$$

$$\mathbb{E}[X] = p$$

$$\text{Var}[X] = p(1 - p)$$

1.1.5 Binomial Distribution

Suppose that n independent experiments are performed with either success ($X = 1$) or failure ($X = 0$). Define p to be the probability of success. Then we say that $X \sim \text{Binomial}(n, p)$, and we may observe the probability of k successes.

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

$$\mathbb{E}[X] = np$$

$$\text{Var}[X] = np(1 - p)$$

1.1.6 Geometric Distribution

Define X such that $X = 1$ when an outcome is a success and $X = 0$ when an outcome is a failure. Define $p = P(X = 1)$. The geometric distribution observes the number of Bernoulli trials k until the first success. We say that $X \sim \text{Geometric}(p)$.

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

$$\mathbb{E}[X] = \frac{1}{p}$$

$$\text{Var}[X] = \frac{1 - p}{p^2}$$

1.1.7 Poisson Distribution

Define X such that $X = k$ for $k \in \{0, 1, 2, 3, \dots\}$. For some $\lambda > 0$, we may say that $X \sim \text{Poisson}(\lambda)$.

$$p(k) = \lambda^k \frac{e^{-\lambda}}{k!}$$
$$\mathbb{E}[X] = \lambda$$
$$\text{Var}[X] = \lambda$$

For large n and small p, the Poisson distribution where $\lambda = n \cdot p$ is a good approximation to the Binomial distribution.

1.1.8 Multinomial Distribution

The multinomial distribution generalizes the binomial distribution. Instead of a binary success or failure, there are k outcomes. n is the number of independent experiments. The pmf $p(\mathbf{x})$ accepts a vector $\mathbf{x}$ of possible outcomes, and each element in $\mathbf{x}$, x_i, has probability p_i of occurring.

$$p(\mathbf{x}) = \frac{n!}{\prod_{i=1}^{n}(x_i!)} \prod_{i=1}^{n} p_i^{x_i}$$
$$\mathbb{E}[X_i] = np_i$$
$$\text{Var}[X_i] = np_i(1 - p_i)$$

2 Continuous Random Variables

Within this section X, Y, Z are continuous random variables and they represent continuous outcomes x, y, z respectively. Some of the distributions for discrete random variables apply to continuous random variables; these distributions will be adjusted appropriately and repeated in this section.

2.1 Distributions

Let A be some continuous set of numbers, then $P(X \in A) = \int_A f_X(x)dx$, where $f_X(x)$ is the probability density function (pdf) of X. The pdf has three properties (assuming $A \in \mathbb{R}$):

1. $f(x) \geq 0 \ \forall x$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$

The third property is the definition of the cumulative density function (cdf), given by $F(x)$, and has the following relationship to the pdf:

$$f(x) = \frac{dF(x)}{dx}$$

Finally, it is important to note that the pdf is not a probability, and therefore $f(x)$ may be greater than one for some x.

2.1.1 Joint Distribution

A joint probability distribution, such as $f_{XY}(x, y)$, represents the probability of X and Y taking on their own outcomes simultaneously. In other words, $f_{XY}(x, y)$ can be used to find the probability of $X = x$ **and** $Y = y$.

2.1.2 Marginal Distribution

A marginal distribution only considers the probability distribution of one random variable X in the presence of other random variables. For two random variables, the marginal distribution for X can be found. Let A be the support of Y, then

$$f_X(x) = \int_A f_{XY}(x, y)dy$$

The marginal distribution exists for more than two variables. Let A be the support of Y and B be the support of Z. Then,

$$f_X(x) = \int_A \int_B f_{XYZ}(x, y, z)dydz$$

2.1.3 Conditional Distribution

The conditional distribution of a random variable gives the probability distribution of a random variable after a different random variable has a known outcome. The distribution is given by

$$f(x \mid y) = \frac{f(x, y)}{f_Y(y)}$$

2.1.4 Uniform Distribution

$X \sim \text{Uniform}(\alpha, \beta)$ if, on the interval $[\alpha, \beta]$,

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha \leq x \leq \beta \\ 0 & x < \alpha, x > \beta \end{cases}$$

$$\mathbb{E}[X] = \frac{\alpha + \beta}{2}$$

$$\text{Var}[X] = \frac{(\beta - \alpha)^2}{12}$$

$$F(a) = \frac{a - \alpha}{\beta - \alpha}$$

2.1.5 Normal (Gaussian) Distribution

If X is distributed normally, then $X \sim \mathcal{N}(\mu, \sigma^2)$.

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

$$\mathbb{E}[X] = \mu$$

$$\text{Var}[X] = \sigma^2$$

$\mathcal{N}(0, 1)$ is known as the standard normal. Note that $F(a)$ does not exist in a closed form solution.

2.1.6 Exponential Distribution

$X \sim \text{Exponential}(\lambda)$ if, given parameter $\lambda > 0$.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

$$\text{Var}[X] = \frac{1}{\lambda^2}$$

$$F(a) = \begin{cases} 1 - e^{-\lambda a} & a \geq 0 \\ 0 & a < 0 \end{cases}$$

3 Properties of Random Variables

This section will present other important properties of random variables. The notation for discrete random variables will be used here, but the principles apply to continuous random variables. In addition, shorthand will be used: $P(X = x)$ is abbreviated to $P(x)$ for brevity.

3.1 Independence

If either of the following two criteria are met $\forall\ x, y, z$ such that $X = x$, $Y = y$, and $Z = z$, then the two random variables X and Y are independent.

1. $P(x, y) = P(x)P(y)$
2. $P(x, y \mid z) = P(x \mid z)P(y \mid z)$ (conditional independence between X and Y)

3.2 More Conditional Probability

3.2.1 Product Rule

The product rule can be obtained by rearranging the conditional probability from 1.1.3. Let $X = x$ and $Y = y$.

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

$$P(x \mid y)P(y) = P(x, y) = P(y \mid x)P(x)$$

3.2.2 Chain Rule

The chain rule is essentially a generalized form of the product rule. Consider $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

$$P(x_1, ..., x_n) = P(x_n \mid x_{n-1}, ..., x_1)P(x_{n-1} \mid x_{n-2}, ..., x_1)...P(x_2 \mid x_1)P(x_1)$$

$$= \prod_i P(x_i \mid x_{i-1}, ..., x_1)$$

3.2.3 Law of Total Probability

Consider the discrete random variable X which represents discrete outcomes $x \in S$, the discrete random variable $Y = y$.

$$P(y) = \sum_{x \in S} P(x, y)$$

Using the product rule,

$$P(y) = \sum_{x \in S} P(y \mid x) P(x)$$

3.2.4 Bayes' Theorem

Given $X = x$ and $Y = y$, Bayes' Theorem can be determined by equating both sides of the product rule.

$$P(x \mid y)P(y) = P(x, y) = P(y \mid x)P(x)$$

$$P(x \mid y)P(y) = P(y \mid x)P(x)$$

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}$$

In this case, $P(x \mid y)$ is the *posterior*, $P(y \mid x)$ is the *likelihood*, $P(x)$ is the *prior*, and $P(y)$ is the *normalization* term. Using the law of total probability, the normalization term can be substituted for:

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{\sum_{x \in S} P(y \mid x)P(x)}$$

3.3 Union and Intersection of Events

The probability of the union of two events $X = x$ and $Y = y$ is given by:

$$P(x \cup y) = P(x) + P(y) - P(x \cap y)$$

The probability of the intersection of two events $X = x$ and $Y = y$ is given by:

$$P(x \cap y) = P(x) + P(y) - P(x \cup y)$$

3.4 Expected Value

The expected value, or mean, of a discrete random variable is given by

$$\mathbb{E}[X] = \sum_{x \in S} x \cdot p(x)$$

where $p(x)$ is the pmf of X. If $p(x)$ is the same for all x and there are n possible outcomes, then the expectation of X may be written as

$$\mathbb{E}[X] = \frac{1}{n} \sum_{i=1}^{n} x$$

3.5 Variance

Let $\mu = \mathbb{E}[X]$. The variance σ^2 of X is given by

$$\sigma^2 = \mathbb{E}[(X-\mu)^2] = \sum_{x \in S}(x-\mu)^2 \cdot p(x)$$

Alternatively,

$$\sigma^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \left(\sum_{x \in S} x^2 \cdot p(x)\right) - \left(\sum_{x \in S} x \cdot p(x)\right)^2$$

The standard deviation is $\sqrt{\sigma^2}$.

3.6 Covariance

Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$, and let n be the number of outcomes for X and Y. The covariance $\text{Cov}(X, Y)$ is given by

$$\text{Cov}(X,Y) = \mathbb{E}(X-\mu_X)]E[(Y-\mu_Y)] = \sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y) \cdot p(x_i, y_i)$$

Note that $\text{Cov}(X, X) = \text{Var}[X]$

3.7 Correlation

The correlation of X and Y is a value between -1 and 1. A correlation of -1 indicates that the random variables are perfectly inversely related, 0 indicates no relationship, and 1 indicates that the two variables vary together. The correlation $\text{Cor}(X, Y)$ is given by

$$\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\text{Var}[X]\text{Var}[Y]}$$

4 MLE and MAP

4.1 Maximum Likelihood Estimation (MLE)

Given a probability distribution $P(X \mid \theta)$, it's often useful to estimate some parameter θ in order to maximize the probabilities of n independent samples $x_1, x_2, ..., x_n$, from some distribution. The likelihood function is defined as

$$L(\theta \mid X) = P(X \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta)$$

Since $P(x_i \mid \theta)$ is always between 0 and 1, this product often results in a very small number. As a result, it is more convenient to consider the loglikelihood:

$$l(\theta \mid X) = \log\left(\prod_{i=1}^{n} P(x_i \mid \theta)\right) = \sum_{i=1}^{n} \log(P(x_i \mid \theta))$$

Using these equations, θ may be computed

$$\theta = \arg\max_{\theta}\ l(\theta \mid X) = \arg\max_{\theta} \sum_{i=1}^{n} \log(P(x_i \mid \theta))$$

4.2 Maximum A Posteriori (MAP) Estimation

MAP Estimation is nearly identical to MLE, except that it utilizes the prior from Bayes' Theorem.

$$L(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)}$$

This is the posterior distribution, and is the thing that needs to be maximized over θ. Since $P(X)$ is just a constant, it is not necessary to include for the optimization calculation.

$$\theta = \arg\max_{\theta} \log\left(\prod_{i=1}^{n} P(x_i \mid \theta)P(\theta)\right) = \arg\max_{\theta} \sum_{i=1}^{n} \log(P(x_i \mid \theta)P(\theta))$$