Events and Probability

Event

A set of outcomes from a random experiment.

Sample Space

Set of all possible outcomes Ω.

Intersection

Outcomes occur in both A and B

$$A \cap B \quad \text{or} \quad AB$$

Disjoint

Two events cannot occur simultaneously or have no common outcomes

$$AB = \varnothing$$

These events are dependent.

Union

Set of outcomes in either A or B

$$A \cup B$$

Complement

Set of all outcomes not in A, but in Ω — $\overline{A} = \Omega \backslash A$.

$$A\overline{A} = \varnothing$$
$$A \cup \overline{A} = \Omega$$

Subset

A is a (non-strict) subset of B if all elements in A are also in B — $A \subset B$.

$$AB = A \quad \text{and} \quad A \cup B = B$$

$$\forall A : A \subset \Omega \wedge \varnothing \subset A$$

Identities

$$A(BC) = (AB)C$$
$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A(B \cup C) = AB \cup AC$$
$$A \cup BC = (A \cup B)(A \cup C)$$

Probability

Measure of the likeliness of an event occurring

$$\Pr(A) \quad \text{or} \quad \mathrm{P}(A)$$

$$0 \le \Pr(E) \le 1$$

where a probability of 0 never happens, and 1 always happens.

$$\Pr(\Omega) = 1$$
$$\Pr\left(\overline{E}\right) = 1 - \Pr(E)$$

Multiplication Rule

For independent events A and B

$$\Pr(AB) = \Pr(A)\Pr(B).$$

For dependent events A and B

$$\Pr(AB) = \Pr(A \mid B)\Pr(B)$$

Addition Rule

For independent A and B

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(AB).$$

If $AB = \varnothing$, then $\Pr(AB) = 0$, so that $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

De Morgan's Laws

$$\overline{A \cup B} = \overline{A}\,\overline{B}$$
$$\overline{AB} = \overline{A} \cup \overline{B}.$$
$$\Pr(A \cup B) = 1 - \Pr\left(\overline{A}\,\overline{B}\right)$$
$$\Pr(AB) = 1 - \Pr\left(\overline{A} \cup \overline{B}\right)$$

Conditional probability

The probability of event A given B has already occurred

$$\Pr(A \mid B) = \frac{\Pr(AB)}{\Pr(B)}.$$

A and B are independent events if

$$\Pr(A \mid B) = \Pr(A)$$
$$\Pr(B \mid A) = \Pr(B)$$

the following statements are also true

$$\Pr\left(A \mid \overline{B}\right) = \Pr(A)$$
$$\Pr\left(\overline{A} \mid B\right) = \Pr\left(\overline{A}\right)$$
$$\Pr\left(\overline{A} \mid \overline{B}\right) = \Pr\left(\overline{A}\right)$$

Probability Rules with Conditional

All probability rules hold when conditioning on another event C.

$$\Pr\left(\overline{A} \mid C\right) = 1 - \Pr(A \mid C)$$
$$\Pr(A \cup B \mid C) = \Pr(A \mid C) + \Pr(B \mid C) - \Pr(AB \mid C)$$
$$\Pr(AB \mid C) = \Pr(A \mid BC)\Pr(B \mid C)$$

Conditional Independence

Given $\Pr(A \mid B) \ne \Pr(A)$ A and B are conditionally dependent given C if

$$\Pr(A \mid BC) = \Pr(A \mid C).$$

Futhermore

$$\Pr(AB \mid C) = \Pr(A \mid C)\Pr(B \mid C).$$

Conversely

$$\Pr(A \mid B) = \Pr(A)$$
$$\Pr(A \mid BC) \ne \Pr(A \mid C)$$
$$\Pr(AB \mid C) = \Pr(A \mid BC)\Pr(B \mid C)$$

Pairwise independence does not imply mutual independence

$$\begin{cases} \Pr(AB) = \Pr(A)\Pr(B) \\ \Pr(AC) = \Pr(A)\Pr(C) \\ \Pr(BC) = \Pr(B)\Pr(C) \end{cases} \not\Rightarrow$$
$$\Pr(ABC) = \Pr(A)\Pr(B)\Pr(C).$$

Independence should not be assumed unless explicitly stated.

Disjoint Events

Given $AB = \varnothing$

$$\Pr(AB) = 0 \implies \Pr(\varnothing) = 0$$
$$\Pr(A \mid B) = 0$$

Subsets

If $A \subset B$ then $\Pr(A) \le \Pr(B)$.

$$\Pr(B \mid A) = 1$$
$$\Pr(A \mid B) = \frac{\Pr(A)}{\Pr(B)}$$

These events are also highly dependent.

Marginal Probability

The probability of an event irrespective of the outcome of another variable.

Total Probability

$A = AB \cup A\overline{B}$

$$\Pr(A) = \Pr(AB) + \Pr\left(A\overline{B}\right)$$
$$\Pr(A) = \Pr(A \mid B)\Pr(B) + \Pr\left(A \mid \overline{B}\right)\Pr\left(\overline{B}\right)$$

In general, partition Ω into disjoint events $B_1, B_2, \ldots, B_n$, such that $\bigcup_{i=1}^{n} B_i = \Omega$

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i)$$

Bayes' Theorem

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$

Combinatorics

Number of outcomes

Let $|A|$ denote the number of outcomes in an event A.
For k disjoint events $\{S_1, \ldots, S_k\}$ where the ith event has n_i possible outcomes,

Addition principle

Number of possible samples from any event

$$\left|\bigcup_{i=0}^{k} S_i\right| = \sum_{i=1}^{k} n_i$$

Multiplication principle

Number of possible samples from every event

$$\left|\bigcap_{i=0}^{k} S_i\right| = \prod_{i=1}^{k} n_i$$

Counting probability

If S_i has equally likely outcomes

$$\Pr(S_i) = \frac{|S_i|}{|S|}$$

Ordered Sampling with Replacement

Number of ways to choose k objects from a set with n elements

$$n^k$$

Ordered Sampling without Replacement

Number of ways to arrange k objects from a set of n elements, or the k-permutation of n-elements

$${}^{n}P_k = \frac{n!}{(n-k)!}$$

for $0 \le k \le n$.
An n-permutation of n elements is the permutation of those elements. In this case, $k = n$, so that

$${}^{n}P_n = n!$$

Unordered Sampling without Replacement

Number of ways to choose k objects from a set of n elements, or the k-combination of n-elements

$$^nC_k = \frac{^nP_k}{k!} = \frac{n!}{k!\,(n-k)!}$$

for $0 \le k \le n$.

Unordered Sampling with Replacement

Number of ways to choose k objects from a set with n elements

$$\binom{n+k-1}{k}$$

Random Variables

A measurable variable whose value holds some uncertainty. An event is when a random variable assumes a certain value or range of values.

Discrete random variables

A discrete random variable takes discrete values.

Continuous random variables

A continuous random variable can take any real value.

Probability Distributions

Probability distribution

The probability distribution of a random variable X is a function that links all outcomes $x \in \Omega$ to the probability that they will occur $\Pr(X = x)$.

Probability mass function

The probability distribution of a discrete random variable X is described by a Probability Mass Function (PMF) p_x.

$$\Pr(X = x) = p_x$$

p_x is a valid PMF provided,

$$\forall x \in \Omega : \Pr(X = x) \ge 0 \quad \text{and} \quad \sum_{x\in\Omega} \Pr(X = x) = 1.$$

Probability density function

The probability distribution of a continuous random variable X is described by a Probability Density Function (PDF) $f(x)$.

The probability that X is exactly equal to a specific value is always 0. Therefore we compute probabilities over intervals:

$$\Pr(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(x)\,\mathrm{d}x$$

$f(x)$ is a valid PDF provided,

$$\forall x \in \Omega : f(x) \ge 0 \quad \text{and} \quad \int_\Omega f(x)\,\mathrm{d}x = 1.$$

Cumulative distribution function

The Cumulative Distribution Function (CDF) computes the probability that the random variable is less than or equal to a particular realisation x. For $U = \{k \in \Omega : k \le x\}$

$$F(x) = \Pr(X \le x) = \begin{cases} \sum_{u\in U} p_u & \text{for discrete random variables} \\ \int_U f(u)\,\mathrm{d}u & \text{for continuous random variables.} \end{cases}$$

$F(x)$ is a valid CDF if:

1. F is monotonically increasing and continuous
2. $\lim_{x\to-\infty} F(x) = 0$
3. $\lim_{x\to\infty} F(x) = 1$

We can recover the PDF given the CDF, by using the Fundamental Theorem of Calculus.

$$\frac{\mathrm{d}F(x)}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}\int_{-\infty}^{x} f(u)\,\mathrm{d}u = f(x)$$

Complementary CDF

For a continuous random variable X the complement function,

$\Pr(X > x) = 1 - \Pr(X \le x) = 1 - F(x)$

is called the complementary CDF, or the survival function.

Quantiles

p-Quantile

For a continuous random variable, the p-quantile, x, is defined such that

$$F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u = p.$$

Median

The median, m, is a special p-quartile defined as the value such that

$$\int_{-\infty}^{m} f(u)\,\mathrm{d}u = \int_{m}^{\infty} f(u)\,\mathrm{d}u = \frac{1}{2}.$$

Lower and upper quartile

Likewise the lower quartile and upper quartiles are two values q_1 and q_2 such that

$$\int_{-\infty}^{q_1} f(u)\,\mathrm{d}u = \frac{1}{4}$$

and

$$\int_{-\infty}^{q_2} f(u)\,\mathrm{d}u = \frac{3}{4}.$$

Quantile function

The quantile function is the inverse of the CDF and can be used to find the x that a certain p provides. I.e.,

$$x = F^{-1}(p) = Q(p)$$

Summary Statistics

Expectation

The expectation $\mathrm{E}(X)$, or $\mathbb{E}(X)$ of a random variable X is its expected value given an infinite number of observations. The expectation is also known as the mean of the X, denoted μ.

$$\mathrm{E}(X) = \begin{cases} \sum_{x\in\Omega} x p_x & \text{for discrete variables} \\ \int_\Omega x f(x)\,\mathrm{d}x & \text{for continuous variables} \end{cases}$$

Theorem 3.15.1. *Using integration by parts, it can be proved that*

$$\mathrm{E}(X) = -\int_{-\infty}^{0} F(x)\,\mathrm{d}x + \int_{0}^{\infty} (1 - F(x))\,\mathrm{d}x$$

Variance

The variance $\mathrm{Var}(X)$, or $\mathbb{V}(X)$ of a random variable X is a measure of spread of the distribution (defined as the average squared distance of each value from the mean). Variance is also denoted as σ^2.

$$\begin{aligned} \mathrm{Var}(X) &= \begin{cases} \sum_{x\in\Omega} (x-\mu)^2 p_x & \text{for discrete v} \\ \int_\Omega (x-\mu)^2 f(x)\,\mathrm{d}x & \text{for continuou} \end{cases} \\ &= \mathrm{E}(X^2) - \mathrm{E}(X)^2 \end{aligned}$$

Standard deviation

The standard deviation is defined as

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

3.17.1 Transformations

For a simple linear function of a random variable

$$\mathrm{E}(aX \pm b) = a\,\mathrm{E}(X) \pm b$$

$$\mathrm{Var}(aX \pm b) = a^2\,\mathrm{Var}(X)$$

Special Discrete Distributions

Discrete Uniform Distribution

A discrete uniform distribution describes the probability distribution of a single trial in a set of equally likely elements.

A discrete random variable X with a discrete uniform distribution is denoted

$$X \sim \mathrm{Uniform}(a,\, b)$$

with

$$\Pr(X = x) = \frac{1}{b-a+1}$$

$$\Pr(X \le x) = \frac{x-a+1}{b-a+1}$$

for outcomes $x \in \{a,\, a+1,\, \ldots,\, b-1,\, b\}$.

We can also summarise the following:

$$\mathrm{E}(X) = \frac{a+b}{2}$$

$$\mathrm{Var}(X) = \frac{(b-a+1)^2 - 1}{12}$$

Bernoulli Distribution

A Bernoulli (or binary) distribution describes the probability distribution of a Boolean-valued outcome, i.e., success (1) or failure (0).

A discrete random variable X with a Bernoulli distribution is denoted

$$X \sim \mathrm{Bernoulli}(p)$$

with

$$\begin{aligned} \Pr(X = x) &= \begin{cases} 1-p & x = 0 \\ p & x = 1 \end{cases} \\ &= p^x (1-p)^{1-x} \end{aligned}$$

$$\Pr(X \le x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \le x < 1 \\ 1 & k \ge 1 \end{cases}$$

for a probability $p \in [0, 1]$ and outcomes $x \in \{0,\, 1\}$. We can also summarise the following:

$$\mathrm{E}(X) = p$$
$$\mathrm{Var}(X) = p(1-p)$$

where $(1-p)$ is sometimes denoted as q.

Binomial Distribution

A binomial distribution describes the probability distribution of the number of successes for n independent trials with the same probability of success p.
A discrete random variable X with a binomial distribution is denoted

$$X \sim \mathrm{B}(n,\, p)$$

with

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$\Pr(X \leq x) = \sum_{u=0}^{x} \binom{n}{u} p^u (1-p)^{n-u}$$

for number of successes $x \in \{0,\, 1,\, \dots,\, n\}$.
Here each individual trial is a Bernoulli trial, so that X can be written as the sum of n *independent and identically distributed* (iid) Bernoulli random variables, $Y_1,\, Y_2,\, \dots,\, Y_n$.

$$X = Y_1 + Y_2 + \cdots + Y_n$$
$$Y_i \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(p) : \forall i \in \{1,\, 2,\, \dots,\, n\}.$$

We can then summarise the following:

$$\mathrm{E}(X) = np$$
$$\mathrm{Var}(X) = np(1-p)$$

Geometric Distribution

A geometric distribution describes the probability distribution of the number of trials up to and including the first success, where each trial is independent and has the same probability of success p.
A discrete random variable N with a geometric distribution is denoted

$$N \sim \mathrm{Geom}(p)$$

with

$$\Pr(N = n) = (1-p)^{n-1} p$$
$$\Pr(N \leq n) = 1 - (1-p)^n$$

for number of trials $n \geq 1$.
We can also summarise the following:

$$\mathrm{E}(N) = \frac{1}{p}$$
$$\mathrm{Var}(N) = \frac{1-p}{p^2}$$

4.4.1 Alternate Geometric Definition

We can alternatively consider the number of failures until a success, Y:

$$Y = N - 1$$

Therefore the PMF and CDF for Y are:

$$\Pr(Y = y) = (1-p)^y p$$
$$\Pr(Y \leq y) = 1 - (1-p)^{y+1}$$

for number of failures $y \geq 0$. This gives the following summary statistics using transformation rules:

$$\mathrm{E}(Y) = \frac{1-p}{p}$$
$$\mathrm{Var}(Y) = \frac{1-p}{p^2}$$

Negative Binomial Distribution

A negative binomial distribution describes the probability distribution of the number of trials until $k \geq 1$ successes, where each trial is independent and has the same probability of success p.
A discrete random variable N with a negative binomial distribution is denoted

$$N \sim \mathrm{NB}(k,\, p)$$

with

$$\Pr(N = n) = \binom{n-1}{k-1} (1-p)^{n-k} p^k$$
$$\Pr(N \leq n) = \sum_{u=k}^{n} \binom{u-1}{k-1} (1-p)^{u-k} p^k$$

for number of trials $n \geq k$. Here each individual trial is a Geometric trial, so that N can be written as the sum of k *independent and identically distributed* (iid) Geometric random variables, $Y_1,\, Y_2,\, \dots,\, Y_k$.

$$N = Y_1 + Y_2 + \cdots + Y_k, \quad Y_i \overset{\text{iid}}{\sim} \mathrm{Geom}(p) : \forall i \in \{1,\, 2,\, \dots,\, k\}.$$

We can then summarise the following:

$$\mathrm{E}(N) = \frac{k}{p}$$
$$\mathrm{Var}(N) = \frac{k(1-p)}{p^2}$$

4.5.1 Alternate Negative Binomial Definition

We can alternatively consider the number of failures Y until k successes:

$$Y = N - k$$

The PMF and CDF for Y are given by:

$$\Pr(Y = y) = \binom{y+k-1}{k-1} (1-p)^y p^k$$
$$\Pr(Y \leq y) = \sum_{u=0}^{y} \binom{u+k-1}{k-1} (1-p)^u p^k$$

for number of failures $y \geq 0$. This gives the following summary statistics using transformation rules:

$$\mathrm{E}(Y) = \frac{k(1-p)}{p}$$
$$\mathrm{Var}(Y) = \frac{k(1-p)}{p^2}$$

Poisson Distribution

A Poisson distribution describes the probability distribution of the number of events N which occur over a fixed interval of time λ.
A discrete random variable N with a Poisson distribution is denoted

$$N \sim \mathrm{Pois}(\lambda)$$

with

$$\Pr(N = n) = \frac{\lambda^n e^{-\lambda}}{n!}$$
$$\Pr(N \leq n) = e^{-\lambda} \sum_{u=0}^{n} \frac{\lambda^u}{u!}$$

for number of events $n \geq 0$. We can also summarise the following:

$$\mathrm{E}(N) = \lambda$$
$$\mathrm{Var}(N) = \lambda$$

Modelling Count Data

If we want to utilise these distributions to model data, we can use the following observations:

- Poisson (mean = variance)
- Binomial (underdispersed, mean > variance)
- Geometric/Negative Binomial (overdispersed, mean < variance)

Special Continuous Distributions

Continuous Uniform Distribution

A continuous uniform distribution describes the probability distribution of an outcome within some interval, where the probability of an outcome in one interval is the same as all other intervals of the same length.
A continuous random variable X with a continuous uniform distribution is denoted

$$X \sim \mathrm{U}(a,\, b)$$

with

$$f(x) = \frac{1}{b-a}$$
$$F(x) = \frac{x-a}{b-a}$$

for outcomes $a < x < b$. We can also summarise the following:

$$\mathrm{E}(X) = \frac{a+b}{2}$$
$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$
$$m = \frac{a+b}{2}$$

Exponential Distribution

An exponential distribution describes the probability distribution of the time between events with rate η.
A continuous random variable T with an exponential distribution is denoted

$$T \sim \mathrm{Exp}(\eta)$$

with

$$f(t) = \eta e^{-\eta t}$$
$$F(t) = 1 - e^{-\eta t}$$

for time $t > 0$. We can also summarise

the following:

$$E(X) = \frac{1}{\eta}$$
$$\mathrm{Var}(X) = \frac{1}{12}$$
$$m = \frac{\ln(2)}{\eta}$$

Memoryless Property

In an exponential distribution with $T \sim \mathrm{Exp}(\eta)$, the distribution of the waiting time $t + s$ until a certain event does not depend on how much time t has already passed.

$$\Pr(T > s + t \,|\, T > t) = \Pr(T > s).$$

The same property also applies in an Geometric distribution with $N \sim \mathrm{Geom}(p)$.

Normal Distribution

The normal distribution is used to represent many random situations, in particular, measurements and their errors. This distribution arises in many statistical problems and can be used to approximate other distributions under certain conditions.

A continuous random variable X with a normal distribution is denoted

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

with

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
$$F(t) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

for $x \in \mathbb{R}$ where $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \, \mathrm{d}t$ is the error function. We can also summarise the following:

$$E(X) = \mu$$
$$\mathrm{Var}(X) = \sigma^2$$

Given the complexity of the analytic expressions for the PDF and CDF of the normal distribution, we often use software to numerically determine probabilities associated with normal distributions.

Standard Normal Distribution

Given $X \sim \mathrm{N}(\mu, \sigma^2)$, consider the transformation

$$Z = \frac{X - \mu}{\sigma}$$

so that $Z \sim \mathrm{N}(0, 1)$. This distribution is called the standard normal distribution. This allows us to deal with the standard normal distribution regardless of μ and σ.

Central Limit Theorem

The central limit theorem states that the sum of independent and identically distributed random variables, when properly standardised, can be approximated by a normal distribution, as the number of elements increases.

Approximating the Average of Random Variables

Given a set of independent and identically distributed random variables $X_1, \ldots, X_n$ from the distribution X, if $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then we can define $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ so that

$$E(\overline{X}) = \mu$$
$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

By standardising $\overline{X}$, we can define

$$Z = \lim_{n\to\infty} \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

so that $Z \to \mathrm{N}(0, 1)$ as $n \to \infty$.

Approximating the Sum of Random Variables

Given a set of independent and identically distributed random variables $X_1, \ldots, X_n$ from the distribution X, if $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then we can define $\overline{Y} = \sum_{i=1}^n X_i$ so that

$$E(Y) = n\mu$$
$$\mathrm{Var}(Y) = n\sigma^2$$

Then for large n

$$Y \sim \mathrm{N}(n\mu, n\sigma^2)$$

Approximating the Binomial Distribution

6.3.1 Normal Distribution

Given a binomial distribution $X \sim \mathrm{B}(n, p)$, we can write X as the sum of n independent and identically distributed Bernoulli random variables $X_1, \ldots, X_n$, so that $X_i \sim \mathrm{Bernoulli}(p)$.

Thus by the central limit theorem, we can use a normal approximation for X, provided that n is large.

$$X \approx Y \sim \mathrm{N}(np, np(1-p))$$

In general, this approximation is sufficient when $np > 5$ and $n(1-p) > 5$.

6.3.2 Poisson Distribution

When $np < 5$ we can use the Poisson distribution to approximate X with the mean np:

$$X \approx Y \sim \mathrm{Pois}(np).$$

When $n(1-p) < 5$ we can consider the number of failures $W = n - X$, so that,

$$W \approx Y \sim \mathrm{Pois}(n(1-p)).$$

6.3.3 Continuity Correction

Given an approximation Y (either Normal or Poisson) for the binomial distribution $X \sim \mathrm{B}(n, p)$ the equality

$$\Pr(X \le x) = \Pr(X < x + 1)$$

must hold for any x. Therefore by adding $\frac{1}{2}$ we apply a continuity correction to the approximate probability:

$$\Pr\left(Y \le x + \frac{1}{2}\right).$$

Approximating a Poisson Distribution

Given a set of independent Poisson distributions $X_1, \ldots, X_n$ where $X_i \sim \mathrm{Pois}(\lambda)$ so that $E(X_i) = \lambda$ and $\mathrm{Var}(X_i) = \lambda$ for all i.

If we consider $X = \sum_{i=1}^n X_i$ then

$$E(X) = n\lambda$$
$$\mathrm{Var}(X) = n\lambda$$

so that by the central limit theorem, we can use the approximation

$$X \approx Y \sim \mathrm{N}(n\lambda, n\lambda).$$

In general, this approximation is sufficient when $n\lambda > 10$, and when an accurate approximation is desired, $n\lambda > 20$.

Bivariate Distributions

Bivariate probability mass function

The distribution over the joint space of two discrete random variables X and Y is given by a bivariate probability mass function:

$$\Pr(X = x, Y = y) = p_{x,\,y}$$

for all pairs of $x \in \Omega_1$ and $y \in \Omega_2$. This function must satisfy

$$\forall x \in \Omega_1 : \forall y \in \Omega_2 : \Pr(X = x, Y = y) \ge 0 \quad \text{ar}$$

The joint probability mass function can be shown using a table:

	y_1	$\cdots$	y_n
x_1	$\Pr(X = x_1, Y = y_1)$	$\cdots$	$\Pr(X = x_1, Y = y_n$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_n	$\Pr(X = x_n, Y = y_1)$	$\cdots$	$\Pr(X = x_n, Y = y_n$

Bivariate probability density function

The distribution over the joint space of two continuous random variables X and Y is given by a bivariate probability density function $f(x, y)$ over the intervals $x \in \Omega_1$ and $y \in \Omega_2$.

$$\Pr(x_1 \le X \le x_2, y_1 \le Y \le y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x$$

This function must satisfy

$$\forall x \in \Omega_1 : \forall y \in \Omega_2 : f(x, y) \ge 0 \quad \text{and} \quad \int_{x\in\Omega_1}$$

Marginal Probability

The marginal probability function can be obtained by calculating the probability function of each random variable. Once the function has been determined, we must specify the range of values that variable can take.

Marginal probability mass function

$$\Pr(X = x) = p_x = \sum_{y\in\Omega_2} \Pr(X = x, Y = y)$$
$$\Pr(Y = y) = p_y = \sum_{x\in\Omega_1} \Pr(X = x, Y = y)$$

Marginal probability density function

$$\Pr(X=x) = f(x) = \int_{y_1}^{y_2} f(x,\, y)\,\mathrm{d}y$$

$$\Pr(Y=y) = f(y) = \int_{x_1}^{x_2} f(x,\, y)\,\mathrm{d}x$$

Conditional Probability

Using the joint probability and marginal probability, we can determine the conditional probability function. Once the function has been determined, we must specify the range of values that variable can take.

Conditional probability mass function

$$\Pr(X=x \,|\, Y=y) = \frac{\Pr(X=x,\, Y=y)}{\Pr(Y=y)}.$$

It follows that

$$\sum_{x\in\Omega_1} \Pr(X=x \,|\, Y=y) = 1$$

Conditional probability density function

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

It follows that

$$\int_{x_1}^{x_2} f(x \,|\, y)\,\mathrm{d}x = 1$$

Independence

Two discrete random variables X and Y are independent if

$$\Pr(X=x \,|\, Y=y) = \Pr(X=x)$$

for all pairs of x and y. From this we can show that

$$\Pr(X=x,\, Y=y) = \Pr(X=x)\Pr(Y=y)$$

for all pairs of x and y. If these random variables are not independent then,

$$\Pr(X=x,\, Y=y) = \Pr(X=x \,|\, Y=y)\Pr(Y=y)$$

To continuous random variables X and Y are independent if we can express $f(x,\, y)$ as

$$f(x,\, y) \propto g(x)\, h(y)$$

and if the joint range of X and Y do not depend on each other. This leads to

$$f(x \,|\, y) = f(x).$$

Law of Total Expectation

Given the conditional distribution of $X \,|\, Y = y$, we can compute its expectation and variance. For discrete random variables, the conditional expectation is

$$\mathrm{E}(X \,|\, Y=y) = \sum_{x\in\Omega_1} x \Pr(X=x \,|\, Y=y)$$

For continuous random variables, the conditional expectation is

$$\mathrm{E}(X \,|\, Y=y) = \int_{x_1}^{x_2} x f(x \,|\, y)\,\mathrm{d}x$$

The conditional variance is given by

$$\mathrm{Var}(X \,|\, Y=y) = \mathrm{E}(X^2 \,|\, Y=y) - \mathrm{E}(X \,|\, Y=y)^2$$

When X and Y are independent,

$$\mathrm{E}(X \,|\, Y=y) = \mathrm{E}(X)$$

$$\mathrm{Var}(X \,|\, Y=y) = \mathrm{Var}(X)$$

By treating $\mathrm{E}(X \,|\, Y)$ as a random variable of Y, then we can calculate its expected value such that

$$\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X \,|\, Y)).$$

This is known as the law of total expectation.

Expectation

The following property holds for both dependent and independent random variables X and Y

$$\mathrm{E}(X \pm Y) = \mathrm{E}(X) \pm \mathrm{E}(Y)$$

If X and Y are independent then

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$$

Variance of Independent Random Variables

If X and Y are independent then

$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

$$\mathrm{Var}(XY) = \mathrm{Var}(X)\,\mathrm{Var}(Y) + \mathrm{E}(X)^2\,\mathrm{Var}(Y) + \mathrm{E}(Y)^2\,\mathrm{Var}(X)$$

Covariance

Covariance

Covariance is a measure of the dependence between two random variables, it can be determined using

$$\begin{aligned}\mathrm{Cov}(X,\, Y) &= \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)\end{aligned}$$

The covariance of X and Y is:

Positive if an increase in one variable is more likely to result in an increase in the other variable.

Negative if an increase in one variable is more likely to result in a decrease in the other variable.

Zero if X and Y are independent. Note that the converse is not true.

The linear transformation of two random variables have the following covariance

$$\mathrm{Cov}(aX + b,\, cY + d) = ac\,\mathrm{Cov}(X,\, Y)$$

for constants a, b, c, and d.

Joint expectation

The joint expectation of two discrete random variables is

$$\mathrm{E}(XY) = \sum_{x\in\Omega_1}\sum_{y\in\Omega_2} xy \Pr(X=x,\, Y=y)$$

and for continuous random variables

$$\mathrm{E}(XY) = \int_{x_1}^{x_2}\int_{x_1}^{x_2} xyf(x,\, y)\,\mathrm{d}y\,\mathrm{d}x.$$

Variance of Dependent Random Variables

If X and Y are dependent then

$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\,\mathrm{Cov}(X,\, Y$$

Correlation

The covariance of two random variables describes the direction of a relationship, however it does not quantify the strength of such a relationship. The correlation explains both the direction and strength of a linear relationship between two random variables.

The correlation of two random variables X and Y is denoted $\rho(X,\, Y)$

$$\rho(X,\, Y) = \frac{\mathrm{Cov}(X,\, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

where $-1 \leq \rho(X,\, Y) \leq 1$.

These value can be interpretted as follows:

- $\rho(X,\, Y) > 0$ iff X and Y have a positive linear relationship.
- $\rho(X,\, Y) < 0$ iff X and Y have a negative linear relationship.
- $\rho(X,\, Y) = 0$ if X and Y are independent. Note that the converse is not true.
- $\rho(X,\, Y) = 1$ iff X and Y have a perfect linear relationship with positive slope.
- $\rho(X,\, Y) = -1$ iff X and Y have a perfect linear relationship with negative slope.

Note that the slope of a perfect linear relationship cannot be obtained from the correlation.