Getting started: Brief Review of Probability & Statistics

The following is a (very) ***brief*** review of concepts needed in the context of Data Science and Machine Learning. For a complete review, please follow the sources provided at the end of this document.

If there are any concepts in this document that you are unfamiliar or *rusty* about, please review them as soon as possible.

Sample Space and Probability

Probability

The concept of **probability** is used to assess uncertainty of a situation, which can broadly be defined in two ways:

- *Frequency of Occurrence* (frequentist view): percentage of successes in a moderately large number of similar situations.
- *Subjective Belief* (Bayesian view): use of personnal/group/cultural beliefs for given experiences of successful decision making.

Experiment

- An **experiment** is a procedure carried out to support, refute a *hypothesis*.

- An **outcome** is the result of a single trial of an **experiment**. For example, when flipping a coin the possible outcomes are Heads (H) or Tails (T).

- An **event** is one or more outcomes of an **experiment**. For example, when flipping a coin 3 times observing the result HHH is an event.

- The **relative frequency** of an event is the number of times that an event occurs divided by the number of times the experiment is conducted.

Simulation Experiments

In many practical situations, the analytical calculation of the probability of some event of interest is very difficult. If we have a physical or computer model that can generate outcomes of a given experiment in accordance with their *true probabilities*, we can use a **simulation experiment** to calculate with high accuracy the probability of any given event.

A **computer simulation** is a computer program that models reality and allows us to conduct experiments that:

- would *require a lot of time* to carry out in real life

- would *require a lot of resources* to carry out in real life
- would *not be possible to repeat* in real life (for instance, simulation of the next day's weather or stock market performance)

Generally we are not only interested in asking about the probabilities of **outcomes** but also the probabilities of **events**, which are combinations of outcomes.

- An event is a **set**.
- The **probability** of an event is a number between 0 and 1 that quantifies how likely that event is to occur. An event that cannot occur has probability 0, and an event that is sure to occur has probability 1. The probabilities of the outcomes sum to 1.
- We say an experiment is **fair** if every outcome is equally likely.

Sets and Set Operations

- The **subset operator** $\subset$ (reads "A is a subset of B") is defined for two sets A and B by:

$$A \subset B \text{ if } x \in A \Rightarrow x \in B$$

- The **union** of A and B (reads "A union B" or "A or B") is a set defined by:

$$A \cup B = \{x | x \in A \text{ or } x \in B\}$$

- The **intersection** of A and B (reads "A intersect B" or "A and B") is a set defined by:

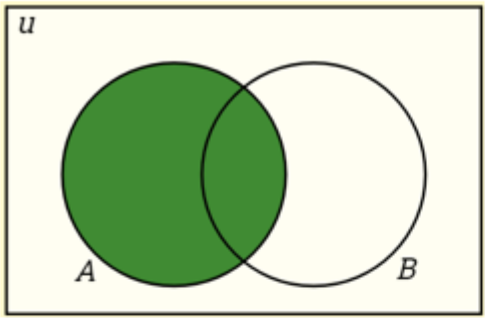
$$A \cap B = AB = \{x | x \in A \text{ and } x \in B\}$$

- The **complement** of a set A (reads "A complement") in a sample space Ω is defined by:

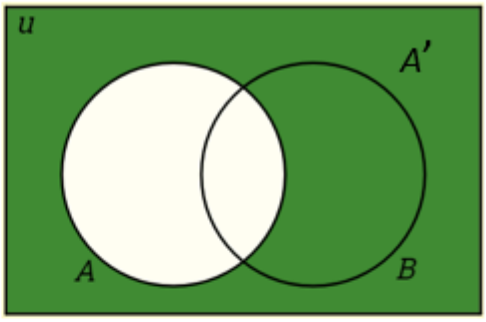
$$A^c = \overline{A} = \{x | x \in \Omega \text{ and } x \notin A\}$$

The **Venn diagram** if a fundamental tools for understanding set operations as it shows all possible logical relations between a finite collection of sets.

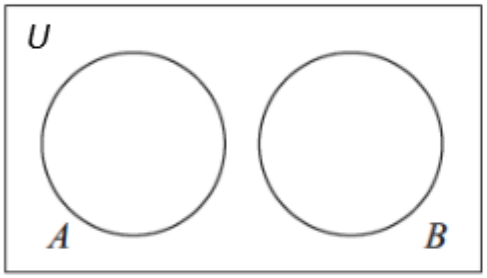
Set Operations and Venn Diagrams



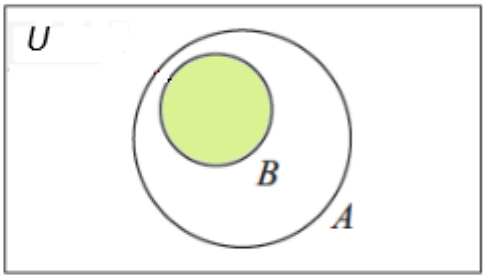
Set A



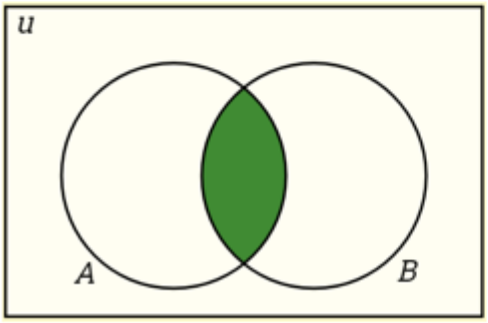
A' the complement of A



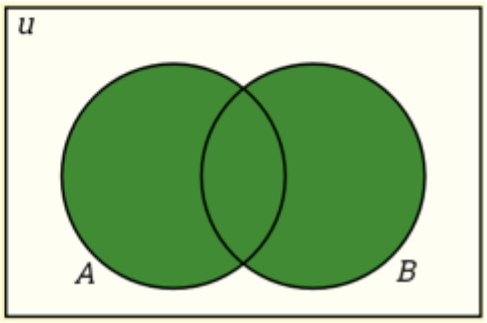
A and B are disjoint sets



B is proper subset of A $\quad B \subset A$



Both A and B
A intersect B $\quad A \cap B$



Either A or B
A union B $\quad A \cup B$

- Two sets A and B are said to be **mutually exclusive** or **disjoint** if and only if (iff) $A \cap B = \emptyset$. (This is a set relation!)
- The **cardinality** of a set S, denoted $|S|$, is the number of elements in that set.
  - A set S is **finite** if $|S| = N < \infty$
  - A set S is **countably infinite** if $|S| = |\mathbb{Z}|$, i.e., it can be put into one-to-one correspondence with the integers.
  - A set S is **uncountably infinite** if $|S| > |\mathbb{Z}|$.
- If a and b ($b > a$) are in an **interval** I, then if $a \leq x \leq b$, $x \in I$.

    - A closed interval $[a, b]$ contains the endpoints of a and b.
    - An open interval (a, b) does not contain the endpoints of a and b.
    - An interval can be half-open, such as $(a, b]$, which does not contain a, or $[a, b)$, which does not contain b.
    - Intervals can also be either finite, infinite, or partially infinite.

- A **discrete set** is either finite or countably infinite.
- A **continuous set** is not countable.

Probabilistic Models

The elements of a probabilistic model are:

- The **sample space** Ω, which is the set of all possible outcomes of an experiment.
- The **event** A, which is a *subset* ($A \subset \Omega$) of the sample space
- The **probability law**, which assigns to a set A of possible outcomes a non-negative probability number $P(A)$

Probability Axioms

1. **Non-negativity**: $P(A) \geq 0$, for every event A.
2. **Additivity**: If A and B are mutually exclusive (disjoint events), then the probability of their union satisfies $P(A \cup B) = P(A) + P(B)$
3. **Normalization**: The probability of the entire sample space Ω is equal to 1, that is, $P(\Omega) = 1$.

(Some) Properties of Probability Laws

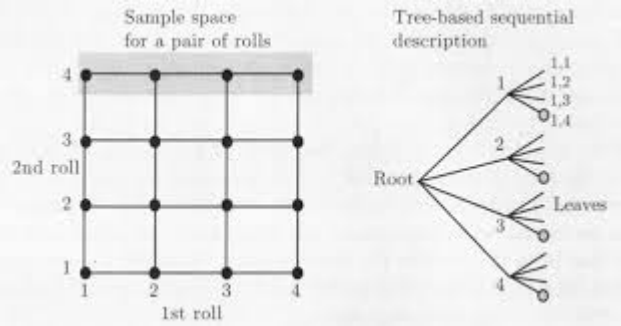
1. If $A \subset B$, then $P(A) \leq P(B)$
2. $P(A^c) = 1 - P(A)$
3. $P(A) \leq 1$
4. $P(\emptyset) = 0$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
6. $P(A \cup B) \leq P(A) + P(B)$
7. $P(A \cup B \cup C) = P(A) + P(A^c \cap B) + P(A^c \cap B^c \cap C)$

Sequential Models

A **combined experiment** or **sequential experiment** is an experiment that consists of a sequence of sub-experiments. When ordered, the sub-experiments may depend on the outcome of previous sub-experiments. (For example, two rolls of a 4-sided die.)

- It is often useful to describe the experiment and the associated sample space by means of a

tree-based sequential description.



Combinatorics

Combinatorics is the mathematics of **counting**. It can be used to find probabilities involving combinations of fair experiments.

There are two types of counting arguments that involve the selection of k objects out of a collection of n objects.

- If the selection order matters, the selection is called a **permutation**
- If the selection order does not matter, it is called a **combination**

The **cartesian product** of two sets A and B is denoted $A \times B$ and is defined by

$$A \times B = \{(a, b) | a \in A \text{ and } b \in B\}$$

That is, it is the set of all two-tuples with the first element from set A and the second element from set B.

The number of **permutations** of n objects is the number of orderings of those n objects, and can be calculated as

$$\begin{aligned} & n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1 \\ & = n! \end{aligned}$$

Sampling with replacement and with ordering

Consider choosing k values from a set of n values. The result is a k-tuple: $(x_1, x_2, \ldots, x_k)$, where $x_i \in A, \forall i = 1, 2, \ldots, k$.

Thus, this is a combined experiment with $|S_1| = |S_2| = \ldots = |S_k| = |A| \equiv n.\backslash$

Therefore the number of distinct ordered k-tuple outcomes is n^k.

Sampling without replacement and with ordering

In general, the number of ways to choose k items from n items **without replacement** and **with ordering** is

$$n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

Sampling without Replacement and without Ordering

The number of ways to choose k items from a set of n items **without replacement** and **without ordering** is

$$\frac{n!}{(n-k)!k!}$$

The value of the equation can also be expressed as

$$\binom{n}{k} = C_k^n$$

and is know as the **binomial coefficient**.

Sampling with Replacement and without Ordering

The number of ways to sample from a set $A = \{a_1, a_2, \cdots, a_n\}$ k times **with replacement** and **without ordering** is

$$C_k^{n+k-1} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!}$$

Conditional Probability

The conditional probability of event A given that event B has occurred is:

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

assuming that $P(B) > 0$. Equivalently we can write:

$$P(A \cap B) = P(B|A)P(A)$$

this is called a **chain rule**.

- Example: intersection of 3 events: $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$

Statistically Independence

Events A and B are **statistically independent (s.i.)** if and only if

$$P(A \cap B) = P(A)P(B)$$

- With this result, we say that A is **statistically independent (s.i.)** of B if

$$P(A|B) = P(A)$$

- Equivalently, we say that B is **statistically independent (s.i.)** of A if

$$P(B|A) = P(B)$$

Total Probability Theorem (or Law of Total Probability)

Let $\{A_i\}$ be disjoint events (partitions) that form the sample space Ω, and assume that $P(A_i) > 0$ for all $i = 1, \ldots, n$. Then, for any event B, by the **Law of Total Probability** we have

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

- The Law of Total Probability is often used in problems where there is a **hidden state**.
- Example: Suppose that a magician has two coins (1 fair coin and 1 2-headed coin). Let H_i denote the event that the outcome of flip i is heads. Then:

$$P(H_1) = P(H_1|F)P(F) + P(H_1|\bar{F})P(\bar{F}) = \frac{1}{2} \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{3}{4}$$

Bayes' Rule

Let $\{A_i\}$ be disjoint events (partitions) that form the sample space Ω, and assume that $P(A_i) > 0$ for all i. Then, for any event B such that $P(B) > 0$, by the **Bayes' Rule** we have

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

where

- $P(A_i|B)$ is called the **posterior**
- $P(B|A_i)$ is called the **likelihood**
- $P(A_i)$ is called the **prior**
- $P(B)$ is called the **evidence**

Using the Law of Total Probability:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}$$

Inference

Bayes's Rule is often used for **inference**, that is, we observe an effect, and we wish to *infer* the cause.

Maximum Likelihood

- Frequentist approach
- Decision Rule:

$$P(\text{event}|A_0) \underset{A_1}{\overset{A_0}{\gtrless}} P(\text{event}|A_1)$$

- Uses observational data (likelihood) only

Maximum A Posteriori

- Bayesian approach
- Decision Rule:

$$P(A_0|\text{event}) \underset{A_1}{\overset{A_0}{\gtrless}} P(A_1|\text{event})$$

- Makes use of posterior, which assumes a prior

Summary Statistics & Resampling

- A **population** is a group of people, objects, events, observations, etc. that is being studied.
- Often we are trying to assess some qualities or properties of that population. We call these **parameters**.
- A **sample** from a population is a subset of the population that can be used to draw inferences about the parameters of interest.
- A **statistic** is a measurement of a quality or property on a sample that is used to assess a parameter of the whole population.
    - Common statistics include: sample mean, sample variance, percentiles.
- **Bootstrapping** is any test or metric that relies on **random sampling with replacement**. Bootstrapping allows assigning measures of accuracy to sample estimates.
    - *The Bootstrap Idea*: the original sample approximates the population from which it was drawn. So resamples from this sample approximate what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on many resamples, approximates the sampling distribution of the statistic, based on many samples.
- **Permutation** is any test or metric that relies on **all possible rearrangements of the observed data points**.
    - When there are too many possible orderings, **Monte Carlo** sampling considered a small random sample of the possible replicates.

Hypothesis Testing

There are broadly two types of hypothesis testing:

1. Classical hypothesis testing, or *Frequentist* hypothesis test
2. Bayesian hypothesis testing

Classical Hypothesis Testing

A *statistical hypothesis* is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. For example: suppose that you toss a coin 10 times and

observe Heads 10 times. Is the coin fair? Or $P(HH\dots H|\text{coin is fair})$?

In **Classical (binary) Hypothesis Testing** we set up two hypothesis:

H_0: (the *null hypothesis*) is that the observed effect is just caused by randomness in the sampling. It is not real in the underlying system or data. For the example above, our null hypothesis is that the coin is actually fair.

H_1: (the *alternative hypothesis*) is that the observed effect is not just caused by random sampling. In this example, the coin is biased toward Heads.

- The next step of a hypothesis test generally involves setting up a simulation to compute the probability of the observed effect (in this case, flipping a coin 10 times and observing 10 Heads).
- The probability of observing an effect of the same size under the null hypothesis is called the p-**value**.
- In classical statistics/hypothesis testing, we say that an effect is **statistically significant** if the p-value is *smaller than* some small value α.
    - α is called the **significance level**.
    - Typical values of α are 0.05 or 0.01, but many argue for even smaller values now.
- The threshold to determine statistical significance,α, must always be determined before the experiment is conducted -- otherwise, there is too much temptation to adjust the threshold based on the observed p-value.

In classical hypothesis testing, we do *not* test the alternative hypothesis directly, nor can we utilize side information that we may already have about the two hypotheses. In fact, when:

- p-value$< \alpha$, we say that the test is statistically significant and therefore we *reject* the null hypothesis.
- p-value$\geq \alpha$, we say that the test is not statistically significant, we *fail to reject* the null hypothesis.

Types of Error

In binary hypothesis testing, there are two types of errors:

1. **Type I Error**, False Alarm or False Positive: occurs if we accept a hypothesis when it is not true, $P_{fa} = P(\text{false alarm}) = \alpha$
2. **Type II Error**, miss or False Negative: occurs if we reject a hypothesis when it is actually true, P_m

Receiver Operating Characteristic (ROC) Curve

The **ROC curve** is a plot that illustrates the diagnostic ability of a binary hypothesis test as its significance threshold is varied.

- the x-axis is **FPR (false positive rate)**

$$\text{FPR} = P_{fa}$$

- the y-axis is **TPR (true positive rate)**

$$\text{TPR} = 1 - P_m$$

The best ROC curve will look like a *step function* and the **area under the curve (AUC)** is 1.

Bayesian Hypothesis Testing

Bayesian hypothesis testing, similar to Bayesian inference and in contrast to frequentist hypothesis testing, is about comparing the prior knowledge about research hypothesis to posterior knowledge about the hypothesis rather than accepting or rejecting a very specific hypothesis based on the experimental data.

For example, suppose that you toss a coin 10 times and observe Heads 10 times. Is the coin fair? If a magician had given you this coin, would it change this probability?

- This *prior* knowledge can be used to compute the posterior probability using the Bayes' Rule:

$$P(\text{coin is fair}|HH\dots H) = \frac{P(HH\dots H|\text{coin is fair})P(\text{coin is fair})}{P(HH\dots H)}$$

- Bayesian hypothesis testing *test hypothesis* whereas Frequentist hypothesis testing assign *probability to hypothesis*.
- For the set of a posterior probabilites, a Bayesian hypothesis testing typically computes a **confidence intervals** for which values the null hypothesis is acceptable.
- Similarly, a significance value α is defined.

Popular tests:

- Z-Test
- T-Test
- χ^2-test (reads chi-squared)
- McNemar's Test

Goodness of Fit

In some cases, we can test on whether a set of observations follows a standard distribution.

- For **discrete distributions**, we can use the χ^2-test
- For **continuous distributions**, we can use **probability plots** or Quantile-Quantile (Q-Q) plots

Random Variables

Given an experiment and the corresponding set of possible outcomes (the sample space), a **random variable** associates a particular *number* with each outcome. We refer to this number as the numerical value or simply the **value** of the RV.

Mathematically, a random variable is a **real-valued function** of the experimental outcome.

We can have three types of random variables:

- discrete
- continuous
- mixture

Discrete Random Variable

A random variable is called **discrete** if its range (the set of values that it can take) is either **finite** or **countably infinite**.

- A discrete RV has an associate **probability mass function**, which gives the probability of each numerical value that the random variable can take.

If x is any real number, the **probability mass function** (or **PMF**) of the random variable X, denoted $p_X(x)$, is the probability of the event $\{X = x\}$ consisting of all outcomes that give rise to a value X equal to x:

$$p_X(x) = P(\{X = x\})$$

where $\sum_x p_X(x) = 1$.

The Bernoulli Random Variable

An event $A \in \Omega$ is considered a "success".

- A **Bernoulli RV** X is defined by

$$X = \begin{cases} 1, & s \in A \\ 0, & s \notin A \end{cases}$$

- We refer to X as the **Bernoulli** RV **with probability of success** p:

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

- The PMF for a Bernoulli RV X is defined by

$$p_X(x) = P(X = x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \\ 0, & x \neq \{0, 1\} \end{cases}$$

The Binomial Random Variable

A **Binomial random variable** can be defined as the sum of n independent Bernoulli RVs.

- Let X be the # of successes.

- We refer to X as the **Binomial** RV **with parameters** n **and** p:

$$X \sim \text{Binomial}(n, p)$$

- The PMF of X is given by

$$p_X(x) = P(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & k = 0, 1, \ldots, n \\ 0, & \text{otherwise (o.w.)} \end{cases}$$

The Geometric Random Variable

A **Geometric random variable** occurs when independent Bernoulli trials are conducted until the first success

- Let X be the number of trials required.
- We refer to X as the **Geometric** RV with **probability** p:

$$X \sim \text{Geometric}(p)$$

- The PMF of X is given by

$$p_X(x) = P(X = x) = \begin{cases} p(1-p)^{k-1}, & k = 1, 2, \ldots, n \\ 0, & \text{o.w.} \end{cases}$$

The Poisson Random Variable

A **Poisson random variable** models events that occur randomly in space or time.

- Let λ = the # of events/(unit of space or time) and consider observing some period of time or space of length t and let $\alpha = \lambda t$.
- Let X= the # events in time (or space) t. We refer to X as the **Poisson** RV with x event:

$$X \sim \text{Poisson}(x)$$

- The PMF of the Poisson random variable is

$$P_N(n) = \begin{cases} \frac{\alpha^n}{n!} e^{-\alpha}, & n = 0, 1, \ldots \\ 0, & \text{o.w.} \end{cases}$$

Continuous Random Variable

A random variable X is called **continuous** if its range (the set of values that it can take) is **uncountably infinite** and its probability law can be described in terms of a *nonnegative* function f_X called **probability density function** (or **PDF**) of X, which satisfies

$$P(X \in B) = \int_B f_X(x) dx,$$

for every subset B of the real line.

Continuous RVs do not have probability at any discrete points, that is,

$$P(X = x) = 0, \forall x \in \mathbb{R}$$

- This only means the event $X = x$ is *extremely unlikely* but not impossible.

PDF Properties

1. $f_X(x) \geq 0$ for all x.
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
3. If δ is small, then $P([x, x + \delta]) \approx f_X(x) \cdot \delta$.
4. For any subset B of the real line,

$$P(X \in B) = \int_B f_X(x)dx.$$

Cumulative Distribution Functions

If $(\Omega, \mathcal{F}, P)$ is a probability space with X a real discrete RV on Ω, the **Cumulative Distribution Function (CDF)** is denoted as $F_X(x)$ and provides the probability $P(X \leq x)$. In particular, for every x we have

$$F_X(x) = P(X \leq x) = \begin{cases} \sum_{k \leq x} p_X(k) & X: \text{ discrete} \\ \int_{-\infty}^{x} f_X(t)dt & X: \text{ continuous} \end{cases}$$

Loosely speaking, the CDF $F_X(x)$ "accumulates" probability "up to" the value x.

- The CDF $F_X(x)$ is a probability measure therefore it inherits all the properties of a probability measure.
- If X is continuous, the PDF and CDF can be obtained from each other by integration or differentiation:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt,$$

$$f_X(x) = \frac{dF_X}{dx}(x)$$

Survival Function

If $(\Omega, \mathcal{F}, P)$ is a probability space with X a real discrete RV on Ω, the **Survival Function (SF)** is denoted as $S_X(x)$ and provides the probability $P(X > x)$. In particular, for every x we have

$$S_X(x) = P(X > x) = 1 - P(X \leq x) = 1 - F_X(x)$$

The Uniform Random Variable

A random variable X that takes values in an interval $[a, b]$, and all subintervals of the same length are equally likely, is called a **uniform** or **uniformly distributed** random variable. Its PDF has the form:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise} \end{cases}$$

The Exponential Random Variable

An **exponential** random variable represents the time between events that occur continuously and independently at a constant average rate λ. Its PDF has the form:

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

- An exponential random variable can be obtainable as a limit of Geometric random variables.

The Gaussian Random Variable

A continuous random variable X is said to be **Gaussian** if it has a PDF form

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

with mean μ and variance $\sigma^2 \geq 0$.

- A random variable is called **Normal** if it is a **Gaussian** random variable with mean $\mu = 0$ and variance $\sigma^2 = 1$.

Limit Theorems

- The **Weak Law of Large Numbers**: asserts that the sample mean of a large number of independent identically distributed random variables is very close to the true mean, with high probability.
- The **Central Limit Theorem**: asserts that the sum of a large number of independent and identically distributed (not necessarily Normal) random variables has an approximately Normal CDF, regardless of the CDF of the individual random variables.

Expected Value

If x_i are samples drawn from a random variable X, then the **expected value** or **mean** of the random variable X is

$$\mu_X = E[X] = \sum_i x_i p_X(x_i), \text{ if } X \text{ is discrete}$$

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx, \text{ if } X \text{ is continuous}$$

Moments

The **moments** of a random variable (or of its distribution) are *expected values of powers* or related functions of the random variable.

The r-th moment of RV X is $E[X^r]$.

- In particular, the first moment is the *mean*, $\mu_X = E[X]$.

Central Moments

The **central moments** of a random variable (or of its distribution) are *expected values of mean-centered powers* or related functions of the random variable.

The r-th central moment of RV X is $E[(X - \mu_X)^r]$, in general, $E[(X - E[X])^r]$.

- In mathematics, a *moment* is a specific quantitative measure of the shape of a function. The most important ones are:

    1. The **mean**, or the first moment of the random variable X: $E[X]$. It measures the average value of the its PDF.
    2. The **variance**, or the second central moment of the random variable X: $E[(X - E[X])^2]$. It provides a measure of how much the PDF spreads away from the mean.
    3. The **skewness**, or the third central moment of the random variable X: $E[(X - E[X])^3]$. It measures the amount of asymmetry of its PDF with respect to the mean value. For example, the skewness of a Normal distribution is 0.
    4. The **kurtosis**, or the fourth central moment of the random variable X: $E[(X - E[X])^4]$. It's a measure of the tailedness of its PDF. For example, the Normal distribution has a kurtosis of 3.

Variance

The variance of a random variable X is computed as the second central moment of its PDF:

$$\begin{aligned} Var[X] = \sigma_X^2 &= E[(X - E[X])^2] && (1) \\ &= E[X^2] - (E[X])^2 && (2) \end{aligned}$$

Properties of Variance:

Let X be a random variable and b and c constant values.

1. $Var[X] = E[X^2] - (E[X])^2 \geq 0$
2. $Var[c] = 0$

3. $Var[X - c] = E[X^2] - (E[X])^2$
4. $Var[cX] = c^2 Var[X]$
5. $Var[cX + b] = c^2 Var[X]$

Covariance

The **covariance** of two random variables X and Y, denoted by $\operatorname{cov}(X, Y)$, is defined by

$$\operatorname{cov}(X, Y) = E\big[(X - E[X])\,(Y - E[Y])\big]$$

- The *unbiased sample* covariance is given by:

$$\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right) \left(y_i - \overline{y}\right)$$

where $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ are the sample means of the RVs X and Y, respectively.

The covariance matrix of two random variables X and Y is a 2×2 matrix:

$$\Sigma = \operatorname{cov}(X, Y) = \begin{bmatrix} \operatorname{cov}(X, X) & \operatorname{cov}(X, Y) \\ \operatorname{cov}(Y, X) & \operatorname{cov}(Y, Y) \end{bmatrix} \tag{3}$$

$$= \begin{bmatrix} \operatorname{var}(X) & \operatorname{cov}(X, Y) \\ \operatorname{cov}(X, Y) & \operatorname{var}(Y) \end{bmatrix} \tag{4}$$

Pearson's Correlation Coefficient

For random variables X and Y, thwe **Pearson's correlation coefficient** (or simply the **correlation coefficient**) is

$$r = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\operatorname{cov}(X, Y)$ is the covariance and σ_X and σ_Y are the square-roots of the corresponding variances.

- **Correlation is not causation.**

Coefficient of Determination

The **coefficient of determination**, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

$$r^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$0 \leq r^2 \leq 1$$

- r^2 is the square of the correlation coefficient r.

Jointly Gaussian Random Variables

Two **Gaussian** random variables X and Y are said to be *jointly Gaussian* if their joint density function (or multivariate Gaussian distribution) can be written as:

$$f_{XY}(x,y) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

where $d = 2$, $\mathbf{x} = [x, y]^T$, $\mu = [\mu_x, \mu_y]^T$ and Σ is the **covariance** matrix.

- Two **Gaussian** random variables X and Y that each have mean 0 and variance 1 are said to be *jointly Gaussian* if their density function can be written as

$$f_{X,Y}(x,y) = \frac{1}{(2\pi)^{d/2}\sqrt{1-r^2}} \exp\left(-\frac{(x^2 - 2rxy + y^2)}{2(1-r)^2}\right)$$

where $|r| < 1$ is the correlation coefficient for X and Y.

Further Review

- "Chapter 3 - Probability and Information Theory" from Deep Learning book by Ian GoodFellow et al., MIT Press, 2016
- "Introduction to Probability" book by Bertsekas and Tsitsiklis, educational resources
- "Seeing Theory", an interactive web-tool for review of probability and statistics