Probability & Statistics

My github: https://github.com/robinyUArizona

Probability

Probabilities A.K.A. Chance → How likely something (an event) is to happen? Probabilities ARE NOT Guarantees! → Probability tells you that **over the long run** there is a certain chance of something happening, not that something will or will not happen at a specific time.

• Probability is a measure of the size of a set.

→ Bridge between descriptive & inferential statistics

→ In **probability**, properties of the **population** are assumed known & questions regarding a **sample** taken from the population are posed and answered.

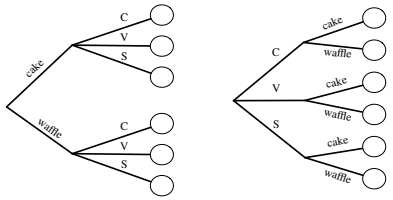
Naive Definition of Probability

If all outcomes are equally likely, the probability of an event A happening is:

$$\boxed{P_{\text{naive}}(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes}}}$$

Counting

Multiplication Rule



Let's say we have a compound experiment (an experiment with multiple components). If the 1st component has n_1 possible outcomes, the 2nd component has n_2 possible outcomes, . . . , and the rth component has n_r possible outcomes, then overall there are $n_1 n_2 \dots n_r$ possibilities for the whole experiment.

Permutation & Combination

Video to add

• **Permutation**: A permutation is an arrangement of r objects from a pool of n objects in a **given order**.

$$\boxed{{}_n\mathrm{P}_r = \frac{n!}{(n-r)!}, \qquad 0 \le r \le n}$$

→ **Permutation when all the Objects are Distinct**

Theorem 1: If the number of permutations of n different objects taken r at a time, satisfying the condition $\boxed{0 < r \le n}$, and where the **repetition is not allowed** is

$$\boxed{{}_n\mathrm{P}_r \to \frac{n!}{(n-r)!} = n(n-1)(n-2)\cdots(n-r+1)}$$

Theorem 2: If the number of permutations of n different objects taken r at a time, where the **repetition is allowed** is

$$\boxed{n^r}$$

→ **Permutation when all the Objects are not Distinct**

Theorem 3: The number of permutations of n objects, and p's are of the same kind and rest is all different kind is

$$\boxed{\frac{n!}{p!}}$$

Theorem 4: The number of permutations of n objects, where n_1 are the objects of one kind, n_2 are of the second kind, ..., n_k is of the k^{th} kind and the rest, if any, are of a different kind, then the permutation is given by

$$\boxed{\frac{n!}{n_1! n_2! \cdots n_k!}}$$

• **Combination**: A combination is an arrangement of r objects from a pool of n objects, where the **order does not matter**.

→ We use combinations to count the number of ways to choose a group of r unordered objects from n possibilities **without replacement** by

$$\boxed{{}_n\mathrm{C}_r = \frac{n!}{(n-r)!\, r!} = \frac{{}_n\mathrm{P}_r}{r!} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!}}$$

→ We use combinations to count the number of ways to choose a group of r unordered objects from n possibilities **with replacement** by

$$\boxed{{}_n\mathrm{C}_r = \frac{(r+n-1)!}{r!\ (n-1)!}}$$

Note: **without replacement** or **without repetition** means that you can **not** pick the same element more than once.

	Permutation	**Combination**
With Replacement	n^k	$\frac{(r+n-1)!}{r!\ (n-1)!}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\frac{n!}{(n-r)!\ r!}$

Thinking Conditionally

Independence

• **Independent Events**: Events A and B are **independent**, if and only if one of the following equivalent statements holds:

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

• Events A, B, and C are **mutually independent** if **both** of the following hold:

(a) A, B, and C are pairwise independent:

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

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

• **Conditional Independence**: Events A and B are **conditionally independent** given C if $\boxed{P(A \cap B|C) = P(A|C)\ P(B|C)}$.

→ Conditional independence does not imply independence, and independence does not imply conditional independence.

Unions, Intersections, and Complements

De Morgan's Laws: A useful identity that can make calculating probabilities of unions easier by relating them to intersections, and vice versa. Analogous results hold with more than two sets.

$$\boxed{(A \cup B)^c = A^c \cap B^c}$$
$$\boxed{(A \cap B)^c = A^c \cup B^c}$$

Complements: The following are true.

$$\boxed{A \cup A^c = \Omega} \quad \boxed{A \cap A^c = \emptyset} \quad \boxed{P(A) = 1 - P(A^c)}$$

Joint Probability

• The **joint probability** $P(A, B)$ or $P(A; B)$ is the probability of both A and B occurring simultaneously (same time).

$$\boxed{P(A, B) = P(A \cap B) = \frac{N(A \cap B)}{N(B)}}$$

Conditional Probability

Conditional Probability, with all outcomes equally likely.

• The **conditional probability** of event A, given that event B occurs is given by

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}} \quad \to \quad \boxed{P(A \cap B) = P(B) \cdot P(A|B)}$$

Note: $\boxed{P(A|B) \neq P(B|A)}$

• **Law of Total Probability**:

$$\boxed{P(B) = \sum_{i=1}^{n} P(B|A_i)\ P(A_i)} \qquad \boxed{\Omega = \bigcup_{i=1}^{n} A_i}$$

Marginal Probability

• The **marginal probability** is a $P(A)$; can compute by *marginalizing* the *joint distribution* $\mathcal{B} = \{B_1, B_2, ..., B_n\}$ is a collection of mutually exclusive events

$$\boxed{P(A) = \sum_{B \in \mathcal{B}} P(A, B) = \sum_{B \in \mathcal{B}} P(A \cap B)}$$

Bayes' Theorem

The chance of you getting hit by lightning at the top of a tall building is high compared to on the ground. Hence information about the location changes the probability. Bayesian tries to interpret such additional information in terms of probability. Given your location, we can have a more precise probability of an event, than a "any location" or a "generic location".

For events A and B such that $P(B) > 0$, we have

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

Note: $P(A|B)$ and $P(B|A)$ are not the same thing!

$$\boxed{P(A_i|B) = \frac{P(B|A_i)\ P(A_i)}{\sum_{j=1}^{n} P(B|A_j)\ P(A_j)}} \qquad \boxed{\Omega = \bigcup_{i=1}^{n} A_i}$$

Likelihood and Posterior

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

- $P(A)$ is the prior
- $P(B|A)$ is the likelihood
- $P(B) = \int P(B|A)P(A)\,dA$ is a scaling factor (constant A)
- $P(A|B)$ is the posterior.

Bayesian probability is like a logic of probability, where we walk from the initial hypothesis (prior) to final beliefs through premises (likelihood) and observations (data).

Probability space

A probability space, **a mathematical concept**, is a ordered triple $(\Omega, \mathcal{B}, P)$, respectively the *sample space*, *event space*, and *probability function*.

Ω: **sample space** $\rightarrow$ Set of outcomes of an experiment.
Example: tossing a coin twice. $\Omega = \{$HH, HT, TH, TT$\}$

$\mathcal{B}$: **event space** $\rightarrow$ all possible **subsets** of an **sample space** Ω. An event is a subset of Ω
Example: (i) "at least one head" is $\{$HH, HT, TH$\}$, (ii) "no more than one head" is $\{$HT, TH, TT$\}$

In **probability theory**, the event space $\mathcal{B}$ is modelled as a σ-**algebra** (or σ-**field**) of Ω, which is a collection of subsets of Ω with the following properties:

(1) $\boxed{\emptyset \in \mathcal{B}}$

(2) If an event $\boxed{A \in \mathcal{B}}$, then $\boxed{A^{\complement} \in \mathcal{B}}$ (closed under complementation)

(3) If $\boxed{A_1, A_2, \ldots \in \mathcal{B}}$, then $\boxed{\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}}$ (closed under countable union).

Additional properties:

(4) (1) + (2) $\rightarrow$ $\boxed{\Omega \in \mathcal{B}}$

(5) (3) + De-Morgan's Laws $\rightarrow$ $\boxed{\bigcap_{i=1}^{\infty} A_i \in \mathcal{B}}$ (closed under countable intersection)

Example: tossing a coin twice
$\Omega = \{$HH, HT, TH, TT$\}$, then $\mathcal{B} = 2^{\Omega} = 16$, and therefore σ-**field** can be given by
$\{\emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, HT, TH\}, \cdots, \Omega\}$
Note: Observable events could be the collection of all intervals, plus what can be obtained by set operations $\cup, \cap, \complement$.

Suppose we cannot distinguish between HT and TH, then we have sample space $\Omega = \{$HH, HT, TT$\}$, and $\mathcal{B} = 2^{\Omega} = 8$, therefore σ-**field** can be given by
$\{\emptyset, \{HH\}, \{HT\}, \{TT\}, \{HH, HT\}, \{HH, TT\}, \{HT, TT\}, \Omega\}$

Example: (i) "at least one head" is $\{$HH, HT, TH$\}$.
If we want the smallest σ-**field** that makes $\{$HH, HT, TH$\}$ measurable, then qualitatively σ-**field** formed by taking complements, finite intersections, and countable unions of members of $\mathcal{B}$, until it satisfies all requirements of a σ-**field**. When we go through this process, we obtain the σ-**algebra**,
$\mathcal{B} = \{\emptyset, \{TT\}, \{HH, HT, TH\}, \cdots, \Omega\})$

P: **probability measure** $\rightarrow$ a function that assigns numbers (called probabilities) between 0 and 1 to each event in $\mathcal{B}$ i.e. $\mathcal{B} \rightarrow [0, 1]$

In other words, a **probability measure** is a function $\boxed{P : \mathcal{B} \rightarrow [0, 1]}$ that maps an event A to a real number in $[0, 1]$. The function must satisfy the **axioms of probability**:

Probability Axioms

1. **Non-negativity**: $\boxed{P(A) \geq 0}$, for all $A \in \mathcal{B} \rightarrow$ ensures that probability is never negative.
2. **Normalization**: $\boxed{P(\Omega) = 1}$. $\rightarrow$ ensures that probability is never greater than 1.
3. Countable **Additivity**: If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint (i.e., $A_i \cap A_j = \emptyset$, if $i \neq j$), then $\boxed{P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)}$

Theorem: Let $(\Omega, \mathcal{B}, P)$ be a probability space.
1. $P\,(A^{\complement}) = 1 - P\,(A)$ for $A \in \mathcal{B}$
2. $P\,(A \cup B) = P\,(A) + P\,(B) - P\,(A \cap B)$. for $A, B \in \mathcal{B}$
3. $P\,(A/B) = P\,(A) - P\,(A \cap B)$. for $A, B \in \mathcal{B}$
4. If $A \subset B$, then $P\,(A) \leq P\,(B)$ for $A, B \in \mathcal{B}$
5. If $A_1, A_2, \ldots \in \mathcal{B}$ are disjoint, then $\boxed{P\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} P(A_j)}$

Example:: Tossing a coin twice. Assuming that the coin is fair, then the *probability function* for the σ-algebra consisting of all subsets of Ω is

Event A	$P(A)$
HH	$\frac{1}{4}$
HT	$\frac{1}{4}$
TH	$\frac{1}{4}$
TT	$\frac{1}{4}$
$\emptyset$	0
Ω	1
(HH, HT, TH)	$\frac{3}{4}$ (using pt. (3) of Def'n above)
(HH,HT)	$\frac{1}{2}$
$\vdots$	$\vdots$

Example:: A coin toss.
Ω : {H, T}
$\mathcal{B}$: $\{\phi, \{$H$\}, \{$T$\}, \{$H, T$\}\}$
P: $P(\phi) = 0$, $P(\{$H$\}) = \frac{1}{4}$, $P(\{$T$\}) = \frac{3}{4}$, $P(\{$H,T$\}) = 1$

Random Variable

Consider an experiment with four outcomes $\Omega = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$. We want to construct the probability space $(\Omega, \mathcal{B}, P)$. The sample space Ω is already defined. The event space $\mathcal{B}$ is the set of all possible subsets in Ω, which, in our case, is a set of 2^4 subsets. For the probability law P, let us assume that the probability of obtaining each outcome is

$$P[\{\clubsuit\}] = \frac{1}{6},\ P[\{\diamondsuit\}] = \frac{2}{6},\ P[\{\heartsuit\}] = \frac{2}{6},\ P[\{\spadesuit\}] = \frac{1}{6}$$

Therefore, we have constructed a **probability space** $(\Omega, \mathcal{B}, P)$ where everything is perfectly defined. So, in principle, they can live together happily forever.
A lazy data scientist comes, and there is a (small) problem. The data scientist does not want to write the symbols $\clubsuit$, $\diamondsuit$, $\heartsuit$, $\spadesuit$. There is nothing wrong with his motivation because all of us want efficiency. How can we help him? Well, the easiest solution is to encode each symbol with a number, for example, $\clubsuit \leftarrow 1$, $\diamondsuit \leftarrow 2$, $\heartsuit \leftarrow 3$, $\spadesuit \leftarrow 4$, where the arrow that we assign a number to the symbol. But we can express this more formally by defining a function $\boxed{X : \Omega \rightarrow \mathbb{R}}$ with

$$X(\clubsuit) = 1,\ X(\diamondsuit) = 2,\ X(\heartsuit) = 3,\ X(\spadesuit) = 4$$

There is nothing new here: we have merely converted the symbols to numbers, with the help of a function X, However, with X defined, the probabilities can be written as

$$P[X = 1] = \frac{1}{6},\ P[X = 2] = \frac{2}{6},\ P[X = 3] = \frac{2}{6},\ P[X = 4] = \frac{1}{6}$$

This is much more convenient, and so the data scientist is happy.

A **random variable** X on a probability space $(\Omega, \mathcal{B}, P)$ is a function that maps every element in a sample space to a real line
i.e., $\boxed{X : \Omega \rightarrow \mathbb{R}}$
In other words, A random variable is a rule that assigns a numerical value to each outcome in a sample space.

Example: Flip a coin 2 times. The sample space Ω is

$$\Omega = \{(\text{head, head}), (\text{head, tail}), (\text{tail, head}), (\text{tail, tail})\}$$

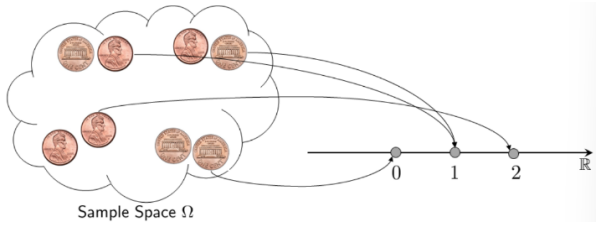
Suppose that X is a random variable that maps an outcome to a number representing the sum of "head", i.e,

$$X(\cdot) = \text{number of heads.}$$

Then, for the four ξ's in the sample space there are only 3 distinct numbers. More precisely, if we let $\xi_1 = (\text{head, head}), \xi_2 = (\text{head, tail}), \xi_3 = (\text{tail, head}), \xi_4 = (\text{tail, tail})$, then, we have

$$X(\xi_1) = 2,\ X(\xi_2) = 1,\ X(\xi_3) = 1,\ X(\xi_4) = 0$$

This shows that the mapping defined by the random variable is not necessarily a one-to-one mapping because multiple outcomes can be mapped to the same number.



Types of Random Variables

(1) **Numerical RV**
  (i) Discrete RV: can take only distinct, separate values
  (ii) Continuous RV: can take any value in some interval (low, high)
(2) **Categorical RV**

Probability Distribution Functions

Probability Distribution: describes how the probability of Ω is distributed along the range of X.
$\rightarrow$ The probability distribution of X says how the total probability of 1 is distributed among the various possible X values.

Command	What it does
`help(distributions)`	shows documentation on distributions
`dbinom(k,n,p)`	PMF $P(X = k)$ for $X \sim \text{Bin}(n, p)$
`dt(x,n)`	PDF $f(x)$ for $X \sim t_n$
`dunif(x,a,b)`	PDF $f(x)$ for $X \sim \text{Unif}(a, b)$

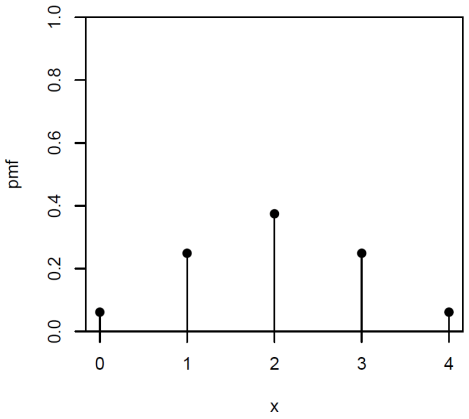
Let X have pmf $f(x)$ then, here $f(x)$ is $\boxed{p_X(x) = P[X = x]}$.

Discrete X	Continuous X
$\mu = E(X) = \sum_{i=1}^{n} x f(x)$	$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$
$\sigma_X^2 = V(X) = \sum_{i=1}^{n} (x-\mu)^2 f(x)$	$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$
$\sigma_X^2 = E\left[(X-\mu)^2\right] = E(X^2) - [E(X)]^2$	fvd

- Properties of Variance
  1. $\boxed{V(aX+b) = a^2 \cdot \sigma^2}$
  2. In particular, $\boxed{\sigma_{aX} = |a| \cdot \sigma_x}$
  3. $\boxed{\sigma_{X+b} = \sigma_X}$

Probability Mass Functions (PMF)

The **PMF** gives the probability of a discrete random variable X that takes on the value x i.e, $X(\xi) = x$. → The set of all possible states of X is denoted as $X(\Omega)$
We denote PMF as $\boxed{p_X(x) = P[X = x]}$



The PMF satisfies,
$p_X(x) \geq 0$, and $\sum_x p_X(x) = 1$
→ The **probabilities** are summarized by a function known as the probability mass function (PMF).
→ PMFs are the ideal histograms of random variables.
→ $\boxed{\text{Expectation} = \text{Mean} = \text{Average computed from a PMF}}$

Example: Flip a coin twice. The sample space is $\Omega = \{\text{HH, HT, TH, TT}\}$. We can assign a random variable X = number of heads. Therefore,

$$X(\text{"HH"}) = 2,\ X(\text{"TH"}) = 1,\ X(\text{"HT"}) = 1,\ X(\text{"TT"}) = 0$$

So the random variable X takes three states: 0,1,2. The PMF is therefore

$$p_X(0) = P[X = 0] = P[\{\text{"TT"}\}] = \frac{1}{4},$$

$$p_X(1) = P[X = 1] = P[\{\text{"TH"}, \text{"HT"}\}] = \frac{1}{2},$$

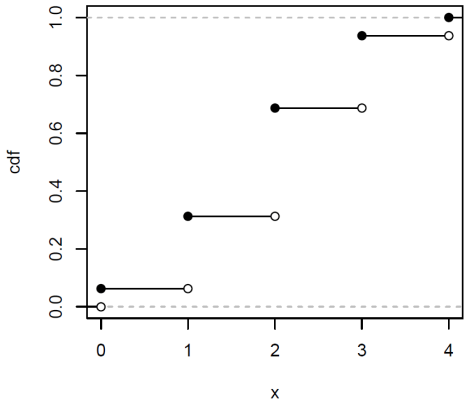
$$p_X(2) = P[X = 2] = P[\{\text{"HH"}\}] = \frac{1}{4}$$

→ The PMF is the weighing function for discrete random variables. Two random variables are different when their PMFs are different because they are constructing two different measures.

Cumulative Density Function (CDF)

The CDF gives the probability that a random variable X is less than or equal to x.

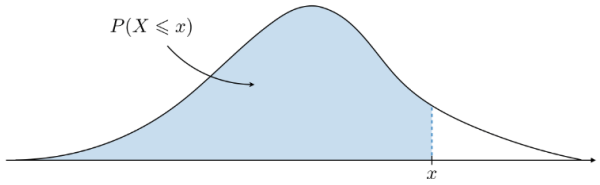
$$F_X(x) = P(X \leq x)$$



The CDF is an increasing, right-continuous function with

$$F_X(x) \to 0 \text{ as } x \to -\infty \text{ and } F_X(x) \to 1 \text{ as } x \to \infty$$

- Continuous probabilities F_X, **CDF**: → $F_X(x) = P[X \leq x]$, monotonic, non-decreasing, x goes to ∞



What's the probability that a Continuous Random Variables is in an interval? Take the difference in CDF values (or use the PDF as described later).

$$\boxed{P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)}$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, this becomes

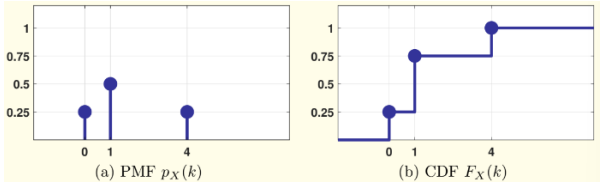
$$P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Example: Consider a random variable X with PMF $p_X(0) = \frac{1}{4}, p_X(1) = \frac{1}{2}$ and $p_X(4) = \frac{1}{4}$. The CDF of X can be computed as

$$F_X(0) = P[X \leq 0] = p_X(0) = \frac{1}{4},$$

$$F_X(1) = P[X \leq 1] = p_X(0) + p_X(1) = \frac{3}{4},$$

$$F_X(4) = P[X \leq 4] = p_X(0) + p_X(1) + p_X(4) = 1$$



- Converting between PMF and CDF: If X is a discrete random variable, then the PMF of X can be obtained from the CDF by

$$\boxed{p_X(x_k) = F_X(x_k) - F_X(x_{k-1})}$$

Continue the above example

$$p_X(0) = F_X(0) - F_X(-\infty) = \frac{1}{4} - 0 = \frac{1}{4},$$

$$p_X(1) = F_X(1) - F_X(0) = \frac{1}{4} - 0 = \frac{3}{4} - \frac{1}{4} = \frac{1}{2},$$

$$p_X(4) = F_X(4) - F_X(1) = 1 - \frac{3}{4} = \frac{1}{4}.$$

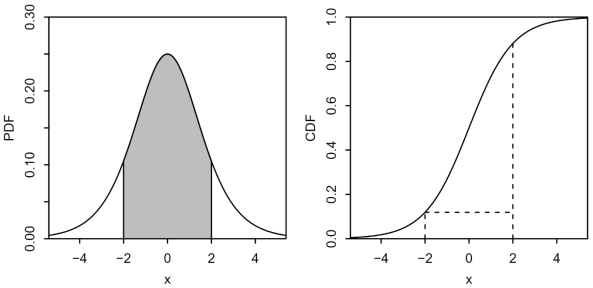
Probability Density Function (PDF)

The PDF f is the derivative of the CDF F.

$$F'(x) = f(x)$$

A PDF is nonnegative and integrates to 1. By the fundamental theorem of calculus, to get from PDF back to CDF we can integrate:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$



Let X be a continuous random variable. The PDF of X is a function $f_X : \Omega \to \mathbb{R}$ that, when integrated over an interval $[a, b]$ yields the probability of obtaining $a \leq X \leq b$:

$$\boxed{\mathbb{P}[a \leq X \leq b] = F(b) - F(a) = \int_a^b f_X(x)\, dx}$$

Two additional properties of a PDF: it must integrate to 1 (because the probability that a CRV falls in the interval $[-\infty, \infty]$ is 1, and the PDF must always be non-negative.

$$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad f(x) \geq 0$$

Example: Let $f_X(x) = 3x^2$ with $\Omega = [0, 1]$. Let $A = [0, 0.5]$. The probability $\mathbb{P}[\{X \in A\}]$ is

$$\mathbb{P}[0 \leq X \leq 0.5] = \int_0^{0.5} 3x^2\, dx = \frac{1}{8}$$

→ Densities can exceed 1, → Densities are **not** probabilities.
Note: *Continous distributions* **does not** *have probability mass functions* instead we use **probability density** because the probability of a continuous variable being any one precise value is 0.
Note: The PDF and the CDF of a given random variable contain exactly the same information.

Classic Statistical Distribution

Binomial Distribution (**Discrete**): Assume X is distributed Bin(n, p). x successes in n independent trials or events, each with p probability. Here, each success occurs with the same probability p and each failure occurs with probability $q = 1 - p$.

$$\text{PDF: } P(X = x) = \binom{n}{x} p^x q^{n-x} \quad : x = 0, 1, 2, \ldots$$

$$\boxed{\mu = np} \qquad \boxed{\sigma^2 = npq}$$

$\rightarrow$ if $n = 1$, this is a Bernoulli distribution.

Negative Binomial: Number of failures before r successes.

Geometric: First success with p probability on the n^{th} trail.

$$\text{PDF: } P(X = x) = q^{n-1}p \qquad \boxed{\mu = \frac{1}{p}} \qquad \boxed{\sigma^2 = \frac{1-p}{p^2}}$$

Hypergeometric: x successes in n draws, no replacement, from a size N population with X items of that feature

$$\text{PDF: } P(X = x) = \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}} \qquad \boxed{\mu = \frac{nX}{N}}$$

Poisson Distribution (**Discrete**): Assume X is distributed Pois(λ). Number of x successes in a fixed interval of time/space, where these success occur independently and with a known constant or average rate λ.

$$\text{PDF: } P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \boxed{\mu = \lambda} \qquad \boxed{\sigma^2 = \lambda}$$

Power Law Distribution (**Discrete**): Many data distributions have much longer tails than the normal or Poisson distributions. In other words, the change in one quantity varies as a power of another quantity. It helps measure the inequality in the world. e.g. wealth, word frequency and Pareto Principle (80/20 Rule)

$$\text{PDF: } P(X = x) = cx^{-\alpha}$$

where α is the law's exponent and c is the normalizing constant.

Normal/Gaussian Distribution (**Continuous**):
Assume X in distributed $\mathcal{N}(\mu, \sigma^2)$. It is a bell-shaped and symmetric distribution. Bulk of the values lie close to the mean and no value is too extreme. Generalization of the binomial distribution as $n \to \infty$.

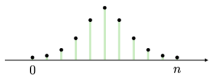
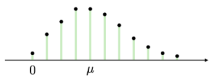
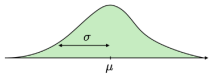
$$\text{PDF: } P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \boxed{\mu = \mu} \qquad \boxed{\sigma^2 = \sigma^2}$$

Implications: 68%-95%-99% rule. 68% of probability mass fall within 1σ of the mean, 95% within 2σ, and 99.7% within 3σ.

$\rightarrow$ Normal Approximation - discrete distributions such as Binomial and Poisson can be approximated using z-scores when np, nq, and λ are greater than 10.

Exponential - memoryless time between independent event occuring at an average rate $\boxed{\lambda \to \lambda e^{\lambda x}}$, with $\boxed{\mu = \frac{1}{\lambda}}$

Gamma - time until n independent events occuring at an average rate λ.

Distribution	PDF	$\psi(\omega)$	$E[X]$	Var(X)	Illustration
$X \sim \mathcal{B}(n,p)$	$\binom{n}{x}p^x q^{n-x}$	$(pe^{i\omega} + q)^n$	np	npq	
$X \sim \text{Po}(\mu)$	$\frac{\mu^x}{x!}e^{-\mu}$	$e^{\mu(e^{i\omega}-1)}$	μ	μ	
$X \sim \mathcal{U}(a,b)$	$\frac{1}{b-a}$	$\frac{e^{i\omega b} - e^{i\omega a}}{(b-a)i\omega}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
$X \sim \mathcal{N}(\mu,\sigma)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$e^{i\omega\mu - \frac{1}{2}\omega^2\sigma^2}$	μ	σ^2	
$X \sim \text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$\frac{1}{1-\frac{i\omega}{\lambda}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	

Joint PDFs and CDFs

Joint Distributions

The **joint CDF** of X and Y is

$$\boxed{F(x, y) = P(X \leq x, Y \leq y)}$$

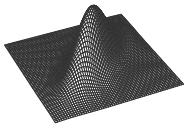
In the discrete case, X and Y have a **joint PMF**

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

In the continuous case, they have a **joint PDF**

$$\boxed{f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)}$$

The joint PMF/PDF must be nonnegative and sum/integrate to 1.



Conditional Distributions

Conditioning and Bayes' rule for discrete r.v.s

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

Conditioning and Bayes' rule for continuous r.v.s

$$\boxed{f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}}$$

Hybrid Bayes' rule

$$\boxed{f_X(x|A) = \frac{P(A|X = x)f_X(x)}{P(A)}}$$

Marginal Distributions

To find the distribution of one (or more) random variables from a joint PMF/PDF, sum/integrate over the unwanted random variables.

Marginal PMF from joint PMF

$$\boxed{P(X = x) = \sum_y P(X = x, Y = y)}$$

Marginal PDF from joint PDF

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy}$$

Independence of Random Variables Random variables X and Y are independent if and only if any of the following conditions holds:

- Joint CDF is the product of the marginal CDFs
- Joint PMF/PDF is the product of the marginal PMFs/PDFs
- Conditional distribution of Y given X is the marginal distribution of Y

Write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent.

Expectation

The expected value of a function $h(X, Y)$ of two jointly distributed random variables is

$$\boxed{E(g(X, Y)) = \sum_{x \in \mathbb{D}_1} \sum_{y \in \mathbb{D}_2} g(x, y)p(x, y)}$$

and can be generalized to the continuous case with integrations.

Correlation

Variance: is a numerical value that describes the **variability** of observations from its arithmetic mean.
Note: The more spread the data, the larger the variance is in relation to the mean.

$$\boxed{\sigma^2 = \frac{\sum(x_i - \mu)^2}{n} = E\left[(X - E[X])^2\right]}$$

Standard deviation: is a measure of the **dispersion** of observations within a data set.

$$\boxed{\sigma = \sqrt{\text{variance}}}$$

Key points:

1. Both variance and standard deviation are always positive.
2. If all the observations in a data set are identical, then the standard deviation and variance will be zero.
3. Standard deviation is preferred over mean as it is expressed in the same units as those of the measurements while the variance is expressed in the units larger than the given data set

Covariance: measures the **direction** of a relationship between two variables.
Covariance is a quantitative measure of the extent to which the deviation of one variable from its mean matches the deviation of the other from its mean.
Ex.: When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative.

$$\boxed{\text{Cov}(X, Y) = E\left[(X - E[X])\,(Y - E[Y])\right]}$$

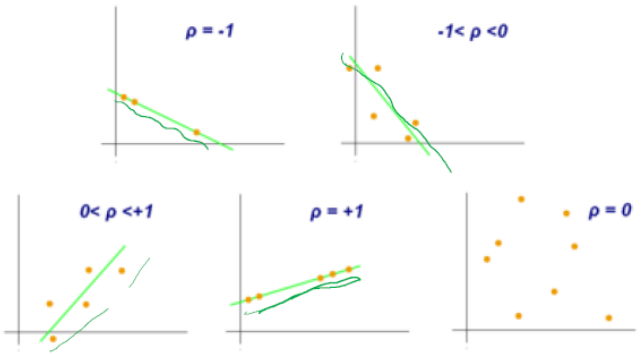
Special case: $\boxed{\sigma^2 = \text{Cov}(X, X)}$

Correlation: measures the **strength** of the linear relationship between two variables. It is normalized version of covariance.

$$\boxed{\text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}}$$

Pearson Correlation: $-1 \leq r \leq 1$, Range [-1, 1] $\rightarrow$ 0 is uncorrelated
For a sample:

$$\boxed{r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}}$$



Correlation and Independence
$\rightarrow$ Two random variables are **independent** are uncorrelated i.e. $\boxed{\text{Corr}(X,Y) = 0}$, if $P(X|Y) = P(X)$. Also, This is because if X and Y are independent, then one property is $E[XY] = E[X]E[Y]$
$\rightarrow$ Reverse is not necessarily true - some uncorrelated variables are dependent

Law of Large Numbers (LLN)

The LLN states that if you sample a random variable independently a large number of times, the measured average value should converge to the random variable's true expectation.

$$\boxed{\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \rightarrow \mu, \quad \text{as } n \rightarrow \infty}$$

This is important in studying the longer-term behavior of random variables over time. As an example, a coin might land on heads 5 times in a row, but over a much larger n we would expect the proportion of heads to be approximately half of the total flips. Similarly, a casino might experience a loss on any individual game, but over the long run should see a predictable profit over time.

Central Limit Theorem (CLT)

The CLT states that if you repeatedly sample a random variable a large number of times, the distribution of the sample mean will approach a normal distribution regardless of the initial distribution of the random variable. The normal distribution takes on

$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp - \left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

with the mean μ and standard deviation σ respectively.
The CLT states that:

$$\boxed{\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \rightarrow \sim N\left(\mu, \frac{\sigma^2}{n}\right);}$$

hence,

$$\boxed{\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)}$$

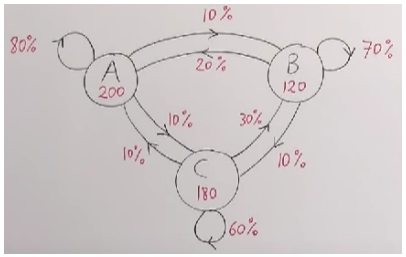
At a very basic level, you can consider the implications of this theorem on coin flipping: the probability of getting some number of heads flipped over a large n should be approximately that of a normal distribution.
Note: Whenever you are asked to reason about any particular distribution over a large sample size, you should remember to think of the CLT, regardless of whether it is Binomial, Poisson, or any other distribution.

Markov Chains

Markov chain is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules. The defining characteristic of a Markov chain is that no matter how the process arrived at its present state, the possible future states are fixed. In other words, the probability of transitioning to any particular state is dependent solely on the current state and time elapsed. The state space, or set of all possible states, can be anything: letters, numbers, weather conditions, baseball scores, or stock performances.
Ref: Markov Chain Tutorial

$$\boxed{[\text{Next State}] = \begin{bmatrix}\text{Matrix of}\\ \text{Transition}\\ \text{Probabilities}\end{bmatrix}[\text{Current State}]}$$

$$X_1 = PX_0$$



$$\begin{bmatrix}A_1\\B_1\\C_1\end{bmatrix} = \begin{matrix}A\\B\\C\end{matrix}\begin{bmatrix}0.8 & 0.2 & 0.1\\0.1 & 0.7 & 0.3\\0.1 & 0.1 & 0.6\end{bmatrix}\begin{matrix}A\\B\\C\end{matrix}\begin{bmatrix}A = 200 \rightarrow 0.4\\B = 120 \rightarrow 0.24\\C = 180 \rightarrow 0.36\end{bmatrix} = \begin{bmatrix}202 \rightarrow 0.404\\158 \rightarrow 0.316\\140 \rightarrow 0.28\end{bmatrix}$$

Why they call it chains ?

$$X_1 = PX_0 \quad X_2 = PX_1 \quad X_3 = PX_2$$

$$X_2 = \begin{bmatrix}0.8 & 0.2 & 0.1\\0.1 & 0.7 & 0.3\\0.1 & 0.1 & 0.6\end{bmatrix}\begin{bmatrix}0.404\\0.316\\0.28\end{bmatrix} = \begin{bmatrix}0.414\\0.346\\0.240\end{bmatrix} = \begin{bmatrix}207\\173\\120\end{bmatrix}$$

$$X_3 = \begin{bmatrix}0.8 & 0.2 & 0.1\\0.1 & 0.7 & 0.3\\0.1 & 0.1 & 0.6\end{bmatrix}\begin{bmatrix}0.414\\0.346\\0.240\end{bmatrix} = \begin{bmatrix}0.424\\0.356\\0.220\end{bmatrix} = \begin{bmatrix}212\\178\\110\end{bmatrix}$$

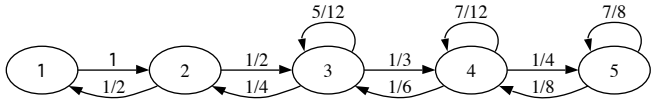
Another method to calculate chain:

$$X_n = P^n X_0 \quad \rightarrow \quad X_2 = P^2 X_0, \quad X_3 = P^3 X_0$$

When chain continues, it may end of up with same stable distribution matrix i.e., $X_s = X_{s-1}$

$$X_s = PX_{s-1} = \begin{bmatrix}0.8 & 0.2 & 0.1\\0.1 & 0.7 & 0.3\\0.1 & 0.1 & 0.6\end{bmatrix}\begin{bmatrix}\text{val1}\\\text{val4}\\\text{val5}\end{bmatrix} = \begin{bmatrix}\text{val1}\\\text{val4}\\\text{val5}\end{bmatrix}$$

Definition



A Markov chain is a random walk in a **state space**, which we will assume is finite, say $\{1, 2, \ldots, M\}$. We let X_t denote which element of the state space the walk is visiting at time t. The Markov chain is the sequence of random variables tracking where the walk is at all points in time, $X_0, X_1, X_2, \ldots$. By definition, a Markov chain must satisfy the **Markov property**, which says that if you want to predict where the chain will be at a future time, if we know the present state then the entire past history is irrelevant. *Given the present, the past and future are conditionally independent*. In symbols,

$$P(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \ldots, X_n = i) = P(X_{n+1} = j | X_n = i)$$

State Properties

A state is either recurrent or transient.

- If you start at a **recurrent state**, then you will always return back to that state at some point in the future. ♪*You can check-out any time you like, but you can never leave.* ♪
- Otherwise you are at a **transient state**. There is some positive probability that once you leave you will never return. ♪*You don't have to go home, but you can't stay here.* ♪

A state is either periodic or aperiodic.

- If you start at a **periodic state** of period k, then the GCD of the possible numbers of steps it would take to return back is $k > 1$.
- Otherwise you are at an **aperiodic state**. The GCD of the possible numbers of steps it would take to return back is 1.

Transition Matrix

Let the state space be $\{1, 2, \ldots, M\}$. The transition matrix Q is the $M \times M$ matrix where element q_{ij} is the probability that the chain goes from state i to state j in one step:

$$q_{ij} = P(X_{n+1} = j | X_n = i)$$

To find the probability that the chain goes from state i to state j in exactly m steps, take the (i,j) element of Q^m.

$$q_{ij}^{(m)} = P(X_{n+m} = j | X_n = i)$$

If X_0 is distributed according to the row vector PMF $\vec{p}$, i.e., $p_j = P(X_0 = j)$, then the PMF of X_n is $\vec{p}Q^n$.

Chain Properties

A chain is **irreducible** if you can get from anywhere to anywhere. If a chain (on a finite state space) is irreducible, then all of its states are recurrent. A chain is **periodic** if any of its states are periodic, and is **aperiodic** if none of its states are periodic. In an irreducible chain, all states have the same period.

A chain is **reversible** with respect to $\vec{s}$ if $s_i q_{ij} = s_j q_{ji}$ for all i, j. Examples of reversible chains include any chain with $q_{ij} = q_{ji}$, with $\vec{s} = (\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M})$, and random walk on an undirected network.

Stationary Distribution

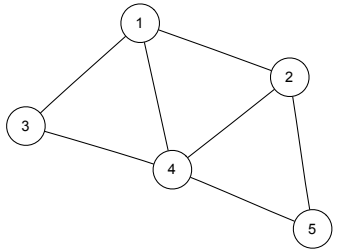
Let us say that the vector $\vec{s} = (s_1, s_2, \ldots, s_M)$ be a PMF (written as a row vector). We will call $\vec{s}$ the **stationary distribution** for the chain if $\vec{s}Q = \vec{s}$. As a consequence, if X_t has the stationary distribution, then all future $X_{t+1}, X_{t+2}, \ldots$ also have the stationary distribution.

For irreducible, aperiodic chains, the stationary distribution exists, is unique, and s_i is the long-run probability of a chain being at state i. The expected number of steps to return to i starting from i is $1/s_i$.

To find the stationary distribution, you can solve the matrix equation $(Q' - I)\vec{s}' = 0$. The stationary distribution is uniform if the columns of Q sum to 1.

Reversibility Condition Implies Stationarity If you have a PMF $\vec{s}$ and a Markov chain with transition matrix Q, then $s_i q_{ij} = s_j q_{ji}$ for all states i, j implies that $\vec{s}$ is stationary.

Random Walk on an Undirected Network



If you have a collection of **nodes**, pairs of which can be connected by undirected **edges**, and a Markov chain is run by going from the current node to a uniformly random node that is connected to it by an edge, then this is a random walk on an undirected network. The stationary distribution of this chain is proportional to the **degree sequence** (this is the sequence of degrees, where the degree of a node is how many edges are attached to it). For example, the stationary distribution of random walk on the network shown above is proportional to $(3, 3, 2, 4, 2)$, so it's $(\frac{3}{14}, \frac{3}{14}, \frac{2}{14}, \frac{4}{14}, \frac{2}{14})$.

Statistics

Descriptive Statistics:

Used to describe, organize and summarize information about an entire population i.e. 90% satisfaction of all customers.

• **Where is the variable centered?** called **measure of central of tendency**. → mean, median, mode

→ Outliers draw the **mean** towards them! (also effect SD), critical for computation and prediction

→ **Median** increases with increase in small values, **good for heavily skewed data** or large outliers

• **How spread out the data is?** called **measure of variability**

→ **Variance** (mean square distance from the mean) and SD

→ **Inter-quartile range:** width of "middle 50%"

Types of Datasets

- **Structured** Ex: Table format i.e. based on row(s) and column(s)
- **Unstructured** Ex: unknown form - music, video, audio
- **Semi-structured** Ex: JSON, CSV, XML, graph data, @email

Types of Variables

The level of measurement of your data determines which statistical tests are appropriate to use.

→ **Quantitative** → Numerical

- **Continuous:** can take on any value (in allowable range).
  Ex: A player's height (176.5cm), distance from KTM to JNK (200km)
  - **Interval** - Measurable, but arbitrary zero point (**Note**: zero point is a point where all values start)
    Ex: credit scores (300-850)
  - **Ratio** - Measurable and meaningful zero point.
    Ex: Weight, length, Temp. in Kelvin (in this scale zero marks the point, cannot go below it)
- **Discrete:** takes on distinct (whole numeric) values.
  → **Integers** are discrete. **Ex:** score in a soccer game (5 goals), how many siblings? (3 siblings),

→ **Qualitative** → Categorical (Descriptive)

- **Nominal**: Categories with no ranking or natural order. All options have the **same** value.
  Ex: Gender, Ethnicity, Personal preferences (favorite meal or color),
- **Ordinal**: Categories with an inherent rank or order. Each options has a **different** value.
  Ex: Income levels (low, medium, high), Levels of agreement (Agree, Neutral, Disagree)

Presentation

• The goal of data presentation to communicate what we learned from the data, the evidence for our conclusions, our goals and/or questions for the analysis

→ Three primary ways to compare statistics: **absolute**, **relative**, and **ratio**

• **Bar chart:** Emphasize *relative difference*. A bar chart is used for **categorical** data, where each bar represents a distinct category with variable heights indicating values like counts or percentages. It has a **categorical x-axis** and a **numerical y-axis**, showing comparisons between different categories (e.g., sales figures for different products).

• **Histogram:** It displays the distribution of **numerical** data across bins or intervals. Bars are contiguous and represent the frequency or count of data points within each bin on the x-axis. It has numerical axes (x for bins, **y for frequencies**), helping visualize data distributions (e.g., exam scores distribution).

• **Box plot:** The distribution of a numeric variable within groups identified by a categorical variable.

• **Scatter plot:** Where observations fall in a two-dimensional space defined by two numeric variables

• **line plot:** How a numeric variable changes from one value to the next of another numeric variable

• **Point plot:** Emphasize *absolute difference*, also relative difference-in-difference

• **Violin plot:** like box, but mean-based

• **Swarm plot:** categorical scatter plot

Inferential Statistics

Used to generalize about a population **based on a sample of data** i.e. 90% satisfaction of a sample of 50 customer → 90% satisfaction of all customers.

Inference is learning about data

→ **Estimating** the value of a parameter

→ **Testing** the data's support for an hypothesis.

→ **Ex:** Mean of 500 penguins is **estimate** (is an estimated value often has a **CI** or similar), mean as the concept is the **estimator**, and the parameter is **estimand**

• **Sampling Distribution**: refers to the distribution of a **sample** statistic (such as the sample mean or sample proportion) obtained from **multiple samples of the same size** from a population.

→ The sampling distribution of the sample mean tends to **Normal** as the sample size increases, regardless of the population distribution, this is **Central limit theorem**

$$\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

→ The CLT is **important** because it allows us to use this bell curve to understand and make predictions about groups of data, even when we don't know the exact details of each individual measurement.

• **Sample statistic**: A metric calculated for a sample of data drawn from a large population.

• **Standard error** The standard error is a single metric that sums up the variability (SD) in the sampling distribution for a statistic.
SE $= \frac{s}{\sqrt{n}}$

• **Bootstrap**: A technique for **estimating sampling distributions** by repeatedly resampling the available (existing) sample **with replacement**.

→ **QQ-Plot**: A plot to visualize how close a sample distribution is to a specified distribution, e.g., the normal distribution

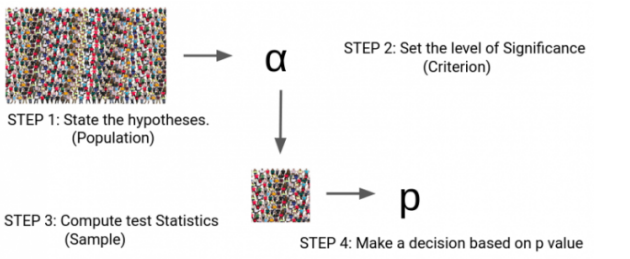
Hypothesis Testing

Hypothesis Test (or significance test) is used to assess & understand the plausibility (reasonable), or likelihood of some assumed viewpoint (a hypothesis) - based upon data.

Example: Let's say we're the coach of an NBA Basketball team - we might find ourselves with the below dilemma...

The New Sensation ≫ Games Played: 2 → Shooting Rate: 60%

The Current Star ≫ Games Played: 102 → Shooting Rate: 49% Here, My new player seems to be a better shooter...but the player only played 2 games. I need more confidence before making a big change! Coach Let's test it!



Example: Since the null hypothesis typically represents a baseline (e.g., the marketing campaign did not increase conversion rates, etc), the goal is to reject the null hypothesis with statistical significance and hope that there is significant outcome.

• In hypothesis testing, the probability that the null hypothesis (H_0) would produce a value as large as the observed value; if the observed statistic is x, and X is a random variable representing the sampling and analysis process, this is $\boxed{P[X > x\,|H_0 \text{ is true}]}$. The **p-value** is the probability of seeing an effect as large as the one observed if there is no true effect to observe.

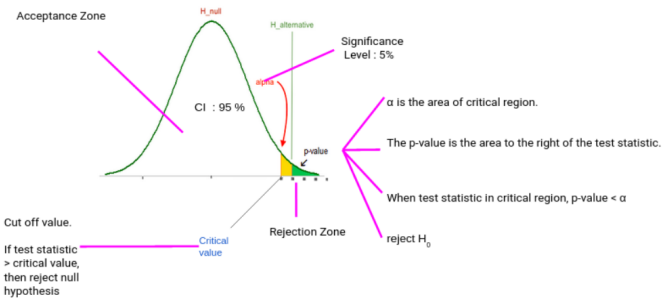
p-value and Confidence Intervals

• **p-value** is the probability of observing the value of the calculated test statistic under the null hypothesis assumptions.

• A p-value is not a probability of an event occurring, it is a probability or likelihood of seeing a different result if we were to sample many times

• A p-value does not tell us how different two samples are.Two samples with the same difference in means, but larger/smaller samples sizes will get different p-values. A p-value is instead telling us how likely it is that they are different (or in other words how confident we can be that they are different)

→ **laymen term p-value**: If I'm living in a world where the pizza delivery time is 30 minutes or less (null hypothesis is true), how surprising is my evidence in real life? P-value answers this question with a number — probability.



• **Bootstrapping the p-value**: Compute statistics t (observed diff.), t^* from each bootstrap sample (diff.) → $\boxed{p = P(|t^*| \geq |t|)}$

→ If $p < 0.05$ then reject H_0, which means there is significantly difference.

• **Pitfalls to p-values:**

Multiple comparisons, Designed for prospective experiemnts, Full validity requires planning before looking at data, Null hypothesis usually not precisely true.

• **Confidence interval** (CI) $\boxed{95\% : \bar{x} \pm 1.96\,(\frac{s}{\sqrt{n}})}$ computed by taking a sample (n) and computing a statistic from that sample, when repeated many times, will **return an interval** containing the true parameter value (mean) 95% of the time.

Note: If the confidence intervals **do not overlap** → This is evidence to reject null hypothesis, i.e. different

→ Hypothesis tests are either one- or two-tailed tests.

• **One-tailed** test:

$\boxed{H_0 : \mu = \mu_0}$ versus $\boxed{H_1 : \mu < \mu_0 \text{ or } H_1 : \mu > \mu_0}$

• **Two-tailed** test:

$\boxed{H_0 : \mu = \mu_0}$ versus $\boxed{H_1 : \mu \neq \mu_0}$

where H_0 is the null hypothesis and H_1 is the alternative hypothesis, and μ is the parameter of interest.

Note: Understanding the hypothesis testing is the basis of A/B testing.

A/B Testing

→ An A/B test is an experiment with two groups to establish which of two treatments, products, procedures, or the like is superior. Two groups:

→ **Treatment group**: A group of subjects exposed to a specific treatment.

→ **Control group**: A group of subjects exposed to no (or standard) treatment.

Ex: Testing two therapies to determine which suppresses cancer more effectively

Ex: Testing two web ads to determine which generates more conversions

• **Effect Size:** Let's say we divide the penguins into two groups, give food1 to group1 and food2 to group2, then compare their growths. Here, the size of the difference is **effect size**.

Note: Effect size and estimates are **usually** more important than hypothesis tests.

Test Statistics

• **z-Test**: Generally, z-Test is used when the **sample size** n is **large** $n \geq 30$ (to invoke the CLT) and when the population **variance** (or standard deviation) is **known**.

→ **One-Sample** z-Test: compare a **sample mean** $\bar{x}$ with the **population mean** μ

$$\boxed{z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)}$$

→ **Two-Sample** z-Test: compare the **mean** of two samples

$$\boxed{z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$

→ **z-score:** gives you an idea of how far from the mean a data point is. **z-score will tell you how many standard errors there are between the sample mean and the population mean.** A z-score of zero tells you the values is exactly average while a score of +3 tells you that the value is much higher than average.

• **Student's t-Distribution**: The *t-distribution* is a normally shaped distribution, except that it is a bit thicker and longer on the tails.

• **t-Test**: used when the **sample size** n is **small** $n < 30$ and when the population **variance** (or standard deviation) is **unknown**. It uses sample variance s^2 in place of population variance σ^2.

1. **One-sample** *t-test* tests whether a single mean is different from zero (or another fixed value μ_0). $\boxed{H_0 : \mu = 0}$

$$\boxed{t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}}$$

2. **Two-sample** *t-test* that tests whether the means of **two independent samples** are the same. $\boxed{H_0 : \mu_1 = \mu_2}$

$$\boxed{t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

3. **Paired** *t-test* that tests, for a **same sample** of paired observations (within-subjects experiment), whether the mean difference between observations for each sample is zero. $\boxed{H_0 : E[x_{i1} - x_{i2}] = 0}$ • $x_{i*} = x_{i1} - x_{i2}$ → $\boxed{H_0 : \mu_* = 0}$

→ Use a **two-sample t-test** to compute $\boxed{p = P(|\bar{X}_G - \bar{X}_C| \text{ this big } |H_0)}$, if p low (often $p < 0.05$), **reject the null** (there is a difference).

• **Ch-Square Distribution**: measure differences between categorical variables, using $\boxed{\chi^2 = \sum \frac{\text{observed} - \text{expected}}{\text{expected}}}$ to test:

- Goodness of fit - if samples of one categorical variable match the population category expectations
- Independence - if being in one category is independent of another, based off two categories
- Homogeneity - if different subgroups come from the same population, based off a single category

• **MLE** and **MAP** : Any probability distribution has parameters, so fitting parameters is an extremely crucial part of data analysis. There are two general methods for doing so. In **maximum likelihood estimation** (MLE), the goal is to estimate the most likely parameters given a likelihood function:

$$\boxed{\theta_{\text{MLE}} = \arg\max L(\theta)}, \quad \text{where } L(\theta) = f_n\,(x_1, \ldots x_n|\theta)$$

Since the values of X are assumed to be i.i.d., then the likelihood function becomes the following:

$$\boxed{L\,(\theta) = \prod_{i=1}^{n} f\,(x_i|\theta)}$$

The natural log of $L(\theta)$ is then taken prior to calculating the maximum; since log is a monotonically increasing function, maximizing the **log-likelihood** log $L(\theta)$ is equivalent to maximizing the likelihood:

$$\boxed{\log\,L(\theta) = \sum_{i=1}^{n} \log\,f\,(x_i|\theta)}$$

Another way of fitting parameters is through **maximum a posterior** estimation (MAP), which assumes a "prior distribution":

$$\boxed{\theta_{\text{MAP}} = \arg\max g(\theta) f(x_1 \dots x_n|\theta)}$$

where the similar **log-likelihood** is again employed, and $g(\theta)$ is a density function of θ.