

# Probability & Statistics

© 2024 Robins Yadav

## 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.

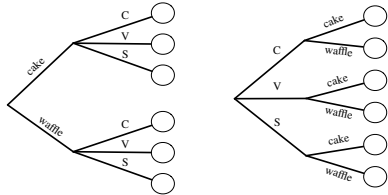
### Naive Definition of Probability

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

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

## 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  $r$ th component has  $n_r$  possible outcomes, then overall there are  $n_1 n_2 \dots n_r$  possibilities for the whole experiment.

### Permutation & Combination

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

$${}_n P_r = \frac{n!}{(n-r)!}, \quad 0 \leq r \leq n$$

→ **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  $0 < r \leq n$ , and where the **repetition is not allowed** is

$${}_n P_r \rightarrow \frac{n!}{(n-r)!} = n(n-1)(n-2) \dots (n-r+1)$$

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

$$n^r$$

→ **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

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

**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  $k^{\text{th}}$  kind and the rest, if any, are of a different kind, then the permutation is given by

$$\frac{n!}{n_1! n_2! \dots n_k!}$$

- **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

$${}_n C_r = \frac{n!}{(n-r)! r!} = \frac{{}_n P_r}{r!} = \frac{n(n-1)(n-2) \dots (n-r+1)}{r!}$$

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

$${}_n C_r = \frac{(r+n-1)!}{r! (n-1)!}$$

**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  $P(A \cap B|C) = P(A|C) P(B|C)$ .

→ 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.

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c$$

**Complements:** The following are true.

$$A \cup A^c = \Omega$$

$$A \cap A^c = \emptyset$$

$$P(A) = 1 - P(A^c)$$

## Joint Probability

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

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

## Conditional Probability

**Conditional Probability**, with all outcomes equally likely.

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

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

**Note:**  $P(A|B) \neq P(B|A)$

- **Law of Total Probability:**

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

$$\Omega = \bigcup_{i=1}^n A_i$$

## Marginal Probability

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

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

## 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

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

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

$$P(A_i|B) = \frac{P(B|A_i) P(A_i)}{\sum_{j=1}^n P(B|A_j) P(A_j)}$$

$$\Omega = \bigcup_{i=1}^n A_i$$

Likelihood and Posterior

P(A|B) = (P(B|A) P(A)) / P(B)

- P(A) is the prior
- P(B|A) is the likelihood
- P(B) = ∫ 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 (Ω, B, P), respectively the sample space, event space, and probability function.

Ω: sample space → Set of outcomes of an experiment.

Example: tossing a coin twice. Ω = {HH, HT, TH, TT}

B: event space → 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 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) ∅ ∈ B
- (2) If an event A ∈ B, then A^c ∈ B (closed under complementation)
- (3) If A1, A2, ... ∈ B, then ∪\_{i=1}^∞ Ai ∈ B (closed under countable union).
- Additional properties:
- (4) (1) + (2) → Ω ∈ B
- (5) (3) + De-Morgan's Laws → ∩\_{i=1}^∞ Ai ∈ B (closed under countable intersection)

Example: tossing a coin twice  
Ω = {HH, HT, TH, TT}, then B = 2^Ω = 16, and therefore σ-field can be given by  
{∅, {HH}, {HT}, {TH}, {TT}, {HH, HT}, {HH, HT, TH}, ... , Ω}  
Note: Observable events could be the collection of all intervals, plus what can be obtained by set operations ∪, ∩, c.

Suppose we cannot distinguish between HT and TH, then we have sample space Ω = {HH, HT, TT}, and B = 2^Ω = 8, therefore σ-field can be given by  
{∅, {HH}, {HT}, {TT}, {HH, HT}, {HH, TT}, {HT, TT}, Ω}

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 B, until it satisfies all requirements of a σ-field. When we go through this process, we obtain the σ-algebra,  
B = {∅, {TT}, {HH, HT, TH}, ... , Ω }

P: probability measure → a function that assigns numbers (called probabilities) between 0 and 1 to each event in B i.e. B → [0, 1]

In other words, a probability measure is a function P : B → [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: P(A) ≥ 0, for all A ∈ B → ensures that probability is never negative.
- 2. Normalization: P(Ω) = 1. → ensures that probability is never greater than 1.
- 3. Countable Additivity: If A1, A2, ... ∈ B are pairwise disjoint (i.e.,

Ai ∩ Aj = ∅, if i ≠ j), then P(∪\_{i=1}^∞ Ai) = ∑\_{i=1}^∞ P(Ai)

Theorem: Let (Ω, B, P) be a probability space.

- 1. P(A^c) = 1 - P(A) for A ∈ B
- 2. P(A ∪ B) = P(A) + P(B) - P(A ∩ B). for A, B ∈ B
- 3. P(A/B) = P(A) - P(A ∩ B). for A, B ∈ B
- 4. If A ⊂ B, then P(A) ≤ P(B) for A, B ∈ B

5. If A1, A2, ... ∈ B are disjoint, then P(∪\_{j=1}^n Aj) = ∑\_{j=1}^n P(Aj)

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	1/4
HT	1/4
TH	1/4
TT	1/4
∅	0
Ω	1
{HH, HT, TH}	3/4 (using pt. (3) of Def'n above)
{HH, HT}	1/2
⋮	⋮

Example:: A coin toss.  
Ω : {H, T}  
B : {∅, {H}, {T}, {H, T}}  
P: P(∅) = 0, P({H}) = 1/4, P({T}) = 3/4, P({H,T}) = 1

Random Variable

Consider an experiment with four outcomes Ω = {♣, ◇, ♥, ♠}. We want to construct the probability space (Ω, B, P). The sample space Ω is already defined. The event space 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[{♣}] = 1/6, P[{◇}] = 2/6, P[{♥}] = 2/6, P[{♠}] = 1/6

Therefore, we have constructed a probability space (Ω, 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 ♣, ◇, ♥, ♠. 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, ♣ ← 1, ◇ ← 2, ♥ ← 3, ♠ ← 4,

where the arrow that we assign a number to the symbol. But we can express this more formally by defining a function X : Ω → ℝ with

X(♣) = 1, X(◇) = 2, X(♥) = 3, X(♠) = 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] = 1/6, P[X = 2] = 2/6, P[X = 3] = 2/6, P[X = 4] = 1/6

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

A random variable X on a probability space (Ω, B, P) is a function that maps every element in a sample space to a real line i.e., X : Ω → 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

Ω = {(head, head), (head, tail), (tail, head), (tail, tail)}

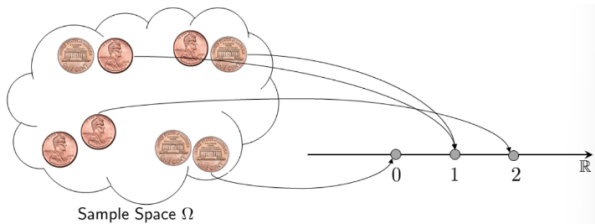
Suppose that X is a random variable that maps an outcome to a number representing the sum of "head", i.e.,

X(.) = number of heads.

Then, for the four ξ's in the sample space there are only 3 distinct numbers. More precisely, if we let ξ1 = (head, head), ξ2 = (head, tail), ξ3 = (tail, head), ξ4 = (tail, tail), then, we have

X(ξ1) = 2, X(ξ2) = 1, X(ξ3) = 1, X(ξ4) = 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.  
→ 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 ~ Bin(n,p)
dt(x,n)	PDF f(x) for X ~ tn
dunif(x,a,b)	PDF f(x) for X ~ Unif(a,b)

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

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

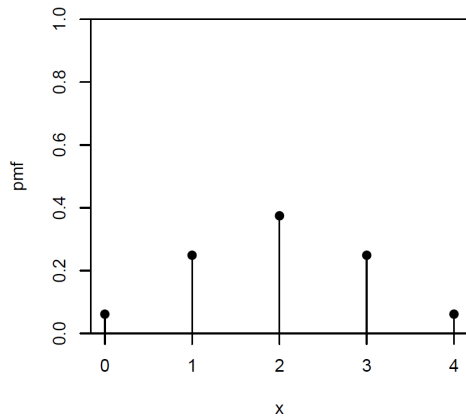
#### • Properties of Variance

- $V(aX + b) = a^2 \cdot \sigma^2$
- In particular,  $\sigma_{aX} = |a| \cdot \sigma_x$
- $\sigma_{X+b} = \sigma_X$

### 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

$$p_X(x) = P[X = x]$$



The PMF satisfies,

$$p_X(x) \geq 0, \text{ and } \sum_x p_X(x) = 1$$

→ The **probabilities** are summarized by a function known as the probability mass function (PMF).

→ PMFs are the ideal histograms of random variables.

→ Expectation = Mean = Average computed from a PMF

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

$$X("HH") = 2, X("TH") = 1, X("HT") = 1, X("TT") = 0$$

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

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

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

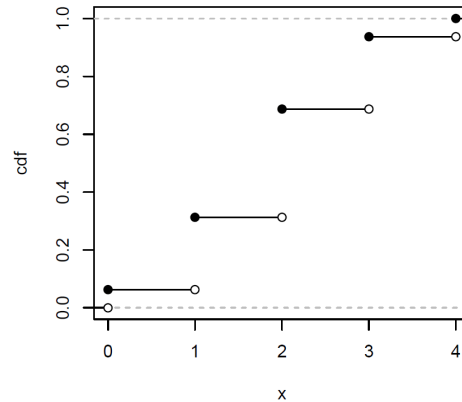
$$p_X(2) = P[X = 2] = P[\{"HH"\}] = \frac{1}{4}$$

→ 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  $x$ .

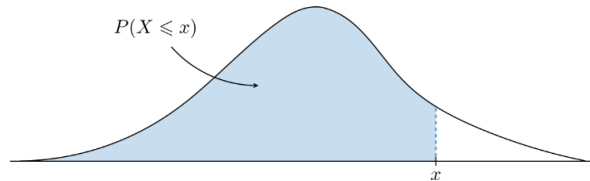
$$F_X(x) = P(X \leq x)$$



The CDF is an increasing, right-continuous function with

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

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



**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).

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

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

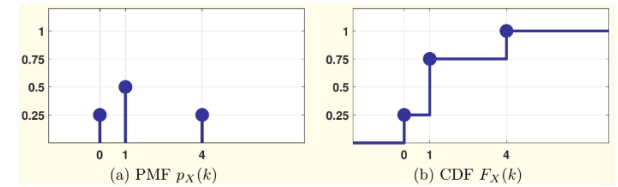
$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

**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

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

Continue the above example

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

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

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

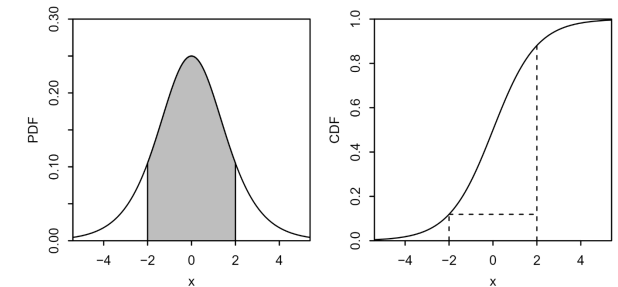
### Probability Density Function (PDF)

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  $f_X : \Omega \rightarrow \mathbb{R}$  that, when integrated over an interval  $[a, b]$  yields the probability of obtaining  $a \leq X \leq b$ :

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

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.

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

**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:** Continuous 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.

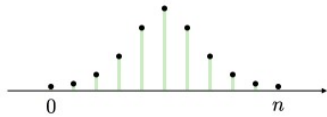
## Classic Statistical Distribution

• **Binomial Distribution (Discrete)**:  $x$  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  $1 - p$ . **Example**: Used in quality control and polling. Assume  $X$  is distributed  $\text{Bin}(n, p)$ .

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

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

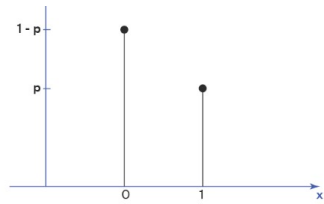


**Negative Binomial**: Number of failures before  $r$  successes.

• **Bernoulli Distribution**: if  $n = 1$ , this is a Bernoulli distribution. **Example**: Binary outcomes like flipping a coin, yes/no surveys

$$\mu = p$$

$$\sigma^2 = p(1-p)$$



• **Geometric**: 1<sup>st</sup> success with  $p$  probability on the  $n$  trials. Tracks visits before the first Ad click

$$\text{PDF: } P(X = x) = q^{n-1}p$$

$$\mu = \frac{1}{p}$$

$$\sigma^2 = \frac{1-p}{p^2}$$

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

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

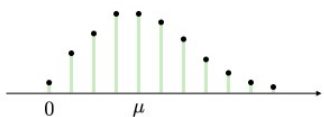
$$\mu = \frac{nX}{N}$$

• **Poisson Distribution (Discrete)**: Counts in fixed intervals. **Example**: No. of calls received per call center. Assume  $X$  is distributed  $\text{Pois}(\lambda)$ . Number of  $x$  successes in a fixed interval of time/space, where these success occur independently and with a known constant or average rate  $\lambda$ .

$$\text{PDF: } P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

$$\mu = \lambda$$

$$\sigma^2 = \lambda$$



• **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  $\alpha$  is the law's exponent and  $c$  is the normalizing constant.

• **Uniform Distribution (Continuous)**: All values have equal probabilities. **Example**: Used in random number generator

$$\text{PDF: } P(x) = \frac{1}{b-a}$$

$$\mu = \frac{a+b}{2}$$

$$\sigma^2 = \frac{(b-a)^2}{12}$$



• **Normal/Gaussian Distribution (Continuous)**:

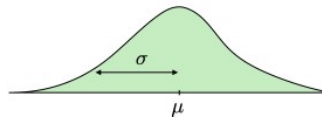
Assume  $X$  is distributed  $\mathcal{N}(\mu, \sigma^2)$ . It is a bell-shaped and symmetric distribution. Bulk of the values lie close to the mean and no value is too extreme. Generalization of the binomial distribution as  $n \rightarrow \infty$ .

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

$$\mu = \mu$$

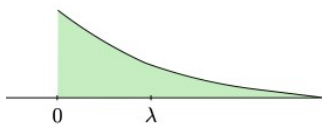
$$\sigma^2 = \sigma^2$$

**Implications**: 68%-95%-99% rule. 68% of probability mass fall within  $1\sigma$  of the mean, 95% within  $2\sigma$ , and 99.7% within  $3\sigma$ .  
→ Normal Approximation - discrete distributions such as Binomial and Poisson can be approximated using  $z$ -scores when  $np$ ,  $nq$ , and  $\lambda$  are greater than 10.



• **Exponential** - Time between events. **Example**: Explains inactivity time among users. Memoryless time between independent event

occurring at an average rate  $\lambda \rightarrow \lambda e^{\lambda x}$ , with  $\mu = \frac{1}{\lambda}$



• **Gamma** - Time between K-th success. **Example**: Explains load on web servers over time. Time until  $n$  independent events occurring at an average rate  $\lambda$ .

## Joint PDFs and CDFs

### Joint Distributions

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

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

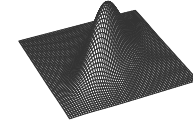
In the discrete case,  $X$  and  $Y$  have a **joint PMF**

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

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

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

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



## Conditional Distributions

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

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

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

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

### Hybrid Bayes' rule

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

## 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**

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

### Marginal PDF from joint PDF

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

**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

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

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.

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

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

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

**Key points:**

- Both variance and standard deviation are always positive.
- If all the observations in a data set are identical, then the standard deviation and variance will be zero.
- 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.

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

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

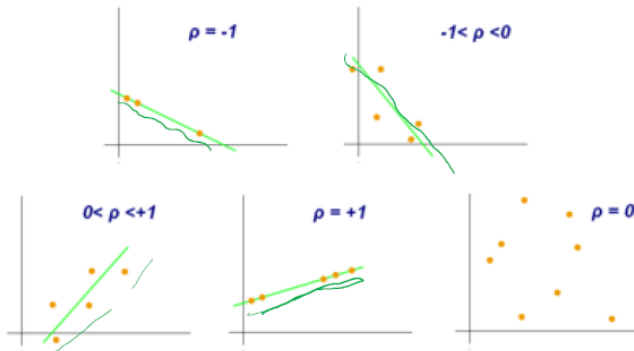
• **Correlation:** measures the **strength** of the linear relationship between two variables. It is normalized version of covariance.

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

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

For a sample:

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



**Correlation and Independence**

$\rightarrow$  Two random variables are **independent** are uncorrelated i.e.

$\text{Corr}(X, Y) = 0$ , 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.

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

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

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

with the mean  $\mu$  and standard deviation  $\sigma$  respectively.

The CLT states that:

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

hence,

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

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.

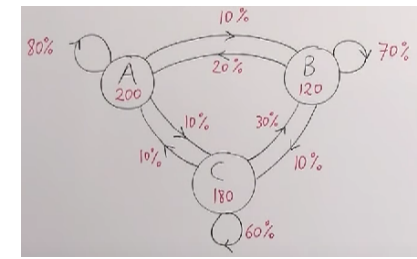
## Markov Chains

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

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

$$X_1 = PX_0$$



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

**Why they call it chains ?**

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

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

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

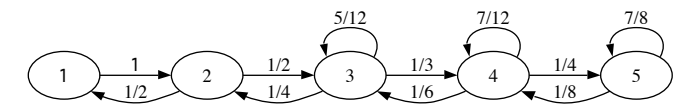
**Another method to calculate chain:**

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

When chain continues, it may end 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  $\{1, 2, \dots, M\}$ . 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,  $X_0, X_1, X_2, \dots$ . 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,

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



## 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  $\{1, 2, \dots, M\}$ . 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  $\vec{s} = (\frac{1}{M}, \frac{1}{M}, \dots, \frac{1}{M})$ , and random walk on an undirected network.

## Stationary Distribution

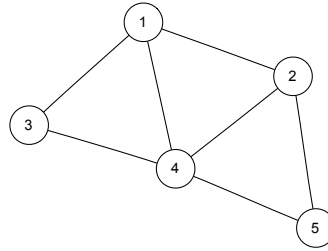
Let us say that the vector  $\vec{s} = (s_1, s_2, \dots, s_M)$  be a PMF (written as a row vector). We will call  $\vec{s}$  the **stationary distribution** for the chain if  $\vec{s}Q = \vec{s}$ . As a consequence, if  $X_t$  has the stationary distribution, then all future  $X_{t+1}, X_{t+2}, \dots$  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:** Outliers draw the **mean** towards them! (also effect SD), critical for computation and prediction. For a dataset  $x_1, x_2, \dots, x_n$

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

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

$$\text{Median} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even} \end{cases}$$

→ **Mode** = value that appears most frequently

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

→ **Variance:** measures how far the values in a dataset spread out from the mean (average).

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

→ **Standard Deviation:** Variance gives extra weight to outliers due to Squared values, an interpretable measure of spread because it is expressed in the same units as the data.

$$\sigma = \sqrt{\sigma^2}$$

→ **Coefficient of Variation (CV):** measures the relative variability of a dataset by comparing the standard deviation to the mean.

$$CV = \frac{\sigma}{\mu} \times 100\%$$

→ **Covariance:** measure of the degree to which two random variables change together. It indicates the **direction** of the linear relationship between the variables.

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y)$$

- A positive covariance indicates that as one variable increases, the other also tends to increase.

- A negative covariance indicates that the other tends to decrease as one variable increases.

- A covariance close to zero suggests that the variables do not have a linear relationship.

→ **Correlation Coefficients:** measures the **strength** and **direction** of a linear relationship between two variables. The formula for Pearson's correlation coefficient is:

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

The value of  $r$  ranges from  $-1$  to  $1$ :

- $r = 1$ : Perfect positive correlation.
- $r = -1$ : Perfect negative correlation.
- $r = 0$ : No correlation.

→ **Inter-quartile range:** width of "middle 50%".  $IQR = Q3 - Q1$   
lower bound =  $Q1 - 1.5 * IQR$ , upper bound =  $Q3 + 1.5 * IQR$

## 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

→ 3 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 from a sample of 50 customer → 90% satisfaction from 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.

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

- **Central Limit Theorem (CLT)** states that the sampling distribution of the sample mean approaches a **normal distribution** as sample size increases, regardless of the population **distribution** (i.e. uniform, exponential etc), given that the sample consists of independent and identically distributed i.e. i.i.d random variables with **finite** mean and variance.

$$Z_n = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0, 1)$$

The standardized sample mean  $Z_n$  converges to a standard normal distribution  $N(0, 1)$  as the sample size  $n$  increases.

→ When we perform experiments or collect data, we often don't know the underlying distribution of the population (it could be **skewed**, **uniform**, or anything else). This is where the **Central Limit Theorem (CLT)** becomes crucial.

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

- **Standard error** Measures the accuracy of the estimate  $SE = \frac{s}{\sqrt{n}}$

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

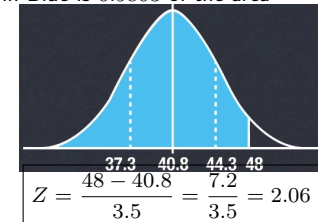
## Normal Distribution

We can compute the Z Score and use the Normal Distribution to calculate probabilities

- **Z-score:** gives the value in standard deviation ( $\sigma$ ) for the Percentile we want. For example, if we want 95% of the data, it tell us how many standard deviations are required.

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

- **Z Table:** Area in Blue is 0.9803 of the area



→ **z-score** will tell you how many standard errors there are between the sample mean and the population mean.

→ 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.

## Hypothesis Testing

→ **Hypothesis:** An educated guess you can test, or likelihood of some assumed viewpoint - based upon data.

→ **Statement:** Restaurants in Los Angeles are Expensive

→ **Hypothesis:** Restaurants in Los Angeles are Expensive vs. Pittsburgh

• **Null Hypothesis:** The Hypothesis you will Test

$$10 < \text{Avg. Restaurant Prices} < 25$$

• **Alternative Hypothesis:** All other Possibilities

$$\text{Avg. Restaurant Prices} < 10 \text{ OR } \text{Avg. Restaurant Prices} > 25$$

• **Significance Level ( $\alpha$ ):** the probability of rejecting the Null Hypothesis ( $H_0$ ) when it is actually true

• Common  $\alpha$ : 0.01, 0.05, 0.

• **Confidence Intervals:** Point Estimates are Inaccurate so alternative is Intervals. Suppose we have 3 sample means

$$\mu_{\bar{x}} = 5.4, \mu_{\bar{x}} = 6.2, \mu_{\bar{x}} = 6.8 \rightarrow \text{Interval} : (5, 7)$$

→ Normal Confidence: 90%, 95%, 99%

→ 90% Confidence: 9/10 of Intervals contain Mean

→ Alpha ( $\alpha$ ): Doubt → 1- Confidence

$$(x, y) = \bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

**Example:** Calculate the probable salary we would receive if we became a player for the Houston Rockets with confidence of 95%.

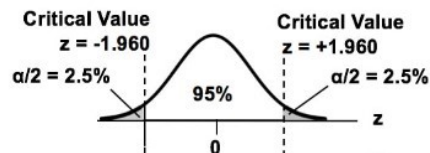
Given  $\mu_{\bar{x}} = 8978814.2$ , **confidence** = 0.95, **significance level**  $\alpha = 0.05$ , **critical probability** =  $1 - \frac{\alpha}{2} = 0.975$ .

Step 1: Look into the Z Table with value 0.975, will give  $Z_{\frac{\alpha}{2}}$  of 1.96

Step 2: Find standard deviation  $\sigma$  is 12471425.74

Step 3:

$$(x, y) = 8978814.2 \pm 1.96 \cdot \frac{12471425.74}{\sqrt{15}} = (2667401.97, 15290226.43)$$



$$95\% : \bar{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

### p-value and Confidence Intervals

• **p-value** is the smallest level of significance at which we Reject the Hypothesis.

• **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.

## Z-Test

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

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

1.  $H_0 : \mu = \text{avg. value}, H_1 : \mu < \text{avg. value}$

2. **Collect Data:**  $\bar{x}, \mu, \sigma, n > 30$

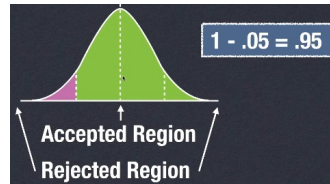
3. Compute **Z-Score:** 
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

4. For a 95% **confidence level** (one-tailed test), **significance level**  $\alpha = 0.05$ , **critical probability** =  $1 - \alpha = 0.95$ , the critical value (from z-table) is -1.645.

5. **Decision Rule:** if Z-score  $< -1.645$ , reject Null Hypothesis  $H_0$

6. Using **p-value** method: Use **Z-Score**  $\pm$  (test statistic), and Look into z-table. Let say **Z-Score** is -2.53, then the row with 2.5 and column with 0.03, the area value is 0.0057. For the area only in left side  $\alpha$  is 0.0057 Since, it's 1-sided test so the p-value is 0.0057.

$p\text{-value} < \alpha \rightarrow 0.0057 < 0.05$ , thus **reject** Null Hypothesis  $H_0$



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

1.  $H_0 : \mu_A = \mu_B, H_1 : \mu_A \neq \mu_B$

2. **Collect Data:**  $\bar{x}_A, \bar{x}_B, \sigma_A, \sigma_B, n_A > 30, n_B > 30$ ,

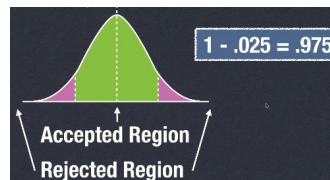
3. Compute **Z-Score:** 
$$z = \frac{(\bar{x}_A - \bar{x}_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

4. For a 95% **confidence level** (two-tailed test), **significance level**  $\alpha = 0.05$ , **critical probability** =  $1 - \frac{\alpha}{2} = 0.975$ , the critical value (from z-table) is  $\pm 1.96$ .

5. **Decision Rule:** if Z-score:  $Z > 1.96$  or  $Z < -1.96$ , reject Null Hypothesis  $H_0$ , thus there is significance difference

6. Using **p-value** method: Use **Z-Score**  $\pm$  (test statistic), and Look into z-table. Let say **Z-Score** is 2.31, then the row with 2.3 and column with 0.01, the area value is 0.98956. For the area in right side  $\frac{\alpha}{2}$  is  $1 - 0.98956 = 0.01044$ . Since, it's 2-sided test so the p-value 2 times 0.01044 is 0.02088.

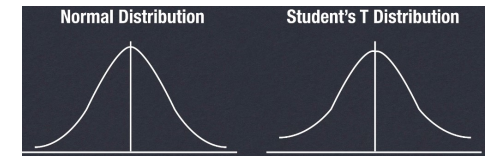
$p\text{-value} < \alpha \rightarrow 0.02088 < 0.05$ , thus **reject** Null Hypothesis  $H_0$



• **Z-test for Proportions z-Test:** compare proportions between two groups

## Student's t-Distribution

The *t-distribution* is a normally shaped distribution, except that it is a bit thicker and longer on the tails.



$$t = \bar{x} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n-1}}$$

**Example:** Let's say a manufacturer is promising brake pads will last for 65,000 km with a .95 confidence level. Our sample mean ( $\bar{x}$ ) is 62,456.2, sample standard deviation ( $S$ ) is 2418.4, Degrees of Freedom ( $n - 1$ ) is 29,  $\alpha$  is  $1 - 0.95 = 0.05$ , critical probability =  $1 - \frac{\alpha}{2} = 0.975$ . The *T*-value is found to be 2.045 by looking into the *T*-table from row no. 29<sup>th</sup> and column no.  $\alpha/2$  with 0.025. Now,

$$\text{Interval} = 62456.2 \pm 2.045 \cdot \frac{2418.4}{\sqrt{29}} = (61537.82, 63374.58)$$

## T-Test

T-test is used when the population **variance** (or standard deviation) is **unknown** or when the **sample size**  $n$  is **small**  $n < 30$  and It uses sample variance  $s^2$  in place of population variance  $\sigma^2$ .

• **One-Sample T-Test:** compare a **sample mean**  $\bar{x}$  with the known value or **population mean**  $\mu$ , Degree of freedom:  $n - 1$

$H_0 : \mu = \text{avg. value}, H_1 : \mu < \text{avg. value}$

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

• **Independent Two-sample t-test:** compares the means of **two independent samples**, Degree of freedom:  $n_A + n_B - 2$

$H_0 : \mu_A = \mu_B, H_1 : \mu_A \neq \mu_B$

$$t = \frac{(\bar{x}_A - \bar{x}_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

• **Paired or Dependent t-test:** Compares the means of two related groups (e.g., measurements before and after a treatment).

1.  $H_0 : \mu_{A1} - \mu_{A2} = 0, H_1 : \mu_{A1} - \mu_{A2} \neq 0$

2. **Collect Data:** Before ( $A1$ ) - [130, 132, 129, 138, 131], After ( $A2$ ) - [125, 128, 126, 127, 130]

3.  $d_i = \text{After}_i - \text{Before}_i$

4. Compute **t-statistic:** 
$$t = \frac{\bar{d}}{\frac{\sigma_d}{\sqrt{n}}}$$



Ch-Square Distribution

measure differences between categorical variables, using

χ² = ∑ (observed - expected)² / expected

- 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

ANOVA

ANOVA (Analysis of Variance) is a statistical method used to determine if there are any statistically significant differences between the means of three or more independent groups.

1. State the Hypotheses

H0 : All group means are equal  
H1 : At least one group mean is different

2. Collect and Prepare Data: **Example:** Consider a study evaluating the effectiveness of three different diets on weight loss.

Diet A : [5, 6, 7, 8, 5]  
Diet B : [6, 9, 8, 7, 6]  
Diet C : [4, 5, 5, 3, 6]

3. Choose the Significance Level (α): Common choices are 0.05 or 0.01.
4. Sum of Squares Between (SSB) = ∑(ni \* (Xi - X̄)²),  
Sum of Squares Within (SSW) = ∑(∑(Xij - Xi)²)  
k: No. of groups  
ni: No. of observations in group i  
Xi: Mean of group i  
X̄: Overall mean (mean of all groups combined).  
Xij: Individual observation in group i
5. Calculate the ANOVA: Calculate the group means and the overall mean.

F = (SSB / (k - 1)) / (SSW / (N - k))

Where:

k = number of groups  
N = total number of observations

6. Determine the Critical Value: Use an F-distribution table to find the critical F-value based on α, k - 1 (degrees of freedom between groups), and N - k (degrees of freedom within groups).
7. Compare F-Statistic to Critical Value: If Fcalculated > Fcritical, reject the null hypothesis H0.
8. If you reject H0, conclude that there is a significant difference between the group means. If you fail to reject H0, conclude that there is no significant difference.

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

A/B Testing

→ An A/B test is an experiment involving two groups to determine which of two treatments, products, or procedures performs better or yields superior results.

**Example:** To determine which of two cancer therapies is more effective at suppressing cancer growth., Therapy A or Therapy B.

1. **Hypothesis:** We hypothesize that Therapy B is more effective than Therapy A in reducing tumor size. Specifically:
- H0 : μA = μB (No difference in effectiveness)
  - H1 : μA < μB (Therapy B is more effective than Therapy A)
2. **Create Variations:**
- Group A (**Control**): Patients receive Therapy A (the standard treatment).
  - Group B (**Variation**): Patients receive Therapy B (the new or experimental treatment).
3. **Random Assignment:** Patients are randomly assigned to one of the two groups:
- **Group A:** Patients receive Therapy A.
  - **Group B:** Patients receive Therapy B.
4. **Data Collection:** Data is collected over a defined period. The key metric is the reduction in tumor size in each group.
5. **Statistical Analysis:** We conduct a statistical test, such as a t-test, to compare the average tumor size reduction between the two groups. If the mean reduction in Group B is significantly greater than in Group A, Therapy B is deemed more effective.
6. **Conclusion:** If the analysis shows that: 

p < α (e.g., α = 0.05), we reject the null hypothesis and conclude that Therapy B is more effective in suppressing cancer growth.

**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.

Dependent Samples vs. Independent Samples

Dependent Samples

- 1 Sample can be used to Determine other Results
- Cause & Effect
- Pairs of Results
- What is Probability a Die Roll is Odd
- What are Weekly Results of Weight Lifting at End of Week

Independent Samples

- Samples have No Relation to Others
- 10 Random Blood Samples at Lab A & 10 Random Samples at Lab B
- 1 Random Group gets Drug & Other Gets a Placebo

Hypothesis Erros

**Type I error:** The null hypothesis H0 is **rejected** when it is **true**.  
**Type II error:** The null hypothesis H0 is **not rejected** when it is **false**.  
→ **False negative (Type I error)** — incorrectly decide **no**  
→ **False positive (Type II error)** — incorrectly decide **yes**

Decision (Prediction)	Truth (Actual)		
		Null Hypothesis (H0)	Alternative Hypothesis (H1)
	Do Not Reject NULL	OK	Type II Error
	Reject NULL	Type I Error	OK

**Ex:** We assume the null hypothesis H0 is true.  
→ H0: Person is not guilty  
→ H1: Person is guilty

Decision (Prediction)	Truth (Actual)		
		Not Guilty (H0) 1	Guilty (H1) 0
	Not Guilty (H0) 1	OK (TP)	Type II Error (FP)
	Guilty (H1) 0	Type I Error (FN)	OK (TN)
		Sensitivity TP / (TP+FN)	Specificity TN / (TN+FP)
		Accuracy (TP+TN) / (TP+TN + FP +FN)	