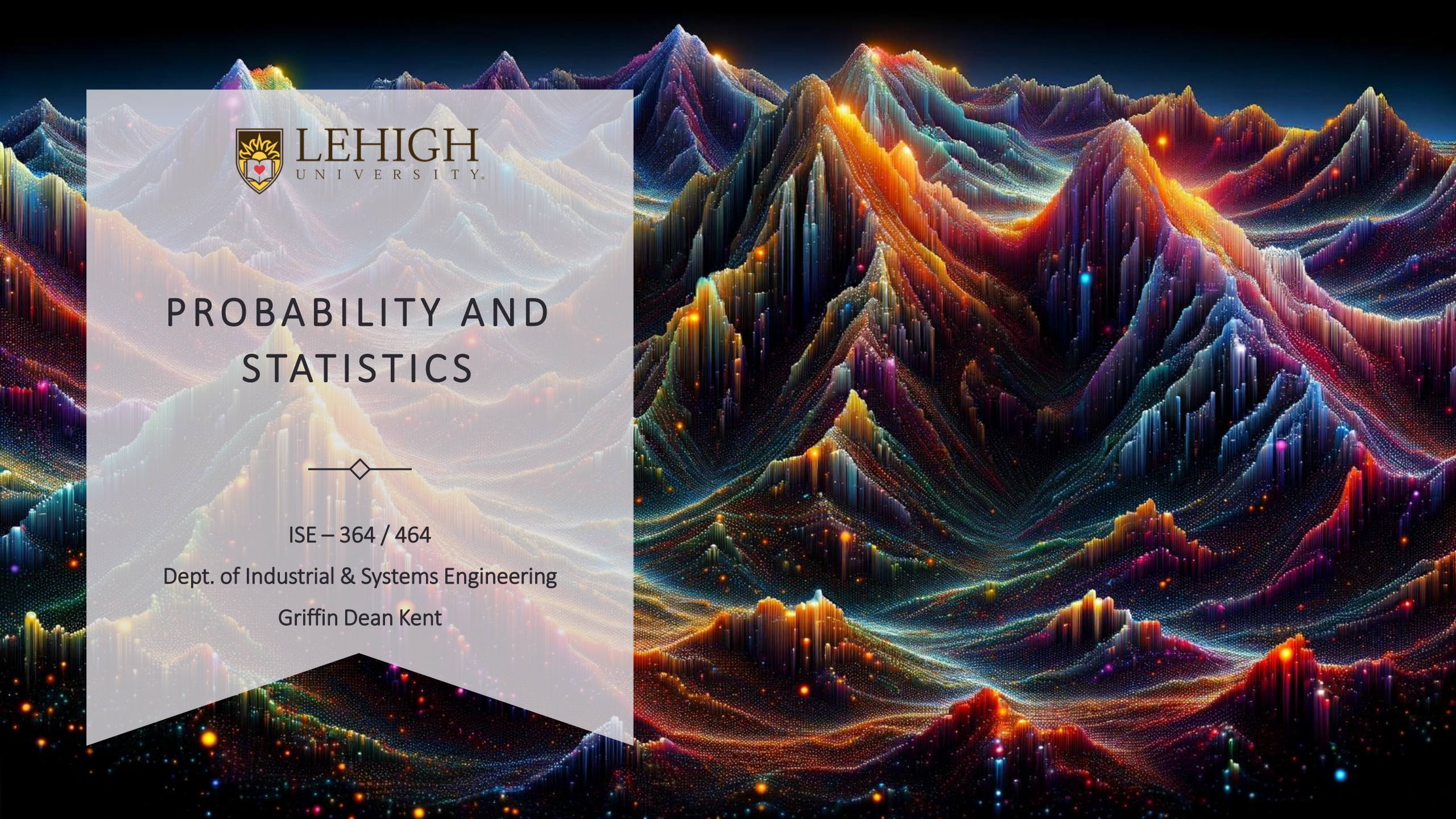


LEHIGH UNIVERSITY

PROBABILITY AND STATISTICS

ISE – 364 / 464

Dept. of Industrial & Systems Engineering

Griffin Dean Kent

What is Statistics?

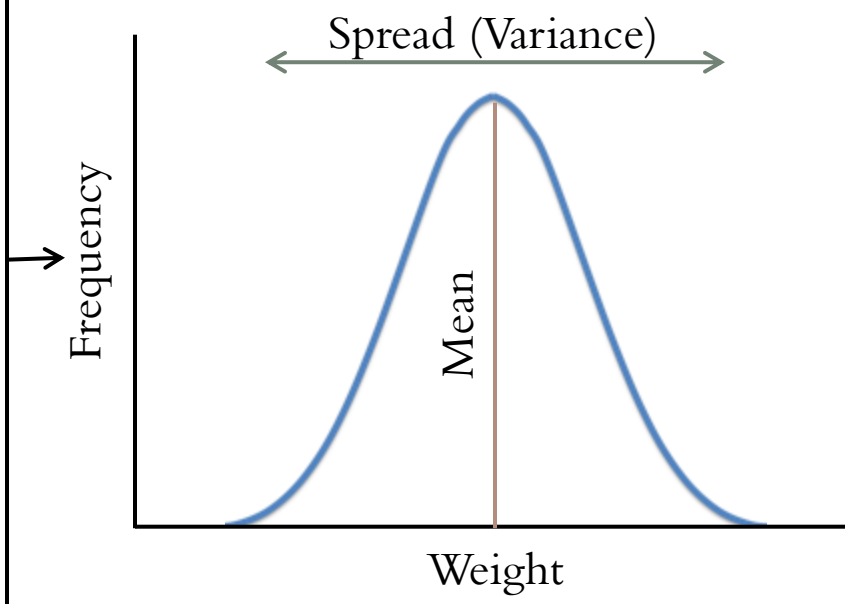
Statistics

Statistics is the discipline that is concerned with the collection, organization, analysis, interpretation, and presentation of data.

"There are lies, damn lies, and statistics." – Mark Twain

"In God we trust. All others bring data." – W. Edwards Demming

X = Average dog weight
10
31
28
17
65
39
84
8
.
.
.



- Statistics is a very useful tool for **summarizing** and talking about **large amounts of data**.
- Instead of talking about every single datapoint, we can talk about the **center of mass** (mean) and the **spread** (variance) of a given random variable. Further, we can talk about different types of distributions which we know have certain properties.
- Statistics is also useful for providing a series of different methods for objectively determining if there is a meaningful difference between two distributions (**Hypothesis Tests**), and as such is an essential tool for measuring the impact that certain variables has on a population.

What is Probability?

"Probability theory is nothing more than common sense reduced to calculation." - Laplace

Experiment

In probability theory, an **experiment** is simply the outcome of some random procedure (such as the outcome of tossing a coin or rolling a dice).

Sample Space

The **sample space** is the set of all possible outcomes of the experiment and is denoted by the set Ω.

- The individual elements of Ω are referred to as **elementary events**.
- Any subset of Ω (either one or many outcomes of the experiment) is referred to as an **event** $\mathcal{X}$.

Event Space

The set of all events in the sample space Ω is referred to as the **event space** and is denoted by the set $\mathcal{F}$.

- The set $\mathcal{F}$ is the collection of all subsets (all events) of Ω.
- The set $\mathcal{F}$ is a specific type of set known as a σ-field (or σ-algebra) which has certain properties.

Probability Measure

Classically, the **probability** of a specific event occurring is defined as the ratio of the total number of ways of that event occurring to the total number of outcomes.

Some Examples

Example 1

- Consider the experiment of rolling a six-sided dice.
- This will result in one of six outcomes with the corresponding sample space $\Omega = \{1,2,3,4,5,6\}$.
- Let's say that the outcome (or event) that we are interested in ascertaining the probability of is the event that an odd number is rolled. Thus, we can write this event as $\mathcal{X} = \{1,3,5\}$. (Notice, per the definition of an event, that $\mathcal{X} \subseteq \Omega$.
- Therefore, the probability of observing event $\mathcal{X}$ is given by $\mathbb{P}(\mathcal{X}) = \frac{\text{\# of ways event } \mathcal{X} \text{ can occur}}{\text{total \# outcomes}} = \frac{|\mathcal{X}|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}$.

Example 2 (Homework)

- What if we instead considered the experiment of rolling the dice twice?
- Further, how would we compute the probability of observing the event of rolling two 1s?
- (Hint: Alter the sample space of this experiment accordingly)

Fundamentals of Probability Theory

Although the notion of probability in some form or another has been around since the mid-16th century, it wasn't until the early 1900s that the concepts were given a rigorous mathematical formalization in the field of **probability theory**.

In probability theory, we define a probability as a way of assigning every possible event $\mathcal{X} \in \mathcal{F}$ a non-negative number $\mathbb{P}(\mathcal{X})$, called the probability of $\mathcal{X}$, where $\mathbb{P}: \mathcal{F} \to \mathbb{R}$ is a set function such that the following set of axioms are satisfied.

Kolmogorov Axioms of Probability

- (1) **Non-Negativity:** $\mathbb{P}(\mathcal{X}) \geq 0$ for any event $\mathcal{X} \in \mathcal{F}$.
- (2) **Normalization:** $\mathbb{P}(\Omega) = 1$.
- (3) **Countable Additivity:** For all countable collections of pairwise disjoint events $\{\mathcal{X}_k\}_{k=1}^{\infty}$ in $\mathcal{F}$, we have that

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} \mathcal{X}_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(\mathcal{X}_k).$$

- These three axioms completely define what is known as a **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ for a given sample space Ω, event space $\mathcal{F}$ (a σ-field), and a probability measure $\mathbb{P}$. A probability space is a specific kind of **measure space** (a basic object in the branch of mathematics known as **measure theory**). Axioms 1 and 3 together form a measure space, but it is Axiom 2 that makes $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.
- The quintessential probability result of $\mathbb{P}: \mathcal{F} \to [0,1]$ can be derived from these three axioms.

Basic Properties of Probability

- **Discrete Law of Probability:** Given the n events $\{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n\}$, then
$$\mathbb{P}(\{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n\}) = \mathbb{P}\left(\bigcup_{i=1}^{n} \mathcal{X}_i\right) = \sum_{i=1}^{n} \mathbb{P}(\mathcal{X}_i)\,.$$

- **Additive Law of Probability:** The probability of the union of two events $(\mathcal{A}, \mathcal{B}) \in \mathcal{F} \times \mathcal{F}$ is given by
$$\mathbb{P}(\mathcal{A} \cup \mathcal{B}) = \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}).$$

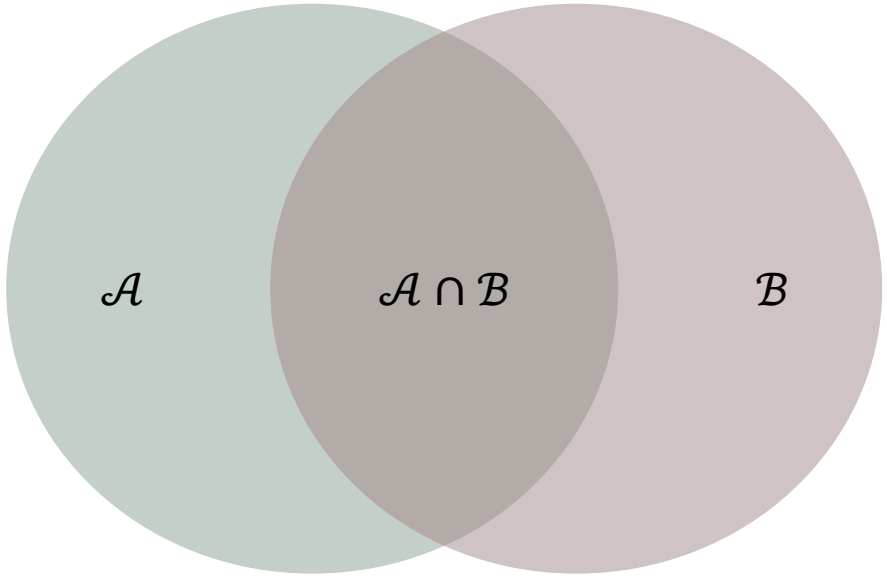
- **Conditional Law of Probability:** Given two events $(\mathcal{A}, \mathcal{B}) \in \mathcal{F} \times \mathcal{F}$, the probability of observing event $\mathcal{A}$ conditioned on having already observed event $\mathcal{B}$ is given by
$$\mathbb{P}(\mathcal{A}|\mathcal{B}) = \frac{\mathbb{P}(\mathcal{A} \cap \mathcal{B})}{\mathbb{P}(\mathcal{B})}.$$

- **Multiplication Law of Probability (Chain Rule) – Generalization of the Conditional Law:** Given n events $\{\mathcal{X}_i\}_{i=1}^{n}$, if all the conditioning events have nonzero probability, then
$$\mathbb{P}\left(\bigcap_{i=1}^{n} \mathcal{X}_i\right) = \prod_{i=1}^{n} \mathbb{P}\left(\mathcal{X}_i \middle| \cap_{j=1}^{n} \mathcal{X}_j\right).$$

Illustration of the Additive Law of Probability

In general, given two sets (or events) $\mathcal{A}$ and $\mathcal{B}$, we can think of their intersection $\mathcal{A} \cap \mathcal{B}$ as "observing $\mathcal{A}$ **and** $\mathcal{B}$". Similarly, we can think of their union $\mathcal{A} \cup \mathcal{B}$ as "observing $\mathcal{A}$ **or** $\mathcal{B}$".

Intersection of two sets



Additive Law of Probability

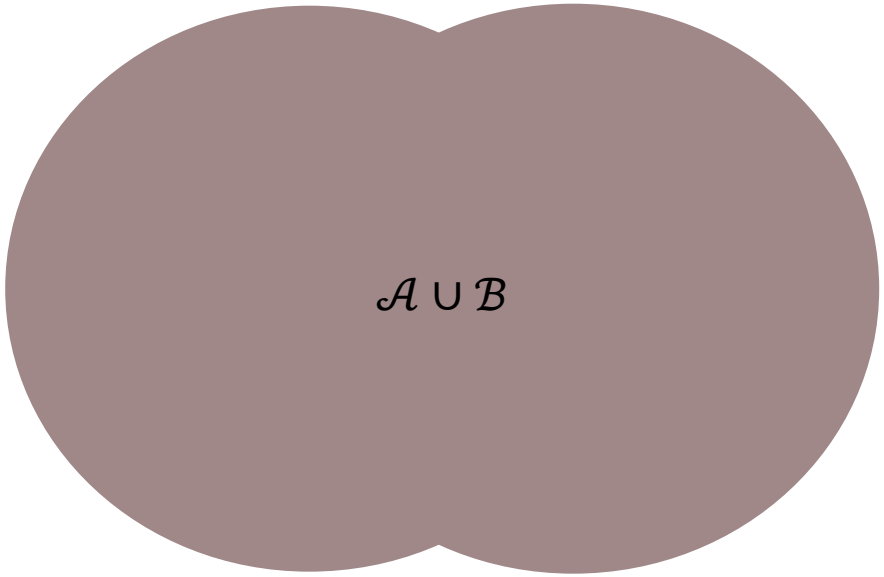
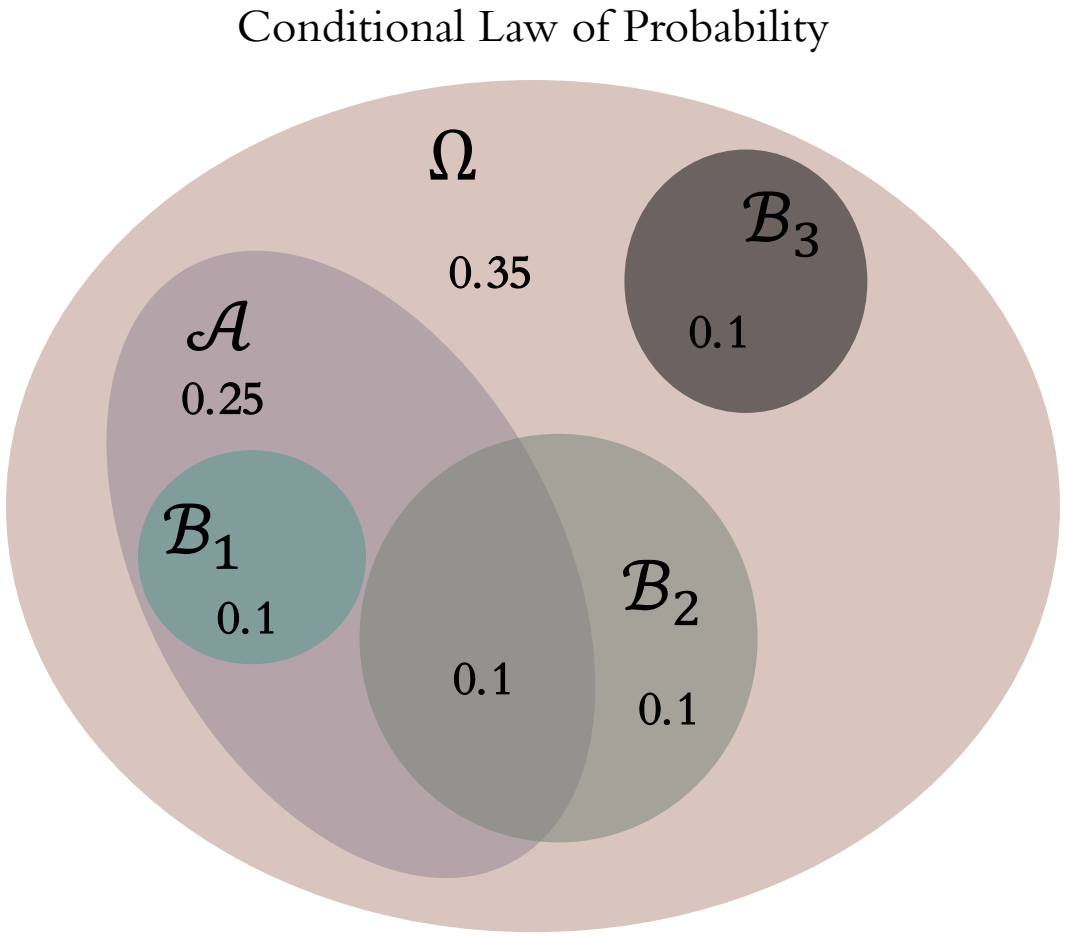


Illustration of the Conditional Law of Probability



Bayes' Theorem

Bayes' Theorem

Given two events $(\mathcal{A}, \mathcal{B}) \in \mathcal{F} \times \mathcal{F}$, the probability of observing the event $\mathcal{B}$ conditioned on having already observed the event $\mathcal{A}$ can be written as

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

Proof

By the definition of conditional probability, we have that

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

and

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

Further, solving the first equation for $\mathbb{P}(\mathcal{A} \cap \mathcal{B})$, we have $\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{B})\mathbb{P}(\mathcal{A}|\mathcal{B})$. Substituting this into the numerator of the second equation, we obtain the desired result

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

The Monty-Hall Problem

Given 5 doors: 4 have nothing behind them, 1 has full student-tuition.

The Monty-Hall Problem

Given 5 doors: 4 have nothing behind them, 1 has full student-tuition.

Probability of choosing correct door = 1/5 = 20%.

The Monty-Hall Problem

Given 5 doors: 4 have nothing behind them, 1 has full student-tuition.

Probability of choosing correct door = 1/5 = 20%.

Given that 2 random doors have been opened (with nothing behand them), do you choose a different door?

The Monty-Hall Problem

Original Probabilities

Given 5 doors: 4 have nothing behind them, 1 has full student-tuition.

20% 20% 20% 20% 20%

Probability of choosing correct door = 1/5 = 20%.

New Probabilities

Given that 2 random doors have been opened (with nothing behand them), do you choose a different door?

20% 40% 40%

Probability of choosing correct door (after switching) = 40%.

The Monty-Hall Problem

Original Probabilities

Given 5 doors: 4 have nothing behind them, 1 has full student-tuition.

20% 20% 20% 20% 20%

Probability of choosing correct door = 1/5 = 20%.

New Probabilities

Given that 2 random doors have been opened (with nothing behand them), do you choose a different door?

20% 40% 40%

Probability of choosing correct door (after switching) = 40%.

How?

Independence

If the occurrence of an event $\mathcal{B}$ provides no information about the likelihood of an event $\mathcal{A}$ (in the sense that knowledge of $\mathcal{B}$ does not affect the probability of $\mathcal{A}$, i.e., $\mathbb{P}(\mathcal{A}|\mathcal{B}) = \mathbb{P}(\mathcal{A})$), then we say that the event $\mathcal{A}$ is independent of event $\mathcal{B}$. Thus, the following definition naturally follows.

Independence

Given two independent events $(\mathcal{A}, \mathcal{B}) \in \mathcal{F} \times \mathcal{F}$ such that $\mathbb{P}(\mathcal{B}) \neq 0$, then we have

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{A})\mathbb{P}(\mathcal{B}).$$

Independence (General Form)

Given the independent events $\{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n\}$, we have

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \mathcal{X}_i\right) = \prod_{i=1}^{n} \mathbb{P}(\mathcal{X}_i)\,.$$

Random variables that are independent and have the same distribution are often said to be **i.i.d.** (or **independent and identically distributed**).

Random Variables

- All the examples we have discussed so far deal with discrete-event random experiments.
- However, as we will soon see, there are random experiments with continuous outcomes (which is incompatible with our current discrete formulation of probability).
- We would like to have a **consistent** way of talking about observing the outcome of a random experiment for both discrete and continuous events. We can begin to go about accomplishing this by encoding discrete non-numerical experimental outcomes to numerical values.

Target Space

- Typically, events of interest are those which are identifiable by numeric values; these are called numeric events (an example might be rolling a six-sided dice with a numeric sample space of $\Omega = \{1,2,3,4,5,6\}$).
- Even if an experimental event does not have numerical values (such as a coin flip with sample space $\Omega = \{\text{heads}, \text{tails}\}$), one can encode the sample space Ω to numerical values (such as $\{0,1\}$ for this example).
- Further, we can define this numerical set of values by the **target space** $\mathcal{T} \subseteq \mathbb{R}$.

- Finally, we can refer to an instantiation of a random experiment (which will have an outcome corresponding to an elementary event in the sample space) as a **random variable** X (since X will vary depending on the outcome of the random experiment), which serves as a mapping from the sample space to the target space.

Random Variable

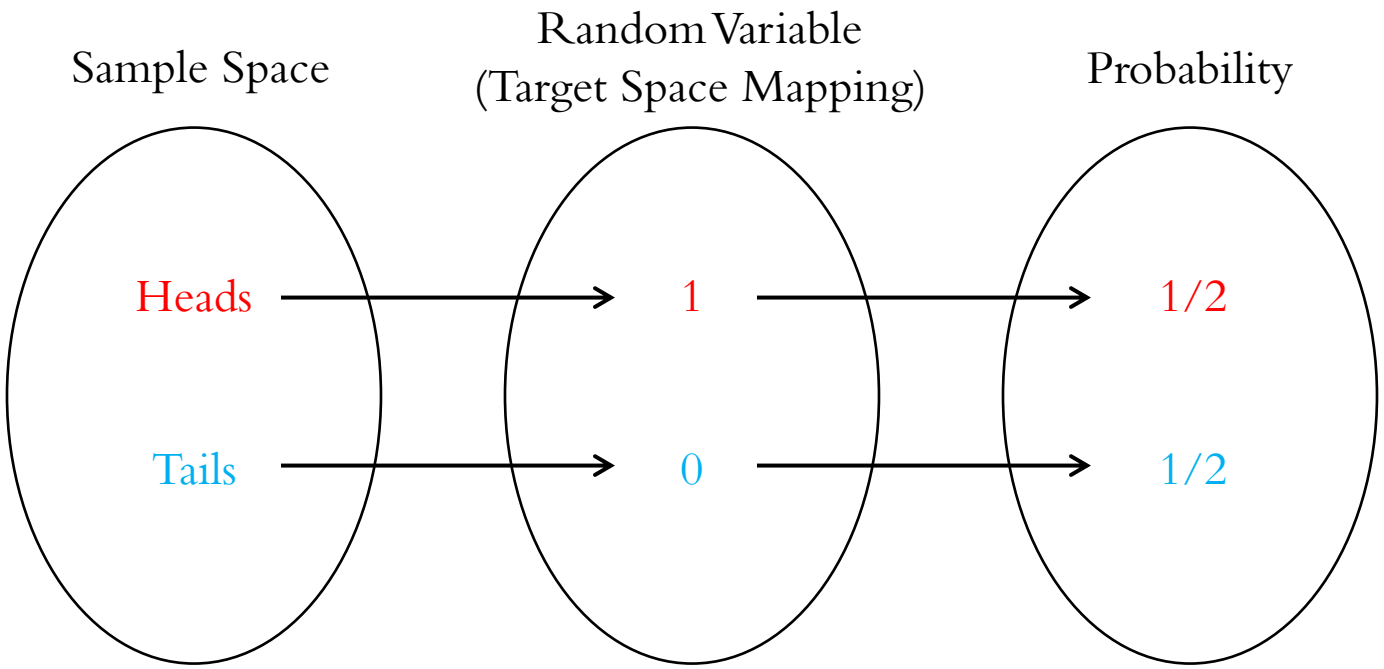
A random variable X is a function that takes an element (an elementary event $\mathcal{E} \in \Omega$) from the sample space Ω and returns an element of the target space $\mathcal{T} \subseteq \mathbb{R}$, i.e., $X: \Omega \to \mathcal{T}$.

Probability with Random Variables

- Notationally, we will use an *upper-case* letter, such as X, to denote a **random variable** and a *lower-case* letter, such as x, to denote a **particular outcome** that a random variable may assume in the target space.
- Further, we will use the expression $(X = x)$ to denote the *set of all elementary events* $\mathcal{E} \in \Omega$ *that are assigned the value of* $x \in \mathcal{T}$ *by the random variable* X, i.e., (in set notation)

$$(X = x) := \{\mathcal{E} \in \Omega | X(\mathcal{E}) = x \in \mathcal{T}\}.$$

- In this way, it is now meaningful to talk about the **probability** that the random variable X takes on some value x, which we denote by $\mathbb{P}(X = x)$.



Probabilities with Random Variables (An Example)

Example

- Consider the experiment of flipping a coin. Then we can define a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where the **sample space** is given by $\Omega = \{\text{heads}, \text{tails}\}$. Further, the **event space** is given by the σ-field $\mathcal{F} = \{\emptyset, \{\text{heads}\}, \{\text{tails}\}, \{\text{heads}, \text{tails}\}\}$.
- Then, the random variable X would map to the two outcomes: $X(\text{heads}) = 0$ and $X(\text{tails}) = 1$. Thus, in this case, the **target space** is given by $\mathcal{T} = \{0,1\}$. With a little abuse of notation, we can redefine the event space using this target encoding as $\mathcal{F} = \{\emptyset, \{0\}, \{1\}, \{0,1\}\}$.
- We can now ask questions such as what is the probability of obtaining heads? (Assuming probabilities are equal for both outcomes)
- $\mathbb{P}(X = 0) = \mathbb{P}(\{\text{heads}\}) = \frac{1}{2}$.
- What is the probability of observing head or tails? (Using the additive Law of Probability)
- $\mathbb{P}(X = 0 \cup X = 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = \mathbb{P}(\{\text{heads}\}) + \mathbb{P}(\{\text{tails}\}) = \frac{1}{2} + \frac{1}{2} = 1$.
- Now, consider a second random variable S that is defined over the same probability space (it is a second coin flip). Clearly, the two coin flips X and S are independent of each other. What is the probability of observing $X(\text{heads})$ and $S(\text{tails})$ (i.e., the first flip coming up heads and the second coming up tails)?
- $\mathbb{P}(X = 0 \cap S = 1) = \mathbb{P}(X = 0, S = 1) = \mathbb{P}(X = 0)\mathbb{P}(S = 1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

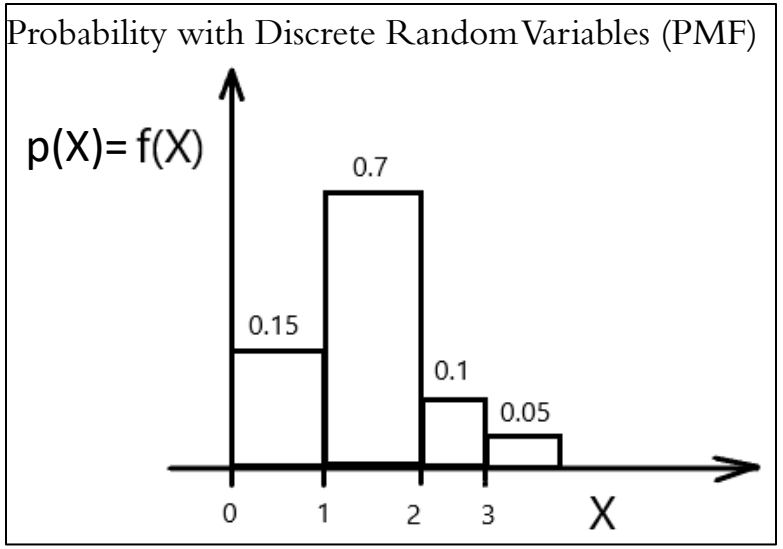
Distribution Functions

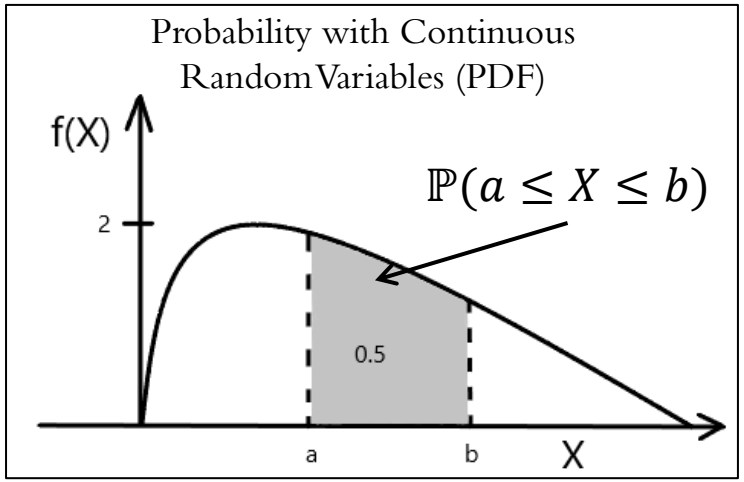
The probabilities that correspond to the values that a random variable X can take are characterized by a **probability distribution function**.

- In the discrete case, X can only take on discrete values (such as $0, 1, 2, \ldots$, etc.) in the coded target space and its distribution is characterized by a **probability mass function** (PMF), which we will denote by $p: \mathcal{T} \rightarrow [0,1]$ (where $\mathcal{T}$ is the target space).

- In the continuous case, X can take on real values (in $\mathbb{R}$) and its distribution is characterized by a **probability density function** (PDF), which we will denote by $f: \mathbb{R} \rightarrow [0,1]$.

*It bears mentioning that the PMF and PDF are often represented using the same notation of f to refer to both. However, this can cause confusion as **they are not the same thing**, so this is why we will try to delineate between the two by using p to represent the PMF and f to represent the PDF.





Probability Mass Functions

(for discrete random variables)

PMF

Given a discrete random variable X, the probability mass function is a function $p: \mathcal{T} \to [0,1]$ (where $\mathcal{T}$ is the target space) that returns the probability associated with X taking on the value x, and is simply given as

$$p(x) = \mathbb{P}(X = x),$$

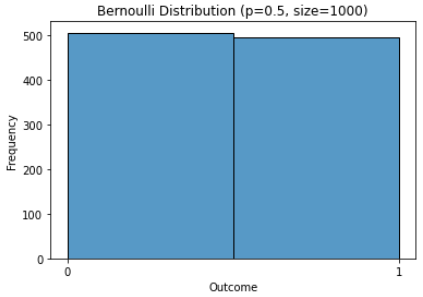
such that p satisfies the normalization property, i.e., $\sum_{x \in \mathcal{T}} p(x) = 1$.

Common PMFs of Discrete Random Variables

Bernoulli Distribution

Models the single outcome of a binary random variable, where the probability of "success" ($x = 1$) is $\alpha \in [0,1]$.

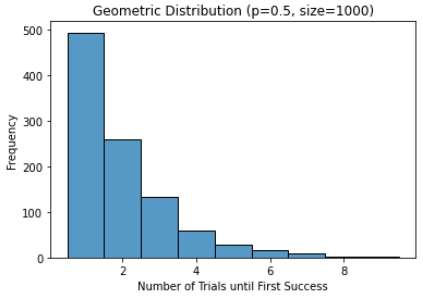
$$p(x) = \begin{cases} \alpha, & \text{if } x = 1 \\ 1 - \alpha, & \text{if } x = 0 \end{cases}.$$



Geometric Distribution

Models a Bernoulli experiment repeatedly until the first "success".

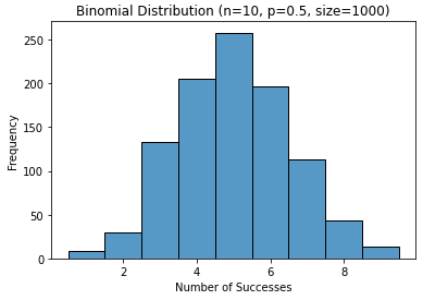
$$p(x) = \alpha(1 - \alpha)^{x-1}, \qquad \forall x.$$



Binomial Distribution

Models the number of "successes" of a Bernoulli experiment that is run n times.

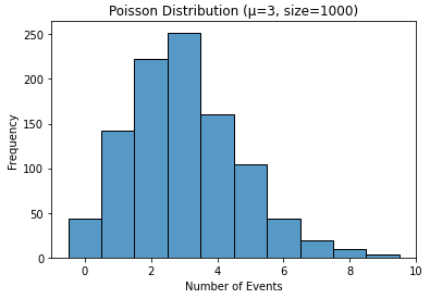
$$p(x) = \binom{n}{x} \alpha^x (1 - \alpha)^{n-x}, \qquad \forall x.$$



Poisson Distribution

Models the number of events that occur within a unit of time given the rate of occurrence λ.

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad \forall x.$$



Probability Density Functions

(for continuous random variables)

PDF

Given a continuous random variable Y, the probability density function is a function $f: \mathbb{R} \to [0,1]$ that returns the density associated with Y taking on the value y. As such, the probability of Y taking on a value $y \in [a, b]$ can be computed as

$$\mathbb{P}(a \leq Y \leq b) = \int_a^b f(y)dy\,,$$

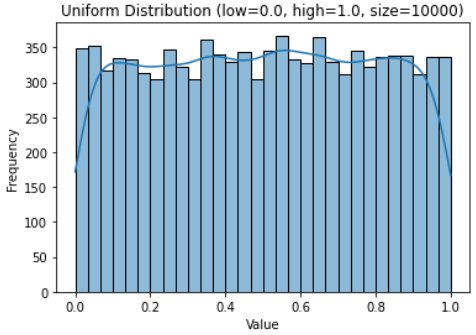
such that f satisfies the normalization property, i.e., $\int_{-\infty}^{\infty} f(y)dy = 1$.

Common PDFs of Continuous Random Variables

Uniform Distribution

Given range $[\underline{x}, \overline{x}]$:

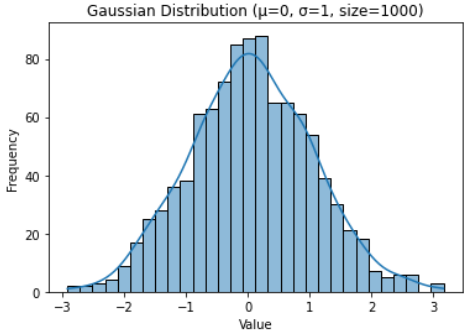
$$f(x) = \begin{cases} (\overline{x} - \underline{x})^{-1}, & x \in [\underline{x}, \overline{x}] \\ 0, & \text{otherwise} \end{cases}.$$



Gaussian (Normal) Distribution

Given mean μ and std σ:

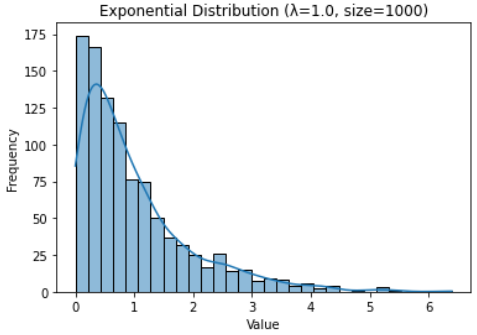
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



Exponential Distribution

Given parameter $\lambda > 0$:

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



An Important Distinction Between PMFs and PDFs

The PMF for a discrete random variable X does return the probability of X taking on a particular value x.

- For example, given that flipping a coin has outcomes $X = 0$ (for heads) and $X = 1$ (for tails), then $p(1) = \mathbb{P}(X = 1) = \frac{1}{2}$.

However, this is not true for probability density functions of continuous random variables, as PDFs do not return probability values! PDFs return a "density" value that tells you "how dense the probabilities are around a particular value".

The intuition for this can be understood when you think about what the probability is of observing a continuous random variable Y taking on any value y. The probability is 0! Why?

The reason why $\mathbb{P}(Y = y) = 0$ for a continuous Y, is because there are an infinite number of possible outcomes in a continuous space… As such, he probability of Y taking on any particular value y will be 0. This is why, for continuous random variables, probabilities are computed for Y taking on a value within a defined range $[a, b]$.

Expected Value

Often, it is convenient to encapsulate and summarize a relevant property (or properties) of a random variable in a single (or few numbers). In particular, the most common descriptive property of a random variable is its **expected value** (or mean), which is a generalization of the weighted average.

Expected Value (Informal Definition)

The expected value (also known as the mean) of a random variable, is a generalization of the weighted average. This can be understood as the value one would expect to get by taking the average of numerous "realizations" of the random variable. It can be established that the empirical average (or sample average) is an unbiased approximator of the expected value.

Expected Value (Formal Definition)

For a discrete or continuous random variable X with probability function f and some general function $g: \mathbb{R} \to \mathbb{R}$, the expected value of $g(X)$ is a function $\mathbb{E}: \mathbb{R} \to \mathbb{R}$ and is defined as

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x)f(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x)f(x)dx, & \text{if } X \text{ is continuous.} \end{cases}$$

Intuitively, the expected value is simply the weighted sum of the values that the random variable X can take and the corresponding probabilities that it will take those values.

Variance & Standard Deviation

Though the expected value of a random variable provides some measure of the "center of mass" of the distribution of the variable, another important quantity is the variance, which provides some measure of the "spread" of the distribution.

Variance & Standard Deviation

The variance of a random variable X is defined as

$$Var(X) = \mathbb{E}\left[\left(X - \mathbb{E}(X)\right)^2\right] = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

The standard deviation of X is the square root of the variance and is denoted by $\sigma = \sqrt{Var(X)}$, i.e., $\sigma^2 = Var(X)$.

The variance and standard deviation both provide measures of the spread of the distribution. What is the difference between the two and what is the advantage of the standard deviation?

- The standard deviation is reported in the same units as the data (the random variable X).

Multivariate Distributions

- The intersection of two or more random variables is frequently of interest (e.g., intersections while sampling data).
- A specific set of outcomes may be expressed in terms of the intersection of the n events $(X_1 = x_1), (X_2 = x_2), \ldots, (X_n = x_n)$, i.e., $\bigcap_{i=1}^{n}(X_i = x_i)$, which we denote by $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, or simply as $(x_1, x_2, \ldots, x_n)$.
- The probability of this intersection is essential for making inferences about populations from which a sample is drawn.

Joint Probability Mass Functions (Discrete)

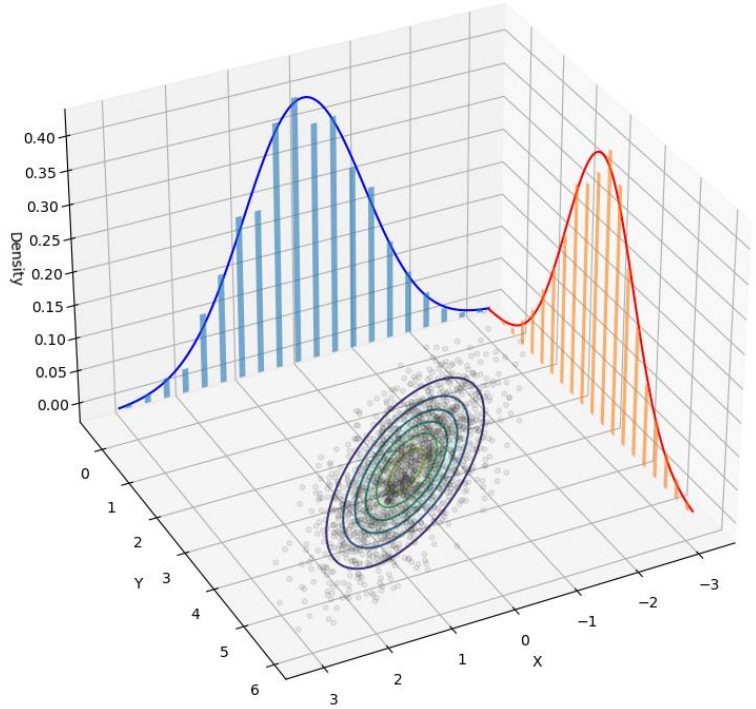
Given that X and Y are discrete random variables, the **joint PMF** of X and Y is a function $p_{X,Y}$ (which satisfies the normalization property) defined as

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

Joint Probability Density Functions (Continuous)

Given that X and Y are continuous random variables, the **joint PDF** of X and Y is a function $f_{X,Y}$ (which satisfies the normalization property) with probabilities that can be computed by

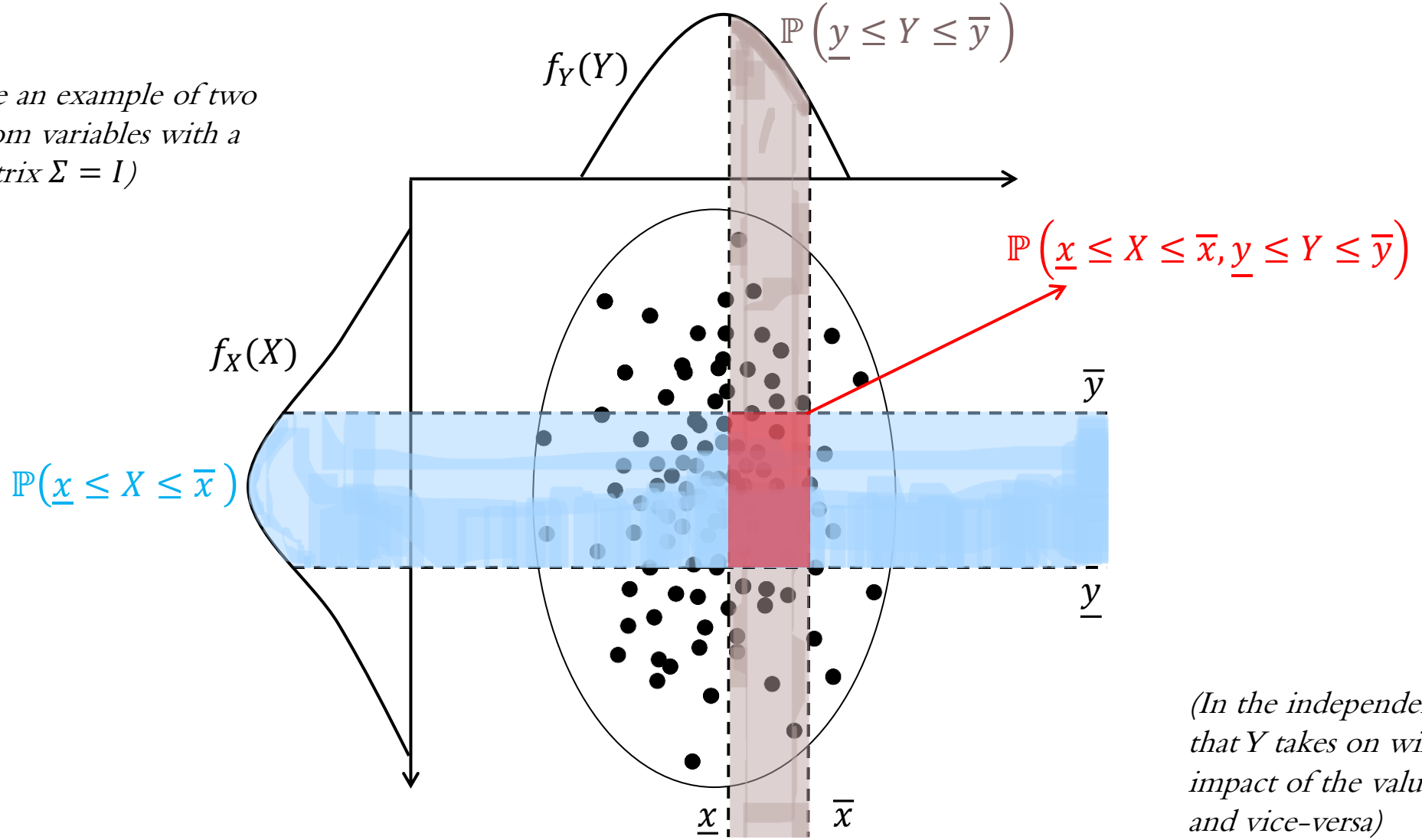
$$\mathbb{P}\left(X \in [\underline{x}, \overline{x}], Y \in \left[\underline{y}, \overline{y}\right]\right) = \mathbb{P}\left(\underline{x} \leq X \leq \overline{x}, \underline{y} \leq Y \leq \overline{y}\right) = \int_{\underline{x}}^{\overline{x}} \int_{\underline{y}}^{\overline{y}} f_{X,Y}(x, y)\, dy\, dx\,.$$



How does one recover the **marginal probability functions** p_X and p_Y (or f_X and f_Y) from the joint functions? To obtain p_X, simply sum (or integrate for f_X) over all y. To obtain p_Y, simply sum (or integrate for f_Y) over all x.

Illustration of Multivariate Probability Density Functions (Independent Random Variables)

(This would be an example of two Gaussian random variables with a covariance matrix $\Sigma = I$)



(In the independent case, the value that Y takes on will not have an impact of the value that X takes on and vice-versa)

Covariance

Given two random variables that are jointly distributed, one may be interested in how they **change together** and the amount of **linear** relationship between the two variables.

Covariance

Given two random variables X and Y that are jointly distributed, the **covariance** between the two variables is a measure of the **linear relationship** between them (the name is given to indicate what the "-variance" between the two "co-"variables is). As such, it is defined as

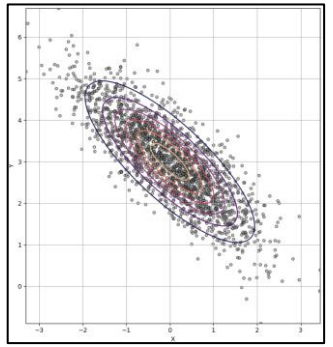
$$Cov(X,Y) = \mathbb{E}\big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Do you notice anything about this equation and its connection to the independence of the variables X and Y?

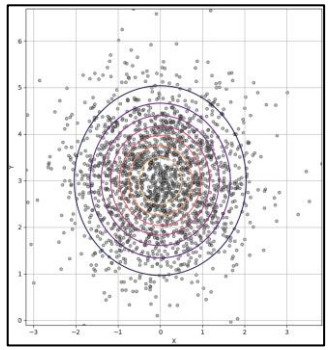
Recall that if X and Y are independent then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. From the definition of covariance, it is clear that $\boldsymbol{Cov(X,Y) = 0}$ **when** $\boldsymbol{X}$**and** $\boldsymbol{Y}$ **are independent**.

When dealing with several jointly distributed random variables $\{X_1, X_2, \dots, X_n\}$, we can define all the covariances between each ith and jth pairs of random variables (for i and j in $\{1,2, \dots, n\}$) as the **covariance matrix** $\boldsymbol{\Sigma}$, which will be a symmetric matrix in $\mathbb{R}^{n\times n}$ with the ith diagonal element denoting the variance of X_i and all the off-diagonal elements defining the covariances between each of the variables.

$\mathrm{Cov}(X,Y) < 0$



$\mathrm{Cov}(X,Y) = 0$



$\mathrm{Cov}(X,Y) > 0$

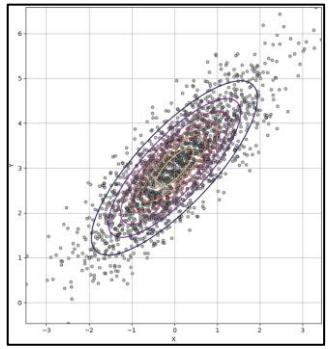
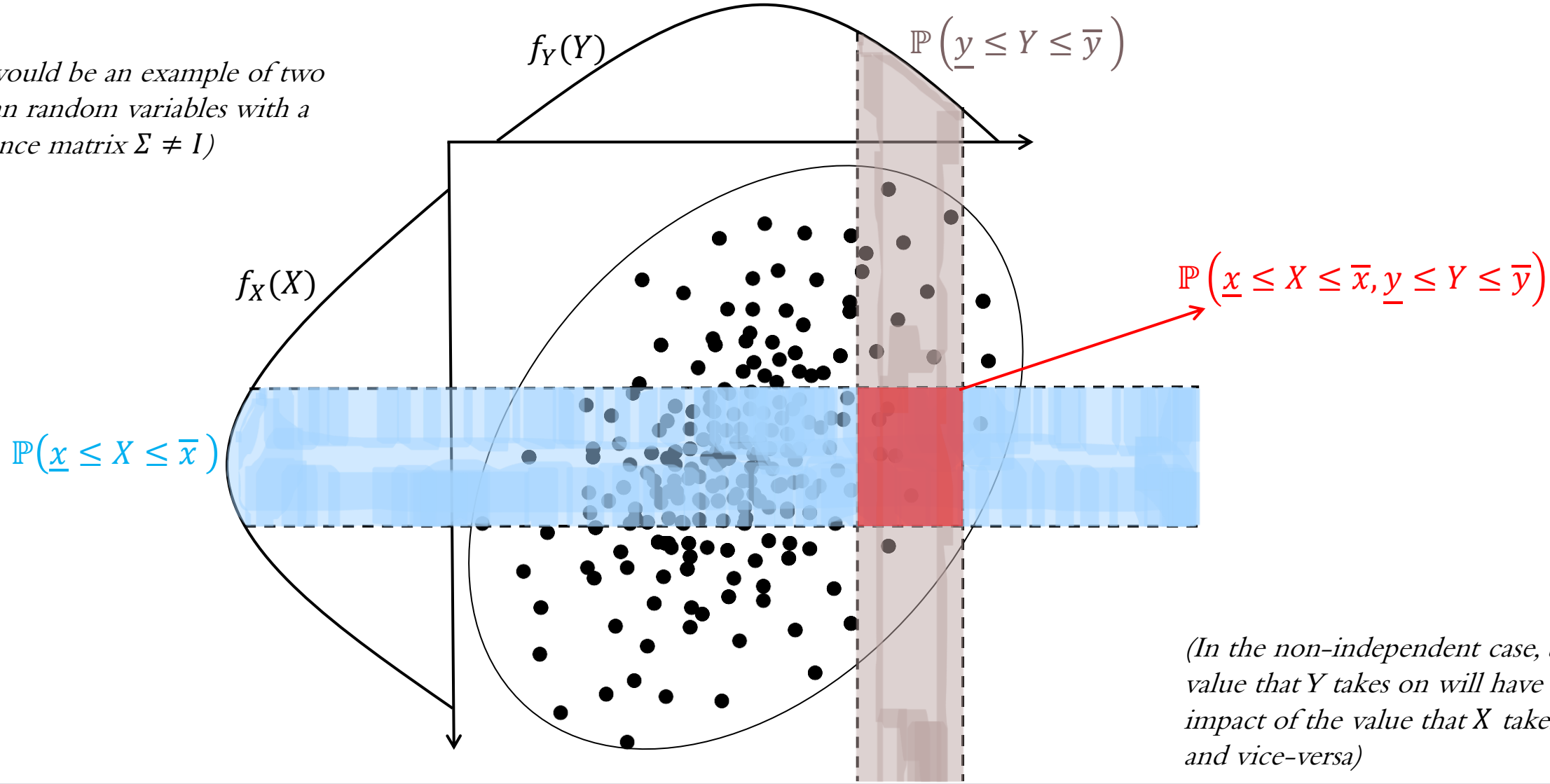


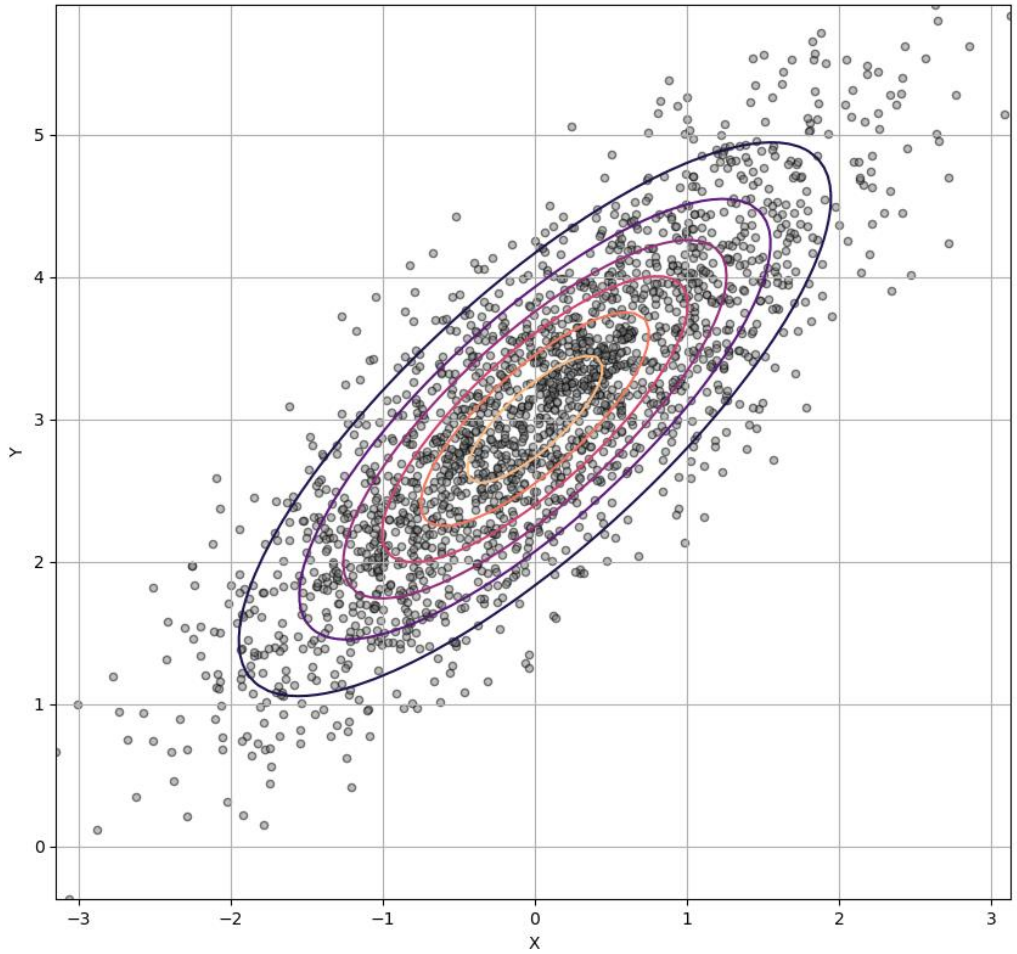
Illustration of Multivariate Probability Density Functions (Non-Independent Random Variables)

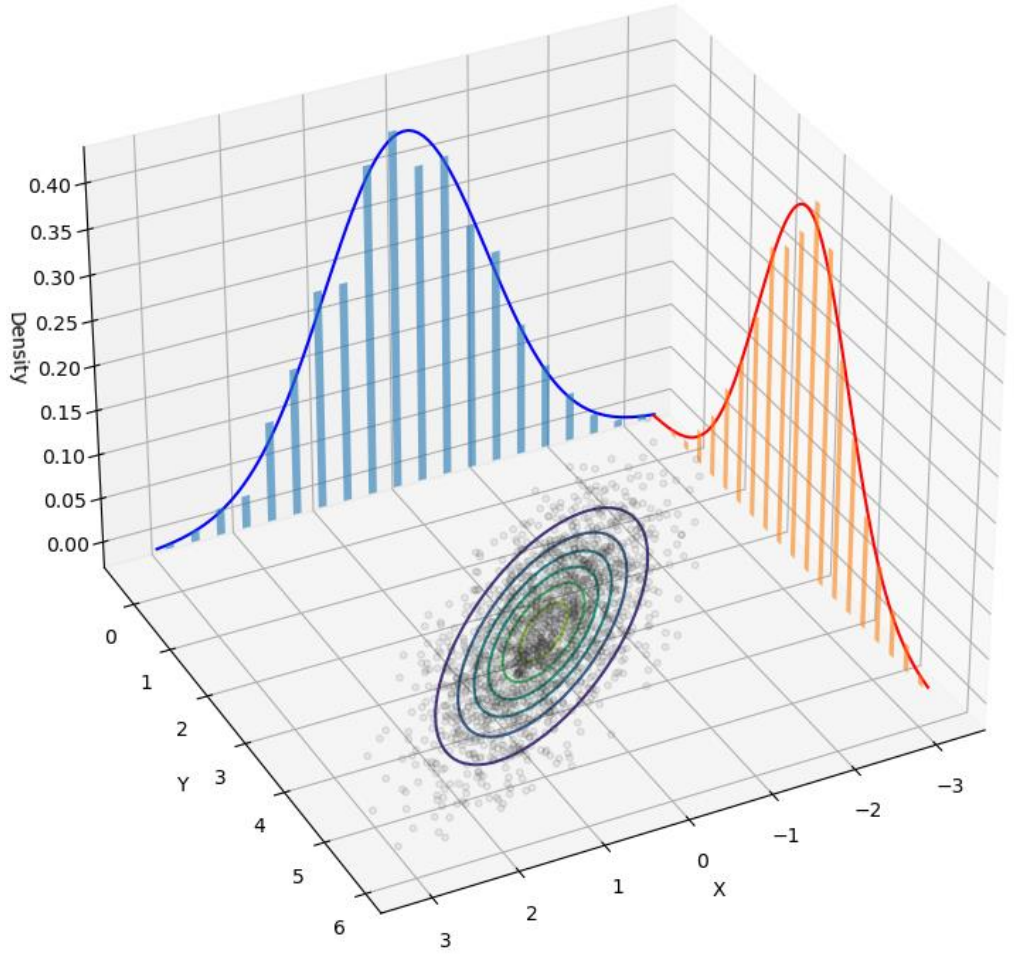
(This would be an example of two Gaussian random variables with a covariance matrix $\Sigma \neq I$)



(In the non-independent case, the value that Y takes on will have an impact of the value that X takes on and vice-versa)

Multivariate Distribution Visualizations in Python





Conditional Distributions

Now that we have seen probability density functions for the intersection of random variables (the joint distribution), it is natural to introduce probability density functions *conditioned on some other event occurring*.

Conditional Probability Mass Functions (Discrete)

Given that X and Y are discrete random variables, the **conditional PMF** of X given Y is a function $p_{X|Y}$ (which satisfies the normalization property) defined as

$$p_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Conditional Probability Density Functions (Continuous)

Given that X and Y are continuous random variables, the **conditional PDF** of X given Y is a function $f_{X|Y}$ (which satisfies the normalization property) with probabilities that can be computed by

$$\mathbb{P}\big(X \in [\underline{x}, \overline{x}]|Y = y\big) = \mathbb{P}\big(\underline{x} \leq X \leq \overline{x}|Y = y\big) = \int_{\underline{x}}^{\overline{x}} f_{X|Y}(x|y)dx\,,$$

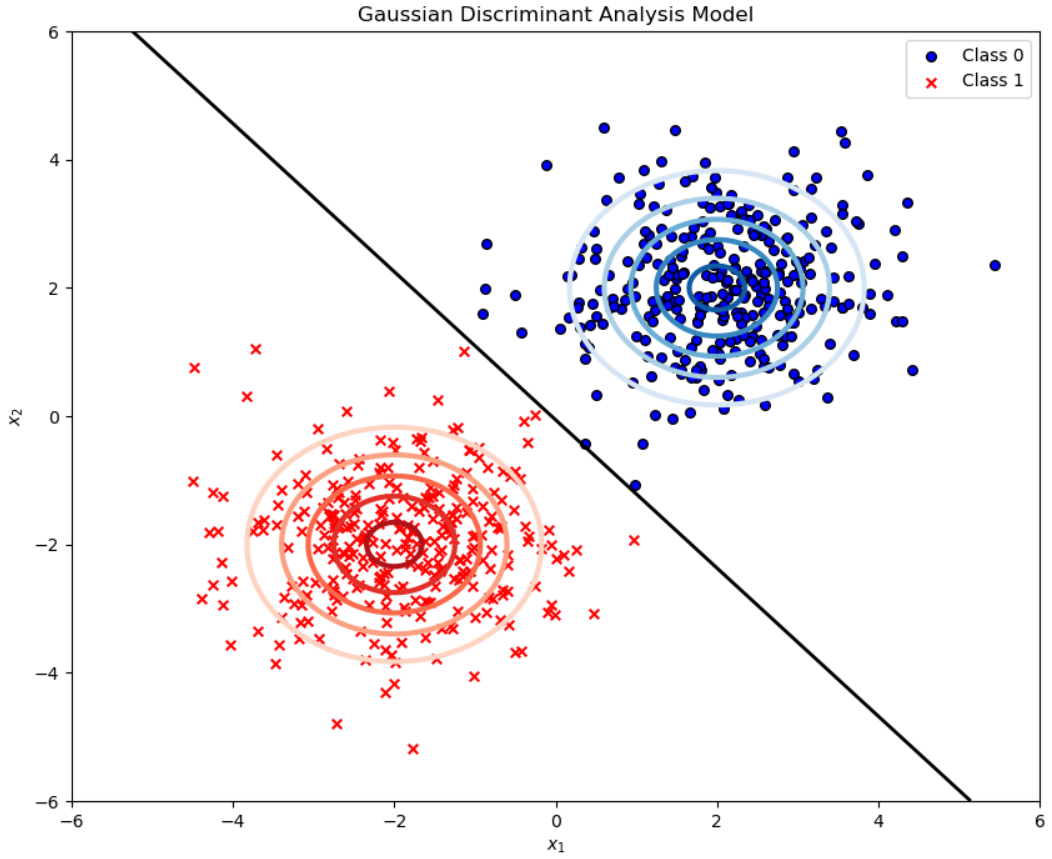
where

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

What does having the marginal functions in the denominator do? It ensures the functions are normalized.

Illustration of Conditional Probability Density Functions

A classic example of probabilities after conditioning on some other random variable. This figure illustrates a Gaussian Discriminative Analysis (GDA) Model when the conditional distribution is a multivariate Gaussian distribution conditioned on a binary variable $Y \in \{0,1\}$, i.e., $f_{X|Y}(X|Y = y) \sim \mathcal{N}(\mu_y, \Sigma)$.



Correlation

"Correlation does not imply causation."

Correlation

In statistics, correlation is any **statistical relationship** between two random variables. As such, there are a variety of ways of mathematically defining correlation to capture some of these relationships. Perhaps the two most popular correlations are the **Pearson** and **Spearman** correlation coefficients; the former providing a way of measuring linear relationships and the latter providing a way of measuring certain types of nonlinear relationships (by assessing how well the two variables can be described by a monotonic function).

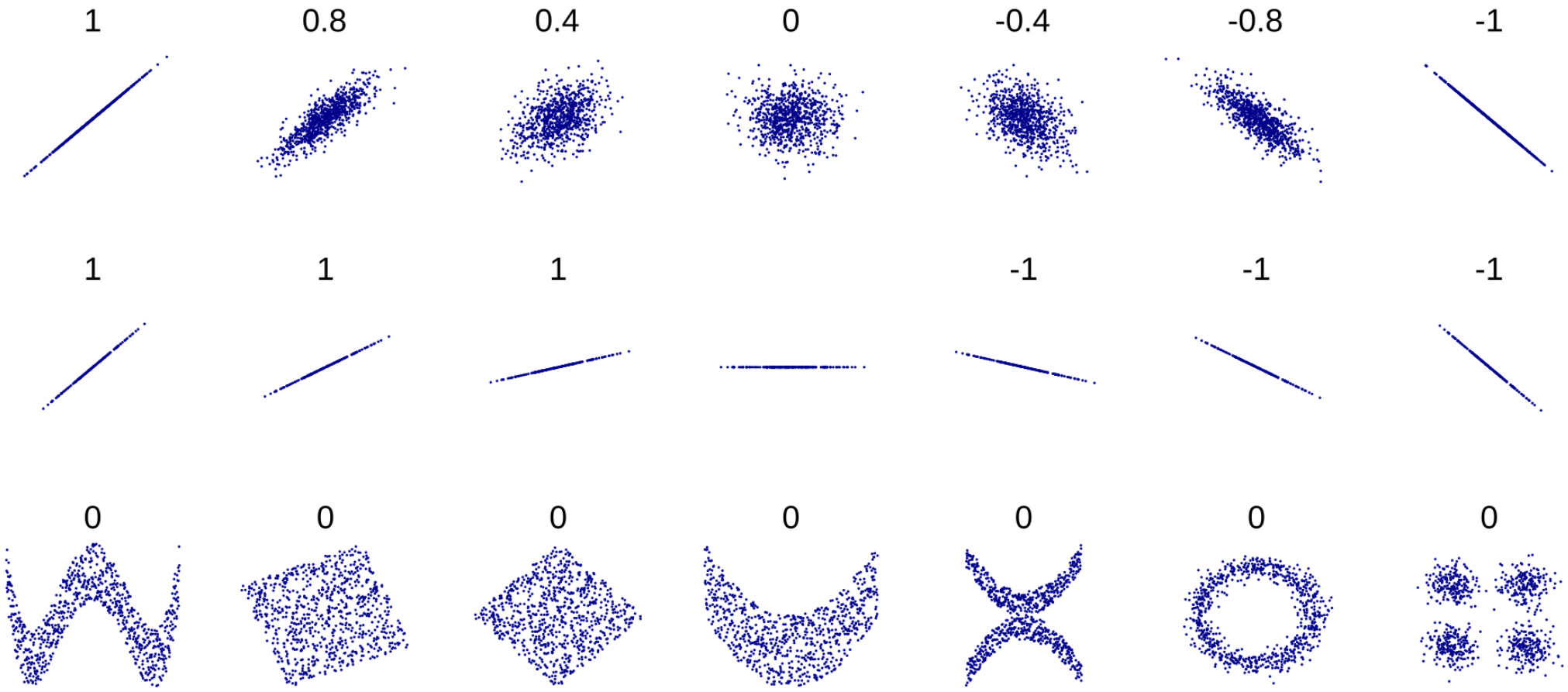
Pearson's Correlation Coefficient

Given two random variables X and Y, the Pearson Correlation Coefficient is a real value $\rho_{X,Y} \in [-1,1]$ which is computed by dividing the covariance of the two random variables by the product of their respective standard deviations, i.e.,

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

As such, this coefficient provides a normalized form of the covariance metric, which only provided a non-standardized measure of how much the two variables varied together.

Illustration of Different Pearson Correlations



The Likelihood Function

One of the most important concepts in statistics (and one of the most-used tools in machine learning) is the idea of the **likelihood function**. Given some number n of i.i.d. random variables $\{X_1, X_2, \dots, X_n\}$, the likelihood function is a function of the parameters θ (could be a scalar or a vector) of a statistical model, given some observed data $\{x_1, x_2, \dots, x_n\}$ (i.e., we have the observations $(X_1 = x_1), (X_2 = x_2), \dots, (X_n = x_n)$).

Likelihood Function

For **i.i.d.** random variables $\{X_i\}_{i=1}^n$, taken from a distribution that is parameterized by θ with corresponding probability function f (which denotes either a PMF or PDF and could also be joint, marginal, or conditional), the likelihood function $L: \Theta \to [0,1]$ (where $\Theta \subseteq \mathbb{R}^m$ is simply the parameter space, i.e., $\theta \in \Theta$) of observing the data $\{x_i\}_{i=1}^n$ is given by

$$L(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)\,.$$

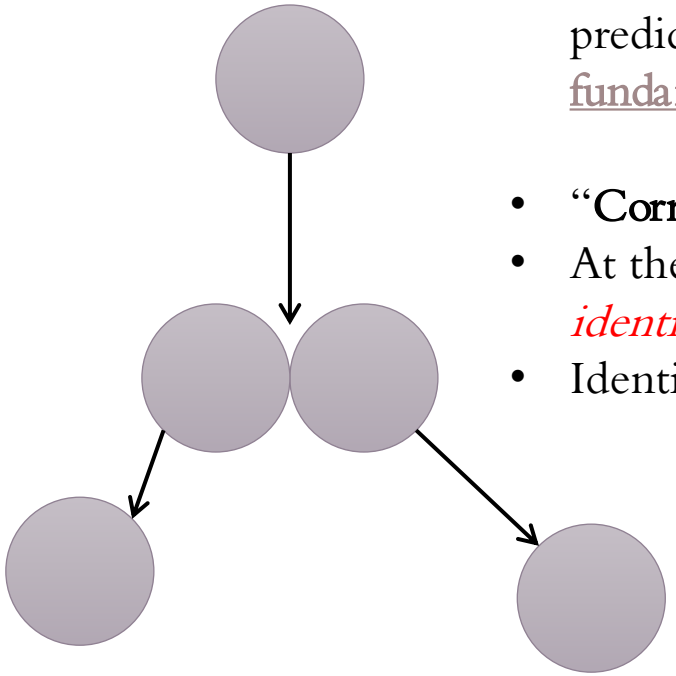
Some Notes

- The likelihood function is a measure of how "likely" it is that a certain statistical model (defined by the parameters θ) describes observed data.
- What does the likelihood function compute in the case of discrete random variables? Probabilities.
- What does the likelihood function compute in the case of continuous random variables? Densities.
- Why in the likelihood function is the joint function $f(\{x_i\}_{i=1}^n; \theta) = \prod_{i=1}^n f(x_i; \theta)$? Independence.


PHILOSOPHY OF STATISTICS

The law of Causality

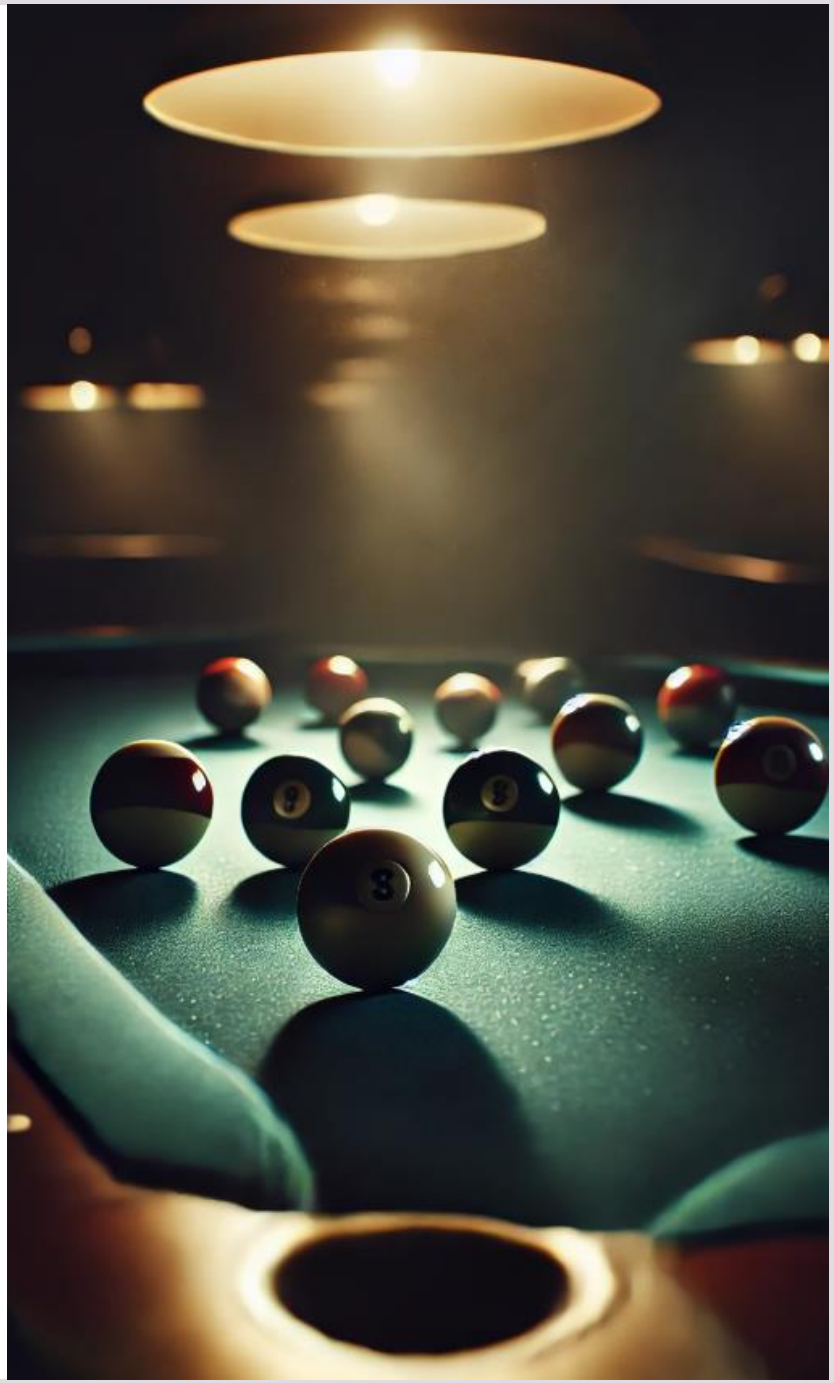
- "**For there to be an effect, there must be a cause…**"
- In Statistics (and as data scientists), the goal is to identify underlying patterns in data. By doing so, one can make better predictions.
- "**Ideally**", this is accomplished by correctly identifying the variables (the underlying causes) that will impact a target. If one can correctly identify the causal relationships between variables, then one can identify what the outcome will be under different instances.



- Thus, machine learning (and any form of making predictions and identifying patterns) fundamentally assumes the **principle of Causality**.

However!

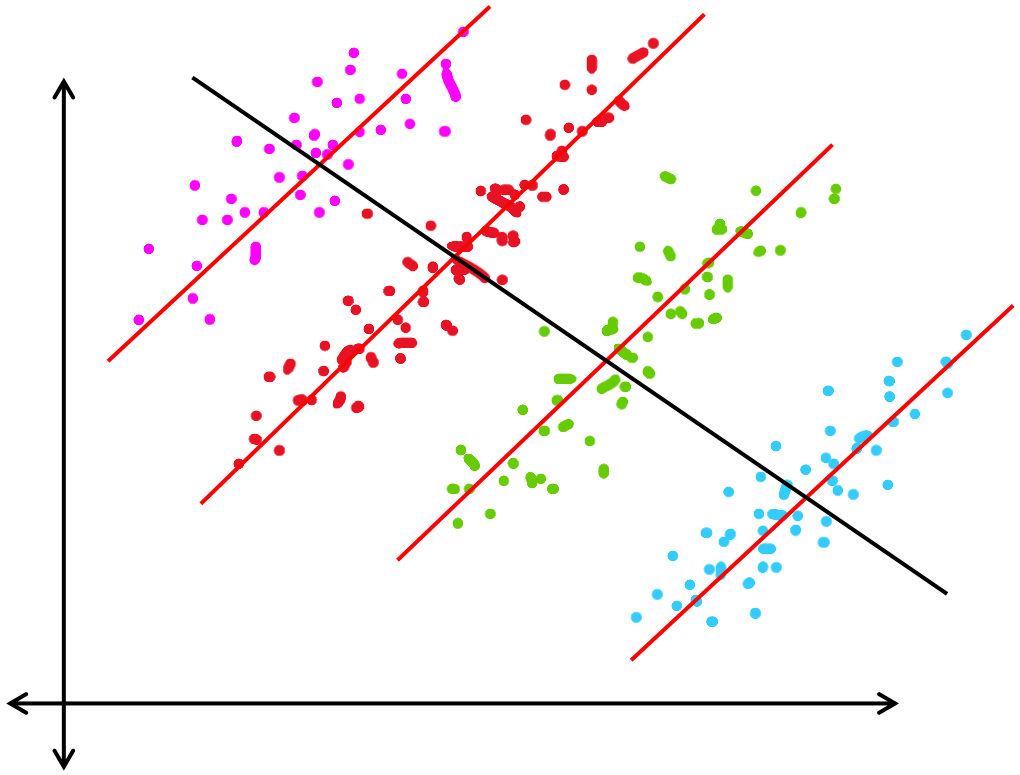
- "**Correlation does not imply causation.**"
- At the end of the day *ML models are built on identifying correlations*… **Not necessarily Causes**.
- Identifying Causes is **difficult**.
  - The **only thing one can say** about a model is that "there is/isn't a strong correlation between features and a target". It cannot be said that one has identified the causes of a target.



Simpson's Paradox

Simpson's Paradox
When an apparent relationship **reverses its sign** when a confounding variable is considered.

Simpson's Paradox is a phenomenon that can occur in any situation but will only be detrimental in circumstances when an analyst / statistician incorrectly identifies a causal relationship. This happens most often when trying to describe complex patterns with too simple of a model.



Classic Example of Simpson's Paradox

In 1973, there seemed to be a very noticeable gender bias in the graduate admission rates at University of California, Berkeley. Looking just at admission rates, out of 8,442 male applicants, 44% were admitted, and out of 4,321 female applicants, 35% were admitted. This seems to be a clear difference in admittance rates.

However, when accounting for the departments that the applicants were applying to (6 departments) this trend reverses in 4 of the 6 departments to showing a statistically significant bias toward female applicants.

Simpson's Paradox is a quintessential reminder for the need to withhold premature judgement about what causes what. It can be a difficult task.

Attribution & License

These lecture slides are part of the "*Introduction to Machine Learning*" course materials created by **Griffin Dean Kent**.

For the latest version, updates, and additional resources, visit the GitHub repository: https://github.com/GdKent/Introduction-to-Machine-Learning.