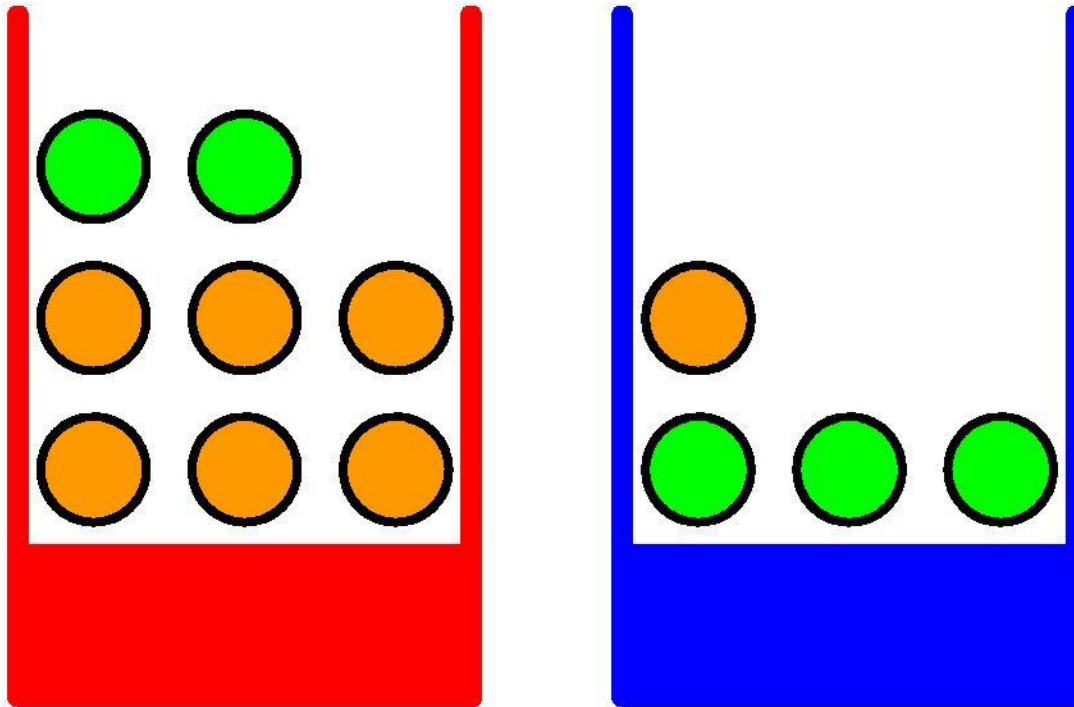


Outline

- Properties and Axioms of Probability
- Conditional Probability, Independence, Bayes' Rule
- Random Variables, Distributions
- Joint Probability, Multivariate
- Likelihood, Maximum Likelihood
- Case Study: Gaussian

Probability Theory

Apples and Oranges



Probability Terminology

Name	What it is	Common Symbols	What it means
Sample Space	Set	Ω, S	"Possible outcomes."
Event Space	Collection of subsets	\mathcal{F}, E	"The things that have probabilities.."
Probability Measure	Measure	P, π	Assigns probabilities to events.
Probability Space	A triple	(Ω, \mathcal{F}, P)	

Remarks: may consider the event space to be the power set of the sample space (for a discrete sample space - more later). e.g., rolling a fair die:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$\mathcal{F} = 2^\Omega = \{\{1\}, \{2\} \dots \{1, 2\} \dots \{1, 2, 3\} \dots \{1, 2, 3, 4, 5, 6\}, \{\}\}$$

$$P(\{1\}) = P(\{2\}) = \dots = \frac{1}{6} \text{ (i.e., a fair die)}$$

$$P(\{1, 3, 5\}) = \frac{1}{2} \text{ (i.e., half chance of odd result)}$$

$$P(\{1, 2, 3, 4, 5, 6\}) = 1 \text{ (i.e., result is "almost surely" one of the faces).}$$

Axioms of Probability

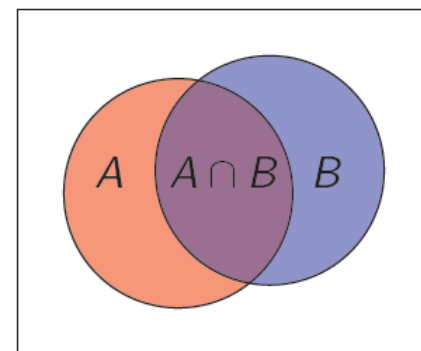
- $P(A) \geq 0 \ \forall A \in F$
- $P(\Omega) = 1$
- If A_1, A_2, \dots are disjoint events, then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

Additional Properties of Probability

- If $A \subseteq B \implies P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$.
- (Union Bound) $P(A \cup B) \leq P(A) + P(B)$.
- $P(\Omega \setminus A) = 1 - P(A)$.
- (Law of Total Probability) If A_1, \dots, A_k are a set of disjoint events such that $\bigcup_{i=1}^k A_i = \Omega$, then

$$\sum_{i=1}^k P(A_i) = 1.$$

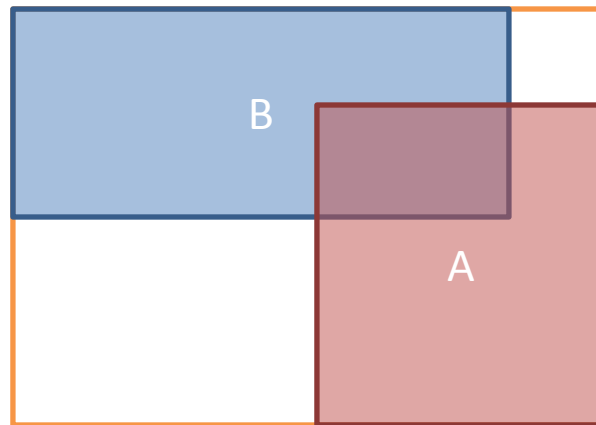


Conditional Probability

For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

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

Interpretation: the outcome is definitely in B , so treat B as the entire sample space and find the probability that the outcome is also in A .

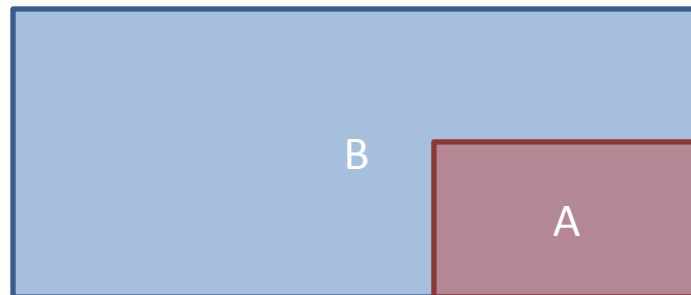


Conditional Probability

For events $A, B \in \mathcal{F}$ with $P(B) > 0$, we may write the **conditional probability of A given B**:

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

Interpretation: the outcome is definitely in B , so treat B as the entire sample space and find the probability that the outcome is also in A .



Conditional Probability Example

Suppose we throw a fair die:

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \mathcal{F} = 2^\Omega, P(\{i\}) = \frac{1}{6}, i = 1 \dots 6$$

$$A = \{1, 2, 3, 4\} \text{ i.e., "result is less than 5,"}$$

$$B = \{1, 3, 5\} \text{ i.e., "result is odd."}$$

What is the probability of A given B?

Probability of B given A?

Conditional Probability Example

Suppose we throw a fair die:

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \mathcal{F} = 2^\Omega, P(\{i\}) = \frac{1}{6}, i = 1 \dots 6$$

$$A = \{1, 2, 3, 4\} \text{ i.e., "result is less than 5,"}$$

$$B = \{1, 3, 5\} \text{ i.e., "result is odd."}$$

What is the probability of A given B?

Probability of B given A?

In English,

What is the probability of the event "result is less than 5" given "result is odd"?

or

What is the probability of the event "result is odd" given "result is less than 5"?

Conditional Probability Example

Suppose we throw a fair die:

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \mathcal{F} = 2^\Omega, P(\{i\}) = \frac{1}{6}, i = 1 \dots 6$$

$$A = \{1, 2, 3, 4\} \text{ i.e., "result is less than 5,"}$$

$$B = \{1, 3, 5\} \text{ i.e., "result is odd."}$$

$$P(A) = \frac{2}{3}$$

$$P(B) = \frac{1}{2}$$

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

$$= \frac{P(\{1, 3\})}{P(B)}$$

$$= \frac{2}{3}$$

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

$$= \frac{1}{2}$$

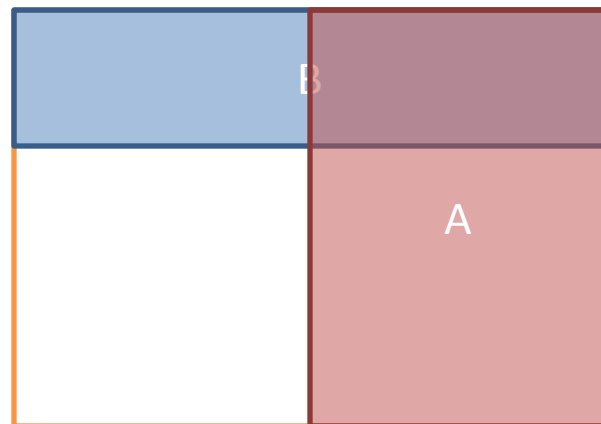
Independence

Two events A, B are called **independent** if $P(A \cap B) = P(A)P(B)$.

When $P(A) > 0$ this may be written $P(B|A) = P(B)$ (why?)
e.g., rolling two dice, flipping n coins etc.

Two events A, B are called **conditionally independent given C** when $P(A \cap B|C) = P(A|C)P(B|C)$.

When $P(A) > 0$ we may write $P(B|A, C) = P(B|C)$
e.g., “the weather tomorrow is independent of the weather yesterday, knowing the weather today.”



Bayes' Rule

Using the chain rule we may see:

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

Rearranging this yields **Bayes' rule**:

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

Often this is written as:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_i P(A|B_i)P(B_i)}$$

Where B_i are a partition of Ω (note the bottom is just the law of total probability).

Random Variables (informal)

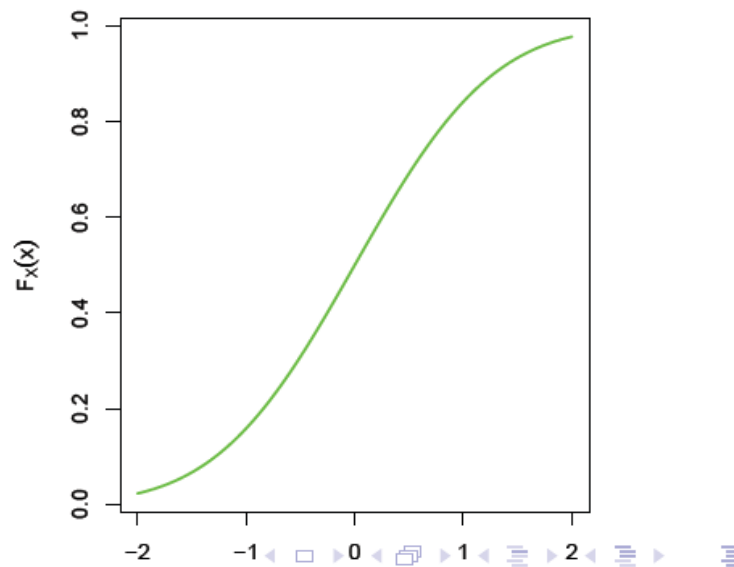
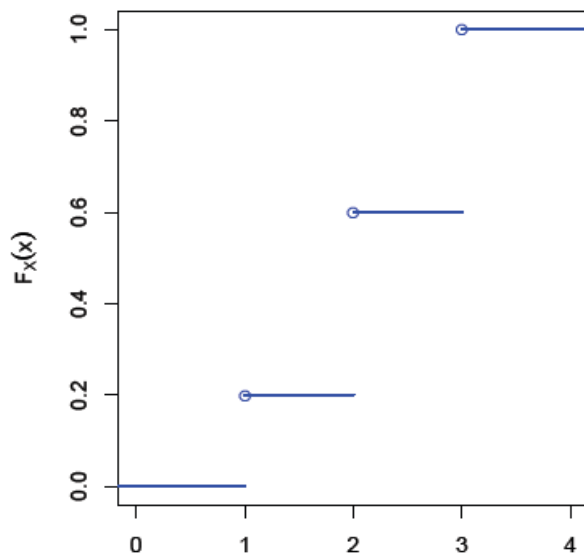
A **random variable** is a function $X : \Omega \rightarrow \mathbb{R}^d$

e.g.,

- ▶ Roll some dice, $X =$ sum of the numbers.
- ▶ Indicators of events: $X(\omega) = 1_A(\omega)$. e.g., toss a coin, $X = 1$ if it came up heads, 0 otherwise. Note relationship between the set theoretic constructions, and binary RVs.
- ▶ Give a few monkeys a typewriter, $X =$ fraction of overlap with complete works of Shakespeare.
- ▶ Throw a dart at a board, $X \in \mathbb{R}^2$ are the coordinates which are hit.

Distributions

- ▶ By considering random variables, we may think of probability measures as functions on the real numbers.
- ▶ Then, the probability measure associated with the RV is completely characterized by its **cumulative distribution function (CDF)**:
 $F_X(x) = P(X \leq x)$.
- ▶ If two RVs have the same CDF we call them **identically distributed**.
- ▶ We say $X \sim F_X$ or $X \sim f_X$ (f_X coming soon) to indicate that X has the distribution specified by F_X (resp, f_X).

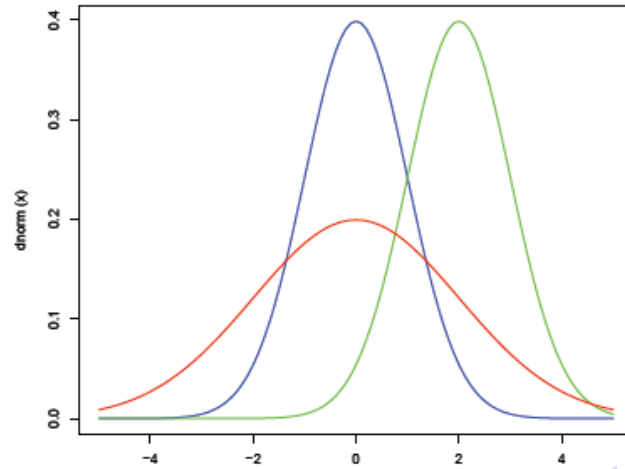


Discrete Distributions

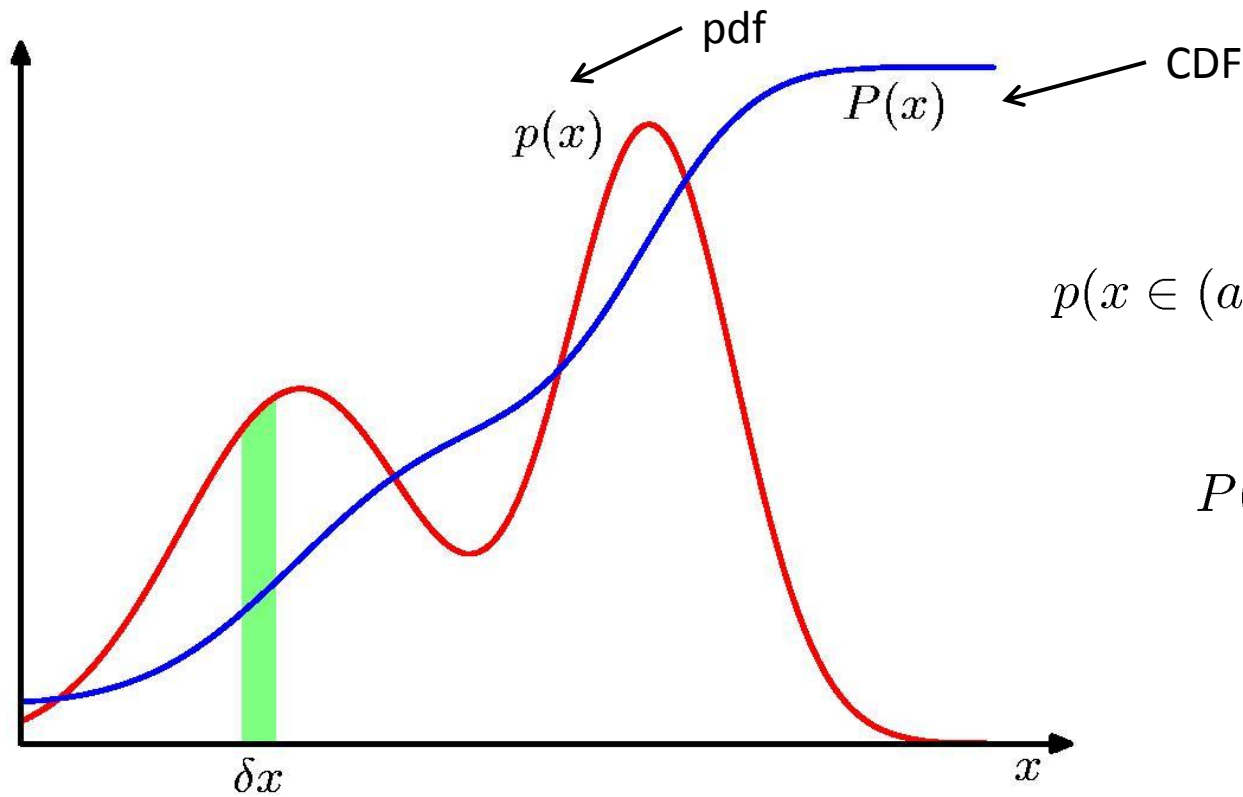
- ▶ If X takes on only a countable number of values, then we may characterize it by a **probability mass function (PMF)** which describes the probability of each value: $f_X(x) = P(X = x)$.
- ▶ We have: $\sum_x f_X(x) = 1$ (why?) – since each ω maps to one x , and $P(\Omega) = 1$.
- ▶ e.g., general discrete PMF: $f_X(x_i) = \theta_i$, $\sum_i \theta_i = 1, \theta_i \geq 0$.

Continuous Distribution

- ▶ When the CDF is continuous we may consider its derivative $f_X(x) = \frac{d}{dx} F_X(x)$.
- ▶ This is called the **probability density function (PDF)**.
- ▶ The probability of an interval (a, b) is given by $P(a < X < b) = \int_a^b f_X(x) dx$.
- ▶ The probability of any specific point c is zero: $P(X = c) = 0$ (why?).
- ▶ e.g., Uniform distribution: $f_X(x) = \frac{1}{b-a} \cdot 1_{(a,b)}(x)$
- ▶ e.g., Gaussian aka “normal:” $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
- ▶ Note that both families give probabilities for every interval on the real line, yet are specified by only two numbers.



Probability Densities



$$p(x \in (a, b)) = \int_a^b p(x) dx$$

$$P(z) = \int_{-\infty}^z p(x) dx$$


$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x) \, dx$$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$


Conditional Expectation
(discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$$

Approximate Expectation
(discrete and continuous)

- Note: in conditional expectation, y should be an instance, not RV.
- What is the expected value of a roll of a fair die?

Properties of Expectation

- $E[a] = a$ for any constant $a \in \mathbb{R}$.
- $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.
- (Linearity of Expectation) $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.
- For a discrete random variable X , $E[1\{X = k\}] = P(X = k)$.

Variance

We may consider the **variance** of a distribution:

$$\text{Var}(X) = E(X - EX)^2$$

This may give an idea of how “spread out” a distribution is.

A useful alternate form is:

$$\begin{aligned} E(X - EX)^2 &= E[X^2 - 2XE(X) + (EX)^2] \\ &= E(X^2) - 2E(X)E(X) + (EX)^2 \\ &= E(X^2) - (EX)^2 \end{aligned}$$

Variance of a coin toss?

- $\text{Var}[a] = 0$ for any constant $a \in \mathbb{R}$.
- $\text{Var}[af(X)] = a^2 \text{Var}[f(X)]$ for any constant $a \in \mathbb{R}$.

Common Discrete Random Variables

- $X \sim \text{Bernoulli}(p)$ (where $0 \leq p \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise.

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

- $X \sim \text{Binomial}(n, p)$ (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p .

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- $X \sim \text{Geometric}(p)$ (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

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

- $X \sim \text{Poisson}(\lambda)$ (where $\lambda > 0$): a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Common Continuous Random Variables

- $X \sim \text{Uniform}(a, b)$ (where $a < b$): equal probability density to every value between a and b on the real line.

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

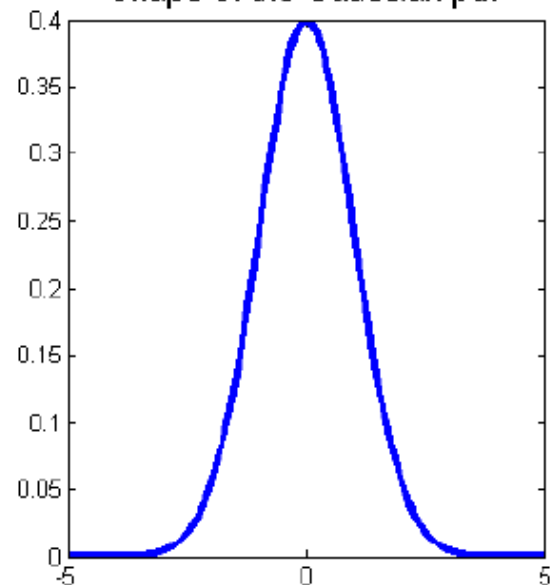
- $X \sim \text{Exponential}(\lambda)$ (where $\lambda > 0$): decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

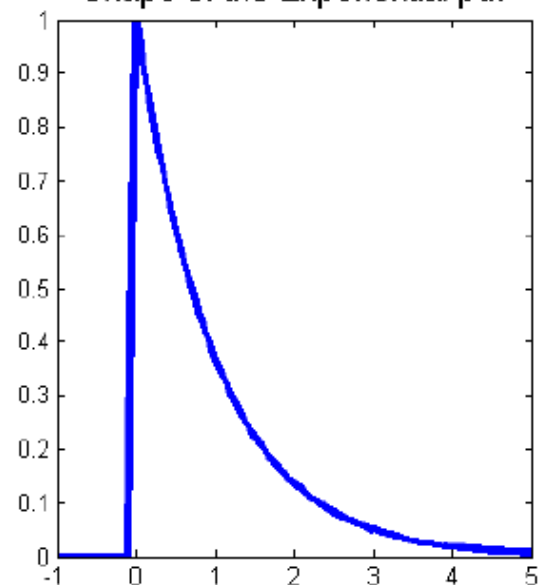
- $X \sim \text{Normal}(\mu, \sigma^2)$: also known as the Gaussian distribution

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

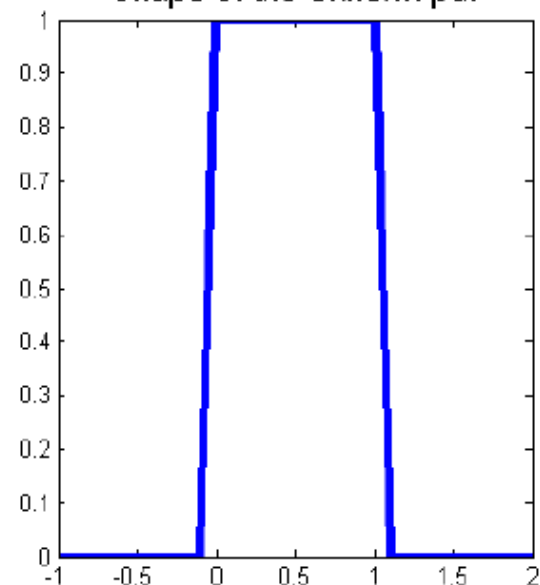
Shape of the Gaussian pdf



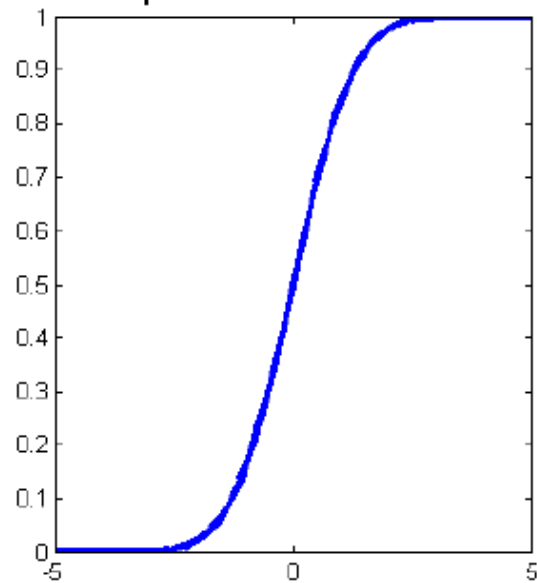
Shape of the Exponential pdf



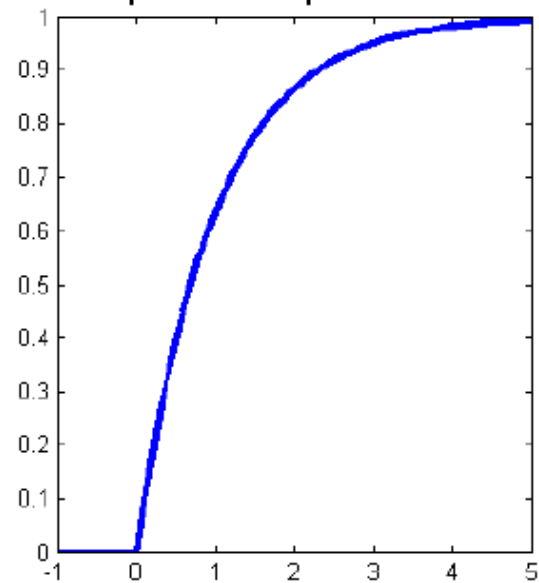
Shape of the Uniform pdf



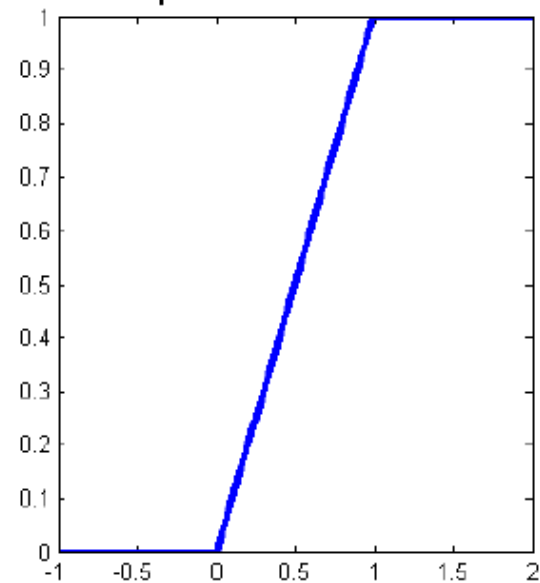
Shape of the Gaussian CDF



Shape of the Exponential CDF



Shape of the Uniform CDF

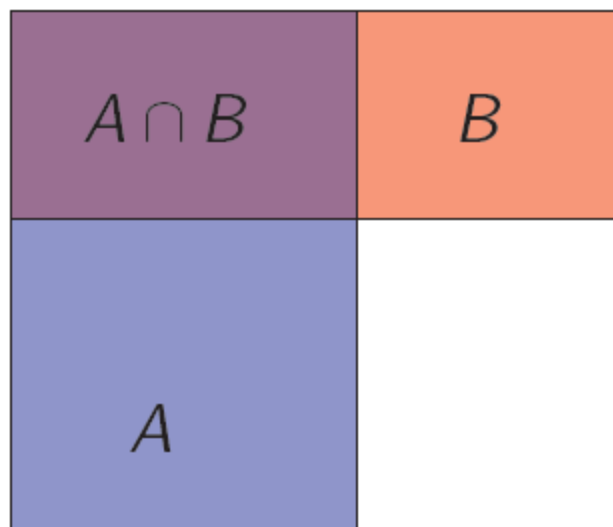


Properties of Random Variables

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0. \end{cases}$	p	$p(1 - p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1 - p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1 - p)^{k-1}$ for $k = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda} \lambda^x / x!$ for $k = 1, 2, \dots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a} \quad \forall x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x} \quad x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Multiple Random Variables

We may consider multiple functions of the same sample space, e.g., $X(\omega) = 1_A(\omega)$, $Y(\omega) = 1_B(\omega)$:



May represent the **joint distribution** as a table:

	$X=0$	$X=1$
$Y=0$	0.25	0.15
$Y=1$	0.35	0.25

We write the joint PMF or PDF as $f_{X,Y}(x,y)$

Marginalizing and Conditioning

We have similar constructions as we did in abstract prob. spaces:

- ▶ **Marginalizing:** $f_X(x) = \int_Y f_{X,Y}(x,y) dy$.

Similar idea to the law of total probability (identical for a discrete distribution).

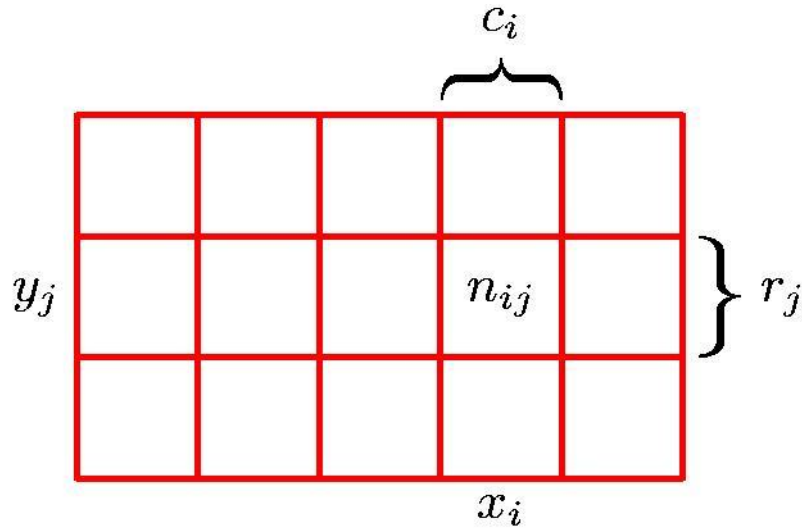
- ▶ **Conditioning:** $f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int_{\mathcal{X}} f_{X,Y}(x,y) dx}$.

Similar to previous definition.

Old?	Blood pressure?	Heart Attack?	P
0	0	0	0.22
0	0	1	0.01
0	1	0	0.15
0	1	1	0.01
1	0	0	0.18
...

How to compute
 $P(\text{heart attack}|\text{old})?$

Joint Probability



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

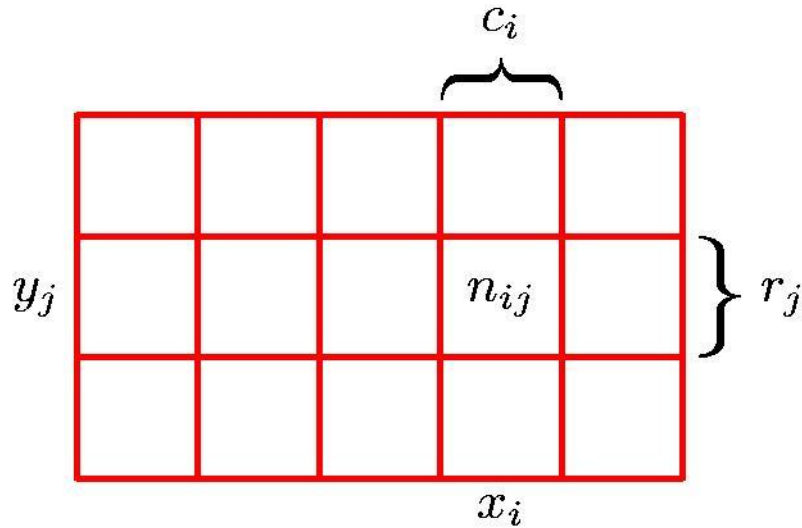
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Joint Probability



Sum Rule

$$\begin{aligned} p(X = x_i) &= \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^L n_{ij} \\ &= \sum_{j=1}^L p(X = x_i, Y = y_j) \end{aligned}$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

The Rules of Probability

Sum Rule

$$p(X) = \sum_Y p(X, Y)$$

Product Rule

$$p(X, Y) = p(Y|X)p(X)$$

Multiple Random Variables

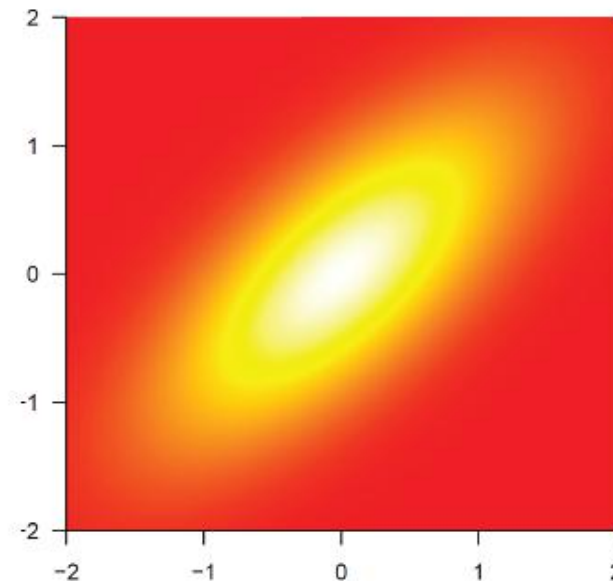
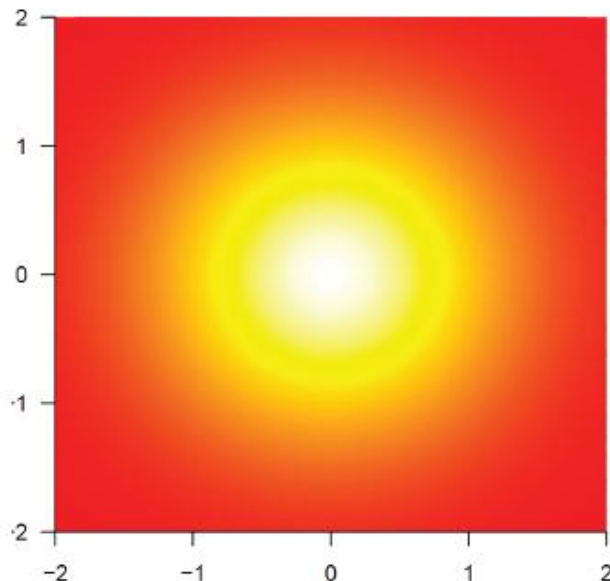
Two random variables are called **independent** when the joint PDF factorizes:

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

When RVs are independent and identically distributed this is usually abbreviated to "i.i.d."

Relationship to independent events: X, Y ind. iff

$\{\omega : X(\omega) \leq x\}, \{\omega : Y(\omega) \leq y\}$ are independent events for all x, y .



Expectation and Covariance

$$E[g(X, Y)] \triangleq \sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} g(x, y) p_{XY}(x, y).$$

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy.$$

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

$$\begin{aligned} \text{Cov}[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

Properties of Expectation and Covariance

- (Linearity of expectation) $E[f(X, Y) + g(X, Y)] = E[f(X, Y)] + E[g(X, Y)]$.
- $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$.
- If X and Y are independent, then $Cov[X, Y] = 0$.
- If X and Y are independent, then $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$.

Multiple Random Variables

We can define the **joint distribution function** of X_1, X_2, \dots, X_n , the **joint probability density function** of X_1, X_2, \dots, X_n , the **marginal probability density function** of X_1 , and the **conditional probability density function** of X_1 given X_2, \dots, X_n , as

$$\begin{aligned}F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) &= P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) \\f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) &= \frac{\partial^n F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)}{\partial x_1 \dots \partial x_n} \\f_{X_1}(X_1) &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_2 \dots dx_n \\f_{X_1|X_2, \dots, X_n}(x_1|x_2, \dots, x_n) &= \frac{f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)}{f_{X_2, \dots, X_n}(x_2, \dots, x_n)}\end{aligned}$$

To calculate the probability of an event $A \subseteq \mathbb{R}^n$ we have,

$$P((x_1, x_2, \dots, x_n) \in A) = \int_{(x_1, x_2, \dots, x_n) \in A} f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

Chain Rule and independence

Chain rule: From the definition of conditional probabilities for multiple random variables, one can show that

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= f(x_n | x_1, x_2, \dots, x_{n-1}) f(x_1, x_2, \dots, x_{n-1}) \\ &= f(x_n | x_1, x_2, \dots, x_{n-1}) f(x_{n-1} | x_1, x_2, \dots, x_{n-2}) f(x_1, x_2, \dots, x_{n-2}) \\ &= \dots = f(x_1) \prod_{i=2}^n f(x_i | x_1, \dots, x_{i-1}). \end{aligned}$$

Independence: For multiple events, A_1, \dots, A_k , we say that A_1, \dots, A_k are **mutually independent** if for any subset $S \subseteq \{1, 2, \dots, k\}$, we have

$$P(\cap_{i \in S} A_i) = \prod_{i \in S} P(A_i).$$

Likewise, we say that random variables X_1, \dots, X_n are independent if

$$f(x_1, \dots, x_n) = f(x_1) f(x_2) \cdots f(x_n).$$

Multivariate Expectation

Expectation: Consider an arbitrary function from $g : \mathbb{R}^n \rightarrow \mathbb{R}$. The expected value of this function is defined as

$$E[g(X)] = \int_{\mathbb{R}^n} g(x_1, x_2, \dots, x_n) f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n, \quad (5)$$

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix},$$

$$E[g(X)] = \begin{bmatrix} E[g_1(X)] \\ E[g_2(X)] \\ \vdots \\ E[g_m(X)] \end{bmatrix}.$$

Multivariate Covariance

Covariance matrix: For a given random vector $X : \Omega \rightarrow \mathbb{R}^n$, its covariance matrix Σ is the $n \times n$ square matrix whose entries are given by $\Sigma_{ij} = \text{Cov}[X_i, X_j]$.

From the definition of covariance, we have

$$\begin{aligned}
 \Sigma &= \begin{bmatrix} \text{Cov}[X_1, X_1] & \cdots & \text{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_n, X_1] & \cdots & \text{Cov}[X_n, X_n] \end{bmatrix} \\
 &= \begin{bmatrix} E[X_1^2] - E[X_1]E[X_1] & \cdots & E[X_1X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix} \\
 &= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] & \cdots & E[X_n^2] \end{bmatrix} - \begin{bmatrix} E[X_1]E[X_1] & \cdots & E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]E[X_1] & \cdots & E[X_n]E[X_n] \end{bmatrix} \\
 &= E[XX^T] - E[X]E[X]^T = \dots = E[(X - E[X])(X - E[X])^T].
 \end{aligned}$$

- $\Sigma \succeq 0$; that is, Σ is positive semidefinite.
- $\Sigma = \Sigma^T$; that is, Σ is symmetric.

Likelihood Functions

- Why is Bayes' so useful in learning? Allows us to evaluate a parameter setting w :

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

$$p(D) = \sum_w p(D|w)p(w)$$

- The likelihood function, $p(D|w)$, is evaluated for observed data D as a function of w . It expresses how probable the observed data set is for various parameter settings w .
- Bayes' in words: posterior \propto likelihood \times prior

Maximum Likelihood

- Maximum likelihood:
 - choose parameter setting w that maximizes likelihood function $p(D|w)$.
 - Choose the value of w that maximizes the probability of observed data.
 - The negative log of the likelihood is called an error function.
 - Because negative logarithm is a monotonically decreasing function, maximizing likelihood is equivalent to minimizing the error.

Case Study: The Gaussian Distribution

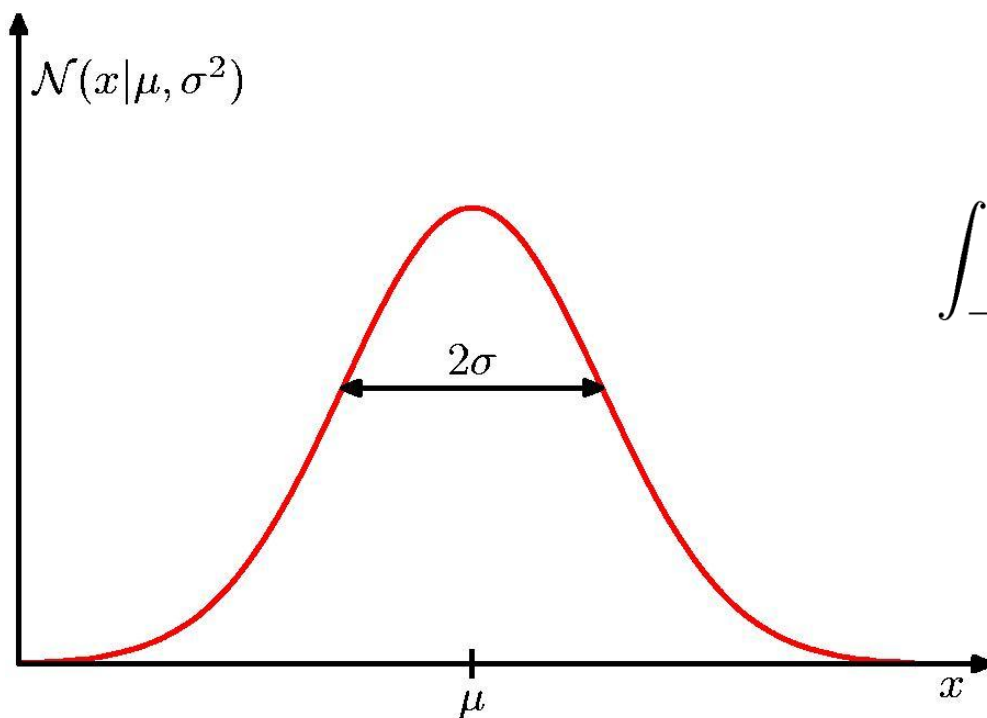
- Data $\{x_1, \dots, x_N\}$
- where
 - drawn from Normal distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

- independent and identically distributed (i.i.d)
- How to estimate the parameter μ and σ that maximizes the log-likelihood?

Case Study: The Gaussian Distribution

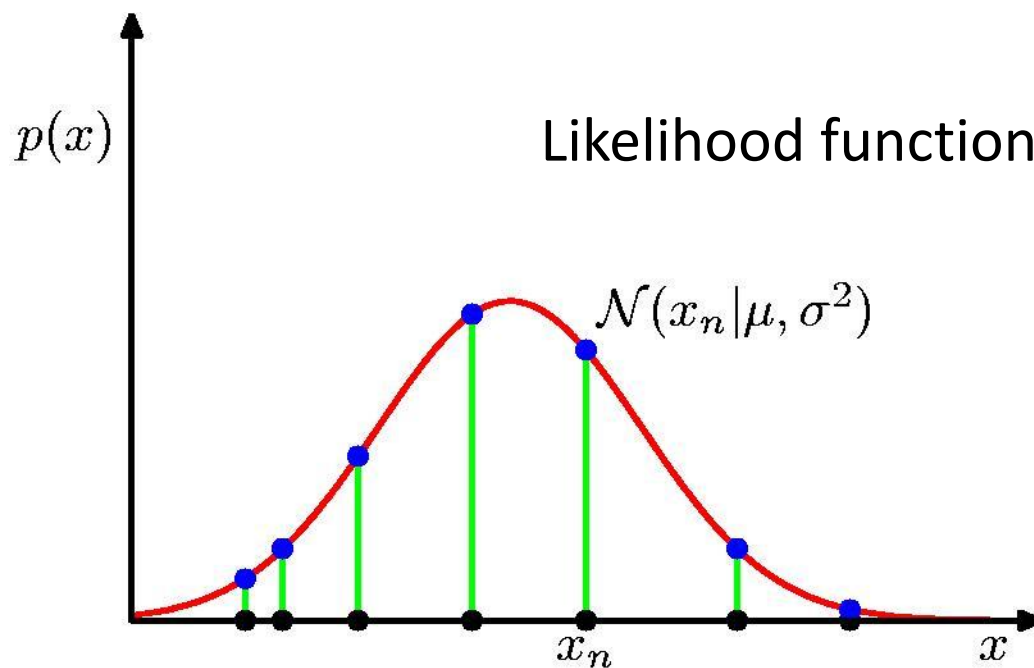
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$



$$\mathcal{N}(x|\mu, \sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, dx = 1$$

Gaussian Parameter Estimation



$$p(\mathbf{x} | \mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n | \mu, \sigma^2)$$

Maximum (Log) Likelihood

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$$

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \qquad \sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$$

- Questions?