Probability

Note: Unless otherwise noted all references including images are from the required textbook, Machine Learning: A Probabilistic Perspective by Kevin P. Murphy.

What is Probability?

There are at least two different interpretations of probability. One is called the **frequentist** interpretation. In this view, probabilities represent long run frequencies of events. For example, the statement “the probability that a coin will land heads is 0.5” means that, if we flip the coin many times, we expect it to land heads about half the time.

What is Probability?

The other interpretation is called the **Bayesian** interpretation of probability. In this view, probability is used to quantify our **uncertainty** about something; hence it is fundamentally related to information rather than repeated trials. In the Bayesian view, “the probability that a coin will land heads is 0.5” means we believe the coin is equally likely to land heads or tails on the next toss.

What is Probability?

One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long term frequencies. For example, we might have received a specific email message, and want to compute the probability it is spam. In this case, the idea of repeated trials does not make sense, but the Bayesian interpretation is valid and indeed quite natural.

Probability

The expression $p(A)$ denotes the probability that the event *A* is true. For example, *A* might be the logical expression "it will rain tomorrow". We require that $0 \leq p(A) \leq 1$, where $p(A) = 0$ means the event definitely will not happen, and $p(A) = 1$ means the event definitely will happen.

Probability

We write $p(\bar{A})$ to denote the probability of the event not A; this is defined as $p(\bar{A}) = 1 - p(A)$. We will often write $A = 1$ to mean the event *A* is true, and $A = 0$ to mean the event A is false.

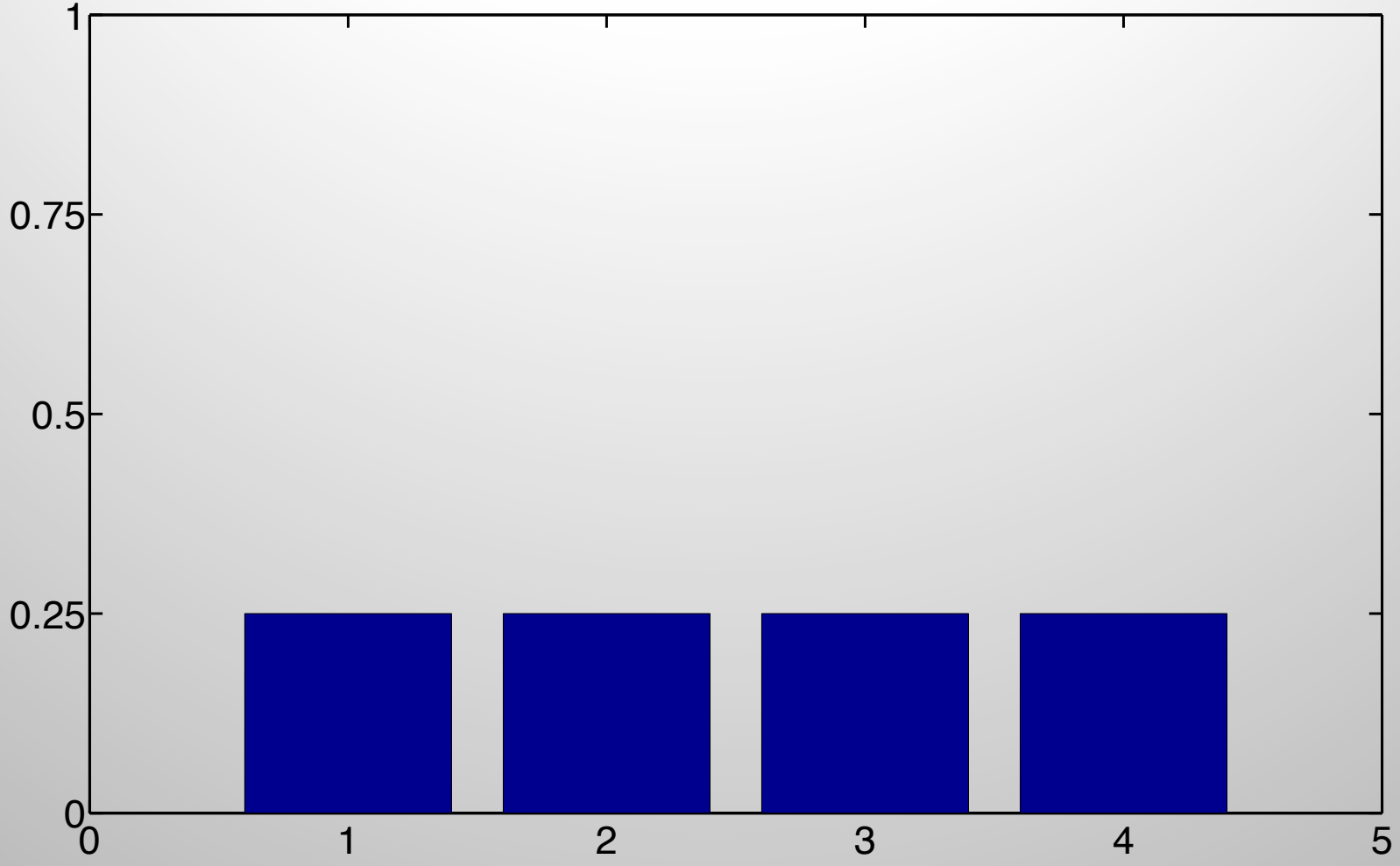
Discrete Random Variables

We can extend the notion of binary events by defining a **discrete random variable** *X*, which can take on any value from a finite or countably infinite set $\mathcal{X}$. We denote the probability of the event that $X = x$ by $p(X = x)$, or just $p(x)$ for short.

Discrete Random Variables

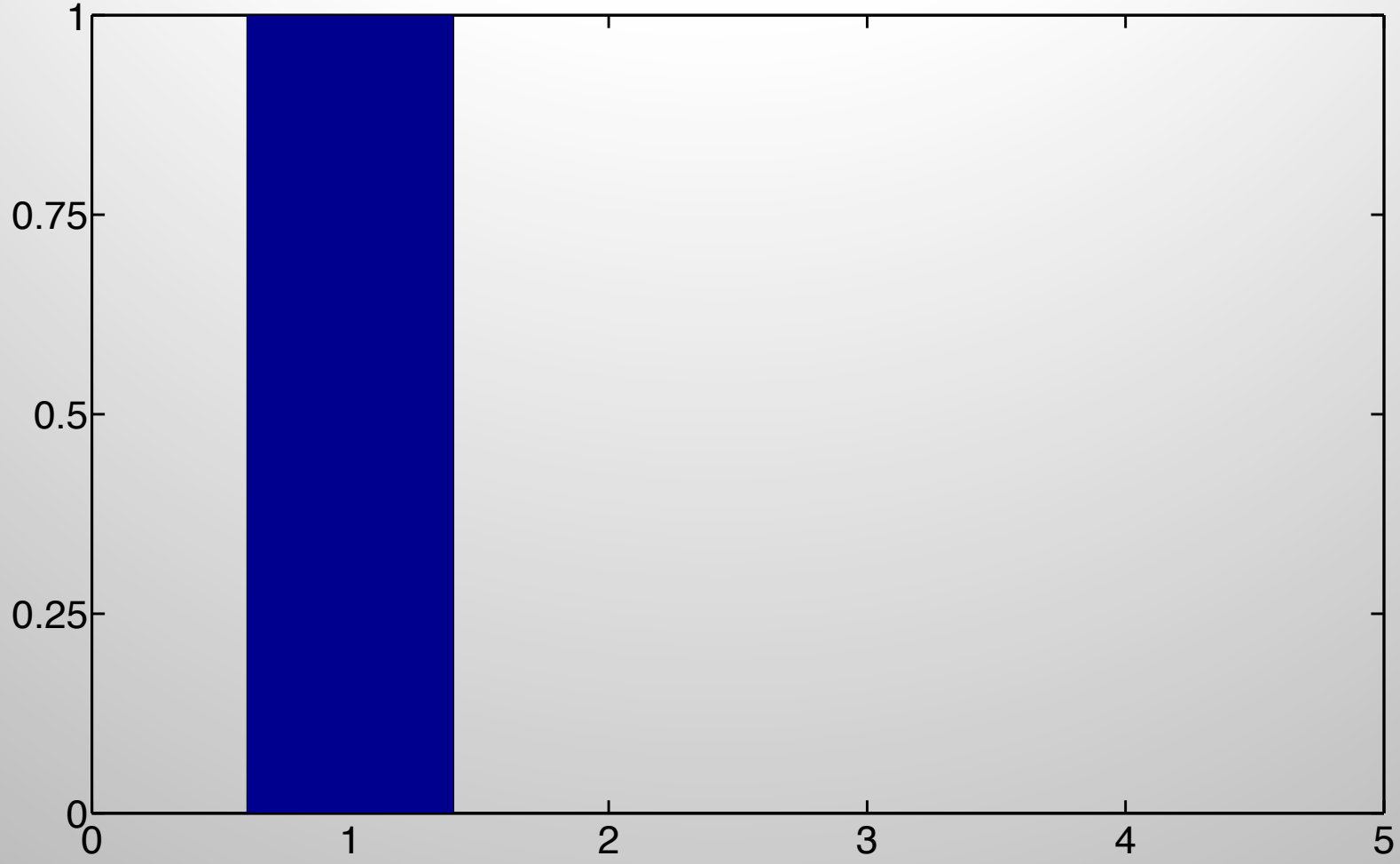
Here $p()$ is called a **probability mass function** or **pmf**. This satisfies the properties $0 \leq p(x) \leq 1$ and $\sum_{x \in \mathcal{X}} p(x) = 1$. The next two slides show two pmf's defined on the finite **state space** $\mathcal{X} = \{1,2,3,4\}$.

Discrete Random Variables



A uniform distribution on $\{1,2,3,4\}$, with $p(x = k) = 1/4$.

Discrete Random Variables



A degenerate distribution $p(x) = 1$ if $x = 1$, and $p(x) = 0$ if $x \in \{2,3,4\}$.

Probability of Union of Two Events

Given two events, *A* and *B*, we define the probability of *A* or *B* as follows:

$$
\begin{aligned}
p(A \vee B) &= p(A) + p(B) - p(A \wedge B) \\
&= p(A) + p(B) \text{ if } A \text{ and } B \text{ are mutually exclusive}
\end{aligned}
$$

Joint Probabilities

We define the probability of the joint event *A* and *B* as follows:

$$p(A, B) = p(A \wedge B) = p(A|B)p(B)$$

This is sometimes called the **product rule**.

Joint Probabilities

Given a **joint distribution** on two events $p(A, B)$, we define the **marginal distribution** as follows:

$$p(A) = \sum_b p(A, B) = \sum_b p(A|B = b)p(B = b)$$

where we are summing over all possible states of *B*. We can define $p(B)$ similarly. This is sometimes called the **sum rule** or the **rule of total probability.**

Joint Probabilities

The product rule can be applied multiple times to yield the **chain rule** of probability:

$$p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2, X_1)p(X_4|X_1, X_2, X_3)\ldots p(X_D|X_{1:D-1})$$

where the Matlab-like notation 1 : *D* denotes the set $\{1,2, \ldots, D\}$.

Conditional Probability

We define the **conditional probability** of event *A*, given that event *B* is true, as follows:

$$p(A|B) = \frac{p(A,B)}{p(B)} \text{ if } p(B) > 0$$

Bayes' Rule

Combining the definition of conditional probability with the product and sum rules yields **Bayes' rule** , also called **Bayes' Theorem**:

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} = \frac{p(X = x)p(Y = y|X = x)}{\sum_{x'} p(X = x')p(Y = y|X = x')}$$

Example: Medical Diagnosis

Suppose you are a woman in your 40s, and you decide to have a medical test for breast cancer called a mammogram. If the test is positive, what is the probability you have cancer? That obviously depends on how reliable the test is.

Example: Medical Diagnosis

Suppose you are told the test has a **sensitivity** of 80%, which means, if you have cancer, the test will be positive with probability 0.8. In other words:

$$p(x = 1|y = 1) = 0.8$$

where $x = 1$ is the event the mammogram is positive, and $y = 1$ is the event you have breast cancer.

Example: Medical Diagnosis

Many people conclude that, if they have a positive mammogram, they are 80% likely to have cancer.

However, this is not correct as it ignores the prior probability of having breast cancer, which is quite low:

$$p(y = 1) = 0.004$$

Example: Medical Diagnosis

Ignoring this prior is called the **base rate fallacy**. We also need to take into account the fact that the test may be a **false positive** or **false alarm**. Unfortunately, such false positives are quite likely (with current screening technology):

$$p(x = 1|y = 0) = 0.1$$

Example: Medical Diagnosis

Combining these three terms using Bayes' rules, we can compute the correct answer as follows:

$$
\begin{aligned}
p(y=1|x=1) &= \frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1)+p(x=1|y=0)p(y=0)} \\
&= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031
\end{aligned}
$$

where $p(y=0) = 1 - p(y=1) = 0.996$.

In other words, if you test positive, you only have about a 3% chance of actually having breast cancer.

Example: Generative Classifiers

We can generalize the medical diagnosis example to classify feature vectors x of arbitrary type as follows:

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(y = c|\boldsymbol{\theta})p(\mathbf{x}|y = c, \boldsymbol{\theta})}{\sum_{c'} p(y = c'|\boldsymbol{\theta})p(\mathbf{x}|y = c', \boldsymbol{\theta})}$$

This is called a **generative classifier**, since it specifies how to generate the data using **class-conditional density** $p(\mathrm{x}|y = c)$ and the class prior $p(y = c)$.

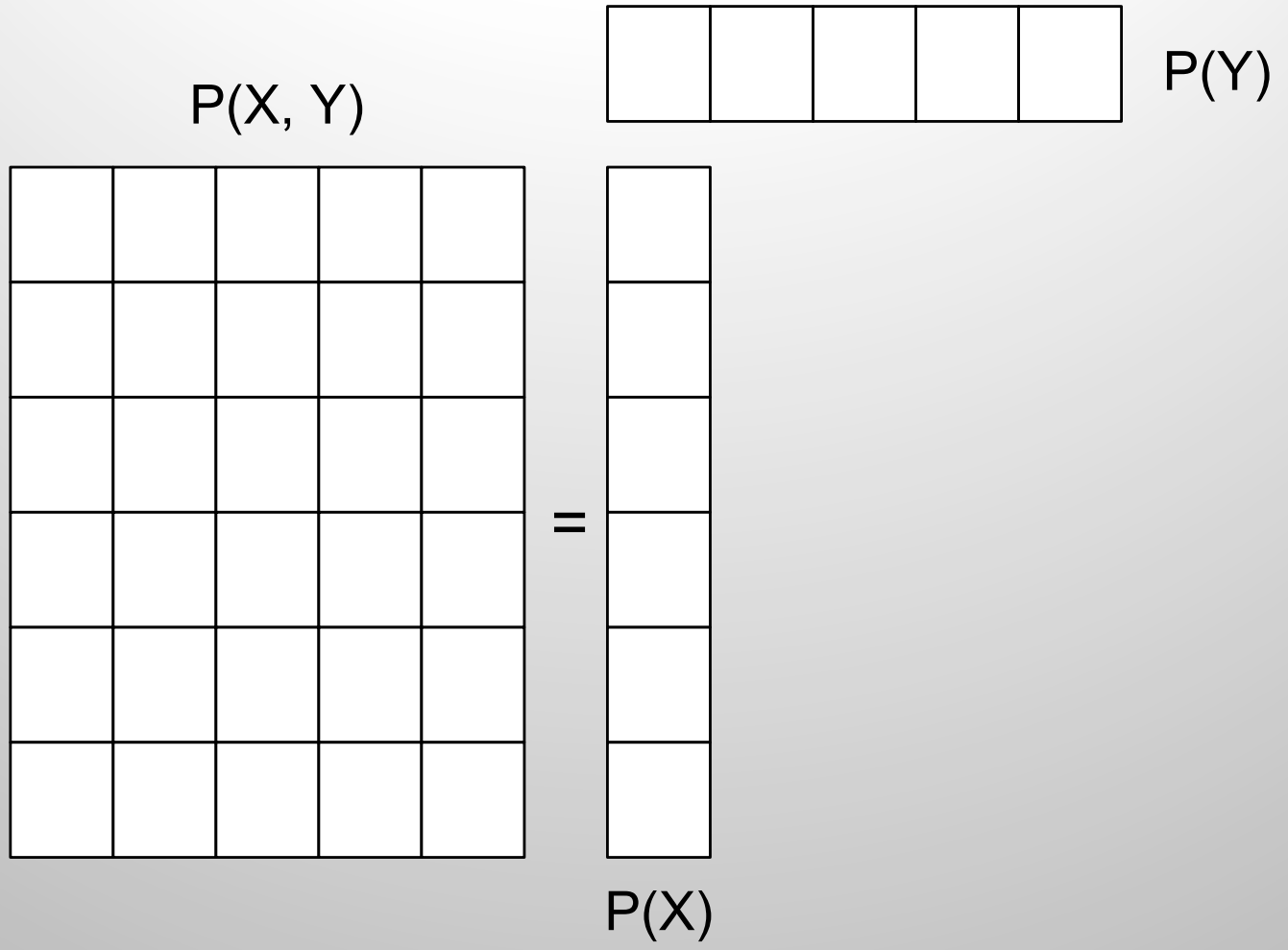
Unconditional Independence

We say *X* and *Y* are **unconditionally independent** or **marginally independent**, denoted $X \perp Y$, if we can represent the joint as the product of the two marginals:

$$X \perp Y \iff p(X,Y) = p(X)p(Y)$$

In general, we say a set of variables is mutually independent if the joint can be written as a product of marginals.

Unconditional Independence



Computing $p(x, y) = p(x)p(y)$, where $X \perp Y$

Conditional Independence

Unconditional independence is rare, because most variables can influence most other variables. However, usually this influence is mediated via other variables rather than being direct. We therefore say *X* and *Y* are **conditionally independent** (CI) given *Z* iff the conditional joint can be written as a product of conditional marginals:

$$X \perp Y|Z \iff p(X,Y|Z) = p(X|Z)p(Y|Z)$$

Continuous Random Variables

Suppose *X* is some uncertain continuous quantity. The probability that *X* lies in any interval $a \leq X \leq b$ can be computed as follows. Define the events

$A = (X \leq a)$,

$B = (X \leq b)$ and

$W = (a < X \leq b\ )$

Continuous Random Variables

We have that $B = A \vee W$, and since *A* and *W* are mutually exclusive, the sum rules gives

$$p(B) = p(A) + p(W)$$

And hence

$$p(W) = p(B) - p(A)$$

Continuous Random Variables

Define the function $F(q) \triangleq p(X \leq q)$. This is called the **cumulative distribution function** or **cdf** of *X*. This is a monotonically non-decreasing function. Using the notation we have

$$p(a < X \leq b) = F(b) - F(a)$$

Continuous Random Variables

Now define $f(x) = \frac{d}{dx} F(x)$ (we assume this derivative exists); this is called the **probability density function** or **pdf**. Given a pdf, we can compute the probability of a continuous variable being in a finite interval as follows:

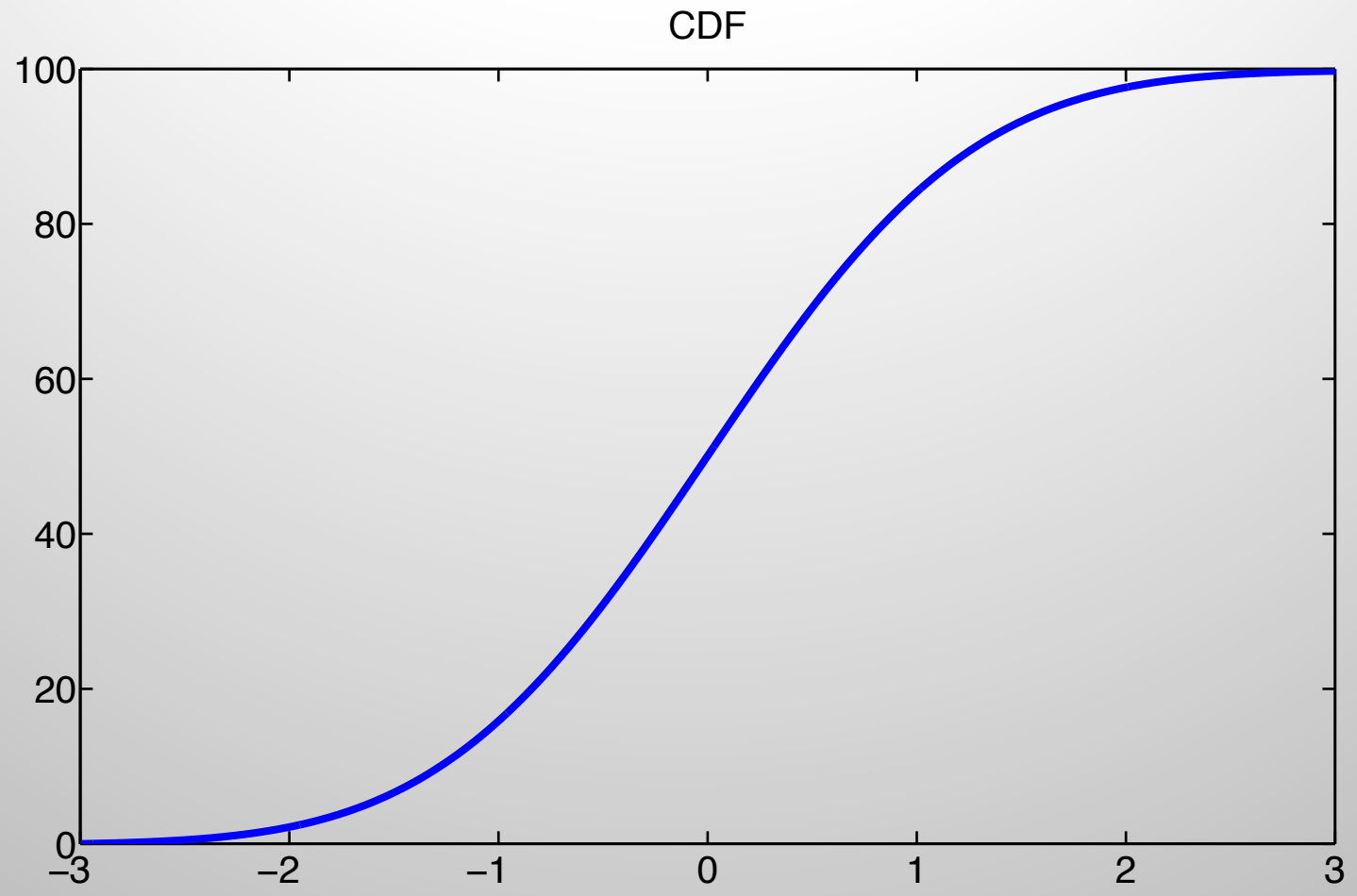
$$P(a < X \leq b) = \int_a^b f(x)dx$$

Continuous Random Variables

As the size of the interval gets smaller, we can write

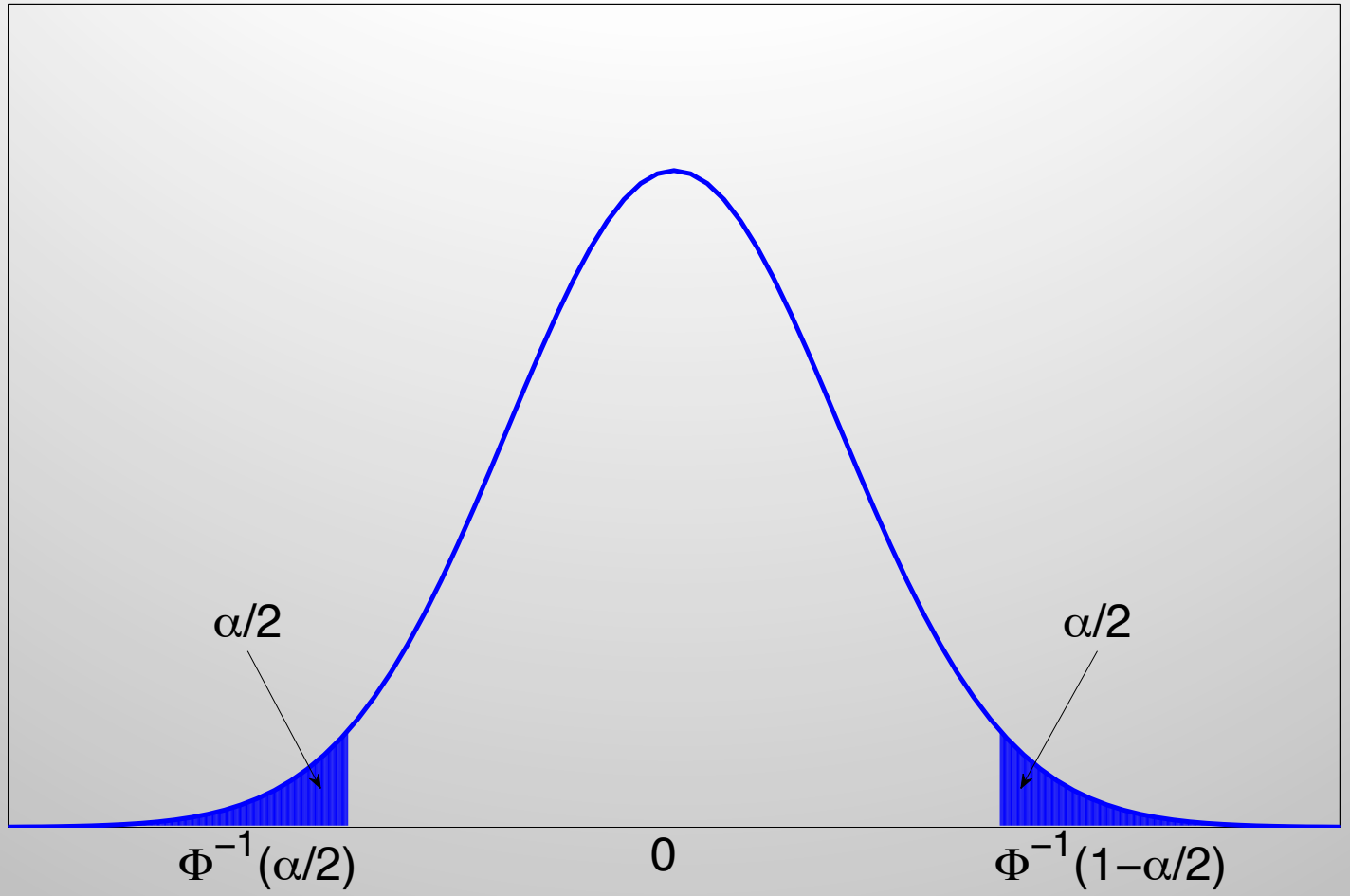
$$P(x \leq X \leq x + dx) \approx p(x)dx$$

Continuous Random Variables



Example of a monotonically non-decreasing function: standard normal $\mathcal{N}(0,1)$.

Continuous Random Variables



Corresponding pdf

Mean

The most familiar property of a distribution is its **mean**, or **expected value**, denoted by μ . For discrete random variables, it is defined as $\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x\, p(x)$, and for continuous random variables, it is defined as

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\, p(x) dx.$$

If this integral is not finite, the mean is not defined.

Variance

The variance is a measure of the "spread" denoted by σ^2. This is defined as follows:

$$\begin{aligned} \mathrm{var}\left[X\right] &\triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x)dx \\ &= \int x^2 p(x)dx + \mu^2 \int p(x)dx - 2\mu \int xp(x)dx = \mathbb{E}\left[X^2\right] - \mu^2 \end{aligned}$$

Variance

From which we derive the useful result

$$\mathbf{E}\left[X^2\right] = \mu^2 + \sigma^2$$

The **standard deviation** is defined as

$$\mathrm{std}\left[X\right] \triangleq \sqrt{\mathrm{var}\left[X\right]}$$

Binomial Distribution

Suppose we toss a coin n times. Let $X \in \{0, \ldots, n\}$ be the number of heads. If the probability of heads is θ, then we say X has a **binomial** distribution, written as $X \sim Bin(n, \theta)$. The pmf is given by

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Binomial Distribution

where

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

is the number of ways to choose k items from n (known as the **binomial coefficient**, and is pronounced “n choose k”). This distribution has the following mean and variance

$$\mathbf{mean} = \theta, \quad \mathbf{var} = n\theta(1-\theta)$$

Binomial Distribution

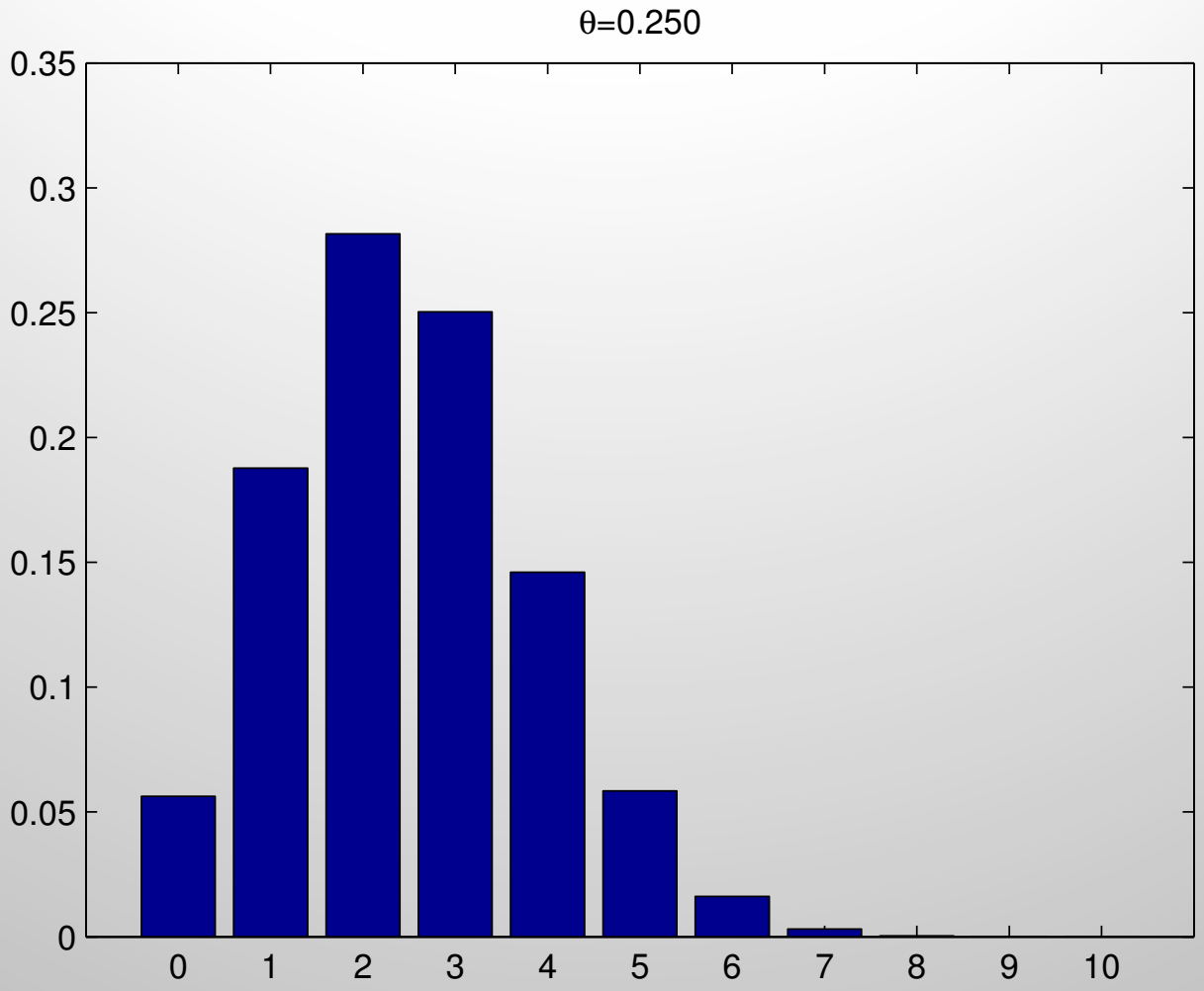


Illustration of the binomial distribution with $n = 10$ and $\theta = 0.25$

Binomial Distribution

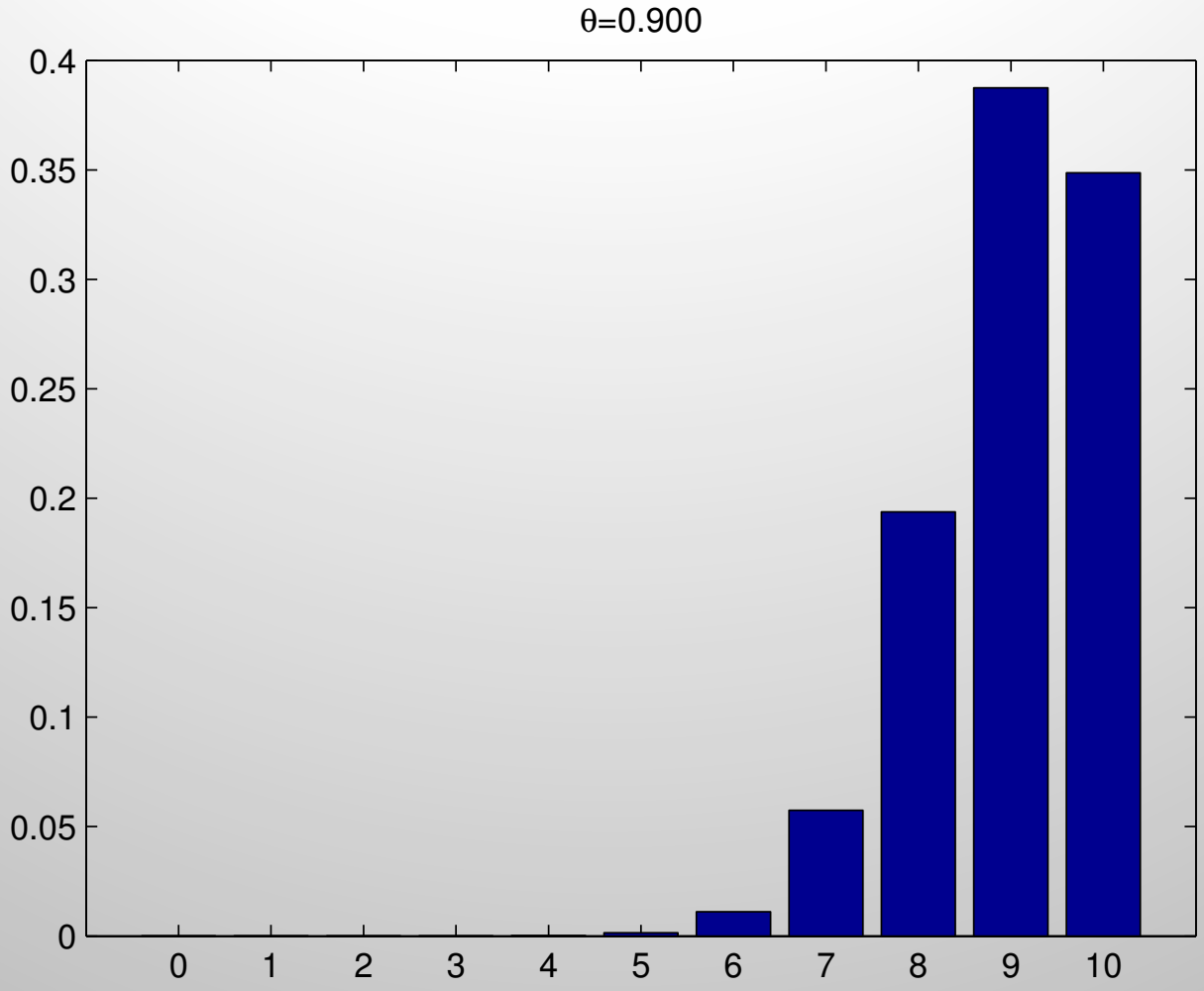


Illustration of the binomial distribution with $n = 10$ and $\theta = 0.9$

Bernoulli Distribution

Now suppose we toss a coin only once. Let $X \in \{0,1\}$ be a binary random variable, with probability of "success" or "heads" of θ. We say that X has a **Bernoulli** distribution. This is written as $X \sim Ber(\theta)$, where the pmf is defined as

$$\mathrm{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$$

Bernoulli Distribution

In other words,

$$\mathrm{Ber}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

This is obviously just a special case of a Binomial distribution with $n = 1$.

Multinomial Distribution

To model the outcomes of tossing a *K*-sided die, we can use the **multinomial** distribution. This is defined as follows: let $\mathrm{x} = (x_1, \ldots, x_k)$ be a random vector, where x_j is the number of times side j of the die occurs. Then x has the following pmf:

$$\mathrm{Mu}(\mathbf{x}|n, \boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

Multinomial Distribution

where θ_j is the probability that side j shows up, and

$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1! x_2! \cdots x_K!}$$

is the **multinomial coefficient** (the number of ways to divide a set of size $n = \sum_{k=1}^{K} x_k$ into subsets with sizes x_1 up to x_k).

Multinoulli Distribution

Now suppose $n = 1$. This is like rolling a *K*-sided die once, so x will be a vector of 0s and 1s (a bit vector), in which only one bit can be turned on. Specifically, if the die shows up as face *k*, then the k^{th} bit will be on.

In this case, we can think of x as being a scalar categorical random variable with *K* states (values), and x is its **dummy encoding**, that is,
$\mathrm{x} = [\mathbb{I}(x = 1), \ldots, \mathbb{I}(x = K)]$.

Multinoulli Distribution

For example, if $K = 3$, we encode the states 1, 2 and 3 as $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. This is also called a **one-hot encoding**, since we imagine that only one of the K "wires" is "hot" or on. In this case, the pmf becomes

$$\mathrm{Mu}(\mathbf{x}|1,\boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}$$

This is special case, which is also known as **categorical** or **discrete** distribution.

Poisson Distribution

We say that $X \in \{0, 1, 2, \ldots\}$ has a **Poisson** distribution with parameter $\lambda > 0$, written $X \sim Poi(\lambda)$, if its pmf is

$$\mathrm{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

$e^{-\lambda}$ is the normalization constant, which ensures the distribution sums to 1. The Poisson distribution is often used as a model for counts of rare events like traffic accidents.

Poisson Distribution

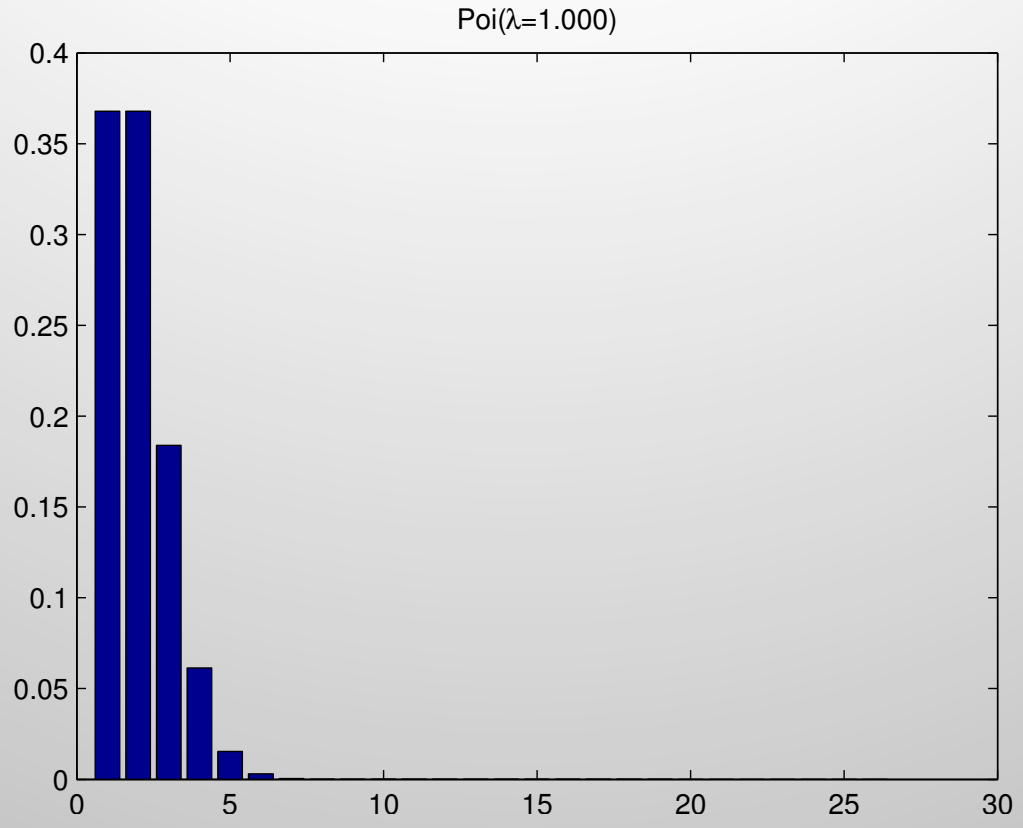


Illustration of Poisson distribution for $\lambda = 1$.

Poisson Distribution

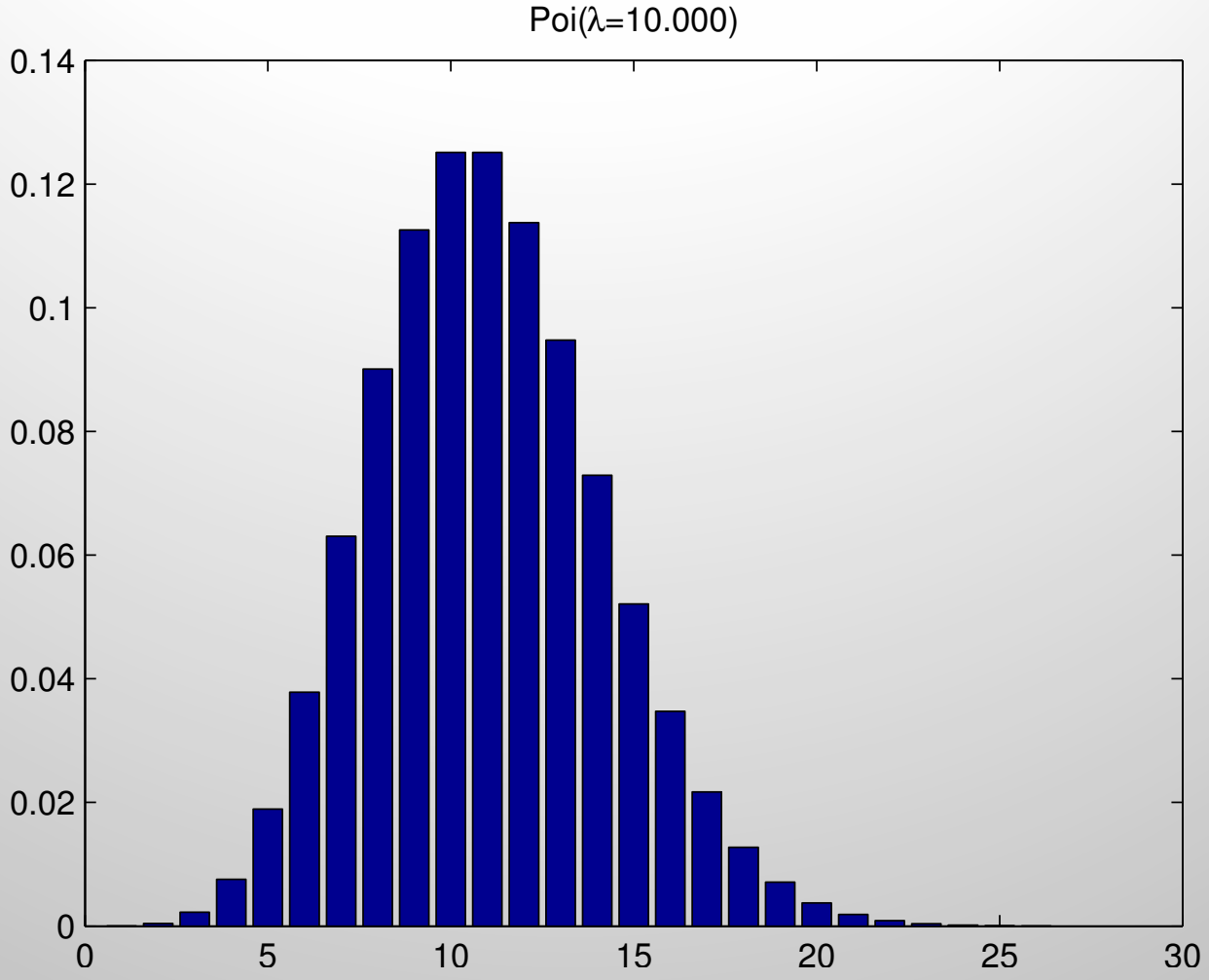


Illustration of Poisson distribution for $\lambda = 10$.

Gaussian (normal) Distribution

The most widely used distribution in statistics and machine learning is the **Gaussian** or **normal** distribution. Its pdf is given by

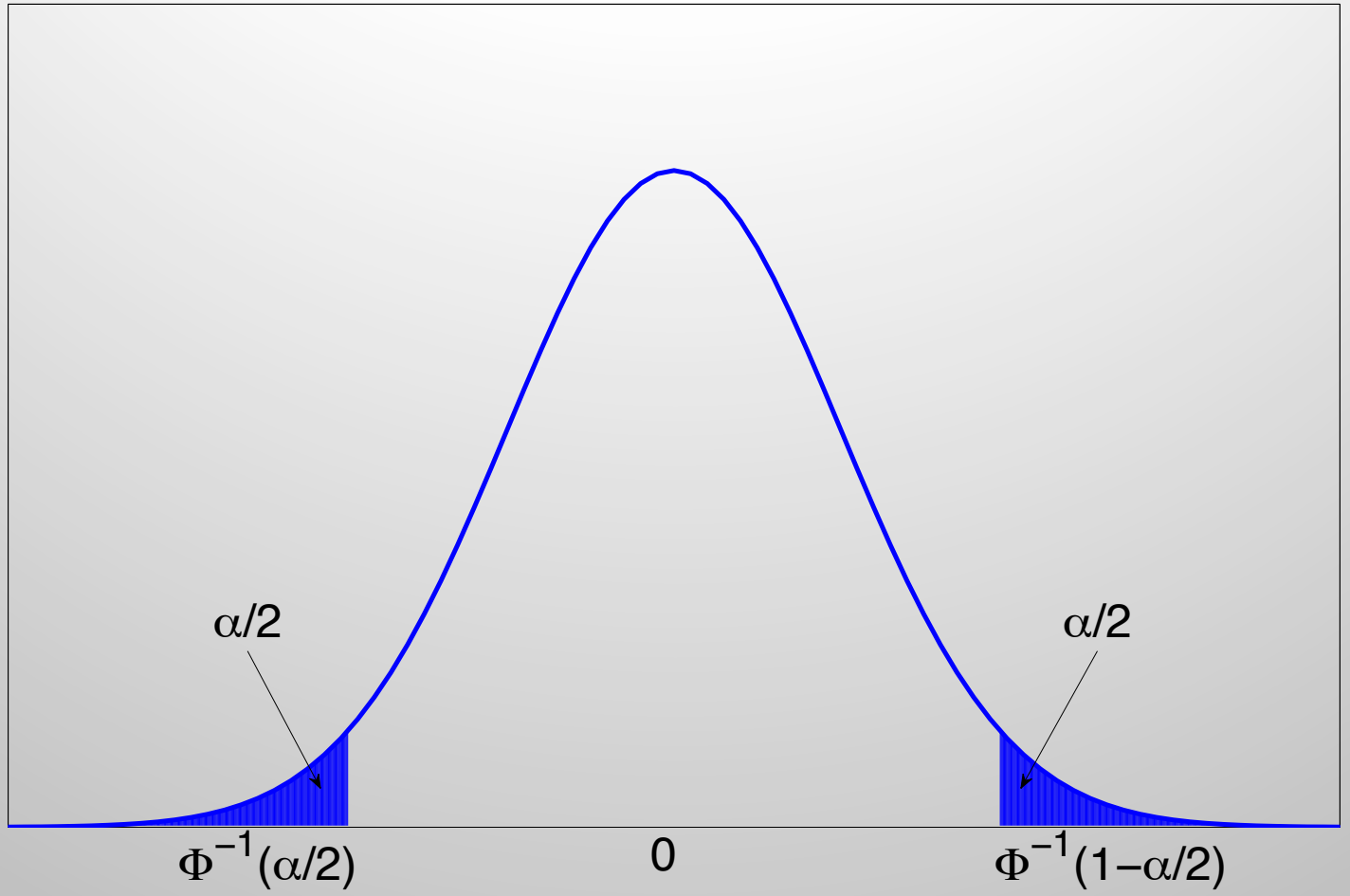
$$\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Gaussian (normal) Distribution

Here $\mu = \mathbb{E}[X]$ is the mean (and mode), and $\sigma^2 = \mathrm{var}[X]$ is the variance. $\sqrt{2\pi\sigma^2}$ is the normalization constant needed to ensure the density integrates to 1.

We write $X \sim \mathcal{N}(\mu, \sigma^2)$ to denote that $p(X = x) = \mathcal{N}(\mu, \sigma^2)$. $If\ X \sim \mathcal{N}(0,1)$, we say *X* follows a **standard normal** distribution.

Gaussian (normal) Distribution



pdf of standard normal distribution $X \sim \mathcal{N}(0,1)$.

Gaussian (normal) Distribution

We will often talk about the **precision** of a Gaussian, by which we mean the inverse variance: $\lambda = 1/\sigma^2$. A high precision means a narrow distribution (low variance) centered on μ.

Joint Probability Distribution

A joint probability distribution has the form $p(x_1, \ldots, x_D)$ for a set of $D > 1$ variables, and models the (stochastic) relationships between the variables. If all the variables are discrete, we can represent the joint distribution as a big multi-dimensional array, with one variable per dimension. However, the number of parameters needed to define such a model is $O(K^D)$, where K is the number of states for each variable.

Covariance

The covariance between two random variables *X* and *Y* measures the degree to which *X* and *Y* are (linearly) related. Covariance is defined as

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

Covariance

If x is a *d*-dimensional random vector, its covariance matrix is defined to be the following symmetric, positive definite matrix:

$$\begin{aligned}\operatorname{cov}[\mathbf{x}] &\triangleq \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T\right] \\ &= \begin{pmatrix} \operatorname{var}[X_1] & \operatorname{cov}[X_1,X_2] & \cdots & \operatorname{cov}[X_1,X_d] \\ \operatorname{cov}[X_2,X_1] & \operatorname{var}[X_2] & \cdots & \operatorname{cov}[X_2,X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}[X_d,X_1] & \operatorname{cov}[X_d,X_2] & \cdots & \operatorname{var}[X_d] \end{pmatrix}\end{aligned}$$

Correlation

Covariances can be between 0 and infinity. Sometimes it is more convenient to work with a normalized measure, with a finite upper bound. The (Pearson) correlation coefficient between *X* and *Y* is defined as

$$\operatorname{corr}[X, Y] \triangleq \frac{\operatorname{cov}[X, Y]}{\sqrt{\operatorname{var}[X] \operatorname{var}[Y]}}$$

Correlation

A **correlation matrix** has the form

$$\mathbf{R} = \begin{pmatrix} \operatorname{corr}[X_1, X_1] & \operatorname{corr}[X_1, X_2] & \cdots & \operatorname{corr}[X_1, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{corr}[X_d, X_1] & \operatorname{corr}[X_d, X_2] & \cdots & \operatorname{corr}[X_d, X_d] \end{pmatrix}$$

Multivariate Gaussian

The **multivariate Gaussian** or **multivariate normal (MVN)** is the most widely used joint probability density function for continuous variables. The pdf of the MVN in D dimensions is defined by the following:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

where $\mu = \mathbb{E}[x] \in \mathbb{R}^D$ is the mean vector, and $\Sigma = \text{cov}[\text{x}]$ is the $D \times D$ covariance matrix.