# Lecture 1.2: Probability and Statistics
# CSC 84020 - Machine Learning

Andrew Rosenberg

January 29, 2009

# Today

- Probability and Statistics

# Background

What exposure have you had to probability and statistics?

- Conditional probabilities?
- Bayes rule?
- The difference between a posterior, a conditional and a prior?

Classical Artificial Intelligence

- Expert Systems
- Theorem Provers
- Shakey
- Chess

Largely characterised by determinism.

# Artificial Intelligence

Modern Artificial Intelligence

- Fingerprint ID
- Internet Search
- Vision – facial ID, etc.
- Speech Recognition
- Asimo
- Jeopardy http://www.research.ibm.com/deepqa/

Statistical modeling to generalize from data.

# Natural Intelligence?

Brief Tangent

Is there a role of probability and statistics in Natural Intelligence?

# Caveats about Probability and Statistics

Black Swans and The Long Tail.

# Black Swans

In the 17th century, all observed swans were *white*.

Therefore, based on evidence, it was deemed "impossible" for a swan to be anything other than white.

# Black Swans

In the 17th century, all observed swans were *white*.

Therefore, based on evidence, it was deemed "impossible" for a swan to be anything other than white.

In the early 18th century, black swans were discovered in Western Australia.

**Black Swans** are rare, sometimes unpredictable, events that have extreme impact.

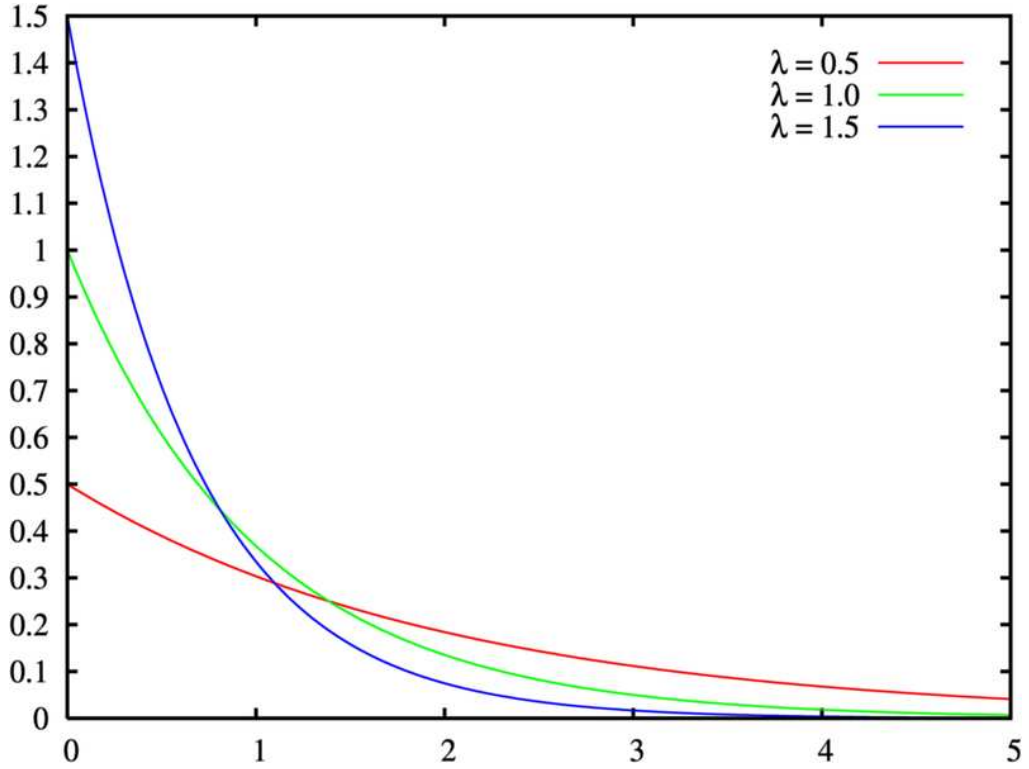
Almost all Statistical models underestimate the likelihood of unseen events.

# The Long Tail

Many events follow an exponential distribution.

These distributions typically have a very long tail. That is, a long region with relatively low probability mass.

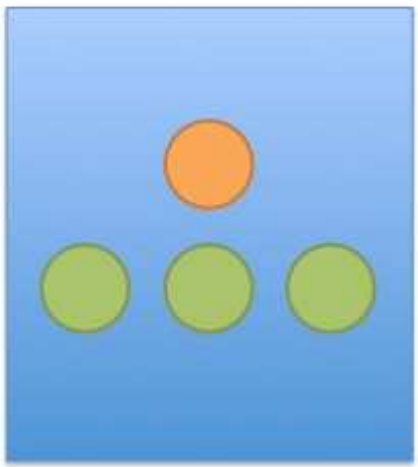
Often, interesting events occur in the Long Tail, but it's difficult to accurately model the behavior in this region of the distribution

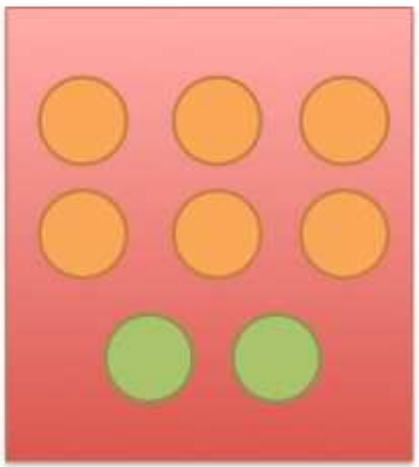


Example: Boxes and Balls.

Two boxes: 1 red, 1 blue

In the red box there are 2 apples and 6 oranges. In the blue box there are 3 apples and 1 orange.





# Boxes and Fruit

Suppose we draw from the Red box 60% of the time and the Blue Box 40% of the time.

We are equally likely to draw any piece of fruit once the box is selected.

The identity of the Box is a **random variable** $B$. The identity of the fruit is a **random variable**, $F$.

$B$ can take one of two values: $r$ (red box) or $b$ (blue box).
$F$ can take one of two values: $a$ (apple) or $o$ (orange).
We want to answer questions like "what is the total probability of picking an apple?" and "given that I chose an orange, what is the probability that it was drawn from the blue box?".

# Some basics

- The **probability** of an event is the fraction of times that an event occurs out of some number of trials, as the number of trials approaches infinity.
- Probabilities lie in the range of [0,1].
- **Mutually exclusive** events are those events that cannot simultaneously occur.
- The sum of the probabilities of all mutually exclusive events must equal 1.
- If two events are independent, $p(X, Y) = p(X)p(Y)$ and $p(X|Y) = p(X)$

# Joint Probability

**Joint probability** table of the example.

| | o | a | |
|---|---|---|---|
| blue | 1 | 3 | 4 |
| red | 6 | 2 | 8 |
| | 7 | 5 | 12 |

Let $n_{ij}$ be the number of times event $i$ and event $j$ simultaneously occur. For example, selecting an orange from the blue box.

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

# Joint Probability

A more generalized representation of a **joint probability**.

| | | | | | |
|---|---|---|---|---|---|
| | | | | | $r_j = \sum_i n_{ij}$ |
| | | | $n_{ij}$ | | |
| | | | | | |
| | $c_i = \sum_j n_{ij}$ | | | | $N = \sum_i \sum_j n_{ij}$ |

Let $n_{ij}$ be the number of times event $i$ and event $j$ simultaneously occur. For example, selecting an orange from the blue box.

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

# Marginalization

Now consider the probability of $X$ irrespective of $Y$.

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

The number of instances in column $i$ is the sum of the instances in each cell.

$$c_i = \sum_{j=1}^{L} n_{ij}$$

Therefore, we can "marginalize" or "sum over" Y:

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

# Conditional Probability

Now consider only instances where $X = x_i$. The fraction of *these* instances where $Y = y_j$ is written $p(Y = y_j|X = x_i)$.

This is a **conditional probability** – "the probability of $y$ *given* $x$".

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

Also,

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

# Sum and Product Rules

In general we will use $p(X)$ to refer to a distribution over a random variable, and $p(x_i)$ to refer to the distribution evaluated at a particular value.

Sum Rule

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

Product Rule

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

# Bayes Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

The denominator can be viewed as a normalization term:

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

# Return to Boxes and Fruit

Now we can return to the question "If an orange was chosen, what box did it come from", or define the distribution, $p(B|F = o)$.

$$
\begin{aligned}
p(B = r|F = o) &= \frac{p(F = o|B = r)p(B = r)}{p(F = o)} \\
&= \frac{\frac{3}{4}\frac{4}{10}}{\frac{9}{20}} = \frac{3}{4} \cdot \frac{4}{10} \cdot \frac{20}{9} \\
&= \frac{2}{3} \\
p(B = b|F = o) &= 1 - \frac{2}{3} = \frac{1}{3}
\end{aligned}
$$

# Interpretation of Bayes Rule

$$p(B|F) = \frac{p(F|B)p(B)}{p(F)}$$

$p(B)$ is called the **prior** of $B$. This is information we have *before* observing anything about the fruit that was drawn.

$p(B|F)$ is call the **posterior probability**, or simply the **posterior**. This is the distribution of $B$ *after* observing $F$.

In our example, the *prior* probability of $B = r$ was $\frac{4}{10}$, but the *posterior* was $\frac{2}{3}$.

The probability that the box was red *increased* after observation of $F$.

# Continuous Probabilities

So far we have been dealing with discrete probabilities, where $X$ can take one of $M$ discrete values. What if $X$ could take continuous values?

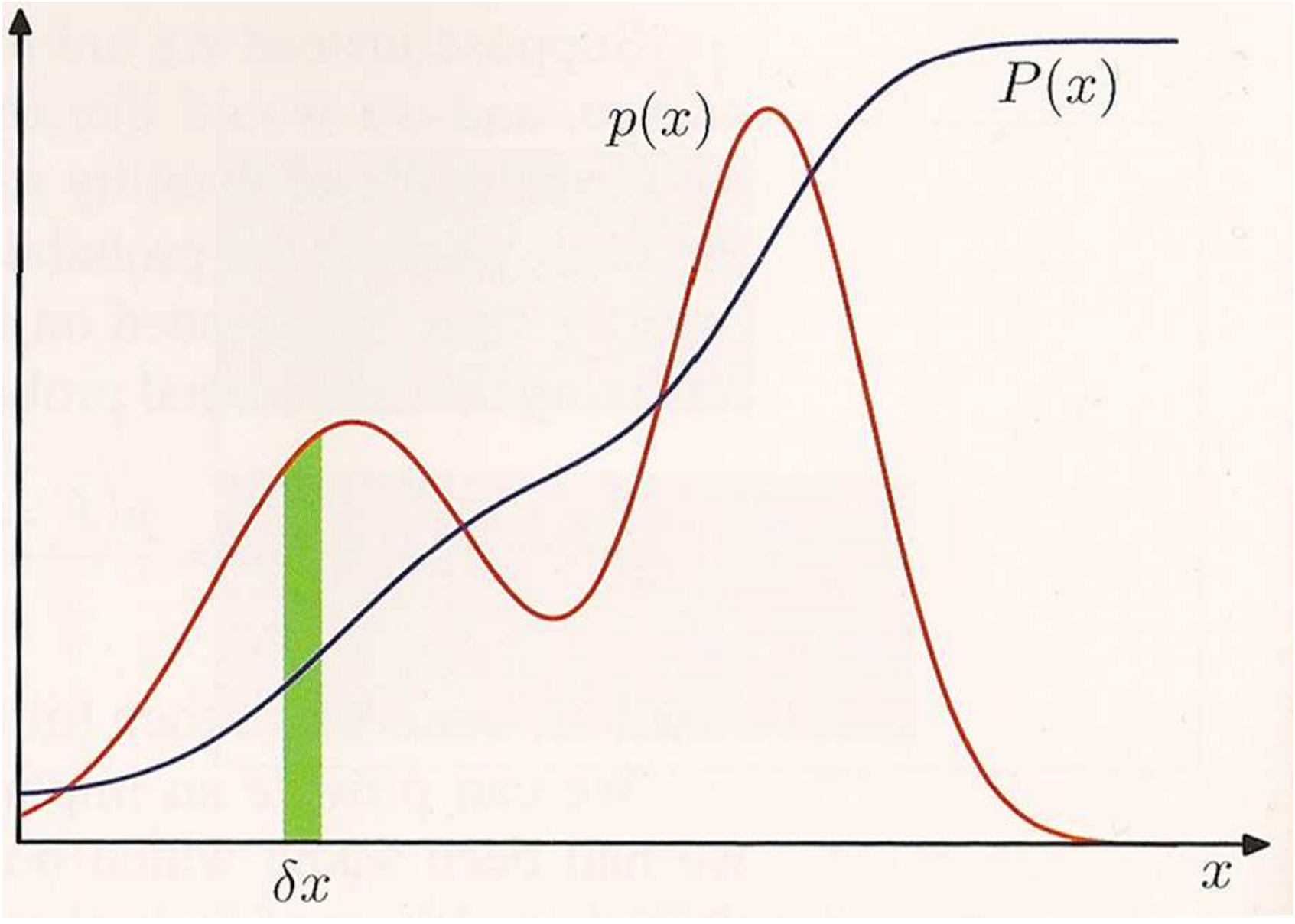
(Enter calculus.)

The probability of a real-valued random variable falling within $(x, x + \delta x)$ is $p(x)\delta x$ as $\delta x \to \infty$.

p(x) is the **probability density** or **probability density function** over $x$.
Thus the *probability* that $x$ will lie in an interval (a,b) is given by:

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

# Graphical Example of continuous probabilities.



# Continuous Probability Identities

$$p(x) \geq 1$$

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

Sum Rule

$$p(x) = \int p(x, y)dy$$

Product Rule

$$p(x, y) = p(y|x)p(x)$$

# Expected Values

Given a random variable $x$ characterized by a distribution $p(x)$, what is the *expected* value of $x$?

# Expected Values

Given a random variable $x$ characterized by a distribution $p(x)$, what is the *expected* value of $x$?

The *expectation* of $x$.

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

or

$$\mathbb{E}[x] = \int p(x)x\mathrm{d}x$$

# Expected Values Example 1

What is the expected value when rolling **one** die?

| x | p(x) |
|---|---|
| 1 | |
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |

# Expected Values Example 1

What is the expected value when rolling **one** die?

| x | p(x) |
|---|---|
| 1 | $\frac{1}{6}$ |
| 2 | $\frac{1}{6}$ |
| 3 | $\frac{1}{6}$ |
| 4 | $\frac{1}{6}$ |
| 5 | $\frac{1}{6}$ |
| 6 | $\frac{1}{6}$ |

# Distribution of Dice values

# Expected Values Example 1

$$
\begin{aligned}
\mathbb{E}[x] &= \sum_x p(x)x \\
&= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\
&= \frac{21}{6} = 3.5
\end{aligned}
$$

# Expected Values Example 1

$$
\begin{aligned}
\mathbb{E}[x] &= \sum_x p(x)x \\
&= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\
&= \frac{21}{6} = 3.5 \\
\mathbb{E}[x] &= \frac{1}{N} \sum_i^N x_i
\end{aligned}
$$

# Expected Values Example 2

What is the expected value when rolling **two** dice?

| x | p(x) |
|---|---|
| 2 | |
| 3 | |
| 4 | |
| 5 | |
| 6 | |
| 7 | |
| 8 | |
| 9 | |
| 10 | |
| 11 | |
| 12 | |

# Expected Values Example 2

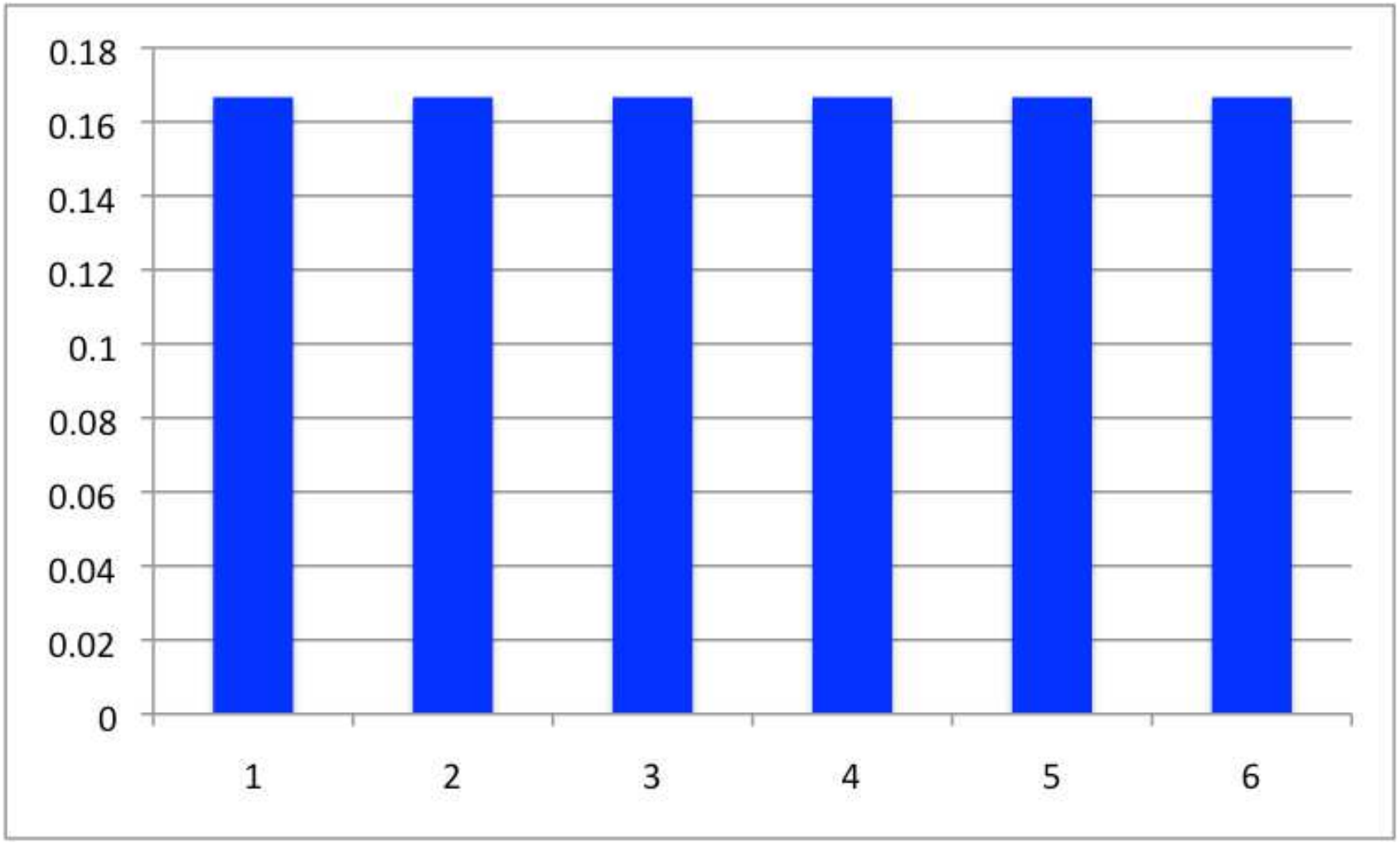
What is the expected value when rolling **two** dice?

| x | p(x) |
|---|---|
| 2 | $\frac{1}{36}$ |
| 3 | $\frac{2}{36}$ |
| 4 | $\frac{3}{36}$ |
| 5 | $\frac{4}{36}$ |
| 6 | $\frac{5}{36}$ |
| 7 | $\frac{6}{36}$ |
| 8 | $\frac{5}{36}$ |
| 9 | $\frac{4}{36}$ |
| 10 | $\frac{3}{36}$ |
| 11 | $\frac{2}{36}$ |
| 12 | $\frac{1}{36}$ |

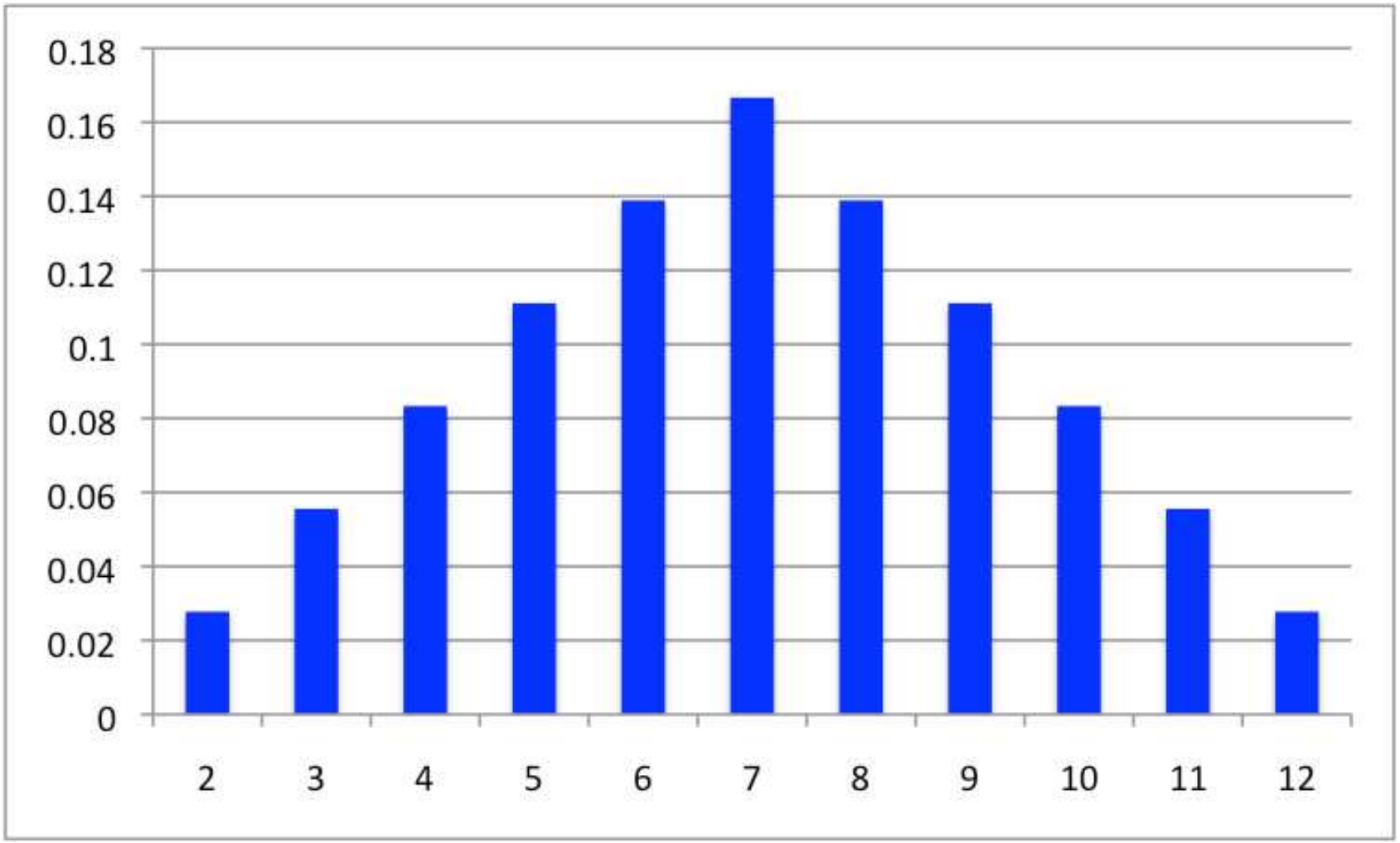
# Expected Values Example 2

$$
\begin{aligned}
\mathbb{E}[x] &= \sum_x p(x)x \\
&= 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} \\
&\quad +8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} \\
&= \frac{252}{36} = 7
\end{aligned}
$$

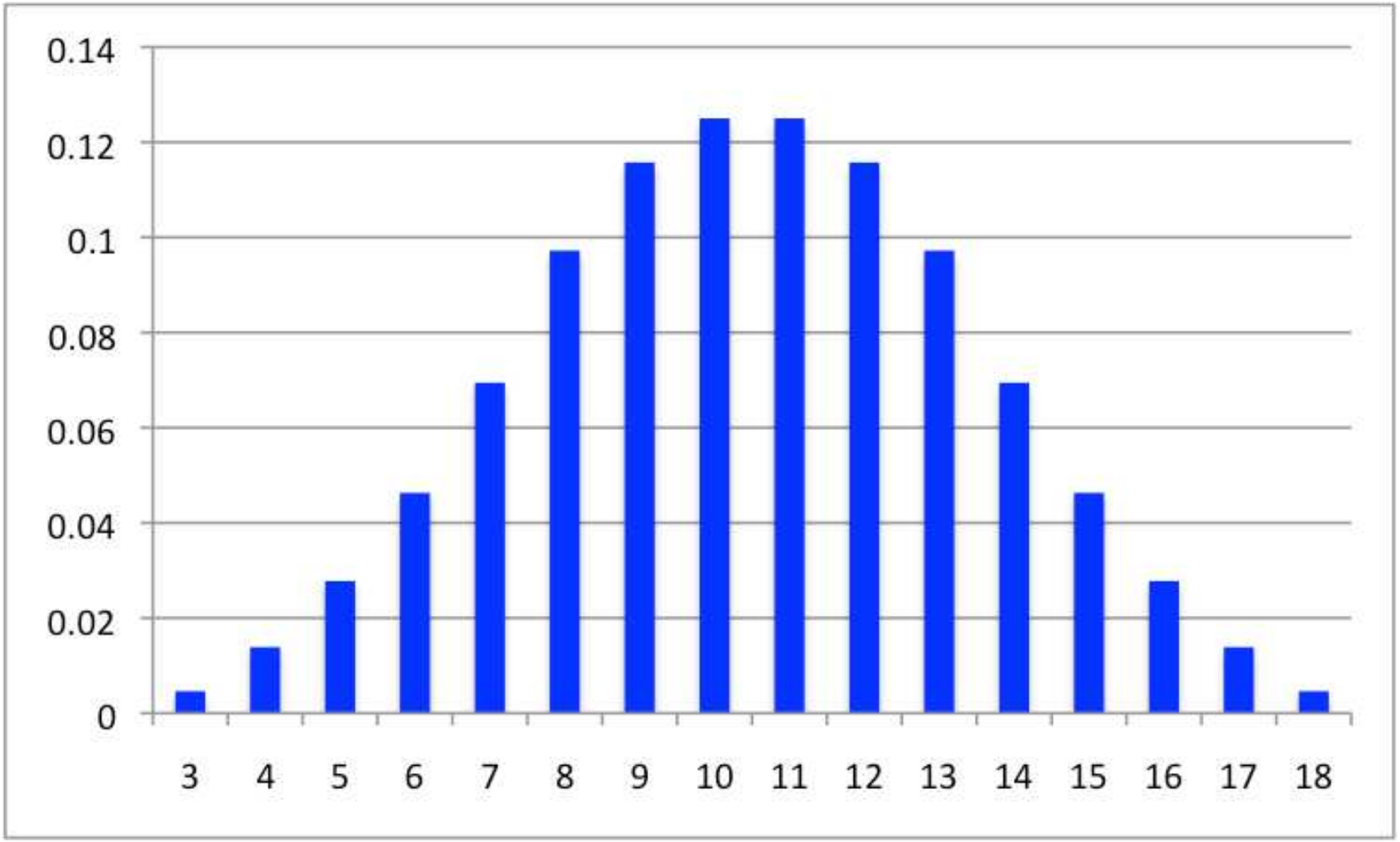
# Distribution of Dice values



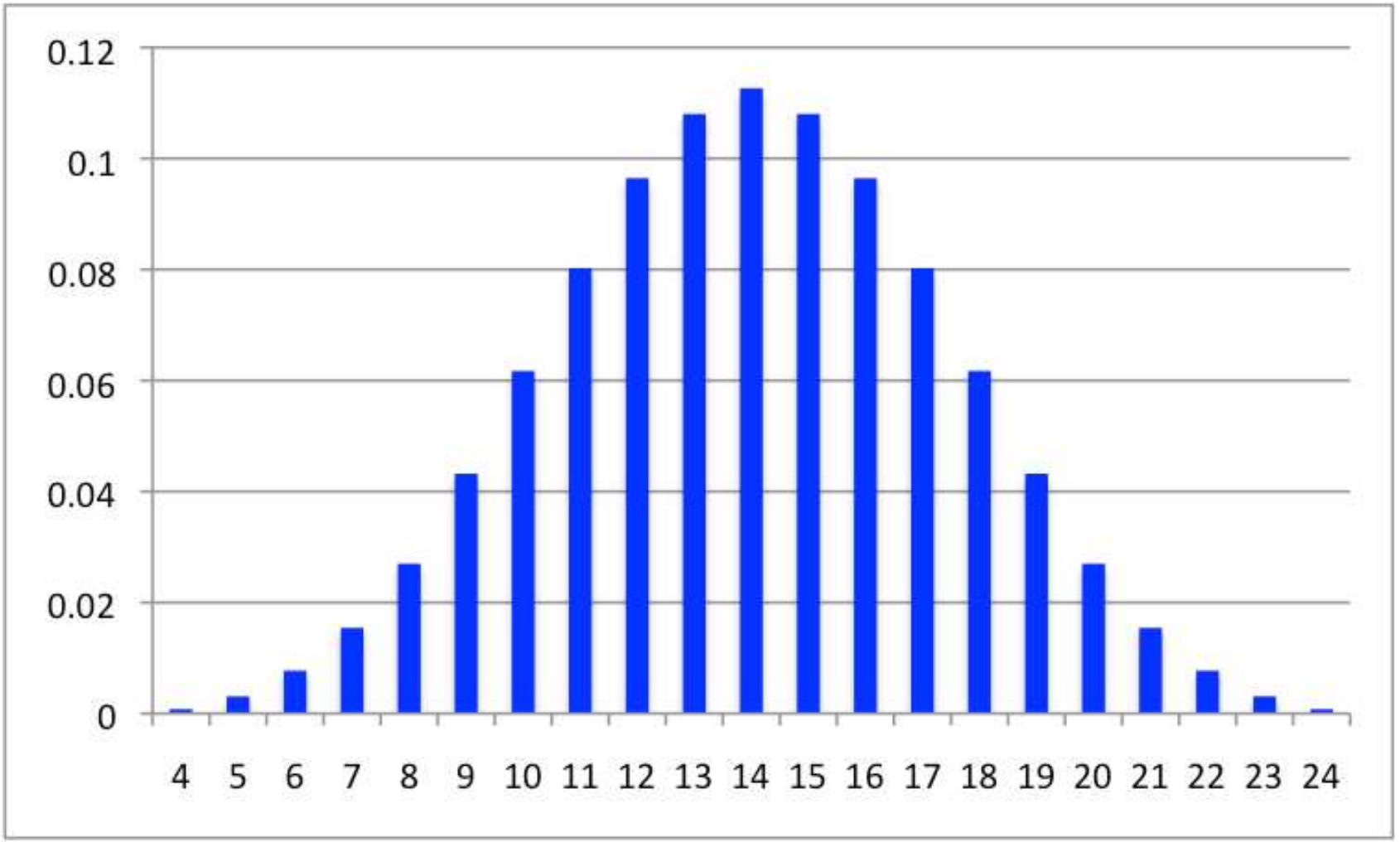
# Distribution of Dice values



# Distribution of Dice values



# Distribution of Dice values



## Multinomial Distribution

If a variable, $x$, can take 1-of-$K$ states, we can represent this variable as being drawn from a **multinomial distribution**.

We say the probability of $x$ being a member of state $k$ is $\mu_k$, elements of a vector $\boldsymbol{\mu}$.

$$\sum_{k=1}^{K} \mu_k = 1$$

$$p(x|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

# Expected Value of a Multinomial Distribution

Expectation

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{x} p(x|\boldsymbol{\mu})x = (\mu_0, \mu_1, \ldots, \mu_{K-1})^T$$

# Gaussian Distribution

As the number of dice increases, the multinomial distribution approaches a **Gaussian Distribution**, or **Normal Distribution**.

One dimensional

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

D-dimensional

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

# Gaussian Example

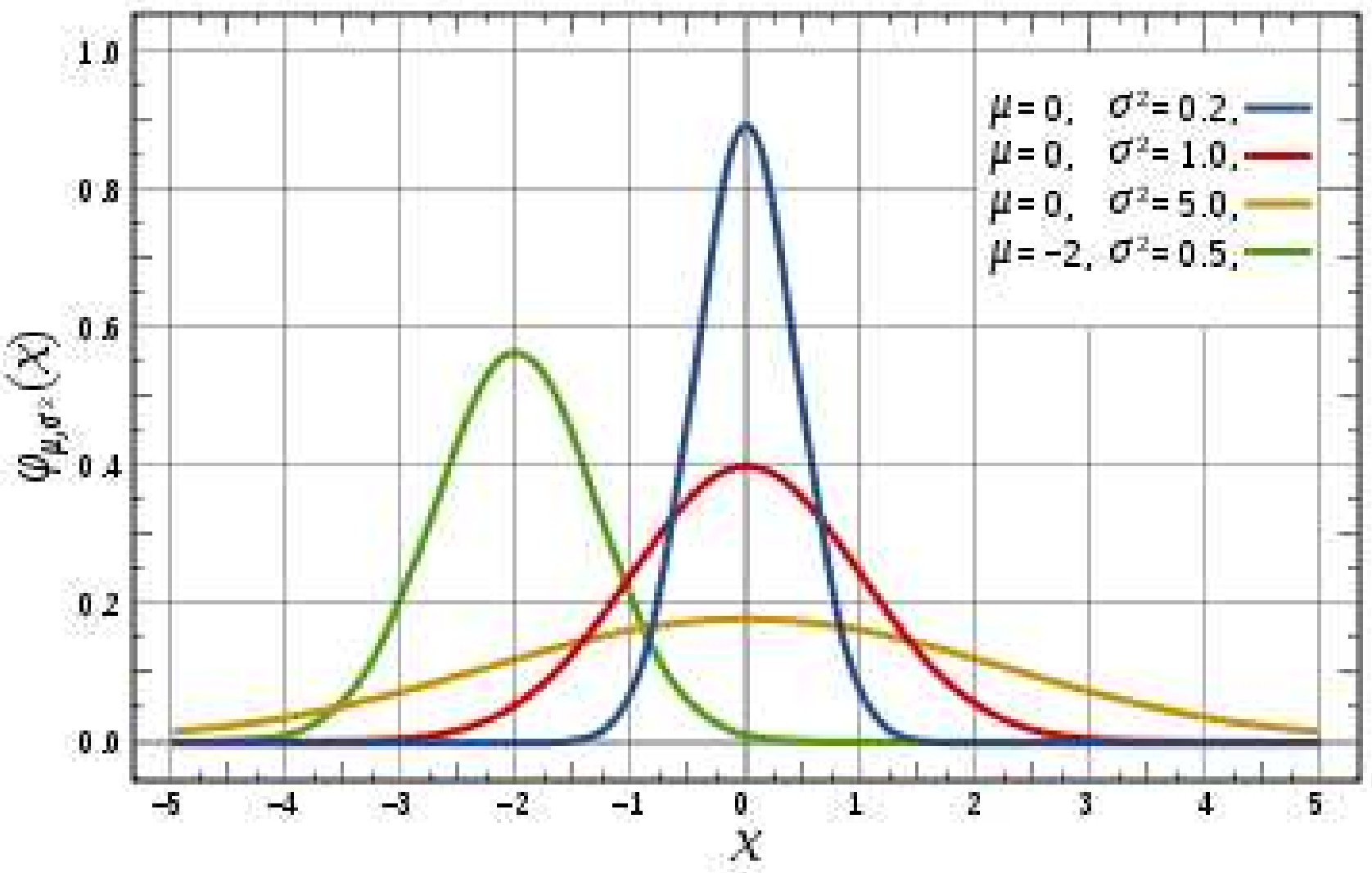



Image from wikipedia.

# Gaussian Example

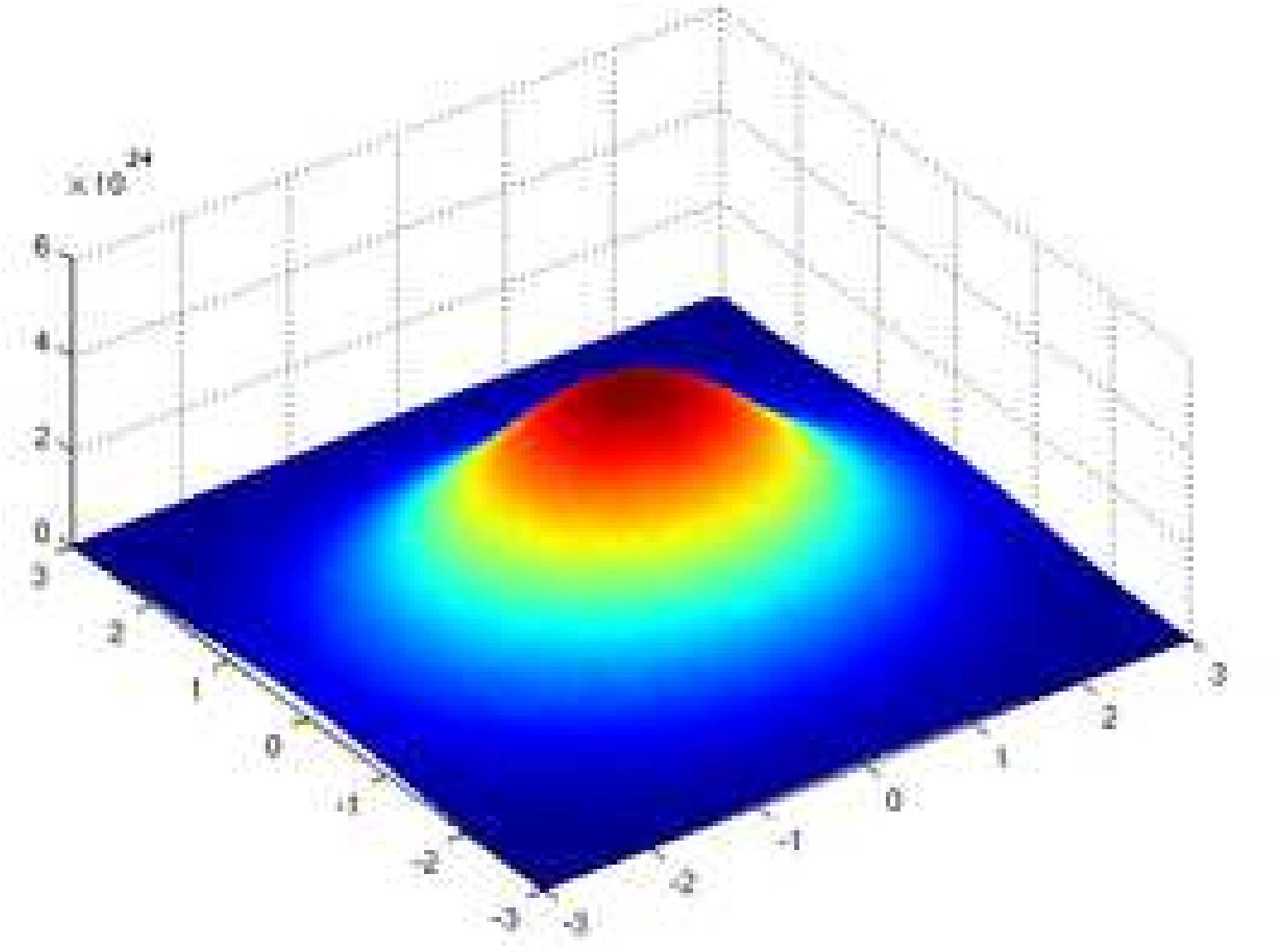



Image from wikipedia.

# Expectation of a Gaussian

$$\begin{aligned}\mathbb{E}[x|\mu,\sigma^2] &= \int N(x|\mu,\sigma^2)x dx \\ &= \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} x dx\end{aligned}$$

or

$$\begin{aligned}\mathbb{E}[\mathbf{x}|\boldsymbol{\mu},\Sigma] &= \int N(\mathbf{x}|\boldsymbol{\mu},\Sigma)\mathbf{x} dx \\ &= \int \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \mathbf{x} dx\end{aligned}$$

We'll need some calculus for this, so next time.

# Variances

The *variance* of $x$ describes how much variability around the mean, $\mathbb{E}[x]$.

$$\text{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

# Covariance

The *covariance* of two random variables, $x$ and $y$, expresses to what extent the two vary together.

$$\begin{aligned} \text{cov}[x, y] &= \mathbb{E}_{x,y}[(x - \mathbb{E}(x))(y - \mathbb{E}[y])] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

If two variables are *independent* their covariance equals zero. (Know how to prove this.)

# How does Machine Learning use Probabilities

- The *expectation* of a function is the guess.
- The *covariance* is the confidence in this guess.

These are simple operations. . .

But how can we find the best estimate of $p(x)$?

# Bye

- Next
  - Linear Algebra
  - Vector Calculus