Math Review

Jordan Boyd-Graber

University of Maryland

Functions

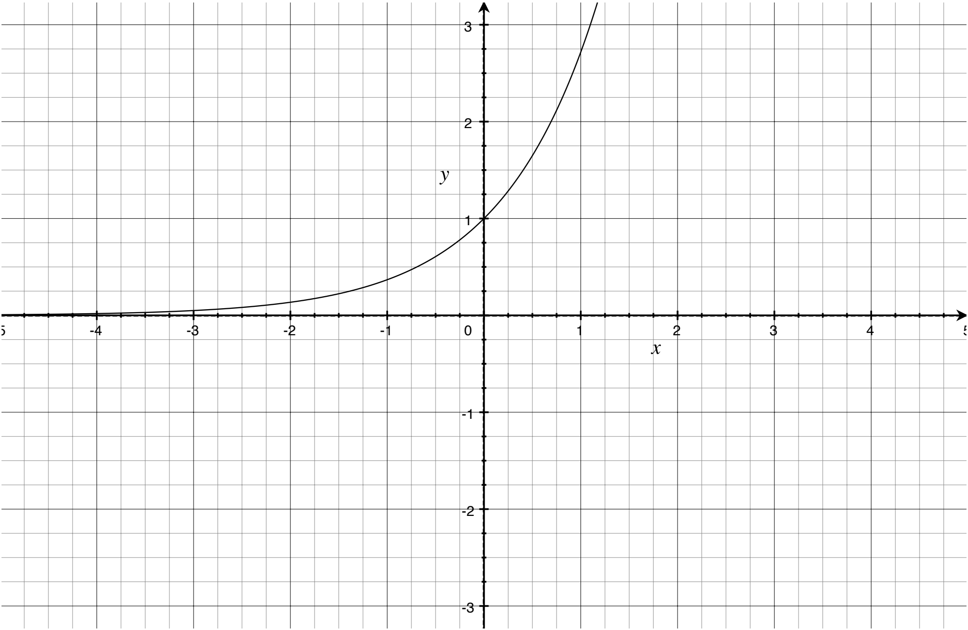
Function Notation

- Take a number a double it
- Mathematical notation

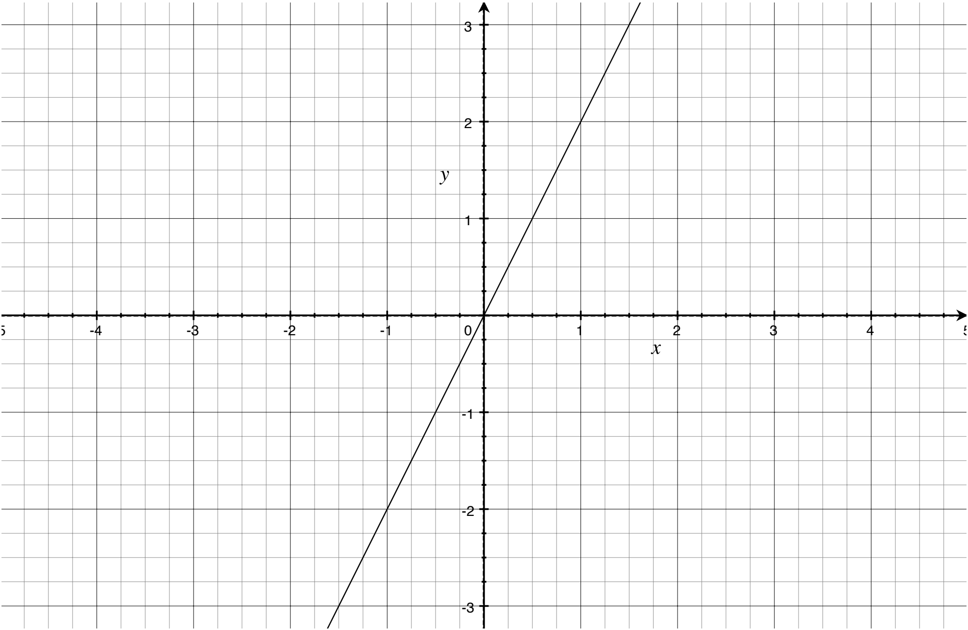
$$f(x) = 2x \tag{1}$$

- Python notation

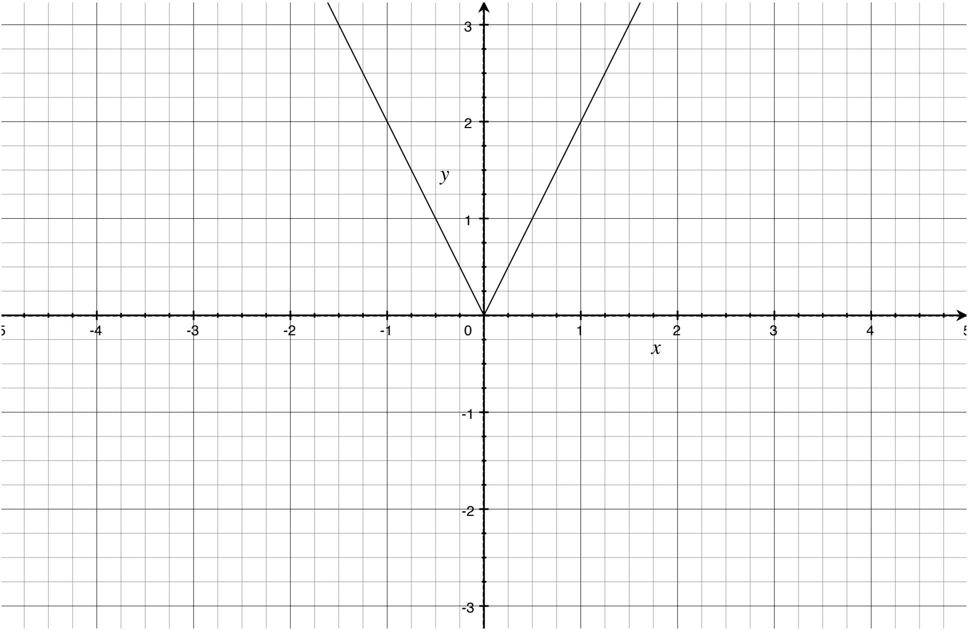
```
def double(x):
    return 2 * x
```



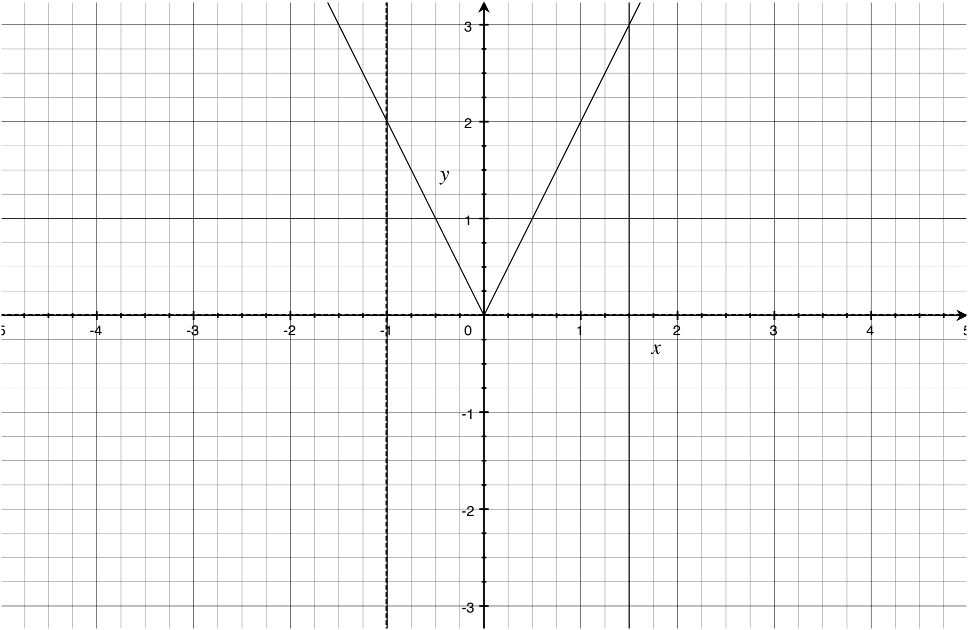
$$f(x) = \exp(x)$$



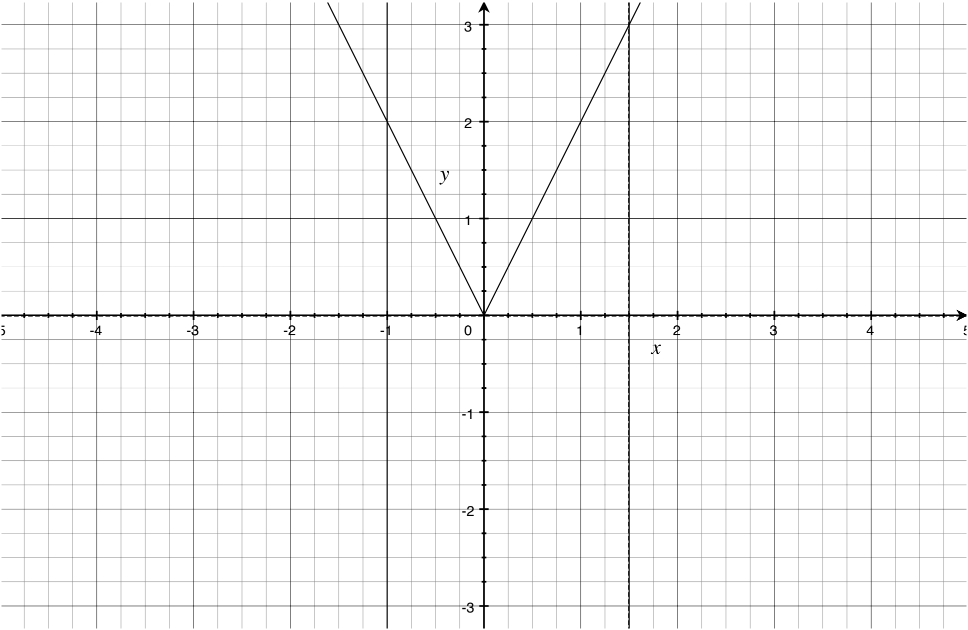
$$f(x) = 2x$$



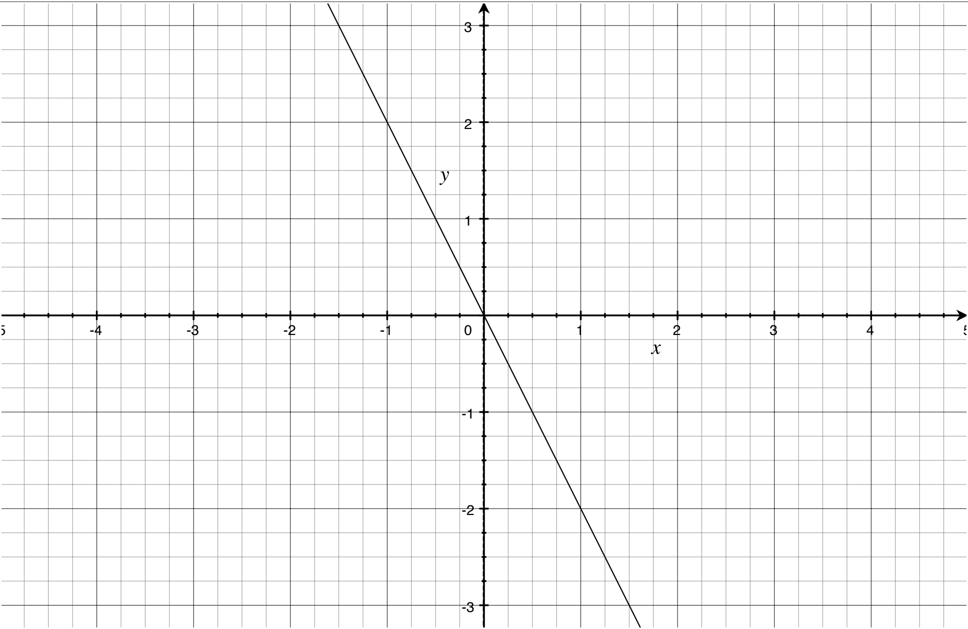
$$f(x) = |x|$$



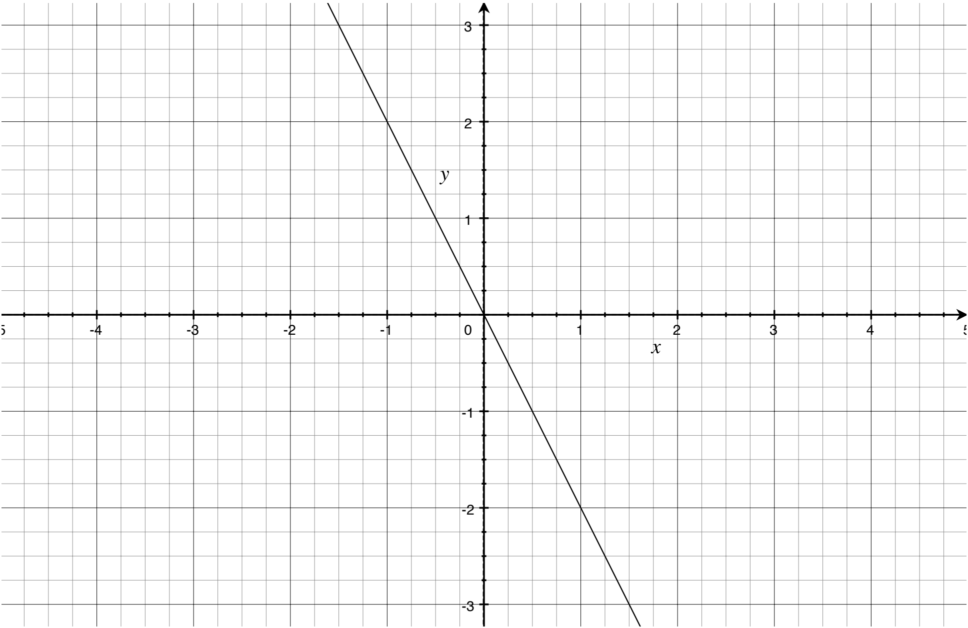
$$f(x) = 2x$$



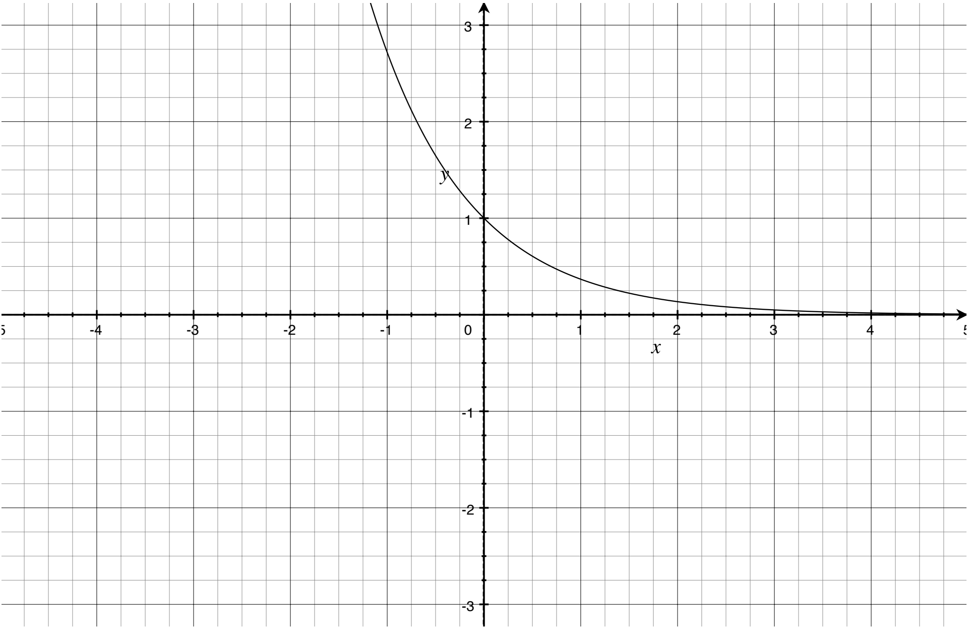
$$f(x) = 2x$$



3
2
y
1
-4
-3
-2
-1
0
1
2
3
4
5
x
-1
-2
-3



$$f(x) = 2(-x)$$



$$f(x) = \exp(-x)$$

Combining functions

$$f(x) = \exp(-x) \tag{2}$$

$$g(x) = 1 + x \tag{3}$$

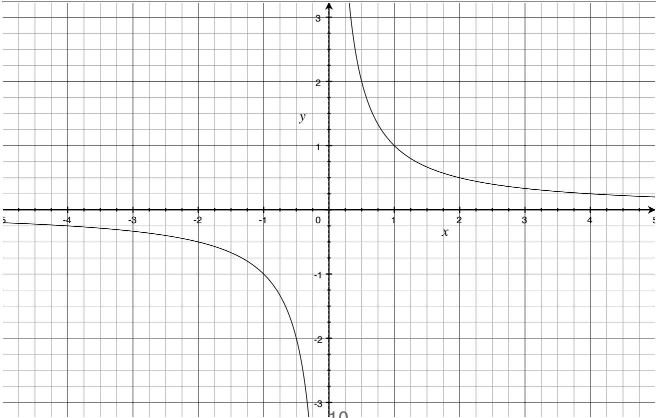
$$h(x) = \frac{1}{x} \tag{4}$$

Combining functions

$$f(x) = \exp(-x) \tag{2}$$

$$g(x) = 1 + x \tag{3}$$

$$h(x) = \frac{1}{x} \tag{4}$$



Combining functions

$$f(x) = \exp(-x) \tag{2}$$

$$g(x) = 1 + x \tag{3}$$

$$h(x) = \frac{1}{x} \tag{4}$$

$$l(x) = g(f(x)) = \frac{1}{\exp(-x)} \tag{5}$$

Combining functions

$$f(x) = \exp(-x) \tag{2}$$
$$g(x) = 1 + x \tag{3}$$
$$h(x) = \frac{1}{x} \tag{4}$$

$$l(x) = g(f(x)) = \frac{1}{\exp(-x)} \tag{5}$$

```
from math import exp

def neg_exp(x):
  return exp(-x)
def composition(x):
  return 1.0 / neg_exp(x)
```

Properties of the Exponential (and log) Function

$$\exp(a+b) = exp(a)exp(b) \tag{6}$$

$$\log(a+b) = \log(a)\log(b) \tag{7}$$

$$\log(a^b) = b \cdot \log(a) \tag{8}$$

Composition didn't do as much as we thought!

$$
\begin{aligned}
I(x) &= g(f(x)) && (9)\\
&= \frac{1}{\exp(-x)} && (10)\\
&= \frac{1}{\exp(x)^{-1}} && (11)\\
&= \frac{1}{\frac{1}{\exp(x)}} && (12)\\
&= \exp x && (13)
\end{aligned}
$$

Logistic Function

$$f(x) = \exp(-x) \tag{14}$$

$$g(x) = 1 + x \tag{15}$$

$$h(x) = \frac{1}{x} \tag{16}$$

Putting them together:

$$\tag{17}$$

Logistic Function

$$f(x) = \exp(-x) \tag{14}$$

$$g(x) = 1 + x \tag{15}$$

$$h(x) = \frac{1}{x} \tag{16}$$

Putting them together:

$$l(x) = h(g(f(x))) \tag{17}$$

$$\tag{18}$$

Logistic Function

$$f(x) = \exp(-x) \tag{14}$$
$$g(x) = 1 + x \tag{15}$$
$$h(x) = \frac{1}{x} \tag{16}$$

Putting them together:

$$
\begin{aligned}
l(x) &= h(g(f(x))) && (17)\\
&= h(g(\exp(-x))) && (18)\\
& && (19)
\end{aligned}
$$

Logistic Function

$$f(x) = \exp(-x) \tag{14}$$
$$g(x) = 1 + x \tag{15}$$
$$h(x) = \frac{1}{x} \tag{16}$$

Putting them together:

$$\begin{aligned} l(x) &= h(g(f(x))) && (17) \\ &= h(g(\exp(-x))) && (18) \\ &= h(1 + \exp(-x)) && (19) \\ & && (20) \end{aligned}$$

Logistic Function

$$f(x) = \exp(-x) \tag{14}$$
$$g(x) = 1 + x \tag{15}$$
$$h(x) = \frac{1}{x} \tag{16}$$

Putting them together:

$$
\begin{aligned}
l(x) &= h(g(f(x))) && (17)\\
&= h(g(\exp(-x))) && (18)\\
&= h(1 + \exp(-x)) && (19)\\
&= \frac{1}{1 + \exp(-x)} && (20)
\end{aligned}
$$

Courses, Lectures, Exercises and More

http://boydgraber.org

Math Review

Jordan Boyd-Graber

University of Maryland

Functions

Engineering rationale behind probabilities

- Encoding uncertainty
  - Data are variables
  - We don't always know the values of variables
  - Probabilities let us reason about variables even when we are uncertain

Engineering rationale behind probabilities

- Encoding uncertainty
  - Data are variables
  - We don't always know the values of variables
  - Probabilities let us reason about variables even when we are uncertain
- Encoding confidence
  - The flip side of uncertainty
  - Useful for decision making: should we trust our conclusion?
  - We can construct probabilistic models to boost our confidence
    - E.g., combining polls

Random variable

- Random variables take on values in a *sample space*.
- They can be *discrete* or *continuous*:
  - Coin flip: $\{H, T\}$
  - Height: positive real values $(0, \infty)$
  - Temperature: real values $(-\infty, \infty)$
  - Number of words in a document: Positive integers $\{1, 2, \ldots\}$
- We call the outcomes *events*.
- Denote the random variable with a capital letter; denote a realization of the random variable with a lower case letter.
  - E.g., X is a coin flip, x is the value (H or T) of that coin flip.

Discrete distribution

- A discrete distribution assigns a probability to every event in the sample space
- For example, if X is a coin, then

$$\begin{aligned} P(X = H) &= 0.5 \\ P(X = T) &= 0.5 \end{aligned}$$

- And probabilities have to be greater than or equal to 0
- The probabilities over the entire space must sum to one

Discrete distribution

- A discrete distribution assigns a probability to every event in the sample space
- For example, if X is a coin, then

$$\begin{aligned} P(X = H) &= 0.5 \\ P(X = T) &= 0.5 \end{aligned}$$

- And probabilities have to be greater than or equal to 0
- The probabilities over the entire space must sum to one

Discrete distribution

- A discrete distribution assigns a probability to every event in the sample space
- For example, if X is a coin, then

$$
\begin{aligned}
P(X=H) &= 0.5 \\
P(X=T) &= 0.5
\end{aligned}
$$

- And probabilities have to be greater than or equal to 0
- The probabilities over the entire space must sum to one

$$\sum P(X=x)=1$$

Discrete distribution

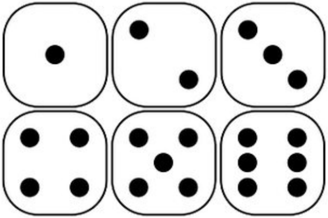
- A discrete distribution assigns a probability to every event in the sample space
- For example, if X is a coin, then

$$\begin{aligned} P(X=H) &= 0.5 \\ P(X=T) &= 0.5 \end{aligned}$$

- And probabilities have to be greater than or equal to 0
- The probabilities over the entire space must sum to one

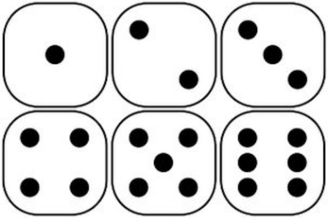
$$\sum_{x} P(X=x) = 1$$

A Fair Die



1	2	3	4	5	6
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

A Fair Die



1	2	3	4	5	6
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

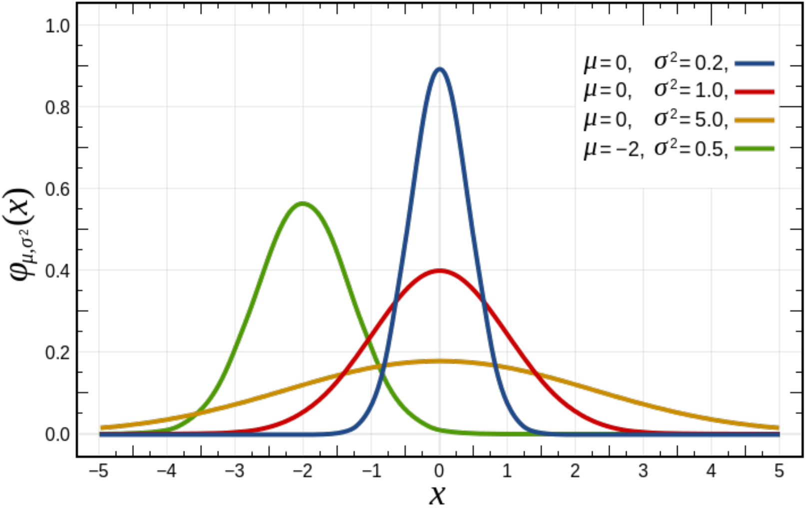
```
def die_prob(x):
  if x in [0, 1, 2, 3, 4, 5, 6]:
    return 1.0 / 6.0
  else:
    return 0.0
```

The normal distribution

- The most common continuous distribution is the normal distribution, also called the Gaussian distribution.
- The density is defined by two parameters:
  - μ: the mean of the distribution
  - σ^2: the variance of the distribution (σ is the standard deviation)
- The normal density has a “bell curve” shape and naturally occurs in many problems.

Carl Friedrich Gauss
1777 – 1855

The normal distribution



The normal distribution

- The probability density of the normal distribution is:

$$f(x) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{Does not depend on } x} \underbrace{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}_{\text{Largest when } x=\mu;\ \text{shrinks as } x \text{ moves away from } \mu}$$

- Notation: $\exp(x) = e^x$
- If X follows a normal distribution, then $\mathbb{E}[X] = \mu$.
- The normal distribution is symmetric around μ.

The normal distribution

- The probability density of the normal distribution is:

$$f(x) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\substack{\text{Does not} \\ \text{depend on } x}} \underbrace{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}_{\substack{\text{Largest when } x=\mu; \\ \text{shrinks as } x \text{ moves} \\ \text{away from } \mu}}$$

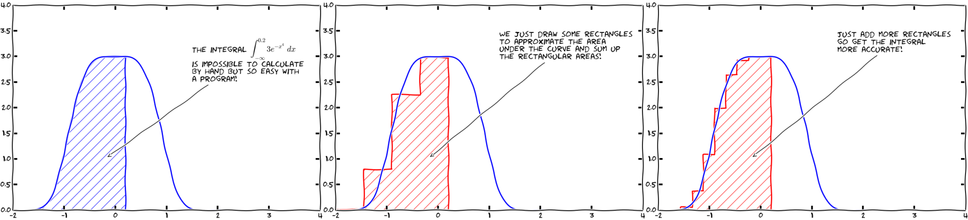
- Notation: $\exp(x) = e^x$
- If X follows a normal distribution, then $\mathbb{E}[X] = \mu$.
- The normal distribution is symmetric around μ.

The normal distribution

- The probability density of the normal distribution is:

$$f(x) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\substack{\text{Does not} \\ \text{depend on } x}} \underbrace{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}_{\substack{\text{Largest when } x=\mu; \\ \text{shrinks as } x \text{ moves} \\ \text{away from } \mu}}$$

- Notation: $\exp(x) = e^x$
- If X follows a normal distribution, then $\mathbb{E}[X] = \mu$.
- The normal distribution is symmetric around μ.



From Svein Linge and Hans Petter Langtangen

The normal distribution

- What is the probability that a value sampled from a normal distribution will be within *n* standard deviations from the mean?
- $P(\mu - n\sigma \leq X \leq \mu + n\sigma) = ?$

The normal distribution

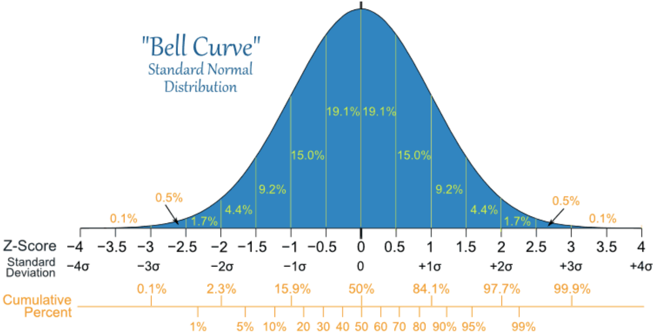
- What is the probability that a value sampled from a normal distribution will be within n standard deviations from the mean?
- $P(\mu - n\sigma \leq X \leq \mu + n\sigma) = ?$
$= \int_{x=\mu-n\sigma}^{\mu+n\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{x=\mu-n\sigma}^{\mu+n\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

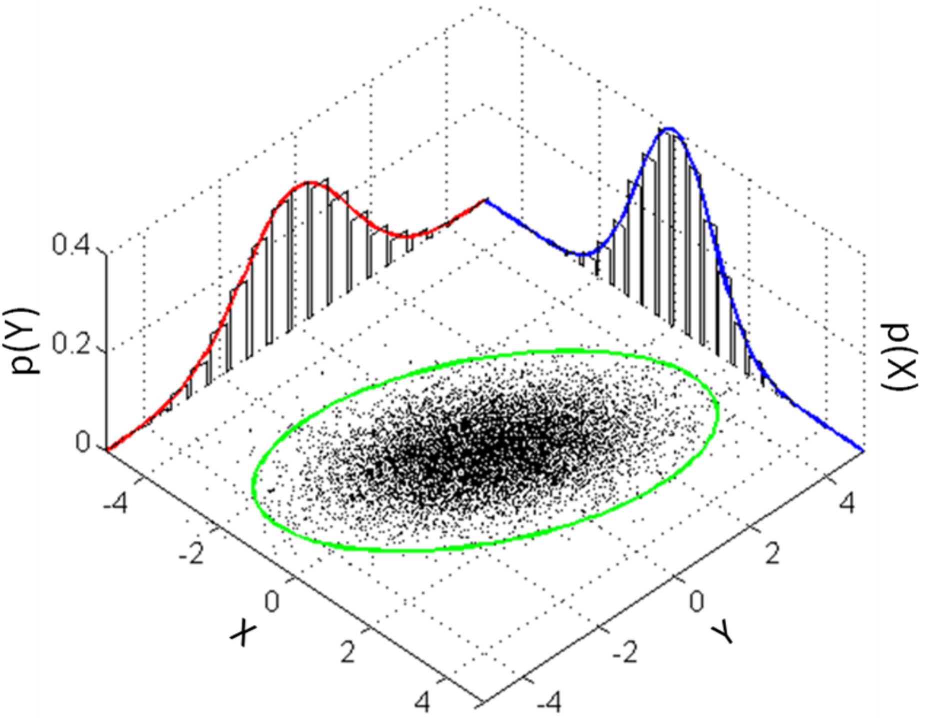
The normal distribution

- What is the probability that a value sampled from a normal distribution will be within n standard deviations from the mean?
- $P(\mu - n\sigma \leq X \leq \mu + n\sigma) = ?$
$= \int_{x=\mu-n\sigma}^{\mu+n\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{x=\mu-n\sigma}^{\mu+n\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

```
>>> from scipy.stats import norm
>>> norm.cdf(1.0) - norm.cdf(-1.0)
0.6826894921370859
```

The normal distribution





0.4
0.2
0
p(Y)
p(X)
-4
-2
0
2
4
X
4
2
0
-2
-4
Y

Courses, Lectures, Exercises and More

http://boydgraber.org

Math Review

Jordan Boyd-Graber

University of Maryland

Functions

Vectors

Row Vector

$$\vec{v} = \begin{bmatrix} 5 & 8 \end{bmatrix} \tag{21}$$

Column Vector

$$\begin{bmatrix} 5 \\ 2 \end{bmatrix} \tag{22}$$

Indexing elements

$v_1 = 5$; $v_2 = 8$

Vector Addition

$$\begin{bmatrix} 5 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ 7 \end{bmatrix} = \begin{bmatrix} 5+3 \\ 2+7 \end{bmatrix} = \begin{bmatrix} 8 \\ 9 \end{bmatrix} \tag{23}$$

Scalar Multiplication

$$3 \cdot \begin{bmatrix} 5 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \cdot 5 \\ 3 \cdot 2 \end{bmatrix} = \begin{bmatrix} 15 \\ 6 \end{bmatrix} \tag{24}$$

Dot Product Example

$$\begin{bmatrix} 4 \\ 3 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 5 \\ 2 \end{bmatrix} =$$

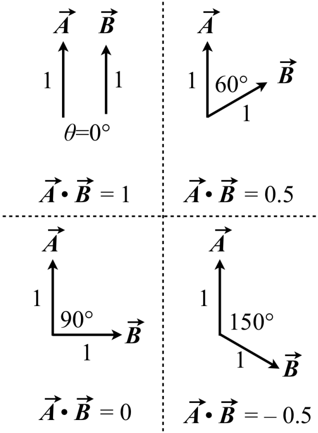
Dot Product Example

$$\begin{bmatrix} 4 \\ 3 \end{bmatrix}^\top \cdot \begin{bmatrix} 5 \\ 2 \end{bmatrix} = 4 \cdot 5 + 3 \cdot 2 = 26 \tag{25}$$

Dot Product Definition

$$\vec{x} \cdot \vec{y} = \sum_{i}^{D} x_i y_i \tag{26}$$

$$\vec{x} \cdot \vec{y} = |\vec{x}||\vec{y}| \cos \theta \tag{27}$$



From Scott Hill

Courses, Lectures, Exercises and More

http://boydgraber.org

Math Review

Jordan Boyd-Graber

University of Maryland

Functions

Dot Product Example

$$\begin{bmatrix}4\\3\end{bmatrix}^\top \cdot \begin{bmatrix}5\\2\end{bmatrix} = \tag{28}$$

$$\begin{bmatrix}4 & 3\end{bmatrix} \cdot \begin{bmatrix}5\\2\end{bmatrix} = \tag{29}$$

$$\begin{bmatrix}4 \cdot 5 + 2 \cdot 3\end{bmatrix} = \begin{bmatrix}26\end{bmatrix} \tag{30}$$

Dot Product Example

$$\begin{bmatrix} 4 \\ 3 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 5 \\ 2 \end{bmatrix} =$$

$$\begin{bmatrix} 26 \end{bmatrix} \tag{30}$$

Dot Product Example

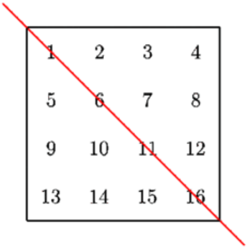
$$\begin{bmatrix}4\\3\end{bmatrix}^\top \cdot \begin{bmatrix}5\\2\end{bmatrix} = \tag{28}$$

$$\begin{bmatrix}4 & 3\end{bmatrix} \cdot \begin{bmatrix}5\\2\end{bmatrix} = \tag{29}$$

$$\begin{bmatrix}4\cdot 5 + 2\cdot 3\end{bmatrix} = \begin{bmatrix}26\end{bmatrix} \tag{30}$$

Transpose

- Turns n by m matrix into m by n matrix
- Swaps element in a_{ij} with element in a_{ji}

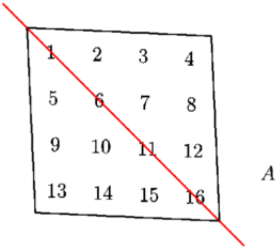


A

From Michael Doob

Transpose

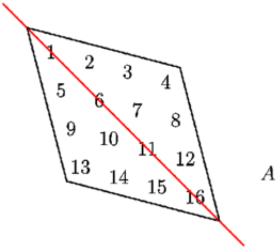
- Turns n by m matrix into m by n matrix
- Swaps element in a_{ij} with element in a_{ji}



From Michael Doob

Transpose

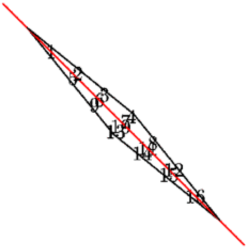
- Turns n by m matrix into m by n matrix
- Swaps element in a_{ij} with element in a_{ji}



From Michael Doob

Transpose

- Turns n by m matrix into m by n matrix
- Swaps element in a_{ij} with element in a_{ji}

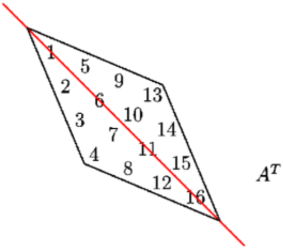


A

From Michael Doob

Transpose

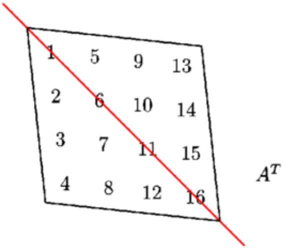
- Turns n by m matrix into m by n matrix
- Swaps element in a_{ij} with element in a_{ji}



From Michael Doob

Transpose

- Turns n by m matrix into m by n matrix
- Swaps element in a_{ij} with element in a_{ji}



From Michael Doob

Transpose

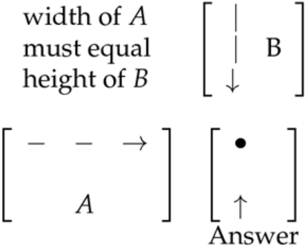
- Turns n by m matrix into m by n matrix
- Swaps element in a_{ij} with element in a_{ji}

$$\begin{bmatrix} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{bmatrix} \quad A^T$$

From Michael Doob

Matrix Multiplication Rules

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot & 0 \\ \cdot & 3 \\ \cdot & 0 \\ \cdot & 2 \end{bmatrix} = \begin{bmatrix} \cdot & 14 \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix}$$



From Denis Auroux

Matrix Multiplication with Identity

General Formula

$$a_{ij} = \sum_k l_{ik} r_{kj} \tag{31}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} \end{bmatrix} \tag{32}$$

Matrix Multiplication with Identity

General Formula

$$a_{ij} = \sum_k l_{ik} r_{kj} \tag{31}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} ? \\ \end{bmatrix} \tag{32}$$

$a_{11} = l_{11} r_{11} + l_{12} r_{21} = 3 + 0 = 3$

Matrix Multiplication with Identity

General Formula

$$a_{ij} = \sum_k l_{ik} r_{kj} \tag{31}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ \end{bmatrix} \tag{32}$$

Matrix Multiplication with Identity

General Formula

$$a_{ij} = \sum_k l_{ik} r_{kj} \tag{31}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ ? \end{bmatrix} \tag{32}$$

$a_{21} = l_{21} r_{11} + l_{22} r_{21} = 0 + 4 = 4$

Matrix Multiplication with Identity

General Formula

$$a_{ij} = \sum_k l_{ik} r_{kj} \tag{31}$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \tag{32}$$

Selecting a Row

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^\top \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} ? & \quad \end{bmatrix} \tag{33}$$

Selecting a Row

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} ? & ? \end{bmatrix} \tag{33}$$

Selecting a Row

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} & ? \end{bmatrix} \tag{33}$$

Selecting a Row

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} 6 & ? \end{bmatrix} \tag{33}$$

Selecting a Row

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} 6 & ? \end{bmatrix} \tag{33}$$

Selecting a Row

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} 6 & 7 \end{bmatrix} \tag{33}$$

Selecting Rows

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} ? & \ \end{bmatrix} \tag{34}$$

Selecting Rows

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} ? & ? \end{bmatrix} \tag{34}$$

Selecting Rows

$$\begin{bmatrix}1\\1\\0\\0\\1\end{bmatrix}^{\top} \cdot \begin{bmatrix}9 & 7\\0 & 8\\6 & 7\\5 & 3\\0 & 9\end{bmatrix} = \begin{bmatrix} & ?\end{bmatrix} \tag{34}$$

Selecting Rows

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}^\top \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} 9+0+0 & ? \end{bmatrix} \tag{34}$$

Selecting Rows

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} 9 & ? \end{bmatrix} \tag{34}$$

Selecting Rows

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}^{\top} \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} 9 & 7+8+9 \end{bmatrix} \tag{34}$$

Selecting Rows

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}^\top \cdot \begin{bmatrix} 9 & 7 \\ 0 & 8 \\ 6 & 7 \\ 5 & 3 \\ 0 & 9 \end{bmatrix} = \begin{bmatrix} 9 & 24 \end{bmatrix} \tag{34}$$

Math Review

Slides adapted from Dave Blei and Lauren Hannah

University of Maryland

Expectations and Entropy

Expectation

An *expectation* of a random variable is a weighted average:

$$\begin{aligned} \mathrm{E}[f(X)] &= \sum_x f(x)\, p(x) && \text{(discrete)} \\ &= \int_{-\infty}^{\infty} f(x)\, p(x)\, dx && \text{(continuous)} \end{aligned}$$

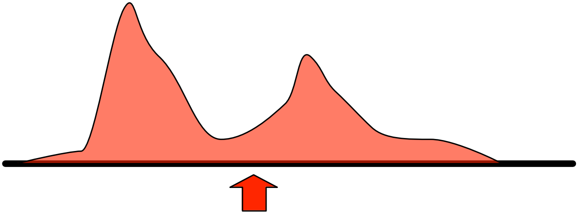
Expectation

Expectations of constants or known values:

- $\mathrm{E}[a] = a$

Expectation Intuition

- $\mathrm{E}[x]$ is most common expectation
- Average outcome (might not be an event: 2.4 children)
- Center of mass



Expectation of die / dice

What is the expectation of the roll of die?

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} =$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

What is the expectation of the sum of two dice?

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

What is the expectation of the sum of two dice?

Two die

$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} =$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

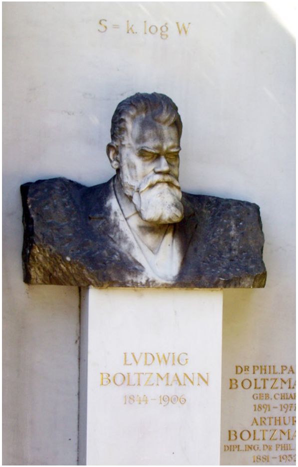
What is the expectation of the sum of two dice?

Two die

$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7$

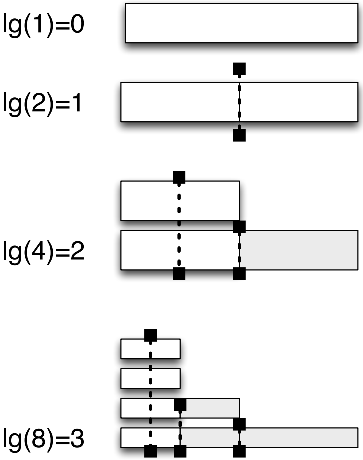
Entropy

- Measure of disorder in a system
- In the real world, entroy in a system tends to increase
- Can also be applied to probabilities:
  - Is one (or a few) outcomes certain (low entropy)
  - Are things equiprobable (high entropy)
- In data science
  - We look for features that allow us to reduce entropy (decision trees)
  - All else being equal, we seek models that have maximum entropy (Occam's razor)



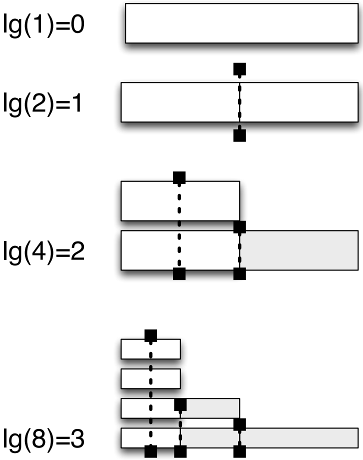
Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Makes big numbers small
- Way to think about them: cutting a carrot



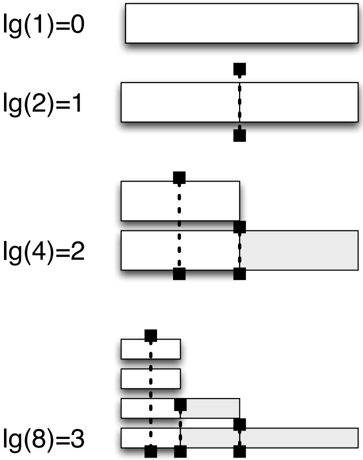
Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Makes big numbers small
- Way to think about them: cutting a carrot
- Negative numbers?



Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Makes big numbers small
- Way to think about them: cutting a carrot
- Negative numbers?
- Non-integers?



Entropy

Entropy is a measure of uncertainty that is associated with the distribution of a random variable:

$$\begin{aligned}
H(X) &= -\mathrm{E}\left[\lg(p(X))\right] \\
&= -\sum_{x} p(x)\lg(p(x)) && \text{(discrete)} \\
&= -\int_{-\infty}^{\infty} p(x)\lg(p(x))\,dx && \text{(continuous)}
\end{aligned}$$

Entropy

Entropy is a measure of uncertainty that is associated with the distribution of a random variable:

$$
\begin{aligned}
H(X) &= -\mathrm{E}\left[\lg(p(X))\right] \\
&= -\sum_x p(x)\lg(p(x)) && \text{(discrete)} \\
&= -\int_{-\infty}^{\infty} p(x)\lg(p(x))\,dx && \text{(continuous)}
\end{aligned}
$$

Does not account for the values of the random variable, only the spread of the distribution.

- $H(X) \geq 0$
- uniform distribution $=$ highest entropy, point mass $=$ lowest
- suppose $P(X=1)=p,\ P(X=0)=1-p$ and $P(Y=100)=p,\ P(Y=0)=1-p$: X and Y have the same entropy

Wrap up

- Probabilities are the language of data science
- You'll need to manipulate probabilities and understand marginalization and independence
- In Class: Working through probability examples
- Next: **Conditional** probabilities

Math Review

Slides adapted from Dave Blei and Lauren Hannah

University of Maryland

Conditional Probability

Context

- Data science is often worried about "if-then" questions
  - If my e-mail looks like this, is it spam?
  - If I buy this stock, will my portfolio improve?
- Since data science uses the language of probabilities, we need <u>conditional</u> probabilities (continuing probability intro)
- Also need to **combine** distributions

Conditional Probabilities

The *conditional probability* of event A given event B is the probability of A when B is known to occur,

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

Conditional Probabilities

The *conditional probability* of event A given event B is the probability of A when B is known to occur,

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

Conditional Probabilities

The *conditional probability* of event A given event B is the probability of A when B is known to occur,

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

Conditional Probabilities

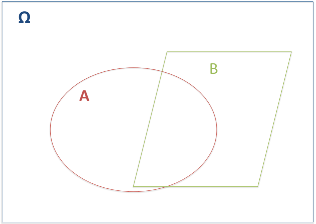
The *conditional probability* of event A given event B is the probability of A when B is known to occur,

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

Conditional Probabilities

The *conditional probability* of event A given event B is the probability of A when B is known to occur,

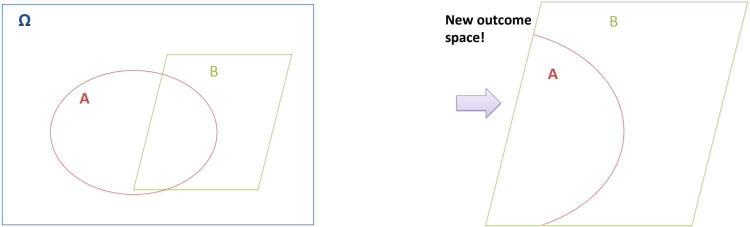
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$



Conditional Probabilities

The *conditional probability* of event A given event B is the probability of A when B is known to occur,

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



Independence (Reminder)

Random variables X and Y are independent if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$. How does this interact with conditional probabilities?

Conditional probabilities equal unconditional probabilities with independence:

- $P(X = x \mid Y) = P(X = x)$
- *Knowing Y tells us nothing about X*

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

$$P(A > 3 \cap B + A = 6) =$$
$$P(A > 3) =$$
$$P(A > 3 | B + A = 6) =$$

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

$$P(A>3 \cap B+A=6) = \frac{2}{36}$$

$$P(A>3) =$$

$$P(A>3 | B+A=6) =$$

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

$$P(A > 3 \cap B + A = 6) = \frac{2}{36}$$

$$P(A > 3) = \frac{3}{6}$$

$$P(A > 3 | B + A = 6) =$$

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

$$P(A > 3 \cap B + A = 6) = \frac{2}{36}$$

$$P(A > 3) = \frac{3}{6}$$

$$P(A > 3 | B + A = 6) = \frac{\frac{2}{36}}{\frac{3}{6}} = \frac{2}{36}\,\frac{6}{3}$$

Conditional Probabilities

Example

What is the probability that the sum of two dice is six given that the first is greater than three?

- $A \equiv$ First die
- $B \equiv$ Second die

	B=1	B=2	B=3	B=4	B=5	B=6
A=1	2	3	4	5	6	7
A=2	3	4	5	6	7	8
A=3	4	5	6	7	8	9
A=4	5	6	7	8	9	10
A=5	6	7	8	9	10	11
A=6	7	8	9	10	11	12

$$P(A > 3 \cap B + A = 6) = \frac{2}{36}$$

$$P(A > 3) = \frac{3}{6}$$

$$P(A > 3 | B + A = 6) = \frac{\frac{2}{36}}{\frac{3}{6}} = \frac{2}{36}\frac{6}{3}$$

$$= \frac{1}{9}$$

Combining Distributions

- Somtimes distributions you have aren't what you need
  - Conditional $\longrightarrow$ joint (chain)
  - Reverse conditional direction (Bayes')

The chain rule

- The definition of conditional probability lets us derive the *chain rule*, which let's us define the joint distribution as a product of conditionals:

$$P(X, Y) \quad = \quad P(X, Y)\frac{P(Y)}{P(Y)}$$

The chain rule

- The definition of conditional probability lets us derive the *chain rule*, which let's us define the joint distribution as a product of conditionals:

$$\begin{aligned} P(X,Y) &= P(X,Y)\frac{P(Y)}{P(Y)} \\ &= P(X|Y)P(Y) \end{aligned}$$

The chain rule

- The definition of conditional probability lets us derive the *chain rule*, which let's us define the joint distribution as a product of conditionals:

$$\begin{aligned} P(X,Y) &= P(X,Y)\frac{P(Y)}{P(Y)} \\ &= P(X|Y)P(Y) \end{aligned}$$

- For example, let Y be a disease and X be a symptom. We may know $P(X|Y)$ and $P(Y)$ from data. Use the chain rule to obtain the probability of having the disease and the symptom.
- In general, for any set of N variables

$$P(X_1,\ldots,X_N) = \prod_{n=1}^{N} P(X_n|X_1,\ldots,X_{n-1})$$

Bayes' Rule

What is the relationship between $P(A|B)$ and $P(B|A)$?

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

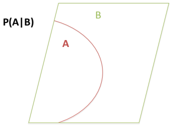
1. Start with $P(A|B)$
2. Change outcome space from B to Ω
3. Change outcome space again from Ω to A

Bayes' Rule

What is the relationship between $P(A|B)$ and $P(B|A)$?

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

1. Start with $P(A|B)$
2. Change outcome space from B to Ω
3. Change outcome space again from Ω to A

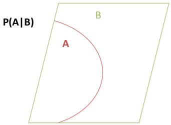


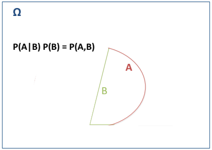
Bayes' Rule

What is the relationship between $P(A|B)$ and $P(B|A)$?

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

1. Start with $P(A|B)$
2. Change outcome space from B to Ω: $P(A|B)P(B)$
3. Change outcome space again from Ω to A





Bayes' Rule

What is the relationship between $P(A|B)$ and $P(B|A)$?

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

1. Start with $P(A|B)$
2. Change outcome space from B to Ω: $P(A|B)P(B)$
3. Change outcome space again from Ω to A: $\frac{P(A|B)P(B)}{P(A)}$