Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

# Mathematical Foundations

## Introduction to Data Science Algorithms

Michael Paul and Jordan Boyd-Graber

JANUARY 23, 2017

## By the end of today . . .

- You’ll be able to apply the concepts of distributions, independence, and conditional probabilities
- You’ll be able to derive joint, marginal, and conditional probabilites from each other
- You’ll be able to compute expectations and entropies

## Preface: Why make us do this?

- Probabilities are the language we use to describe data
- A reasonable (but geeky) definition of data science is how to get probabilities we care about from data
- Later classes will be about how to do this for different probability models
- But first, we need key definitions of probability (and it makes more sense to do it all at once)

## Preface: Why make us do this?

- Probabilities are the language we use to describe data
- A reasonable (but geeky) definition of data science is how to get probabilities we care about from data
- Later classes will be about how to do this for different probability models
- But first, we need key definitions of probability (and it makes more sense to do it all at once)
- So pay attention!

# Mathematical Foundations

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

JANUARY 23, 2017

## Preface: Why make us do this?

- Probabilities (along with statistics) are the language we use to describe data
- A reasonable (but geeky) definition of data science is how to get probabilities we care about from data
- Later classes will be about how to do this for different probability models
- But first, we need key definitions of probability (and it makes more sense to do it all at once)

## Preface: Why make us do this?

- Probabilities (along with statistics) are the language we use to describe data
- A reasonable (but geeky) definition of data science is how to get probabilities we care about from data
- Later classes will be about how to do this for different probability models
- But first, we need key definitions of probability (and it makes more sense to do it all at once)
- So pay attention!

## By the end of today . . .

- You'll be able to apply the concepts of distributions, independence, and conditional probabilities
- You'll be able to derive joint, marginal, and conditional probabilities from each other
- You'll be able to compute expectations and entropies

## 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

College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

# Mathematical Foundations

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

SLIDES ADAPTED FROM DAVE BLEI AND LAUREN HANNAH

## Random variable

- Probability is about *random variables.*
- A random variable is any “probabilistic” outcome.
- Examples of variables:
  - Yesterday’s high temperature
  - The height of someone
- Examples of random variables:
  - Tomorrow’s high temperature
  - The height of someone chosen randomly from a population

## Random variable

- Probability is about *random variables.*
- A random variable is any "probabilistic" outcome.
- Examples of variables:
  - Yesterday's high temperature
  - The height of someone
- Examples of random variables:
  - Tomorrow's high temperature
  - The height of someone chosen randomly from a population
- We'll see that it's sometimes useful to think of quantities that are not strictly probabilistic as random variables.
  - The high temperature on 03/04/1905
  - The number of times "streetlight" appears in a document

## 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
- Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

$$P(D>3)=P(D=4)+P(D=5)+P(D=6)$$

- 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
- Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

$$P(D>3)=P(D=4)+P(D=5)+P(D=6)$$

- 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
- Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

$$P(D>3)=P(D=4)+P(D=5)+P(D=6)$$

- 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
- Probabilities of disjunctions are sums over part of the space. E.g., the probability that a die is bigger than 3:

$$P(D>3)=P(D=4)+P(D=5)+P(D=6)$$

- The probabilities over the entire space must sum to one

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

## Events

An *event* is a set of outcomes to which a probability is assigned

- drawing a black card from a deck of cards
- drawing a King of Hearts

Intersections and unions:

- Intersection: drawing a red and a King

$$P(A \cap B) \tag{1}$$

- Union: drawing a spade or a King

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2}$$

## Events

An *event* is a set of outcomes to which a probability is assigned

- drawing a black card from a deck of cards
- drawing a King of Hearts

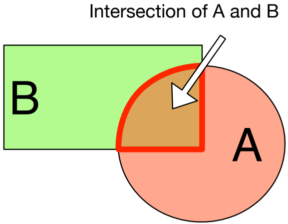
Intersections and unions:

- Intersection: drawing a red and a King

$$P(A \cap B) \tag{1}$$

- Union: drawing a spade or a King

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2}$$



## Events

An *event* is a set of outcomes to which a probability is assigned

- drawing a black card from a deck of cards
- drawing a King of Hearts

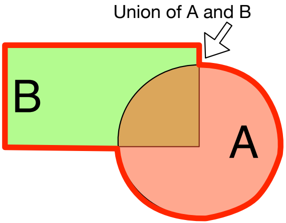
Intersections and unions:

- Intersection: drawing a red and a King

$$P(A \cap B) \quad (1)$$

- Union: drawing a spade or a King

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (2)$$



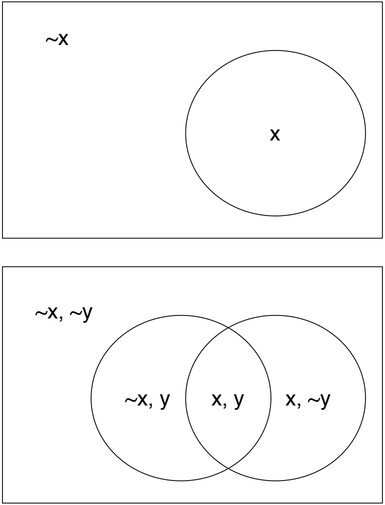
## Joint distribution

- Typically, we consider collections of random variables.
- The *joint distribution* is a distribution over the configuration of all the random variables in the ensemble.
- For example, imagine flipping 4 coins. The joint distribution is over the space of all possible outcomes of the four coins.

$$
\begin{aligned}
P(HHHH) &= 0.0625 \\
P(HHHT) &= 0.0625 \\
P(HHTH) &= 0.0625 \\
&\dots
\end{aligned}
$$

- You can think of it as a single random variable with 16 values.

## Visualizing a joint distribution



# Mathematical Foundations

## INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

SLIDES ADAPTED FROM DAVE BLEI AND LAUREN HANNAH

## Marginalization

If we know a joint distribution of multiple variables, what if we want to know the distribution of only one of the variables?

We can compute the distribution of $P(X)$ from $P(X, Y, Z)$ through *marginalization*:

$$\begin{aligned}\sum_y \sum_z P(X, Y = y, Z = z) &= \sum_y \sum_z P(X) P(Y = y, Z = z \mid X) \\ &= P(X) \sum_y \sum_z P(Y = y, Z = z \mid X) \\ &= P(X)\end{aligned}$$

## Marginalization

If we know a joint distribution of multiple variables, what if we want to know the distribution of only one of the variables?

We can compute the distribution of $P(X)$ from $P(X, Y, Z)$ through *marginalization*:

$$\begin{aligned}\sum_y \sum_z P(X, Y = y, Z = z) &= \sum_y \sum_z P(X) P(Y = y, Z = z \mid X) \\ &= P(X) \sum_y \sum_z P(Y = y, Z = z \mid X) \\ &= P(X)\end{aligned}$$

We'll explain this notation more next week for now the formula is the most important part.

## Marginalization (from Leyton-Brown)

**Joint distribution**

temperature (T) and weather (W)

| | T=Hot | T=Mild | T=Cold |
|---|---|---|---|
| W=Sunny | .10 | .20 | .10 |
| W=Cloudy | .05 | .35 | .20 |

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather
- Marginalize out temperature

## Marginalization (from Leyton-Brown)

### Joint distribution

temperature (T) and weather (W)

| | T=Hot | T=Mild | T=Cold |
|---|---|---|---|
| W=Sunny | .10 | .20 | .10 |
| W=Cloudy | .05 | .35 | .20 |

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X, Y) = \sum_z P(X, Y, Z = z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

| T=Hot | T=Mild | T=Cold |
|---|---|---|
| | | |

- Marginalize out temperature

## Marginalization (from Leyton-Brown)

### Joint distribution

temperature (T) and weather (W)

| | T=Hot | T=Mild | T=Cold |
|---|---|---|---|
| W=Sunny | .10 | .20 | .10 |
| W=Cloudy | .05 | .35 | .20 |

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

| T=Hot | T=Mild | T=Cold |
|---|---|---|
| | | |

- Marginalize out temperature

## Marginalization (from Leyton-Brown)

**Joint distribution**

temperature (T) and weather (W)

| | T=Hot | T=Mild | T=Cold |
|---|---|---|---|
| W=Sunny | .10 | .20 | .10 |
| W=Cloudy | .05 | .35 | .20 |

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

| T=Hot | T=Mild | T=Cold |
|---|---|---|
| .15 | | |

- Marginalize out temperature

## Marginalization (from Leyton-Brown)

### Joint distribution

temperature (T) and weather (W)

| | T=Hot | T=Mild | T=Cold |
|---|---|---|---|
| W=Sunny | .10 | .20 | .10 |
| W=Cloudy | .05 | .35 | .20 |

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X, Y) = \sum_z P(X, Y, Z = z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

| T=Hot | T=Mild | T=Cold |
|---|---|---|
| .15 | .55 | .30 |

- Marginalize out temperature

## Marginalization (from Leyton-Brown)

### Joint distribution

temperature (T) and weather (W)

| | T=Hot | T=Mild | T=Cold |
|---|---|---|---|
| W=Sunny | .10 | .20 | .10 |
| W=Cloudy | .05 | .35 | .20 |

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

| T=Hot | T=Mild | T=Cold |
|---|---|---|
| .15 | .55 | .30 |

- Marginalize out temperature

| |
|---|
| W=Sunny |
| W=Cloudy |

## Marginalization (from Leyton-Brown)

**Joint distribution**

temperature (T) and weather (W)

| | T=Hot | T=Mild | T=Cold |
|---|---|---|---|
| W=Sunny | .10 | .20 | .10 |
| W=Cloudy | .05 | .35 | .20 |

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

| T=Hot | T=Mild | T=Cold |
|---|---|---|
| .15 | .55 | .30 |

- Marginalize out temperature

| |
|---|
| W=Sunny |
| W=Cloudy |

## Marginalization (from Leyton-Brown)

### Joint distribution

temperature (T) and weather (W)

| | T=Hot | T=Mild | T=Cold |
|---|---|---|---|
| W=Sunny | .10 | .20 | .10 |
| W=Cloudy | .05 | .35 | .20 |

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X, Y) = \sum_z P(X, Y, Z = z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

| T=Hot | T=Mild | T=Cold |
|---|---|---|
| .15 | .55 | .30 |

- Marginalize out temperature

| | |
|---|---|
| W=Sunny | .40 |
| W=Cloudy | |

## Marginalization (from Leyton-Brown)

### Joint distribution

temperature (T) and weather (W)

| | T=Hot | T=Mild | T=Cold |
|---|---|---|---|
| W=Sunny | .10 | .20 | .10 |
| W=Cloudy | .05 | .35 | .20 |

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

| T=Hot | T=Mild | T=Cold |
|---|---|---|
| .15 | .55 | .30 |

- Marginalize out temperature

| | |
|---|---|
| W=Sunny | .40 |
| W=Cloudy | .60 |

# Mathematical Foundations

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

SLIDES ADAPTED FROM DAVE BLEI AND LAUREN HANNAH

## Independence

Random variables $X$ and $Y$ are *independent* if and only if $P(X=x,Y=y)=P(X=x)P(Y=y)$.

Mathematical examples:

- If I flip a coin twice, is the second outcome independent from the first outcome?

## Independence

Random variables $X$ and $Y$ are *independent* if and only if $P(X=x, Y=y) = P(X=x)P(Y=y)$.

Mathematical examples:

- If I flip a coin twice, is the second outcome independent from the first outcome?
- If I draw two socks from my (multicolored) laundry, is the color of the first sock independent from the color of the second sock?

## Independence

Intuitive Examples:

- Independent:
  - you use a Mac / the Hop bus is on schedule
  - snowfall in the Himalayas / your favorite color is blue

## Independence

Intuitive Examples:

- Independent:
  - you use a Mac / the Hop bus is on schedule
  - snowfall in the Himalayas / your favorite color is blue
- Not independent:
  - you vote for Mitt Romney / you are a Republican
  - there is a traffic jam on 25 / the Broncos are playing

## Independence

Sometimes we make convenient assumptions.

- the values of two dice (ignoring gravity!)
- the value of the first die and the sum of the values
- whether it is raining and the number of taxi cabs
- whether it is raining and the amount of time it takes me to hail a cab
- the first two words in a sentence

# Mathematical Foundations

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

SLIDES ADAPTED FROM DAVE BLEI AND LAUREN HANNAH

## Expectation

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

$$\mathrm{E}[f(X)] = \sum_x f(x)\, p(x) \qquad \text{(discrete)}$$

$$= \int_{-\infty}^{\infty} f(x)\, p(x)\, dx \qquad \text{(continuous)}$$

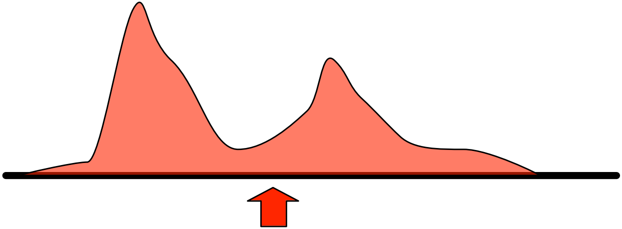
# Expectation

Expectations of constants or known values:

- $\mathrm{E}[a] = a$
- $\mathrm{E}[Y \mid Y = y] = y$

## Expectation Intuition

- 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?

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**

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

# Mathematical Foundations

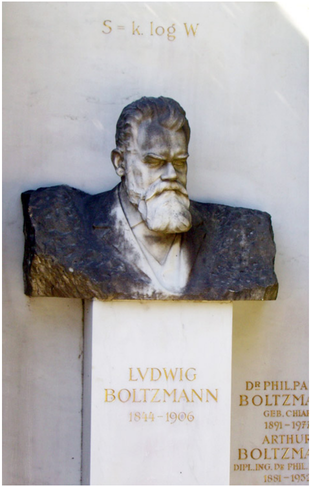
INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

SLIDES ADAPTED FROM DAVE BLEI AND LAUREN HANNAH

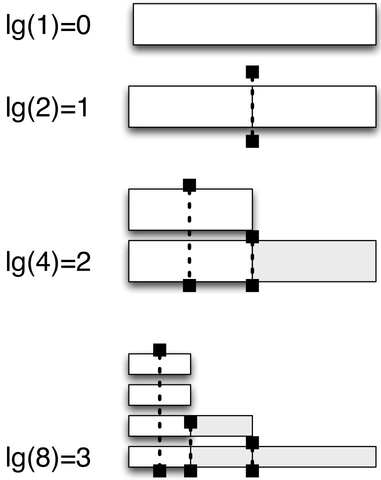
## Entropy

- 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)



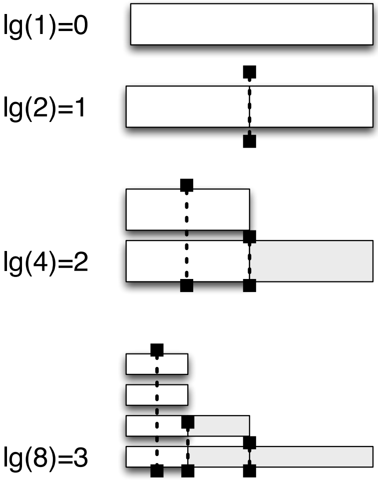
## Aside: Logarithms

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



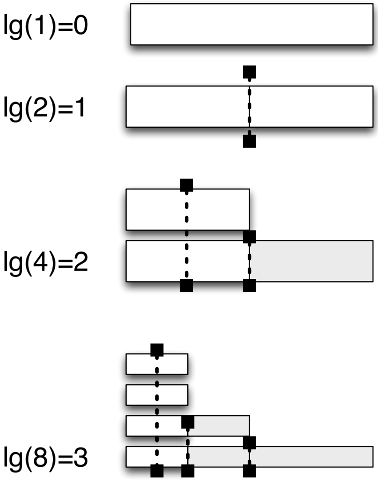
## Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- 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
- Thursday: Working through probability examples
- Next week: **Conditional** probabilities