Probability and Information

Roman V. Belavkin

Faculty of Science and Technology
Middlesex University, London NW4 4BT, UK

July 19, 2021
ACDL 2021

Introduction

Data frequency and probability

- We have considered frequent itemsets to infer association rules (i.e. discover knowledge) from a transactional database (TDB).

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

Data frequency and probability

- We have considered frequent itemsets to infer association rules (i.e. discover knowledge) from a transactional database (TDB).
- How is frequency related to probability?

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

Data frequency and probability

- We have considered frequent itemsets to infer association rules (i.e. discover knowledge) from a transactional database (TDB).
- How is frequency related to probability?

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

Data frequency and probability

- We have considered frequent itemsets to infer association rules (i.e. discover knowledge) from a transactional database (TDB).
- How is frequency related to probability?

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

The support of itemset A in TDB is the fraction of transactions with A:

$$\mathrm{supp}(A) = \frac{\#\text{transactions}(A)}{\#\text{transactions}} = \frac{n(A)}{n}$$

where $\#$ means 'the number n of' (e.g. $\mathrm{supp}(\text{bread}) = 4/5$).

Data frequency and probability

- We have considered frequent itemsets to infer association rules (i.e. discover knowledge) from a transactional database (TDB).
- How is frequency related to probability?

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

The support of itemset A in TDB is the fraction of transactions with A:

$$\mathrm{supp}(A) = \frac{\#\text{transactions}(A)}{\#\text{transactions}} = \frac{n(A)}{n}$$

where $\#$ means 'the number n of' (e.g. $\mathrm{supp}(\text{bread}) = 4/5$).

Laws of large numbers

The frequency of observing event E in n independent and identically distributed (i.i.d.) experiments converges (in some sense) to the probability of E:

$$\frac{n(E)}{n} \to P(E) \quad \text{as} \quad n \to \infty$$

Brief history of probability theory

1654 Blaise Pascal and Pierre Fermat discuss games of chance.

Brief history of probability theory

1654 Blaise Pascal and Pierre Fermat discuss games of chance.

1657 Christian Huygens publishes *On Ratiocination in Dice Games*.

Brief history of probability theory

1654 Blaise Pascal and Pierre Fermat discuss games of chance.

1657 Christian Huygens publishes *On Ratiocination in Dice Games*.

1760 Thomas Bayes defines conditional probability.

Brief history of probability theory

1654 Blaise Pascal and Pierre Fermat discuss games of chance.

1657 Christian Huygens publishes *On Ratiocination in Dice Games*.

1760 Thomas Bayes defines conditional probability.

1812 Pierre-Simon Laplace (principle of insufficient reason).

Brief history of probability theory

1654 Blaise Pascal and Pierre Fermat discuss games of chance.

1657 Christian Huygens publishes *On Ratiocination in Dice Games*.

1760 Thomas Bayes defines conditional probability.

1812 Pierre-Simon Laplace (principle of insufficient reason).

1932 John von Neumann's *Mathematical Foundations of Quantum Mechanics*.

Brief history of probability theory

1654 Blaise Pascal and Pierre Fermat discuss games of chance.

1657 Christian Huygens publishes *On Ratiocination in Dice Games*.

1760 Thomas Bayes defines conditional probability.

1812 Pierre-Simon Laplace (principle of insufficient reason).

1932 John von Neumann's *Mathematical Foundations of Quantum Mechanics*.

1933 Andrey Kolmogorov's formulates axioms of probability.

Brief history of probability theory

- 1654 Blaise Pascal and Pierre Fermat discuss games of chance.
- 1657 Christian Huygens publishes *On Ratiocination in Dice Games*.
- 1760 Thomas Bayes defines conditional probability.
- 1812 Pierre-Simon Laplace (principle of insufficient reason).
- 1932 John von Neumann's *Mathematical Foundations of Quantum Mechanics*.
- 1933 Andrey Kolmogorov's formulates axioms of probability.
- 1920–1940 Ronald Fisher, Abraham Wald (work statistics).

Brief history of probability theory

1654 Blaise Pascal and Pierre Fermat discuss games of chance.

1657 Christian Huygens publishes *On Ratiocination in Dice Games*.

1760 Thomas Bayes defines conditional probability.

1812 Pierre-Simon Laplace (principle of insufficient reason).

1932 John von Neumann's *Mathematical Foundations of Quantum Mechanics*.

1933 Andrey Kolmogorov's formulates axioms of probability.

1920–1940 Ronald Fisher, Abraham Wald (work statistics).

1948 Claude Shannon (information theory).

Brief history of probability theory

- 1654 Blaise Pascal and Pierre Fermat discuss games of chance.
- 1657 Christian Huygens publishes *On Ratiocination in Dice Games*.
- 1760 Thomas Bayes defines conditional probability.
- 1812 Pierre-Simon Laplace (principle of insufficient reason).
- 1932 John von Neumann's *Mathematical Foundations of Quantum Mechanics*.
- 1933 Andrey Kolmogorov's formulates axioms of probability.
- 1920–1940 Ronald Fisher, Abraham Wald (work statistics).
- 1948 Claude Shannon (information theory).
- 1965 Value of information theory (Stratonovich).

Brief history of probability theory

1654 Blaise Pascal and Pierre Fermat discuss games of chance.

1657 Christian Huygens publishes *On Ratiocination in Dice Games*.

1760 Thomas Bayes defines conditional probability.

1812 Pierre-Simon Laplace (principle of insufficient reason).

1932 John von Neumann's *Mathematical Foundations of Quantum Mechanics*.

1933 Andrey Kolmogorov's formulates axioms of probability.

1920–1940 Ronald Fisher, Abraham Wald (work statistics).

1948 Claude Shannon (information theory).

1965 Value of information theory (Stratonovich).

1970-80 Information geometry (e.g. Chentsov, Amari).

Sources of uncertainty

Complexity : the number of possible states of a system in question can be too large (e.g. predict how a chess game can develop after 10 moves?)

Sources of uncertainty

Complexity : the number of possible states of a system in question can be too large (e.g. predict how a chess game can develop after 10 moves?)

Ignorance : some important information about the system may not be available.

Sources of uncertainty

Complexity : the number of possible states of a system in question can be too large (e.g. predict how a chess game can develop after 10 moves?)

Ignorance : some important information about the system may not be available.

Randomness : the system may be random by nature, and thus the uncertainty is irreducible.

What is probability?

Definition (Probability of event E)

the measure $P(E)$ of certainty that event E will occur and ranging from $P(E) = 0$ (impossible) to $P(E) = 1$ (certain):

$$\text{(Impossible)} \qquad 0 \leq P(E) \leq 1 \qquad \text{(Certain)}$$

What is probability?

Definition (Probability of event E)

the measure $P(E)$ of certainty that event E will occur and ranging from $P(E) = 0$ (impossible) to $P(E) = 1$ (certain):

$$\text{(Impossible)} \qquad 0 \leq P(E) \leq 1 \qquad \text{(Certain)}$$

Example (Fair coin)

For a fair coin, $P(\text{heads}) = \frac{1}{2} = 0.5$

What is probability?

Definition (Probability of event E)

the measure $P(E)$ of certainty that event E will occur and ranging from $P(E) = 0$ (impossible) to $P(E) = 1$ (certain):

$$\text{(Impossible)} \qquad 0 \leq P(E) \leq 1 \qquad \text{(Certain)}$$

Example (Fair coin)

For a fair coin, $P(\text{heads}) = \frac{1}{2} = 0.5$

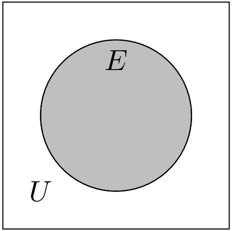
Example (Dice)

For a fair die, $P(6) = \frac{1}{6}$

Set-theoretic intuition

- Events E are considered as subsets of the universal set U:

$$E \subseteq U$$

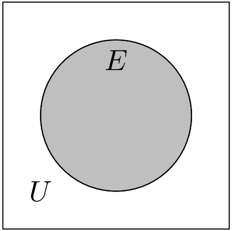


Set-theoretic intuition

- Events E are considered as subsets of the universal set U:

$$E \subseteq U$$

- Probability of E is a measure of a subset $E \subseteq U$.



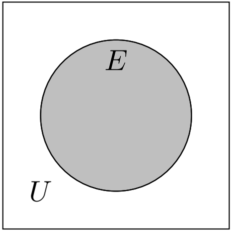
Set-theoretic intuition

- Events E are considered as subsets of the universal set U:

$$E \subseteq U$$

- Probability of E is a measure of a subset $E \subseteq U$.
- Probabilities of negation (not E), disjunction (A or B) and conjunction (A and B):

$$P(\bar{E}) = P(U - E)\,, \quad P(A \cup B)\,, \quad P(A \cap B)$$



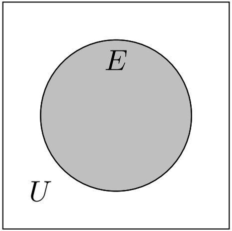
Set-theoretic intuition

- Events E are considered as subsets of the universal set U:

$$E \subseteq U$$

- Probability of E is a measure of a subset $E \subseteq U$.
- Probabilities of negation (not E), disjunction (A or B) and conjunction (A and B):

$$P(\bar{E}) = P(U - E)\,, \quad P(A \cup B)\,, \quad P(A \cap B)$$



Universal set

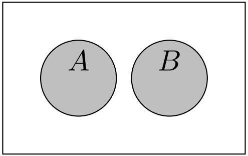
Because the unvierse is certain, we set

$$P(U) = 1$$

Additivity of probabilities

- For disjoint events $A \cap B = \varnothing$:

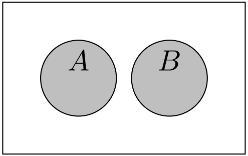
$$P(A \text{ or } B) = P(A) + P(B)$$



Additivity of probabilities

- For disjoint events $A \cap B = \varnothing$:

$$P(A \text{ or } B) = P(A) + P(B)$$



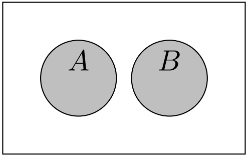
- For n disjoint events such that $E_1 \cup E_2 \cup \cdots \cup E_n = U$

$$P(E_1)+P(E_2)+\cdots+P(E_n) = P(U) = 1$$

Additivity of probabilities

- For disjoint events $A \cap B = \varnothing$:

$$P(A \text{ or } B) = P(A) + P(B)$$



- For n disjoint events such that $E_1 \cup E_2 \cup \cdots \cup E_n = U$

$$P(E_1)+P(E_2)+\cdots+P(E_n) = P(U) = 1$$

Example

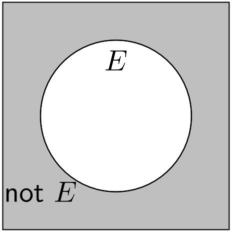
For a fair coin and a fair dice we have

$$\frac{1}{2} + \frac{1}{2} = 1 \qquad \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

Probability of negation

- Probability of E not happening, is the measure of the complement of E:

$$P(\mathrm{not}E) = P(U - E)$$



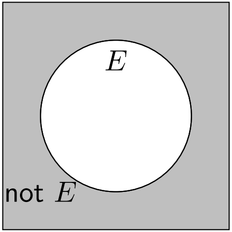
Probability of negation

- Probability of E not happening, is the measure of the complement of E:

$$P(\text{not}E) = P(U - E)$$

- We can show that

$$P(\text{not}E) = 1 - P(E)$$



Probability of negation

- Probability of E not happening, is the measure of the complement of E:

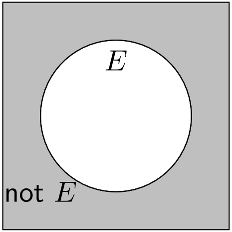
$$P(\text{not}E) = P(U - E)$$

- We can show that

$$P(\text{not}E) = 1 - P(E)$$

- Because

$$P(E \text{ or not } E) = P(U) = 1$$
$$P(E \text{ or not } E) = P(E) + P(\text{not}E)$$



Probability of negation

- Probability of E not happening, is the measure of the complement of E:

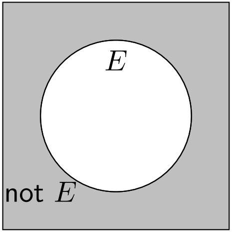
$$P(\text{not}E) = P(U - E)$$

- We can show that

$$P(\text{not}E) = 1 - P(E)$$

- Because

$$P(E \text{ or not } E) = P(U) = 1$$
$$P(E \text{ or not } E) = P(E) + P(\text{not}E)$$

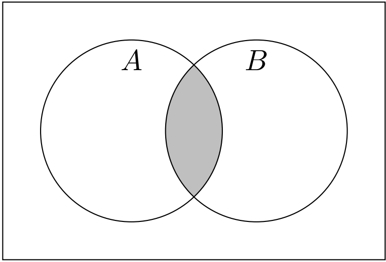


Empty set

$$P(\varnothing) = P(\text{not}U) = 1 - P(U) = 0$$

Joint probability

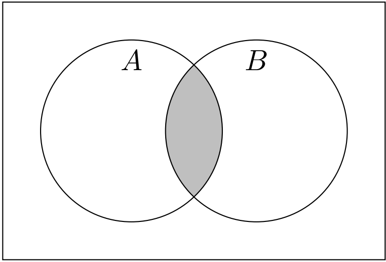
- Co-occurrence of events A and B together (e.g. clouds and rain) is their set intersection: $A \cap B$.



Joint probability

- Co-occurrence of events A and B together (e.g. clouds and rain) is their set intersection: $A \cap B$.
- Probability of $A \cap B$ is called joint probability:

$$P(A \cap B)$$

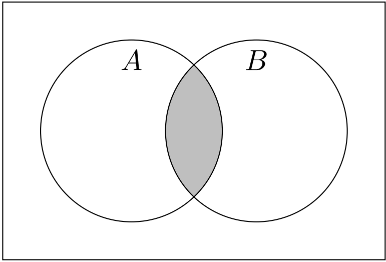


Joint probability

- Co-occurrence of events A and B together (e.g. clouds and rain) is their set intersection: $A \cap B$.
- Probability of $A \cap B$ is called joint probability:

$$P(A \cap B)$$

- Often denoted simply $P(A, B)$.

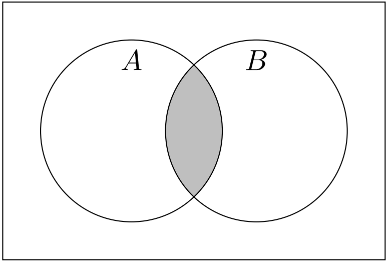


Joint probability

- Co-occurrence of events A and B together (e.g. clouds and rain) is their set intersection: $A \cap B$.
- Probability of $A \cap B$ is called joint probability:

$$P(A \cap B)$$

- Often denoted simply $P(A, B)$.



Example (Two coins)

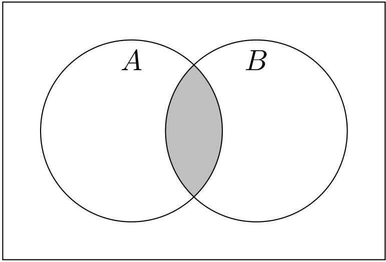
A	B
heads	heads
heads	tails
tails	heads
tails	tails

Joint probability

- Co-occurrence of events A and B together (e.g. clouds and rain) is their set intersection: $A \cap B$.
- Probability of $A \cap B$ is called joint probability:

$$P(A \cap B)$$

- Often denoted simply $P(A, B)$.



Example (Two coins)

A	B
heads	heads
heads	tails
tails	heads
tails	tails

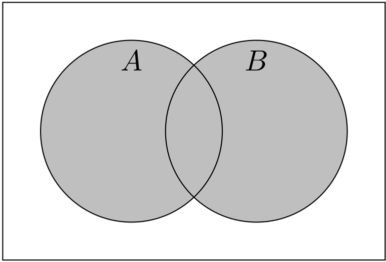
Example (Bread and milk)

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

Probability of union

- Probability of A or B is:

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

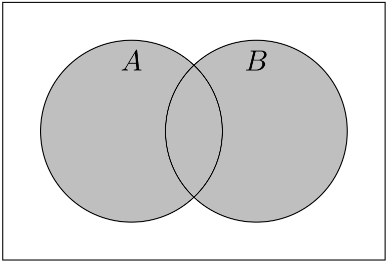


Probability of union

- Probability of A or B is:

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

- We subtract $P(A \cap B)$, because otherwise we count it twice.

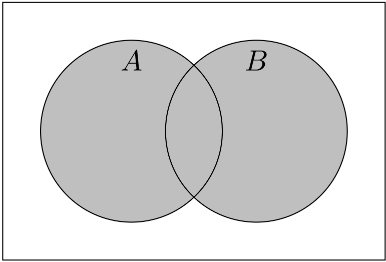


Probability of union

- Probability of A or B is:

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

- We subtract $P(A \cap B)$, because otherwise we count it twice.
- Check for $P(A \cap B) = \varnothing$.

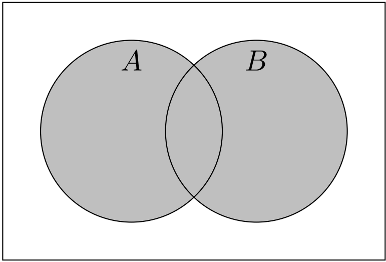


Probability of union

- Probability of A or B is:

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

- We subtract $P(A \cap B)$, because otherwise we count it twice.
- Check for $P(A \cap B) = \varnothing$.



Example (Bread or milk)

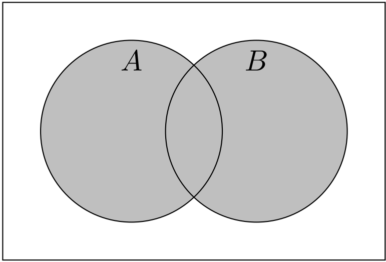
- $P(\text{bread}) = 4/5$ and $P(\text{milk}) = 4/5$

Probability of union

- Probability of A or B is:

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

- We subtract $P(A \cap B)$, because otherwise we count it twice.
- Check for $P(A \cap B) = \varnothing$.



Example (Bread or milk)

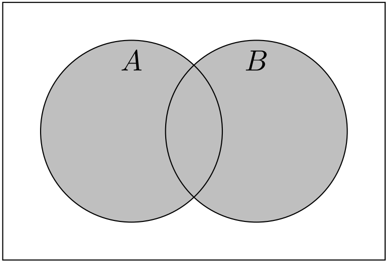
- $P(\text{bread}) = 4/5$ and $P(\text{milk}) = 4/5$
- What is $P(\text{bread} \cup \text{milk})$?

Probability of union

- Probability of A or B is:

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

- We subtract $P(A \cap B)$, because otherwise we count it twice.
- Check for $P(A \cap B) = \varnothing$.



Example (Bread or milk)

- $P(\text{bread}) = 4/5$ and $P(\text{milk}) = 4/5$
- What is $P(\text{bread} \cup \text{milk})$?
- Using $P(\text{bread} \cap \text{milk}) = 3/5$ we have

$$P(\text{bread} \cup \text{milk}) = \frac{4}{5} + \frac{4}{5} - \frac{3}{5} = 1$$

Probability distributions

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

- Consider events: '*item x belongs to a transaction T*'

Probability distributions

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

- Consider events: '*item x belongs to a transaction T*'
- What are their probabilities?

Probability distributions

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

- Consider events: '*item x belongs to a transaction T*'
- What are their probabilities?
- For example, coke appears 2 out of 18 items bought

$$P(\text{coke} \in T) = \frac{2}{18} = \frac{1}{6}$$

Probability distributions

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

- Consider events: '*item x belongs to a transaction T*'
- What are their probabilities?
- For example, coke appears 2 out of 18 items bought

$$P(\text{coke} \in T) = \frac{2}{18} = \frac{1}{6}$$

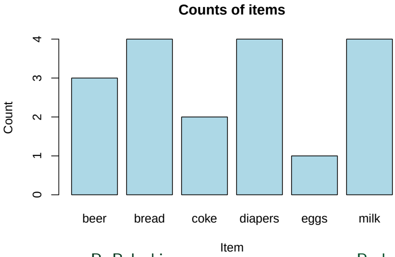
Probability distributions

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

- Consider events: '*item x belongs to a transaction T*'
- What are their probabilities?
- For example, coke appears 2 out of 18 items bought

$$P(\text{coke} \in T) = \frac{2}{18} = \frac{1}{6}$$

Probability distribution is the collection of probabilities of all such events:



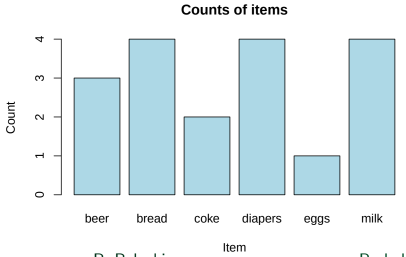
Probability distributions

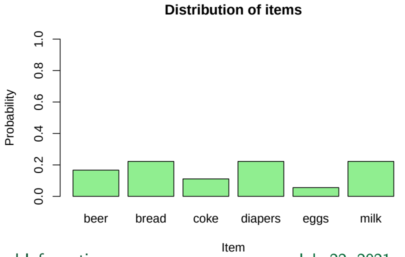
TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

- Consider events: '*item x belongs to a transaction T*'
- What are their probabilities?
- For example, coke appears 2 out of 18 items bought

$$P(\text{coke} \in T) = \frac{2}{18} = \frac{1}{6}$$

Probability distribution is the collection of probabilities of all such events:





Random variables and their distributions

- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \leq 100\,, \qquad x \geq 10\,, \quad x \in [10, 100]$$

Random variables and their distributions

- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \leq 100\,, \qquad x \geq 10\,, \quad x \in [10, 100]$$

- We can find probabilties of these events from their distributions.

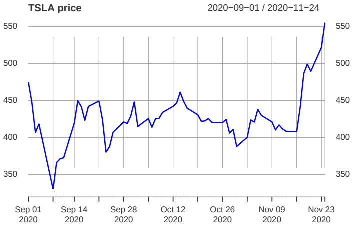
Random variables and their distributions

- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \leq 100\,, \qquad x \geq 10\,, \quad x \in [10, 100]$$

- We can find probabilties of these events from their distributions.

Distribution of Tesla prices



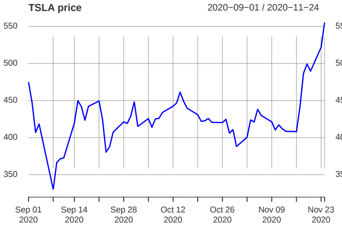
Random variables and their distributions

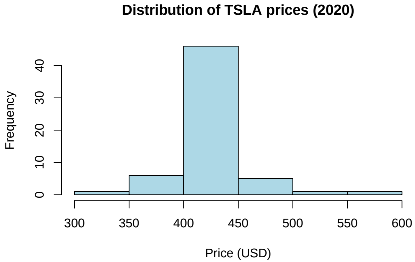
- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \leq 100\,, \qquad x \geq 10\,, \quad x \in [10, 100]$$

- We can find probabilties of these events from their distributions.

Distribution of Tesla prices





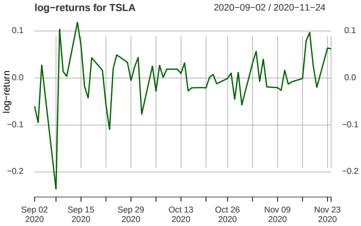
Random variables and their distributions

- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \leq 100\,, \qquad x \geq 10\,, \quad x \in [10, 100]$$

- We can find probabilties of these events from their distributions.

Distribution of Tesla returns



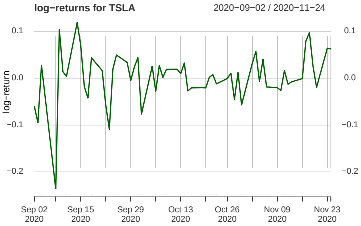
Random variables and their distributions

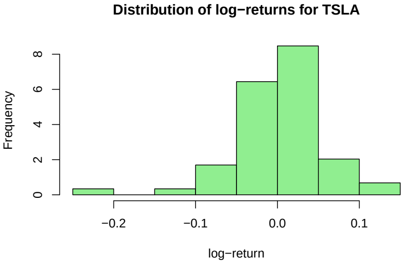
- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \le 100\,, \qquad x \ge 10\,, \quad x \in [10, 100]$$

- We can find probabilties of these events from their distributions.

Distribution of Tesla returns





Random variables and their distributions

- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \leq 100\,, \qquad x \geq 10\,, \quad x \in [10, 100]$$

- We can find probabilties of these events from their distributions.

Distribution of Bitcoin prices

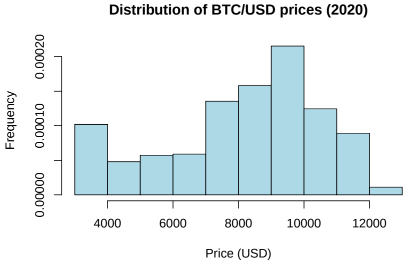
Random variables and their distributions

- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \le 100\,, \qquad x \ge 10\,, \quad x \in [10, 100]$$

- We can find probabilties of these events from their distributions.

Distribution of Bitcoin prices



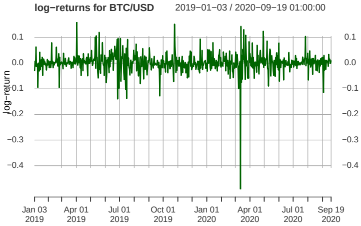
Random variables and their distributions

- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \leq 100\,, \qquad x \geq 10\,, \quad x \in [10, 100]$$

- We can find probabilties of these events from their distributions.

Distribution of Bitcoin returns



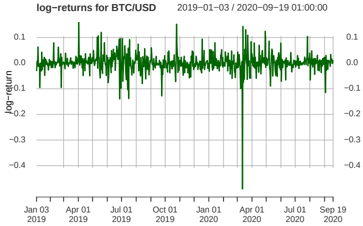
Random variables and their distributions

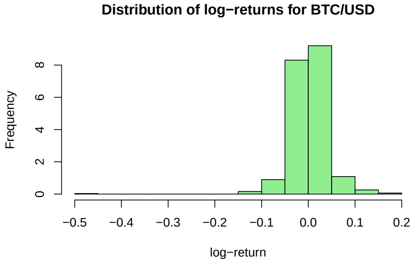
- For a random variable x, such as 'stock price' or 'return', we can consider events:

$$x \le 100 \,, \qquad x \ge 10 \,, \quad x \in [10, 100]$$

- We can find probabilties of these events from their distributions.

Distribution of Bitcoin returns

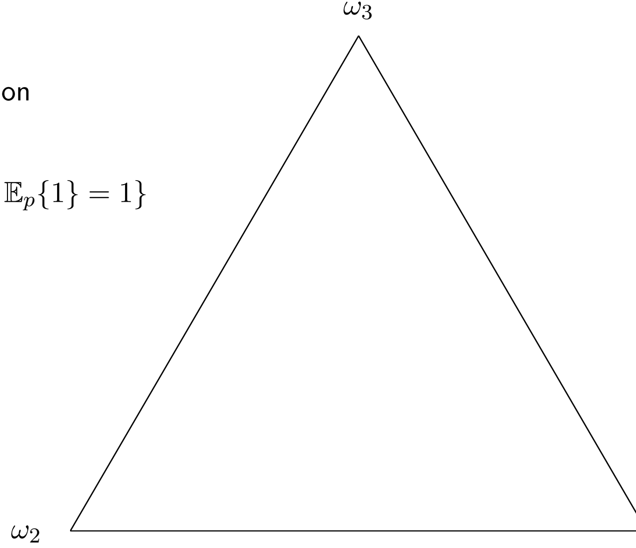




Information-geometruc view

- The set $\mathcal{P}(\Omega)$ of all probability measures on Ω is a simplex:

  $\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$

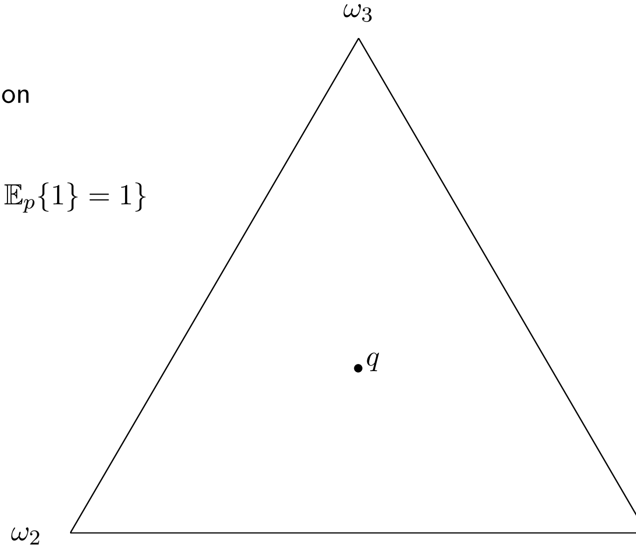


Information-geometruc view

- The set $\mathcal{P}(\Omega)$ of all probability measures on Ω is a simplex:

  $$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Can be defined for infinite Ω.

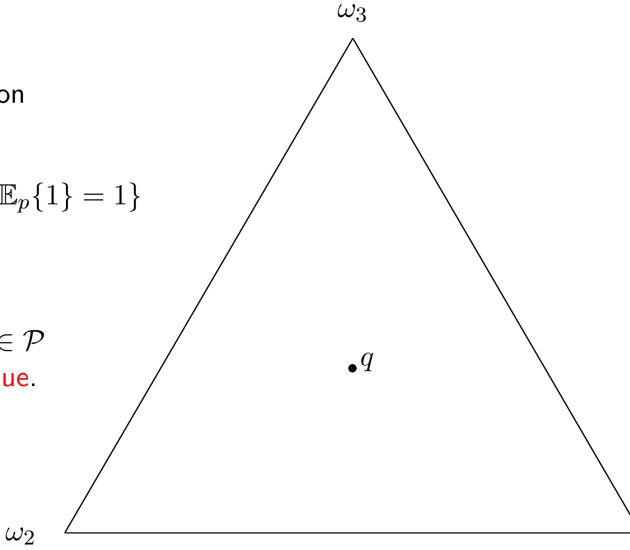


Information-geometruc view

- The set $\mathcal{P}(\Omega)$ of all probability measures on Ω is a simplex:

  $$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Can be defined for infinite Ω.
- Representations of $p \in \mathcal{P}$ by $\delta \in \operatorname{ext}\mathcal{P}$ are unique.

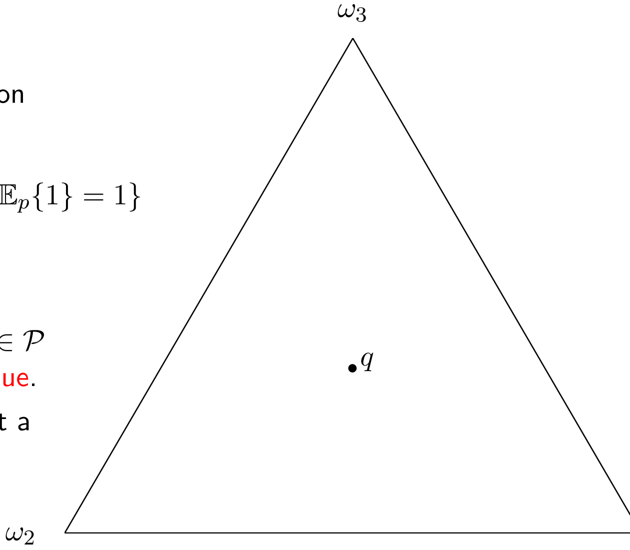


Information-geometruc view

- The set $\mathcal{P}(\Omega)$ of all probability measures on Ω is a simplex:

$$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Can be defined for infinite Ω.
- Representations of $p \in \mathcal{P}$ by $\delta \in \operatorname{ext}\mathcal{P}$ are unique.
- Quantum $\mathcal{P}(\Omega)$ is not a simplex.

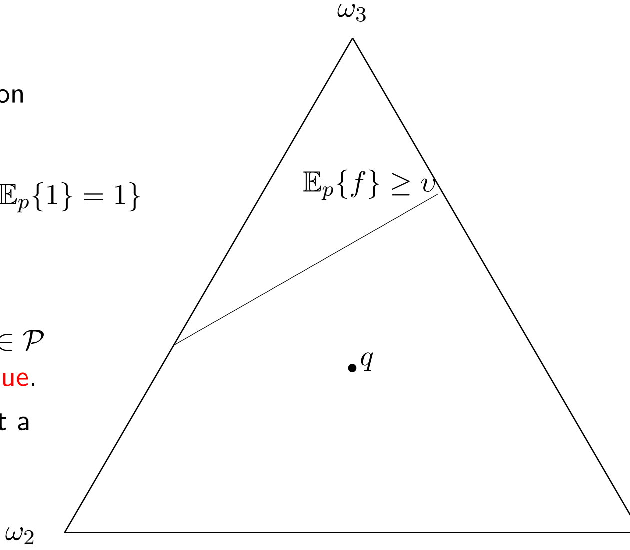


Information-geometruc view

- The set $\mathcal{P}(\Omega)$ of all probability measures on Ω is a simplex:

$$\mathcal{P}(\Omega) := \{p : p \geq 0\,,\ \mathbb{E}_p\{1\} = 1\}$$

- Can be defined for infinite Ω.
- Representations of $p \in \mathcal{P}$ by $\delta \in \operatorname{ext} \mathcal{P}$ are unique.
- Quantum $\mathcal{P}(\Omega)$ is not a simplex.



Measures of location

- Answer questions such as '*What is the most probable value?*', '*What is the most typical value?*', *What value should I expect in the long term?*'

Measures of location

- Answer questions such as '*What is the most probable value?*', '*What is the most typical value?*', *What value should I expect in the long term?*'
- If variable x has n possible values X_1, X_2,...,X_n with probabilities $P(X_1)$, $P(X_2)$,...,$P(X_n)$, then the expected value is

$$\mathbb{E}\{x\} = X_1 P(X_1) + X_2 P(X_2) + \cdots + X_n P(X_n) = \sum_{i=1}^{n} X_i \, P(X_i)$$

Measures of location

- Answer questions such as '*What is the most probable value?*', '*What is the most typical value?*', *What value should I expect in the long term?*'
- If variable x has n possible values X_1, X_2,...,X_n with probabilities $P(X_1)$, $P(X_2)$,...,$P(X_n)$, then the expected value is

$$\mathbb{E}\{x\} = X_1 P(X_1) + X_2 P(X_2) + \cdots + X_n P(X_n) = \sum_{i=1}^{n} X_i \, P(X_i)$$

- If all $P(x) = \frac{1}{n}$, then $\mathbb{E}\{x\}$ is the same as mean value (i.e. average).

Measures of location

- Answer questions such as '*What is the most probable value?*', '*What is the most typical value?*', *What value should I expect in the long term?*'
- If variable x has n possible values X_1, X_2,...,X_n with probabilities $P(X_1)$, $P(X_2)$,...,$P(X_n)$, then the expected value is

$$\mathbb{E}\{x\} = X_1 P(X_1) + X_2 P(X_2) + \cdots + X_n P(X_n) = \sum_{i=1}^{n} X_i \, P(X_i)$$

- If all $P(x) = \frac{1}{n}$, then $\mathbb{E}\{x\}$ is the same as mean value (i.e. average).

Measures of location

- Answer questions such as '*What is the most probable value?*', '*What is the most typical value?*', *What value should I expect in the long term?*'
- If variable x has n possible values X_1, X_2,...,X_n with probabilities $P(X_1)$, $P(X_2)$,...,$P(X_n)$, then the expected value is

$$\mathbb{E}\{x\} = X_1 P(X_1) + X_2 P(X_2) + \cdots + X_n P(X_n) = \sum_{i=1}^{n} X_i \, P(X_i)$$

- If all $P(x) = \frac{1}{n}$, then $\mathbb{E}\{x\}$ is the same as mean value (i.e. average).

Example

Let $\texttt{Age} = \{21, 18, 50, 23, 40\}$ and $P(\texttt{Age}) = \frac{1}{5}$. Then the mean age is

$$E\{\texttt{Age}\} = \frac{21 + 18 + 50 + 23 + 40}{5} = 30,4$$

Measures of Dispersion

- Answer questions such as '*What is the range of the variable?*', '*How far can it deviate from the mean?*', '*What risk is associated with the variable?*'

Measures of Dispersion

- Answer questions such as '*What is the range of the variable?*', '*How far can it deviate from the mean?*', '*What risk is associated with the variable?*'
- An absolute and squared deviation from the mean is respectively:

$$|x - \mathbb{E}\{x\}| \qquad \text{and} \qquad |x - \mathbb{E}\{x\}|^2$$

Measures of Dispersion

- Answer questions such as '*What is the range of the variable?*', '*How far can it deviate from the mean?*', '*What risk is associated with the variable?*'
- An absolute and squared deviation from the mean is respectively:

$$|x - \mathbb{E}\{x\}| \qquad \text{and} \qquad |x - \mathbb{E}\{x\}|^2$$

- We can compute the mean values of these deviations.

Measures of Dispersion

- Answer questions such as '*What is the range of the variable?*', '*How far can it deviate from the mean?*', '*What risk is associated with the variable?*'
- An absolute and squared deviation from the mean is respectively:

$$|x - \mathbb{E}\{x\}| \qquad \text{and} \qquad |x - \mathbb{E}\{x\}|^2$$

- We can compute the mean values of these deviations.
- The average squared deviation is called variance:

$$\mathrm{Var}\{x\} = \mathbb{E}\Big\{|x - \mathbb{E}\{x\}|^2\Big\}$$

Measures of Dispersion

- Answer questions such as '*What is the range of the variable?*', '*How far can it deviate from the mean?*', '*What risk is associated with the variable?*'
- An absolute and squared deviation from the mean is respectively:

$$|x - \mathbb{E}\{x\}| \qquad \text{and} \qquad |x - \mathbb{E}\{x\}|^2$$

- We can compute the mean values of these deviations.
- The average squared deviation is called variance:

$$\mathrm{Var}\{x\} = \mathbb{E}\Big\{|x - \mathbb{E}\{x\}|^2\Big\}$$

- Its square root is called standard deviation: $\mathrm{Sd}\{x\} = \sqrt{\mathrm{Var}\{x\}}$

Distribution vs Moments

- $P(\omega)$ gives us all moments of $x : \mathcal{A} \subseteq 2^\Omega \to \mathbb{R}$:

$$\mathbb{E}_P\{x\}\,,\ \mathbb{E}_P\{x^2\}\,,\ \mathbb{E}_P\{x^3\}\ldots$$

Distribution vs Moments

- $P(\omega)$ gives us all moments of $x : \mathcal{A} \subseteq 2^{\Omega} \to \mathbb{R}$:

$$\mathbb{E}_P\{x\}\,,\ \mathbb{E}_P\{x^2\}\,,\ \mathbb{E}_P\{x^3\}\ldots$$

- Note that

$$\mathbb{E}\{x^n\} = \frac{1}{i^n}\frac{\partial^n\Theta(u)}{\partial u^n}\bigg|_{u=0}$$

of the characteristic function $\Theta(u) = \mathbb{E}_P\{e^{iux}\}$.

Distribution vs Moments

- $P(\omega)$ gives us all moments of $x : \mathcal{A} \subseteq 2^{\Omega} \to \mathbb{R}$:

$$\mathbb{E}_P\{x\}\,,\ \mathbb{E}_P\{x^2\}\,,\ \mathbb{E}_P\{x^3\}\dots$$

- Note that

$$\mathbb{E}\{x^n\} = \frac{1}{i^n}\frac{\partial^n \Theta(u)}{\partial u^n}\bigg|_{u=0}$$

  of the characteristic function $\Theta(u) = \mathbb{E}_P\{e^{iux}\}$.

- $\Theta(u)$ is Fourier transform of P, so that

$$P(x) = \frac{1}{2\pi}\int_U \Theta(u)e^{-ixu}\,du$$

Distribution vs Moments

- $P(\omega)$ gives us all moments of $x : \mathcal{A} \subseteq 2^\Omega \to \mathbb{R}$:

$$\mathbb{E}_P\{x\}\,,\ \mathbb{E}_P\{x^2\}\,,\ \mathbb{E}_P\{x^3\}\dots$$

- Note that

$$\mathbb{E}\{x^n\} = \frac{1}{i^n}\frac{\partial^n \Theta(u)}{\partial u^n}\bigg|_{u=0}$$

  of the characteristic function $\Theta(u) = \mathbb{E}_P\{e^{iux}\}$.

- $\Theta(u)$ is Fourier transform of P, so that

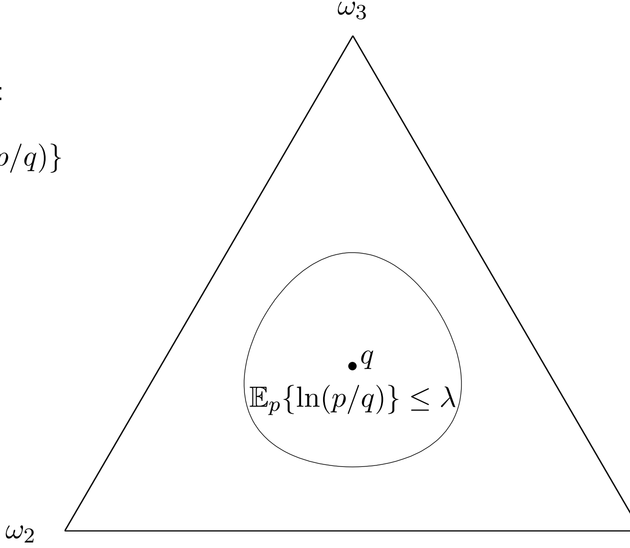
$$P(x) = \frac{1}{2\pi}\int_U \Theta(u)e^{-ixu}\,du$$

- What is better: To know $P(x)$ or to know moments $\mathbb{E}_P\{x^n\}$?

KL-divergence and $\Gamma(u) = \ln \Theta(u)$

- The KL-divergence between p, $q \in \mathcal{P}(\Omega)$:

  $D_{KL}[p, q] := \mathbb{E}_P\{\ln(p/q)\}$



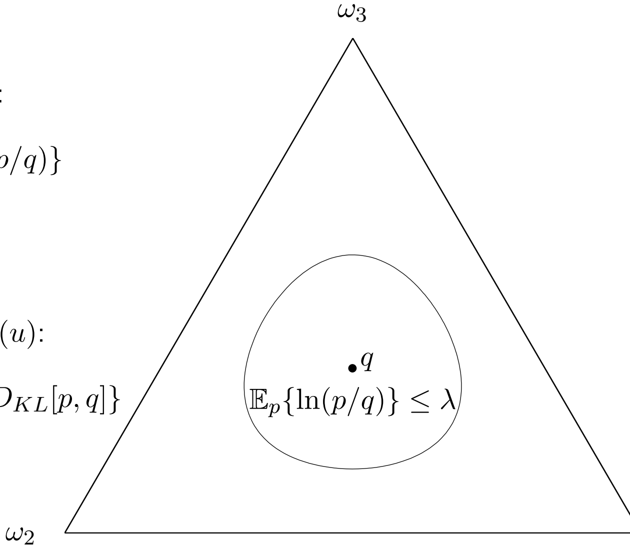
KL-divergence and $\Gamma(u) = \ln \Theta(u)$

- The KL-divergence between p, $q \in \mathcal{P}(\Omega)$:

  $$D_{KL}[p, q] := \mathbb{E}_P\{\ln(p/q)\}$$

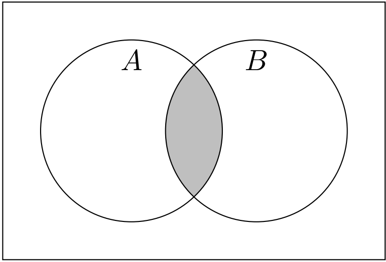
- Its Legendre-Fenchel transform is the kumulant generating function $\Gamma[u] := \ln \Theta(u)$:

  $$\Gamma[u] = \sup_p \{\mathbb{E}_p\{u\} - D_{KL}[p, q]\}$$



Joint probability

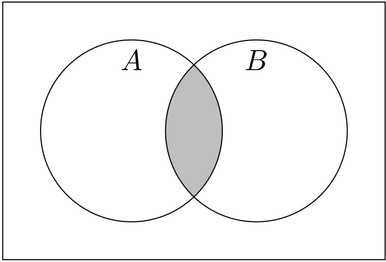
- Co-occurrence of events A and B together (e.g. clouds and rain) is their set intersection: $A \cap B$.



Joint probability

- Co-occurrence of events A and B together (e.g. clouds and rain) is their set intersection: $A \cap B$.
- Probability of $A \cap B$ is called joint probability:

$$P(A \cap B)$$

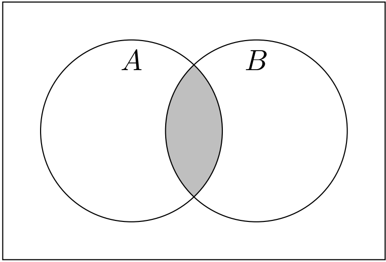


Joint probability

- Co-occurrence of events A and B together (e.g. clouds and rain) is their set intersection: $A \cap B$.
- Probability of $A \cap B$ is called joint probability:

$$P(A \cap B)$$

- Often denoted simply $P(A, B)$.

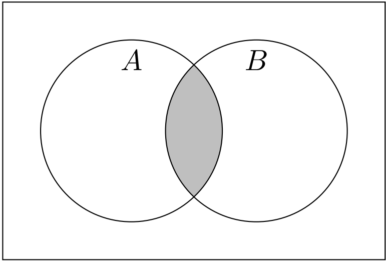


Joint probability

- Co-occurrence of events A and B together (e.g. clouds and rain) is their set intersection: $A \cap B$.
- Probability of $A \cap B$ is called joint probability:

$$P(A \cap B)$$

- Often denoted simply $P(A, B)$.



Example (Two independent fair coins)

Coin $A = \{\text{head}, \text{tail}\}$
Coin $B = \{\text{head}, \text{tail}\}$

$P(A \cap B) =$

	head	tail
head	1/4	1/4
tail	1/4	1/4

Marginal probabilities

Probabilities $P(A)$ or $P(B)$ are sometimes called marginal, because they can be obtained from joint probability $P(A \cap B)$ by summation:

$$P(A) = \sum_{b \in B} P(A \cap B)\,, \qquad P(B) = \sum_{a \in A} P(A \cap B)$$

Marginal probabilities

Probabilities $P(A)$ or $P(B)$ are sometimes called marginal, because they can be obtained from joint probability $P(A \cap B)$ by summation:

$$P(A) = \sum_{b \in B} P(A \cap B)\,, \qquad P(B) = \sum_{a \in A} P(A \cap B)$$

Example (Two independent fair coins)

$$P(A \cap B) = \left[\begin{array}{r|cc} & \text{head} & \text{tail} \\ \hline \text{head} & 1/4 & 1/4 \\ \text{tail} & 1/4 & 1/4 \end{array}\right]$$

Marginal probabilities

Probabilities $P(A)$ or $P(B)$ are sometimes called marginal, because they can be obtained from joint probability $P(A \cap B)$ by summation:

$$P(A) = \sum_{b \in B} P(A \cap B)\,, \qquad P(B) = \sum_{a \in A} P(A \cap B)$$

Example (Two independent fair coins)

$$P(A \cap B) = \left[\begin{array}{r|cc} & \text{head} & \text{tail} \\ \hline \text{head} & 1/4 & 1/4 \\ \text{tail} & 1/4 & 1/4 \end{array}\right] \quad P(A) = \left[\begin{array}{r|c} & \\ \hline \text{head} & 1/2 \\ \text{tail} & 1/2 \end{array}\right]$$

Marginal probabilities

Probabilities $P(A)$ or $P(B)$ are sometimes called marginal, because they can be obtained from joint probability $P(A \cap B)$ by summation:

$$P(A) = \sum_{b \in B} P(A \cap B)\,, \qquad P(B) = \sum_{a \in A} P(A \cap B)$$

Example (Two independent fair coins)

$$P(A \cap B) = \left[\begin{array}{r|cc} & \text{head} & \text{tail} \\ \hline \text{head} & 1/4 & 1/4 \\ \text{tail} & 1/4 & 1/4 \end{array}\right] \quad P(A) = \left[\begin{array}{r|c} & \\ \hline \text{head} & 1/2 \\ \text{tail} & 1/2 \end{array}\right]$$

$$P(B) = \left[\begin{array}{r|cc} & \text{head} & \text{tail} \\ \hline & 1/2 & 1/2 \end{array}\right]$$

Marginal probabilities

Probabilities $P(A)$ or $P(B)$ are sometimes called marginal, because they can be obtained from joint probability $P(A \cap B)$ by summation:

$$P(A) = \sum_{b \in B} P(A \cap B)\,, \qquad P(B) = \sum_{a \in A} P(A \cap B)$$

Example (Clouds and rain)

$$P(A \cap B) = \left[\begin{array}{r|cc} & \text{no rain} & \text{rain} \\ \hline \text{no clouds} & 1/2 & 0 \\ \text{clouds} & 3/10 & 1/5 \end{array}\right]$$

Marginal probabilities

Probabilities $P(A)$ or $P(B)$ are sometimes called marginal, because they can be obtained from joint probability $P(A \cap B)$ by summation:

$$P(A) = \sum_{b \in B} P(A \cap B)\,, \qquad P(B) = \sum_{a \in A} P(A \cap B)$$

Example (Clouds and rain)

$$P(A \cap B) = \left[\begin{array}{r|cc} & \text{no rain} & \text{rain} \\ \hline \text{no clouds} & 1/2 & 0 \\ \text{clouds} & 3/10 & 1/5 \end{array}\right] \quad P(A) = \left[\begin{array}{r|c} & \\ \hline \text{no clouds} & 1/2 \\ \text{clouds} & 1/2 \end{array}\right]$$

Marginal probabilities

Probabilities $P(A)$ or $P(B)$ are sometimes called marginal, because they can be obtained from joint probability $P(A \cap B)$ by summation:

$$P(A) = \sum_{b \in B} P(A \cap B), \qquad P(B) = \sum_{a \in A} P(A \cap B)$$

Example (Clouds and rain)

$P(A \cap B) =$

	no rain	rain
no clouds	$1/2$	0
clouds	$3/10$	$1/5$

$P(A) =$

no clouds	$1/2$
clouds	$1/2$

$P(B) =$

	no rain	rain
	$4/5$	$1/5$

Conditional probability

Question

How likely is it to rain if you see clouds?

Conditional probability

Question

How likely is it to rain if you see clouds?

Definition (Conditional probability)

- The probability of event A conditioned on the outcome of B:

$$P(A \mid B)$$

- The condition on B can be understood as 'B *has already happened*' or as an assumption that you '*know the outcome of* B'.

Conditional probability

Question

How likely is it to rain if you see clouds?

Definition (Conditional probability)

- The probability of event A conditioned on the outcome of B:

$$P(A \mid B)$$

- The condition on B can be understood as 'B *has already happened*' or as an assumption that you '*know the outcome of* B'.

Conditional probability

Question

How likely is it to rain if you see clouds?

Definition (Conditional probability)

- The probability of event A conditioned on the outcome of B:

$$P(A \mid B)$$

- The condition on B can be understood as 'B *has already happened*' or as an assumption that you '*know the outcome of* B'.

Example (Clouds and rain)

- For $A = \{\text{clouds}, \text{clear sky}\}$ and $B = \{\text{rain}, \text{no rain}\}$, we can consider

$$P(\text{rain} \mid \text{clouds})$$

- Is it the same as $P(\text{rain})$?

Independence

- Conditional probability is used to define the statistical dependence.

Independence

- Conditional probability is used to define the statistical dependence.
- Events A and B are independent if (and only if):

$$P(A \mid B) = P(A) \qquad \text{or} \qquad P(B \mid A) = P(B)$$

Independence

- Conditional probability is used to define the statistical dependence.
- Events A and B are independent if (and only if):

$$P(A \mid B) = P(A) \qquad \text{or} \qquad P(B \mid A) = P(B)$$

- This means B does not change the chance of A (and vice versa).

Independence

- Conditional probability is used to define the statistical dependence.
- Events A and B are independent if (and only if):

$$P(A \mid B) = P(A) \qquad \text{or} \qquad P(B \mid A) = P(B)$$

- This means B does not change the chance of A (and vice versa).
- Knowledge about B does not add any information about A (i.e. does not reduce uncertainty about A).

Independence

- Conditional probability is used to define the statistical dependence.
- Events A and B are independent if (and only if):

$$P(A \mid B) = P(A) \qquad \text{or} \qquad P(B \mid A) = P(B)$$

- This means B does not change the chance of A (and vice versa).
- Knowledge about B does not add any information about A (i.e. does not reduce uncertainty about A).
- Otherwise, if

$$P(A \mid B) \neq P(A) \qquad \text{or} \qquad P(B \mid A) \neq P(B)$$

events A and B are said to be statistically dependent.

Independence

- Conditional probability is used to define the statistical dependence.
- Events A and B are independent if (and only if):

$$P(A \mid B) = P(A) \qquad \text{or} \qquad P(B \mid A) = P(B)$$

- This means B does not change the chance of A (and vice versa).
- Knowledge about B does not add any information about A (i.e. does not reduce uncertainty about A).
- Otherwise, if

$$P(A \mid B) \neq P(A) \qquad \text{or} \qquad P(B \mid A) \neq P(B)$$

events A and B are said to be statistically dependent.

Example (Clouds and rain)

If you believe that rain is not possible without clouds, then

$$P(\text{clouds} \mid \text{rain}) = 1$$

Independence

- Conditional probability is used to define the statistical dependence.
- Events A and B are independent if (and only if):

$$P(A \mid B) = P(A) \qquad \text{or} \qquad P(B \mid A) = P(B)$$

- This means B does not change the chance of A (and vice versa).
- Knowledge about B does not add any information about A (i.e. does not reduce uncertainty about A).
- Otherwise, if

$$P(A \mid B) \neq P(A) \qquad \text{or} \qquad P(B \mid A) \neq P(B)$$

events A and B are said to be statistically dependent.

Example (Clouds and rain)

If you believe that rain is not possible without clouds, then

$$P(\text{clouds} \mid \text{rain}) = 1\,, \quad \text{but} \qquad P(\text{clouds}) \neq 1$$

Conditional probability formula

- From the decomposition of joint probability $P(A \cap B)$:

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

Conditional probability formula

- From the decomposition of joint probability $P(A \cap B)$:

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

- We also obtain the formulae for the conditional probabilities:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{and} \qquad P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Conditional probability formula

- From the decomposition of joint probability $P(A \cap B)$:

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

- We also obtain the formulae for the conditional probabilities:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{and} \qquad P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

- Compare with rule's confidence: $\mathrm{conf}(B \impliedby A) = \frac{\mathrm{supp}(A) \cap \mathrm{supp}(B)}{\mathrm{supp}(A)}$

Conditional probability formula

- From the decomposition of joint probability $P(A \cap B)$:

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

- We also obtain the formulae for the conditional probabilities:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{and} \qquad P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

- Compare with rule's confidence: $\mathrm{conf}(B \Longleftarrow A) = \frac{\mathrm{supp}(A) \cap \mathrm{supp}(B)}{\mathrm{supp}(A)}$

Example

In the TDB example, we saw

$$P(\text{Milk} \mid \text{Bread}) = \frac{3}{4} = \frac{3/5}{4/5}$$

Bayes' rule

- Look at these two decompositions of joint probability $P(A \cap B)$

$$P(A \mid B)P(B) = P(B \mid A)P(B)$$

Bayes' rule

- Look at these two decompositions of joint probability $P(A \cap B)$

$$P(A \mid B)P(B) = P(B \mid A)P(B)$$

- Divide both sides by $P(B)$ or by $P(A)$ and obtain the formula:

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

Bayes' rule

- Look at these two decompositions of joint probability $P(A \cap B)$

$$P(A \mid B)P(B) = P(B \mid A)P(B)$$

- Divide both sides by $P(B)$ or by $P(A)$ and obtain the formula:

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

- It is called the Bayes' rule (due to Thomas Bayes, 1763).

Bayes' rule

- Look at these two decompositions of joint probability $P(A \cap B)$

$$P(A \mid B)P(B) = P(B \mid A)P(B)$$

- Divide both sides by $P(B)$ or by $P(A)$ and obtain the formula:

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

- It is called the Bayes' rule (due to Thomas Bayes, 1763).
- It relates two conditional probabilities $P(A \mid B)$ with $P(B \mid A)$.

Bayes' rule

- Look at these two decompositions of joint probability $P(A \cap B)$

$$P(A \mid B)P(B) = P(B \mid A)P(B)$$

- Divide both sides by $P(B)$ or by $P(A)$ and obtain the formula:

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

- It is called the Bayes' rule (due to Thomas Bayes, 1763).
- It relates two conditional probabilities $P(A \mid B)$ with $P(B \mid A)$.
- It is important, because often one is easier to estimate than the other.

Bayes' rule

- Look at these two decompositions of joint probability $P(A \cap B)$

$$P(A \mid B)P(B) = P(B \mid A)P(B)$$

- Divide both sides by $P(B)$ or by $P(A)$ and obtain the formula:

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

- It is called the Bayes' rule (due to Thomas Bayes, 1763).
- It relates two conditional probabilities $P(A \mid B)$ with $P(B \mid A)$.
- It is important, because often one is easier to estimate than the other.

Example (Clouds and rain)

- What is $P(\text{rain} \mid \text{clouds}) = ?$

Bayes' rule

- Look at these two decompositions of joint probability $P(A \cap B)$

$$P(A \mid B)P(B) = P(B \mid A)P(B)$$

- Divide both sides by $P(B)$ or by $P(A)$ and obtain the formula:

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

- It is called the Bayes' rule (due to Thomas Bayes, 1763).
- It relates two conditional probabilities $P(A \mid B)$ with $P(B \mid A)$.
- It is important, because often one is easier to estimate than the other.

Example (Clouds and rain)

- What is $P(\text{rain} \mid \text{clouds}) =?$
- Assuming $P(\text{clouds} \mid \text{rain}) = 1$ and $P(\text{rain}) = 1/5$, $P(\text{clouds}) = 1/2$

$$P(\text{rain} \mid \text{clouds}) = \frac{1 \times 1/5}{1/2} = \frac{2}{5}$$

Another view on independence

- The joint probability $P(A \cap B)$ (i.e. probability of A and B)

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

Another view on independence

- The joint probability $P(A \cap B)$ (i.e. probability of A and B)

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

- for independent A and B becomes simply

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

Another view on independence

- The joint probability $P(A \cap B)$ (i.e. probability of A and B)

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

- for independent A and B becomes simply

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

- Because $P(A \mid B) = P(A)$ if A and B are independent.

Another view on independence

- The joint probability $P(A \cap B)$ (i.e. probability of A and B)

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

- for independent A and B becomes simply

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

- Because $P(A \mid B) = P(A)$ if A and B are independent.

Example (Two independent fair coins)

$$P(A \cap B) = \left[\begin{array}{r|cc} & \text{head} & \text{tail} \\ \hline \text{head} & 1/4 & 1/4 \\ \text{tail} & 1/4 & 1/4 \end{array}\right] \quad P(A) = \left[\begin{array}{r|c} & \\ \hline \text{head} & 1/2 \\ \text{tail} & 1/2 \end{array}\right]$$

$$P(B) = \left[\begin{array}{r|cc} & \text{head} & \text{tail} \\ \hline & 1/2 & 1/2 \end{array}\right] \quad P(\text{head}, \text{tail}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Another view on independence

- The joint probability $P(A \cap B)$ (i.e. probability of A and B)

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

- for independent A and B becomes simply

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

- Because $P(A \mid B) = P(A)$ if A and B are independent.

Example (Clouds and rain)

$P(A \cap B) =$

	no rain	rain
no clouds	$1/2$	0
clouds	$3/10$	$1/5$

$P(A) =$

no clouds	$1/2$
clouds	$1/2$

$P(B) =$

	no rain	rain
	$4/5$	$1/5$

Another view on independence

- The joint probability $P(A \cap B)$ (i.e. probability of A and B)

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

- for independent A and B becomes simply

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

- Because $P(A \mid B) = P(A)$ if A and B are independent.

Example (Clouds and rain)

$P(A \cap B) =$

	no rain	rain
no clouds	$1/2$	0
clouds	$3/10$	$1/5$

$P(A) =$

no clouds	$1/2$
clouds	$1/2$

$P(B) =$

	no rain	rain
	$4/5$	$1/5$

$$P(\text{no clouds}, \text{rain}) = 0 \neq \frac{1}{2} \times \frac{1}{5}$$

Probability and surprise

- If $P(E)$ is the probability of event E, then the logarithm of $1/P(E)$ is a measure of surprise associated with observing E:

Probability and surprise

- If $P(E)$ is the probability of event E, then the logarithm of $1/P(E)$ is a measure of surprise associated with observing E:

Probability and surprise

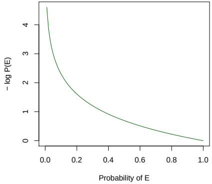
- If $P(E)$ is the probability of event E, then the logarithm of $1/P(E)$ is a measure of surprise associated with observing E:

$$h(E) = \log \frac{1}{P(E)}$$

Probability and surprise

- If $P(E)$ is the probability of event E, then the logarithm of $1/P(E)$ is a measure of surprise associated with observing E:

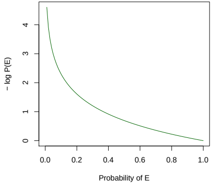
$$h(E) = \log \frac{1}{P(E)}$$



Probability and surprise

- If $P(E)$ is the probability of event E, then the logarithm of $1/P(E)$ is a measure of surprise associated with observing E:

$$h(E) = \log \frac{1}{P(E)}$$

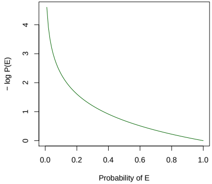


- It also represents the amount of information associated with E.

Probability and surprise

- If $P(E)$ is the probability of event E, then the logarithm of $1/P(E)$ is a measure of surprise associated with observing E:

$$h(E) = \log \frac{1}{P(E)}$$

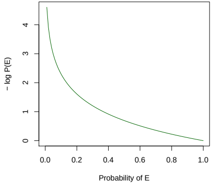


- It also represents the amount of information associated with E.
- When probability $P(E)$ is high (e.g. $P(E) = 1$), then there is no surprise (not much information), and vice versa.

Probability and surprise

- If $P(E)$ is the probability of event E, then the logarithm of $1/P(E)$ is a measure of surprise associated with observing E:

$$h(E) = \log \frac{1}{P(E)}$$



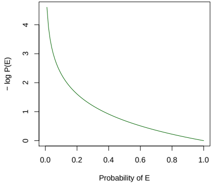
- It also represents the amount of information associated with E.
- When probability $P(E)$ is high (e.g. $P(E) = 1$), then there is no surprise (not much information), and vice versa.
- The average (expected) surprise is called entropy:

$$H(E) = -\mathbb{E}\{\log P(E)\}$$

Probability and surprise

- If $P(E)$ is the probability of event E, then the logarithm of $1/P(E)$ is a measure of surprise associated with observing E:

$$h(E) = \log \frac{1}{P(E)}$$



- It also represents the amount of information associated with E.
- When probability $P(E)$ is high (e.g. $P(E) = 1$), then there is no surprise (not much information), and vice versa.
- The average (expected) surprise is called entropy:

$$H(E) = -\mathbb{E}\{\log P(E)\}$$

- It is usually thought of as a measure of uncertainty, but it is also a measure of potential information.

Mutual information

- The logarithm of the ratio of $P(A \mid B)$ and $P(A)$ is called random mutual information:

$$i(A, B) = \log \frac{P(A \mid B)}{P(A)}$$

Mutual information

- The logarithm of the ratio of $P(A \mid B)$ and $P(A)$ is called random mutual information:

$$i(A, B) = \log \frac{P(A \mid B)}{P(A)}$$

- Notice that $i(A, B) = 0$ iff $P(A \mid B) = P(A)$ (because $\log 1 = 0$).

Mutual information

- The logarithm of the ratio of $P(A \mid B)$ and $P(A)$ is called random mutual information:

$$i(A,B) = \log \frac{P(A \mid B)}{P(A)}$$

- Notice that $i(A,B) = 0$ iff $P(A \mid B) = P(A)$ (because $\log 1 = 0$).
- The conditional probability can be expressed as

$$P(A \mid B) = P(A)\, e^{i(A,B)}$$

Mutual information

- The logarithm of the ratio of $P(A \mid B)$ and $P(A)$ is called random mutual information:

$$i(A, B) = \log \frac{P(A \mid B)}{P(A)}$$

- Notice that $i(A, B) = 0$ iff $P(A \mid B) = P(A)$ (because $\log 1 = 0$).
- The conditional probability can be expressed as

$$P(A \mid B) = P(A)\, e^{i(A,B)}$$

- $e^{i(A,B)}$ represents dependency between A and B (because $e^0 = 1$).

Mutual information

- The logarithm of the ratio of $P(A \mid B)$ and $P(A)$ is called random mutual information:

$$i(A, B) = \log \frac{P(A \mid B)}{P(A)}$$

- Notice that $i(A, B) = 0$ iff $P(A \mid B) = P(A)$ (because $\log 1 = 0$).
- The conditional probability can be expressed as

$$P(A \mid B) = P(A)\, e^{i(A,B)}$$

- $e^{i(A,B)}$ represents dependency between A and B (because $e^0 = 1$).
- The average (expected) value of $i(A, B)$ is called (Shannon's) mutual information:

$$I(A, B) = \mathbb{E}\left\{\log \frac{P(A \mid B)}{P(A)}\right\}$$

Mutual information

- The logarithm of the ratio of $P(A \mid B)$ and $P(A)$ is called random mutual information:

$$i(A,B) = \log \frac{P(A \mid B)}{P(A)}$$

- Notice that $i(A,B) = 0$ iff $P(A \mid B) = P(A)$ (because $\log 1 = 0$).
- The conditional probability can be expressed as

$$P(A \mid B) = P(A)\, e^{i(A,B)}$$

- $e^{i(A,B)}$ represents dependency between A and B (because $e^0 = 1$).
- The average (expected) value of $i(A,B)$ is called (Shannon's) mutual information:

$$I(A,B) = \mathbb{E}\left\{\log \frac{P(A \mid B)}{P(A)}\right\} = \underbrace{H(A)}_{\text{prior uncert.}} - \underbrace{H(A \mid B)}_{\text{posterior uncert. (after } B)}$$

- Thus, information is the amount by which uncertainty is reduced.