MATH 323: Probability

Fall 2018

David A. Stephens

Department of Mathematics and Statistics, McGill University

david.stephens@mcgill.ca

Burnside Hall, Room 1225.

Introduction

Syllabus

1. The basics of probability.
   - Review of set theory notation.
   - Sample spaces and events.
   - The probability axioms and their consequences.
   - Probability spaces with equally likely outcomes.
   - Combinatorial probability.
   - Conditional probability and independence.
   - The Theorem of Total Probability.
   - Bayes Theorem.

Syllabus (cont.)

2. Random variables and probability distributions.
   - Random variables.
   - Discrete and continuous univariate distributions: cdfs, pmfs and pdfs.
   - Moments: expectation and variance.
   - Moment generating functions (mgfs): derivation and uses.
   - Named distributions:
     - discrete uniform,
     - hypergeometric,
     - binomial,
     - Poisson,
     - continuous uniform,
     - gamma,
     - exponential,
     - chi-squared,
     - beta,
     - Normal.

Syllabus (cont.)

3. Probability calculation methods.
   - Transformations in one dimension.
   - Techniques for sums of random variables.

Syllabus (cont.)

4. Multivariate distributions.
   - Marginal cdfs and pdfs.
   - Conditional cdfs and pdfs.
   - Conditional expectation.
   - Independence of random variables.
   - Covariance and correlation.

Syllabus (cont.)

5. Probability inequalities and theorems.
   - Markov's inequality.
   - Chebychev's inequality.
   - Definition of convergence in probability.
   - The Weak Law of Large Numbers.
   - The Central Limit Theorem and applications.

Introduction

Introduction

This course is concerned with developing mathematical concepts and techniques for modelling and analyzing situations involving **uncertainty**.

- *Uncertainty* corresponds to a lack of complete or perfect knowledge.
- Assessment of uncertainty in such real-life problems is a complex issue which requires a rigorous mathematical treatment.
- This course will develop the probability framework in which questions of practical interest can be posed and resolved.

Introduction (cont.)

Uncertainty could be the result of

- incomplete observation of a system;
- unpredictable variation;
- simple lack of knowledge of the "state of nature".

Introduction (cont.)

"State of nature": some aspect of the real world.

- could be the *current* state that is imperfectly observed;
- could be the *future* state, that is, the result of an experiment yet to be carried out.

Introduction (cont.)

Example (Uncertain states of nature)

- the outcome of a coin toss;
- the millionth digit of π;

  3.14159265358979323846264338327950288419716939937 5....

- the height of this building;
- the number of people in this room;
- the temperature at noon tomorrow.

Note

The millionth digit of π is a fixed number: however, unless we know its value, we are still in a state of uncertainty when asked to assess it, as we have a lack of perfect knowledge.

Introduction (cont.)

Example (Coin tossing)

If I take a coin and decide to carry out an experiment where I toss the coin once and see which face is turned upward when the coin comes to rest, then I can assess that there are two possible results of the toss:





The outcome is of the toss is uncertain *before* I toss the coin, and *after* I toss the coin *until I see the result.*

Example (Thumbtack tossing)

If I take a thumbtack and decide to carry out an experiment where I toss the thumbtack once, then I can assess that there are two possible results of the toss:





The outcome of the toss is uncertain *before* I toss the thumbtack, and *after* I toss the thumbtack *until I see the result.*

Chapter 1: The basics of probability

What is probability ?

By probability, we generally mean the chance of a particular event occurring, given a particular set of circumstances. The probability of an event is generally expressed as a quantitative measurement.

What is probability ?

That is,

(i) the chance of a particular event occurring ...

What is probability ?

That is,

(i) the chance of a particular event occurring ...

(ii) ... given a particular set of circumstances.

What is probability ?

It has seemed to me that the theory of probabilities ought to serve as the basis for the study of all the sciences.

Adolphe Quetelet

The probable is what usually happens.

Aristotle

Probability is the very guide of life.

Cicero, *De Natura*, 5, 12

`http://probweb.berkeley.edu/quotes.html`

What is probability ? (cont.)

My thesis, paradoxically, and a little provocatively, but nonetheless genuinely, is simply this:

PROBABILITY DOES NOT EXIST

The abandonment of superstitious beliefs about the existence of the Phlogiston, the Cosmic Ether, Absolute Space and Time, ... or Fairies and Witches was an essential step along the road to scientific thinking.

Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs.

B. de Finetti, *Theory of Probability*, 1970.

What is probability ? (cont.)

That is,

- probability is not some *objectively* defined, universal quantity – it has a *subjective* interpretation.
- this subjective element may be commonly agreed upon in certain circumstances, but in general is always present.
- recall
  - coin example,
  - thumbtack example.

Constructing the mathematical framework

General strategy for constructing the probability framework: for a given 'experiment', we

- identify the possible outcomes;
- assign numerical values to collections of possible values to represent the chance that they will coincide with the actual outcome;
- lay out the rules for how the numerical values can be manipulated.

Basic concepts in set theory

A set S is a collection of individual elements, such as s. We write

$$s \in S$$

to mean "s is one of the elements of S".

- $S = \{0, 1\}$; $0 \in S, 1 \in S$.
- $S = \{\text{'dog'}, \text{'cat'}, \text{'mouse'}, \text{'elephant'}\}$; 'elephant' $\in S$
- $S = [0, 1]$ (that is, the closed interval from zero to one); $0.3428 \in S$

Basic concepts in set theory (cont.)

S may be

- *finite* if it contains a finite number of elements;
- *countable* if it contains a countably infinite number of elements;
- *uncountable* if it contains an uncountably infinite number of elements.

Basic concepts in set theory (cont.)

A is a *subset* of S if it contains some, all, or none of the elements of S

$$\text{some: } A \subset S \qquad \text{some or all: } A \subseteq S.$$

That is, A is a subset of S if every element of A is also an element of S

$$s \in A \quad \Longrightarrow \quad s \in S.$$

- $S = \{0, 1\}$; $A = \{0\}$, or $A = \{1\}$, or $A = \{0, 1\}$.
- $S = \{\text{'dog'}, \text{'cat'}, \text{'mouse'}, \text{'elephant'}\}$; $A = \{\text{'cat'}, \text{'mouse'}\}$
- $S = [0, 1]$; $A = (0.25, 0.3)$.

Special case: the *empty* set, $\emptyset$, is a subset that contains no elements.

Basic concepts in set theory (cont.)

Note

In probability, it is necessary to think of *sets of subsets* of S. For example:

- $S = \{1, 2, 3, 4\}$.
- Consider all the subsets of S:

  One element: $\{1\}, \{2\}, \{3\}, \{4\}$

  Two elements: $\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}$

  Three elements: $\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{2, 3, 4\}$

  Four elements: $\{1, 2, 3, 4\}$.

- Add to this list the empty set (zero elements), $\emptyset$.

Thus there is a collection of 16 (that is, 2^4) subsets of S.

Basic concepts in set theory (cont.)

Note

This is a bit trickier if S is an interval ...

Set operations

To manipulate sets, we use three basic operations: we may focus on the case where two sets A and B are subsets of a set S.

- intersection: the *intersection* of two sets A and B is the collection of elements that are elements of both A and B.

$$s \in A \cap B \qquad \Longleftrightarrow \qquad s \in A \text{ and } s \in B.$$

Note

We have for any $A \subseteq S$ that

$$A \cap \emptyset = \emptyset$$

and

$$A \cap S = A.$$

Set operations (cont.)

Note

It is clear that

$$A \cap B \subseteq A$$

and

$$A \cap B \subseteq B.$$

Set operations (cont.)

- union: the *union* of two sets A and B is the set of distinct elements that are either in A, or in B, or in both A and B.

$$s \in A \cup B \qquad \Longleftrightarrow \qquad s \in A \text{ or } s \in B \text{ (or } s \in A \cap B).$$

Note

We have for any $A \subseteq S$ that

$$A \cup \emptyset = A$$

and

$$A \cup S = S.$$

Set operations (cont.)

Note

It is clear that

$$A \subseteq A \cup B$$

and

$$B \subseteq A \cup B.$$

Set operations (cont.)

- complement: the *complement* of set A, a subset of S, is the collection of elements of S that are not elements of A.

$$s \in A' \qquad \Longleftrightarrow \qquad s \in S \text{ but } s \notin A.$$

We have that

$$A \cap A' = \emptyset \qquad \text{and} \qquad A \cup A' = S.$$

Set operations (cont.)

Note

The notations $\overline{A}$ and A^c are also used for the complement of A.

The textbook uses

$$\overline{A} \qquad \overline{(A \cap B)}$$

etc., but I find this leads to confusion.

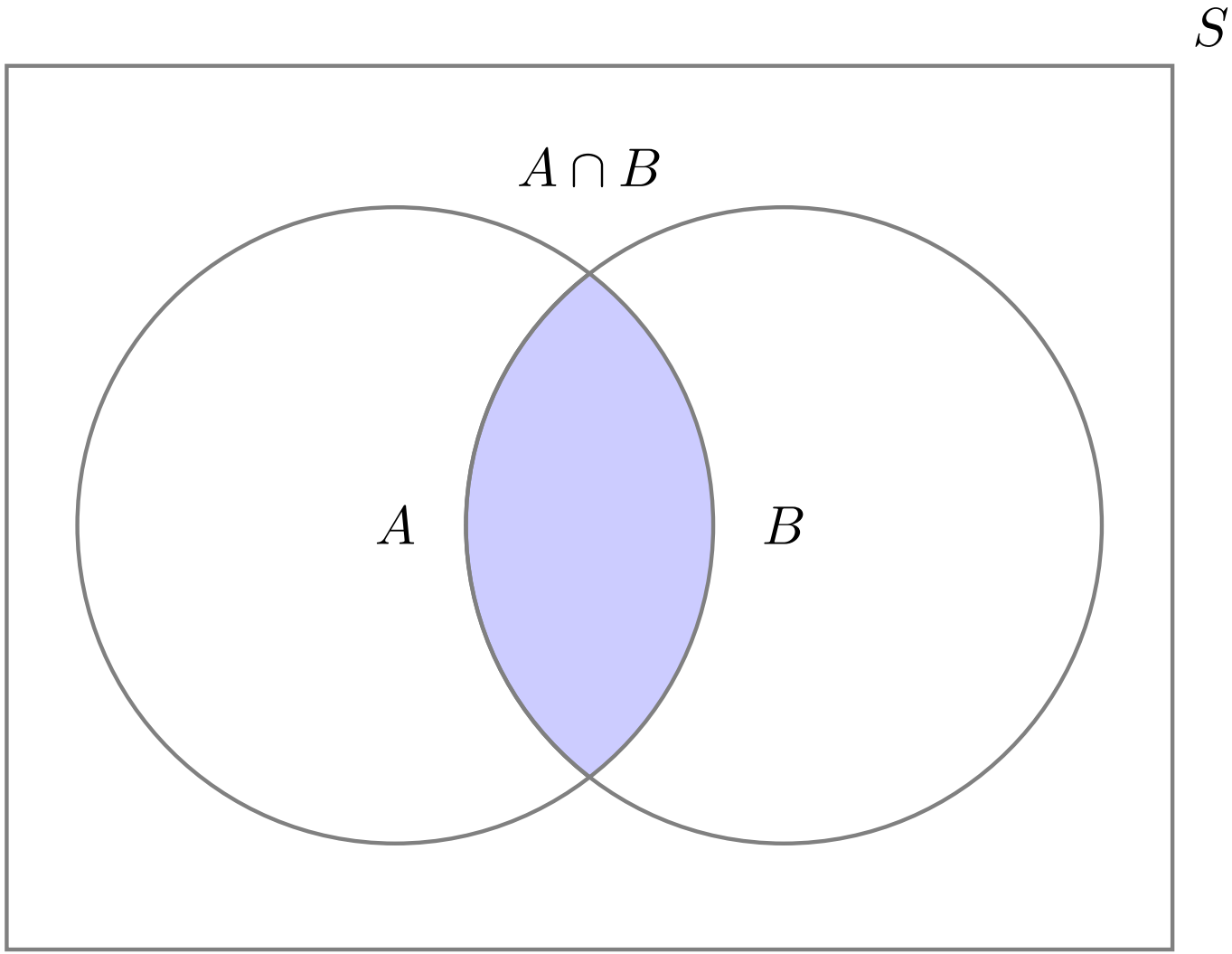
Note

We have for any $A \subseteq S$ that

$$(A')' = A.$$

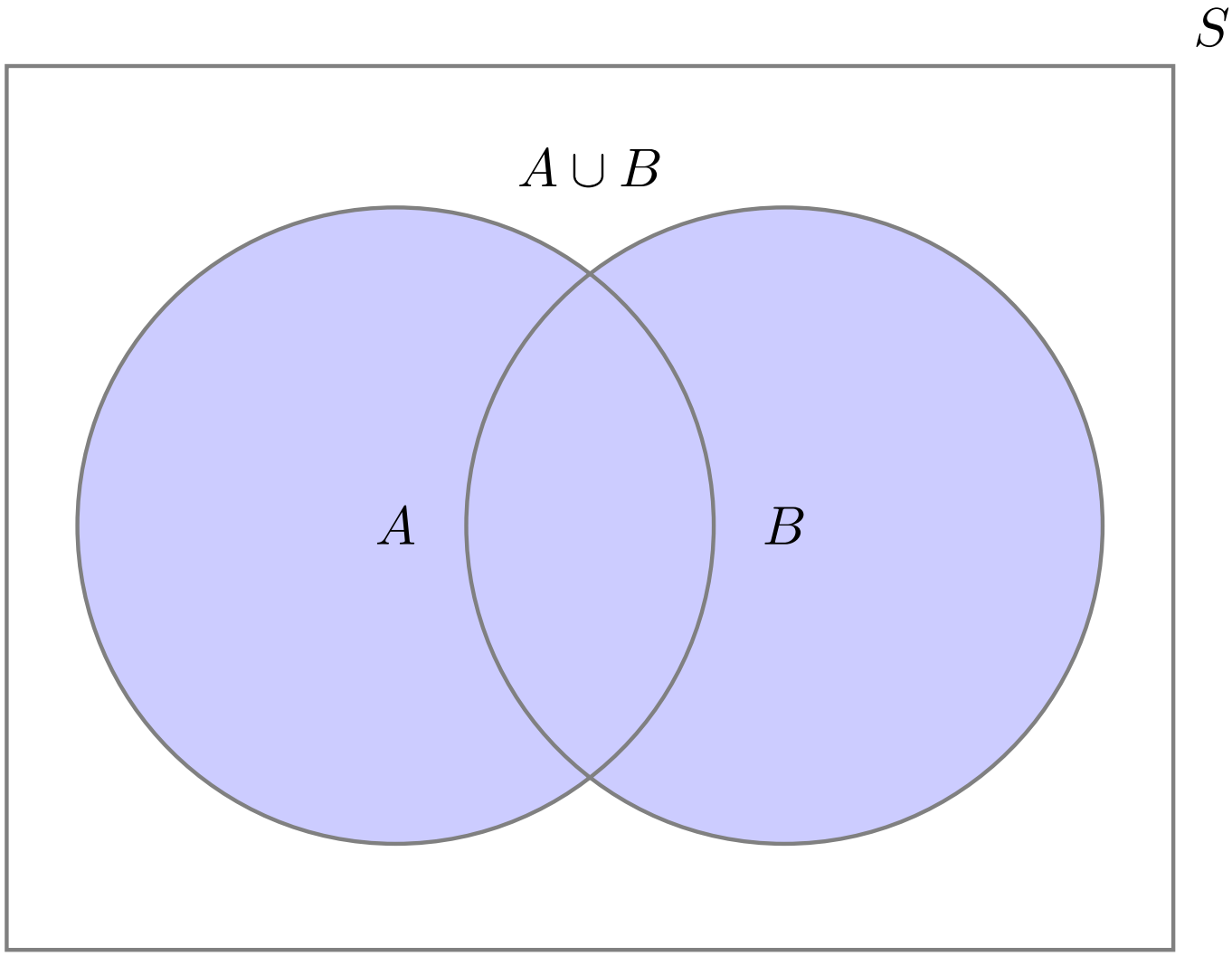
Set operations: in pictures

Intersection:



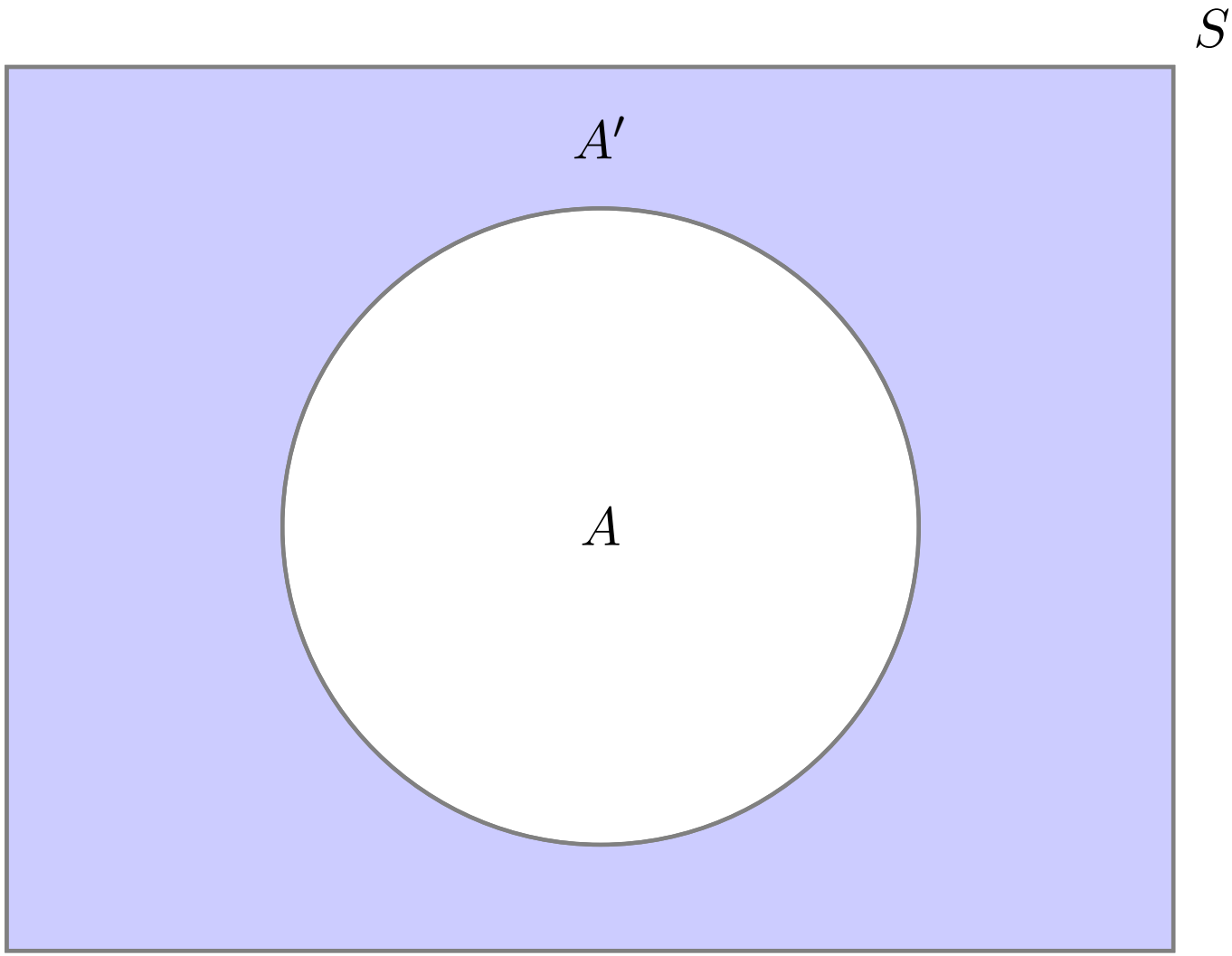
Set operations: in pictures (cont.)

Union:



Set operations: in pictures (cont.)

Complement:



Example

Example (Finite set)

- $S = \{1, 2, 3, \ldots, 9, 10\}$
- $A = \{2, 4, 6\}$
- $B = \{1, 2, 5, 7, 9\}$

Then

- $A \cap B = \{2\}$
- $A \cup B = \{1, 2, 4, 5, 6, 7, 9\}$
- $A' = \{1, 3, 5, 7, 8, 9, 10\}$

Set operations: extensions

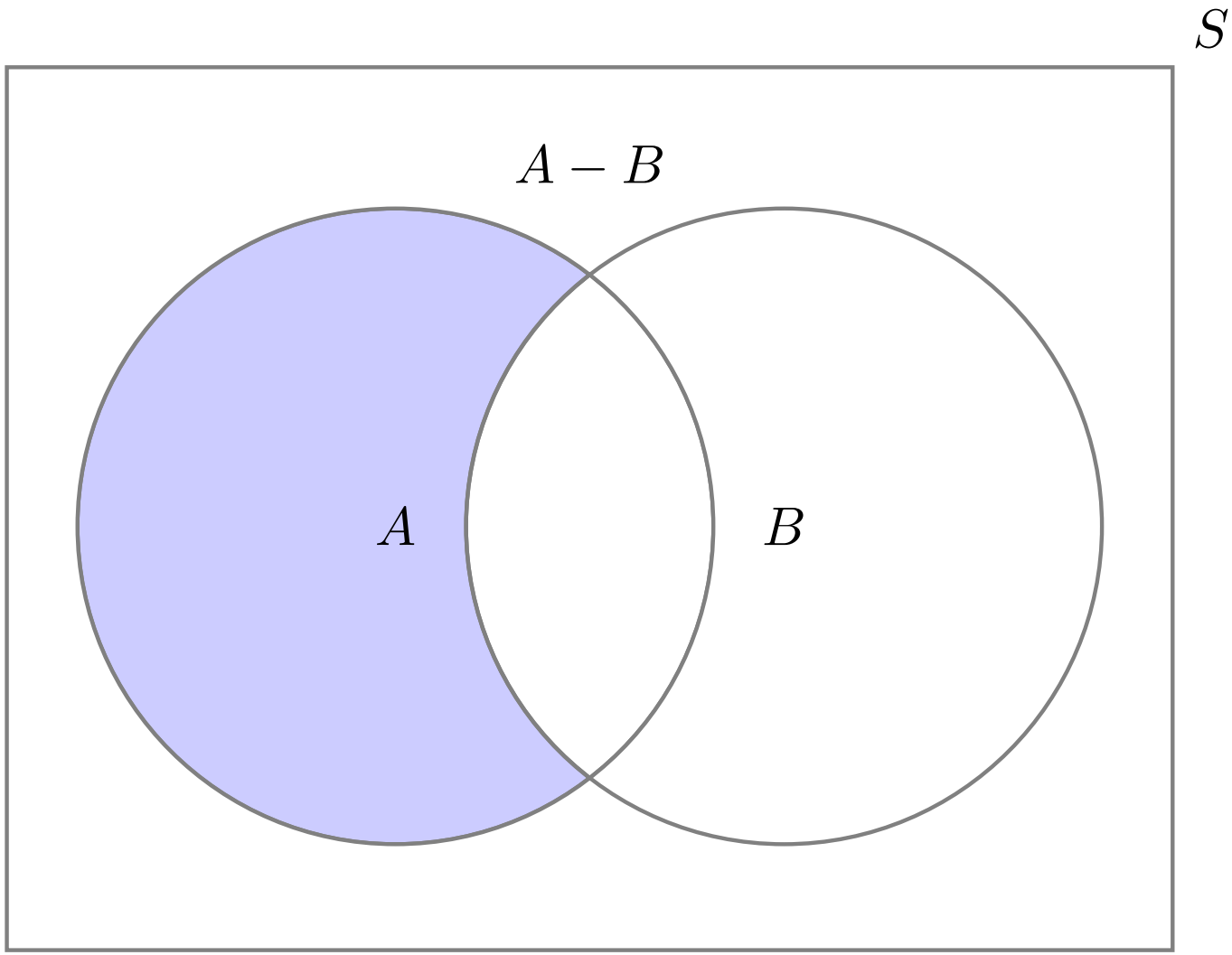
- Set difference: $A - B$ (or $A \backslash B$)

$$s \in A - B \quad \Longleftrightarrow \quad s \in A \text{ and } s \in B'$$

that is

$$A - B \equiv A \cap B'.$$

Set operations: extensions (cont.)



Set operations: extensions (cont.)

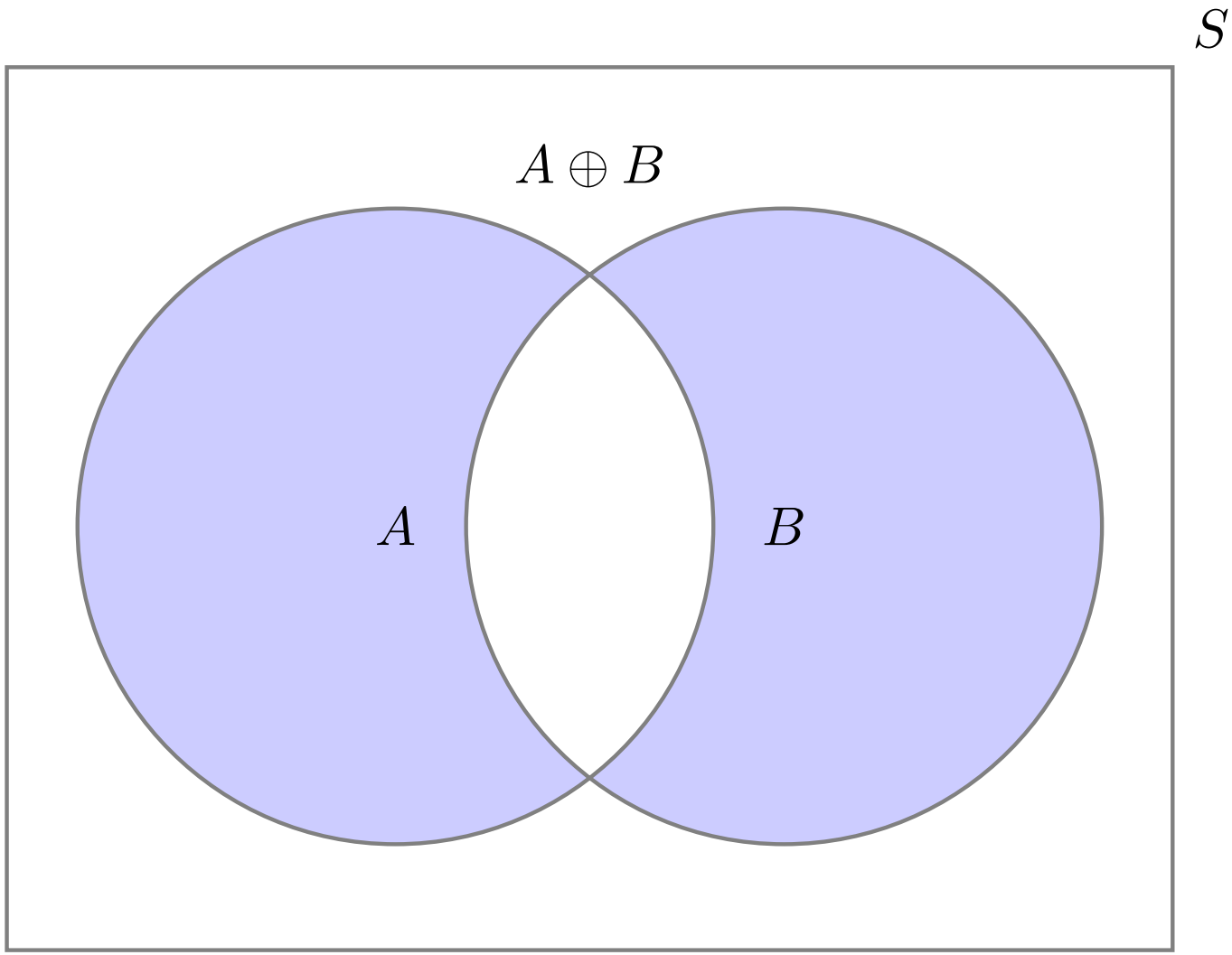
- A or B but not both: $A \oplus B$

$$s \in A \oplus B \quad \Longleftrightarrow \quad s \in A \text{ or } s \in B, \text{ but } s \notin A \cap B$$

aka

$$(A \cup B) - (A \cap B)$$

Set operations: extensions (cont.)



Combining set operations

Both intersection and union operators are *binary* operators, that is, take two arguments:

$$A \cap B \qquad A \cup B.$$

We have immediately from the definitions that

$$A \cap B = B \cap A$$

and that

$$A \cup B = B \cup A$$

(that is, the order does not matter).

Combining set operations (cont.)

However, A and B are arbitrary, so suppose there is a third set $C \subseteq S$, and consider

$$(A \cap B) \cap C$$

as $A \cap B$ is a set in its own right. Then

$$\begin{aligned} s \in (A \cap B) \cap C \quad &\iff \quad s \in (A \cap B) \text{ and } s \in C \\ &\iff \quad s \in A \text{ and } s \in B \text{ and } s \in C. \end{aligned}$$

Combining set operations (cont.)

That is

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

and by the same logic

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

and so on.

Combining set operations (cont.)

This extends to the case of any finite collection of sets

$$A_1, A_2, \ldots, A_K$$

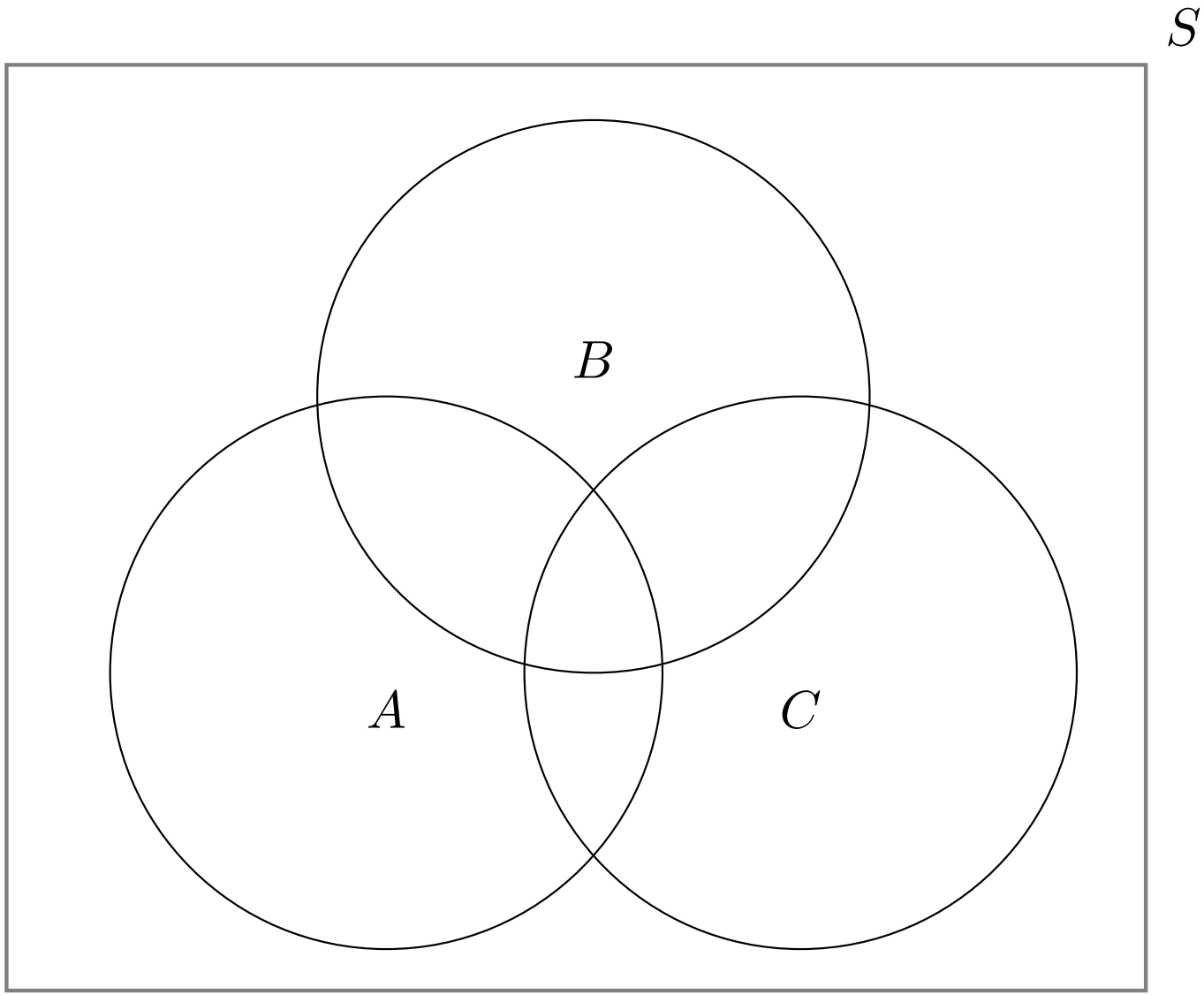
and we may write

$$A_1 \cap A_2 \cap \ldots \cap A_K \equiv \bigcap_{k=1}^{K} A_k$$

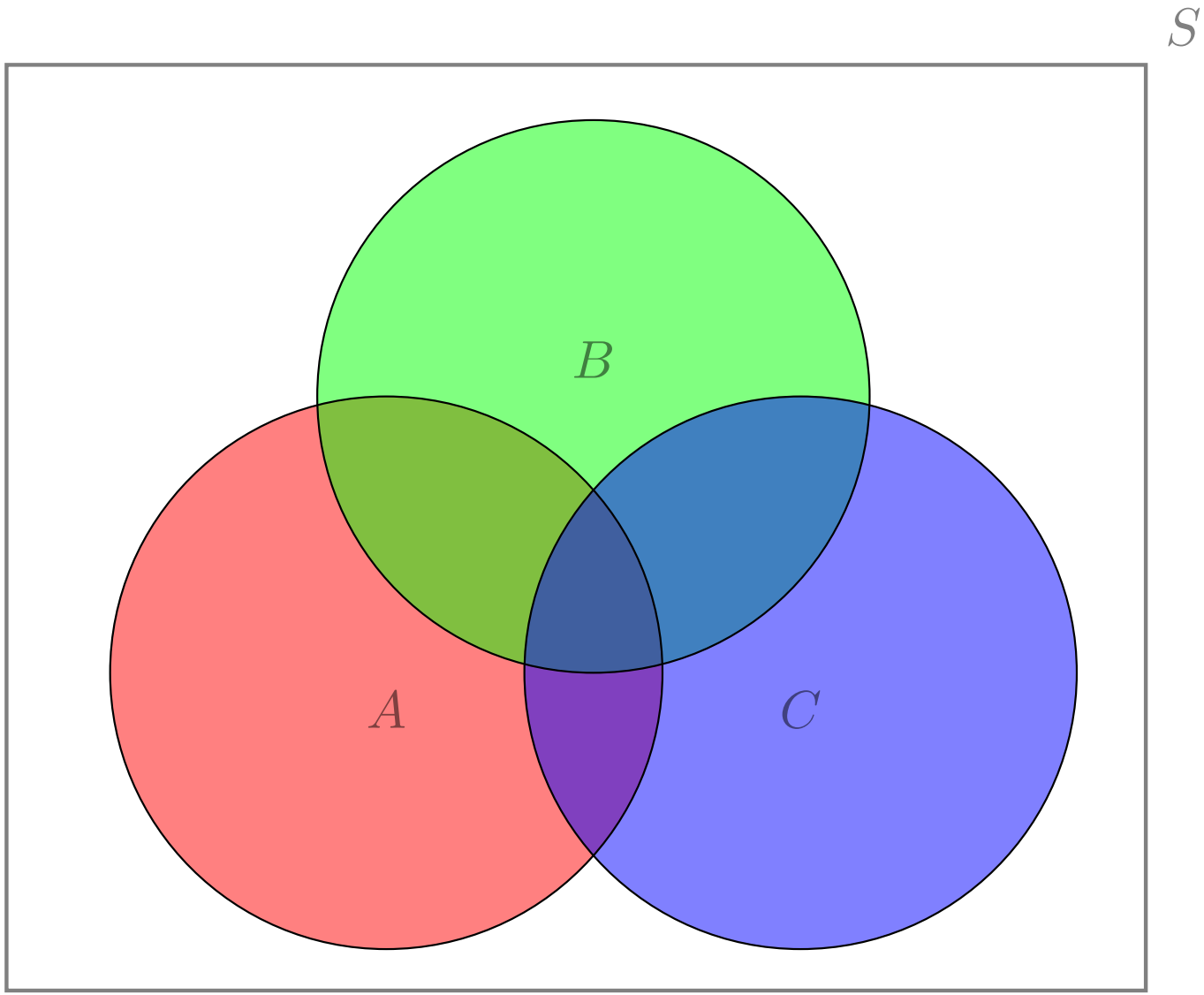
where

$$s \in \bigcap_{k=1}^{K} A_k \qquad \Longleftrightarrow \qquad s \in A_k \text{ for all } k$$

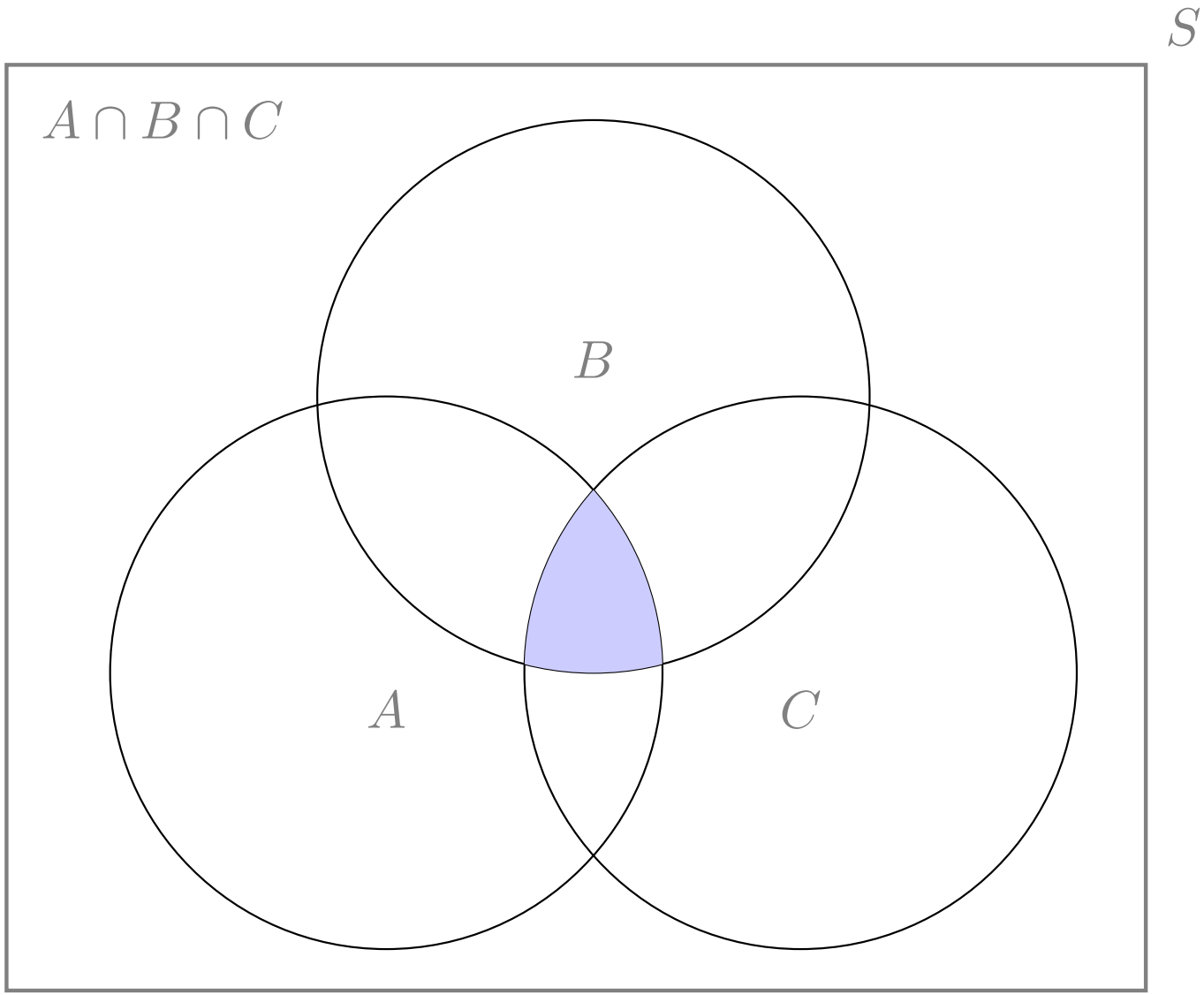
Combining set operations (cont.)



Combining set operations (cont.)



Combining set operations (cont.)



Combining set operations (cont.)

The same logic holds for the union operator:

$$(A \cup B) \cup C = A \cup B \cup C = A \cup (B \cup C)$$

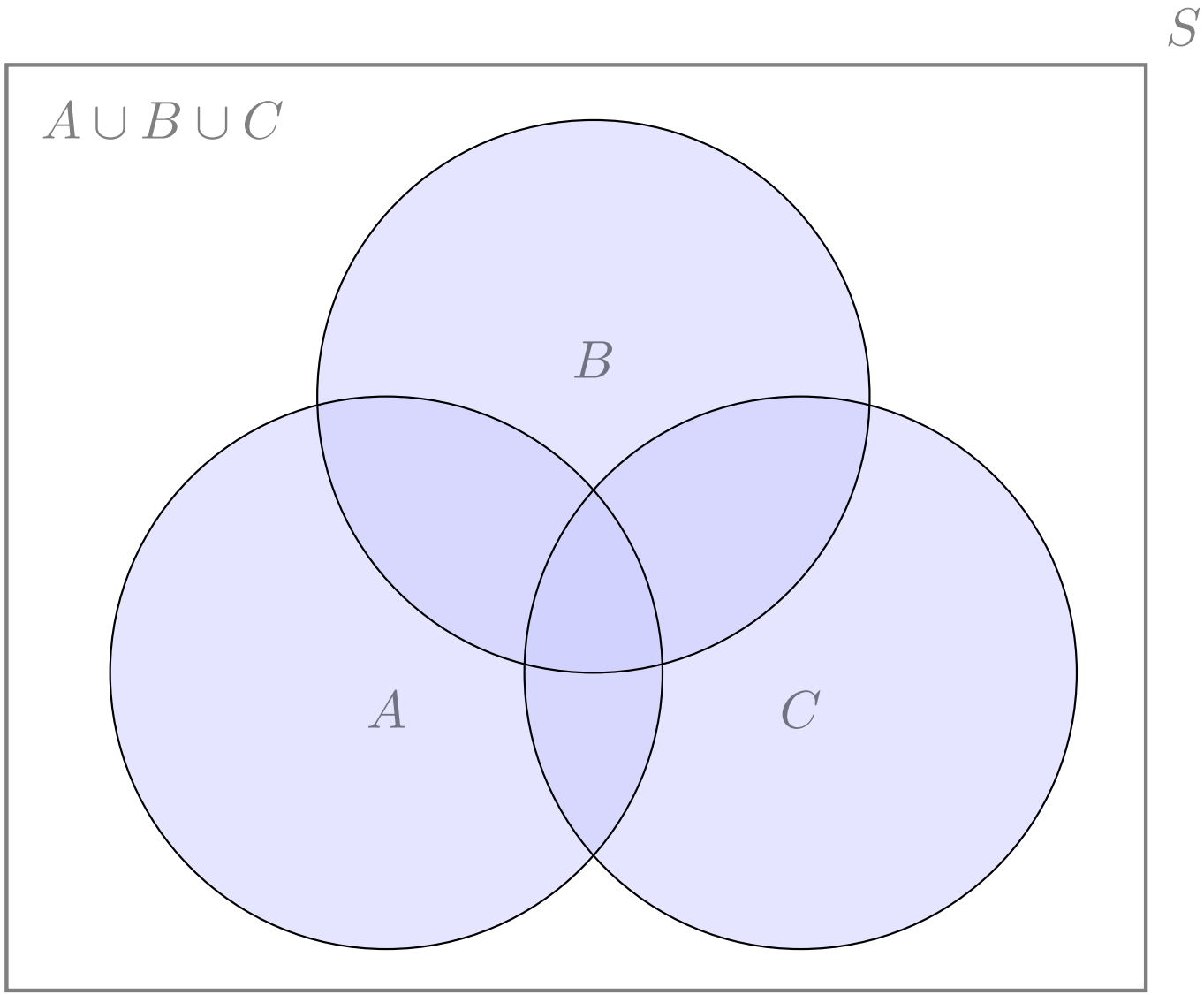
and

$$A_1 \cup A_2 \cup \ldots \cup A_K \equiv \bigcup_{k=1}^{K} A_k$$

where

$$s \in \bigcup_{k=1}^{K} A_k \qquad \Longleftrightarrow \qquad s \in A_k \text{ for at least one } k$$

Combining set operations (cont.)



Combining set operations (cont.)

Note

The intersection and union operations can be extended to work with a countably infinite number of sets

$$A_1, A_2, \ldots, A_k, \ldots$$

and we can consider

$$\bigcap_{k=1}^{\infty} A_k \quad : \quad s \in \bigcap_{k=1}^{\infty} A_k \iff s \in A_k \text{ for all } k$$

$$\bigcup_{k=1}^{\infty} A_k \quad : \quad s \in \bigcup_{k=1}^{\infty} A_k \iff s \in A_k \text{ for at least one } k$$

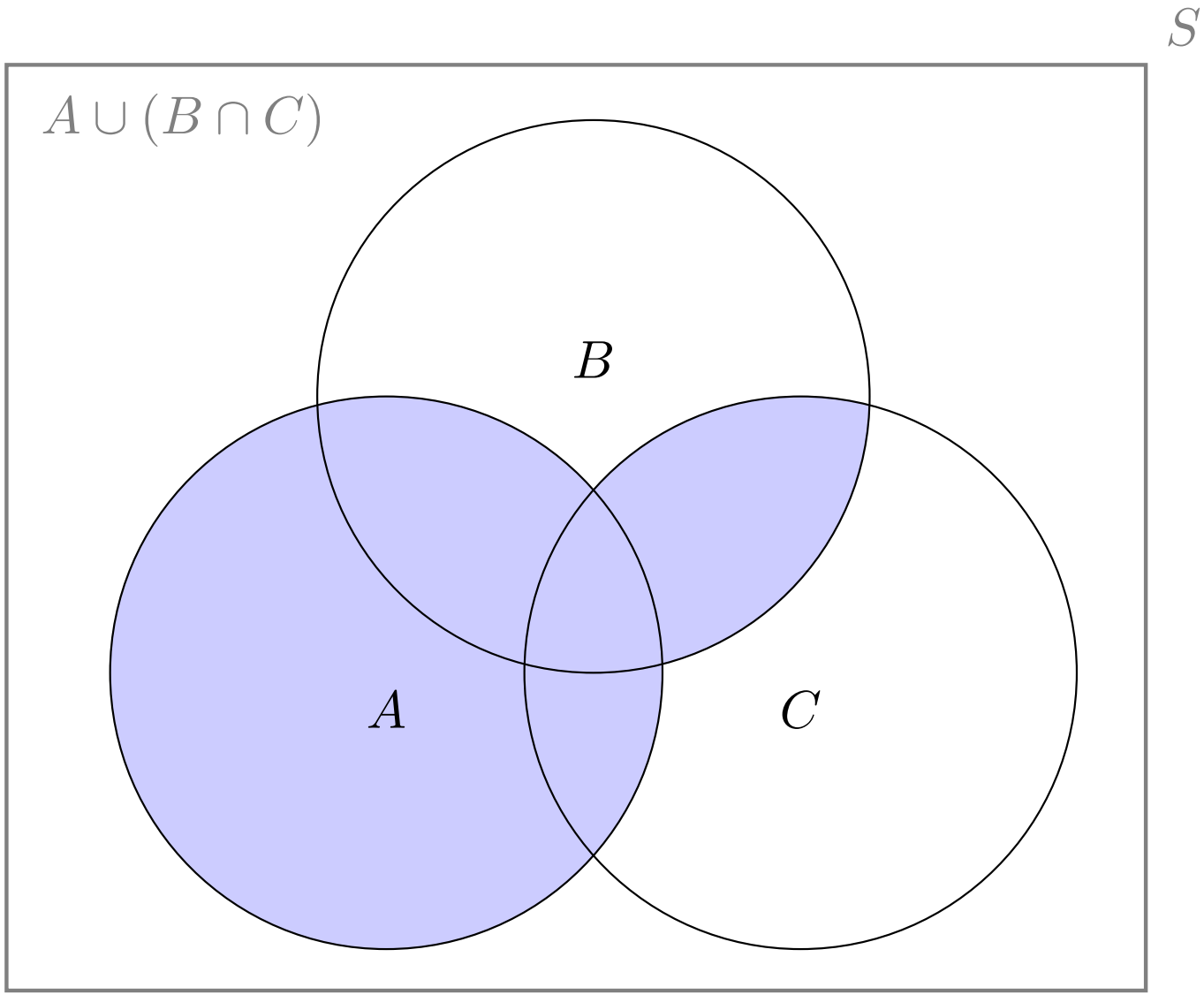
Combining set operations (cont.)

Consider now

$$A \cup (B \cap C)$$

To be an element in this set, you need to be an element of A or an element of $B \cap C$ (or both).

Combining set operations (cont.)



Combining set operations (cont.)

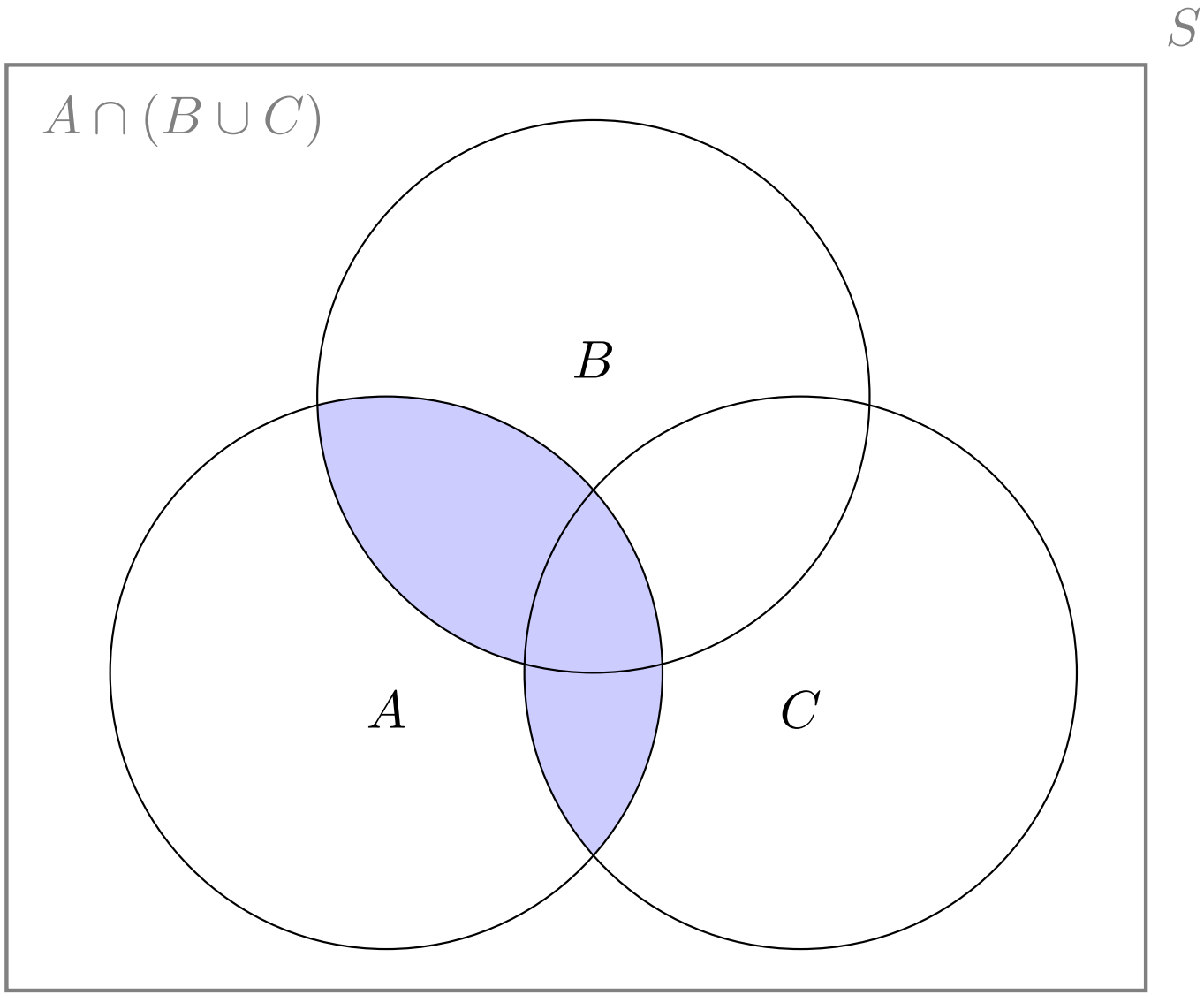
We can therefore deduce that

$$A \cup (B \cap C) \equiv (A \cup B) \cap (A \cup C).$$

Similarly

$$A \cap (B \cup C) \equiv (A \cap B) \cup (A \cap C).$$

Combining set operations (cont.)



Partitions

We call $A_1, A_2, \ldots, A_K$ a *partition* of S if these sets are

- pairwise disjoint (or mutually exclusive):

$$A_j \cap A_k = \emptyset \quad \text{for all } j \neq k$$

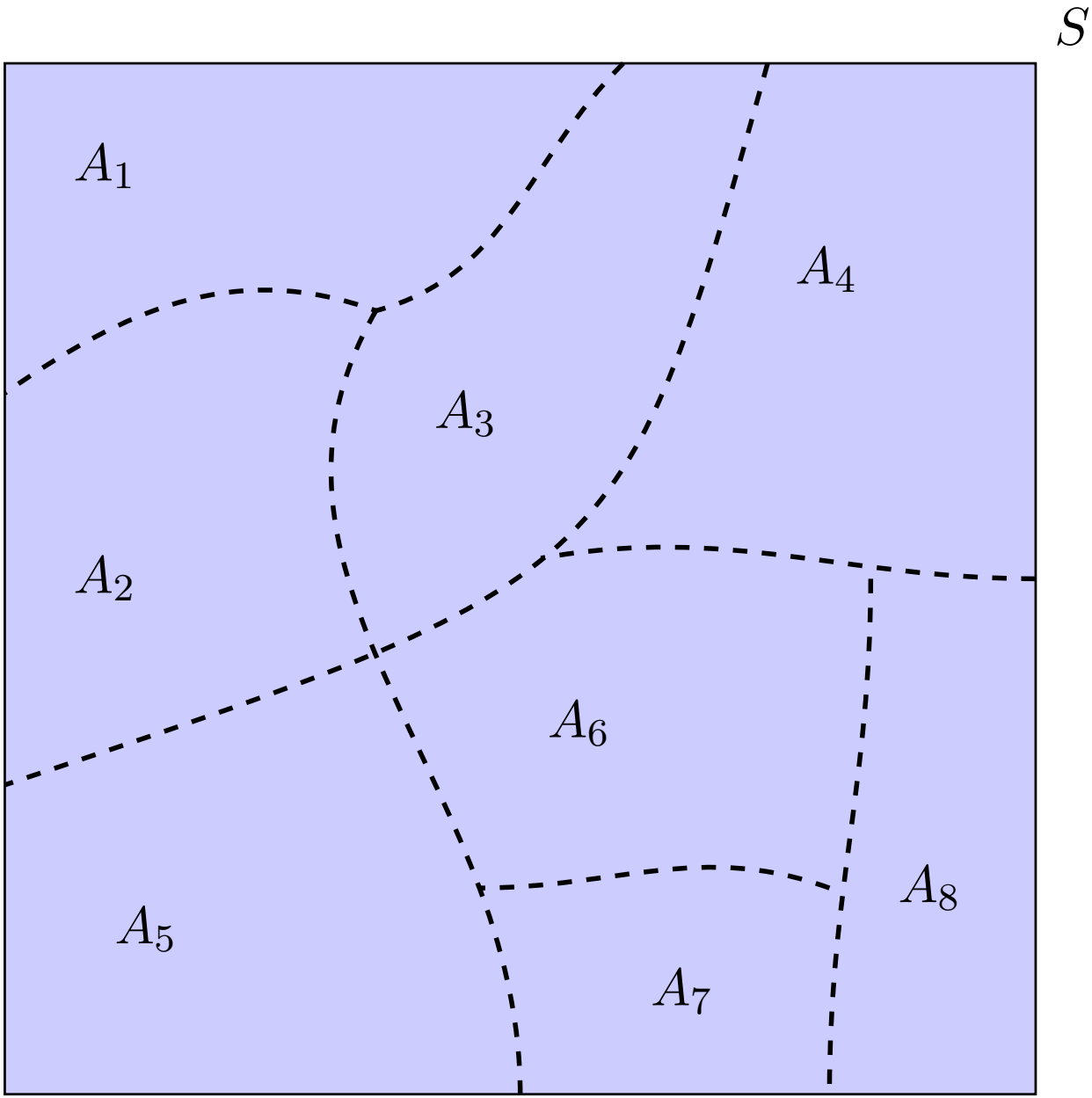
- exhaustive:

$$\bigcup_{k=1}^{K} A_k = S.$$

(that is, the As cover the whole of S, but do not overlap).

For every $s \in S$, s is an element of precisely one of the A_k.

Partitions (cont.)

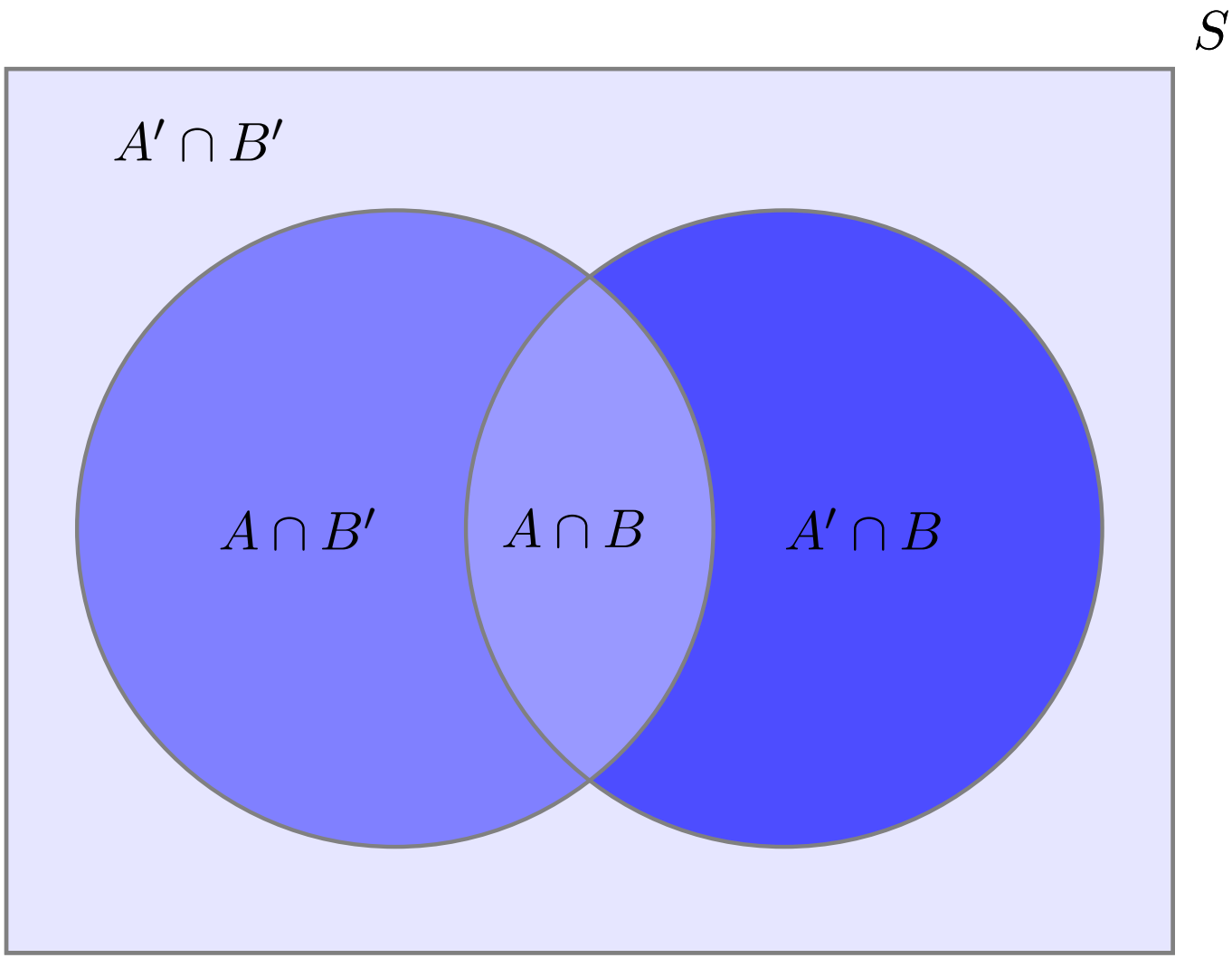


Some simple theorems

For any two subsets A and B of S, we have the partition of S via the four sets

$$A \cap B \qquad A \cap B' \qquad A' \cap B \qquad A' \cap B'$$

Some simple theorems (cont.)



Some simple theorems (cont.)

This picture implies that

$$(A \cup B)' = A' \cap B'.$$

Some simple theorems (cont.)

We can prove this as follows: let

$$A_1 = A \cup B \qquad\qquad A_2 = A' \cap B'.$$

We need to show that

$$A_1' = A_2$$

that is,

(i) $A_1 \cap A_2 = \emptyset$;

(ii) $A_1 \cup A_2 = S$.

Some simple theorems (cont.)

(i) Disjoint:

$$\begin{aligned} A_1 \cap A_2 &= (A \cup B) \cap A_2 \\ &= (A \cap A_2) \cup (B \cap A_2) \\ &= (A \cap A' \cap B') \cup (B \cap A' \cap B') \\ &= \emptyset \cup \emptyset \\ &= \emptyset \end{aligned}$$

Some simple theorems (cont.)

(ii) Exhaustive:

$$\begin{aligned} A_1 \cup A_2 &= A_1 \cup (A' \cap B') \\ &= (A_1 \cup A') \cap (A_1 \cup B') \\ &= (A \cup B \cup A') \cap (A \cup B \cup B') \\ &= S \cap S \\ &= S \end{aligned}$$

Some simple theorems (cont.)

Similarly

$$A' \cup B' = (A \cap B)'$$

which is equivalent to saying

$$(A' \cup B')' = A \cap B.$$

Some simple theorems (cont.)

The first result

$$(A \cup B)' = A' \cap B'.$$

holds for arbitrary sets A and B. In particular it holds for the sets

$$C = A' \qquad\qquad D = B'$$

that is we have

$$(C \cup D)' = C' \cap D'.$$

Some simple theorems (cont.)

But

$$C' = (A')' = A \qquad\qquad D' = (B')' = B$$

Therefore

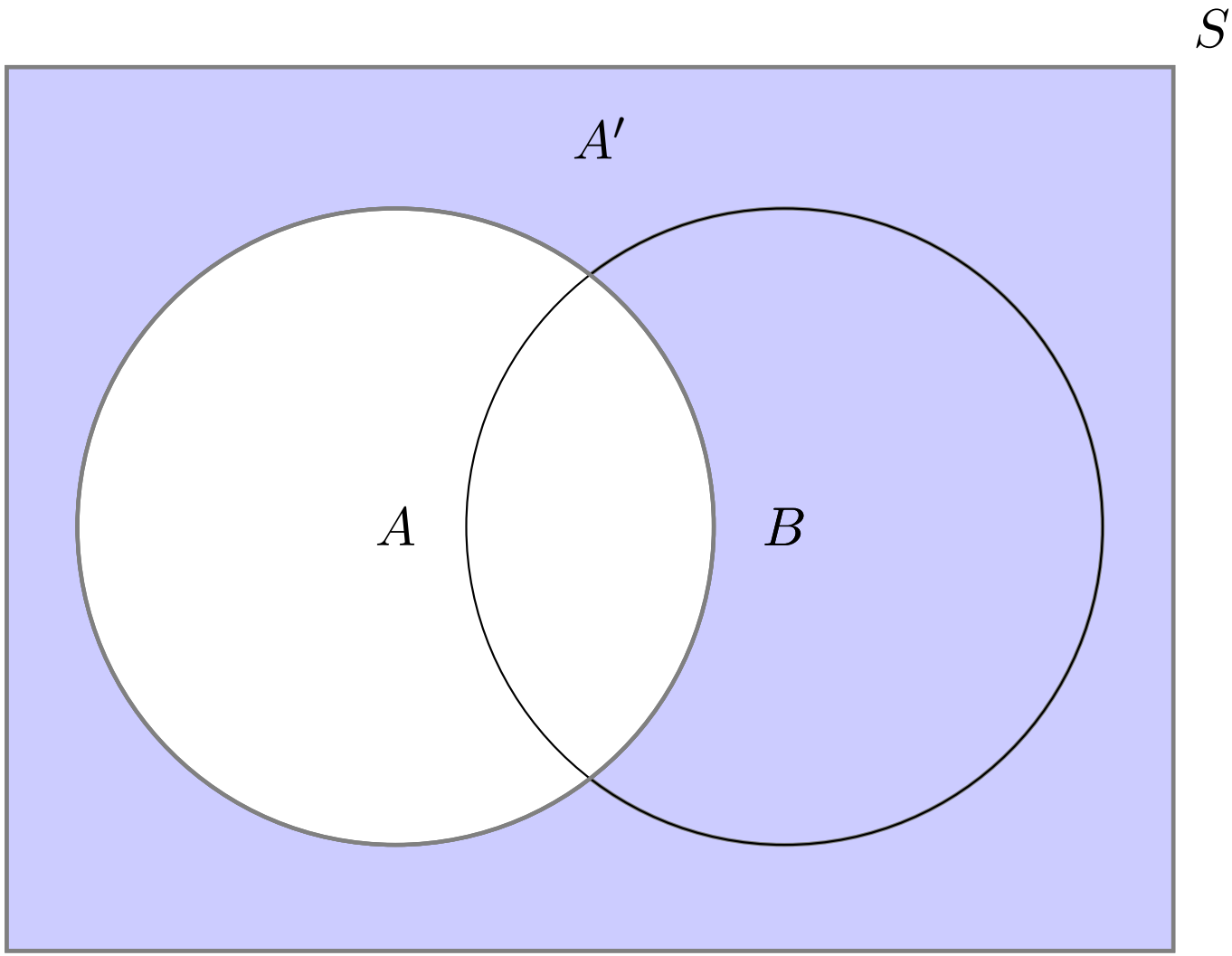
$$(A' \cup B')' = A \cap B$$

or equivalently

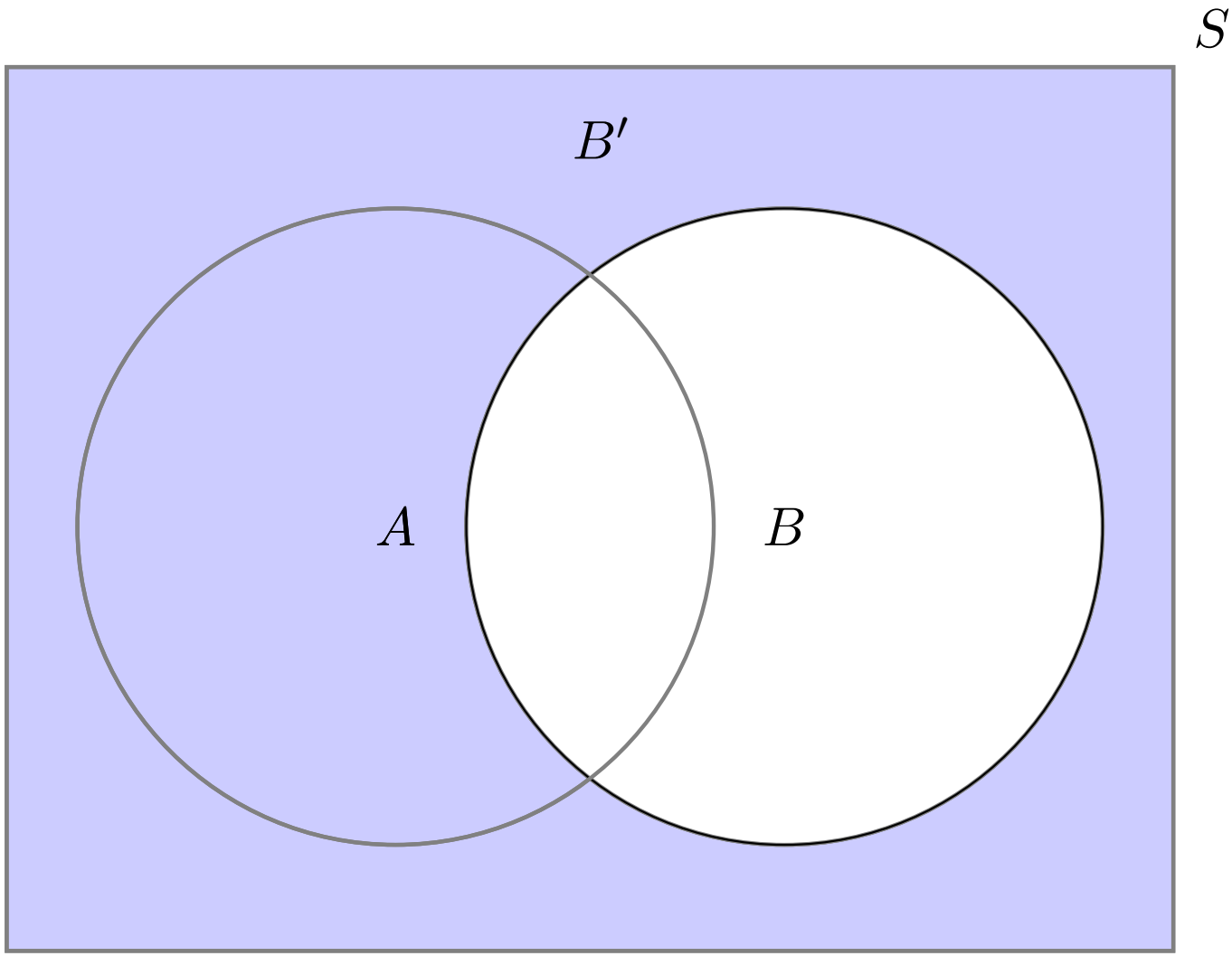
$$A' \cup B' = (A \cap B)'$$

as required.

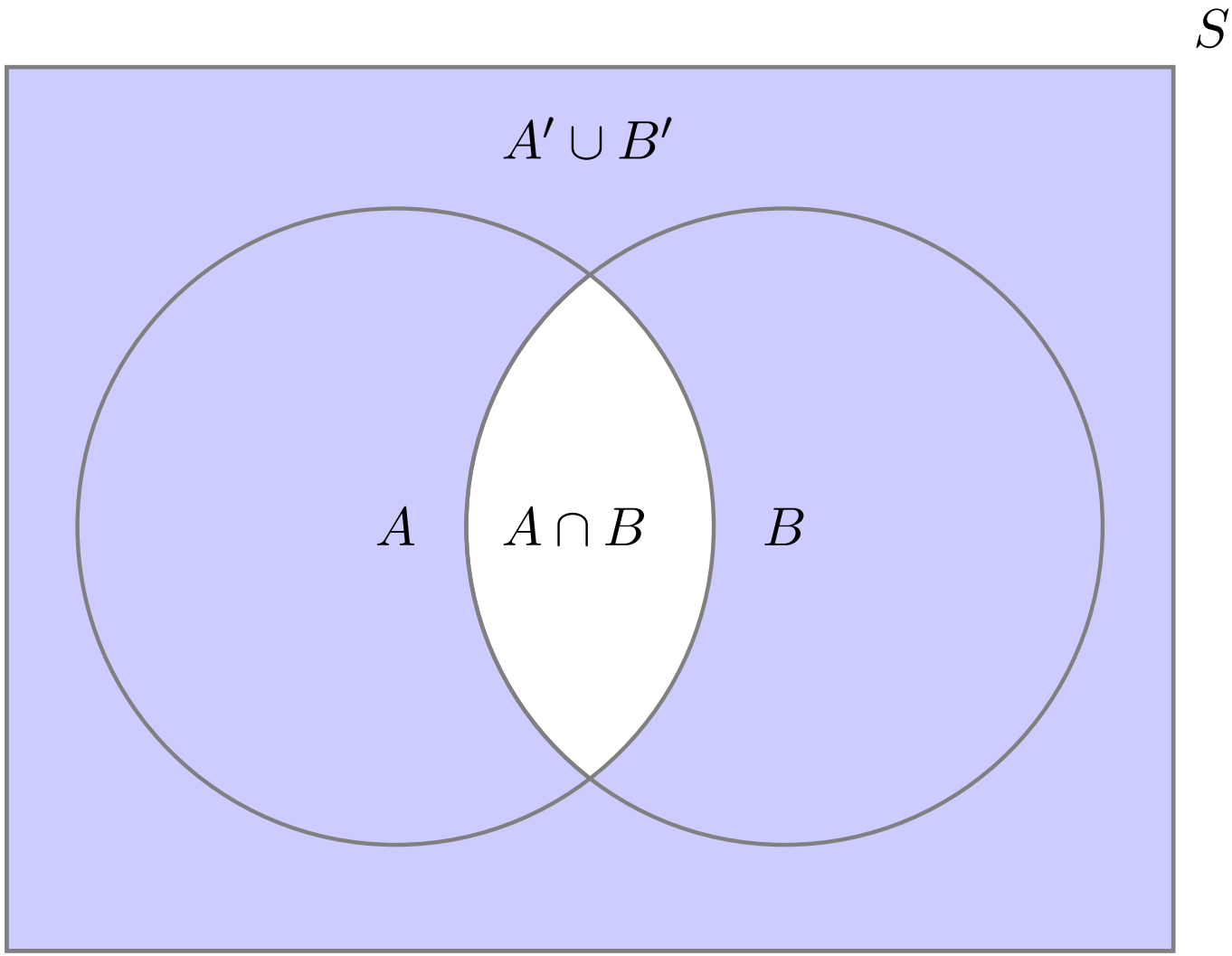
Some simple theorems (cont.)



Some simple theorems (cont.)



Some simple theorems (cont.)



Some simple theorems (cont.)

These two results

$$A' \cap B' = (A \cup B)'$$

$$A' \cup B' = (A \cap B)'$$

are sometimes known as de Morgan's Laws.

Sample spaces and Events

We now utilize the set theory formulation and notation in the probability context.

Recall the earlier informal definition:

> *By probability, we generally mean the chance of a particular event occurring, given a particular set of circumstances. The probability of an event is generally expressed as a quantitative measurement.*

We need to carefully define what an 'event' is, and what constitutes a 'particular set of circumstances'.

Sample spaces and Events (cont.)

We will consider the general setting of an *experiment*:

- this can be interpreted as any setting in which an uncertain consequence is to arise;
- could involve observing an outcome, taking a measurement etc.

1. **Consider the possible outcomes of the experiment:** Make a 'list' of the outcomes that can arise, and denote the corresponding set by S.
   - $S = \{0, 1\}$;
   - $S = \{\text{'head'}, \text{'tail'}\} \equiv \{\text{H}, \text{T}\}$;
   - $S = \{\text{'cured'}, \text{'not cured'}\}$;
   - $S = \{\text{'Arts'}, \text{'Engineering'}, \text{'Medicine'}, \text{'Science'}\}$;
   - $S = \{1, 2, 3, 4, 5, 6\}$;
   - $S = \mathbb{R}^+$.

   The set S is termed the *sample space* of the experiment. The individual elements of S are termed *sample points* (or *sample outcomes*).

Constructing the framework (cont.)

2. **Events**: An *event* A is a collection of sample outcomes. That is, A is a subset of S,

$$A \subseteq S.$$

For example

- $A = \{0\}$;
- $A = \{\text{'tail'}\} \equiv \{\text{T}\}$;
- $A = \{\text{'cured'}\}$;
- $A = \{\text{'Arts'}, \text{'Engineering'}\}$;
- $A = \{1, 3, 5\}$;
- $A = [2, 3)$.

The individual sample outcomes are termed *simple* (or *elementary*) events, and may be denoted

$$E_1, E_2, \ldots, E_K, \ldots.$$

Constructing the framework (cont.)

3. **Terminology**: We say that event A *occurs* if the actual outcome, s, is an element of A.

   For two events A and B

   - $A \cap B$ occurs if and only if A occurs and B occurs, that is

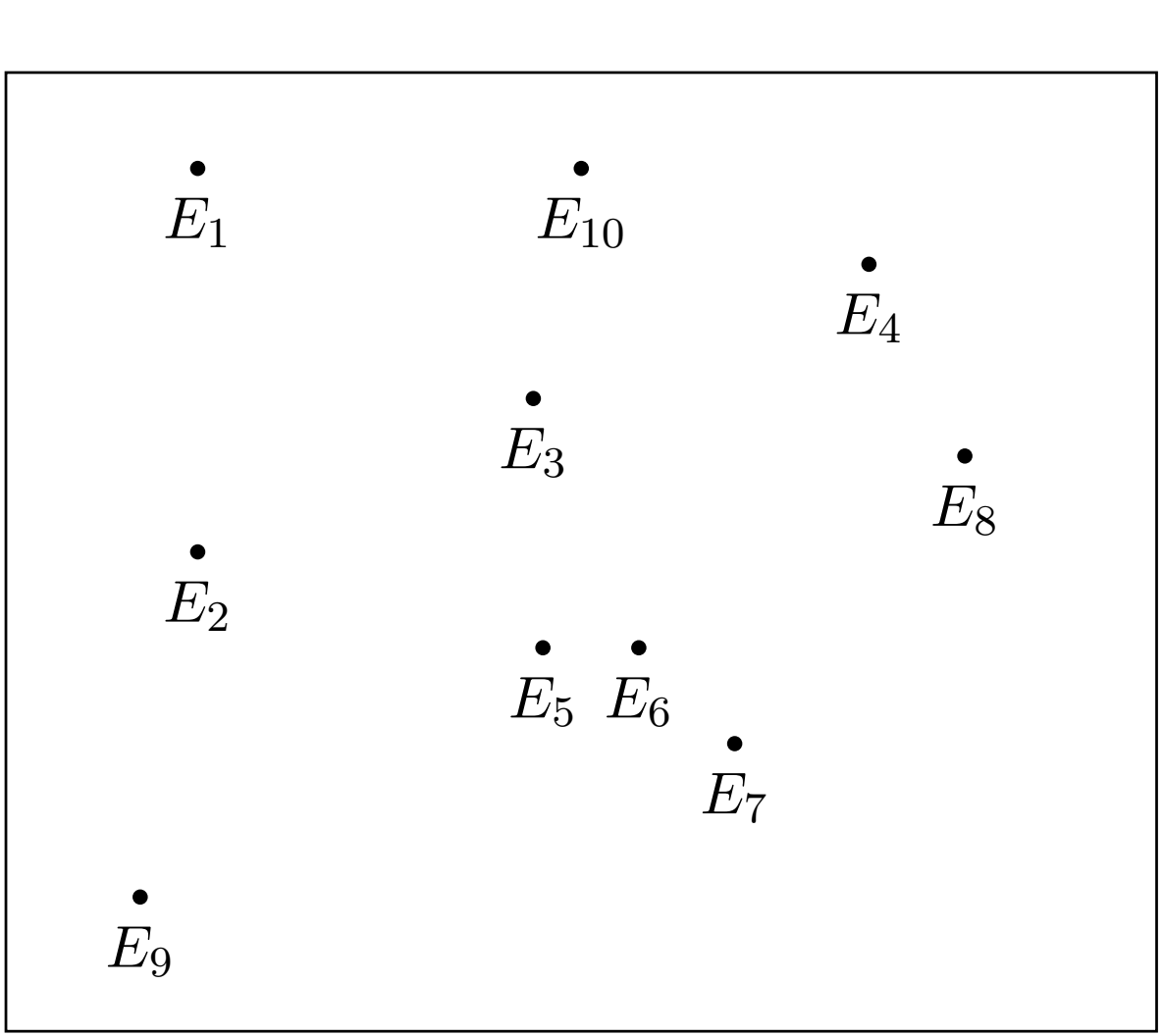
     $$s \in A \cap B.$$

   - $A \cup B$ occurs if A occurs or if B occurs, or if both A and B occur, that is

     $$s \in A \cup B$$

   - if A occurs, then A' does not occur.

Constructing the framework (cont.)

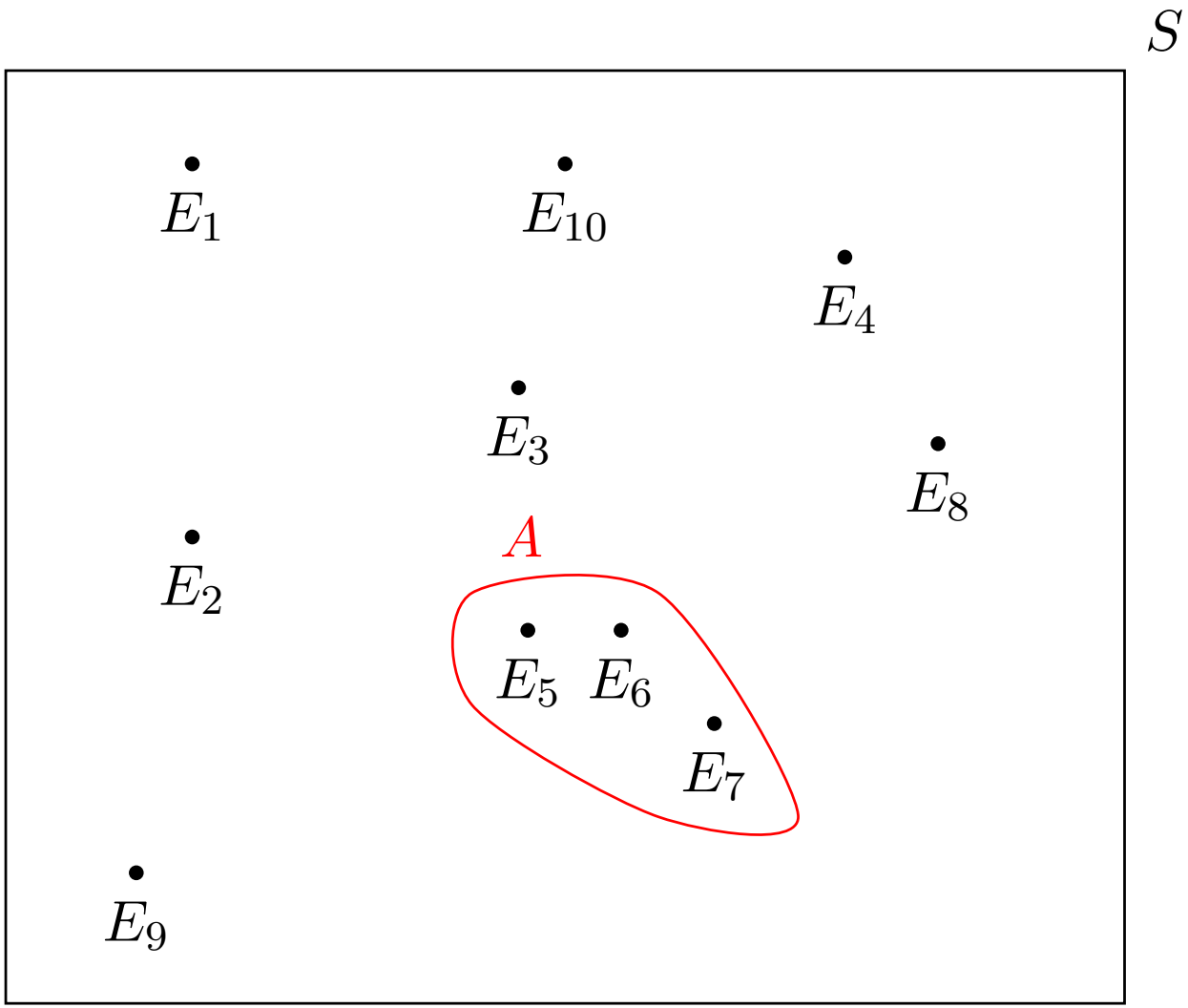


Constructing the framework (cont.)

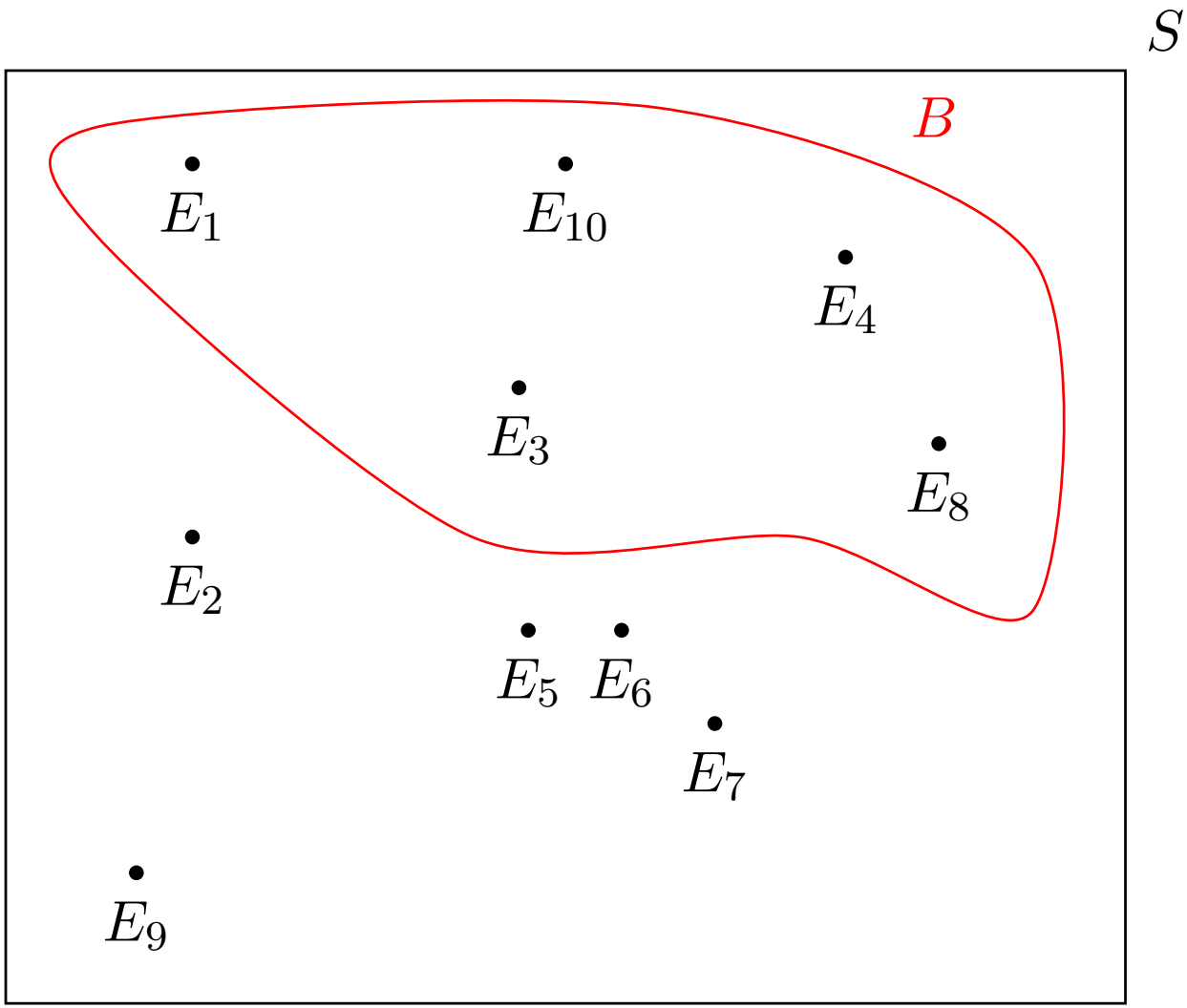
In this case

$$S = \bigcup_{k=1}^{10} E_k.$$

Constructing the framework (cont.)



Constructing the framework (cont.)



Constructing the framework (cont.)

In this case

$$A = E_5 \cup E_6 \cup E_7$$

and

$$B = E_1 \cup E_3 \cup E_4 \cup E_8 \cup E_{10}.$$

Constructing the framework (cont.)

Note

- The event S is termed the certain event;
- The event $\emptyset$ is termed the impossible event.

Mathematical definition of probability

Probability is a means of assigning a quantitative measure of uncertainty to events in a sample space.

Formally, the *probability function*, $P(.)$ is a function that assigns numerical values to events.

$$P : \mathcal{A} \longrightarrow \mathbb{R}$$

$$A \longmapsto p$$

that is

$$P(A) = p$$

that is, the probability assigned to event A (a set) is p (a numerical value).

Mathematical definition of probability (cont.)

Note

(i) The function $P(.)$ is a set function (that is, it takes a set as its argument).

(ii) $\mathcal{A}$ is a "set of subsets of S"; we pick a subset $A \in \mathcal{A}$ and consider its probability.

(iii) $\mathcal{A}$ has certain nice properties that ensure that the probability function can operate successfully.

Mathematical definition of probability (cont.)

This definition is too general:

- what properties does $P(.)$ have ?
- how does $P(.)$ assign numerical values; that is, how do we compute

  $$P(A)$$

  for a given event A in sample space S ?

The probability axioms

Suppose S is a sample space for an experiment, and A is an event, a subset of S. Then we assign $P(A)$, the probability of event A, so that the following axioms hold:

(I) $P(A) \geq 0$.

(II) $P(S) = 1$.

(III) If $A_1, A_2, \ldots$ form a (countable) sequence of events such that

$$A_j \cap A_k = \emptyset \qquad \text{for all } j \neq k$$

then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

that is, we say that $P(.)$ is *countably additive*.

The probability axioms (cont.)

Note

Axiom (III) immediately implies that $P(.)$ is *finitely additive*, that is, for all n, $1 \leq n < \infty$, if $A_1, A_2, \ldots, A_n$ form a sequence of events such that $A_j \cap A_k = \emptyset$ for all $j \neq k$, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

To see this, fix n and define $A_i = \emptyset$ for $i > n$.

The probability axioms (cont.)

Example (Partitions)

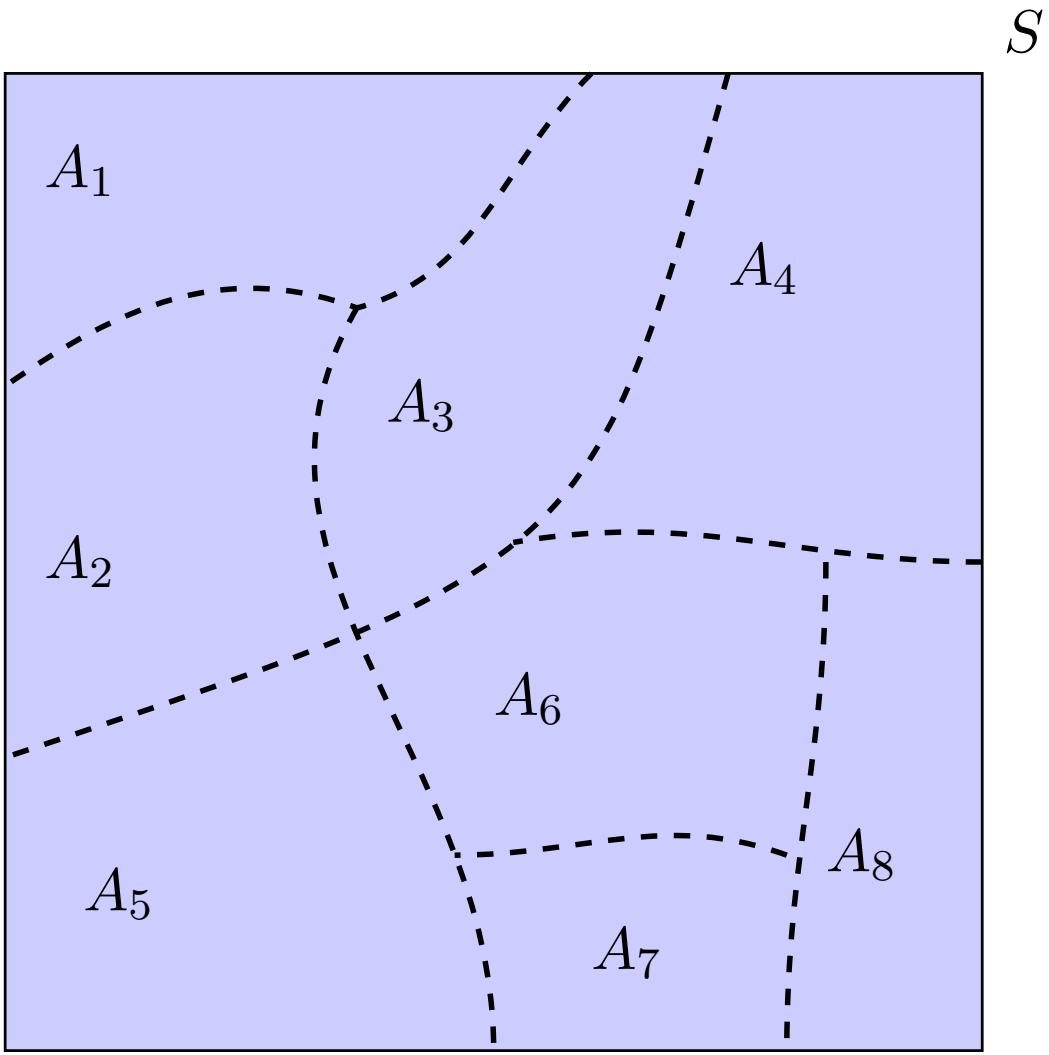
If $A_1, A_2, \ldots, A_n$ form a partition of S, and

$$P(A_i) = p_i \qquad i = 2, \ldots, n$$

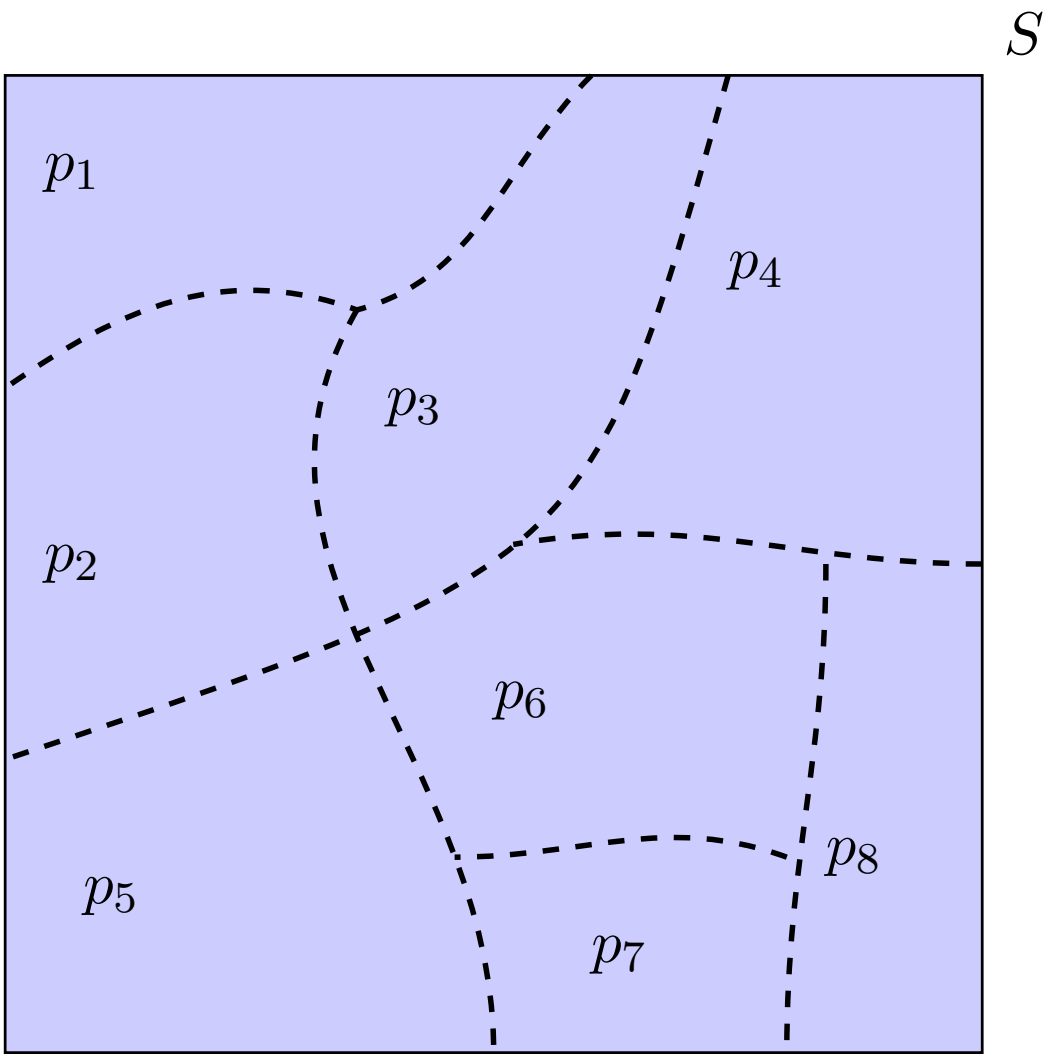
say, then

$$\sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{n} p_i = 1.$$

The probability axioms (cont.)



The probability axioms (cont.)



The probability axioms (cont.)

Some immediate corollaries of the axioms:

(i) For any A, $P(A') = 1 - P(A)$.

- We have that $S = A \cup A'$. By Axiom (III) we have therefore that
$$P(S) = P(A) + P(A').$$
- By Axiom (II), $P(S) = 1$, so therefore
$$1 = P(A) + P(A') \qquad \therefore \qquad P(A') = 1 - P(A).$$

The probability axioms (cont.)

(ii) $P(\emptyset) = 0$.

- Apply the result from point (i) to the set $A \equiv S$.
- Note that by Axiom (II), $P(S) = 1$.

The probability axioms (cont.)

(iii) For any A, $P(A) \leq 1$.

- By Axiom (III) and point (i) we have

$$1 = P(S) = P(A) + P(A') \geq P(A)$$

as $P(A') \geq 0$ by Axiom (I).

The probability axioms (cont.)

(iv) For any two events A and B, if $A \subseteq B$, then

$$P(A) \leq P(B)$$

- We may write in this case

$$B = A \cup (A' \cap B)$$

and, as the two events on the right hand side are mutually exclusive, by Axiom (III)

$$P(B) = P(A) + P(A' \cap B) \geq P(A)$$

as $P(A' \cap B) \geq 0$ by Axiom (I).

The probability axioms (cont.)

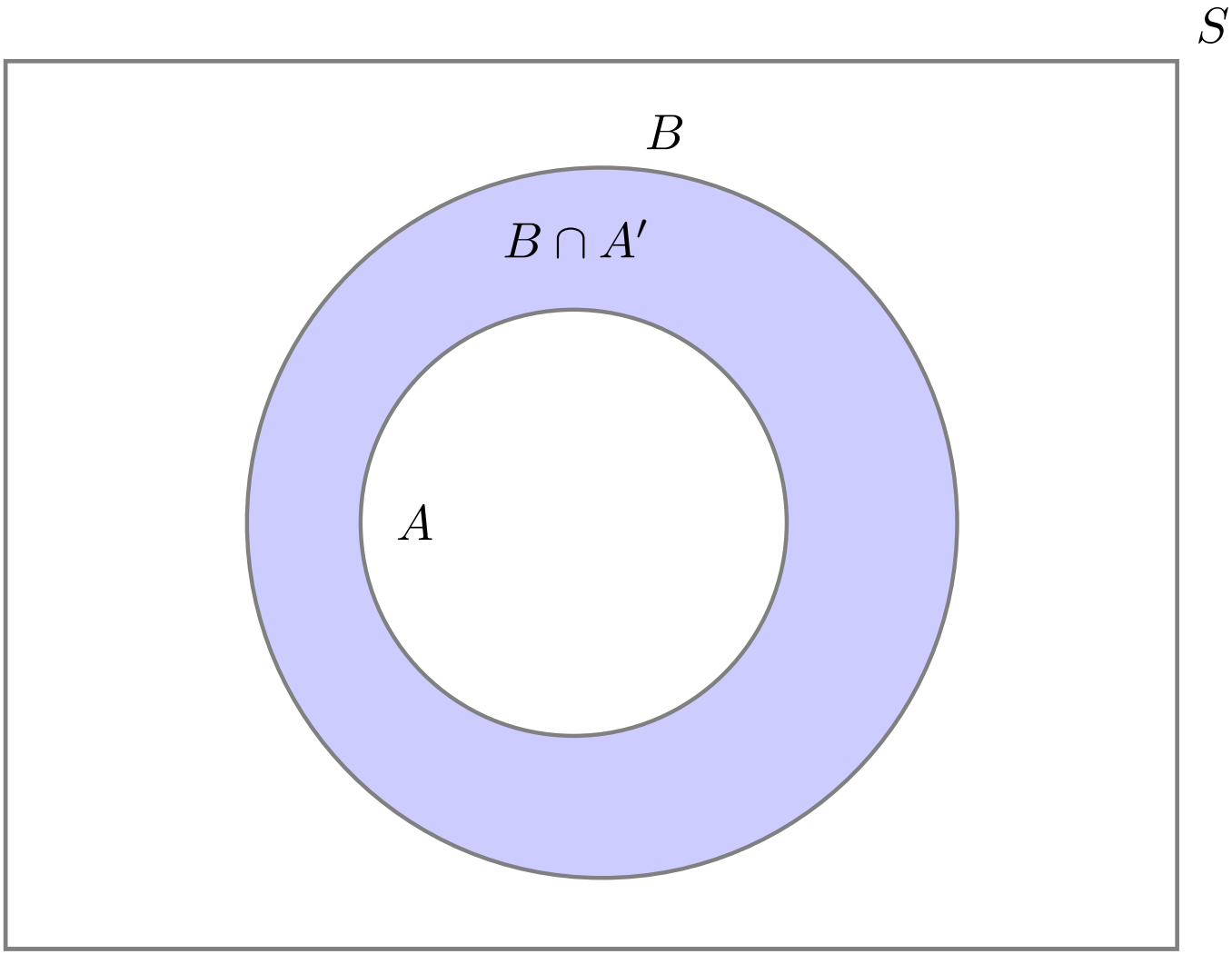
Therefore, for example,

$$P(A \cap B) \leq P(A) \qquad \text{and} \qquad P(A \cap B) \leq P(B)$$

that is

$$P(A \cap B) \leq \min\{P(A), P(B)\}.$$

The probability axioms (cont.)

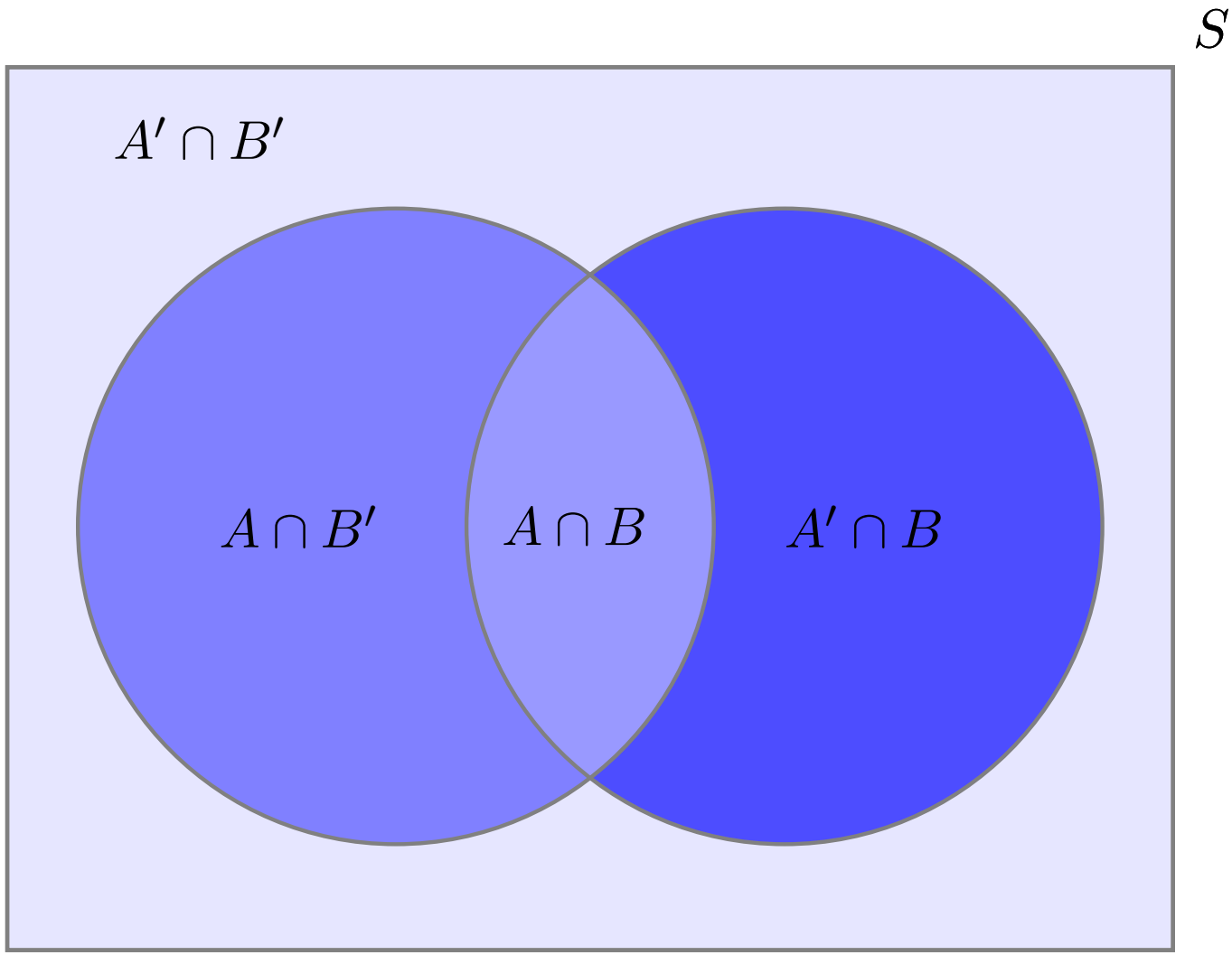


The probability axioms (cont.)

(v) **General Addition Rule:** For two arbitrary events A and B,

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

The probability axioms (cont.)



The probability axioms (cont.)

- We have that

$$(A \cup B) = A \cup (A' \cap B)$$

so that by Axiom (III)

$$P(A \cup B) = P(A) + P(A' \cap B).$$

But

$$B = (A \cap B) \cup (A' \cap B)$$

so that

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

and therefore

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

The probability axioms (cont.)

Substituting back in, we have

$$\begin{aligned} P(A \cup B) &= P(A) + P(A' \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

as required.

- We may deduce from this that

$$P(A \cup B) \leq P(A) + P(B)$$

as $P(A \cap B) \geq 0$.

The probability axioms (cont.)

In general, for events $A_1, A_2, \ldots, A_n$, we can construct a formula for

$$P\left(\bigcup_{i=1}^{n} A_i\right)$$

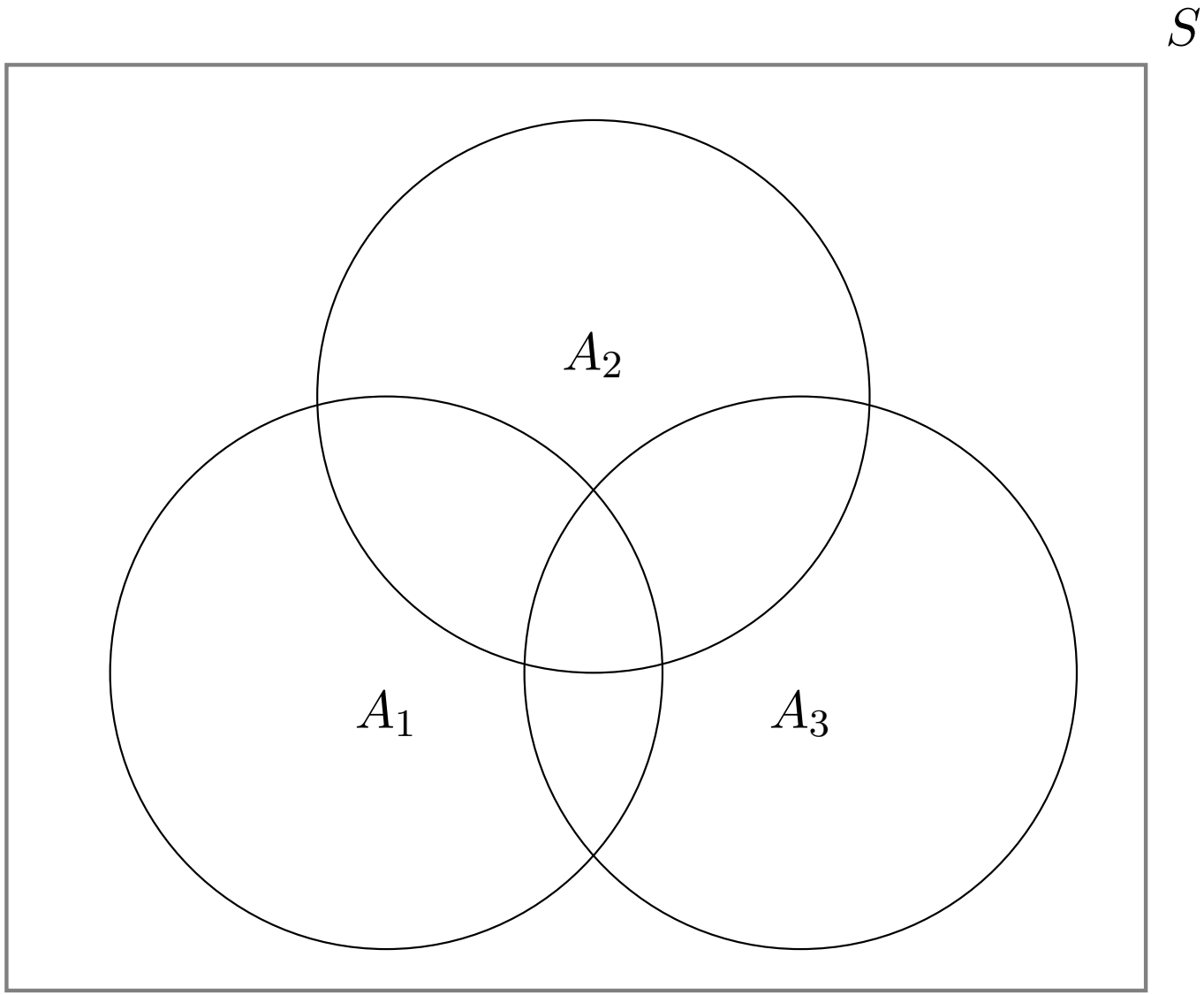
using inductive arguments.

The probability axioms (cont.)

For example, with $n = 3$, we have that

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) = \; & P(A_1) + P(A_2) + P(A_3) \\ & - P(A_1 \cap A_2) \\ & - P(A_1 \cap A_3) \\ & - P(A_2 \cap A_3) \\ & + P(A_1 \cap A_2 \cap A_3). \end{aligned}$$

The probability axioms (cont.)



The probability axioms (cont.)

A more straightforward general result is Boole's Inequality

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P\left(A_i\right).$$

Proof of this result follows by a simple inductive argument.

Probability tables

It is sometimes helpful to lay out probabilities in table form:

	A	A'	Total
B	$p_{A\cap B}$	$p_{A'\cap B}$	p_B
B'	$p_{A\cap B'}$	$p_{A'\cap B'}$	$p_{B'}$
Total	p_A	$p_{A'}$	p_S

that is, for example, $p_{A\cap B} = P(A \cap B)$, and

$$A = (A \cap B) \cup (A \cap B')$$

so that

$$p_A = P(A) = P(A \cap B) + P(A \cap B') = p_{A\cap B} + p_{A\cap B'}$$

and so on.

Probability tables (cont.)

Example

Suppose

$$P(A) = \frac{1}{3} \qquad P(B) = \frac{1}{2} \qquad P(A \cup B) = \frac{3}{4}.$$

	A	A'	Total
B			$\frac{1}{2}$
B'			$\frac{1}{2}$
Total	$\frac{1}{3}$	$\frac{2}{3}$	1

Example

We have

$$\begin{aligned} P(A \cap B) &= P(A) + P(B) - P(A \cup B) \\ &= \frac{1}{3} + \frac{1}{2} - \frac{3}{4} \\ &= \frac{1}{12} \end{aligned}$$

Probability tables (cont.)

Example

Hence

	A	A'	Total
B	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{1}{2}$
B'	$\frac{3}{12}$	$\frac{3}{12}$	$\frac{1}{2}$
Total	$\frac{1}{3}$	$\frac{2}{3}$	1

Probability tables (cont.)

Not all configurations of entries in the table are valid: we need to ensure that

$$0 \leq p_* \leq 1$$

for all entries, with

$$p_A = p_{A \cap B} + p_{A \cap B'}$$

and

$$p_B = p_{A \cap B} + p_{A' \cap B}$$

Specifying probabilities

The earlier discussion defines mathematically how the probability function behaves; it does not tell us how to assign numerical values to the probabilities of events.

It is usually thought that there are three ways via which we could specify the probability of an event.

Specifying probabilities (cont.)

1. **Equally likely sample outcomes:** Suppose that sample space S is finite, with N sample outcomes in total that are considered to be equally likely to occur. Then for the elementary events $E_1, E_2, \ldots, E_N$, we have

$$P(E_i) = \frac{1}{N} \qquad i = 1, \ldots, N$$

Specifying probabilities (cont.)

Every event $A \subseteq S$ can be expressed

$$A = \bigcup_{i=1}^{n} E_{i_A}$$

for some $n \leq N$, and collection $E_{1_A}, \ldots, E_{n_A}$, with indices

$$i_A \in \{1, \ldots, n\}.$$

Then

$$P(A) = \frac{n}{N} = \frac{\text{Number of sample outcomes in } A}{\text{Number of sample outcomes in } S}.$$

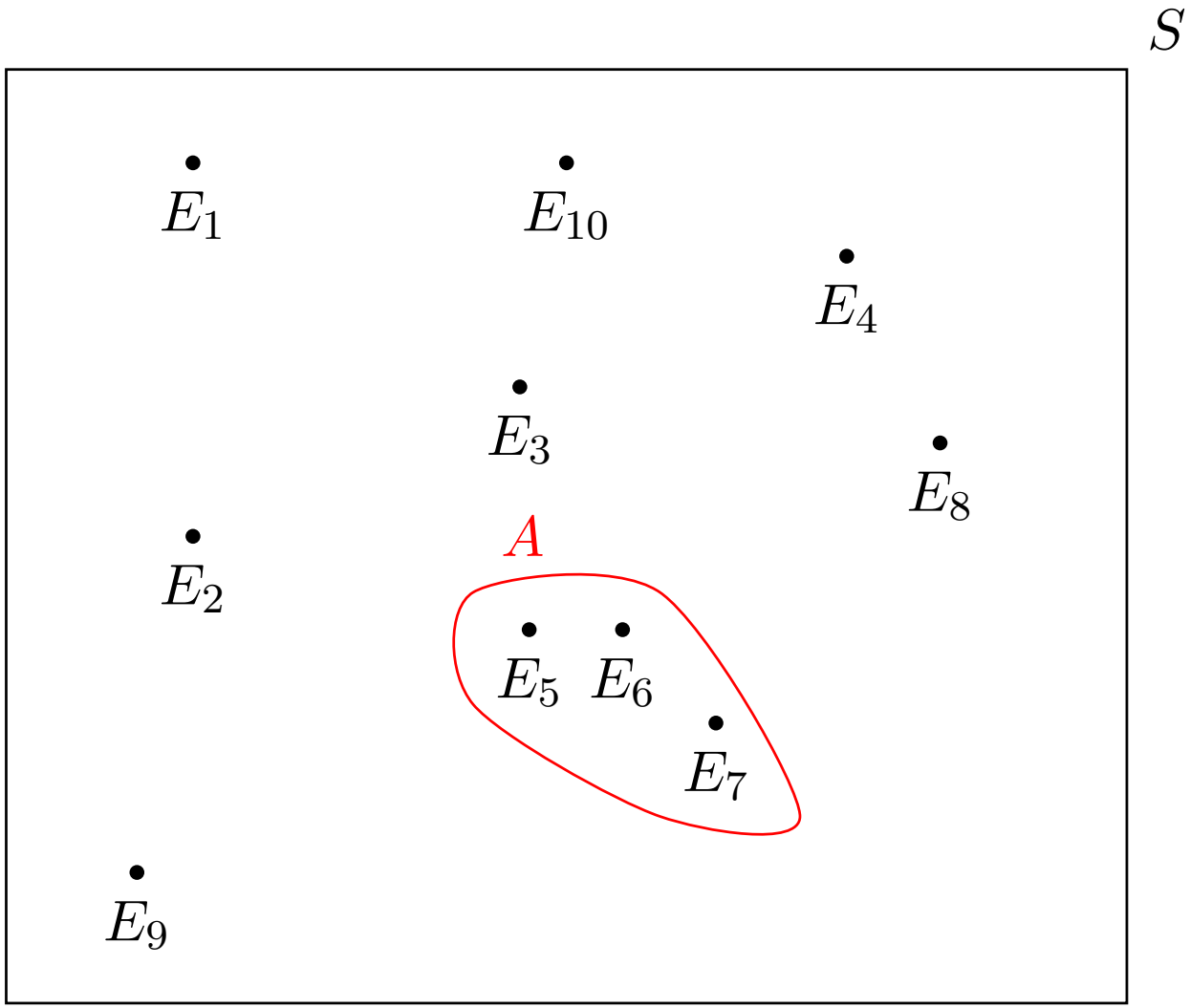
Specifying probabilities (cont.)

"**Equally likely outcomes**": this is a strong (and subjective) assumption, but if it holds, then it leads to a straightforward calculation.

Suppose we have $N = 10$:

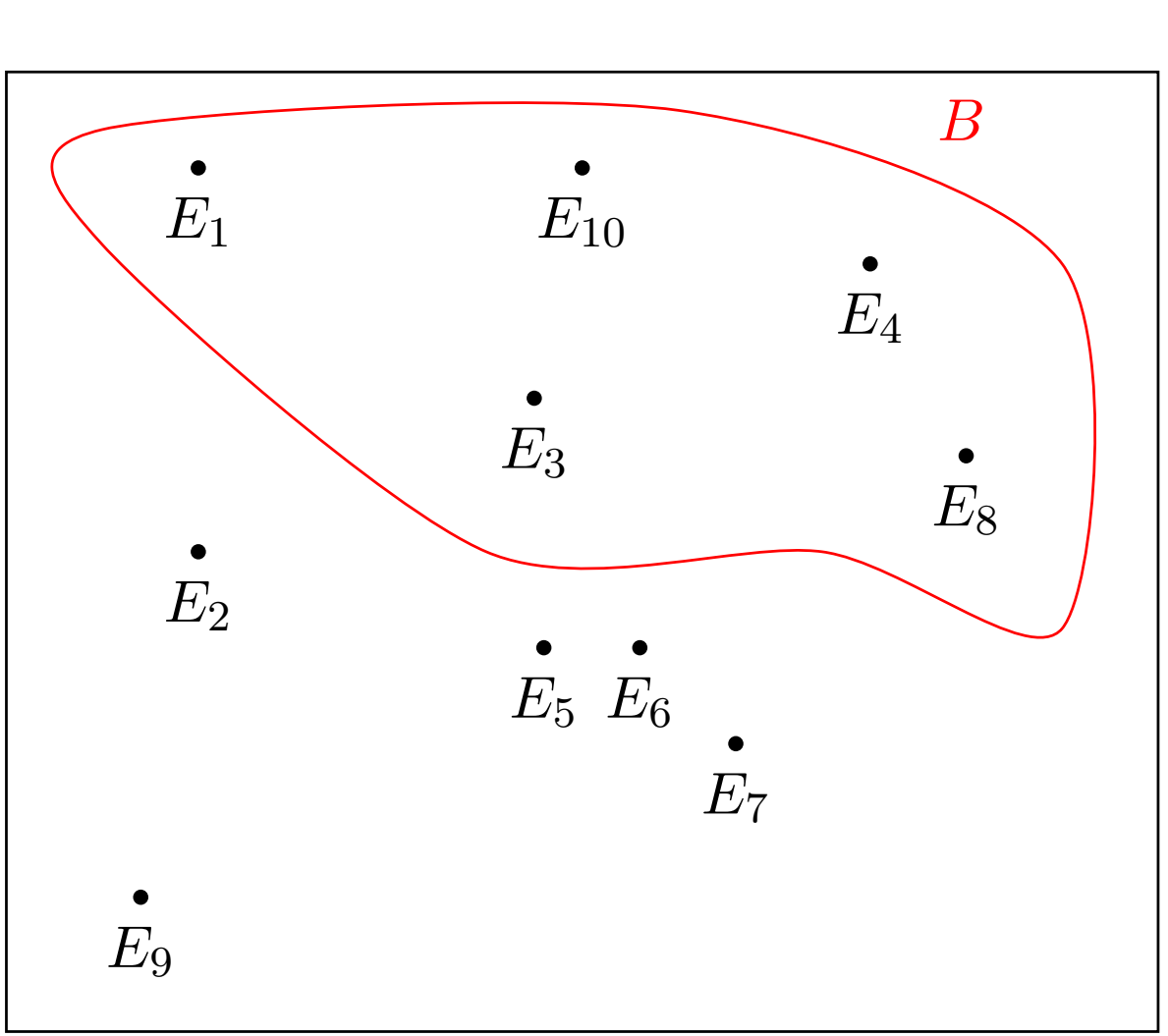
Specifying probabilities (cont.)

$$P(A) = \frac{3}{10}\text{: } 1_A = 5, 2_A = 6, 3_A = 7.$$



Specifying probabilities (cont.)

$$P(B) = \frac{5}{10} = \frac{1}{2}\colon\ 1_B = 1, 2_B = 3, 3_B = 4, 4_B = 8, 5_B = 10.$$



Specifying probabilities (cont.)

We usually apply the logic of "equally likely outcomes" in certain mechanical settings where we can appeal to some form of symmetry argument.

(a) **Coins:** It is common to assume the existence of a 'fair' coin, and an experiment involving a single flip. Here the sample space is

$$S = \{\text{Head}, \text{Tail}\}$$

with elementary events $E_1 = \{\text{Head}\}$ and $E_2 = \{\text{Tail}\}$, and it is usual to assume that

$$P(E_1) = P(E_2) = \frac{1}{2}.$$

(b) **Dice:** Consider a single roll of a 'fair' die, with the outcome being the upward face after the roll is complete. Here the sample space is

$$S = \{1, 2, 3, 4, 5, 6\}$$

with elementary events $E_i = \{i\}$ for $i = 1, 2, 3, 4, 5, 6$, and it is usual to assume that

$$P(E_i) = \frac{1}{6}.$$

Let A be the event that the outcome of the roll is an even number. Then

$$A = E_2 \cup E_4 \cup E_6$$

and

$$P(A) = \frac{3}{6} = \frac{1}{2}.$$

Specifying probabilities (cont.)

(c) **Cards:** A standard deck of cards contains 52 cards, comprising four suits (Hearts, Clubs, Diamonds, Spades) with thirteen cards in each suit, with cards denominated

$$2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace.$$

Thus each card has a suit and a denomination. An experiment involves selecting a card from the deck after it has been well shuffled.

There are 52 elementary outcomes, and

$$P(E_i) = \frac{1}{52} \quad i = 1, \ldots, 52.$$

If A is the event Ace is selected, then

$$P(A) = \frac{\text{Total number of } Aces}{\text{Total number of cards}} = \frac{4}{52} = \frac{1}{13}.$$

Specifying probabilities (cont.)

In a second experiment, five cards are to be selected without replacement from the deck. The elementary outcomes correspond to all sequences of five cards that could be obtained. All such sequences are equally likely.

Let A be the event that the five cards contain three cards of one denomination, and two cards of another denomination.

What is $P(A)$?

Need some rules to help in counting the elementary outcomes.

Specifying probabilities (cont.)

The concept of "equally likely outcomes" can be extended to the uncountable sample space case:

- pick a point from the interval $(0, 1)$ with each point equally likely; if A is the event

  "point picked is in the interval $(0, 0.25)$"

  then we can set $P(A) = 0.25$.

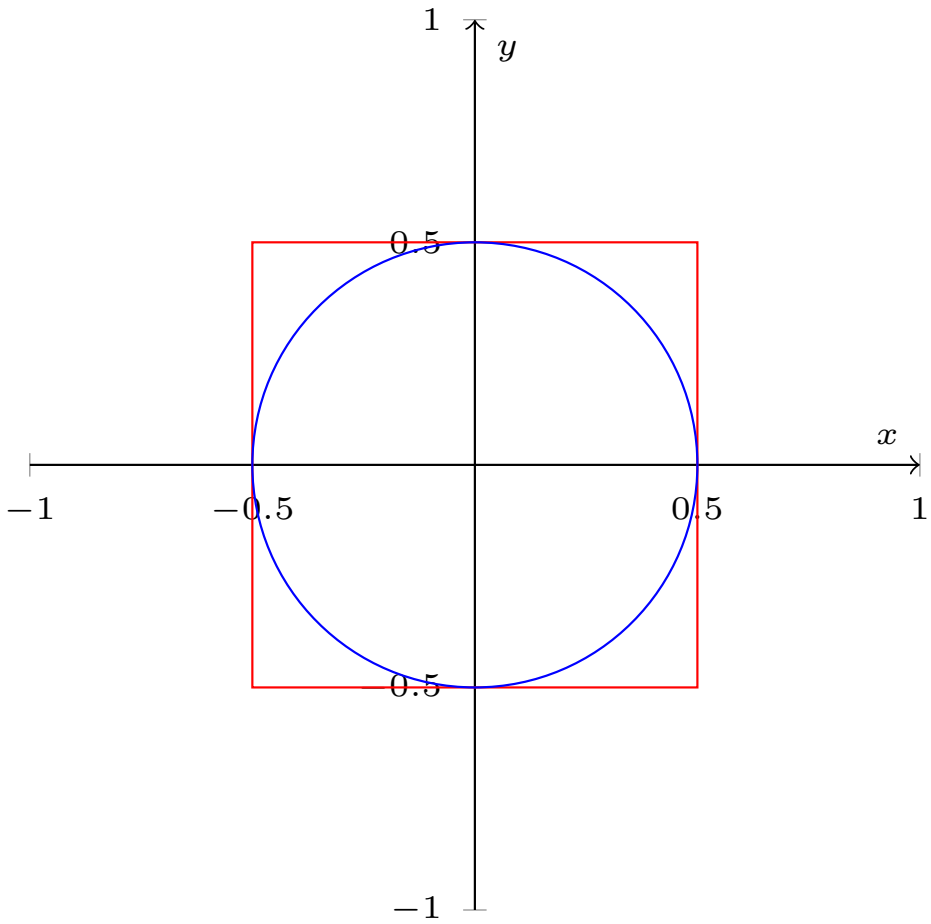
- a point is picked internal to the square centered at $(0, 0)$ with side length one, with each point equally likely. Let A be the event that

  "point lies in the circle centered at $(0, 0)$ with radius 0.5."

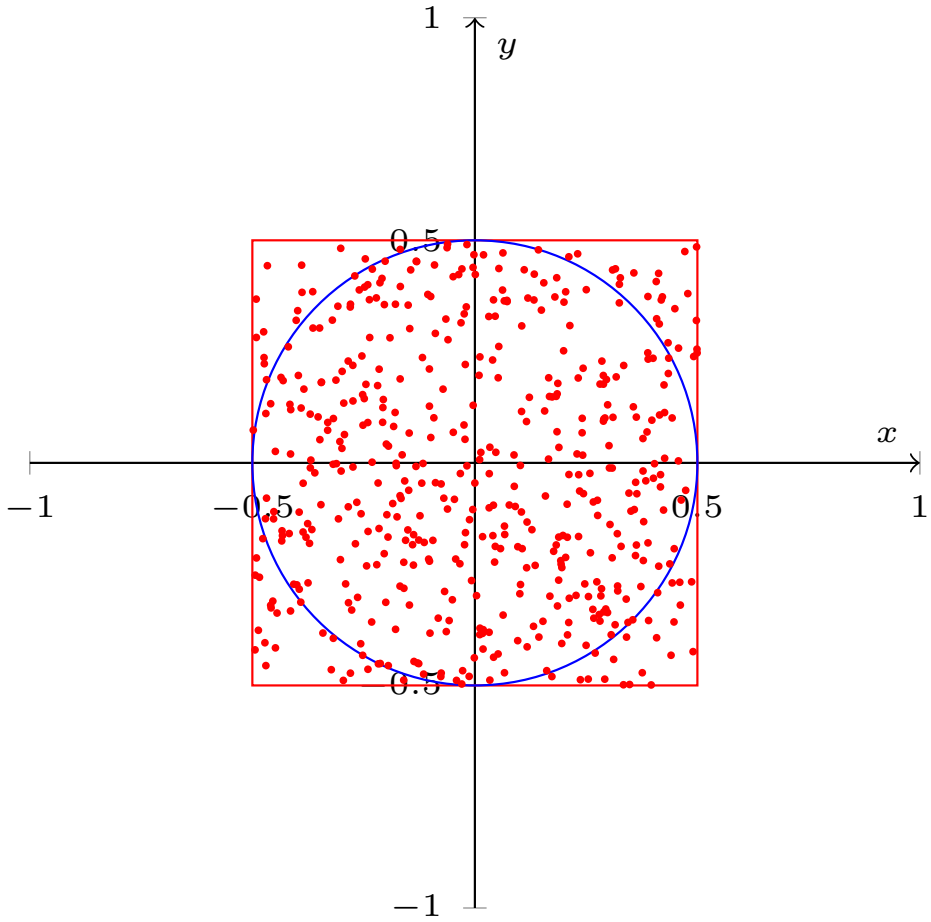
  Then

$$P(A) = \frac{\text{Area of circle}}{\text{Area of Square}} = \frac{\pi}{4} = 0.7853982.$$

Specifying probabilities (cont.)

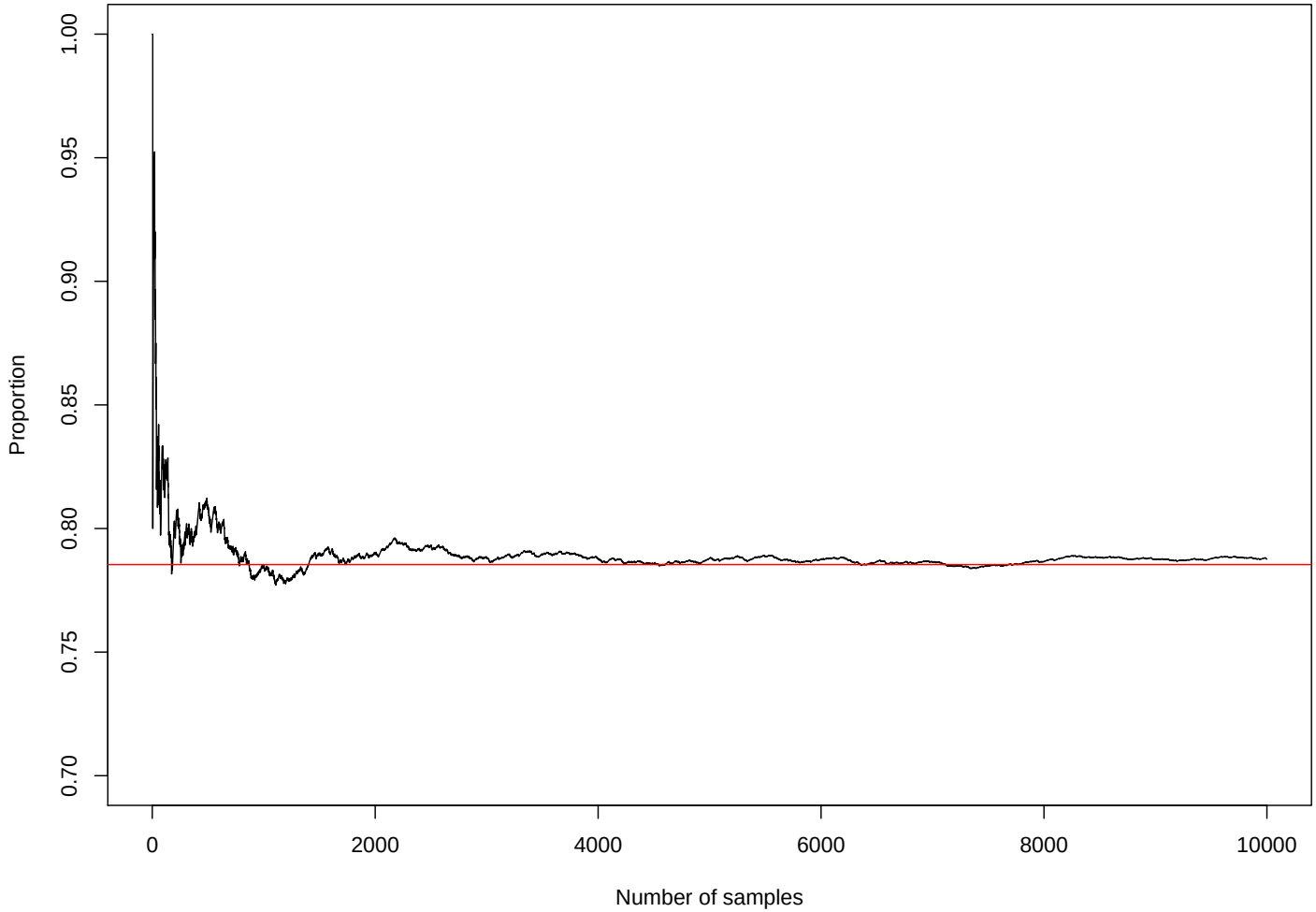


Specifying probabilities (cont.)



Specifying probabilities (cont.)

A simulation: 10000 points picked from the square.



Specifying probabilities (cont.)

Example (Thumbtack tossing)





- $S = \{\text{Up}, \text{Down}\}$.
- $A = \{\text{Up}\}$.

What is $P(A)$?

Specifying probabilities (cont.)

2. **Relative frequencies:** Suppose S is the sample space, and A is the event of interest. Consider an infinite sequence of repeats of the experiment under identical conditions. Then we may define $P(A)$ by considering the relative frequency with which event A occurs in the sequence.

   - Consider a finite sequence of N repeat experiments.
   - Let n be the number of times (out of N) that A occurs.
   - Define
   $$P(A) = \lim_{N \longrightarrow \infty} \frac{n}{N}.$$

   The *frequentist* definition of probability; it generalizes the "equally likely outcomes" version of probability.

Specifying probabilities (cont.)

The frequentist definition would cover the thumbtack example:



Can we always envisage an infinite sequence of repeats ?

Specifying probabilities (cont.)

3. **Subjective assessment:** The subjective definition of probability is that for a given experiment with sample space S, the probability of event A is

   - a numerical representation of your own personal

     degree of belief

     that the actual outcome lies in A;
   - you are *rational* and *coherent* (that is, internally consistent in your assessment).

   This definition generalizes the "equally likely outcomes" and "frequentist" versions of probability.

Specifying probabilities (cont.)

- Especially useful for 'one-off' experiments.
- Assessment of 'odds' on an event can be helpful:

$$\text{Odds}: \quad \frac{P(A)}{P(A')} = \frac{P(A)}{1 - P(A)}$$

eg Odds = 10:1, then $P(A) = 10/11$ etc.

Rules for counting outcomes

Multiplication principle: A sequence of k operations, in which operation i can result in n_i possible outcomes, can result in

$$n_1 \times n_2 \times \cdots \times n_k$$

possible sequences of outcomes.

Rules for counting outcomes (cont.)

Example (Two dice)

Two dice are rolled: there are $6 \times 6 = 36$ possible outcomes.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Rules for counting outcomes (cont.)

Example

We are to pick $k = 5$ numbers sequentially from the set

$$\{1, 2, 3, \ldots, 100\},$$

where every number is available on every pick. The number of ways of doing that is

$$100 \times 100 \times 100 \times 100 \times 100 = 100^5.$$

Rules for counting outcomes (cont.)

Example

We are to pick $k = 11$ players sequentially from a squad of 25 players. The number of ways of doing that is

$$25 \times 24 \times 23 \times \cdots \times 15.$$

as players cannot be picked twice.

Rules for counting outcomes (cont.)

Selection principles: When selecting repeatedly from a finite set $\{1, 2, \ldots, N\}$ we may select

- with replacement, that is,
  - each successive selection can be one of the original set, irrespective of previous selections,

  or
- without replacement, that is,
  - the set is depleted by each successive selection,

Rules for counting outcomes (cont.)

Ordering: When examining the result of a sequence of selections, it may be required that

- order is important, that is

$$13456 \quad \text{is considered distinct from} \quad 54163$$

or

- order is unimportant, that is

$$13456 \quad \text{is considered identical to} \quad 54163$$

Rules for counting outcomes (cont.)

Example (Two dice)

Two dice are rolled: there are $(7 \times 6)/2 = 21$ possible unordered outcomes.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	-	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	-	-	(3,3)	(3,4)	(3,5)	(3,6)
4	-	-	-	(4,4)	(4,5)	(4,6)
5	-	-	-	-	(5,5)	(5,6)
6	-	-	-	-	-	(6,6)

Rules for counting outcomes (cont.)

Distinguishable items: It may be that the objects being selected are

- distinguishable, that is, individually uniquely labelled.
  - eg lottery balls.
- indistinguishable, that is, labelled according to a type, but not labelled individually.
  - eg a bag containing 7 red balls, 2 green balls, and 3 yellow balls.

Permutations

An ordered arrangement of r distinct objects is called a *permutation.*

The number of ways of ordering n distinct objects taken r at a time is denoted P_r^n, and by the multiplication rule we have that

$$P_r^n = n \times (n-1) \times (n-2) \times \cdots (n-r+2) \times (n-r+1) = \frac{n!}{(n-r)!}$$

Multinomial coefficients

The number of ways of partitioning n distinct objects into k disjoint subsets of sizes

$$n_1, n_2, \ldots, n_k$$

where

$$\sum_{i=1}^{k} n_i = n$$

is

$$N = \binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! \times n_2! \ldots \times n_k!}.$$

Multinomial coefficients (cont.)

To see this, consider first selecting the n objects in order; there are

$$P_n^n = n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$$

ways of doing this. Then designate

- objects selected 1 to n_1 as Subset 1,
- objects selected $n_1 + 1$ to $n_1 + n_2$ as Subset 2,
- objects selected $n_1 + n_2 + 1$ to $n_1 + n_2 + n_3$ as Subset 3,
- ...
- objects selected $n_1 + n_2 + \cdots + n_{k-1} + 1$ to $n_1 + n_2 + \cdots + n_k$ as Subset k.

Multinomial coefficients (cont.)

Then note that the specific ordered selection that achieves the partition is only one of several that yield the same partition; there are

- $n_1!$ ways of permuting Subset 1,
- $n_2!$ ways of permuting Subset 2,
- $n_3!$ ways of permuting Subset 3,
- ...
- $n_k!$ ways of permuting Subset k,

that yield the same partition.

Multinomial coefficients (cont.)

Therefore, we must have that

$$P_n^n = n! = N \times (n_1! \times n_2! \ldots \times n_k!)$$

and hence

$$N = \frac{n!}{n_1! \times n_2! \times \ldots \times n_k!}$$

Multinomial coefficients (cont.)

Example ($n = 8$, $k = 3$, $(n_1, n_2, n_3) = (4, 2, 2)$)

- One specific ordered selection:

$$4, 6, 2, 1, 8, 5, 3, 7$$

  – there are $8!$ ways of obtaining such a selection.

- Partition into subsets with (4,2,2) elements:

$$(4, 6, 2, 1), (8, 5), (3, 7)$$

Multinomial coefficients (cont.)

Example ($n = 8$, $k = 3$, $(n_1, n_2, n_3) = (4, 2, 2)$)

- Consider all within-subset permutations:

$$(4, 6, 2, 1), (8, 5), (3, 7)$$

$$(4, 6, 1, 2), (8, 5), (3, 7)$$

$$(4, 1, 2, 6), (8, 5), (3, 7)$$

$$\vdots$$

Multinomial coefficients (cont.)

Example ($n = 8$, $k = 3$, $(n_1, n_2, n_3) = (4, 2, 2)$)

- There are

$$\frac{8!}{4! \times 2! \times 2!} = \frac{40320}{24 \times 2 \times 2} = 420$$

possible distinct partitions;

$$(4, 6, 2, 1), (8, 5), (3, 7)$$

is regarded as identical to

$$(1, 2, 4, 6), (5, 8), (3, 7).$$

Multinomial coefficients (cont.)

Example (Department committees)

A University Department comprises 40 faculty members. Two committees of 12 faculty members, and two committees of 8 faculty members, are needed.

The number of distinct committee configurations that can be formed is

$$N = \binom{40}{12, 12, 8, 8} = \frac{40!}{12! \times 12! \times 8! \times 8!} = 1785474512.$$

Multinomial coefficients (cont.)

Note: Binomial coefficients

Recall the binomial theorem:

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}$$

where

$$\binom{n}{j} = \frac{n!}{j!(n-j)!} \equiv \binom{n}{j, n-j}$$

Multinomial coefficients (cont.)

Note: Binomial coefficients

Here we are solving a partitioning problem; partition n elements into one subset of size j and one subset of size $n - j$.

We can use this result and the multiplication principle to derive the general multinomial coefficient result.

Note: Binomial coefficients

For a collection of n elements:

1. Partition into two subsets of n_1 and $n - n_1$ elements;
2. Take the $n - n_1$ elements and partition them into two subsets of n_2 and $n - n_1 - n_2$ elements;
3. Repeat until the final partition step; at this stage we have a set of $n - n_1 - \cdots - n_{k-2}$ elements, which we partition into a subset of n_{k-1} elements and a subset of

$$n_k = n - n_1 - \cdots - n_{k-2} - n_{k-1}$$

elements.

Note: Binomial coefficients

By the multiplication rule, the number of ways this sequence of partitioning steps can be carried out is

$$\binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \cdots \times \binom{n-n_1-\cdots-n_{k-2}}{n_{k-1}}$$

that is

$$\frac{n!}{n_1!(n-n_1)!} \times \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \times \cdots \times \frac{(n-n_1-\cdots-n_{k-2})!}{n_{k-1}!n_k!}$$

Note: Binomial coefficients

Cancelling terms top and bottom in successive factors, we see that this equals

$$\frac{n!}{n_1 \times n_2! \times \ldots \times n_{k-1}! \times n_k!}$$

as required.

Combinations

The number *combinations* n objects taken r at a time is the number of subsets, each of size r, that can be formed from *objects*.

This number is denoted

$$C_r^n = \binom{n}{r}$$

where

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Combinations (cont.)

We first consider sequential selection of r objects: the number of possible selections (without replacement) is

$$n \times (n-1) \times ... \times (n-r+1) = \frac{n!}{(n-r)!} = P_r^n$$

leaving $(n-r)$ objects unselected.

We then remember that the order of selected objects is not important in identifying a combination, so therefore we must have that

$$P_r^n = r! \times C_r^n.$$

as there are $r!$ equivalent combinations that yield the same permutation. The result follows.

Binary sequences

$$101000100101010$$

arise in many probability settings; if we take '1' to indicate inclusion and '0' to indicate exclusion, then we can identify combinations with binary sequences.

- the number of binary sequences of length n containing r 1s is

$$\binom{n}{r}$$

Combinatorial probability

The above results are used to compute probabilities in the case of equally likely outcomes.

- S: complete list of possible sequences of selections.
- A: sequences having property of interest.
- We have

$$P(A) = \frac{\text{Number of elements in } A}{\text{Number of elements in } S} = \frac{n_A}{n_S}$$

Combinatorial probability (cont.)

Example (Cards)

Five cards are selected without replacement from a standard deck. What is the probability they are all Hearts ?

Combinatorial probability (cont.)

Example (Cards)

- Number of elements in S:

$$n_S = \binom{52}{5}$$

- Number of elements in A:

$$n_A = \binom{13}{5}$$

so therefore

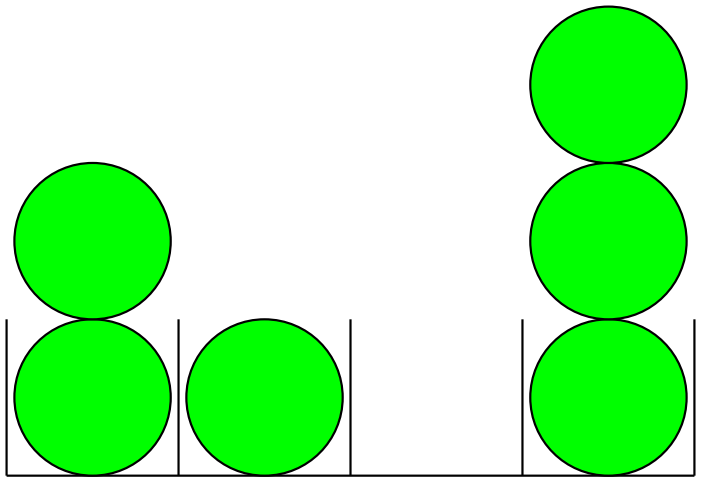
$$P(A) = \frac{\binom{13}{5}}{\binom{52}{5}} = 0.0004951981.$$

Other combinatorial problems

- Poker hands;
- Hypergeometric selection;
- Occupancy (or allocation) problems – allocate r objects to n boxes and identify
  - the occupancy pattern;
  - the occupancy of a specific box.

Other combinatorial problems (cont.)

Allocate $r = 6$ indistinguishable balls to $n = 4$ boxes: how many distinct allocation patterns are there ?



Conditional probability

We now consider probability assessments in the presence of

partial knowledge.

In the case of 'ordinary' probability, we have (by assumption) that the outcome **must** be an element in the set S, and proceed to assess the probability that the outcome is an element of the set $A \subseteq S$.

That is, the only 'certain' knowledge we have is that the outcome, s, is in S.

Conditional probability (cont.)

Now suppose we have the 'partial' knowledge that, in fact,

$$s \in B \subseteq S$$

for some B;

- that is, we know that event B occurs.

Conditional probability (cont.)

How does this change our probability assessment concerning A ?

- in light of the information that B occurs, what do we now think about the probability that A occurs also ?

Conditional probability (cont.)

First, we are considering the event that **both** A and B occur, that is

$$s \in A \cap B$$

so $P(A \cap B)$ must play a role in the calculation.

Secondly, with the knowledge that event B occurs, we restrict the parts of the sample space that should be considered; we are **certain** that the sample outcome must lie in B.

Conditional probability (cont.)

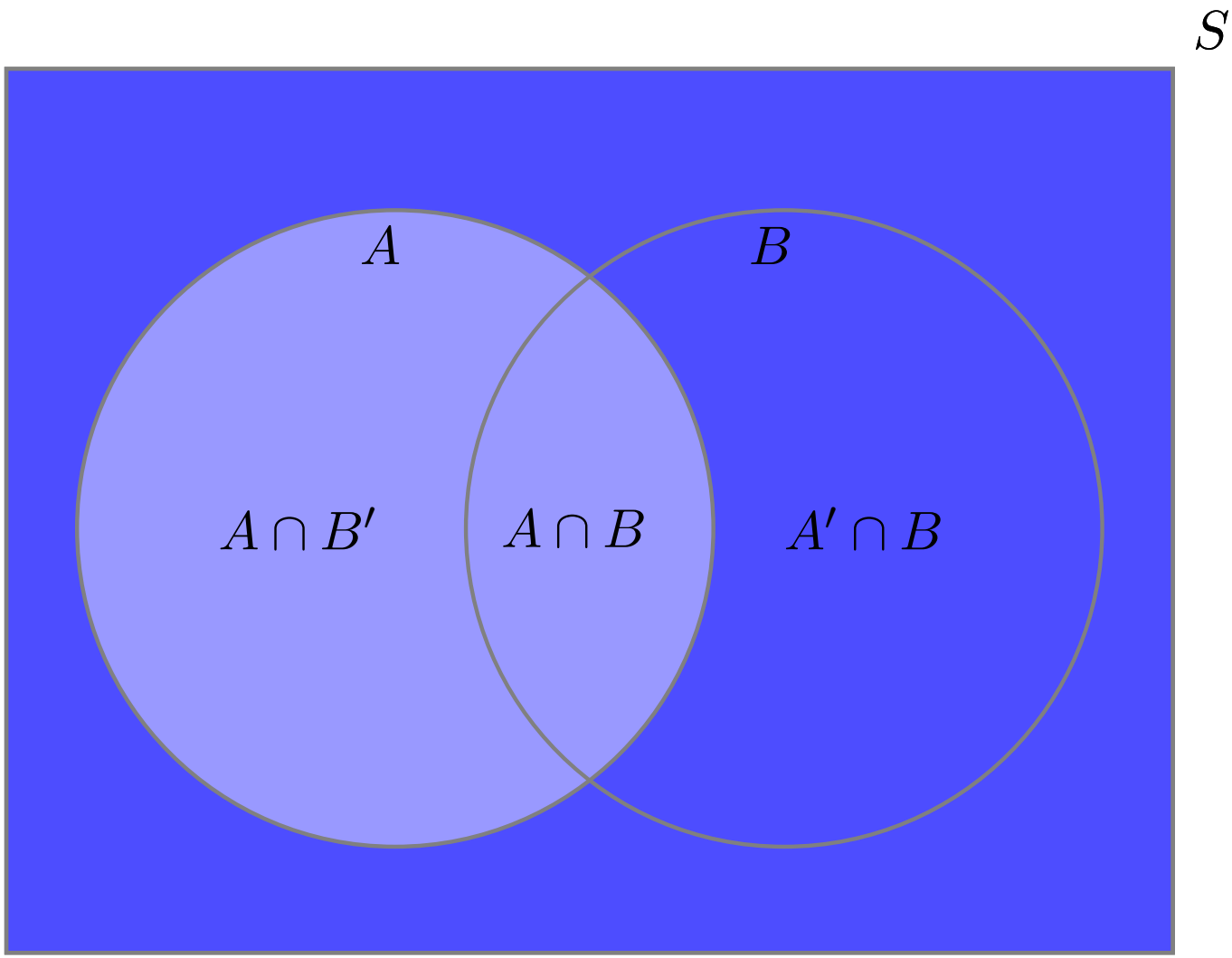
For two events A and B, the *conditional probability* of A *given* B is denoted $P(A|B)$, and is defined by

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

Note that we consider this **only** in cases where $P(B) > 0$.

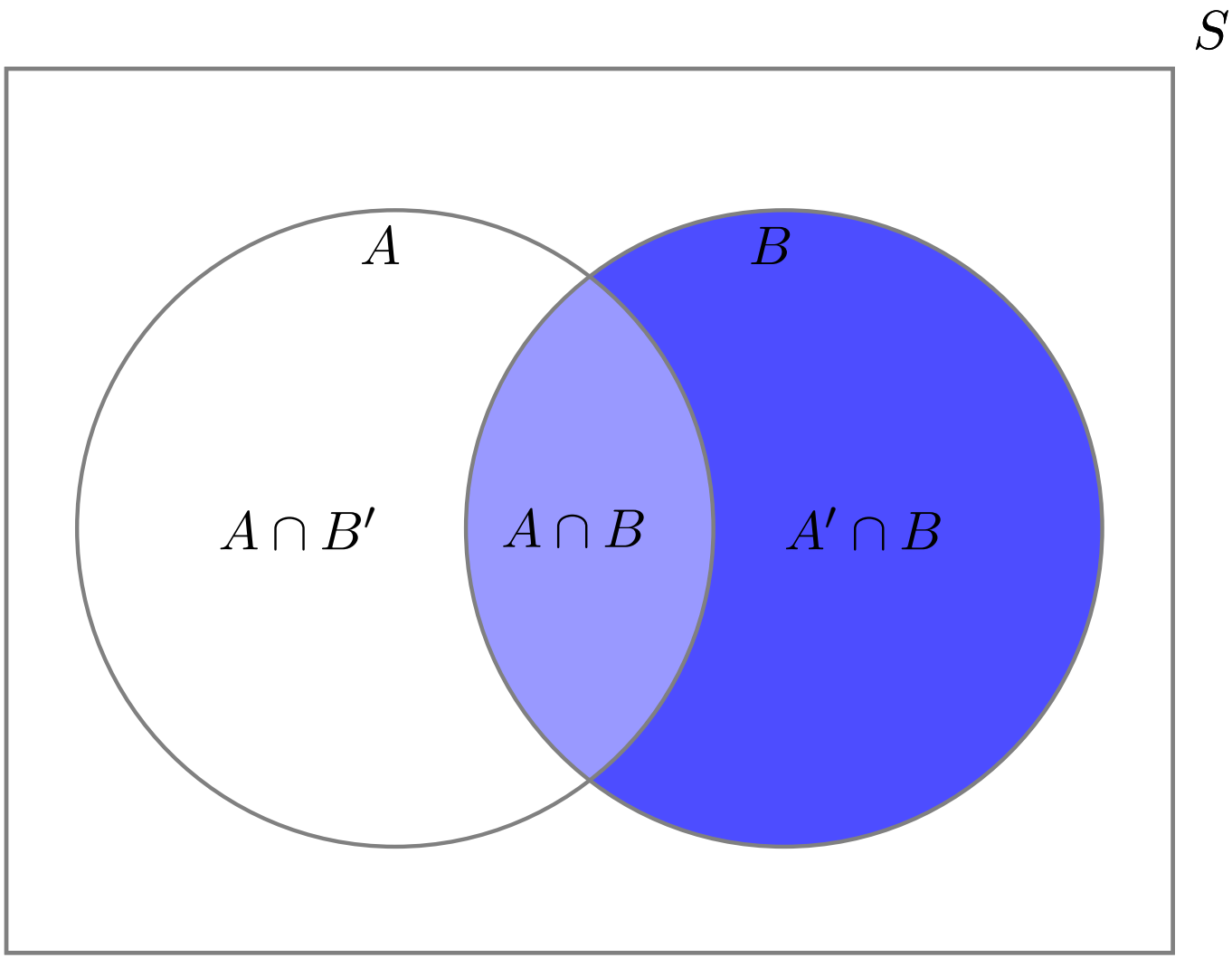
Conditional probability (cont.)

Ordinary Probability: $P(A)$.



Conditional probability (cont.)

Conditional Probability: $P(A|B) = P(A \cap B)/P(B)$.



Conditional probability (cont.)

Example

Experiment: roll a fair die, record the score.

- $S = \{1, 2, 3, 4, 5, 6\}$, all outcomes equally likely.
- $A = \{1, 3, 5\}$ (score is odd).
- $B = \{4, 5, 6\}$ (score is more than 3).
- $A \cap B = \{5\}$.

$$P(A) = \frac{3}{6} = \frac{1}{2} \qquad P(B) = \frac{3}{6} = \frac{1}{2} \qquad P(A \cap B) = \frac{1}{6}$$

so

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Conditional probability (cont.)

Example

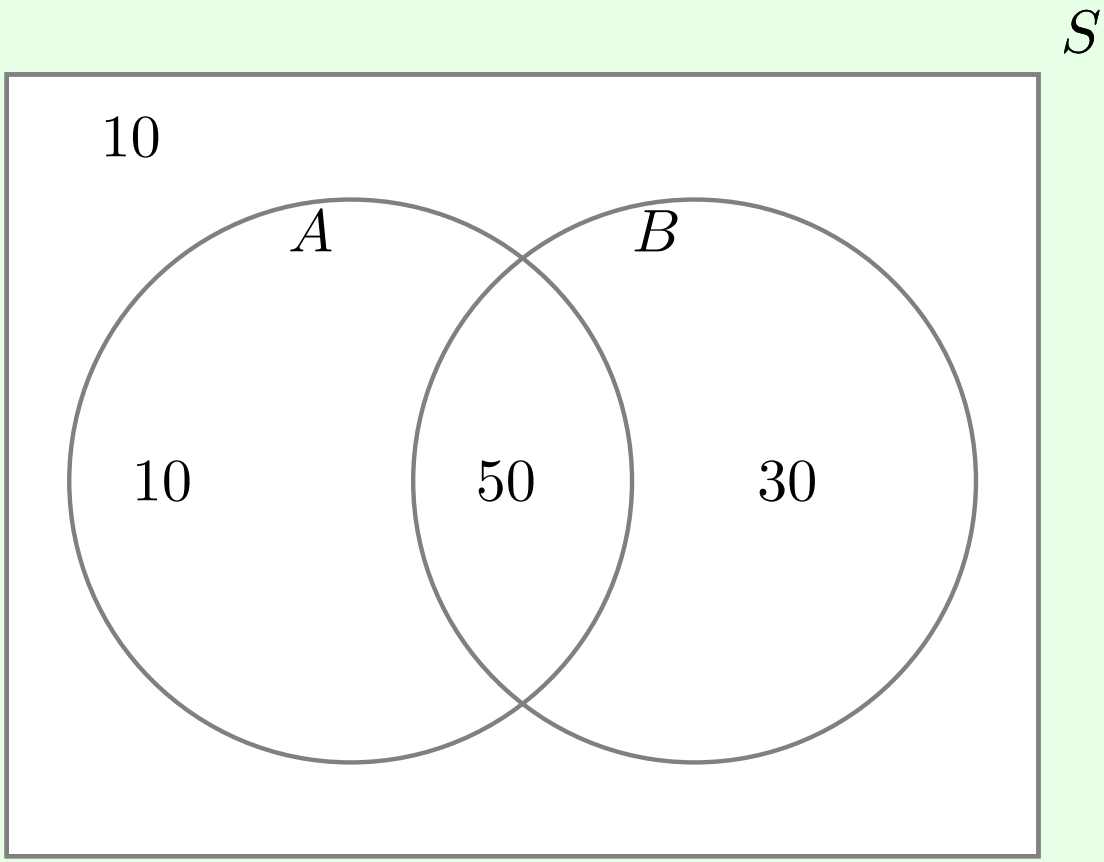
In a class of 100 students:

- 60 are from Faculty of Science, 40 are from Faculty of Arts.
- 80 are in a Major program, 20 are in another program.
- 50 of the Science students are in a Major program.

Conditional probability (cont.)

Example

A – Faculty of Science; B – Major program.



Conditional probability (cont.)

Example

A student is selected from the class, with all students equally likely to be selected.

- What is the probability that the selected student is from Science, $P(A)$?:

$$P(A) = \frac{60}{100} = \frac{3}{5}$$

Conditional probability (cont.)

Example

- What is the probability that the selected student is from Science and in a Major program, $P(A \cap B)$?:

$$P(A \cap B) = \frac{50}{100} = \frac{1}{2}$$

Conditional probability (cont.)

Example

- If the selected student is **known** to be in a Major program, what is the probability that the student is from Science, $P(A|B)$?:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{50/100}{80/100} = \frac{50}{80} = \frac{5}{8}.$$

Note: some consequences

- If $B \equiv S$, then

$$P(A|S) = \frac{P(A \cap S)}{P(S)} = \frac{P(A)}{1} = P(A).$$

- If $B \equiv A$, then

$$P(A|A) = \frac{P(A \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1.$$

- We have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \leq \frac{P(B)}{P(B)} = 1.$$

Conditional probability (cont.)

Direct from the definition, we have that

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

(recall $P(B) > 0$).

Conditional probability (cont.)

Note

It is important to understand the distinction between

$$P(A \cap B) \qquad \text{and} \qquad P(A|B).$$

- $P(A \cap B)$ records the chance of A **and** B occurring relative to S. A and B are treated symmetrically in the calculation.
- $P(A|B)$ records the chance of A **and** B occurring relative to B. A and B are **not** treated symmetrically; from the definition, we see that in general

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

Note

From the above calculation, we see that we could have

$$P(A|B) \leq P(A)$$

or

$$P(A|B) \geq P(A)$$

or

$$P(A|B) = P(A).$$

That is, certain knowledge that B occurs could decrease, increase or leave unchanged the probability that A occurs.

Conditional probability (cont.)

Example

A box of 500 light bulbs is purchased from each of two factories (labelled 1 and 2).

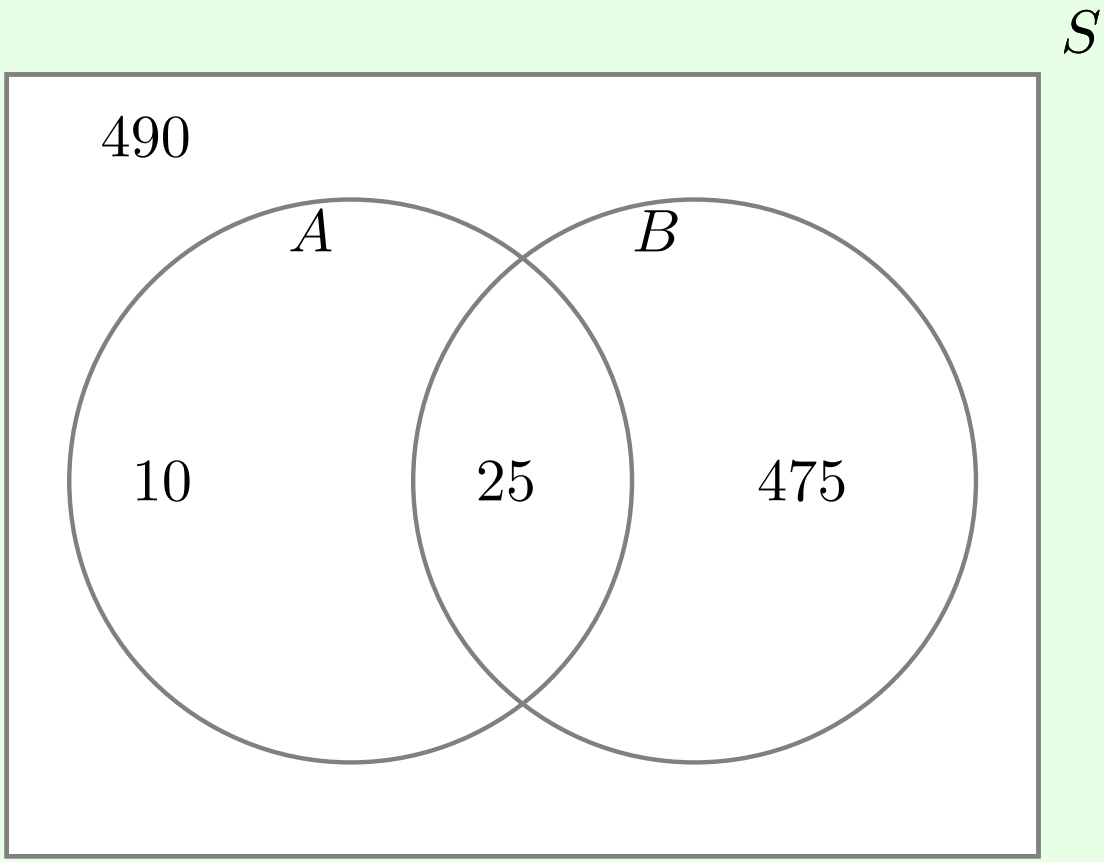
- In the box from Factory 1, there are 25 defective bulbs.
- In the box from Factory 2, there are 10 defective bulbs.

The bulbs are unpacked and placed in storage.

Conditional probability (cont.)

Example

A – Defective; B – Factory 1.



Conditional probability (cont.)

Example

At the time of the next installation, a bulb is selected from the store.

- What is the probability that the bulb is defective, $P(A)$?:

$$P(A) = \frac{35}{1000} = 0.035.$$

Conditional probability (cont.)

Example

- What is the probability that the bulb came from Factory 1, $P(B)$?:

$$P(B) = \frac{500}{1000} = 0.5.$$

Conditional probability (cont.)

Example

- If the selected bulb **came from Factory 1**, what is the probability it is **defective** $P(A|B)$?:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{25/1000}{500/1000} = \frac{25}{500} = 0.05.$$

Conditional probability (cont.)

Example

- If the selected bulb is **defective**, what is the probability it came from Factory 1, $P(B|A)$?:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{25/1000}{35/1000} = \frac{25}{35} = \frac{5}{7}.$$

Conditional probability (cont.)

Example

Experiment: We measure the failure time of an electrical component.

- $S = R^+$.
- $A = [10, \infty)$ (fails after 10 hours or more).
- $B = [5, \infty)$ (fails after 5 hours or more).

Suppose we assess

$$P(A) = 0.25 \qquad P(B) = 0.4.$$

Now here $A \subset B \subset S$, so therefore $P(A \cap B) \equiv P(A)$, and

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{0.25}{0.4} = 0.625.$$

Conditional probability (cont.)

The conditional probability function satisfies the Probability Axioms: suppose B is an event in S such that $P(B) > 0$.

(I) Non-negativity:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \geq 0.$$

(II) For $P(S|B)$ we have

$$P(S|B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

Conditional probability (cont.)

(III) Countable additivity: if $A_1, A_2, \ldots$ form a (countable) sequence of events such that $A_j \cap A_k = \emptyset$ for all $j \neq k$.

Conditional probability (cont.)

Then

$$P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \frac{P\left(\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right)}{P(B)}$$

$$= \frac{P\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right)}{P(B)}$$

$$= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)}$$

$$= \sum_{i=1}^{\infty} P(A_i|B).$$

Conditional probability (cont.)

$$
\begin{aligned}
P\left(A_1 \cup A_2 \middle| B\right) &= \frac{P\left((A_1 \cup A_2) \cap B\right)}{P(B)} \\
&= \frac{P\left((A_1 \cap B) \cup (A_2 \cap B)\right)}{P(B)} \\
&= \frac{P(A_1 \cap B) + P(A_1 \cap B)}{P(B)} \\
&= P(A_1|B) + P(A_2|B).
\end{aligned}
$$

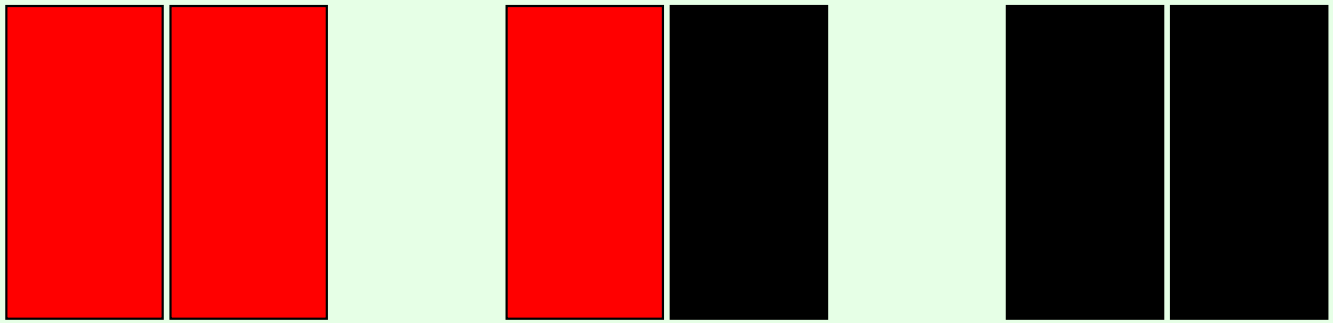
as

$$
A_1, A_2 \text{ disjoint} \quad \Longrightarrow \quad (A_1 \cap B), (A_2 \cap B) \text{ disjoint}.
$$

Conditional probability (cont.)

Example (Three cards)

Three two sided cards:



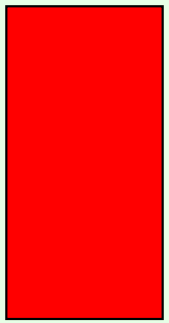
1 2 3

One of the three cards is selected, with all cards being equally likely to be chosen.

Conditional probability (cont.)

Example (Three cards)

One side is displayed:



What is the probability that the other side of this card is red ?

Conditional probability (cont.)

Example (Three cards)

Let

- $S = \{R_1, R_2, R, B, B_1, B_2\}$ be the possible exposed sides. The outcomes in S are equally likely.
- A – Card 1 is selected.

$$A = \{R_1, R_2\}$$

and $P(A) = 2/6$.

- B – A red side is exposed.

$$B = \{R_1, R_2, R\}$$

and $P(B) = 3/6$.

Example (Three cards)

Then

$$A \cap B = \{R_1, R_2\} \equiv A$$

and hence

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{2/6}{3/6} = \frac{2}{3}.$$

Conditional probability (cont.)

Example (Drug testing)

A drug testing authority tests 100000 athletes to assess whether they are using performance enhancing drugs (PEDs). The drug test is not perfect as it produces

- *False positives:* declares an athlete to be using PEDs, when in reality they are not.
- *False negatives:* declares an athlete not to be using PEDs, when in reality they are.

Conditional probability (cont.)

Example (Drug testing)

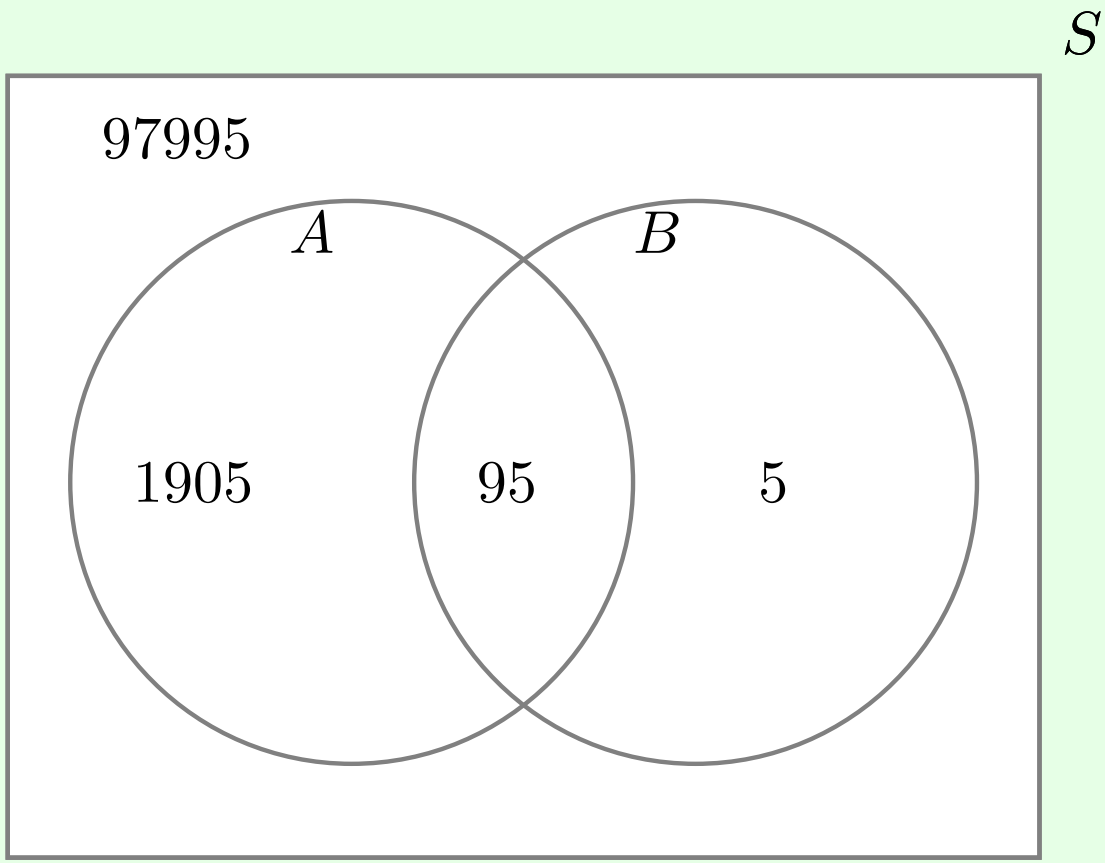
Suppose that after detailed investigation it is discovered that

- 2000 athletes gave positive test results.
- 100 athletes were using PEDs.
- 95 athletes who were using PEDs tested positive.

Conditional probability (cont.)

Example (Drug testing)

A – Drug test is positive; B – Athlete is using PED.



Conditional probability (cont.)

Example (Drug testing)

However, in the original analysis, an athlete is selected at random from the 100000, and is observed to test positive.

What is the probability that they actually were using PEDs ?

Conditional probability (cont.)

Example (Drug testing)

The required conditional probability is

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{95}{2000} = 0.0475.$$

Note that is very different from

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{95}{100} = 0.95.$$

Independence

Two events A and B in sample space S are **independent** if

$$P(A|B) = P(A);$$

equivalently, they are independent if

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

Note that if $P(A|B) = P(A)$, then

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

Independence (cont.)

Note

Independence is not a property of the events A and B themselves, it is a property of the probabilities assigned to them.

Independence (cont.)

Note

Independence is **not** the same as mutual exclusivity.

$$\text{Independent: } P(A \cap B) = P(A)P(B)$$

$$\text{Mutually exclusive: } P(A \cap B) = 0.$$

Independence (cont.)

Example

A fair coin is tossed twice, with the outcomes of the two tosses independent. We have

$$S = \{HH, HT, TH, TT\}$$

with all four outcomes equally likely.

(a) If the first toss result is H, what is the probability that the second toss is also H ?

(b) If one result is H, what is the probability that the other is H ?

Independence (cont.)

Example

Let

- $A = \{HH\}$ (both heads).
- $B = \{HH, HT, TH\}$ (at least one head).
- $C = \{HH, HT\}$ (first result is head).

(a) We want $P(A|C)$:

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(A)}{P(C)} = \frac{1/4}{2/4} = \frac{1}{2}.$$

(b) We want $P(A|B)$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Independence for multiple events

Independence as defined above is a statement concerning two events.

What if we have more than two events ?

- Events A_1, A_2, A_3 in sample space S.
- We can consider independence pairwise:

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

$$P(A_1 \cap A_3) = P(A_1)P(A_3)$$

$$P(A_2 \cap A_3) = P(A_2)P(A_3)$$

Independence for multiple events (cont.)

- What about
$$P(A_1 \cap A_2 \cap A_3);$$
Can we deduce
$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)?$$
- In general, no.

Independence for multiple events (cont.)

Example (Two dice)

Suppose we roll two dice with the outcomes being independent.

- A_1 – first roll outcome is odd.
- A_2 – second roll outcome is odd.
- A_3 – total score is odd.

We have

$$P(A_1) = P(A_2) = \frac{1}{2}$$

and also

$$P(A_1|A_3) = P(A_2|A_3) = \frac{1}{2}$$

– can compute this by identifying all 36 pairs of scores, which are all equally likely, and counting the relevant sample outcomes.

Thus A_1, A_2 and A_3 are *pairwise* independent.

Independence for multiple events (cont.)

Example (Two dice)

However,

$$P(A_1 \cap A_2 \cap A_3) = 0$$

and

$$P(A_1 | A_2 \cap A_3) = 0$$

etc.

Independence for multiple events (cont.)

Mutual Independence: Events $A_1, A_2, \ldots, A_K$ are *mutually independent* if

$$P\left(\bigcap_{k \in \mathcal{I}} A_k\right) = \prod_{k \in \mathcal{I}} P(A_k)$$

for **all subsets** $\mathcal{I}$ of $\{1, 2, \ldots, K\}$.

For example, if $K = 3$, we require that

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

$$P(A_1 \cap A_3) = P(A_1)P(A_3)$$

$$P(A_2 \cap A_3) = P(A_2)P(A_3)$$

and $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$.

Conditional independence: Consider events A_1, A_2 and B in sample space S, with $P(B) > 0$. Then A_1 and A_2 are *conditionally independent* given B if

$$P(A_1|A_2 \cap B) = P(A_1|B)$$

or equivalently

$$P(A_1 \cap A_2|B) = P(A_1|B)P(A_2|B).$$

General Multiplication Rule

For events $A_1, A_2, \ldots, A_K$, we have the general result that

$$P(A_1 \cap A_2 \cap \cdots \cap A_K) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\ldots$$
$$\ldots P(A_K|A_1 \cap A_2 \cap \cdots \cap A_{K-1})$$

This follows by the recursive calculation

$$\begin{aligned} P(A_1 \cap A_2 \cap \cdots \cap A_K) &= P(A_1)P(A_2 \cap \cdots \cap A_K|A_1) \\ &= P(A_1)P(A_2|A_1) \\ &\qquad P(A_3 \cap \cdots \cap A_K|A_1 \cap A_2) \\ &= \ldots \end{aligned}$$

Also known as the *Chain Rule* for probabilities.

General Multiplication Rule (cont.)

If the events are mutually independent, then we have that

$$P(A_1 \cap A_2 \cap \cdots \cap A_K) = \prod_{k=1}^{K} P(A_k)$$

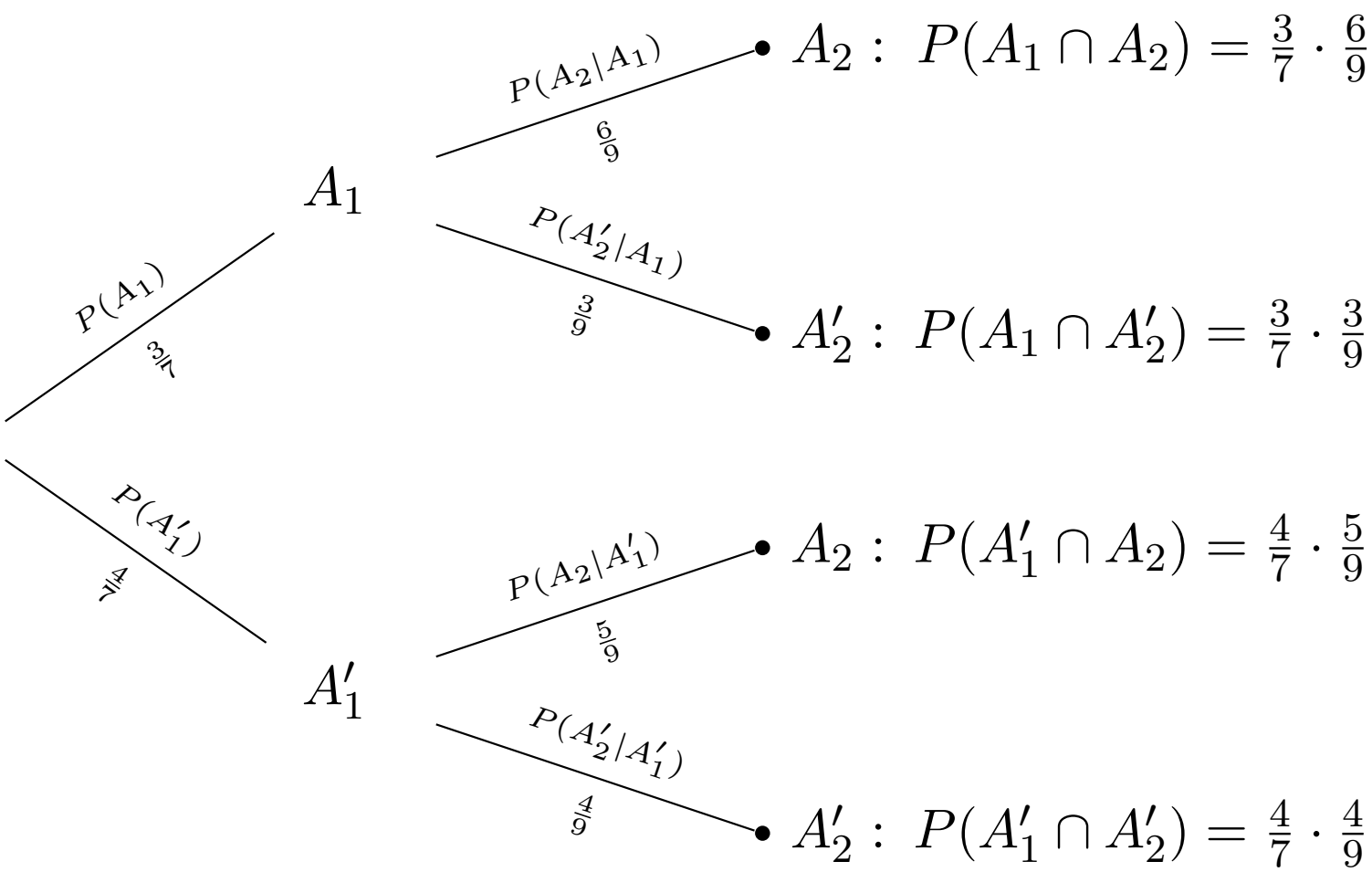
Probability Trees

Probability trees are simple ways to display joint probabilities of multiple events. They comprise

- Junctions: corresponding to the multiple events
- Branches: corresponding to paths displaying the sequence of (conditional) choices of events, given the previous choices.

We multiply along the branches to get the joint probabilities.

Probability Trees (cont.)



Probability Trees (cont.)

Note that each junction has two possible branches coming from it, corresponding to

$$A_k \quad \text{and} \quad A_k'$$

respectively.

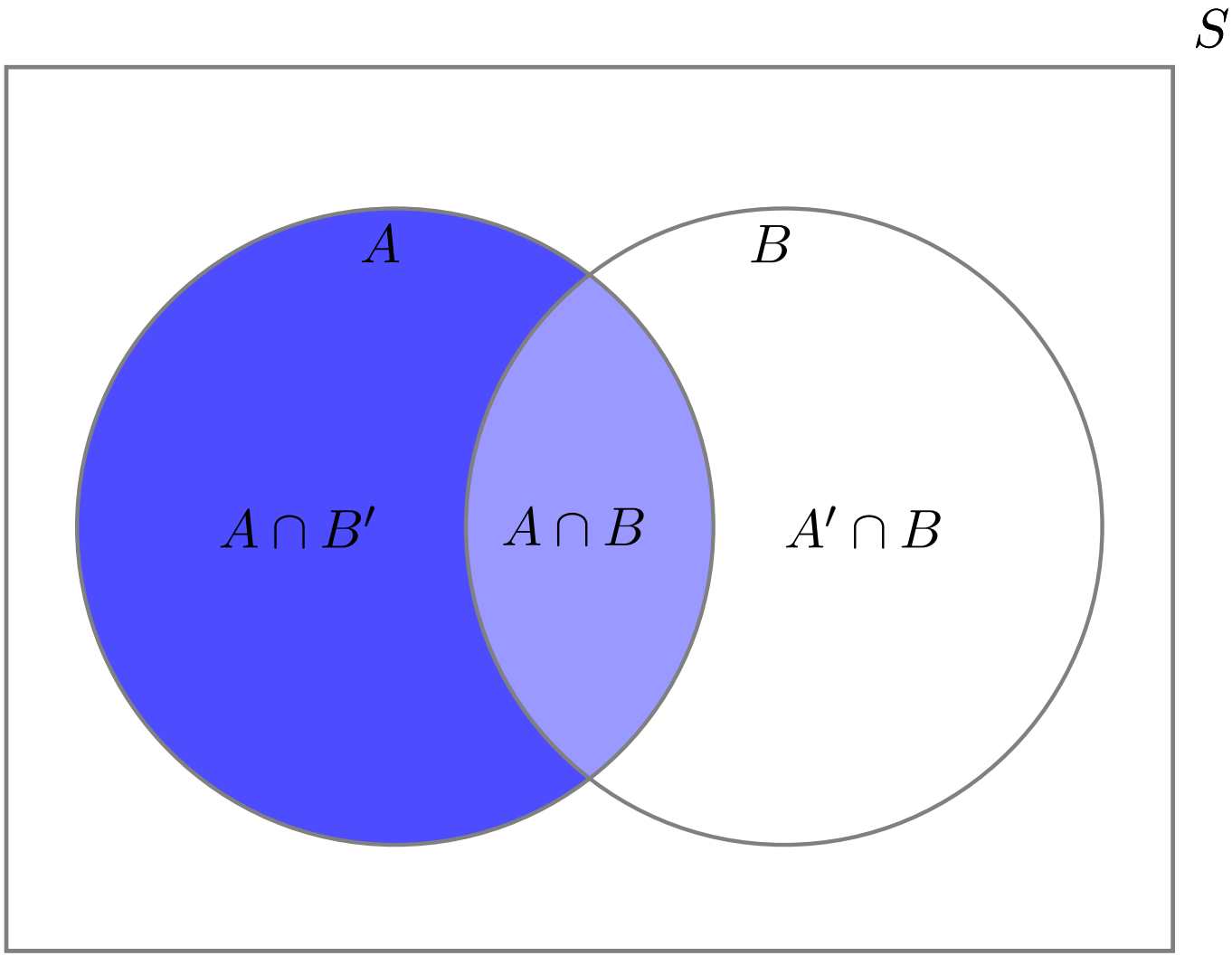
Such a tree extends to as many events as we need.

The Theorem of Total Probability

For two events A and B in sample space S, we have the partition of A as

$$A = (A \cap B) \cup (A \cap B').$$

The Theorem of Total Probability (cont.)



The Theorem of Total Probability (cont.)

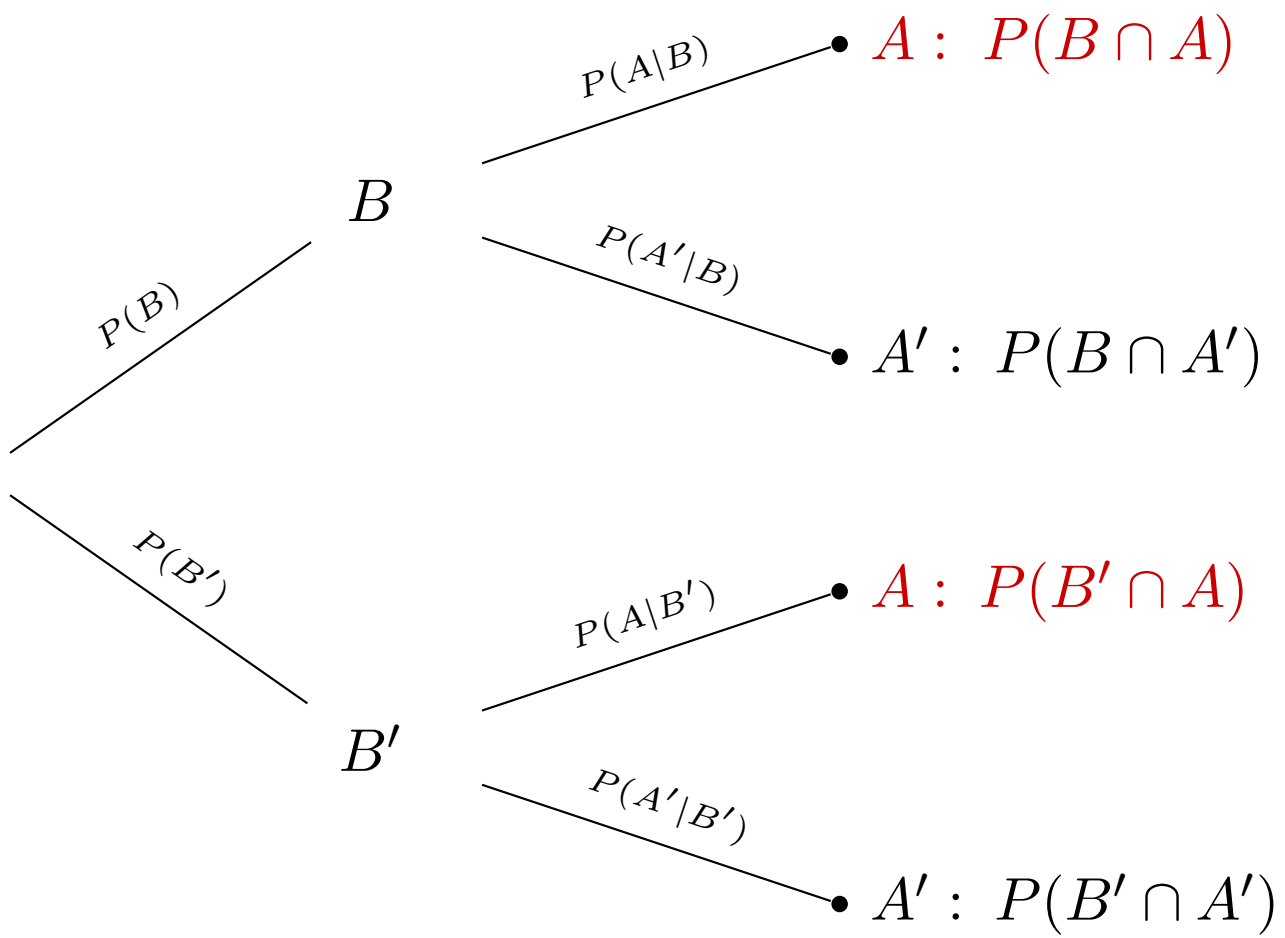
Therefore

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

and using the multiplication rule, we may rewrite this as

$$P(A) = P(A|B)P(B) + P(A|B')P(B').$$

The Theorem of Total Probability (cont.)



The Theorem of Total Probability (cont.)

That is, there are two ways to get to A:

- via B;
- via B'.

To compute $P(A)$ we add up all the probabilities on paths that end up at A.

Note that B and B' together form a partition of S.

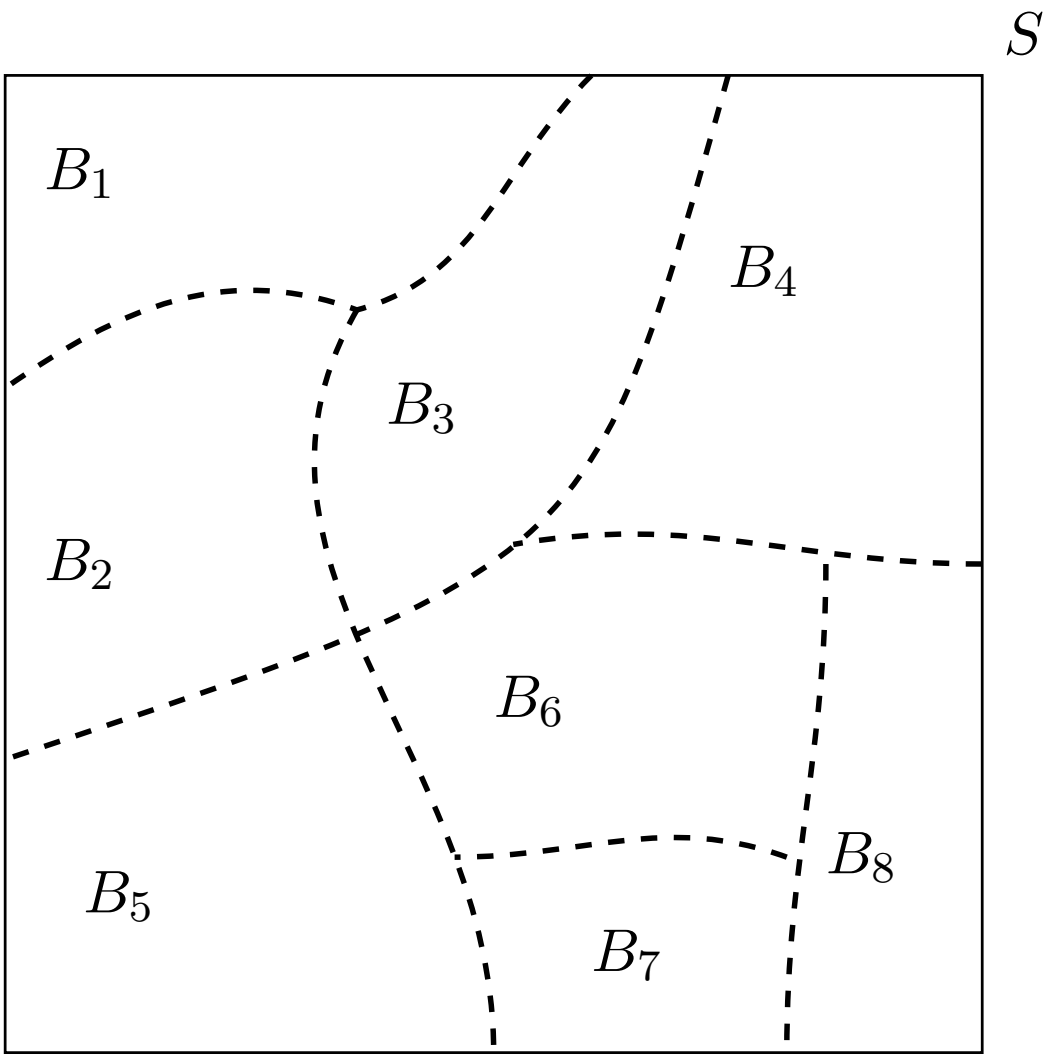
The Theorem of Total Probability (cont.)

Now suppose we have a partition into n subsets

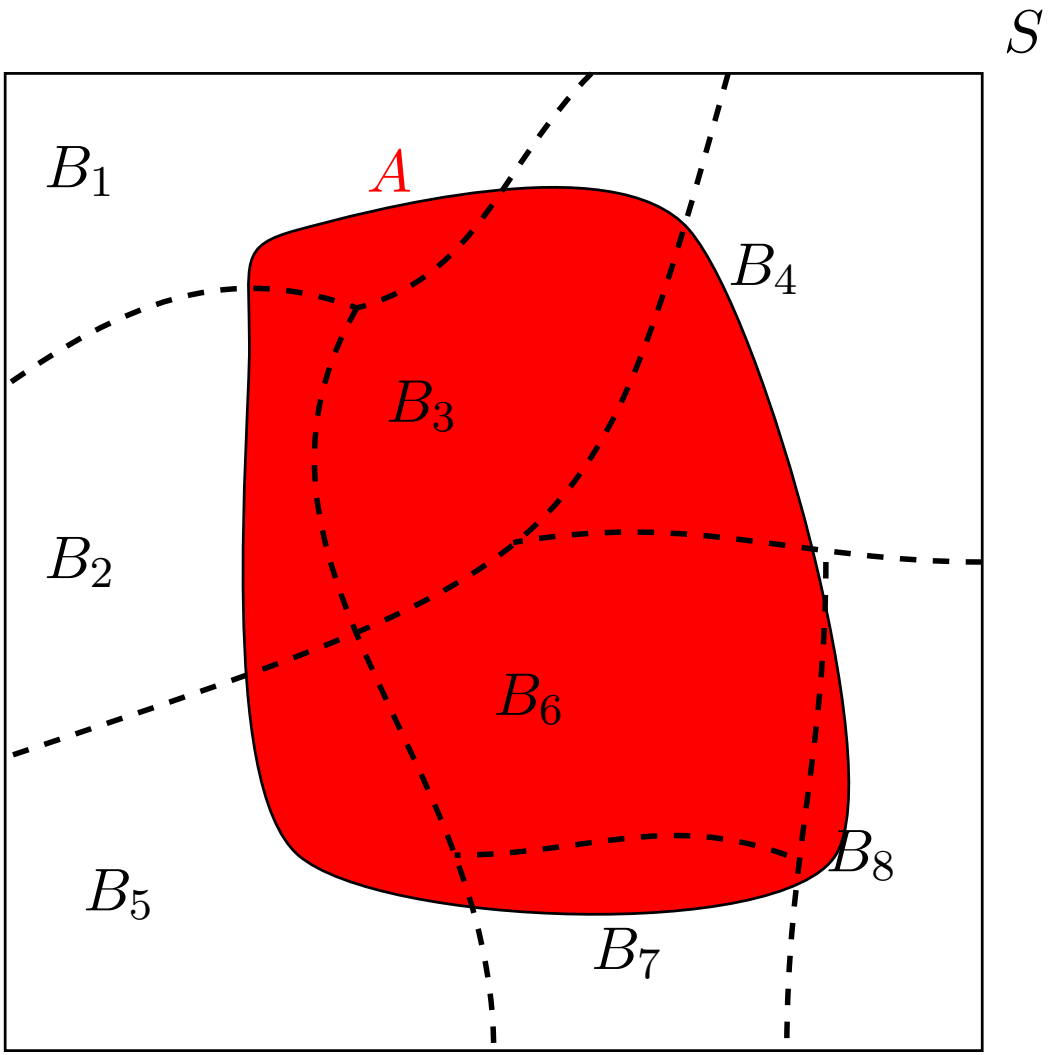
$$B_1, B_2, \ldots, B_n.$$

Again, these events also partition A.

The Theorem of Total Probability (cont.)



The Theorem of Total Probability (cont.)



The Theorem of Total Probability (cont.)

That is, we have that

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n)$$

and

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n).$$

The Theorem of Total Probability (cont.)

Using the definition of conditional probability, we therefore have

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n).$$

that is

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

This result is known as the *Theorem of Total Probability*.

The Theorem of Total Probability (cont.)

Notes

- The formula assumes that $P(B_i) > 0$ for $i = 1, \ldots, n$.
- It might be that

$$P(A \cap B_i) = P(A|B_i) = 0.$$

  for some i.

- This formula is a mathematical encapsulation of the probability tree when it is used to compute $P(A)$.

The Theorem of Total Probability (cont.)

Example (Three bags)

Suppose that an experiment involves selecting a ball from one of three bags. Let

- Bag 1: 4 red and 4 white balls.
- Bag 2: 1 red and 10 white balls.
- Bag 3: 7 red and 11 white balls.

A bag is selected (with all bags equally likely), and then a ball is selected from that bag (with all balls equally likely).

What is the probability that the ball selected is red ?

The Theorem of Total Probability (cont.)

Example (Three bags)

Let

- S: all possible selections of balls.
- B_1: bag 1 selected.
- B_2: bag 2 selected.
- B_2: bag 3 selected.
- A: ball selected is red.

The Theorem of Total Probability (cont.)

Example (Three bags)

We have that

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\ &= \frac{4}{8} \times \frac{1}{3} + \frac{1}{11} \times \frac{1}{3} + \frac{7}{18} \times \frac{1}{3} \\ &= \frac{97}{297} \end{aligned}$$

Example (Three bags)

Note that this is **not** equal to

$$\frac{\text{Total number of red balls}}{\text{Total number of balls}}$$

that is

$$\frac{4+1+7}{8+11+18} = \frac{12}{37}$$

as the red balls are not equally likely to be selected – this is only true conditional on the bag selection.

Bayes Theorem

The second great theorem of probability is *Bayes Theorem*.

- Named after Reverend Thomas Bayes (1701 – 1761)

  `https://en.wikipedia.org/wiki/Thomas_Bayes`

- Could be written *Bayes's Theorem*.
- It is not really a theorem.

Bayes Theorem (cont.)

Bayes Theorem: For two events A and B in sample space S, with $P(A) > 0$ and $P(B) > 0$,

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

If $0 < P(B) < 1$, we may write by the Theorem of Total Probability.

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

Bayes Theorem (cont.)

"Proof": By the definition of conditional probability

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

Then as $P(A) > 0$ and $P(B) > 0$ we can cross multiply and write

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

If $0 < P(B) < 1$, then we can legitimately write

$$P(A) = P(A|B)P(B) + P(A|B')P(B')$$

and substitute this expression in the denominator.

Bayes Theorem (cont.)

General version: Suppose that $B_1, B_2, \ldots, B_n$ form a partition of S, with $P(B_j) > 0$ for $j = 1, \ldots, n$. Suppose that A is an event in S with $P(A) > 0$.

Then for $i = 1, \ldots, n$

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}.$$

Bayes Theorem (cont.)

Probability tree interpretation: Given that we end up at A, what is the probability that we got there via branch B_i ?

Bayes Theorem (cont.)

Example (Three bags)

Suppose that an experiment involves selecting a ball from one of three bags. Let

- Bag 1: 4 red and 4 white balls.
- Bag 2: 1 red and 10 white balls.
- Bag 3: 7 red and 11 white balls.

A bag is selected (with all bags equally likely), and then a ball is selected from that bag (with all balls equally likely).

The ball selected is red. What is the probability it came from Bag 2 ?

Bayes Theorem (cont.)

Example (Three bags)

Let

- S: all possible selections of balls.
- B_1: bag 1 selected.
- B_2: bag 2 selected.
- B_2: bag 3 selected.
- A: ball selected is red.

Example (Three bags)

Recall that

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\ &= \frac{4}{8} \times \frac{1}{3} + \frac{1}{11} \times \frac{1}{3} + \frac{7}{18} \times \frac{1}{3} \\ &= \frac{97}{297} \end{aligned}$$

Example (Three bags)

Then

$$P(B_2|A) = \frac{P(A|B_2)P(B_2)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)}$$

$$= \frac{\frac{1}{11} \times \frac{1}{3}}{\frac{4}{8} \times \frac{1}{3} + \frac{1}{11} \times \frac{1}{3} + \frac{7}{18} \times \frac{1}{3}}$$

$$= \frac{1/33}{97/297}$$

$$= \frac{9}{97} \doteq 0.09278.$$

Example (Three bags)

We can also compute that

$$P(B_1|A) = \frac{4/24}{97/297} = \frac{297}{582} \simeq 0.51031$$

and

$$P(B_3|A) = \frac{21/54}{97/297} = \frac{1746}{2079} \simeq 0.39691.$$

Bayes Theorem (cont.)

Note

We have that

$$P(B_1|A) + P(B_2|A) + P(B_3|A) = 1$$

but note that

$$P(A|B_1) + P(A|B_2) + P(A|B_3) \neq 1.$$

In the first formula, we are conditioning on A everywhere; in the second, we have different conditioning sets.

Using Bayes Theorem

Bayes theorem is often used to make probability statements concerning an event B that has not been observed, given an event A that has been observed.

Using Bayes Theorem (cont.)

Example (Medical screening)

A health authority tests individuals from a population to assess whether they are sufferers from some disease. The screening test is not perfect as it produces

- *False positives:* declares someone to be a sufferer, when in reality they are a non-sufferer.
- *False negatives:* declares someone to be a non-sufferer, when in reality they are a sufferer.

Using Bayes Theorem (cont.)

Example (Medical screening)

Suppose that we denote by

- A – Screening test is positive;
- B – Person is actually a sufferer.

Using Bayes Theorem (cont.)

Example (Medical screening)

Suppose that

- $P(B) = p$;
- $P(A|B) = 1 - \alpha$ (true positive rate);
- $P(A|B') = \beta$ (false positive rate).

for probabilities p, α, β.

Using Bayes Theorem (cont.)

Example (Medical screening)

We have for the rate of positive tests

$$P(A) = P(A|B)P(B) + P(A|B')P(B') = (1-\alpha)p + \beta(1-p)$$

by the Theorem of Total Probability, and by Bayes theorem

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

$$= \frac{(1-\alpha)p}{(1-\alpha)p + \beta(1-p)}.$$

Using Bayes Theorem (cont.)

Example (Medical screening)

Similarly, for the false negative rate, we have

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

$$= \frac{\alpha p}{\alpha p + (1-\beta)(1-p)}.$$

Using Bayes Theorem (cont.)

Example (Medical screening)

We have

$$P(A|B) = (1-\alpha) \qquad P(B|A) = (1-\alpha)\frac{p}{(1-\alpha)p + \beta(1-p)}$$

and these two values are potentially very different.

Using Bayes Theorem (cont.)

Example (Medical screening)

That is, in general,

$$P(\text{“Spots”}|\text{“Measles”}) \neq P(\text{“Measles”}|\text{“Spots”})$$

Using Bayes Theorem (cont.)

Note

This phenomenon is sometimes known as the *Prosecutor's Fallacy*:

$$P(\text{“Evidence”}|\text{“Guilt”}) \neq P(\text{“Guilt”}|\text{“Evidence”})$$

Bayes Theorem and Odds

Recall that the odds on event B is defined by

$$\frac{P(B)}{P(B')} = \frac{P(B)}{1 - P(B)}$$

The *conditional odds* given event A (with $P(A) > 0$) is defined by

$$\frac{P(B|A)}{P(B'|A)} = \frac{P(B|A)}{1 - P(B|A)}$$

Bayes Theorem and Odds (cont.)

We have that

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

that is, the odds change by a factor

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

Bayes Theorem with multiple events

If A_1 and A_2 are two events in a sample space S so that

$$P(A_1 \cap A_2) > 0$$

and there is a partition of S via $B_1, \ldots, B_n$, with $P(B_i) > 0$, then

$$P(B_i|A_1 \cap A_2) = \frac{P(A_1 \cap A_2|B_i)P(B_i)}{\sum_{j=1}^{n} P(A_1 \cap A_2|B_j)P(B_j)}$$

- we use the previous version of Bayes Theorem with

$$A \equiv A_1 \cap A_2.$$

- this extends to conditioning events $A_1, A_2, \ldots, A_n$.

Bayes Theorem with multiple events (cont.)

Note that we have by earlier results

$$P(A_1 \cap A_2|B_i) = P(A_1|B_i)P(A_2|A_1 \cap B_i)$$

that is, the chain rule applies to conditional probabilities also.

Chapter 2: Random variables and probability distributions

Random variables

Up to this point we have dealt with probability assignments for events defined on a general sample space S.

We now focus on the specific case where S is a subset of the real numbers, $\mathbb{R}$. We could have S

- finite or countable $\{0, 1, 2, \ldots\}$;
- uncountable.

Events in such a sample space can be termed *numerical events*.

Random variables (cont.)

For most experiments defined on an arbitrary space S, we can usually consider a simple mapping, Y, defined on S that maps the sample outcomes onto $\mathbb{R}$, that is

$$Y : S \longrightarrow \mathbb{R}$$

$$s \longmapsto y$$

We are essentially saying that

$$Y(s) = y \qquad \text{for } s \in S.$$

The mapping Y is termed a *random variable*.

Example (Two coins)

Two coins are tossed with the outcomes being independent. The sample space is the list of sequences of possible outcomes

$$S = \{HH, HT, TH, TT\}.$$

with all sequences equally likely.

Define the random variable Y by

$$Y(s) = \text{“the number of heads in sequence } s\text{”}$$

Random variables (cont.)

Example (Two coins)

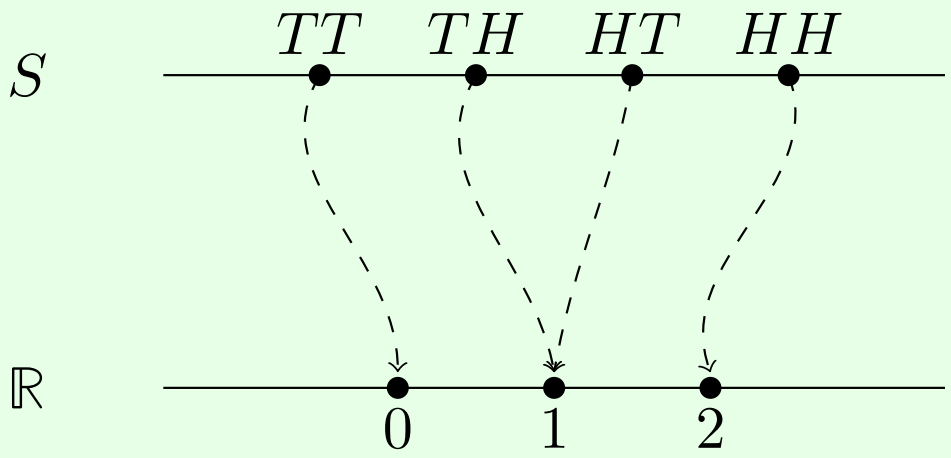
In the original probability specification for events in S,

$$P(\{HH\}) = P(\{HT\}) = P(\{TH\}) = P(\{TT\}) = \frac{1}{4}$$

Each $s \in S$ carries its probability with it through the mapping via $Y(.)$.

Random variables (cont.)

Example (Two coins)



Example (Two coins)

Then

$$Y(HH) = 2 \qquad Y(HT) = Y(TH) = 1 \qquad Y(TT) = 0$$

and $Y(.)$ maps S onto the set $\{0, 1, 2\}$. We have that

$$P(Y = 0) \equiv P(s \in \{TT\}) = \frac{1}{4}$$

$$P(Y = 1) \equiv P(s \in \{HT, TH\}) = \frac{2}{4}$$

$$P(Y = 2) \equiv P(s \in \{HH\}) = \frac{1}{4}$$

with $P(Y(s) = y) = 0$ for all other real values y.

Random variables (cont.)

Notes

- We typically consider the image of S under the mapping $Y(.)$ first
  - that is, we identify the possible values that Y can take.
- In general, $Y(.)$ is a many-to-one mapping.
- We usually suppress the dependence of $Y(.)$ on its argument, and merely write

$$P(Y = y).$$

Random variables (cont.)

Notes

We typically use upper case letters

$$X, Y, Z$$

etc. to denote random variables, and lower case letters

$$x, y, z$$

etc. to denote the real values that the random variable takes.

Discrete Random variables

A random variable Y is termed *discrete* if the possible values that Y can take is a **countable** set

$$y_1, y_2, \ldots, y_n, \ldots$$

Could be a

- finite set: eg $\{1, 2, \ldots, 100\}$
- countably infinite set: eg $\mathbb{Z}$.

I will denote the set of values that carry positive probability using the notation $\mathcal{Y}$.

Discrete Random variables (cont.)

In this case, the probabilities assigned to the individual points in the set $\mathcal{Y}$ in must carry all the probability:

$$\sum_{y \in \mathcal{Y}} P(Y = y) = 1.$$

This follows by the probability axioms as all the probability from S has been mapped into $\mathcal{Y}$.

Probability mass function

The *probability mass function* (pmf) for Y, $p(.)$, is the mathematical function that records how probability is distributed across points in $\mathbb{R}$.

That is,

$$p(y) = P(Y = y)$$

- If $y \in \mathcal{Y}$, $p(y) > 0$,
- otherwise $p(y) = 0$.

Probability mass function (cont.)

Note

The notation $f(y)$ is also commonly used, and also it is sometimes helpful to write

$$p_Y(y) \quad \text{or} \quad f_Y(y)$$

to indicate that we are dealing with the pmf for Y.

Probability mass function (cont.)

The pmf could be defined pointwise: for example

$$p(y) = \begin{cases} 0.2 & y = -2 \\ 0.1 & y = 0 \\ 0.2 & y = 3 \\ 0.5 & y = 5 \\ 0 & \text{otherwise} \end{cases}$$

Probability mass function (cont.)

Alternatively it could be defined using some functional form:

$$p(y) = \begin{cases} \frac{y}{15} & y \in \{1, 2, 3, 4, 5\} \\ \\ 0 & \text{otherwise} \end{cases}$$

Probability mass function (cont.)

The pmf **must** be specified to satisfy two conditions:

(i) The function must specify probabilities: that is

$$0 \leq p(y) \leq 1 \qquad \text{for all } y$$

(ii) The function must sum to 1 when evaluated at points in $\mathcal{Y}$:

$$\sum_{y \in \mathcal{Y}} P(Y = y) = 1.$$

Probability mass function (cont.)

Example (Combinatorial pmf)

A bag contains 7 red balls and 5 black balls. Four balls are selected, with all sequences of selections being equally likely.

Let Y be the random variable recording the number of red balls selected.

What is

$$p(y) = P(Y = y)$$

for $y \in \mathbb{R}$?.

Probability mass function (cont.)

Example (Combinatorial pmf)

First consider $\mathcal{Y}$; we have that

$$\mathcal{Y} = \{0, 1, 2, 3, 4\}$$

- Y must be a non-negative integer;
- there are only four balls in the selection;
- it is possible to select four black balls.

Probability mass function (cont.)

Example (Combinatorial pmf)

What is $P(Y = y)$ for $y \in \{0, 1, 2, 3, 4\}$?

$$(Y = y) \equiv A = \{y \text{ red balls selected out of four}\}$$

That is,

$$(Y = y) \equiv A = \{y \text{ red balls, and } 4 - y \text{ black balls selected}\}$$

Example (Combinatorial pmf)

We compute $P(A)$ as follows: for $y \in \{0, 1, 2, 3, 4\}$

$$n_A = \binom{7}{y} \times \binom{5}{4-y}$$

$\binom{7}{y}$: number of ways of selecting the y red balls from the 7 available.

$\binom{5}{4-y}$: number of ways of selecting the $4-y$ black balls from the 5 available.

Example (Combinatorial pmf)

Also

$$n_S = \binom{12}{4}$$

which is the number of ways of selecting the 4 balls.

Therefore

$$P(Y = y) = P(A) = \frac{n_A}{n_S} = \frac{\binom{7}{y}\binom{5}{4-y}}{\binom{12}{4}} \quad y = 0, 1, 2, 3, 4$$

and zero for all other values of y.

Probability mass function (cont.)

Example (Combinatorial pmf)

y	0	1	2	3	4
Prob.	$\frac{\binom{7}{0}\binom{5}{4}}{\binom{12}{4}}$	$\frac{\binom{7}{1}\binom{5}{3}}{\binom{12}{4}}$	$\frac{\binom{7}{2}\binom{5}{2}}{\binom{12}{4}}$	$\frac{\binom{7}{3}\binom{5}{1}}{\binom{12}{4}}$	$\frac{\binom{7}{4}\binom{5}{0}}{\binom{12}{4}}$
Value	0.01010	0.14141	0.42424	0.35354	0.07071

Probability mass function (cont.)

Note

The collection of y values for which $p(y) > 0$ may be a countably infinite set, say

$$\mathcal{Y} = \{y_1, y_2, \ldots\}.$$

In this case we would require the infinite sum

$$\sum_{i=1}^{\infty} p(y_i)$$

be equal to 1.

Probability mass function (cont.)

Note

All pmfs take the form

$$\text{constant} \times \text{function of } y$$

or

$$cg(y)$$

say, where $g(y)$ also contains information on the range of values, $\mathcal{Y}$, for which $p(y) > 0$. We have

$$\sum_{y \in \mathcal{Y}} p(y) = 1 \qquad \Longrightarrow \qquad \sum_{y \in \mathcal{Y}} g(y) = \frac{1}{c}$$

Expectations for discrete random variables

The *expectation* (or *expected value*) of a discrete random variable Y, denoted $\mathbb{E}[Y]$, is defined as

$$\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} y\, p(y)$$

whenever this sum is finite; if it is not finite, we say that the expectation does not exist.

Expectations for discrete random variables (cont.)

Notes

- $\mathbb{E}[Y]$ is the centre of mass of the probability distribution.
- If $\mathcal{Y} = \{y_1, y_2, \ldots, y_n, \ldots\}$, we have that

$$\mathbb{E}[Y] = y_1 p(y_1) + y_2 p(y_2) + \cdots + y_n p(y_n) + \cdots$$

- It follows that $\mathbb{E}[Y]$ is finite if

$$\sum_{y \in \mathcal{Y}} |y| \, p(y) < \infty$$

Expectations for discrete random variables (cont.)

Example (Expectation and Mean)

Suppose that $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$, and

$$p(y) = \frac{1}{n} \qquad y \in \mathcal{Y}$$

and zero otherwise. Then

$$\begin{aligned}\mathbb{E}[Y] &= y_1 p(y_1) + y_2 p(y_2) + \cdots + y_n p(y_n) \\ &= \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = \overline{y}.\end{aligned}$$

This is an example of a *discrete uniform distribution.*

Expectations for discrete random variables (cont.)

The discrete uniform distribution is a distribution with constant $p(y)$, that is,

$$p(y) = c \qquad y \in \mathcal{Y}$$

for some finite set $\mathcal{Y}$. It follows that

$$\frac{1}{c}$$

equals the number of elements in $\mathcal{Y}$.

The expectation of a function

Suppose $g(.)$ is a real-valued function, and consider $g(Y)$. For example, if Y takes the values $\{-2, -1, 0, 1, 2\}$, and

$$g(t) = t^2$$

then $g(Y)$ takes the values $\{0, 1, 4\}$. If $p(y)$ denotes the pmf of Y, then

$$P(g(Y) = t) = \begin{cases} p(0) & t = 0 \\ p(-1) + p(1) & t = 1 \\ p(-2) + p(2) & t = 4 \\ 0 & \text{otherwise} \end{cases}$$

The expectation of a function (cont.)

Suppose that we denote the set of possible values of $g(Y)$ by

$$\mathcal{G} = \{g_i, i = 1, \ldots, m : g_i = g(y), \text{some } y \in \mathcal{Y}\}.$$

We can define the expectation of $g(Y)$ as

$$\mathbb{E}[g(Y)] = \sum_{y \in \mathcal{Y}} g(y)p(y).$$

The expectation of a function (cont.)

We see that this is the correct formula by considering the random variable $g(Y)$; this is a discrete random variable itself, with pmf given by

$$P(g(Y) = g_i) = \sum_{\substack{y \in \mathcal{Y}: \\ g(y)=g_i}} p(y) = p^*(g)$$

say.

The expectation of a function (cont.)

Then

$$E[g(Y)] = \sum_{i=1}^{m} g_i p^*(g_i)$$

$$= \sum_{i=1}^{m} g_i \left\{ \sum_{\substack{y \in \mathcal{Y}: \\ g(y)=g_i}} p(y) \right\}$$

$$= \sum_{y \in \mathcal{Y}} g(y)p(y).$$

Variance

The *variance* of random variable Y is denoted $\mathbb{V}[Y]$, and is defined by

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mu)^2]$$

where $\mu = \mathbb{E}[Y]$. The *standard deviation* of Y is

$$\sqrt{\mathbb{V}[Y]}.$$

Note that it is possible that $\mathbb{V}[Y]$ is not finite, even if $\mathbb{E}[Y]$ is finite.

Variance (cont.)

The variance is a measure of how "dispersed" the probability distribution is.

Note that

$$\mathbb{V}[Y] = \mathbb{E}[(Y - \mu)^2] = \sum_y (y - \mu)^2 p(y) \geq 0$$

as $p(y) \geq 0$ and $(y - \mu)^2 \geq 0$ for all y.

Variance (cont.)

In fact, $\mathbb{V}[Y] > 0$ unless the distribution of Y is *degenerate*, that is, for some constant c

$$p(c) = 1$$

in which case $\mathbb{V}[Y] = 0$.

Properties of expectations

For discrete random variable Y with expectation $\mathbb{E}[Y]$, and real constants a and b, we have that

$$\mathbb{E}[aY + b] = a\mathbb{E}[Y] + b.$$

This result follows from the previous result for $g(Y)$ with

$$g(t) = at + b.$$

Properties of expectations (cont.)

This follows as

$$\begin{aligned}\mathbb{E}[aY+b] &= \sum_y (ay+b)p(y) \\ &= a\sum_y yp(y) + b\sum_y p(y) \\ &= a\mathbb{E}[Y] + b\end{aligned}$$

as

$$\sum_y p(y) = 1.$$

Properties of expectations (cont.)

Now suppose we have

$$g(Y) = ag_1(Y) + bg_2(Y)$$

for two separate functions $g_1(.)$ and $g_2(.)$. Then

$$\begin{aligned}
\mathbb{E}[g(Y)] &= \mathbb{E}[ag_1(Y) + bg_2(Y)] \\
&= \sum_y (ag_1(y) + bg_2(y))p(y) \\
&= a\sum_y g_1(y)p(y) + b\sum_y g_2(y)p(y) \\
&= a\mathbb{E}[g_1(Y)] + b\mathbb{E}[g_2(Y)].
\end{aligned}$$

Properties of expectations (cont.)

Note

These results merely follow standard mathematical concepts of the *linearity* of summations:

$$\sum_i (a_i + b_i) = \sum_i a_i + \sum_i b_i.$$

Properties of expectations (cont.)

For the variance, if we denote $\mathbb{E}[Y] = \mu$, then

$$\begin{aligned}\mathbb{E}[(Y-\mu)^2] &= \mathbb{E}[Y^2 - 2\mu Y + \mu^2] \\ &= \mathbb{E}[Y^2] - 2\mu\mathbb{E}[Y] + \mu^2 \\ &= \mathbb{E}[Y^2] - 2\mu^2 + \mu^2 \\ &= \mathbb{E}[Y^2] - \mu^2.\end{aligned}$$

In this formula

$$\mathbb{E}[Y^2] = \sum_y y^2\, p(y).$$

Properties of expectations (cont.)

From this we can deduce that

$$\mathbb{V}[aY + b] = a^2\mathbb{V}[Y].$$

We have that

$$\mathbb{E}[aY + b] = a\mathbb{E}[Y] + b = a\mu + b$$

so

$$\mathbb{V}[aY + b] = \mathbb{E}[(aY + b - (a\mu + b))^2]$$

which reduces to

$$\mathbb{E}[(aY - a\mu)^2] = \mathbb{E}[a^2(Y - \mu)^2] = a^2\mathbb{E}[(Y - \mu)^2].$$

The binomial distribution

We can now consider specific pmfs that correspond to particular experimental settings.

On simple pmf is obtained if we consider an experiment with two possible outcomes, which we can refer to as 'success' and 'failure' respectively. We can define the random variable Y by

$$Y = 1 \text{ if outcome is 'success'} \qquad Y = 0 \text{ if outcome is 'failure'}$$

and then consider

$$P(Y = 0) \text{ and } P(Y = 1).$$

The binomial distribution (cont.)

Note however that

$$P(Y = 0) + P(Y = 1) = 1$$

so if we specify

$$P(Y = 1) = p$$

then automatically

$$P(Y = 0) = 1 - p = q$$

say.

The binomial distribution (cont.)

We may then write

$$p(y) = p^y(1-p)^{1-y} = p^y q^{1-y} \qquad y \in \{0, 1\}$$

with $p(y) = 0$ for all other values of y.

This distribution is known as the *Bernoulli* distribution with parameter p.

Note

To avoid the trivial case, we typically assume that $0 < p < 1$.

The binomial distribution (cont.)

This model covers many experimental settings:

- coin tosses;
- thumbtack tosses;
- pass/fail;
- cure/not cure

etc.

The binomial distribution (cont.)

We have for this model that

$$\mu \equiv \mathbb{E}[Y] = \sum_{i=0}^{1} yp(y) = (0 \times p(0) + 1 \times p(1))$$

$$= (0 \times q + 1 \times p) = p$$

The binomial distribution (cont.)

Similarly

$$\mathbb{E}[Y^2] = \sum_{i=0}^{1} y^2 p(y) = (0^2 \times p(0) + 1^2 \times p(1))$$

$$= (0 \times q + 1 \times p) = p$$

so therefore

$$\mathbb{V}[Y] = \mathbb{E}[Y^2] - \mu^2 = p - p^2 = p(1-p).$$

The binomial distribution (cont.)

We now consider an extended version of the experiment where a sequence of n identical binary experiments (or *'trials'*) with success probability p whose outcomes are independent is carried out.

$$0101011101$$

We consider the total number of successes in the sequence.

Provided $0 < p < 1$, we have that

$$\mathcal{Y} = \{0, 1, 2, \ldots, n\}$$

The binomial distribution (cont.)

Now if Y is the random variable that records the number of successes in the sequence, we have that

$$(Y = y) \quad \Longleftrightarrow \quad \text{“}y \text{ successes in } n \text{ trials”}$$

Consider a sequence that results in y successes and $n - y$ failures: for example, if $n = 8$ and $y = 3$, we might have

10101000
01010001
11100000

etc.

The binomial distribution (cont.)

All such sequences are equally likely, under the 'identical and independent' assumption. The probability of any such sequence is, for $y \in \{0, 1, 2, \ldots, n\}$,

$$p^y q^{n-y}$$

that is, in the examples above

$$pqpqpqqq$$
$$qpqpqqqp$$
$$pppqqqqq$$

etc. all yield $p^3 q^5$.

The binomial distribution (cont.)

However for the total number of successes, we do not care about the order of the outcomes; we must count the number of all such sequences, as the individual sequences partition the event of interest.

By the previous combinatorial arguments, we see that the number of such sequences is

$$C_y^n = \binom{n}{y} = \frac{n!}{y!(n-y)!}.$$

The binomial distribution (cont.)

Therefore we have that

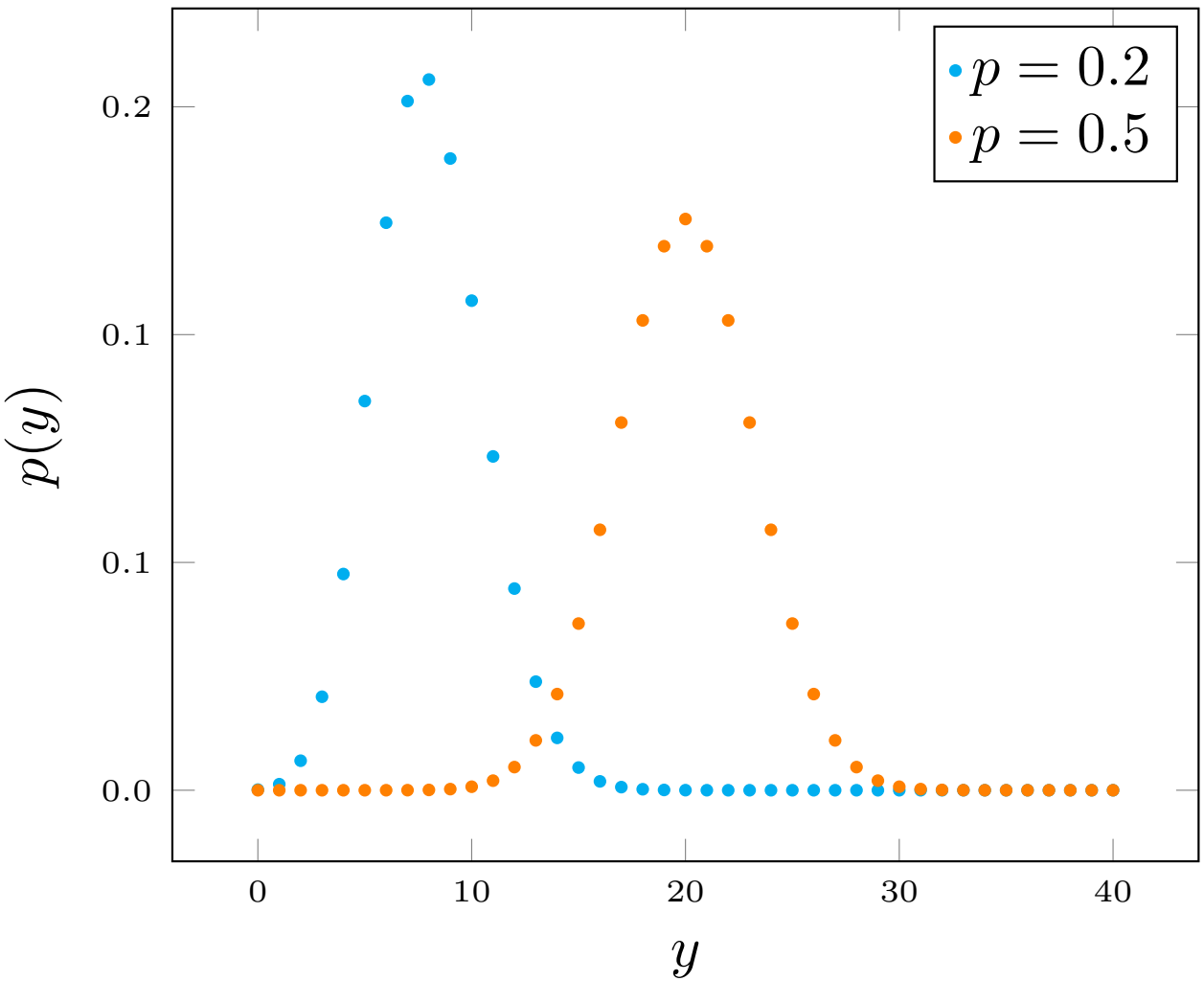
$$p(y) \equiv P(Y = y) = \binom{n}{y} p^y q^{n-y} \qquad y \in \{0, 1, 2, \ldots, n\}$$

and $p(y) = 0$ for all other y.

This is the pmf for the *Binomial distribution*.

The binomial distribution (cont.)

Two examples: $n = 40$, with $p = 0.2$ and $p = 0.5$.



The binomial distribution (cont.)

Expectation : For the expectation of a binomial distribution, we may compute directly as follows:

$$\begin{aligned}
\mathbb{E}[Y] = \sum_{y=0}^{n} yp(y) &= \sum_{y=0}^{n} y\binom{n}{y}p^y q^{n-y} \\
&= \sum_{y=1}^{n} y\frac{n!}{y!(n-y)!}p^y q^{n-y} \\
&= n\sum_{y=1}^{n} \frac{(n-1)!}{(y-1)!(n-y)!}p^y q^{n-y}
\end{aligned}$$

The binomial distribution (cont.)

Now we may write

$$(n-y) = (n-1) - (y-1)$$

so therefore the sum can be written

$$\sum_{y=1}^{n} \frac{(n-1)!}{(y-1)!((n-1)-(y-1))!} p^y q^{(n-1)-(y-1)}.$$

The binomial distribution (cont.)

Rewriting the terms in the sum by setting $j = y - 1$, this equals

$$\sum_{j=0}^{n-1} \frac{(n-1)!}{j!((n-1)-j)!} p^{j+1} q^{(n-1)-j}.$$

or equivalently, by the binomial theorem, this equals

$$p \sum_{j=0}^{n-1} \frac{(n-1)!}{j!((n-1)-j)!} p^{j} q^{(n-1)-j} = p(p+q)^{n-1}.$$

But $p + q = 1$, so therefore

$$\mathbb{E}[Y] = np.$$

The binomial distribution (cont.)

Variance : For the variance, we have by the previous result that

$$\mathbb{E}[Y^2] = \mathbb{E}[Y(Y-1)] + \mathbb{E}[Y].$$

Now, we have

$$\begin{aligned}
\mathbb{E}[Y(Y-1)] &= \sum_{y=0}^{n} y(y-1)p(y) \\
&= \sum_{y=0}^{n} y(y-1)\binom{n}{y}p^y q^{n-y} \\
&= \sum_{y=2}^{n} y(y-1)\frac{n!}{y!(n-y)!}p^y q^{n-y} \\
&= n(n-1)\sum_{y=2}^{n} \frac{(n-2)!}{(y-2)!(n-y)!}p^y q^{n-y}
\end{aligned}$$

The binomial distribution (cont.)

Using the same trick as for the expectation, and writing

$$(n-y) = (n-2) - (y-2)$$

we see that we can rewrite the sum

$$n(n-1)\sum_{y=2}^{n} \frac{(n-2)!}{(y-2)!((n-2)-(y-2))!} p^y q^{(n-2)-(y-2)}$$

$$= n(n-1)\sum_{j=0}^{n-2} \frac{(n-2)!}{j!((n-2)-j)!} p^{j+2} q^{(n-2)-j}$$

$$= n(n-1)p^2 \sum_{j=0}^{n-2} \frac{(n-2)!}{j!((n-2)-j)!} p^j q^{(n-2)-j}$$

$$= n(n-1)p^2(p+q)^{n-2} = n(n-1)p^2.$$

The binomial distribution (cont.)

Therefore

$$\mathbb{E}[Y(Y-1)] = n(n-1)p^2$$

and so

$$\mathbb{E}[Y^2] = \mathbb{E}[Y(Y-1)+Y] = \mathbb{E}[Y(Y-1)] + \mathbb{E}[Y] = n(n-1)p^2 + np.$$

Hence

$$\begin{aligned}\mathbb{V}[Y] &= \mathbb{E}[Y^2] - \mu^2 = n(n-1)p^2 + np - (np)^2 \\ &= -np^2 + np \\ &= np(1-p) \\ &= npq.\end{aligned}$$

The binomial distribution (cont.)

Note

Recall that for a single binary experiment, for the total number of successes Y we have that

$$\mathbb{E}[Y] = p \qquad \mathbb{V}[Y] = pq$$

whereas for n identical and independent binary experiments, for the total number of successes Y we have that

$$\mathbb{E}[Y] = np \qquad \mathbb{V}[Y] = npq.$$

The binomial distribution (cont.)

If $n = 1$, we have that

$$p(y) = p^y q^{1-y} \qquad y \in \{0, 1\}$$

and $p(y) = 0$ for all other y. This special case is known as the *Bernoulli* distribution with parameter p.

In this case we consider the number of successes in single binary trial. The expectation is p and the variance is pq.

The geometric distribution

We now consider another experimental setting based on identical and independent binary trials. Here, rather than counting the total number of successes, we count the number of trials carried out until the first success.

Using the 0/1 notation, we consider sequences such as

0001
01
000000001
1
000001

etc.

The geometric distribution (cont.)

If random variable Y records the number of trials carried out up to an including the first success, then

$$p(y) = P(Y = y) = q^{y-1}p \qquad y \in \{1, 2, 3, \ldots\}$$

and zero for other values of y. This follows as

$$Y = y \quad \Longleftrightarrow \quad \text{“}y - 1 \text{ failures, followed by a success”}$$

that is, for example, we observe a sequence 00001 which occurs with probability

$$q \times q \times q \times q \times p = q^4 p.$$

This is the pmf of the *geometric distribution*.

The geometric distribution (cont.)

Notes

- Note that

$$\sum_{y=1}^{\infty} q^{y-1}p = p\sum_{j=0}^{\infty} q^j = p\frac{1}{1-q} = p\frac{1}{(1-(1-p))} = 1.$$

- The geometric distribution is an example of a waiting-time distribution.

The geometric distribution (cont.)

Notes

- Sometimes the geometric distribution is written

$$p(y) = q^y p \qquad y \in \{0, 1, 2, \ldots\}$$

– this is the same distribution, but for the random variable that records the number of failures observed up to the first success, that is, for the random variable

$$Y^* = Y - 1.$$

.

The geometric distribution (cont.)

Expectation:

$$\mathbb{E}[Y] = \sum_{y=1}^{\infty} yp(y) = \sum_{y=1}^{\infty} yq^{y-1}p = p\sum_{y=1}^{\infty} yq^{y-1}.$$

We have that

$$\sum_{y=1}^{\infty} yq^{y-1} = \frac{1}{(1-q)^2} = \frac{1}{(1-(1-p))^2} = \frac{1}{p^2}.$$

Therefore

$$\mathbb{E}[Y] = p\frac{1}{p^2} = \frac{1}{p}.$$

The geometric distribution (cont.)

Variance: Using the same approach as for the binomial distribution, we have

$$\mathbb{E}[Y(Y-1)] = \sum_{y=1}^{\infty} y(y-1)p(y) = pq \sum_{y=2}^{\infty} y(y-1)q^{y-2}.$$

Now, we have the following calculus results for a convergent sum

$$\frac{d}{dt}\left\{\sum_{j=0}^{\infty} t^j\right\} = \sum_{j=1}^{\infty} jt^{j-1}$$

$$\frac{d^2}{dt^2}\left\{\sum_{j=0}^{\infty} t^j\right\} = \sum_{j=2}^{\infty} j(j-1)t^{j-2}.$$

The geometric distribution (cont.)

But

$$\sum_{j=0}^{\infty} t^j = \frac{1}{1-t}$$

so therefore we must have that

$$\frac{d}{dt}\left\{\sum_{j=0}^{\infty} t^j\right\} = \frac{d}{dt}\left\{\frac{1}{1-t}\right\} = \frac{1}{(1-t)^2}$$

$$\frac{d^2}{dt^2}\left\{\sum_{j=0}^{\infty} t^j\right\} = \frac{d^2}{dt^2}\left\{\frac{1}{1-t}\right\} = \frac{2}{(1-t)^3}.$$

The geometric distribution (cont.)

Thus

$$\sum_{y=2}^{\infty} y(y-1)q^{y-2} = \frac{2}{(1-q)^3} = \frac{2}{p^3}$$

and hence

$$\mathbb{E}[Y(Y-1)] = \sum_{y=1}^{\infty} y(y-1)pq^{y-1} = pq\frac{2}{p^3} = \frac{2(1-p)}{p^2}.$$

The geometric distribution (cont.)

Therefore

$$\begin{aligned}\mathbb{V}[Y] &= \mathbb{E}[Y(Y-1)] + \mathbb{E}[Y] - \{\mathbb{E}[Y]\}^2 \\ &= \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} \\ &= \frac{(1-p)}{p^2}.\end{aligned}$$

The geometric distribution (cont.)

Now consider the probability

$$P(Y > y)$$

for $y \in \{1, 2, 3, \ldots\}$. Using a partitioning argument, we see that

$$P(Y > y) = \sum_{j=y+1}^{\infty} P(Y = j).$$

That is

$$P(Y > y) = \sum_{j=y+1}^{\infty} q^{j-1}p = pq^y \sum_{j=y+1}^{\infty} q^{j-y-1} = pq^y \frac{1}{1-q}.$$

The geometric distribution (cont.)

Therefore

$$P(Y > y) = q^y \qquad y \in \{1, 2, 3, \ldots\}$$

and consequently

$$P(Y \leq y) = 1 - q^y = 1 - (1 - p)^y \qquad y \in \{1, 2, 3, \ldots\}.$$

The geometric distribution (cont.)

We can define the function

$$F(y) = P(Y \leq y) = 1 - (1-p)^y \qquad y \in \{1, 2, 3, \ldots\}.$$

$F(y)$ is known as the *cumulative distribution function* (cdf) of the distribution.

The geometric distribution (cont.)

We can define all such 'interval' probabilities such as

$$P(Y \leq a) \qquad P(Y \geq b) \qquad P(c \leq Y \leq d)$$

etc. via $F(y)$. This is because for **all** discrete distributions and sets $\mathcal{X}$

$$P(Y \in \mathcal{X}) = \sum_{y \in \mathcal{X}} p(y).$$

The geometric distribution (cont.)

Note

We can compute

$$F(y) = P(Y \leq y)$$

for any real value y. However, we must remember that

$$P(Y \leq y) = P(Y \leq \lfloor y \rfloor)$$

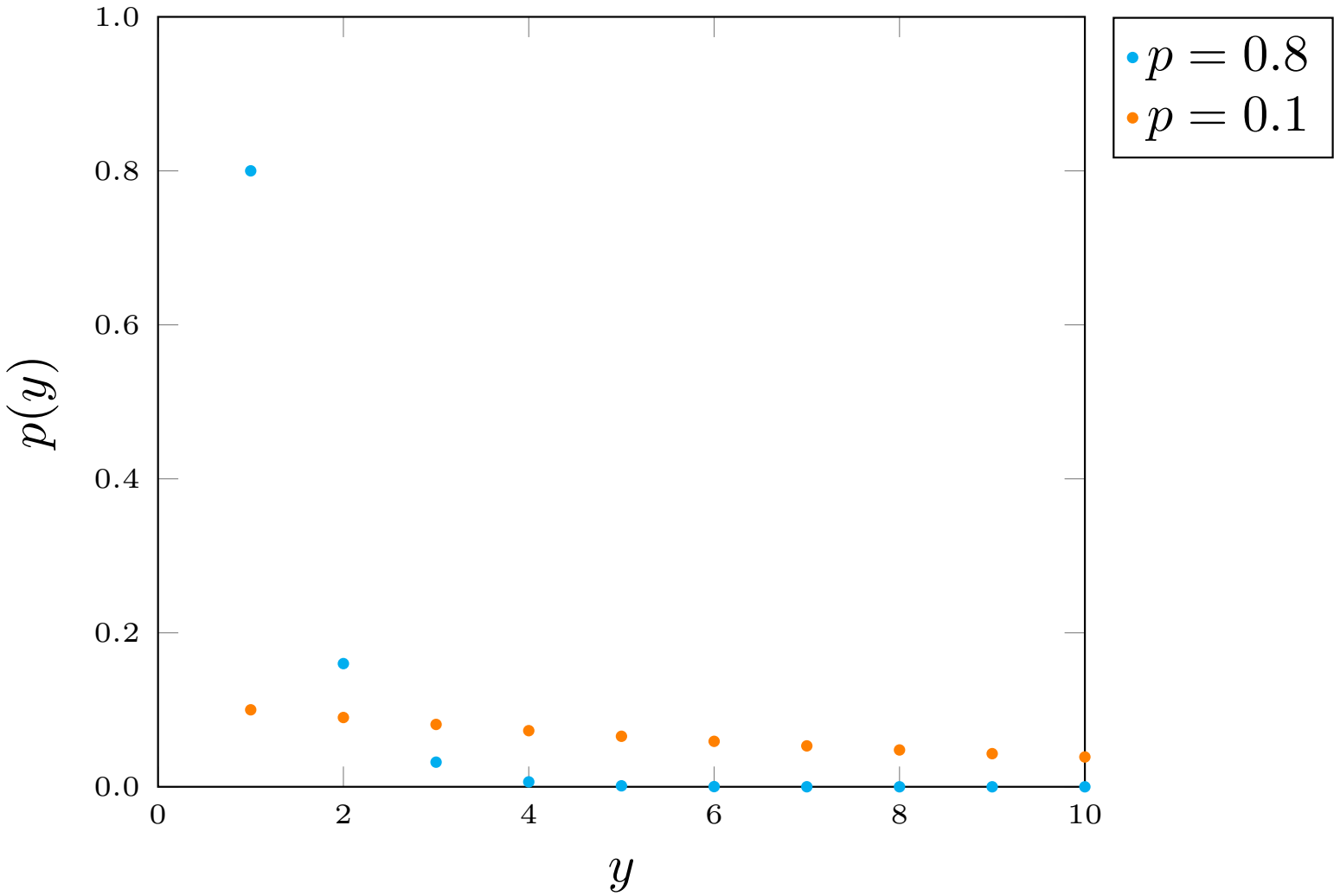
where $\lfloor y \rfloor$ is the largest integer that is no greater than y. For example,

$$F(1.234) = F(1) \qquad F(4.789) = F(4)$$

etc.

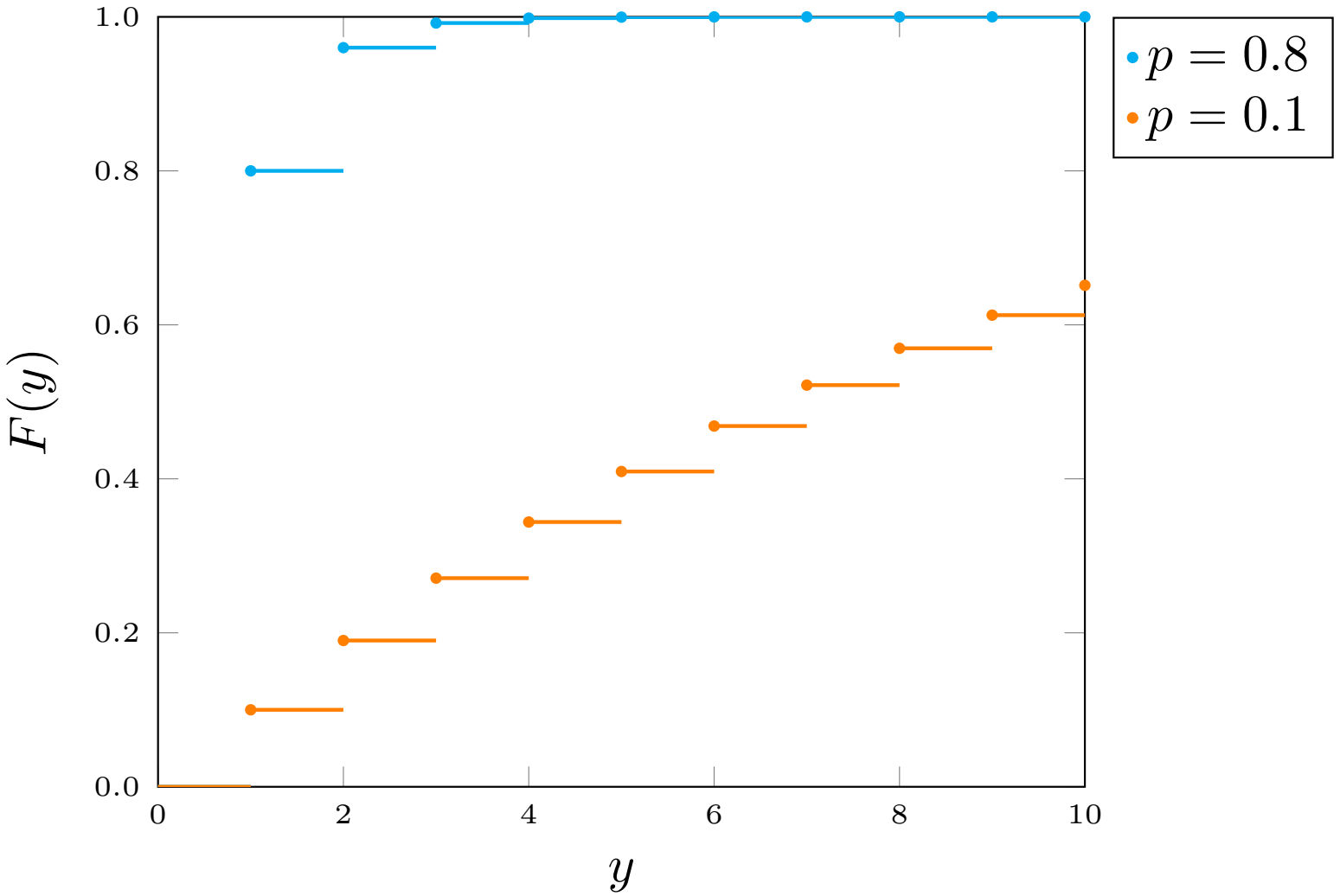
The geometric distribution (cont.)

Geometric $p(y)$: with $p = 0.8$ and $p = 0.1$.



The geometric distribution (cont.)

Geometric $F(y)$: with $p = 0.8$ and $p = 0.1$.



The negative binomial distribution

As an extension to the geometric distribution setting, we now consider a sequence of identical and independent binary trials, but we continue until we observe the rth success, where $r = \{1, 2, \ldots, \}$ is a fixed number. That is, if $r = 3$, we consider sequences such as

00010101
0111
010000011
10101
0000010101

etc.

The negative binomial distribution (cont.)

If random variable Y records the number of trials carried out up to an including the rth success, then we see that

$$(Y = y) \equiv A \cap B$$

where

$$A = \text{“}(r-1) \text{ successes in the first } (y-1) \text{ trials”}$$
$$B = \text{“a success on the } y\text{th trial”}$$

with A and B independent events.

The negative binomial distribution (cont.)

We have that Y can take values on the set

$$r, r+1, \ldots$$

Now, we have from the binomial distribution that

$$P(A) = \binom{y-1}{r-1} p^{r-1} q^{(y-1)-(r-1)} \qquad y = \{r, r+1, r+2, \ldots\}$$

and

$$P(B) = p$$

The negative binomial distribution (cont.)

Therefore, as

$$(y-1)-(r-1)=y-r$$

we have that

$$p(y)=P(Y=y)=\binom{y-1}{r-1}p^r q^{y-r} \qquad y=r,r+1,r+2,\ldots$$

and zero for other values of y.

This is the pmf of the *negative binomial distribution.*

The negative binomial distribution (cont.)

Notes

- We can show (Exercise)

$$\sum_{y=r}^{\infty} p(y) = 1.$$

- The negative binomial distribution is another example of a waiting-time distribution.

The negative binomial distribution (cont.)

Notes

- Sometimes the negative binomial distribution is written

$$p(y) = \binom{y+r-1}{r-1} p^r q^y \qquad y = 0, 1, 2, \ldots$$

– this is the same distribution, but for the random variable that records the number of failures observed up to the rth success, that is, for the random variable

$$Y^* = Y - r.$$

- Note that by definition

$$\binom{y+r-1}{r-1} = \binom{y+r-1}{y}$$

The negative binomial distribution (cont.)

For random variable Y, we have the following results:

Expectation:

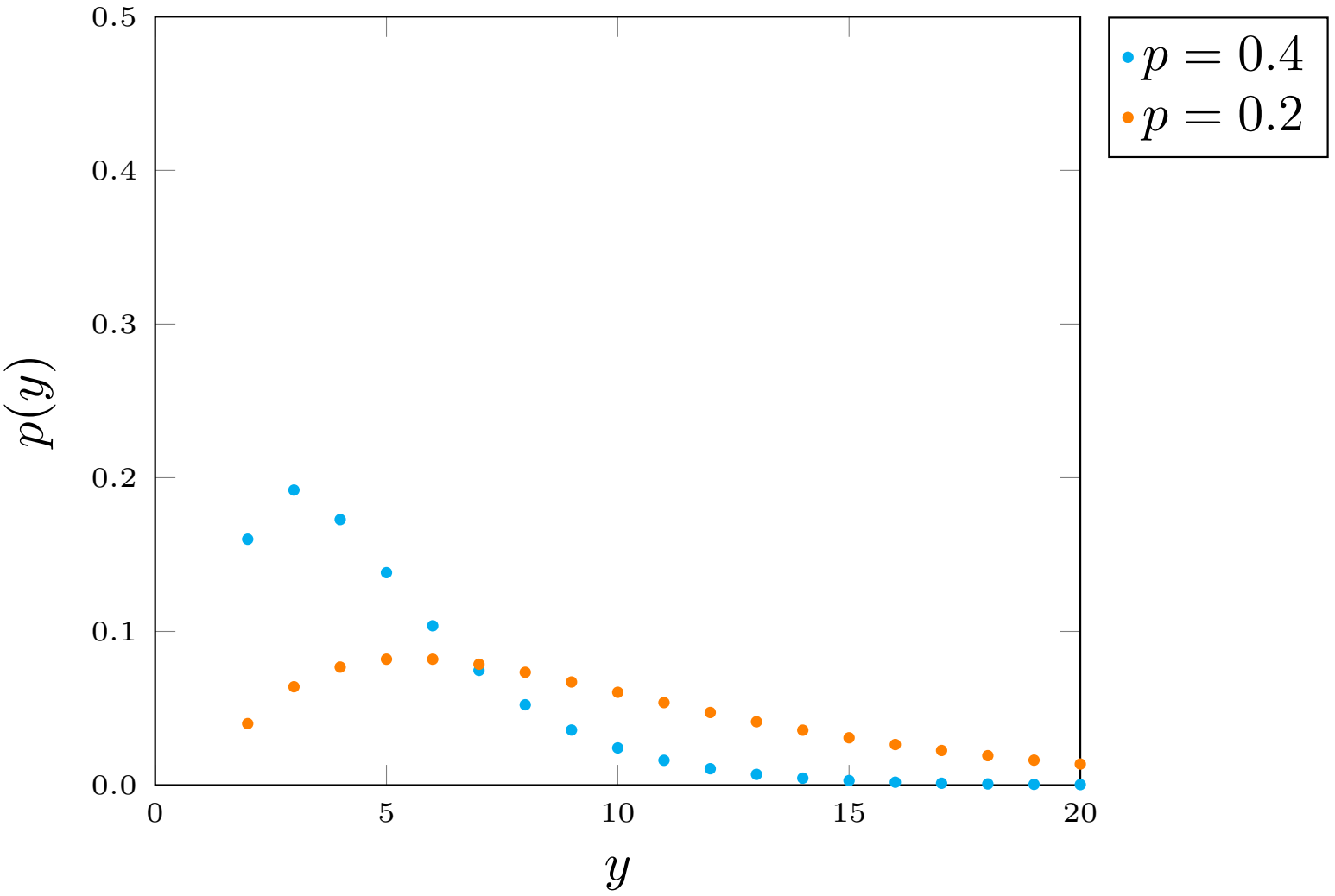
$$\mathbb{E}[Y] = \frac{r}{p}$$

Variance:

$$\mathbb{V}[Y] = \frac{r(1-p)}{p^2}$$

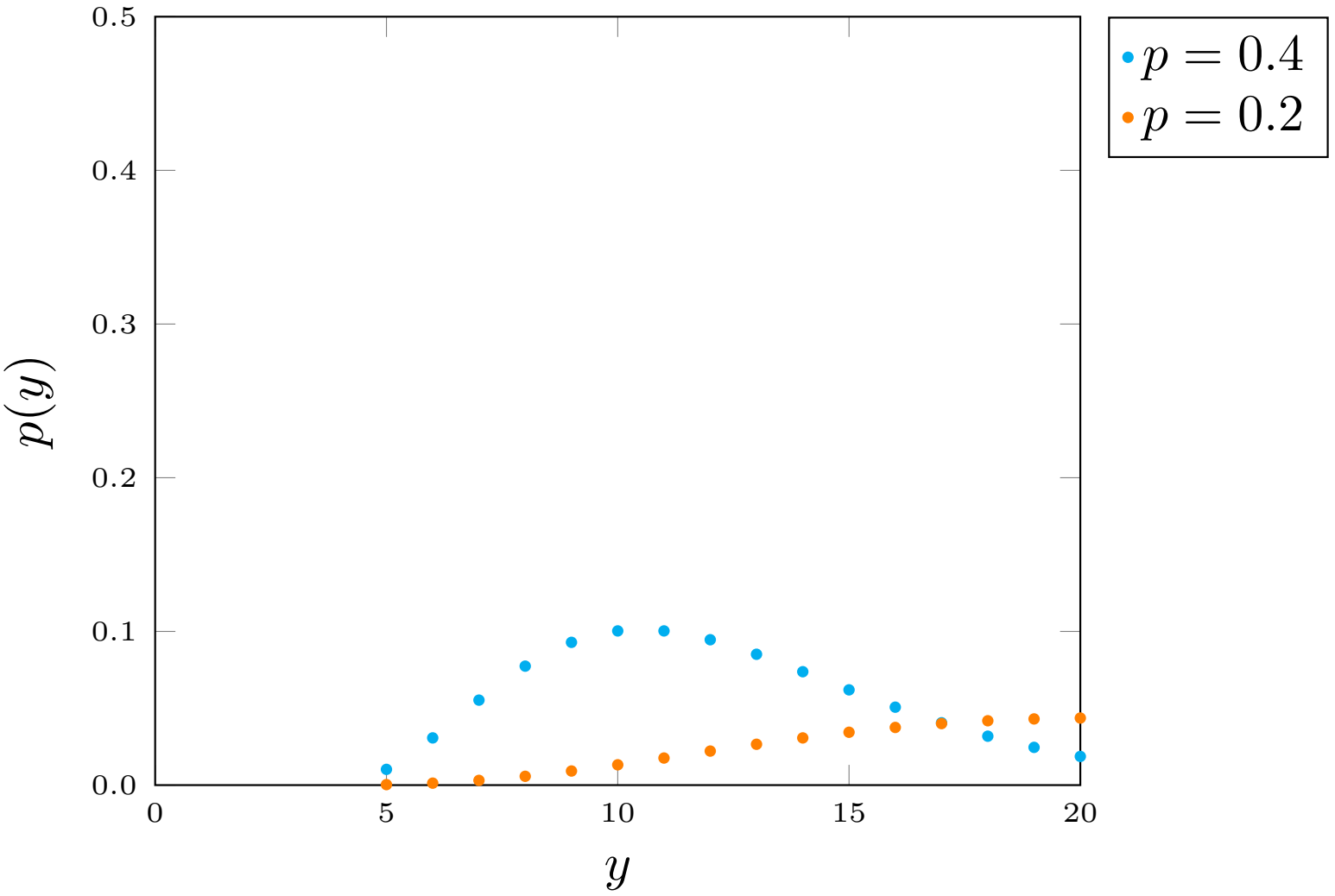
The negative binomial distribution $r = 2$

Negative binomial $p(y)$: with $p = 0.4$ and $p = 0.2$.



The negative binomial distribution $r = 5$

Negative binomial $p(y)$: with $p = 0.4$ and $p = 0.2$.



The hypergeometric distribution

The experimental setting for another discrete distribution known as the *hypergeometric distribution* relates to the earlier combinatorial exercises.

Suppose we have a finite 'population' (set) of N items which comprise

- r Type I items,
- $N - r$ Type II objects.

The experiment consists of selecting n items from the population, without replacement, such that all such selections are equally likely.

The hypergeometric distribution (cont.)

Define the random variable Y such that

$$(Y = y) \equiv A = \text{“}y \text{ Type I items in the selection”}$$

First, consider the feasible values for Y; we know that

$$0 \leq y \leq r$$

but also that

$$0 \leq n - y \leq N - r$$

The hypergeometric distribution (cont.)

Putting these two facts together, we must have that

$$\max\{0, n - N + r\} \leq y \leq \min\{n, r\}$$

so

$$\mathcal{Y} = \{\max\{0, n - N + r\}, \ldots, \min\{n, r\}\}.$$

The hypergeometric distribution (cont.)

Secondly, we compute $P(A)$ as follows: for $y \in \mathcal{Y}$

$$n_A = \binom{r}{y} \times \binom{N-r}{n-y}$$

$\binom{r}{y}$: number of ways of selecting the y Type I items from the r available.

$\binom{N-r}{n-y}$: number of ways of selecting the $n-y$ Type II items from the $N-r$ available.

The hypergeometric distribution (cont.)

Also

$$n_S = \binom{N}{n}$$

which is the number of ways of selecting the n items.

Therefore

$$p(y) = P(Y = y) = P(A) = \frac{n_A}{n_S} = \frac{\binom{r}{y} \times \binom{N-r}{n-y}}{\binom{N}{n}} \quad y \in \mathcal{Y}$$

and zero for all other values of y.

This is the pmf of the *hypergeometric distribution*.

The hypergeometric distribution (cont.)

Note

An equivalent expression for $p(y)$ is

$$p(y) = \frac{\binom{n}{y} \times \binom{N-n}{r-y}}{\binom{N}{r}} \quad y \in \mathcal{Y}$$

which we can obtain by simple manipulation of the combinatorial terms, or justify by considering

- the ways of arranging the y Type I items in the sample of size n,
- the $r - y$ Type I items in the $N - n$ 'non-sampled' items,
- the number of ways of arranging the Type I and Type II items.

The hypergeometric distribution (cont.)

Note

When N and r are very large, $p(y)$ is well approximated by a binomial pmf

$$\binom{n}{y}\left(\frac{r}{N}\right)^{y}\left(1-\frac{r}{N}\right)^{n-y}$$

The Poisson distribution

We now examine a limiting case of the binomial experimental setting; recall that binomial distribution arises when we consider n identical and independent trials each with success probability p.

$$10001000101010101010101$$

etc. We have

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y} \qquad y \in \{0, 1, 2, \ldots, n\}$$

with $p(y) = 0$ for all other values of y.

The Poisson distribution (cont.)

Suppose we allow n to get larger, and p to get smaller, but in such a way that

$$n \times p$$

remains constant at the value λ say. That is, we may write

$$p = \frac{\lambda}{n}$$

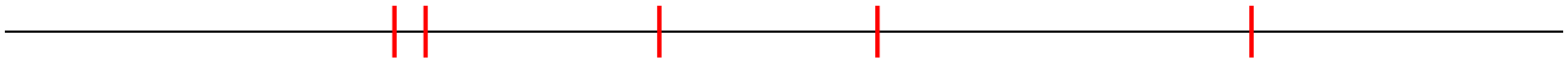
and consider n becoming large.

The Poisson distribution (cont.)

Note

The parameter λ corresponds to the 'rate' at which successes occur in a continuous approximation to the experiment.

The Poisson distribution (cont.)



The Poisson distribution (cont.)

We then have that

$$p^y(1-p)^{n-y} = \left(\frac{\lambda}{n}\right)^y \left(1-\frac{\lambda}{n}\right)^{n-y}$$

and

$$\binom{n}{y} = \frac{n!}{y!(n-y)!} = \frac{1}{y!}\frac{n!}{(n-y)!}$$

so we may write the binomial pmf

$$\frac{\lambda^y}{y!}\left(1-\frac{\lambda}{n}\right)^n \frac{n(n-1)\dots(n-y+1)}{n^y}\left(1-\frac{\lambda}{n}\right)^{-y}$$

The Poisson distribution (cont.)

Now we may rewrite

$$\frac{n(n-1)\dots(n-y+1)}{n^y} = \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\dots\left(\frac{n-y+1}{n}\right)$$

and deduce from this that for fixed y as $n \longrightarrow \infty$

$$\frac{n(n-1)\dots(n-y+1)}{n^y}\left(1-\frac{\lambda}{n}\right)^{-y} \longrightarrow 1$$

as each term in the product converges to 1.

The Poisson distribution (cont.)

Finally, by a standard representation

$$\lim_{n \longrightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Recall that

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

The Poisson distribution (cont.)

Therefore, as $n \longrightarrow \infty$ with $p = \lambda/n$ for fixed λ, we have that the binomial pmf converges to

$$p(y) = \frac{\lambda^y}{y!} e^{-\lambda} \qquad y \in \{0, 1, 2, \ldots\}$$

with $p(y) = 0$ for other values of y.

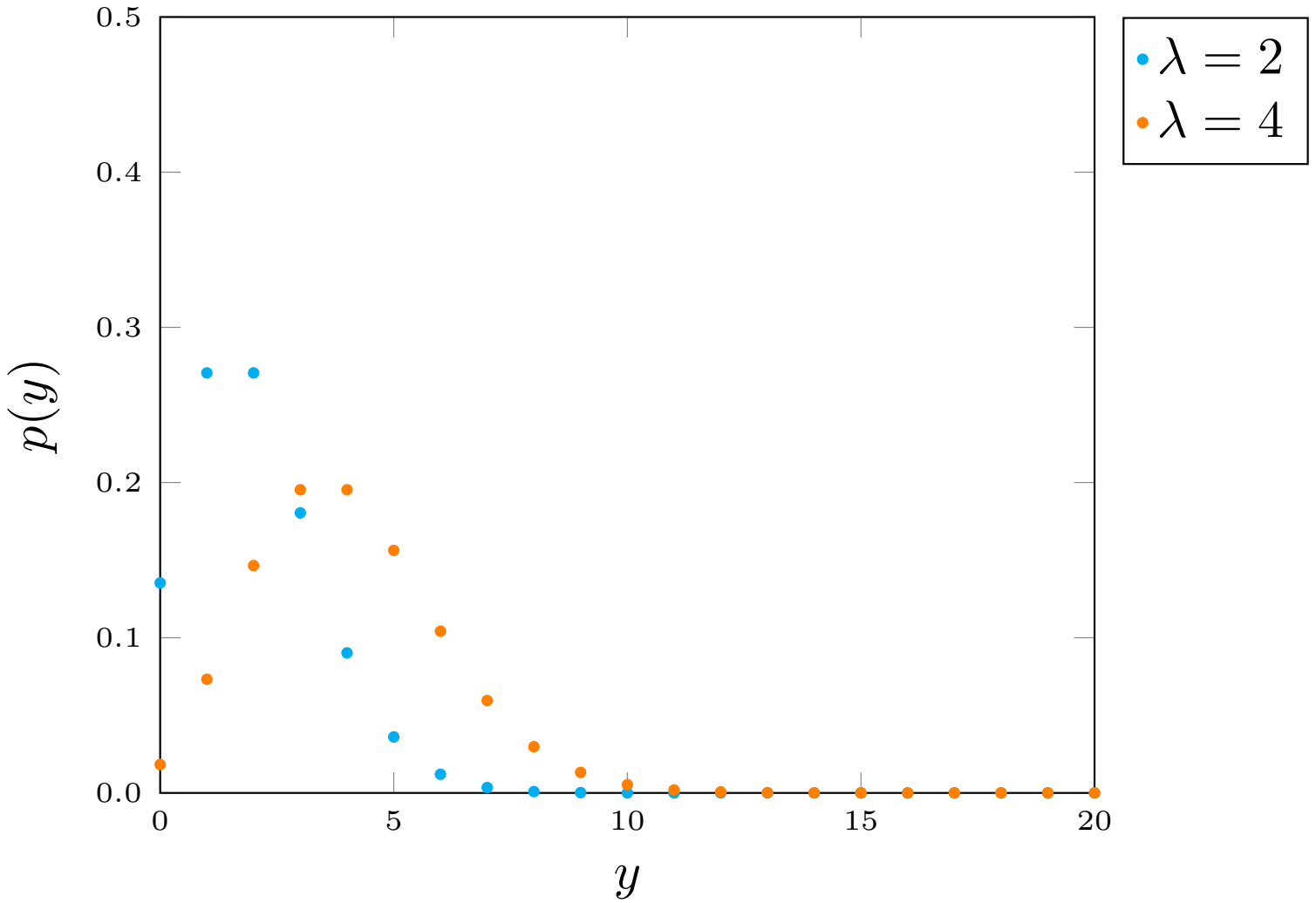
This is the pmf for the *Poisson distribution*.

Note that

$$\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{-\lambda} e^{\lambda} = 1.$$

The Poisson distribution (cont.)

Poisson $p(y)$: with $\lambda = 2$ and $\lambda = 4$.



The Poisson distribution (cont.)

Expectation : For the expectation of a Poisson distribution:

$$\begin{aligned}
\mathbb{E}[Y] = \sum_{y=0}^{\infty} yp(y) &= \sum_{y=0}^{\infty} y\frac{\lambda^y}{y!}e^{-\lambda} \\
&= e^{-\lambda}\sum_{y=1}^{\infty}\frac{\lambda^y}{(y-1)!} \\
&= e^{-\lambda}\lambda\sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!} \\
&= e^{-\lambda}\lambda\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} \\
&= e^{-\lambda}\lambda e^{\lambda} = \lambda.
\end{aligned}$$

The Poisson distribution (cont.)

Variance : For the variance of a Poisson distribution: using the usual trick

$$\mathbb{E}[Y(Y-1)] = \sum_{y=0}^{\infty} y(y-1)p(y) = \sum_{y=2}^{\infty} y(y-1)\frac{\lambda^y}{y!}e^{-\lambda}$$

$$= e^{-\lambda}\lambda^2 \sum_{y=2}^{\infty} \frac{\lambda^{y-2}}{(y-2)!}$$

$$= e^{-\lambda}\lambda^2 \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}$$

$$= e^{-\lambda}\lambda^2 e^{\lambda} = \lambda^2.$$

The Poisson distribution (cont.)

Therefore

$$\mathbb{V}[Y] = \mathbb{E}[Y(Y-1)] + \mathbb{E}[Y] - \{\mathbb{E}[Y]\}^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Thus for the Poisson distribution

$$\mathbb{E}[Y] = \mathbb{V}[Y] = \lambda.$$

The Poisson distribution (cont.)

Note

The Poisson distribution is widely used to model the distribution of count data.

It corresponds to a model where incidents ('successes') occur independently in continuous time at a constant rate.

Moments

We may generalize the concept of expectation and variance by considering the following quantities: if $\mu = \mathbb{E}[Y]$,

$$k^{\text{th}} \text{ moment}: \quad \mu'_k = \mathbb{E}[Y^k] = \sum_y y^k p(y)$$

$$k^{\text{th}} \text{ central moment}: \quad \mu_k = \mathbb{E}[(Y-\mu)^k] = \sum_y (y-\mu)^k p(y)$$

These quantities help to characterize $p(y)$.

Moment-generating functions

In mathematics, the *generating function*, $g(t)$, for the real-valued sequence $\{a_j\}$ is the function defined by the sum

$$g(t) = \sum_j a_j t^j.$$

For example,

$$g(t) = (1+t)^n = \sum_{j=0}^{n} \binom{n}{j} t^j$$

is the generating function for the binomial coefficients

$$a_j = \binom{n}{j} \qquad j = 0, 1, \ldots, n.$$

Moment-generating functions (cont.)

We can also define an (exponential) generating function as

$$g(t) = \sum_j a_j \frac{t^j}{j!}.$$

Moment-generating functions (cont.)

For pmf $p(y)$, the *moment-generating function* (mgf), $m(t)$, is the (exponential) generating function for the moments of $p(y)$:

$$m(t) = \sum_{j=0}^{\infty} \mu'_j \frac{t^j}{j!}$$

whenever this sum is convergent.

More specifically, we require this sum to be finite at least in the neighbourhood of $t = 0$, for $|t| \leq b$, say.

Moment-generating functions (cont.)

Provided the sum is finite, we have for fixed t,

$$\begin{aligned} m(t) &= \sum_{j=0}^{\infty} \mu'_j \frac{t^j}{j!} \\ &= \sum_{j=0}^{\infty} \left\{ \sum_y y^j p(y) \right\} \frac{t^j}{j!} \\ &= \sum_y \left\{ \sum_{j=0}^{\infty} \frac{(ty)^j}{j!} \right\} p(y) \end{aligned}$$

Moment-generating functions (cont.)

But

$$\sum_{j=0}^{\infty} \frac{(ty)^j}{j!} = e^{ty}$$

so therefore

$$m(t) = \sum_y e^{ty} p(y) = \mathbb{E}\left[e^{tY}\right]$$

Moment-generating functions (cont.)

Note also that if we write

$$m^{(k)}(t) = \frac{d^k}{dt^k}\{m(t)\}$$

for the kth derivative of $m(t)$ with respect to t, we have

$$\begin{aligned}\frac{d^k}{dt^k}\{m(t)\} &= \frac{d^k}{dt^k}\left\{1 + \mu'_1 t + \mu'_2\frac{t^2}{2!} + \mu'_3\frac{t^3}{3!} + \cdots\right\} \\ &= \mu'_k + \mu'_{k+1}\frac{2t}{2!} + \mu'_{k+2}\frac{3t^2}{3!} + \cdots\end{aligned}$$

Moment-generating functions (cont.)

Therefore if we evaluate at $t = 0$, we have

$$m^{(k)}(0) = \mu'_k \qquad k = 1, 2, \ldots.$$

This result indicates that we can use the moment-generating function, if it is given, to compute the moments of $p(y)$.

This is the first main use of moment generating function.

Moment-generating functions (cont.)

The second main use is to identify distributions; suppose $m_1(t)$ and $m_2(t)$ are two moment-generating functions for pmfs $p_1(y)$ and $p_2(y)$. Then

$$m_1(t) = m_2(t) \text{ for all } |t| \leq b$$

for some b if and only if

$$p_1(y) = p_2(y) \qquad \text{for all } y.$$

Moment-generating functions (cont.)

Example (Poisson mgf)

We have for the Poisson distribution with parameter λ

$$\begin{aligned} m(t) = \sum_{y=0}^{\infty} e^{ty} p(y) &= \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y}{y!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!} \\ &= e^{-\lambda} \exp\{\lambda e^t\} \\ &= \exp\{\lambda(e^t - 1)\} \end{aligned}$$

Note that this holds for all t.

Moment-generating functions (cont.)

Example (Binomial mgf)

We have for the Binomial distribution with parameters n and p

$$m(t) = \sum_{y=0}^{n} e^{ty} p(y) = \sum_{y=0}^{n} e^{ty} \binom{n}{y} p^y q^{n-y}$$

$$= \sum_{y=0}^{n} \binom{n}{y} (pe^t)^y q^{n-y}$$

$$= (pe^t + q)^n.$$

Moment-generating functions (cont.)

Example (Binomial mgf)

If $p = \lambda/n$, then

$$\begin{aligned}(pe^t + q)^n &= \left(\frac{\lambda}{n}e^t + \left(1 - \frac{\lambda}{n}\right)\right)^n \\ &= \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n \\ &\longrightarrow \exp\{\lambda(e^t - 1)\}\end{aligned}$$

as $n \longrightarrow \infty$, by definition of the exponential function.

Moment-generating functions (cont.)

Example (Geometric mgf)

We have for the geometric distribution with parameter p

$$m(t) = \sum_{y=1}^{\infty} e^{ty} p(y) = \sum_{y=1}^{\infty} e^{ty} q^{y-1} p$$

$$= pe^t \sum_{y=1}^{\infty} (qe^t)^{y-1}$$

$$= pe^t \sum_{j=0}^{\infty} (qe^t)^{j}$$

$$= pe^t \frac{1}{1-qe^t} = \frac{pe^t}{1-qe^t}$$

Moment-generating functions (cont.)

Example

Suppose we are told that the mgf of a distribution takes the form

$$m(t) = \exp\{3(e^t - 1\};$$

then we may deduce that the corresponding distribution is a Poisson distribution with parameter $\lambda = 3$.

Probability-generating functions

A related function is the *probability-generating function* (pgf); this can be defined for a random variable Y that takes values on

$$\{0, 1, 2, \ldots\}$$

where

$$P(Y = i) = p_i.$$

The pgf, $G(t)$, is defined by

$$G(t) = \mathbb{E}\left[t^Y\right] = p_0 + p_1 t + p_2 t^2 + \cdots = \sum_{i=0}^{\infty} p_i t^i$$

whenever this sum is finite. Note that $G(1) = 1$, so we typically focus on evaluating $G(t)$ in a neighbourhood of 1, say $(1-b, 1+b)$ for some $b > 0$.

Probability-generating functions (cont.)

Note that

$$G^{(k)}(t) = \frac{d^k}{dt^k}\{G(t)\} = \frac{d^k}{dt^k}\left\{\sum_{i=0}^{\infty} p_i t^i\right\}$$

$$= \sum_{i=k}^{\infty} i(i-1)\ldots(i-k+1)p_i t^{i-k}$$

so therefore

$$G^{(k)}(1) = \sum_{i=k}^{\infty} i(i-1)\ldots(i-k+1)p_i$$

$$= \mathbb{E}[Y(Y-1)(Y-2)\ldots(Y-k+1)].$$

Probability-generating functions (cont.)

The quantity

$$\mu_{[k]} = \mathbb{E}[Y(Y-1)(Y-2)\dots(Y-k+1)]$$

is termed the kth factorial moment.

Recall that we previously computed

$$\mu_{[2]} = \mathbb{E}[Y(Y-1)]$$

to calculate the variance of a distribution.

Probability-generating functions (cont.)

Note

As we have the identity

$$t = \exp\{\ln t\}$$

it follows that

$$G(t) \equiv \mathbb{E}\left[t^Y\right] = \mathbb{E}\left[\exp\{Y \ln t\}\right] = m(\ln t)$$

so once we have computed $m(t)$ we may immediately deduce $G(t)$.

Probability-generating functions (cont.)

Note

	$m(t)$	$G(t)$
Binomial	$(pe^t + q)^n$	$(pt + q)^n$
Poisson	$\exp\{\lambda(e^t - 1)\}$	$\exp\{\lambda(t - 1)\}$
Geometric	$\dfrac{pe^t}{1 - qe^t}$	$\dfrac{pt}{1 - qt}$

Continuous Random variables

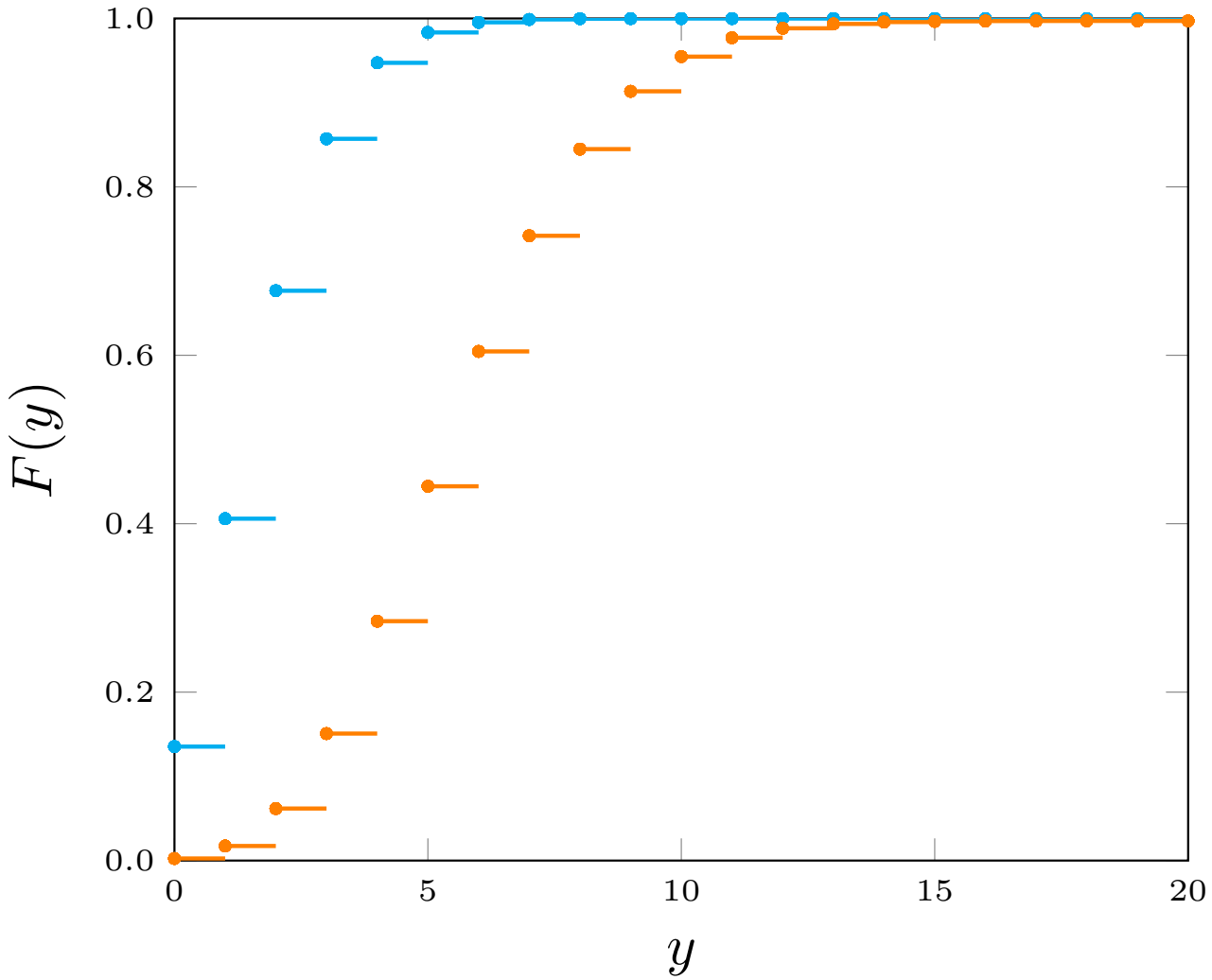
Recall the previous definition of the cumulative distribution function (or cdf), $F(y)$, for a random variable:

$$F(y) = P(Y \leq y)$$

which can be considered for any real value y. For a discrete random variable, this function is termed a *step* function, with steps at only a countable number of points.

Continuous Random variables (cont.)

Poisson $F(y)$: with $\lambda = 2$ and $\lambda = 6$.



Continuous Random variables (cont.)

Directly from the probability axioms, we can deduce three properties of the function $F(y)$:

1. "Starts at zero":

$$F(-\infty) \equiv \lim_{y \longrightarrow -\infty} F(y) = 0.$$

2. "Ends at 1":

$$F(\infty) \equiv \lim_{y \longrightarrow \infty} F(y) = 1.$$

3. "Non-decreasing in between": if $y_1 < y_2$, then

$$F(y_1) \leq F(y_2).$$

Continuous Random variables (cont.)

Now suppose that the random variable Y is most accurately thought of as varying continuously, that is, can potentially take any real value in an interval

- height/weight,
- temperature,
- time,

and so on.

Continuous Random variables (cont.)

In this case it seems that the probability that Y lies in an interval $(y_1, y_2]$ for $y_1 < y_2$ must be positive, but should converge to zero the closer y_1 moves to y_2. That is

$$P(Y \in (y_1, y_2]) = P(Y \leq y_2) - P(Y \leq y_1) = F(y_2) - F(y_1)$$

should be positive, but decrease to zero as $y_1 \longrightarrow y_2$.

Continuous Random variables (cont.)

This is not the case for a discrete random variable: suppose

$$p(0) = \frac{1}{2} \qquad p(1) = \frac{1}{4} \qquad p(2) = \frac{1}{4}$$

and $p(y) = 0$ for all other values of y. Then for $0 < \epsilon < 1$

$$F(0) = P(Y \leq 0) = \frac{1}{2}$$

$$F(1 - \epsilon) = \frac{1}{2}$$

$$F(1) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

Continuous Random variables (cont.)

A random variable Y with distribution function $F(y)$ is called *continuous* if $F(y)$ is continuous for all real y. That is, for any y,

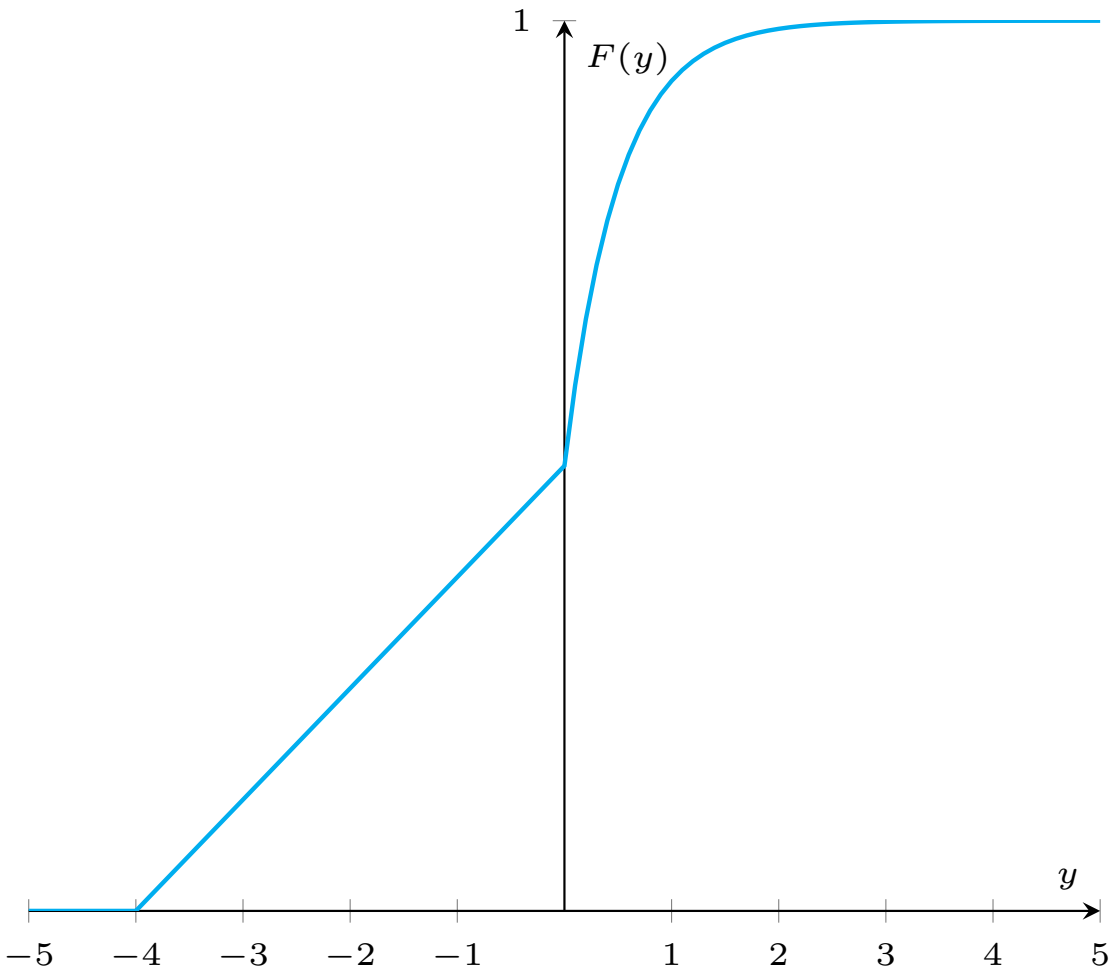
$$\lim_{\epsilon \longrightarrow 0+} F(y-\epsilon) = \lim_{\epsilon \longrightarrow 0+} F(y+\epsilon) = F(y).$$

Continuous Random variables (cont.)

We often also assume that $F(y)$ has a continuous first derivative for all $y \in \mathbb{R}$, except perhaps at a finite number of points.

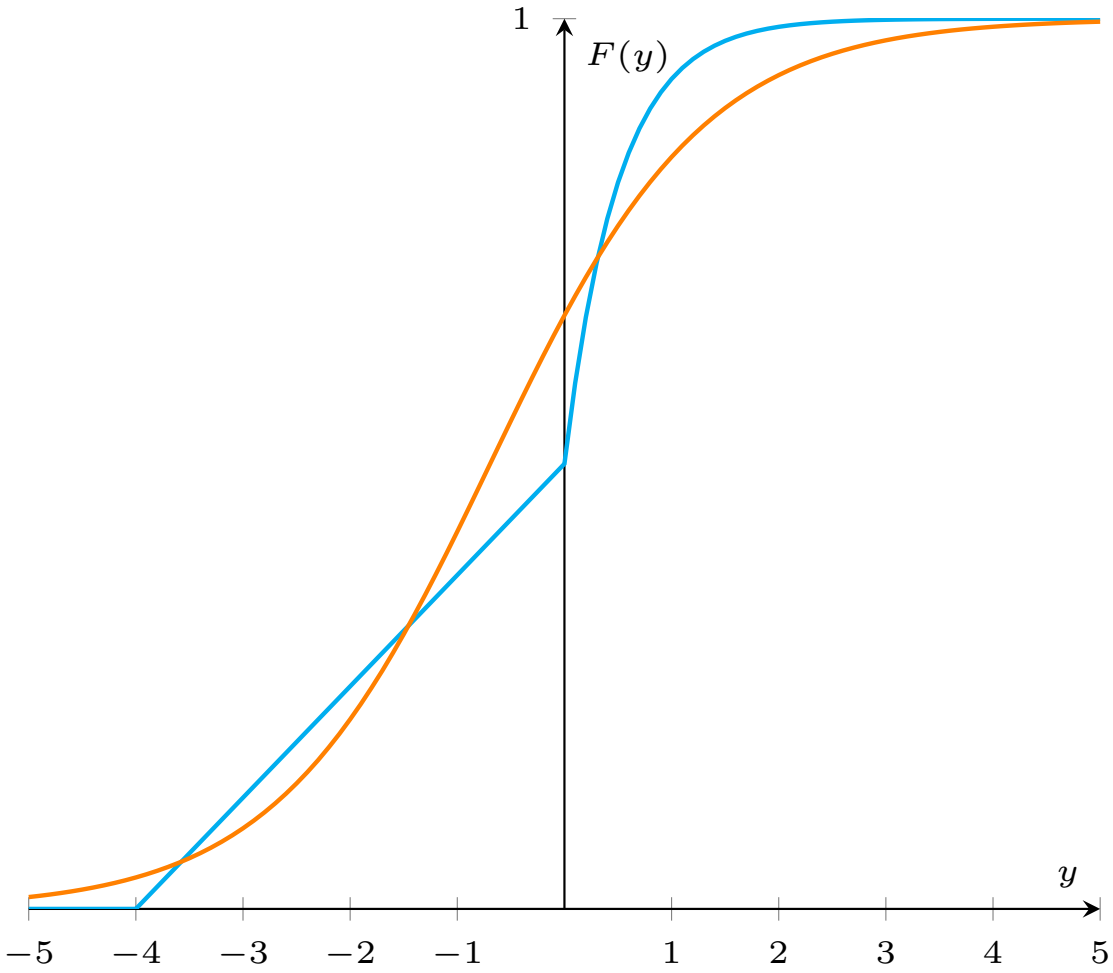
Continuous Random variables (cont.)

$F(y)$ continuous but not differentiable everywhere.



Continuous Random variables (cont.)

$F(y)$ continuous and differentiable.



Continuous Random variables (cont.)

Note

Even if Y is discrete, $F(y)$ is right-continuous. That is, for any y,

$$\lim_{\epsilon \longrightarrow 0+} F(y+\epsilon) = F(y).$$

Continuous Random variables (cont.)

Important Note

If $F(y)$ is continuous, then

$$P(Y = y) = 0$$

for all $y \in \mathbb{R}$. If there were to exist a y_0 such that

$$P(Y = y_0) = p_0 > 0$$

then $F(y)$ would have a step of size p_0 at y_0, and would be discontinuous.

Probability density functions

For a continuous cdf $F(y)$, denote by $f(y)$ the first derivative of $F(.)$ at y

$$f(y) = \frac{dF(y)}{dy}$$

wherever this derivative exists.

The function $f(y)$ is known as the *probability density function* (pdf).

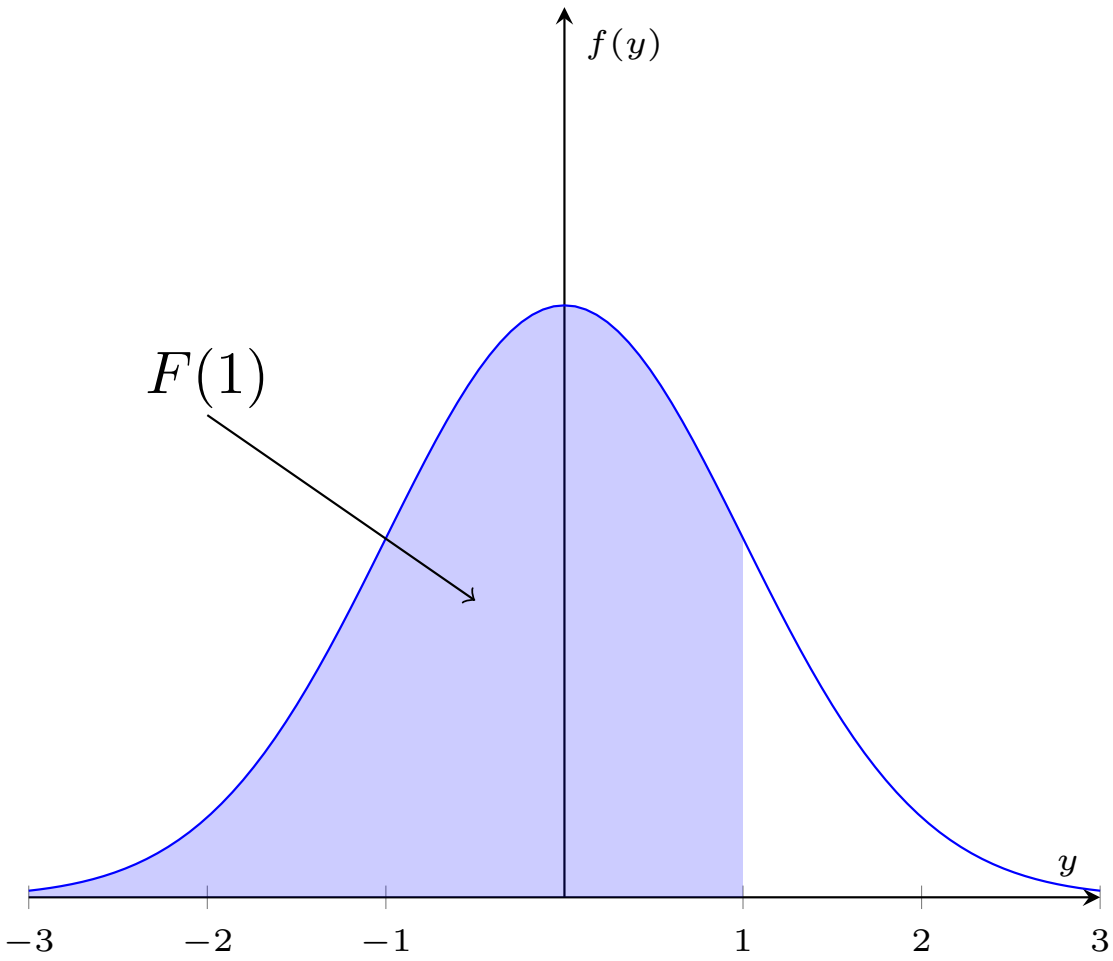
Probability density functions (cont.)

It follows that

$$F(t) = \int_{-\infty}^{y} f(t)\, dt$$

Probability density functions (cont.)

$$F(1) = P(Y \leq 1) = \int_{-\infty}^{1} f(y)\, dy.$$



Properties of density functions

From the definition, we have two properties of the pdf $f(y)$:

1. Non-negative:

$$f(y) \geq 0 \qquad -\infty < y < \infty$$

   – this follows as $F(y)$ is non-decreasing.

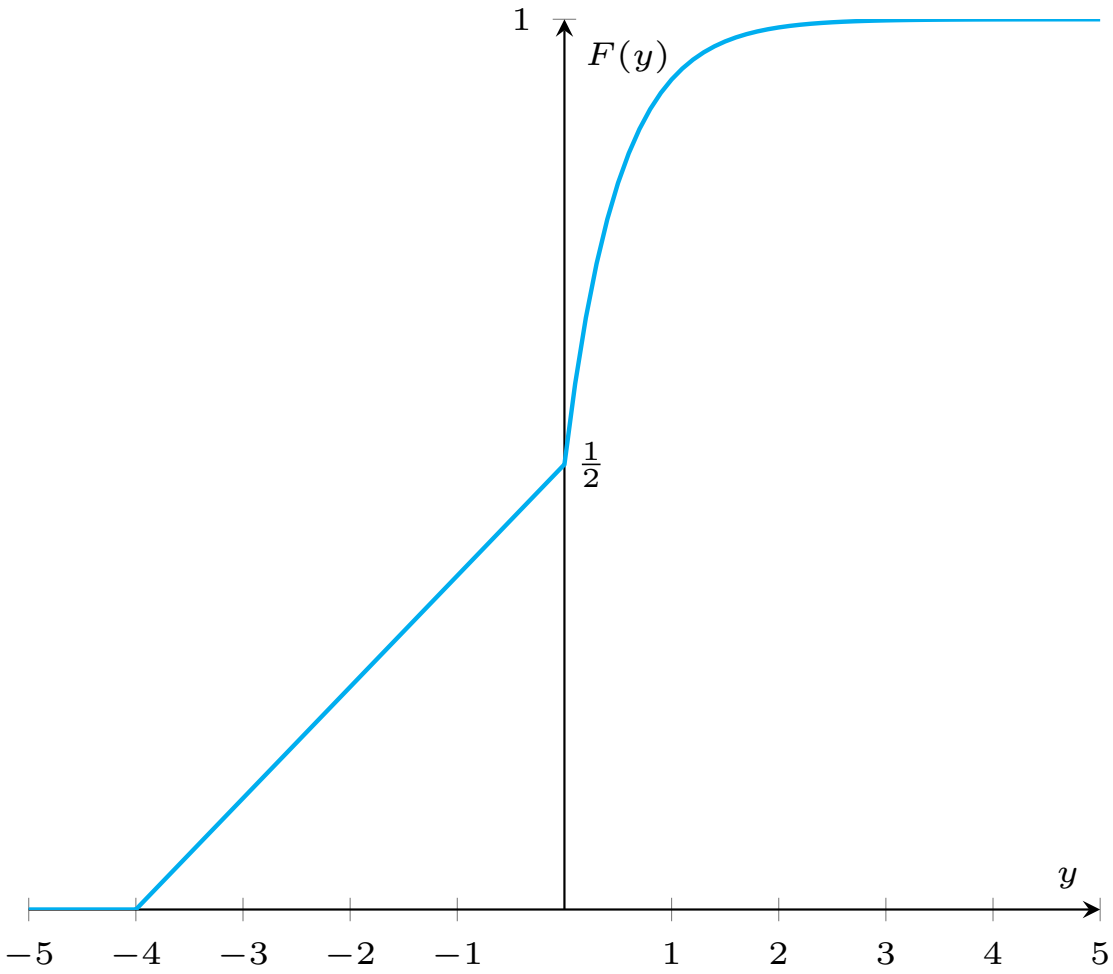
2. Integrates to 1:

$$\int_{-\infty}^{\infty} f(y)\, dy = 1$$

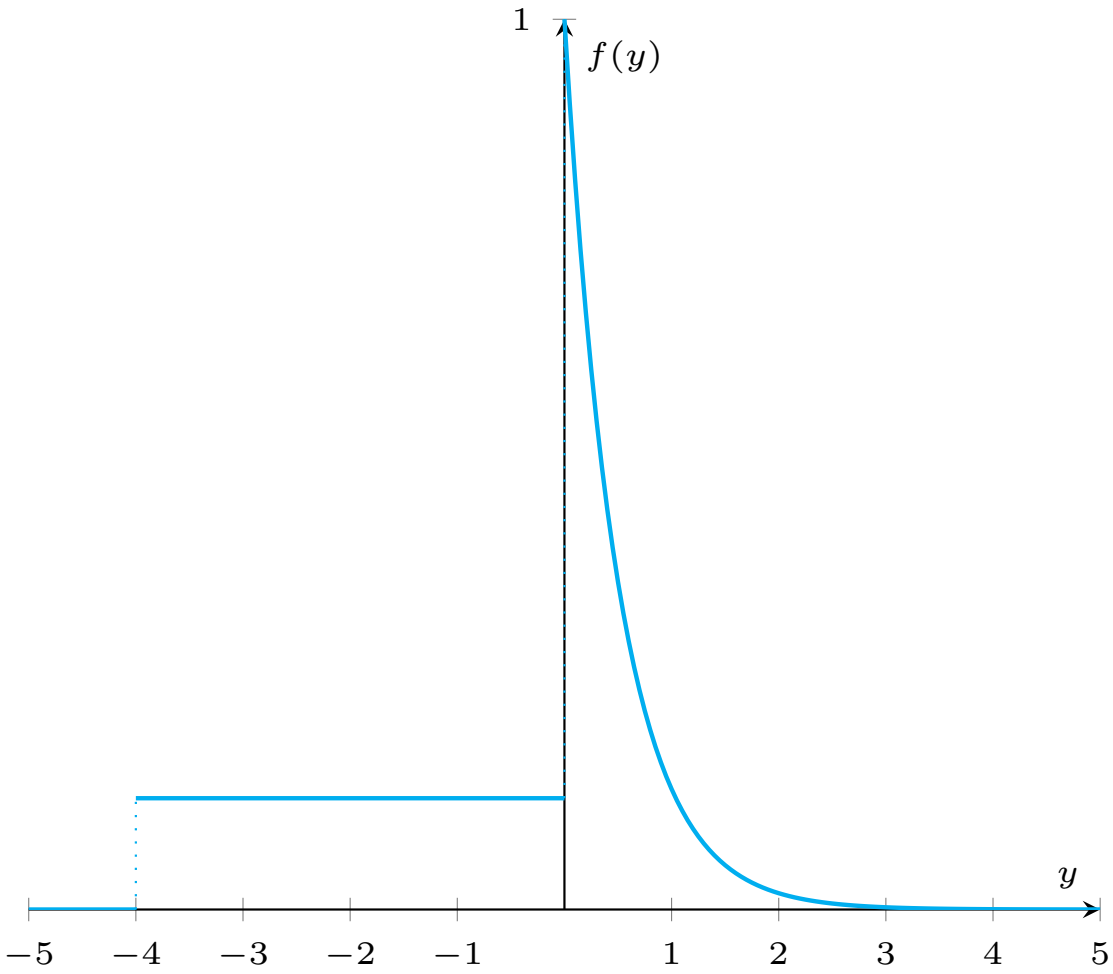
   – this follows as

$$F(\infty) \equiv \lim_{y \longrightarrow \infty} F(y) = 1.$$

Properties of density functions (cont.)



Properties of density functions (cont.)



Properties of density functions (cont.)

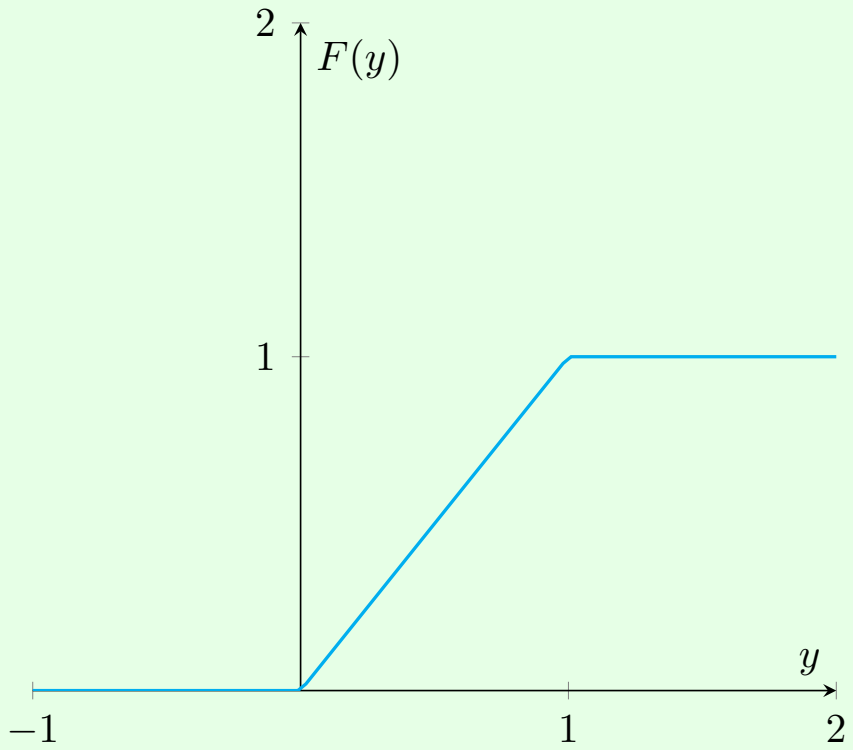
Example

Suppose

$$F(y) = \begin{cases} 0 & y \leq 0 \\ y & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases}$$

Properties of density functions (cont.)

Example



Properties of density functions (cont.)

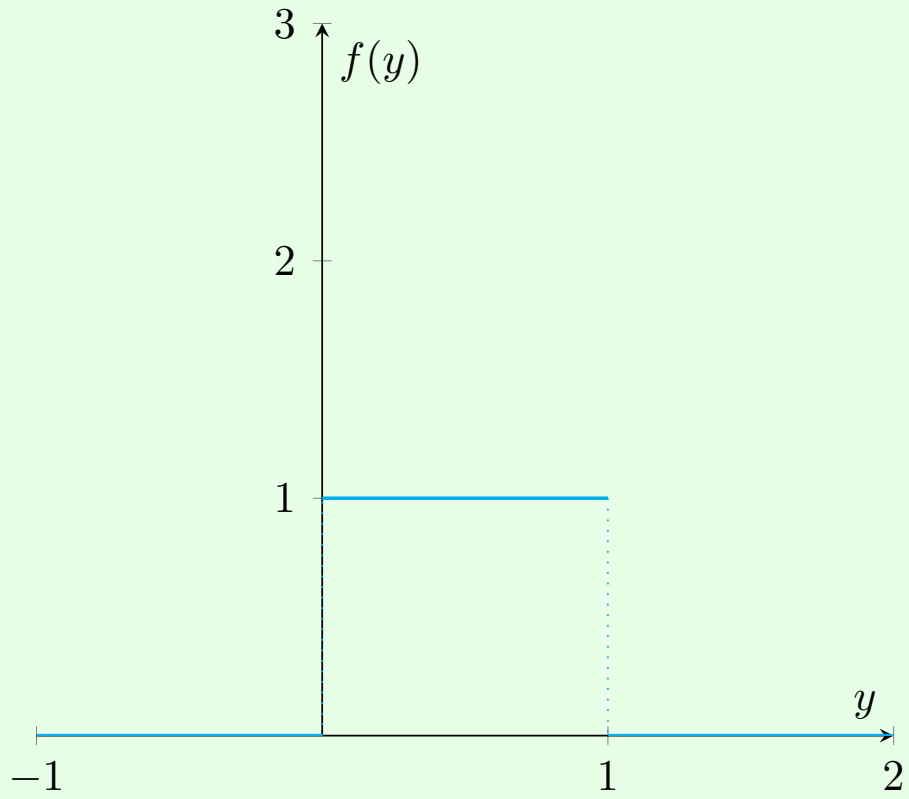
Example

Then

$$f(y) = \begin{cases} 0 & y < 0 \\ 1 & 0 < y < 1 \\ 0 & y > 1 \end{cases}$$

Properties of density functions (cont.)

Example



Properties of density functions (cont.)

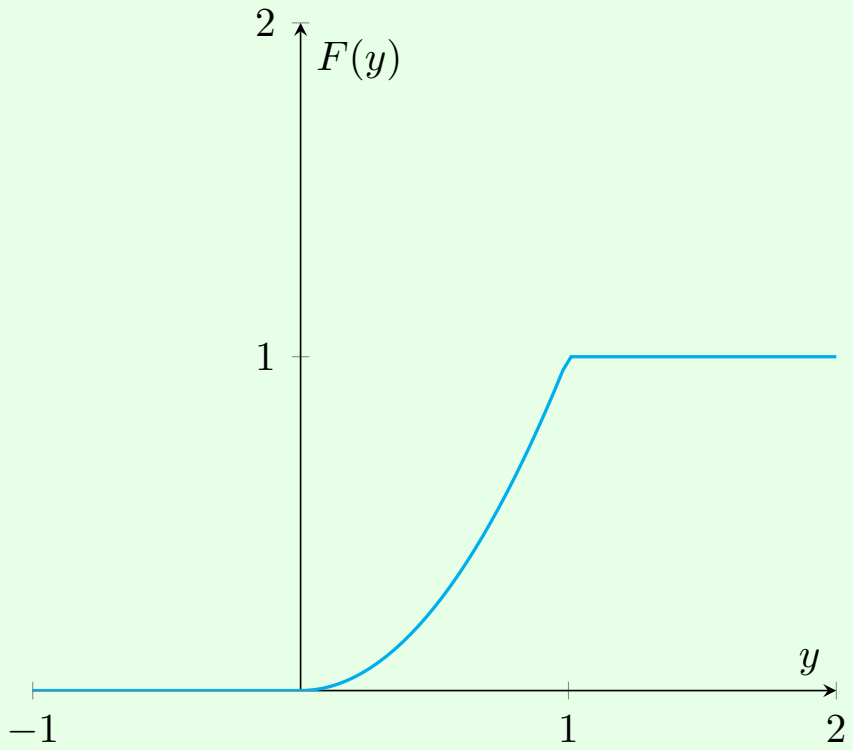
Example

Suppose

$$F(y) = \begin{cases} 0 & y \leq 0 \\ y^2 & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases}$$

Properties of density functions (cont.)

Example



Properties of density functions (cont.)

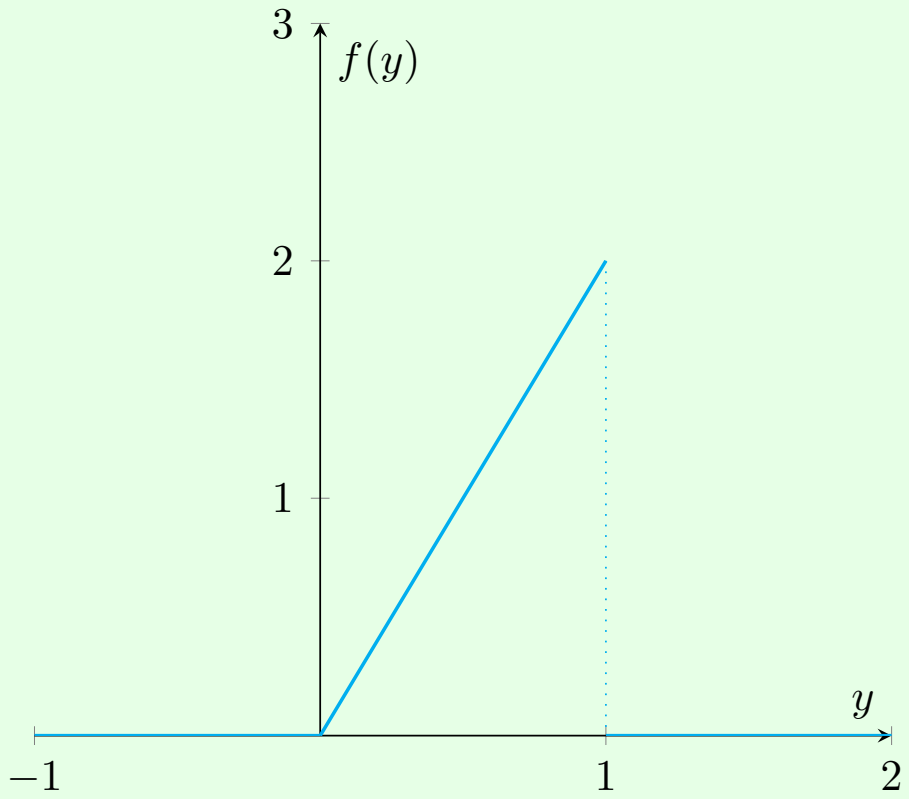
Example

Then

$$f(y) = \begin{cases} 0 & y < 0 \\ 2y & 0 < y < 1 \\ 0 & y > 1 \end{cases}$$

Properties of density functions (cont.)

Example



Properties of density functions (cont.)

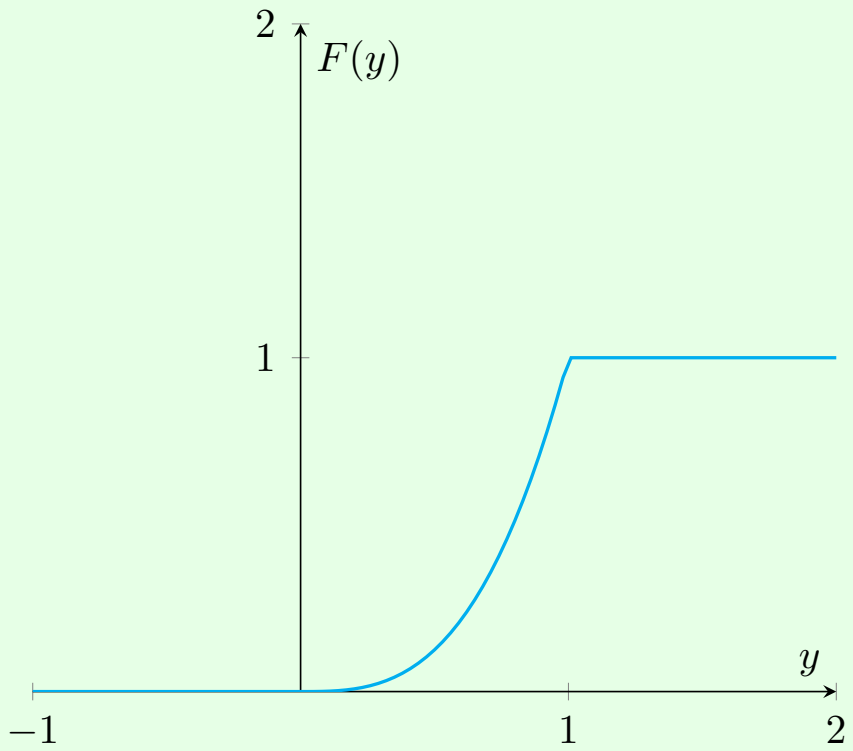
Example

Suppose

$$F(y) = \begin{cases} 0 & y \leq 0 \\ y^3 & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases}$$

Properties of density functions (cont.)

Example



Properties of density functions (cont.)

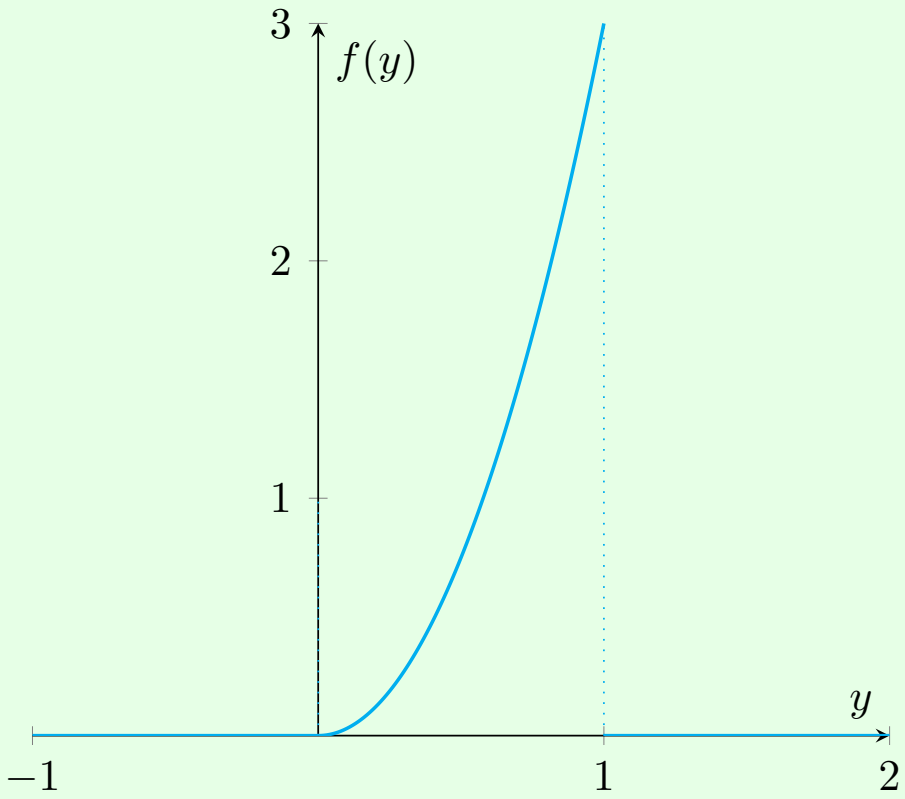
Example

Then

$$f(y) = \begin{cases} 0 & y < 0 \\ 3y^2 & 0 < y < 1 \\ 0 & y > 1 \end{cases}$$

Properties of density functions (cont.)

Example



Interval probabilities

To compute

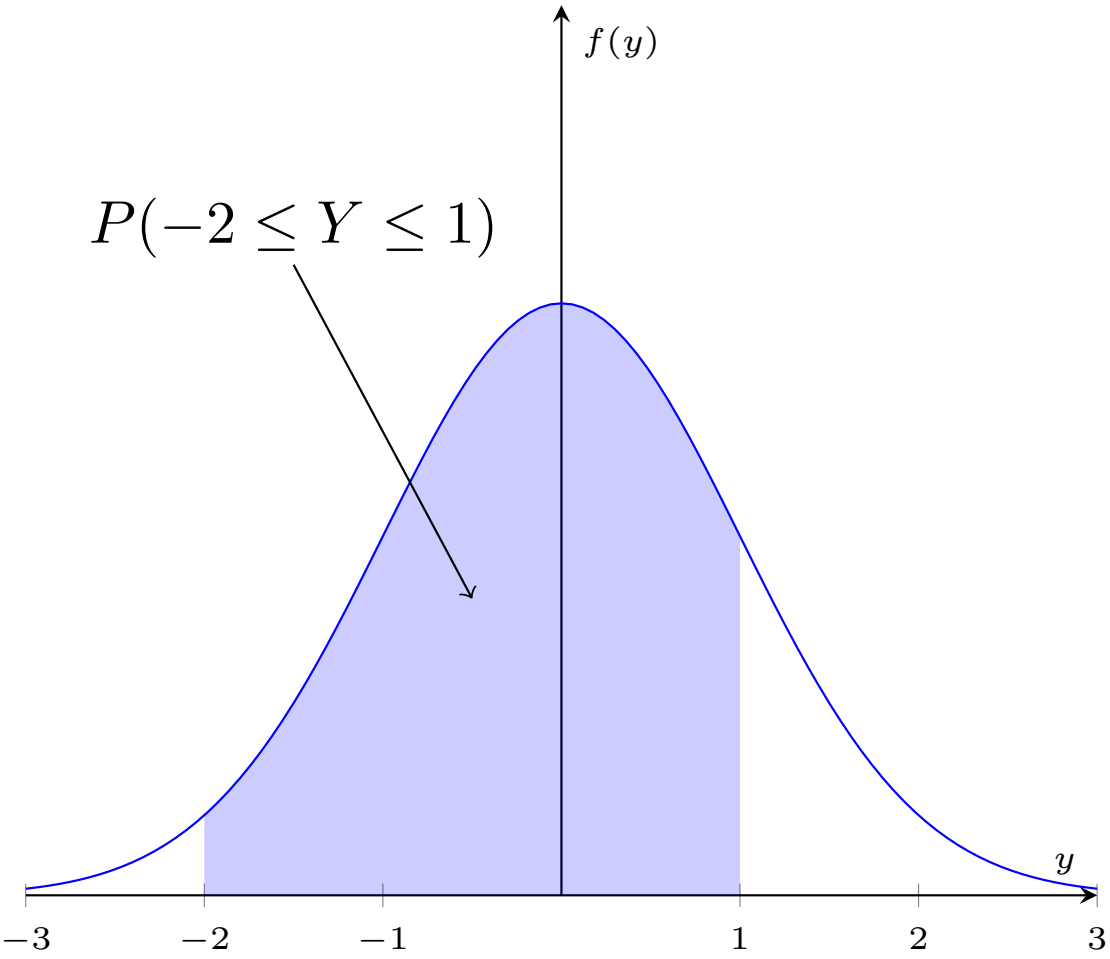
$$P(a \leq Y \leq b)$$

we integrate $f(y)$ over the interval (a, b) as

$$\begin{aligned} P(a \leq Y \leq b) &= P(Y \leq b) - P(Y \leq a) \\ &= \int_{-\infty}^{b} f(t)\, dt - \int_{-\infty}^{a} f(t)\, dt \\ &= \int_{a}^{b} f(t)\, dt. \end{aligned}$$

Interval probabilities (cont.)

$$P(-2 \leq Y \leq 1) = \int_{-2}^{1} f(y)\, dy = F(1) - F(-2).$$



Interval probabilities (cont.)

Example

Suppose

$$f(y) = \begin{cases} 0 & y \leq 0 \\ cy^2 & 0 \leq y \leq 2 \\ 0 & y > 2 \end{cases}$$

for some $c > 0$.

Interval probabilities (cont.)

Example

Then

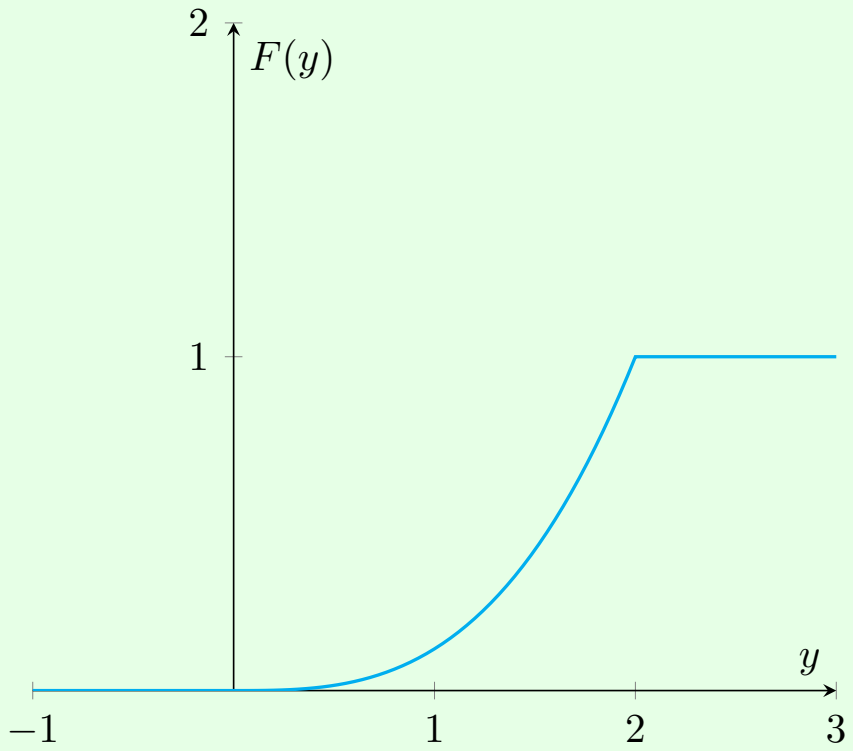
$$F(y) = \begin{cases} 0 & y \leq 0 \\ c\dfrac{y^3}{3} & 0 \leq y \leq 2 \\ c\dfrac{8}{3} & y \geq 2 \end{cases}$$

Therefore, as we require $F(\infty) = 1$, we must have

$$c = \frac{3}{8}.$$

Interval probabilities (cont.)

Example



Interval probabilities (cont.)

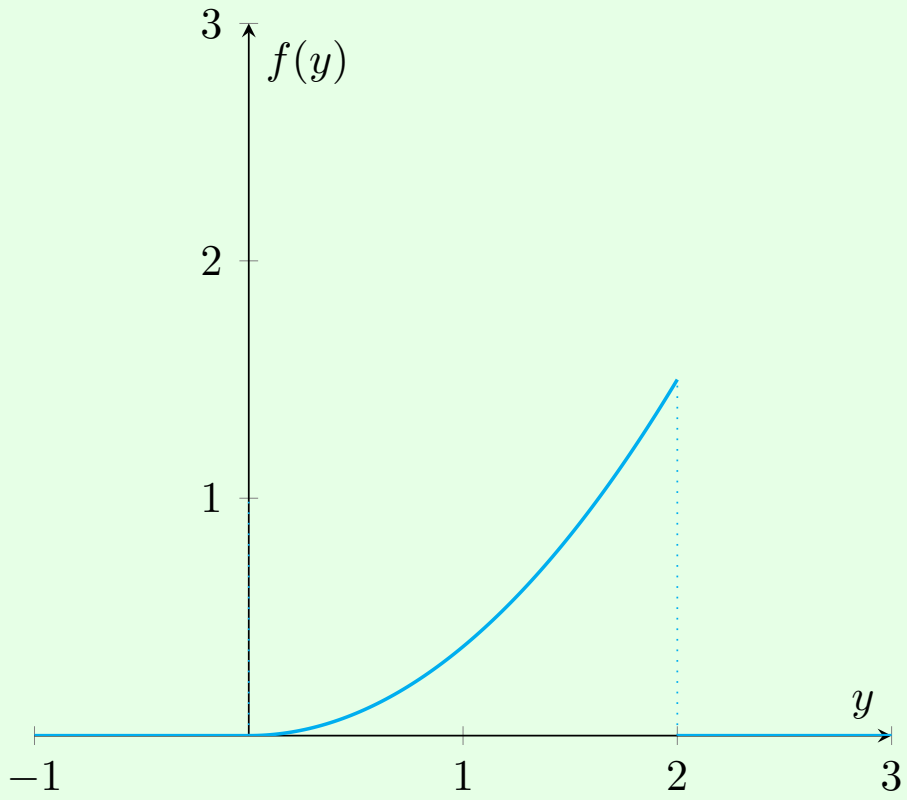
Example

Then

$$f(y) = \begin{cases} 0 & y < 0 \\ \frac{3}{8}y^2 & 0 < y < 2 \\ 0 & y > 2 \end{cases}$$

Interval probabilities (cont.)

Example



Interval probabilities (cont.)

To compute

$$P(1/2 \leq Y \leq 3/2)$$

we use the cdf to show that this is equal to

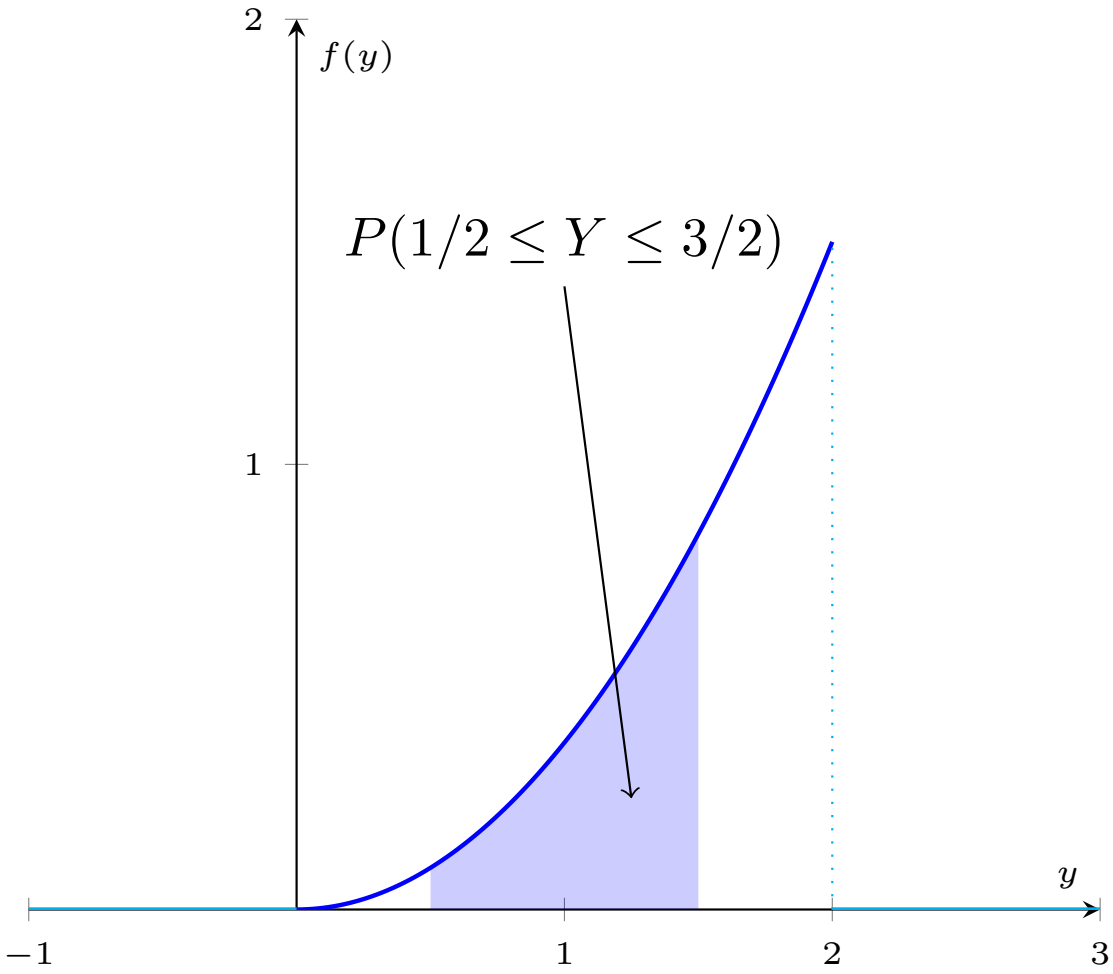
$$F(3/2) - F(1/2) = \frac{3}{8}\frac{1}{3}\left(\frac{3}{2}\right)^3 - \frac{3}{8}\frac{1}{3}\left(\frac{1}{2}\right)^3 = 0.40625$$

Interval probabilities (cont.)

This calculation is merely

$$P(1/2 \leq Y \leq 3/2) = \int_{1/2}^{3/2} \frac{3}{8} y^2 \, dy.$$

Interval probabilities (cont.)



Interval probabilities (cont.)

Note

For a continuous random variable

$$P(a \leq Y \leq b) = P(a < Y \leq b) = P(a \leq Y < b) = P(a < Y < b)$$

Interval probabilities (cont.)

Note

We often have that for some $y \in \mathbb{R}$

$$f(y) = 0$$

and so we again denote by $\mathcal{Y}$ the values for which $f(y) > 0$, that is

$$\mathcal{Y} = \{y : f(y) > 0\}$$

Interval probabilities (cont.)

Note

All pdfs take the form

$$\text{constant} \times \text{function of } y$$

that is

$$f(y) = cg(y)$$

say, where $g(y)$ also contains information on the range of values, $\mathcal{Y}$, for which $f(y) > 0$. We have

$$\int_{-\infty}^{\infty} f(y)\, dy = 1 \qquad \Longrightarrow \qquad \int_{-\infty}^{\infty} g(y)\, dy = \frac{1}{c}.$$

Expectations

The expectation of a continuous random variable, Y, is defined by

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} yf(y)\,dy = \int_{\mathcal{Y}} yf(y)\,dy = \mu$$

say, if this integral is finite; if it is not finite, we say that the expectation does not exist.

If

$$\int_{-\infty}^{\infty} |y| f(y)\,dy < \infty$$

then it follows that $\mathbb{E}[Y]$ exists.

Expectations (cont.)

The expectation of a function of Y, $g(Y)$ say, is defined by

$$\mathbb{E}[g(Y)] = \int_{-\infty}^{\infty} g(y) f(y)\, dy$$

if this integral is finite.

Properties of expectations

The key properties of expectations for discrete random variables also hold for continuous random variables. Specifically, if $g_1(.)$ and $g_2(.)$ are two functions, and a and b are constants then

$$\mathbb{E}[ag_1(Y) + bg_2(y)] = a\mathbb{E}[g_1(Y)] + b\mathbb{E}[g_2(Y)]$$

provided $\mathbb{E}[g_1(Y)]$ and $\mathbb{E}[g_2(Y)]$ exist.

Properties of expectations (cont.)

Note

This result follows by the linearity properties of integrals.

$$\int (h_1(y) + h_2(y))\, dy = \int h_1(y)\, dy + \int h_2(y)\, dy$$

Properties of expectations (cont.)

We have identical definitions to the discrete case of the following quantities:

- Variance:

$$\mathbb{V}[X] = \mathbb{E}\left[(Y-\mu)^2\right] = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\, dy$$

Properties of expectations (cont.)

- Moments: for $k = 1, 2, \ldots$

$$\mu'_k = \mathbb{E}\left[Y^k\right] = \int_{-\infty}^{\infty} y^k f(y)\, dy$$

- Central moments: $k = 1, 2, \ldots$

$$\mu_k = \mathbb{E}\left[(Y - \mu)^k\right] = \int_{-\infty}^{\infty} (y - \mu)^k f(y)\, dy$$

Properties of expectations (cont.)

- Factorial moments: $k = 1, 2, \ldots$

$$\begin{aligned}\mu_{[k]} &= \mathbb{E}\left[Y(Y-1)(Y-2)\cdots(Y-k+1)\right] \\ &= \int_{-\infty}^{\infty} y(y-1)(y-2)\cdots(y-k+1)f(y)\,dy\end{aligned}$$

Properties of expectations (cont.)

- Moment-generating function: for $|t| \leq b$, some $b > 0$

$$m(t) = \mathbb{E}\left[e^{tY}\right] = \int_{-\infty}^{\infty} e^{ty} f(y)\, dy$$

In addition, if

$$m^{(k)}(t) = \frac{d^k m(t)}{dt^k}$$

then as in the discrete case

$$m^{(k)}(0) = \mu'_k.$$

Properties of expectations (cont.)

- Factorial moment-generating function: for $1 - b \leq t \leq 1 + b$, some $b > 0$

$$G(t) = \mathbb{E}\left[t^Y\right] = \int_{-\infty}^{\infty} t^y f(y)\, dy$$

In addition, if

$$G^{(k)}(t) = \frac{d^k G(t)}{dt^k}$$

then as in the discrete case

$$G^{(k)}(1) = \mu_{[k]}.$$

Note: We do not use the terminology probability-generating function in the continuous case.

Example

Suppose for some $\alpha > 0$, we have

$$f(y) = \begin{cases} 0 & y \leq 0 \\ \dfrac{c}{(1+y)^{\alpha+1}} & y > 0 \end{cases}$$

Properties of expectations (cont.)

Example

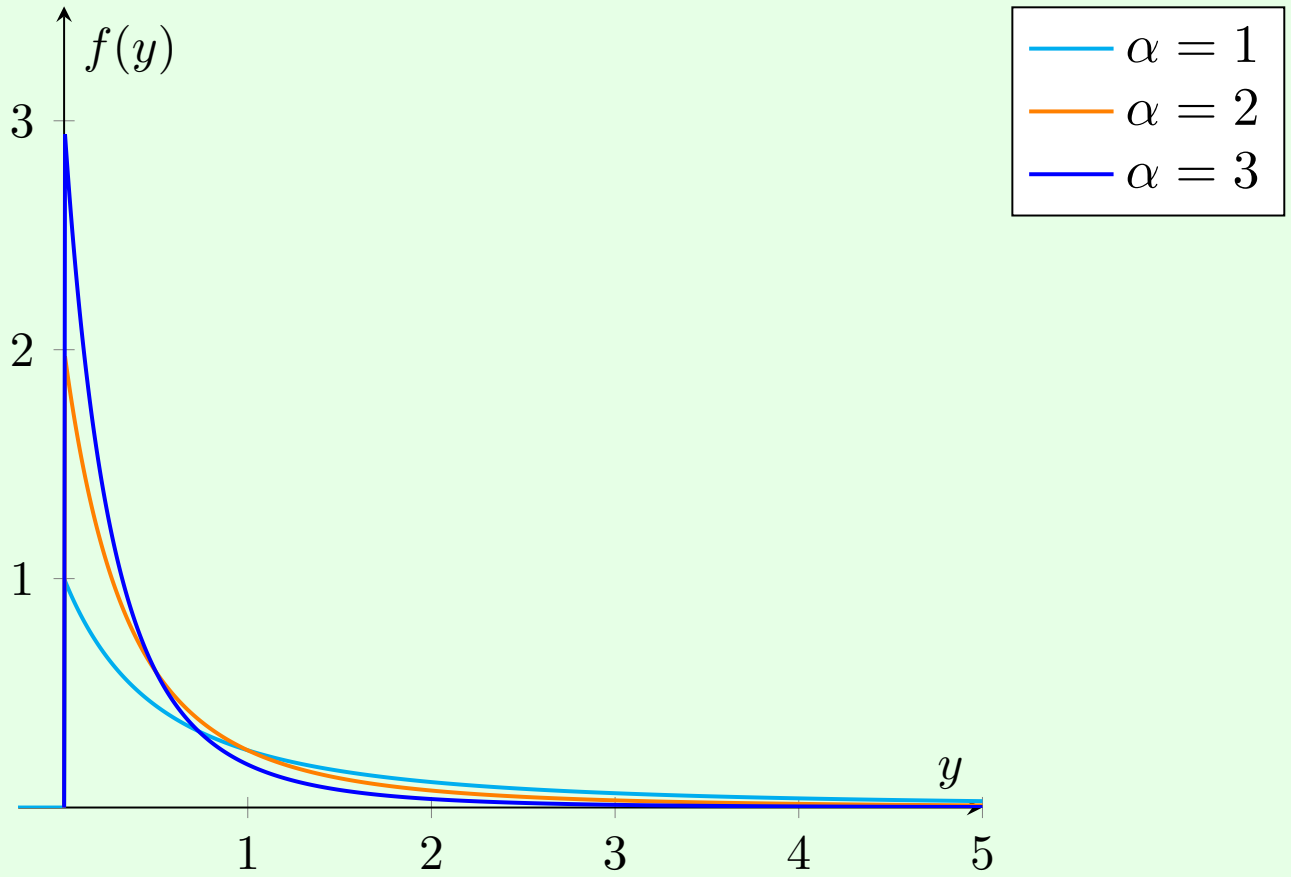
Then

$$F(y) = \begin{cases} 0 & y \leq 0 \\ 1 - \dfrac{c}{(1+y)^{\alpha}} & y > 0 \end{cases}$$

and as we need $F(\infty) = 1$, we must have $c = \alpha$.

Properties of expectations (cont.)

Example



Properties of expectations (cont.)

Example

Then

$$\mu_k' = \mathbb{E}\left[Y^k\right] = \int_{-\infty}^{\infty} y^k f(y)\, dy$$

$$= \int_{0}^{\infty} y^k \frac{\alpha}{(1+y)^{\alpha+1}}\, dy$$

Example

We inspect this integral and note that it is only finite if

$$\alpha > k.$$

If this holds, we can find μ'_k using integration by parts. For example

$$\mu'_1 = \mathbb{E}[Y] = \int_0^\infty y^k \frac{\alpha}{(1+y)^{\alpha+1}}\, dy = \frac{1}{\alpha - 1}$$

provided $\alpha > 1$.

Properties of expectations (cont.)

The quantity α in this pdf is termed a *parameter* of the model. It determines the properties of the model such as

- $f(y)$ and $F(y)$;
- $\mathbb{E}[Y]$ and other moments

The continuous uniform distribution

The continuous uniform distribution has a pdf that is constant on some interval (θ_1, θ_2) say (with $\theta_1 < \theta_2$):

$$f(y) = \frac{1}{\theta_2 - \theta_1} \qquad \theta_1 \leq y \leq \theta_2$$

with $f(y) = 0$ otherwise.

The continuous uniform distribution (cont.)

It is straightforward to show that

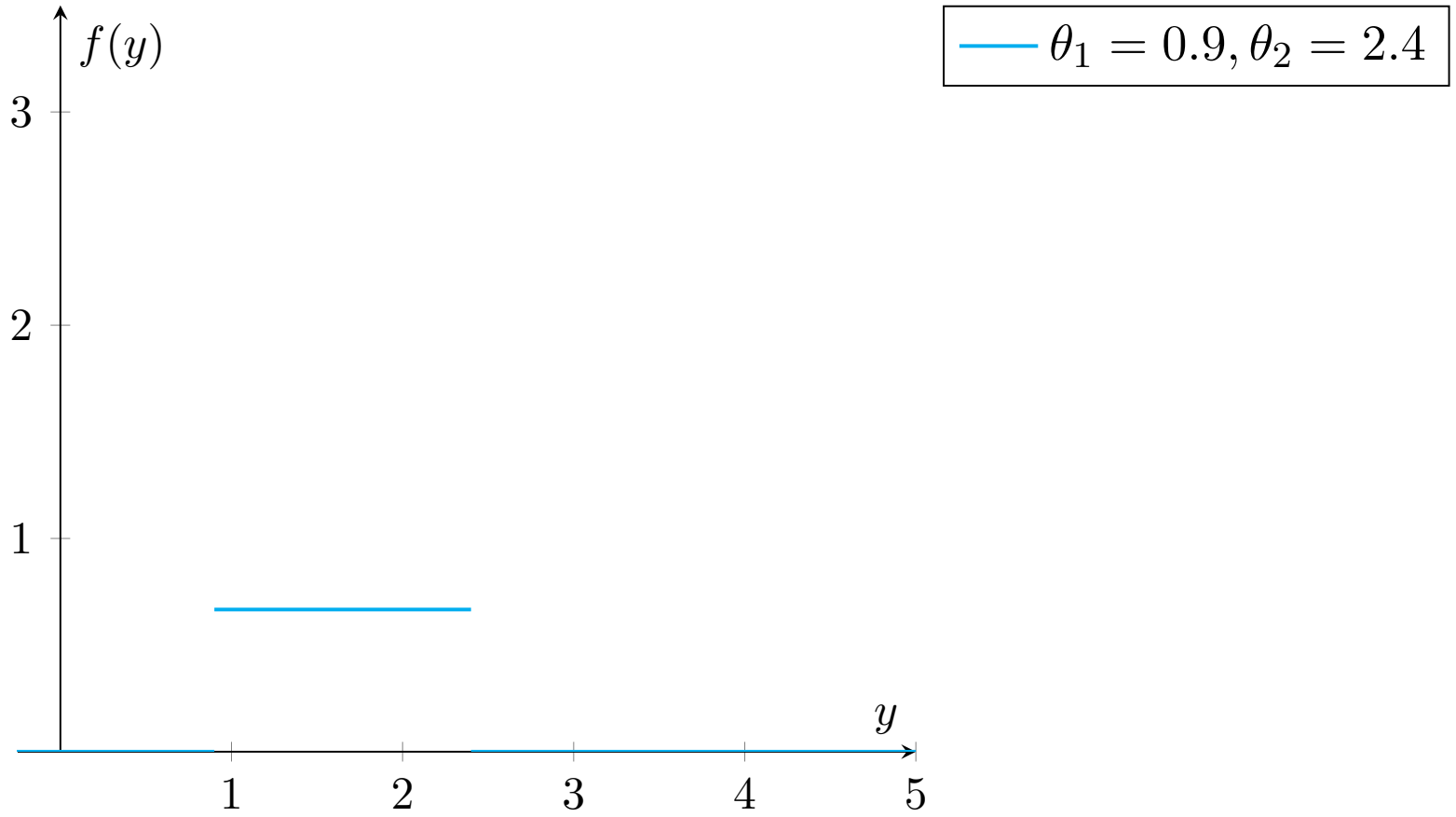
$$F(y) = \begin{cases} 0 & y \leq \theta_1 \\ \dfrac{y - \theta_1}{\theta_2 - \theta_1} & \theta_1 \leq y \leq \theta_2 \\ 1 & y \geq \theta_2 \end{cases} .$$

The continuous uniform distribution (cont.)

By direct calculation, we can compute that

$$\mathbb{E}[Y] = \frac{\theta_1 + \theta_2}{2} \qquad \mathbb{V}[Y] = \frac{(\theta_2 - \theta_1)^2}{12}.$$

The continuous uniform distribution



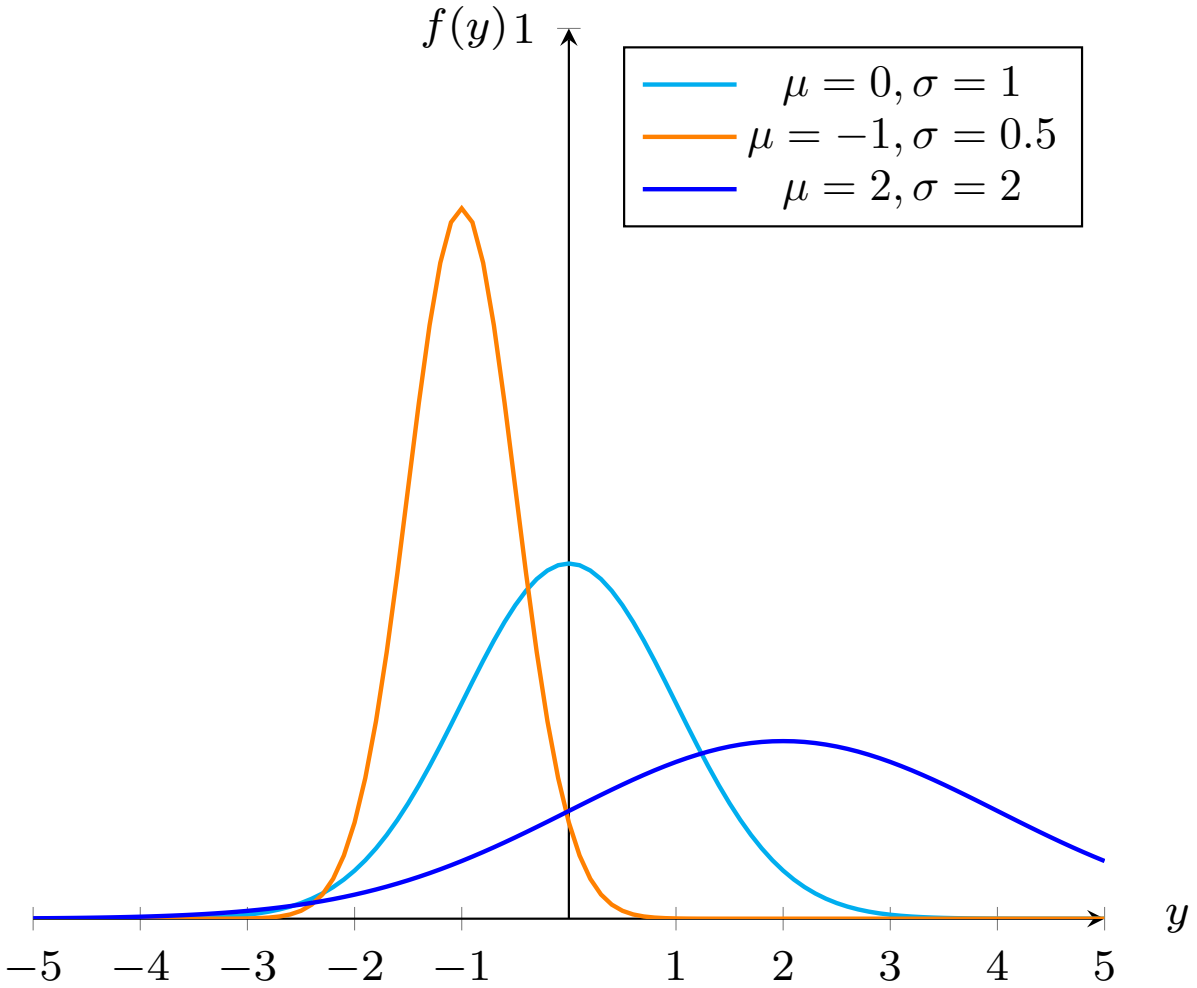
The Normal distribution

The Normal (or Gaussian) distribution has pdf defined on the whole real line by

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \qquad -\infty < y < \infty$$

where μ and $\sigma > 0$ are parameters.

The Normal distribution (cont.)



The Normal distribution (cont.)

Notes

- μ is the centre of the distribution;
  - $f(y)$ is symmetric about μ.
- σ measures the spread of the distribution;
  - as σ decreases, $f(y)$ becomes more concentrated around μ.
- It can be shown that

$$\mathbb{E}[Y] = \mu \qquad \mathbb{V}[Y] = \sigma^2.$$

The Normal distribution (cont.)

The case where $\mu = 0$ and $\sigma = 1$ yields the standard Normal pmf

$$f(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \qquad -\infty < y < \infty$$

The Normal distribution (cont.)

The cdf, $F(y)$, in the general case is given by

$$F(y) = \int_{-\infty}^{y} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(t-\mu)^2\right\}\, dt$$

but this cannot be computed analytically.

The Normal distribution (cont.)

However, if Y has a Normal distribution with parameters μ and σ, and we consider the random variable Z where

$$Z = \frac{Y - \mu}{\sigma}$$

then we see that for real value z

$$\begin{aligned} F_Z(z) = P(Z \leq z) &= P\left(\sigma Z + \mu \leq \sigma z + \mu\right) \\ &= P\left(Y \leq \sigma z + \mu\right) = F_Y\left(\sigma z + \mu\right) \end{aligned}$$

The Normal distribution (cont.)

The quantity on the right-hand side is

$$\int_{-\infty}^{\sigma z+\mu} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(t-\mu)^2\right\}\, dt$$

but if we make the substitution $u = (t-\mu)/\sigma$ in the integral, we see that this becomes

$$\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\}\, du$$

The Normal distribution (cont.)

This is the integral of the standard Normal pdf, so we may conclude that the random variable Z

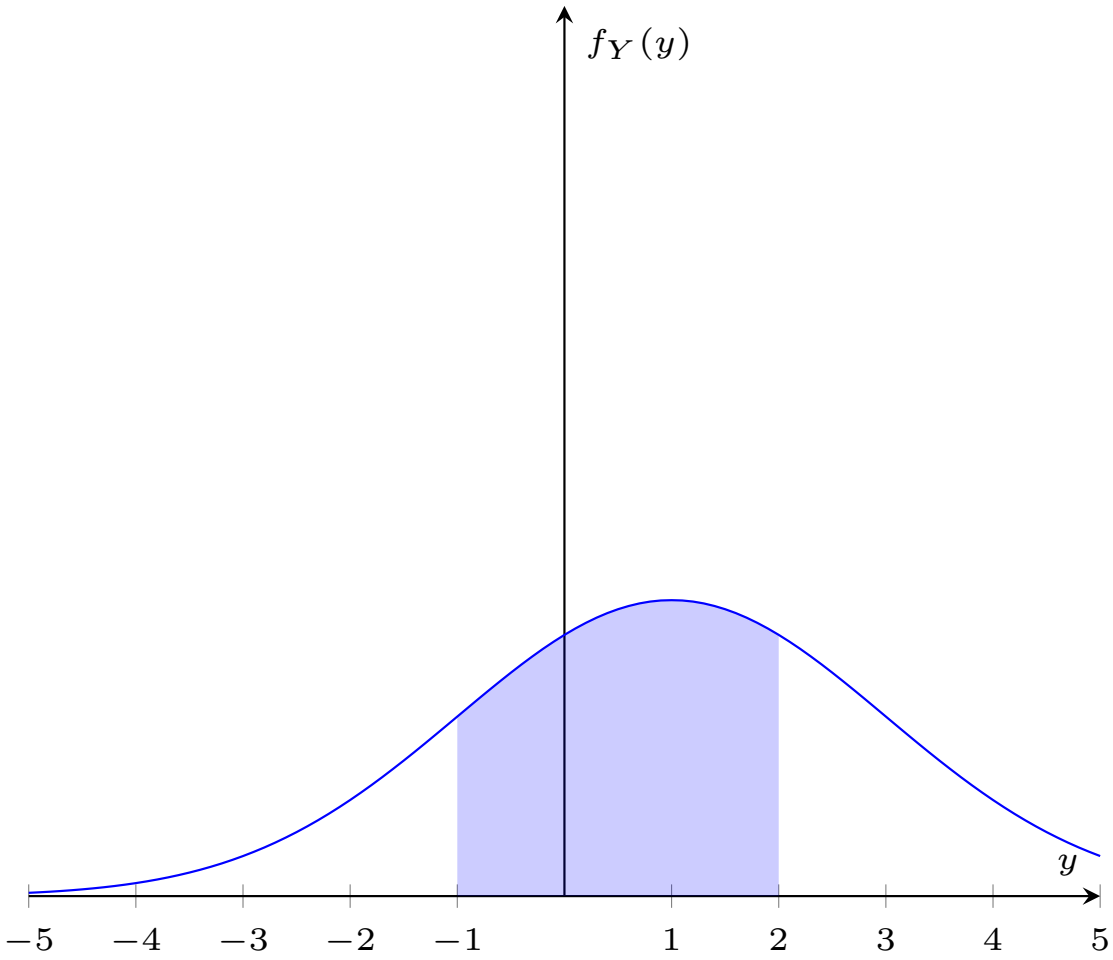
$$Z = \frac{Y - \mu}{\sigma}$$

has a *standard Normal distribution.*

This result tells us that we can always transform a general Normal random variable Y into a standard Normal random variable Z.

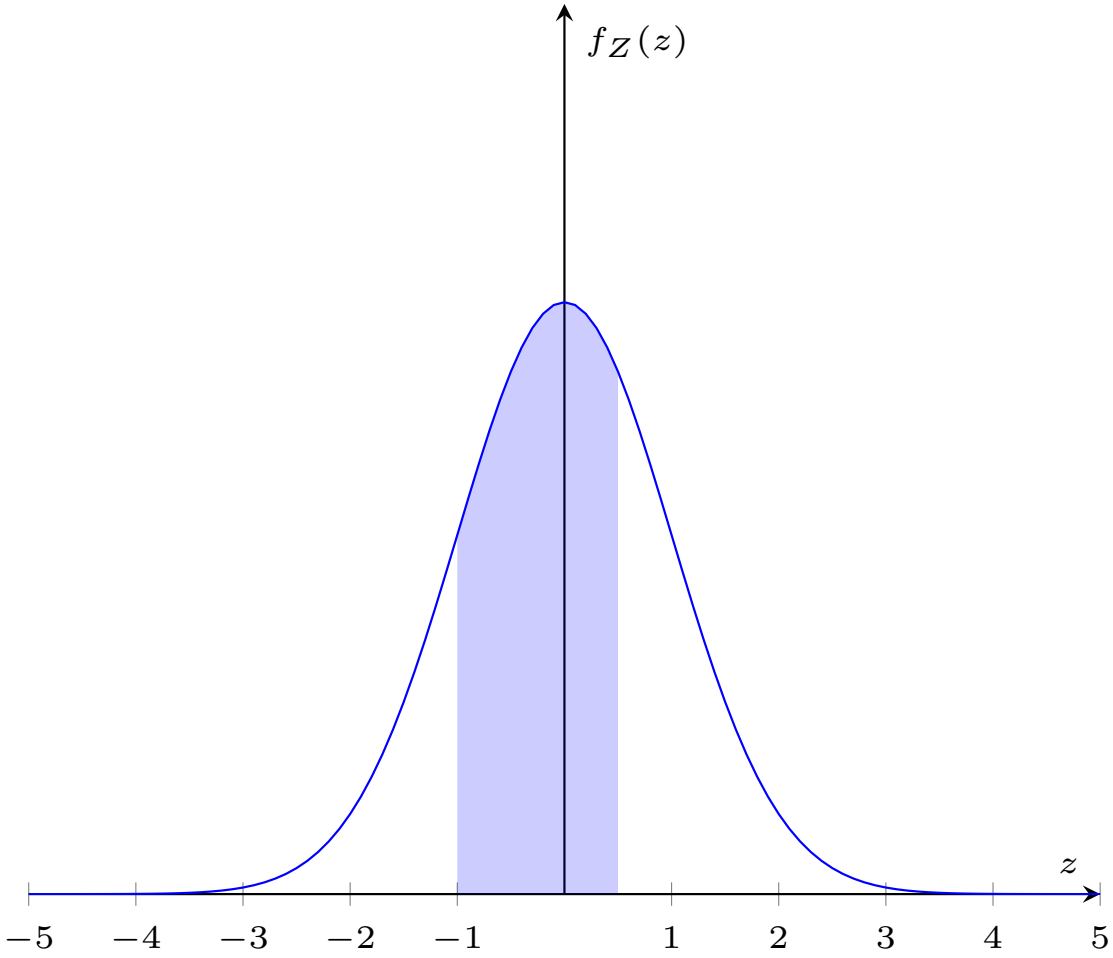
The Normal distribution (cont.)

$$Y : \mu = 1, \sigma = 2 : P(-1 \leq Y \leq 2)$$



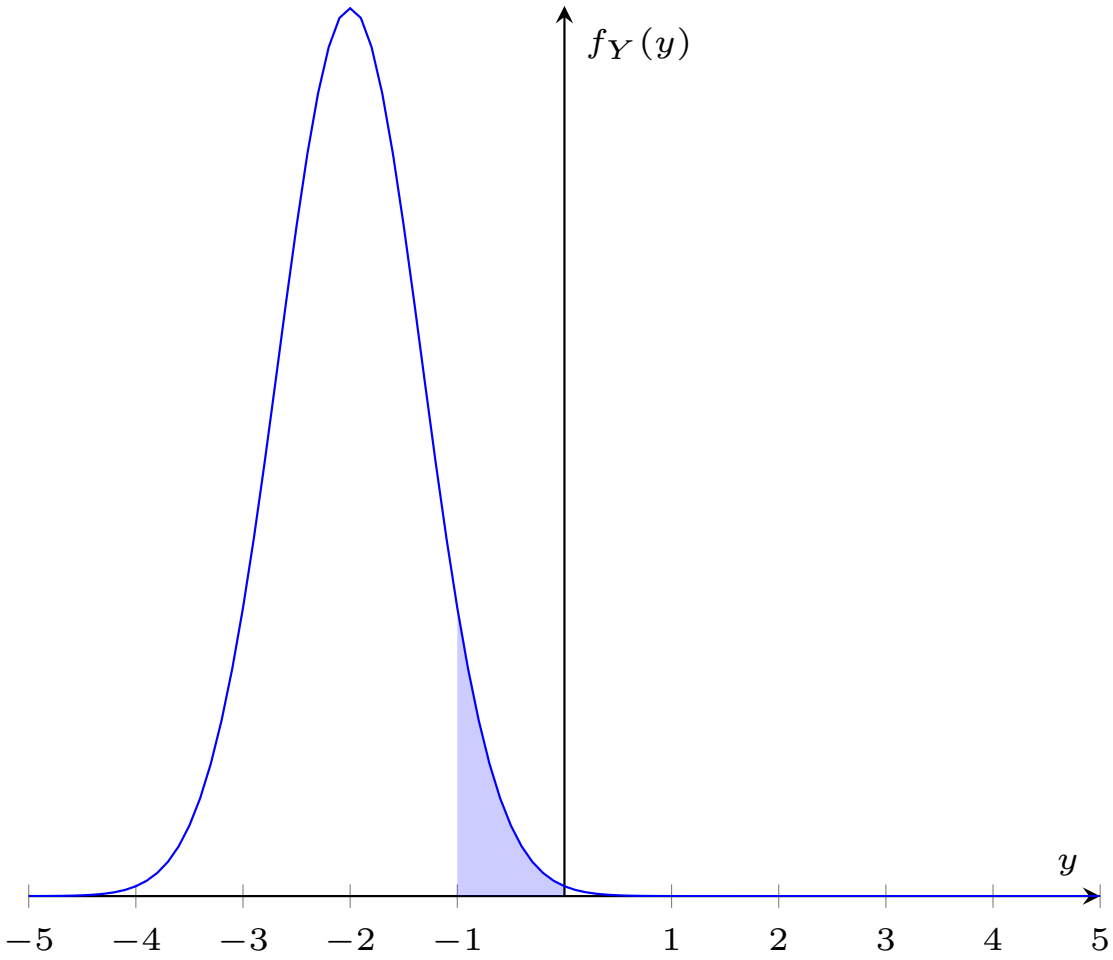
The Normal distribution (cont.)

$$Z : \mu = 0, \sigma = 1 : P(-1 \leq Z \leq 1/2)$$



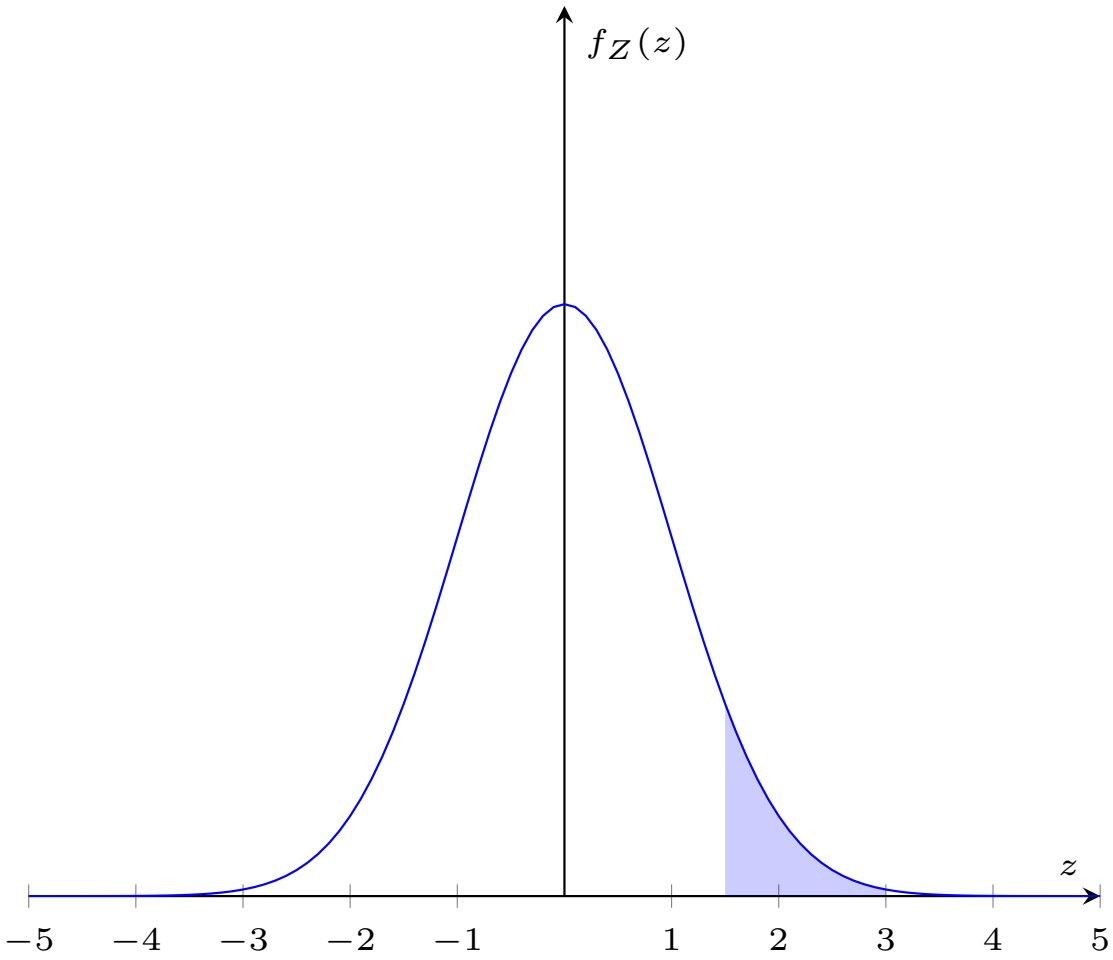
The Normal distribution (cont.)

$$Y : \mu = -2, \sigma = 2/3 : P(-1 \leq Y \leq 0)$$



The Normal distribution (cont.)

$$Z : \mu = 0, \sigma = 1 : P(3/2 \leq Z \leq 3)$$



Note: Change of variable

We have by the change of variable rule that

$$\int_a^b g(t)dt = \int_c^d g(t(u))\frac{dt(u)}{du}du$$

where $t \longrightarrow u$ in the integral.

Moment-generating function

For the mgf of the Normal distribution, first consider the standard Normal case: we have for $t \in \mathbb{R}$ that

$$\begin{aligned} m(t) = \mathbb{E}\left[e^{tY}\right] &= \int_{-\infty}^{\infty} e^{ty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{ty - \frac{1}{2}y^2\right\}\, dy \\ &= \exp\left\{\frac{t^2}{2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y-t)^2}{2}\right\}\, dy \\ &= \exp\left\{\frac{t^2}{2}\right\} \end{aligned}$$

as the integral is the integral of a Normal pdf where the expectation is $\mu = t$.

The Gamma distribution

The Gamma distribution has pdf defined on the whole real line by

$$f(y) = \begin{cases} 0 & y < 0 \\ \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta} & 0 \leq y < \infty \end{cases}$$

where $\alpha, \beta > 0$ are parameters.

The Gamma distribution (cont.)

In this expression, $\Gamma(\alpha)$ is the *Gamma function*

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt.$$

This expression cannot be computed analytically, unless α is a positive integer. However, by integration by parts, we have that

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1).$$

Hence, if $n = 2, 3, 4, \ldots$, we have that

$$\Gamma(n) = (n - 1)!$$

with $\Gamma(1) = 1$ by direct calculation.

The Gamma distribution (cont.)

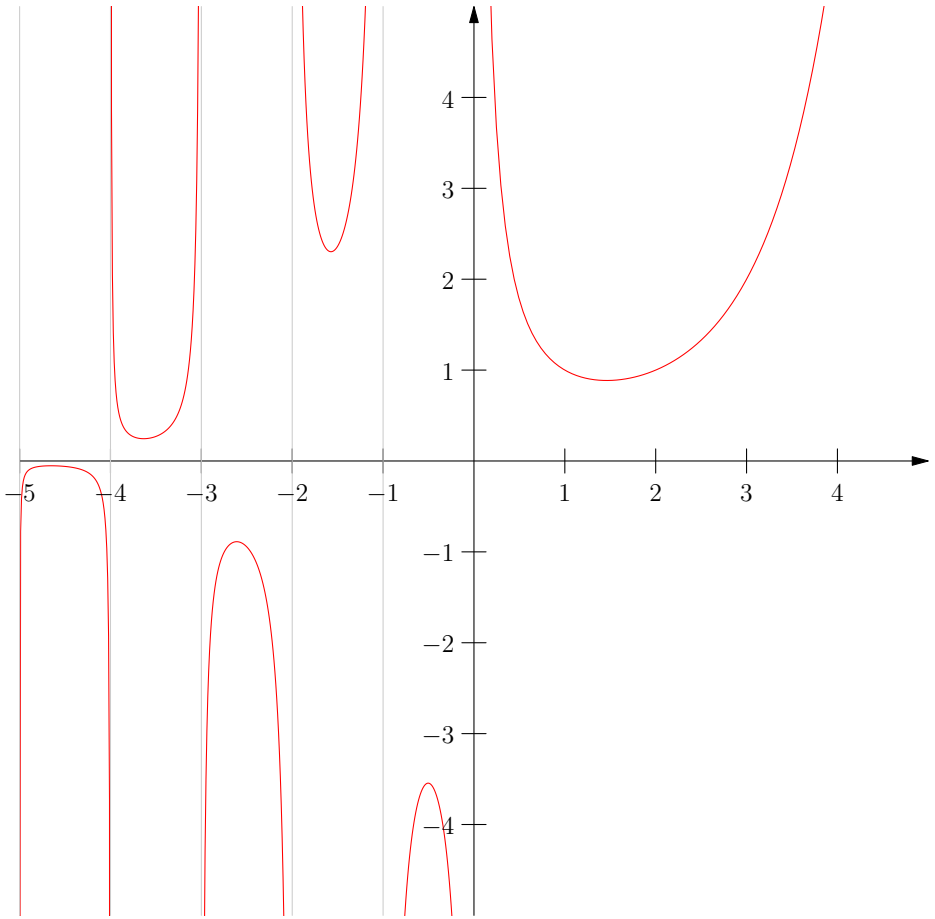
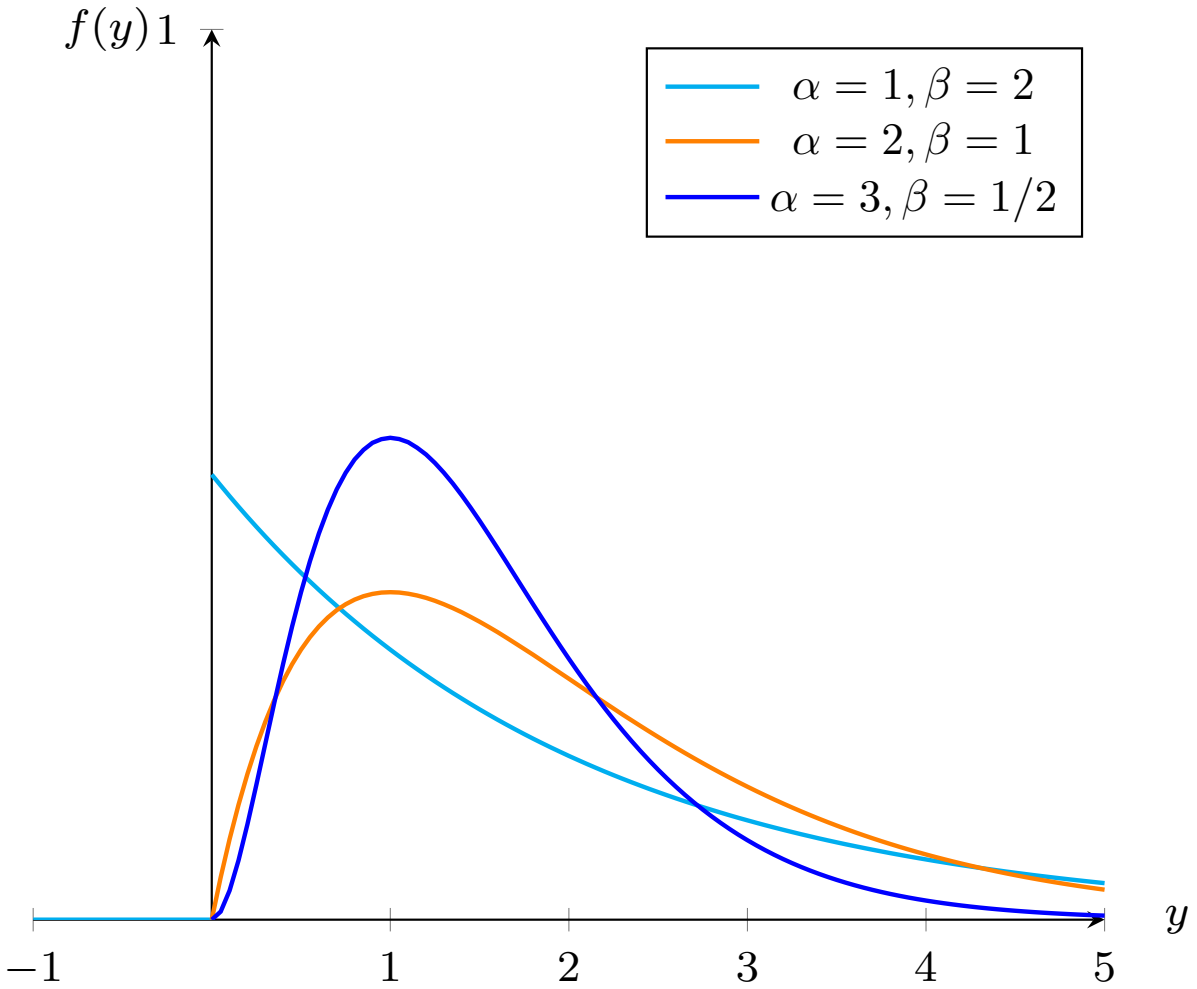


Figure: $\Gamma(\alpha)$ for $-5 \leq \alpha < 5$

The Gamma distribution (cont.)



The Gamma distribution (cont.)

Notes

- α is the shape parameter of the distribution;
- β is the scale parameter of the distribution;
- It can be shown that

$$\mathbb{E}[Y] = \alpha\beta \qquad \mathbb{V}[Y] = \alpha\beta^2.$$

The Gamma distribution (cont.)

For the calculation of the kth moment of the distribution, we have to compute

$$\mathbb{E}[Y^k] = \int_{-\infty}^{\infty} y^k f(y)\, dy = \int_0^{\infty} y^k \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}\, dy.$$

The Gamma distribution (cont.)

Now, as we require $f(y)$ integrate to 1, this implies that

$$\int_0^\infty \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}\, dy = 1$$

for any $\alpha > 0$, so

$$\int_0^\infty y^{\alpha-1} e^{-y/\beta}\, dy = \beta^\alpha \Gamma(\alpha).$$

The Gamma distribution (cont.)

Returning to the moment calculation, we therefore deduce that

$$\begin{aligned}\int_0^\infty y^k \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}\, dy \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty y^{k+\alpha-1} e^{-y/\beta}\, dy \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \beta^{k+\alpha} \Gamma(\alpha+k)\end{aligned}$$

from the previous result.

The Gamma distribution (cont.)

Furthermore

$$\Gamma(\alpha+k) = (\alpha+k-1)\Gamma(\alpha+k-1) = (\alpha+k-1)(\alpha+k-2)\Gamma(\alpha+k-2)$$

etc. until we have the result

$$\Gamma(\alpha+k) = (\alpha+k-1)(\alpha+k-2)\cdots(\alpha+1)\alpha\Gamma(\alpha).$$

The Gamma distribution (cont.)

Therefore

$$E[Y^k] = \frac{1}{\beta^\alpha \Gamma(\alpha)} \beta^{k+\alpha}(\alpha+k-1)(\alpha+k-2)\cdots(\alpha+1)\alpha\Gamma(\alpha)$$

$$= \beta^k(\alpha+k-1)(\alpha+k-2)\cdots(\alpha+1)\alpha$$

so that

$$\mathbb{E}[Y] = \alpha\beta \qquad \mathbb{E}[Y^2] = \alpha(\alpha+1)\beta^2$$

and so on.

The Gamma distribution (cont.)

Note

Sometimes the Gamma distribution is written

$$\frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \quad 0 \leq y < \infty$$

that is, using a different parameterization.

The Gamma distribution (cont.)

Special Cases

1. $\alpha = \nu/2$, $\beta = 2$, where $\nu = 1, 2, \ldots$ is a positive integer. This is the *chi-square distribution* with ν degrees of freedom.

2. $\alpha = 1$. This is the *exponential distribution*

$$f(y) = \frac{1}{\beta} e^{-y/\beta} \quad 0 \leq y < \infty.$$

as $\Gamma(1) = 1$.

The Gamma distribution (cont.)

For the moment-generating function of the Gamma distribution:

$$m(t) = \mathbb{E}\left[e^{tY}\right] = \int_0^\infty e^{ty} \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}\, dy$$

$$= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty y^{\alpha-1} e^{-y(1/\beta - t)}\, dy$$

$$= \frac{1}{\beta^\alpha \Gamma(\alpha)} \Gamma(\alpha) \left(\frac{1}{\beta} - t\right)^{-\alpha} = \left(\frac{1}{1-\beta t}\right)^\alpha$$

as the integral is proportional to a Gamma pdf with parameters α and

$$\left(\frac{1}{\beta} - t\right)^{-1}.$$

The Beta distribution

The *Beta distribution* is suitable for a random variable defined on the range $[0, 1]$; its pdf takes the form

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \qquad 0 \leq y \leq 1$$

and $f(y) = 0$ for all other values of y. The constant

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

ensures that the pdf integrates to 1.

The Beta distribution (cont.)

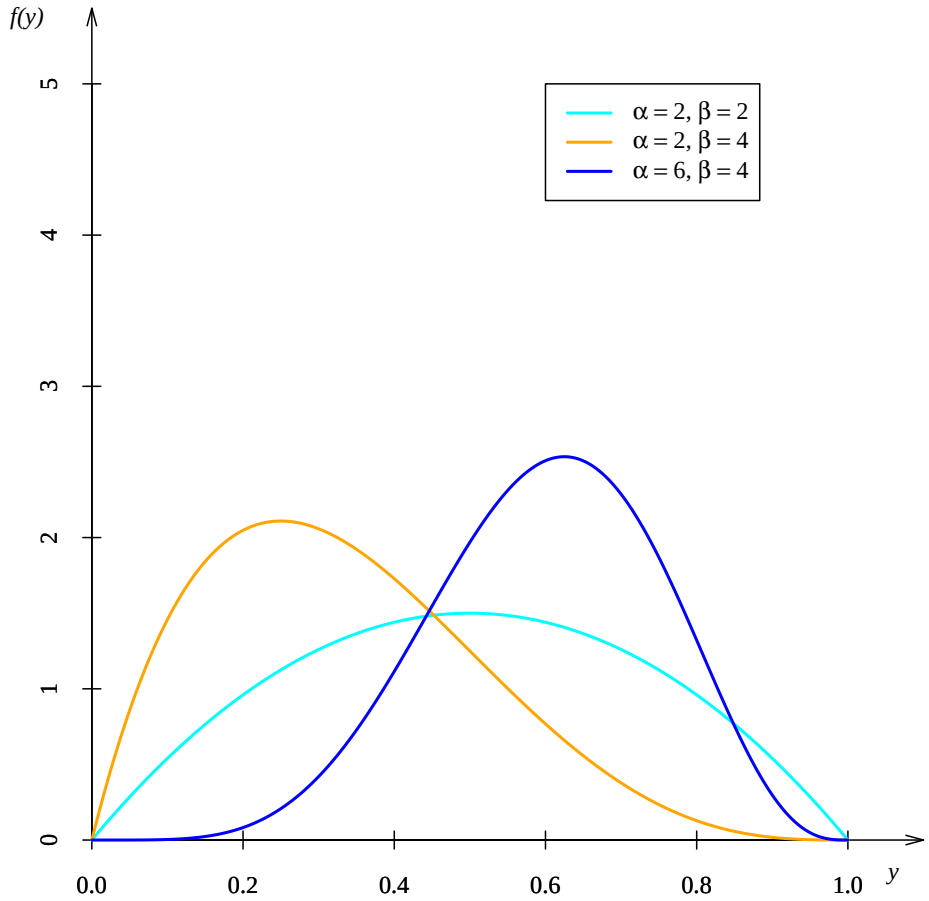
The *Beta function* is defined

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\, dy.$$

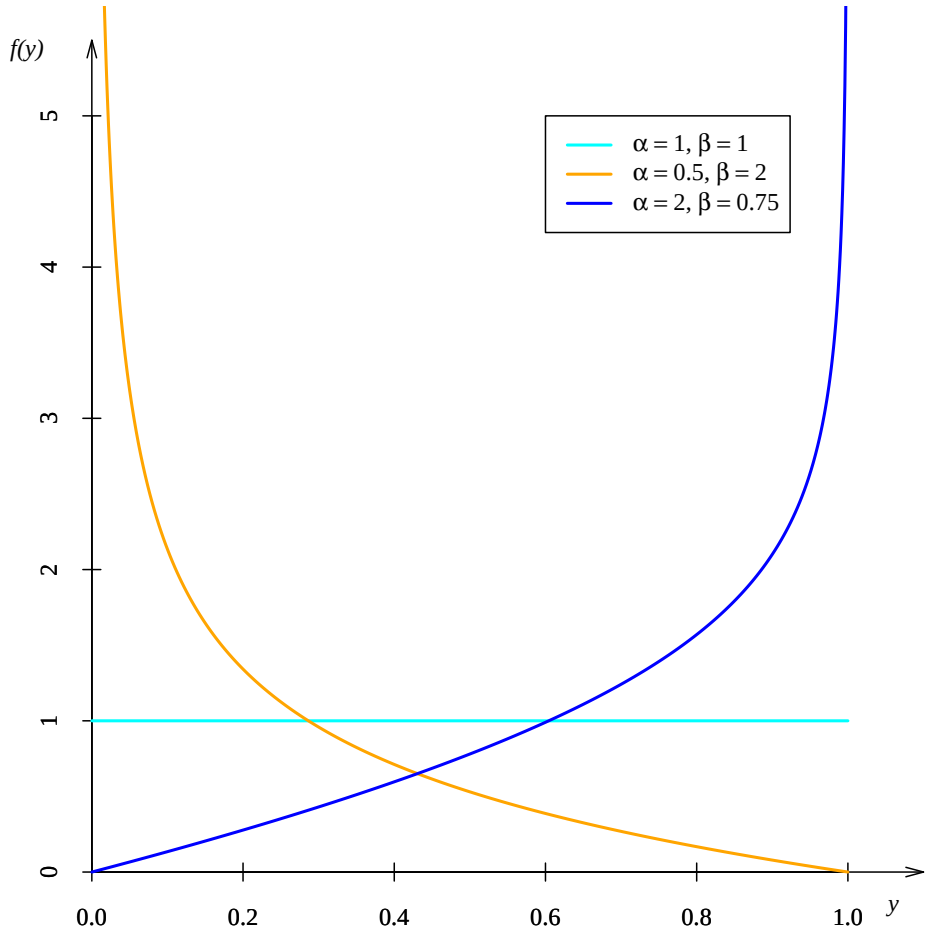
so we may write for the Beta pdf

$$f(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} \qquad 0 \leq y \leq 1.$$

The Beta distribution (cont.)



The Beta distribution (cont.)



The Beta distribution (cont.)

Note

The case $\alpha = \beta = 1$ has

$$B(\alpha, \beta) = \frac{\Gamma(1)\Gamma(1)}{\Gamma(2)} = \frac{1 \times 1}{1} = 1$$

and the pdf is

$$f(y) = 1 \qquad 0 \leq y \leq 1$$

with $f(y) = 0$ for other values of y.

This is the continuous uniform distribution with parameters $\theta_1 = 0$ and $\theta_2 = 1$.

The Beta distribution (cont.)

Note that for $k = 1, 2, \ldots$

$$\mathbb{E}\left[Y^k\right] = \int_{-\infty}^{\infty} y^k f(y)\, dy = \int_0^1 y^k \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}\, dy$$

$$= \frac{1}{B(\alpha,\beta)} \int_0^1 y^{\alpha+k-1}(1-y)^{\beta-1}\, dy$$

$$= \frac{1}{B(\alpha,\beta)} B(\alpha+k,\beta)$$

Chapter 3: Probability calculation methods

Probability calculation methods

We now study two specific calculation methods relevant for probability and distribution calculations:

- transformations;
- sums of random variables.

Transformations

For a given random variable Y, we are interested in calculating the probability distribution of the *transformed* random variable

$$U = h(Y)$$

where $h(.)$ is a real-valued function with domain that includes the region $\mathcal{Y}$ on which Y takes values with positive probability.

Transformations (cont.)

Discrete case: In the discrete case, we consider how the probability attached to individual values is transported under the transformation.

Suppose Y takes values on the (countable) set

$$\{y_1, y_2, \ldots, y_n, \ldots\}.$$

Under the transformation $h(.)$, suppose the individual values map to the (countable) set

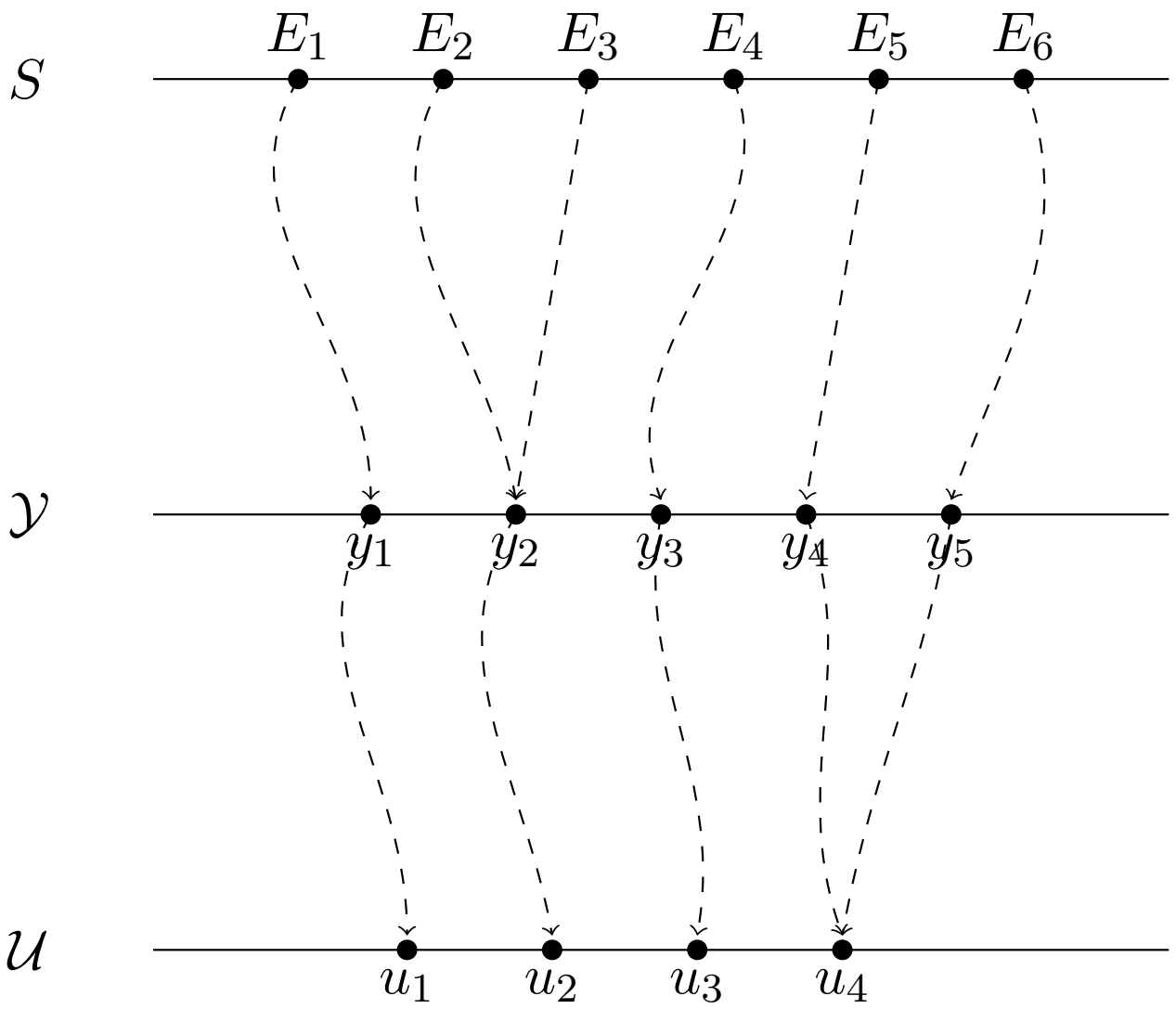
$$\{u_1, u_2, \ldots, u_n, \ldots\}$$

where

$$u_i = h(y_j)$$

for at least one j.

Transformations (cont.)



Transformations (cont.)

In the above example

$$u_1 = h(y_1) \quad u_2 = h(y_2) \quad u_3 = h(y_3) \quad u_4 = h(y_4) \text{ and } h(y_5)$$

We typically deal with simple transformations, for example

$$U = 2Y \qquad U = \exp(-Y) \qquad U = \frac{Y}{Y+1} \qquad U = Y^2$$

Transformations (cont.)

To derive the probability distribution of U, we simply see how the probability accumulates on individual values under the transformation.

In the picture above, we have four values $\{u_1, u_2, u_3, u_4\}$ which receive probability as follows:

$$p_U(u) = P(U = u) = \begin{cases} p_Y(y_1) & u = u_1 \\ p_Y(y_2) & u = u_2 \\ p_Y(y_3) & u = u_3 \\ p_Y(y_4) + p_Y(y_5) & u = u_4 \end{cases}$$

Transformations (cont.)

For 1-1 transformations, this is extremely straightforward to apply; we simply

- identify the set of values $\{u_1, u_2, \ldots, u_n, \ldots\}$ that the values $\{y_1, y_2, \ldots, y_n, \ldots\}$ map onto – in this case there is a one-to-one correspondence between the sets, and

$$u_i = h(y_i);$$

- for each i, assign the probability $p_Y(y_i)$ to u_i to obtain $p_U(u_i)$

Example (Poisson case)

Suppose that Y has a Poisson distribution with parameter λ. Let $U = 3Y$; then only the values in the set

$$\mathcal{U} = \{0, 3, 6, 9, \ldots, 3n, \ldots\}$$

receive any probability, and

$$p_U(u) = \frac{\lambda^{u/3}}{(u/3)!} e^{-\lambda} \qquad u \in \{0, 3, 6, 9, \ldots\}$$

with $p_U(u)$ equal to zero for all other values of u.

Transformations (cont.)

If the transformation is not 1-1, we essentially have to identify the set $\mathcal{U}$, and compute for each case the probability assigned to the values in the set.

Transformations (cont.)

Continuous case: In the continuous random variable case, we try to adopt the same strategy and follow how values in the set $\mathcal{Y}$ carry their probability to values in a set $\mathcal{U}$.

The obstacle is that for a continuous random variable Y we have

$$P(Y = y) = 0$$

for all real values y.

Transformations (cont.)

To get around this, we can instead consider

$$P(Y \in \mathcal{A})$$

for a subset (eg interval) of the real line. For example, we could consider $\mathcal{A} = [2, 5]$, and the probability

$$P(Y \in \mathcal{A}) \equiv P(2 \leq Y \leq 5).$$

Transformations (cont.)

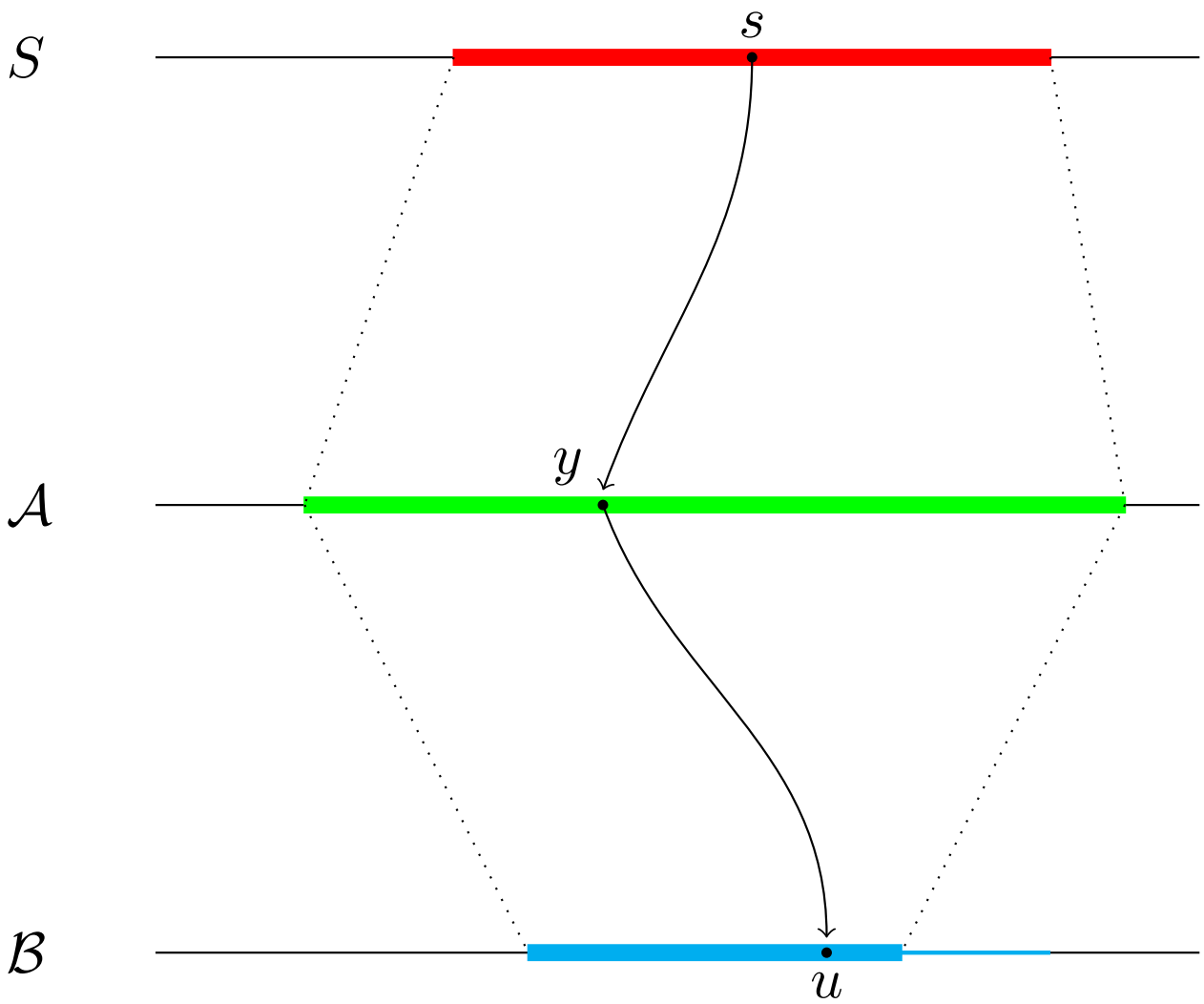
Suppose that $\mathcal{B}$ represents the set of values containing all the points that points in $\mathcal{A}$ can be mapped to under the transformation $h(.)$. That is,

$$\mathcal{B} = \{u : u = h(y), \text{some } y \in \mathcal{A}\}$$

Then we have that probability is preserved under the mapping

$$P(Y \in \mathcal{A}) = P(U \in \mathcal{B}).$$

Transformations (cont.)



Example

Suppose Y is a continuous random variable that can take values on $[0, \infty)$, and $U = h(Y) = 2Y$. This means that we need to consider the set

$$\mathcal{U} = \{u : u = 2y, \text{ some } y \in [0, \infty)\} \equiv [0, \infty)$$

that is, every point y gets mapped to $u = 2y$.

Example

Consider interval $\mathcal{A} = [a, b]$ for $a < b$; if we let $c = h(a) = 2a$ and $d = 2b$, then we have that

$$P(a \leq Y \leq b) = P(c \leq U \leq d).$$

In terms of the cdfs, we have that

$$F_Y(b) - F_Y(a) = F_U(d) - F_U(c)$$

and letting $a \longrightarrow 0$, we have that

$$F_Y(b) = P(Y \leq b) = P(U \leq d) = F_U(d).$$

Example

That is, for $b \in \mathbb{R}$.

$$F_Y(b) = F_U(2b)$$

Now, given F_Y, we wish to compute F_U. We can do this by noting that, for arbitrary u

$$F_U(u) = P(U \leq u) = P(U/2 \leq u/2) = P(Y \leq u/2) = F_Y(u/2).$$

Example

To get to the pdf, we need to differentiate both sides with respect to u:

$$f_U(u) = \frac{dF_U(u)}{du} = \frac{dF_Y(u/2)}{du} = \frac{1}{2} f_Y(u/2)$$

by the chain rule. For example, if $F_Y(y) = 1 - \exp\{-y\}$, then for $u \geq 0$,

$$f_U(u) = \frac{d}{du} \{1 - \exp\{-u/2\}\} = \frac{1}{2} \exp\{-u/2\}.$$

Transformations (cont.)

The key relationship here is that

$$P(Y \in \mathcal{A}) = P(U \in \mathcal{B}).$$

We want to compute $F_U(.)$ or $f_U(.)$, which means we want to fix $\mathcal{B}$ first. Therefore, write

$$h^{-1}(\mathcal{B})$$

for the set of y values that map onto $\mathcal{B}$. Then we have

$$P(U \in \mathcal{B}) = P(Y \in h^{-1}(\mathcal{B}))$$

Transformations (cont.)

If $h(.)$ is a 1–1 transformation, then $h^{-1}(.)$ is the inverse transformation which we can often compute explicitly.

$h(y)$	$h^{-1}(y)$
$y+\theta$	$y-\theta$
λy	y/λ
e^y	$\ln y$
$\dfrac{y}{1+y}$	$\dfrac{y}{1-y}$

etc.

Transformations (cont.)

For 1–1 (or monotonic) transformations, we consider two cases

1. $h(y)$ increasing: eg $U = 2X$
2. $h(y)$ decreasing: eg $U = 2/X$

Transformations (cont.)

For a 1–1 increasing transformation, we have for the cdf

$$F_U(u) = P(U \leq u) = P(h(Y) \leq u) = P(Y \leq h^{-1}(u))$$
$$= F_Y(h^{-1}(u))$$

and therefore for the pdf, by the chain rule

$$f_U(u) = \frac{dF_U(u)}{du} = \frac{dF_Y(h^{-1}(u))}{du} = f_Y(h^{-1}(u))\frac{dh^{-1}(u)}{du}.$$

Transformations (cont.)

For a 1–1 decreasing transformation, we have instead

$$F_U(u) = P(U \leq u) = P(h(Y) \leq u) = P(Y \geq h^{-1}(u))$$
$$= 1 - F_Y(h^{-1}(u))$$

and therefore for the pdf, by the chain rule

$$f_U(u) = \frac{dF_U(u)}{du} = -\frac{dF_Y(h^{-1}(u))}{du} = -f_Y(h^{-1}(u))\frac{dh^{-1}(u)}{du}.$$

Therefore we can write

$$f_U(u) = f_Y(h^{-1}(u)) \left| \frac{dh^{-1}(u)}{du} \right|$$

to cover both cases. The term

$$\left| \frac{dh^{-1}(u)}{du} \right|$$

is the *Jacobian* of the transformation.

Transformations (cont.)

Note

The identity

$$P(U \in \mathcal{B}) = P(Y \in h^{-1}(\mathcal{B}))$$

can be re-written

$$\int_{\mathcal{B}} f_U(u)\, du = \int_{h^{-1}(\mathcal{B})} f_Y(y)\, dy$$

Note (cont.)

Changing variables in the right-hand integral

$$u = h(y) \quad \Longleftrightarrow \quad y = h^{-1}(u)$$

we have by the rules of integration

$$\int_{\mathcal{B}} f_U(u)\, du = \int_{\mathcal{B}} f_Y(h^{-1}(u)) \left| \frac{dh^{-1}(u)}{du} \right| du$$

and the same result follows by equating the integrands.

Transformations (cont.)

If $h(.)$ is not 1–1 (non-monotonic), then the identity

$$P(U \in \mathcal{B}) = P(Y \in h^{-1}(\mathcal{B}))$$

still holds, we just need more care in computing the right-hand side.

Example

Suppose that Y has a continuous uniform distribution on the range $[0, 1]$, and that

$$U = Y(1 - Y)$$

that is $U = h(Y)$ with $h(t) = t(1 - t)$.

First, as $\mathcal{Y} = [0, 1]$, we have that

$$\mathcal{U} = [0, 1/4].$$

Example

Then for $0 \leq u \leq 1/4$, we have that

$$P(U \leq u) = P(Y(1-Y) \leq u) = P(Y - Y^2 - u \leq 0)$$

That is

$$F_U(u) = P(Y^2 - Y + u \geq 0) = P((Y - y_1)(Y - y_2) \geq 0)$$

where, factorizing the quadratic, we have

$$(y_1, y_2) = \left(\frac{1 - \sqrt{1-4u}}{2}, \frac{1 + \sqrt{1-4u}}{2} \right)$$

Example

Now, in the quadratic

$$(y - y_1)(y - y_2) \geq 0 \qquad \Longleftrightarrow y \leq y_1 \quad \text{or} \quad y \geq y_2$$

so for $0 \leq u \leq 1/4$

$$\begin{aligned} F_U(u) &= P((Y - y_1)(Y - y_2) \geq 0) \\ &= P(Y \leq y_1) + P(Y \geq y_2) \\ &= F_Y(y_1) + (1 - F_Y(y_2)) \\ &= y_1 + 1 - y_2 \\ &= 1 - \sqrt{1 - 4u}. \end{aligned}$$

Example

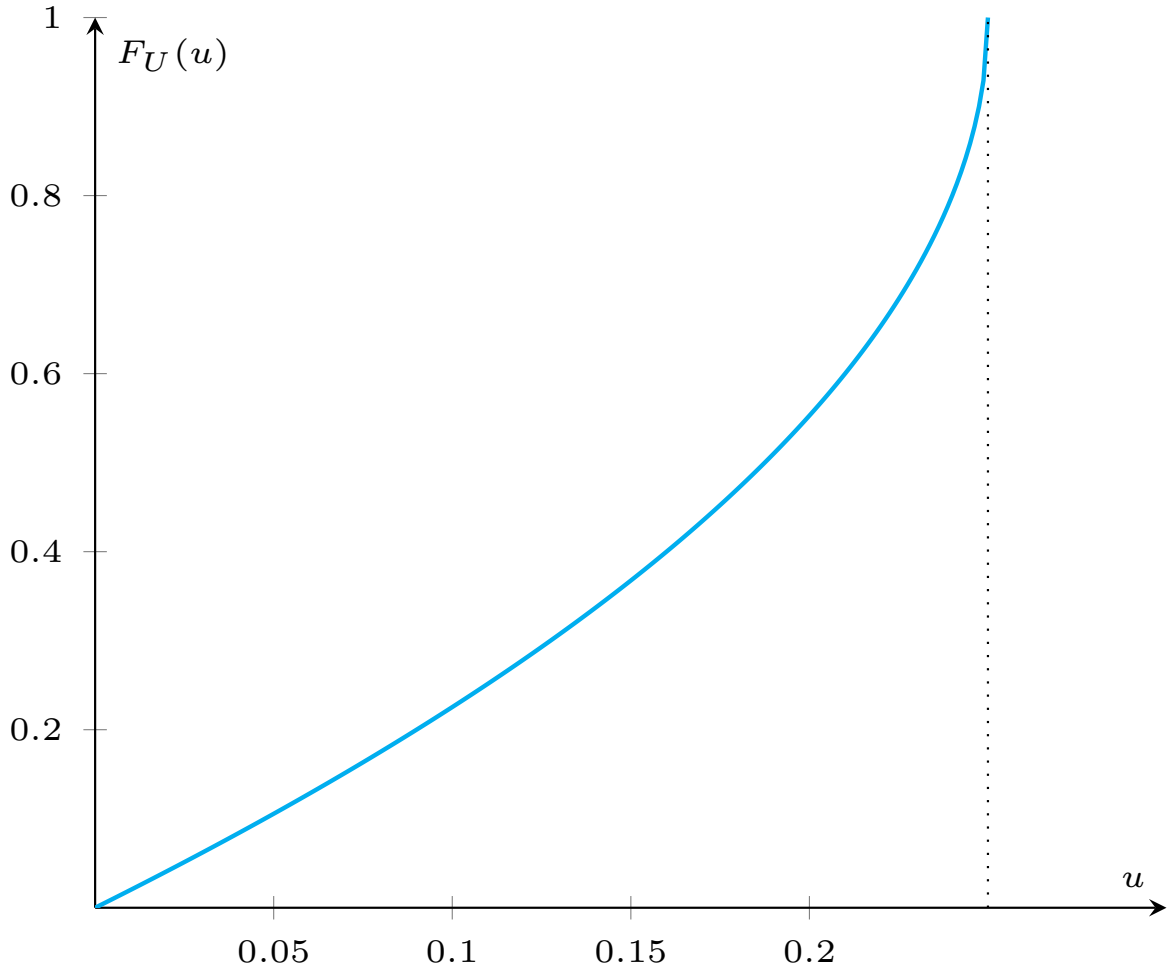
Differentiating, we have

$$f_U(u) = \frac{2}{\sqrt{1-4u}} \qquad 0 \leq u \leq 1/4.$$

For $u < 0$ and $u > 1/4$, $f_U(u) = 0$.

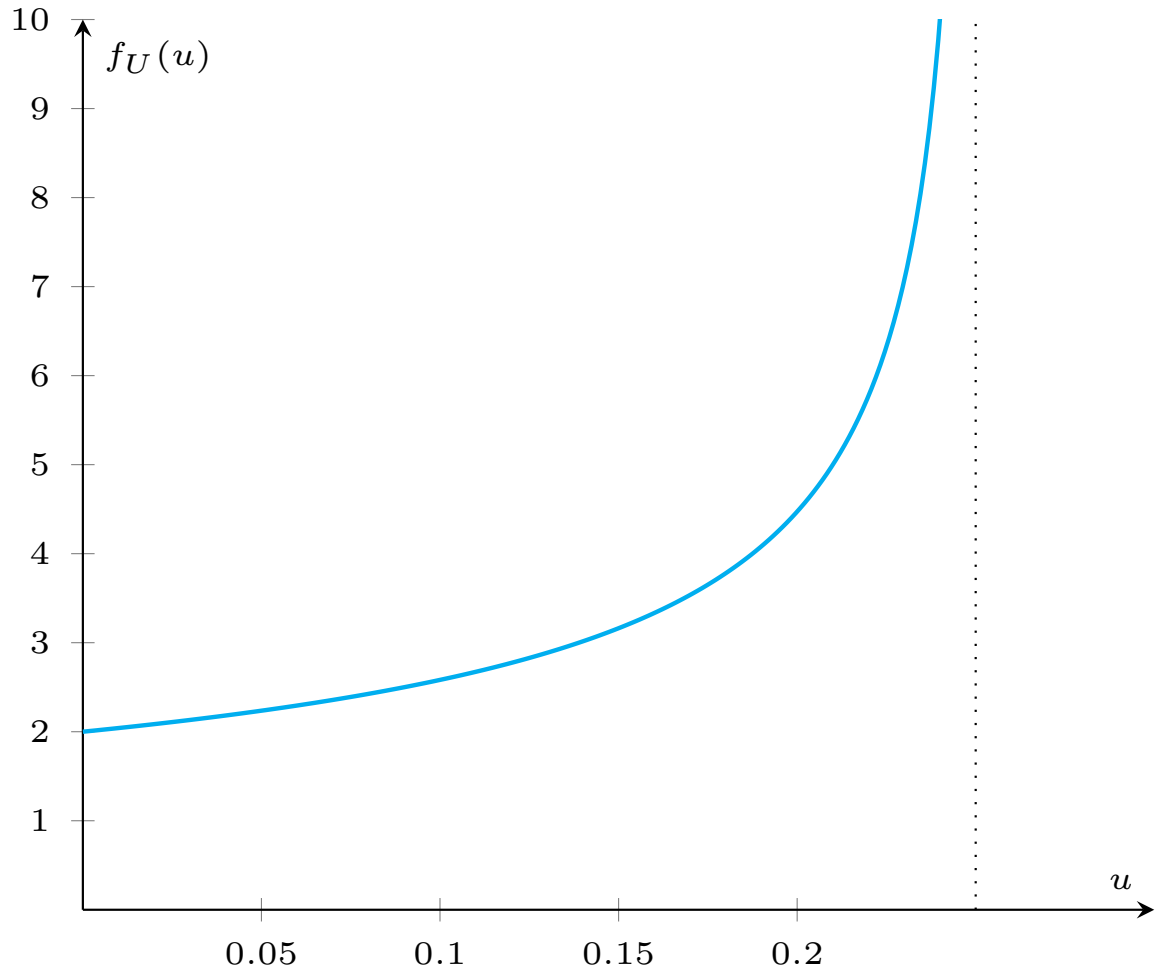
Transformations (cont.)

$$F_U(u) = 1 - \sqrt{1 - 4u}.$$



Transformations (cont.)

$$f_U(u) = 2/\sqrt{1-4u}.$$



Transformations (cont.)

Note

For non 1-1 transformations, it is usually necessary to go back to 'first principles' and investigate

$$P(U \in \mathcal{B}) = P(Y \in h^{-1}(\mathcal{B}))$$

for $\mathcal{B} = (-\infty, u]$ and then carefully construct $h^{-1}(\mathcal{B})$ from it.

Expectations

We previously saw for continuous random variable Y and function $g(.)$

$$\mathbb{E}[g(Y)] = \int_{-\infty}^{\infty} g(y) f_Y(y)\, dy.$$

However, $U = g(Y)$ is a random variable in its own right. We have

$$\mathbb{E}[U] = \int_{-\infty}^{\infty} u f_Y(U)\, du.$$

From the above argument, we see that

$$\mathbb{E}[U] = \mathbb{E}[g(Y)]$$

that is, we can regard the computation in two equivalent ways.

Moment Generating Functions

Recall the definition of the moment generating function (mgf): for $-b \leq t \leq b$, $b > 0$,

$$m_Y(t) = \mathbb{E}[e^{tY}].$$

If $U = cY + d$, for constants $c > 0$ and d, we have that

$$m_U(t) = \mathbb{E}\left[e^{tU}\right] = \mathbb{E}\left[e^{t(cY+d)}\right] = \mathbb{E}\left[e^{(ct)Y}e^{dt}\right] = e^{dt}\mathbb{E}\left[e^{(ct)Y}\right]$$

as e^{dt} is a constant. Therefore

$$m_U(t) = e^{dt}m_Y(ct).$$

Moment Generating Functions: Normal case

If Y has a standard Normal distribution,

$$m_Y(t) = \exp\left\{\frac{t^2}{2}\right\} \qquad t \in \mathbb{R}.$$

If $U = \sigma Y + \mu$, then

$$m_U(t) = e^{\mu t} m_Y(\sigma t) = \exp\left\{\mu t + \frac{(\sigma t)^2}{2}\right\}.$$

Thus, by the earlier result for the transformations of standard Normal random variables

$$m(t) = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$$

is the mgf of the $Normal(\mu, \sigma^2)$ distribution.

Sums of random variables

Example (Two dice)

Two fair dice are rolled, with the outcomes of the rolls independent events.

What is the probability that the sum of the scores of the two dice exceeds 10 ?

Sums of random variables (cont.)

Example (Two dice)

Let Y_1 and Y_2 be the random variables recording the scores on the two dice: Y_1 and Y_2 have discrete uniform distributions on $\{1, 2, 3, 4, 5, 6\}$.

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

The outcome combinations in this table are equally likely, and occur with probability 1/36.

Sums of random variables (cont.)

Example (Two dice)

Let Y denote the total score. The possible total scores are

$$\mathcal{Y} = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

and to compute $P(Y = y)$ we can sum down the antidiagonals of the table to obtain

$$p(y) = P(Y = y) = \frac{\min\{6, y-1\} - \max\{1, y-6\} + 1}{36} \quad y \in \mathcal{Y}$$

Example (Two dice)

That is

$$p(y) = \begin{cases} 1/36 & y = 2, 12 \\ 2/36 & y = 3, 11 \\ 3/36 & y = 4, 10 \\ 4/36 & y = 5, 9 \\ 5/36 & y = 6, 8 \\ 6/36 & y = 7. \end{cases}$$

Therefore

$$P[Y > 10] = P[Y = 11] + P[Y = 12] = \frac{3}{36} = \frac{1}{12}.$$

Sums of random variables (cont.)

In the example, we are computing the probability distribution of

$$Y = Y_1 + Y_2$$

under the assumption that Y_1 and Y_2 are *independent random variables*, that is,

- they are connected to events that are independent in the original sample space.

Sums of random variables (cont.)

The assumption of independence of the events means that we must have

$$P((Y_1 = y_1) \cap (Y_2 = y_2)) = P(Y_1 = y_1) \times P(Y_2 = y_2)$$

for all pairs of possible realizations (y_1, y_2), each value taken from the set $\{1, 2, 3, 4, 5, 6\}$.

Sums of random variables (cont.)

Now consider the event $(Y = y)$ for the set of possible values

$$y \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\};$$

we have the partition given by result of the first roll

$$(Y = y) \equiv \bigcup_{y_1=1}^{6} (Y = y \cap Y_1 = y_1).$$

– that is, when we observe $Y = 5$, say, that could arise as the unordered pairs

$$(1, 4), (2, 3), (3, 2), (4, 1).$$

Sums of random variables (cont.)

However the event

$$(Y = y \cap Y_1 = y_1)$$

can be re-written

$$(Y_2 = y - y_1 \cap Y_1 = y_1)$$

as if $Y = y$ and $Y_1 = y_1$, we **must** have $Y_2 = y - y_1$.

Sums of random variables (cont.)

Using Axiom III, we can then write

$$P(Y = y) = \sum_{y_1=1}^{6} P((Y = y) \cap (Y_1 = y_1))$$

$$= \sum_{y_1=1}^{6} P((Y_2 = y - y_1) \cap (Y_1 = y_1))$$

$$= \sum_{y_1=1}^{6} P(Y_1 = y_1) P(Y_2 = y - y_1)$$

Sums of random variables (cont.)

For the range of the summation, we require for fixed y that

$$1 \leq y_1 \leq 6$$

so that $P(Y_1 = y_1) > 0$ and

$$1 \leq y - y_1 \leq 6$$

in order to make

$$P(Y_2 = y - y_1) > 0.$$

Sums of random variables (cont.)

That is we need to have

$$1 \leq y_1 \leq 6$$

$$y - 6 \leq y_1 \leq y - 1$$

in order for a term in the sum to be non-zero.

Sums of random variables (cont.)

Therefore we need

$$\max\{1, y-6\} \leq y_1 \leq \min\{6, y-1\}$$

That is for example,

$$y = 3: \qquad 1 \leq y_1 \leq 2$$

$$y = 7: \qquad 1 \leq y_1 \leq 6$$

$$y = 10: \qquad 4 \leq y_1 \leq 6$$

and so on.

Sums of random variables (cont.)

Therefore

$$P(Y=y) = \sum_{y_1=\max\{1,y-6\}}^{\min\{6,y-1\}} P(Y_1=y_1)P(Y_2=y-y_1)$$

$$= \sum_{y_1=\max\{1,y-6\}}^{\min\{6,y-1\}} \frac{1}{6}\times\frac{1}{6}$$

$$= \frac{\min\{6,y-1\}-\max\{1,y-6\}+1}{36}$$

Sums of random variables (cont.)

The crucial components of this calculation are

- the partition of the event $(Y = y)$ according to the possible values of Y_1;
- the theorem of total probability;
- the independence of the outcomes of the two rolls, equivalently, the independence of the random variables Y_1 and Y_2.

Using these elements, we can derive a general result in the discrete case.

Sums of random variables (cont.)

Suppose Y_1 and Y_2 are discrete random variables, with Y_1 and Y_2 independent. Suppose that Y_1 and Y_2 take values on the sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ respectively. Let $Y = Y_1 + Y_2$. Then

$$\begin{aligned} P(Y = y) &= P\left(\bigcup_{y_1 \in \mathcal{Y}_1} (Y = y) \cap (Y_1 = y_1) \right) \\ &= \sum_{y_1 \in \mathcal{Y}_1} P((Y = y) \cap (Y_1 = y_1)) \\ &= \sum_{y_1 \in \mathcal{Y}_1} P((Y_2 = y - y_1) \cap (Y_1 = y_1)) \\ &= \sum_{y_1 \in \mathcal{Y}_1} P(Y_2 = y - y_1) P(Y_1 = y_1) \end{aligned}$$

Sums of random variables (cont.)

That is

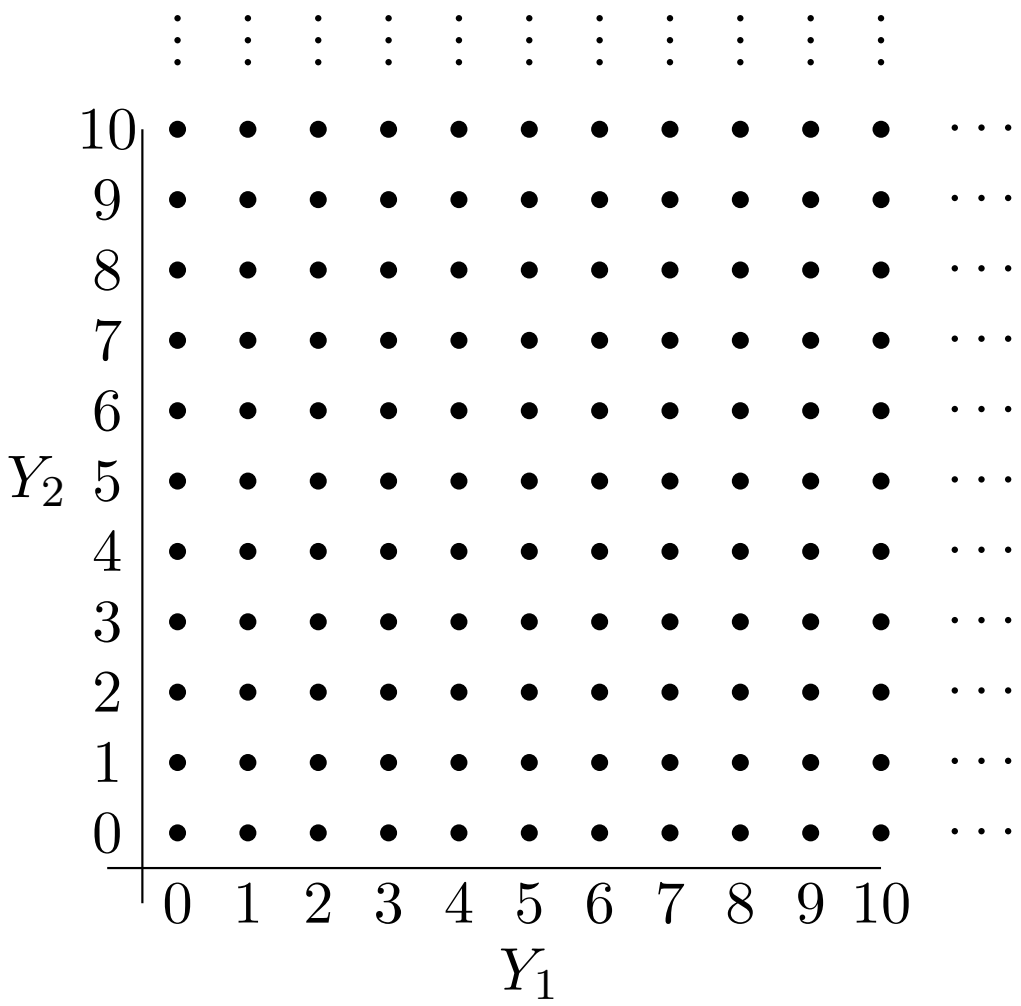
$$P(Y = y) = \sum_{y_1 \in \mathcal{Y}_1^*} P(Y_2 = y - y_1) P(Y_1 = y_1)$$

where

$$\mathcal{Y}_1^* = \{y_1 : \ y_1 \in \mathcal{Y}_1 \text{ and } y - y_1 \in \mathcal{Y}_2\}$$

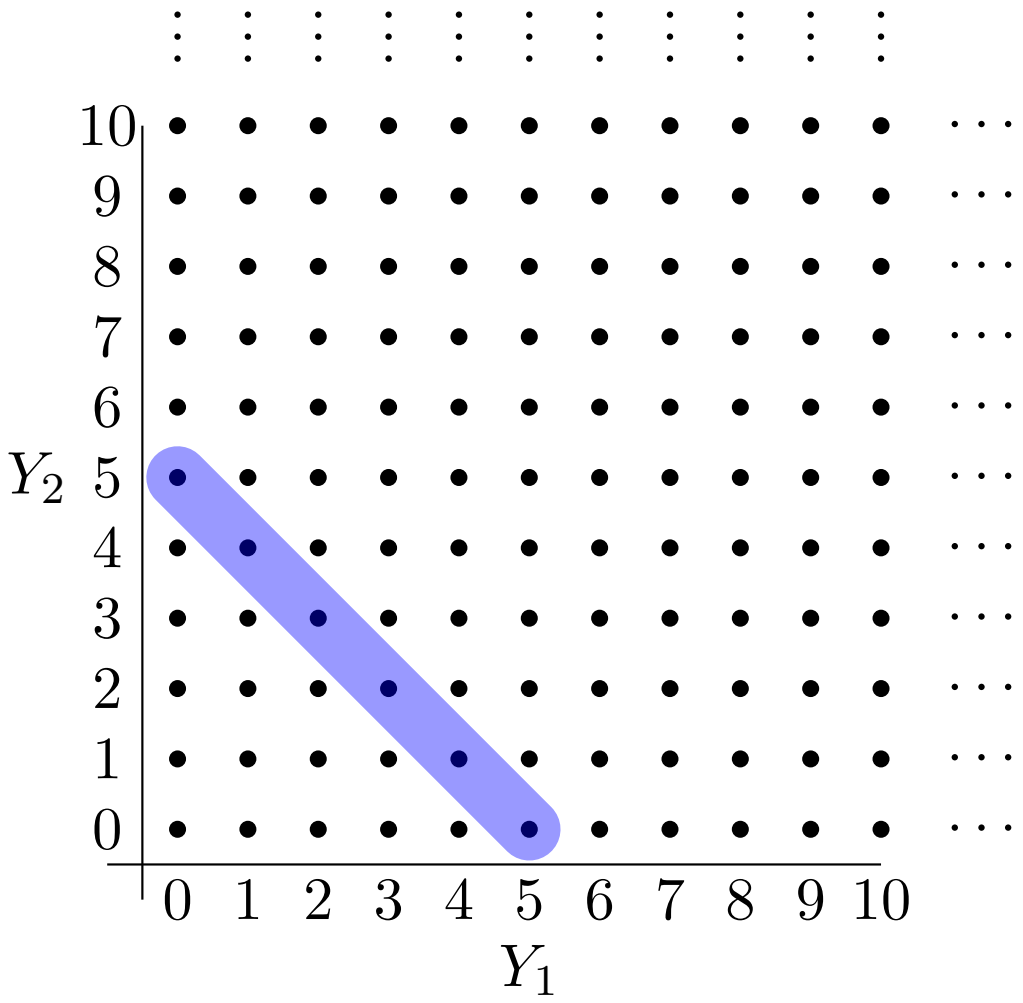
Sums of random variables (cont.)

Y_1, Y_2 take values on $\{0, 1, 2, \ldots\}$.



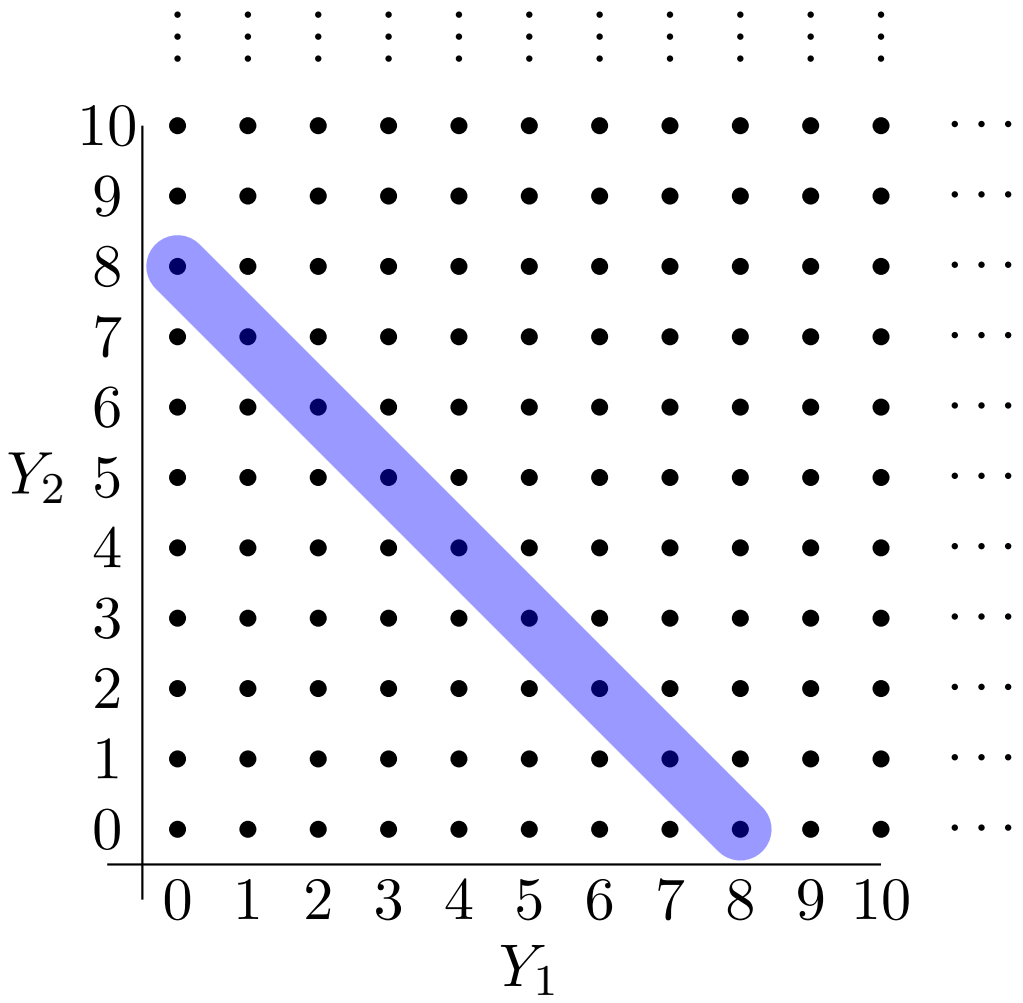
Sums of random variables (cont.)

$Y_1 + Y_2 = 5$



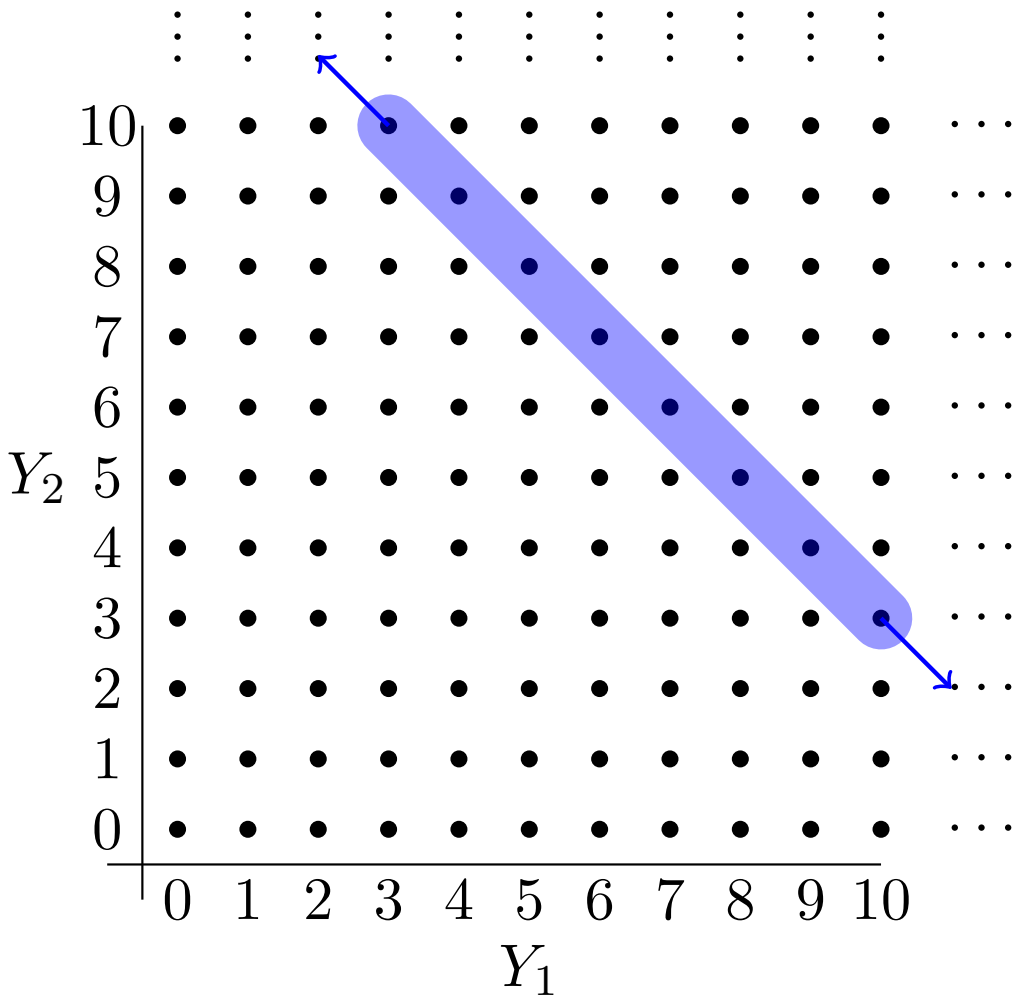
Sums of random variables (cont.)

$Y_1 + Y_2 = 8$



Sums of random variables (cont.)

$Y_1 + Y_2 = 13$



Sums of random variables (cont.)

In this example, for $y = 0, 1, 2, \ldots$

$$P(Y = y) = \sum_{y_1=0}^{y} P(Y_1 = y_1) P(Y_2 = y - y_1)$$

Sums of random variables (cont.)

If Y_1 and Y_2 are independent Poisson random variables with parameters λ_1 and λ_2, then if $Y = Y_1 + Y_2$,

$$P(Y = y) = \sum_{y_1=0}^{y} P(Y_1 = y_1)P(Y_2 = y - y_1)$$

$$= \sum_{y_1=0}^{y} \frac{\lambda_1^{y_1} e^{-\lambda_1}}{y_1!} \frac{\lambda_2^{y-y_1} e^{-\lambda_2}}{(y-y_1)!}$$

$$= e^{-(\lambda_1+\lambda_2)} \lambda_2^y \sum_{y_1=0}^{y} \left(\frac{\lambda_1}{\lambda_2}\right)^{y_1} \frac{1}{y_1!(y-y_1)!}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)} \lambda_2^y}{y!} \left(1 + \frac{\lambda_1}{\lambda_2}\right)^y$$

Sums of random variables (cont.)

But

$$\lambda_2^y \left(1 + \frac{\lambda_1}{\lambda_2}\right)^y = (\lambda_1 + \lambda_2)^y$$

so therefore

$$P(Y = y) = \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1 + \lambda_2)^y}{y!} \qquad y = 0, 1, 2, \ldots$$

that is, Y has a Poisson distribution with parameter $\lambda_1 + \lambda_2$.

Sums of random variables (cont.)

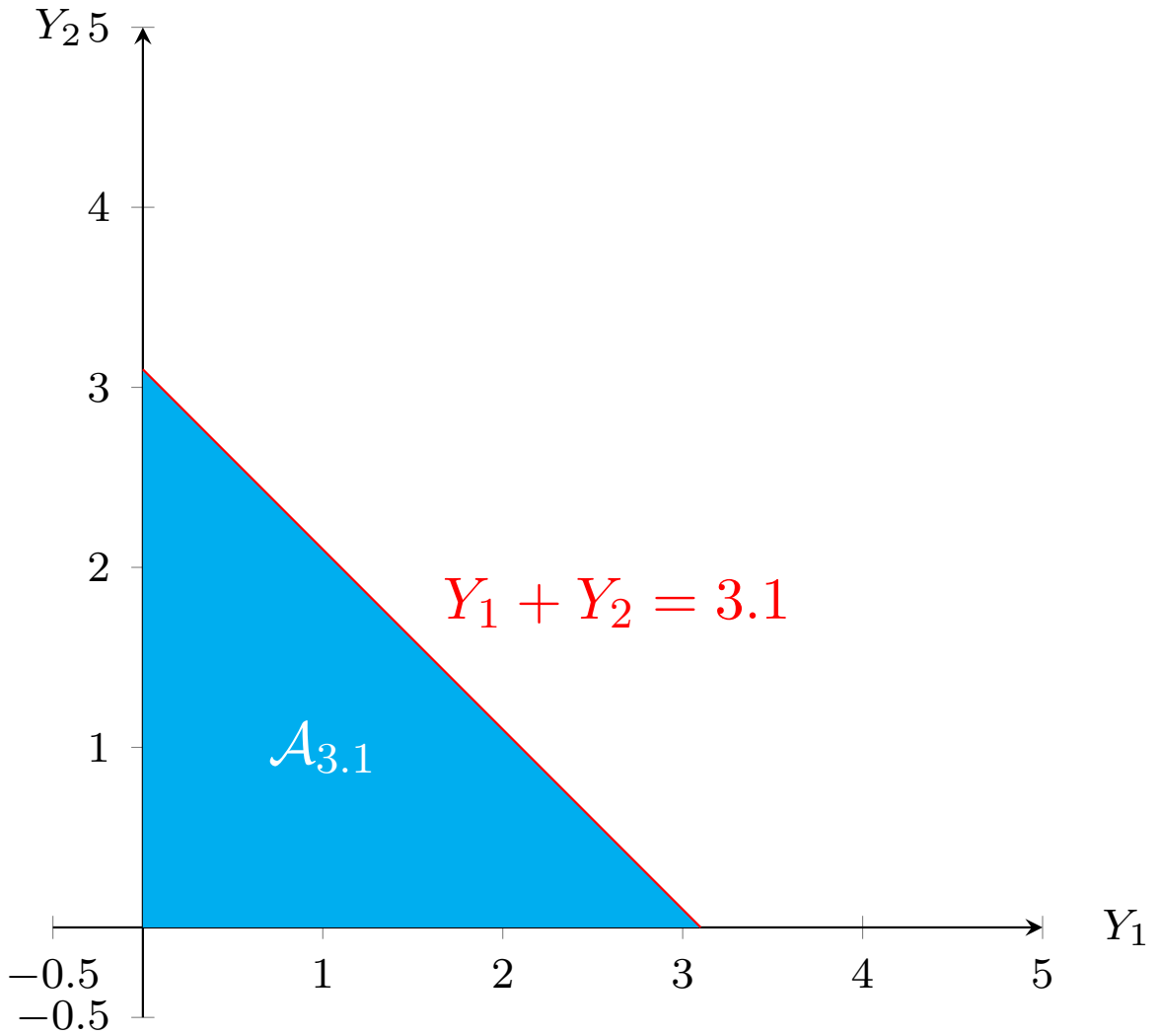
If Y_1 and Y_2 are independent continuous random variables, a similar strategy can be used. As before for continuous random variables, we start with the cdf and aim to compute

$$F_Y(y) = P(Y \leq y)$$

If Y_1 and Y_2 are continuous, then it follows that Y is continuous.

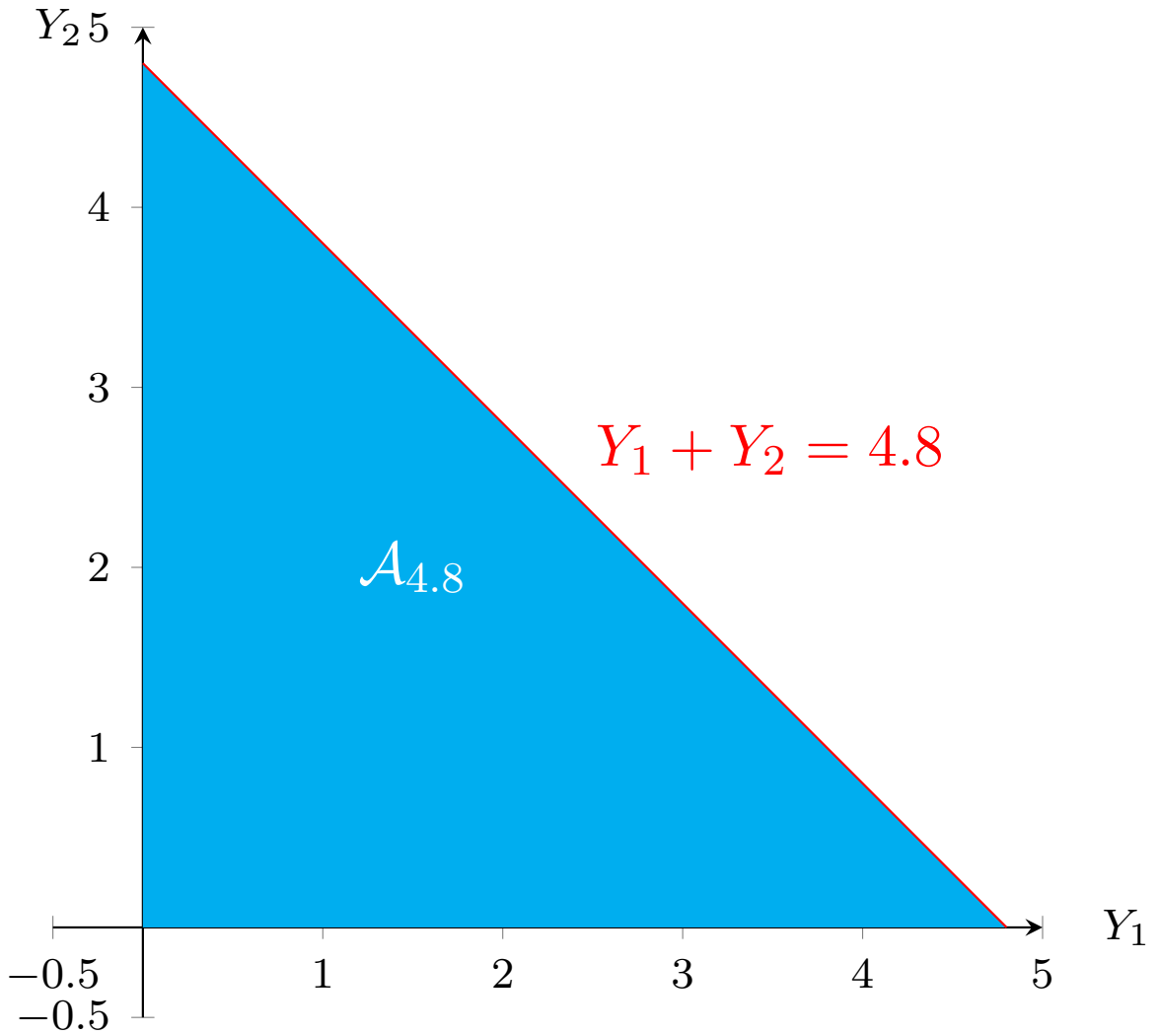
Sums of random variables (cont.)

Y_1, Y_2 taking values on $\mathbb{R}^+$; region $\mathcal{A}_{3.1}$ is where $Y_1 + Y_2 \leq 3.1$.



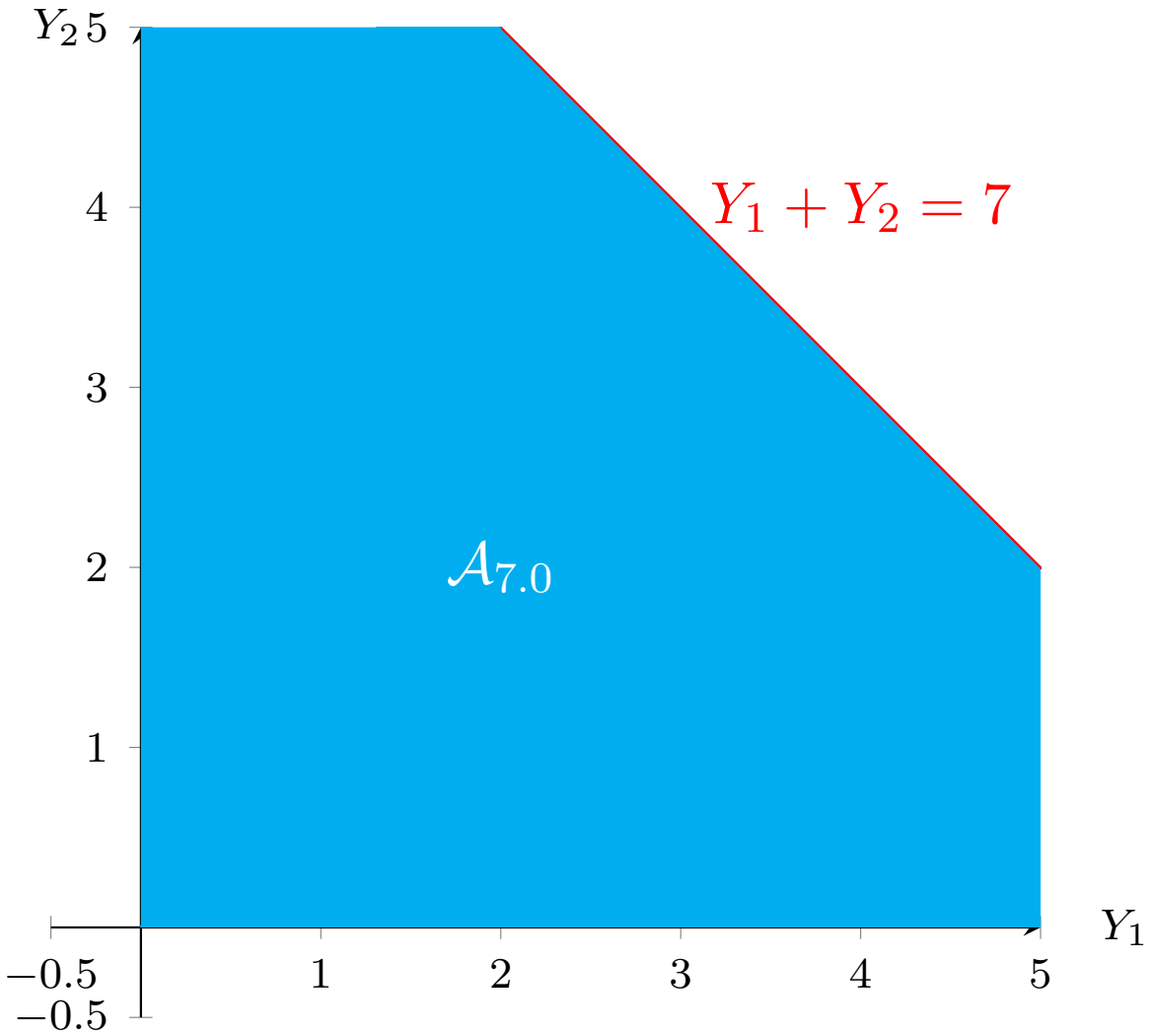
Sums of random variables (cont.)

Y_1, Y_2 taking values on $\mathbb{R}^+$; region $\mathcal{A}_{4.8}$ is where $Y_1 + Y_2 \leq 4.8$.



Sums of random variables (cont.)

Y_1, Y_2 taking values on $\mathbb{R}^+$; region $\mathcal{A}_{7.0}$ is where $Y_1 + Y_2 \leq 7$.



Sums of random variables (cont.)

Then we have by definition that

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(t)\, dt$$

where $f_Y(y)$ is the pdf for Y. However,

$$(Y \leq y) = (Y_1 + Y_2 \leq y)$$

Therefore to compute $F_Y(y) = P(Y \leq y)$ we need to aggregate all the probability across the blue shaded region, denoted $\mathcal{A}_y$, in the above pictures.

Sums of random variables (cont.)

In the discrete case in which

$$p_Y(y) = P(Y = y) = \sum_{y_1 \in \mathcal{Y}_1^*} P(Y_2 = y - y_1) P(Y_1 = y_1)$$

$$= \sum_{y_1 \in \mathcal{Y}_1^*} p_{Y_2}(y - y_1) p_{Y_1}(y_1)$$

where

- $p_Y(.)$ is the pmf for Y,
- $p_{Y_1}(.)$ is the pmf for Y_1,
- $p_{Y_2}(.)$ is the pmf for Y_2.

Sums of random variables (cont.)

we may write

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y_2}(y - y_1) f_{Y_1}(y_1)\, dy_1$$

where

- $f_Y(.)$ is the pdf for Y,
- $f_{Y_1}(.)$ is the pdf for Y_1,
- $f_{Y_2}(.)$ is the pdf for Y_2.

Sums of random variables (cont.)

Note

This is termed a *convolution* formula.

Example (Exponential random variables)

Suppose

$$Y_1 : f_{Y_1}(y_1) = \frac{1}{\beta_1} e^{-y_1/\beta_1} \qquad y_1 \geq 0$$

$$Y_2 : f_{Y_2}(y_2) = \frac{1}{\beta_2} e^{-y_2/\beta_2} \qquad y_2 \geq 0$$

are independent random variables, and let $Y = Y_1 + Y_2$. For convenience, write

$$\lambda_1 = \frac{1}{\beta_1} \qquad \lambda_2 = \frac{1}{\beta_2}.$$

Example (Exponential random variables)

Then from the convolution formula, for $y \geq 0$,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{Y_2}(y - y_1) f_{Y_1}(y_1)\, dy_1 \\ &= \int_0^y \lambda_2 e^{-\lambda_2 (y - y_1)} \lambda_1 e^{-\lambda_1 y_1}\, dy_1 \\ &= \lambda_1 \lambda_2 e^{-\lambda_2 y} \int_0^y e^{-(\lambda_1 - \lambda_2) y_1}\, dy_1 \end{aligned}$$

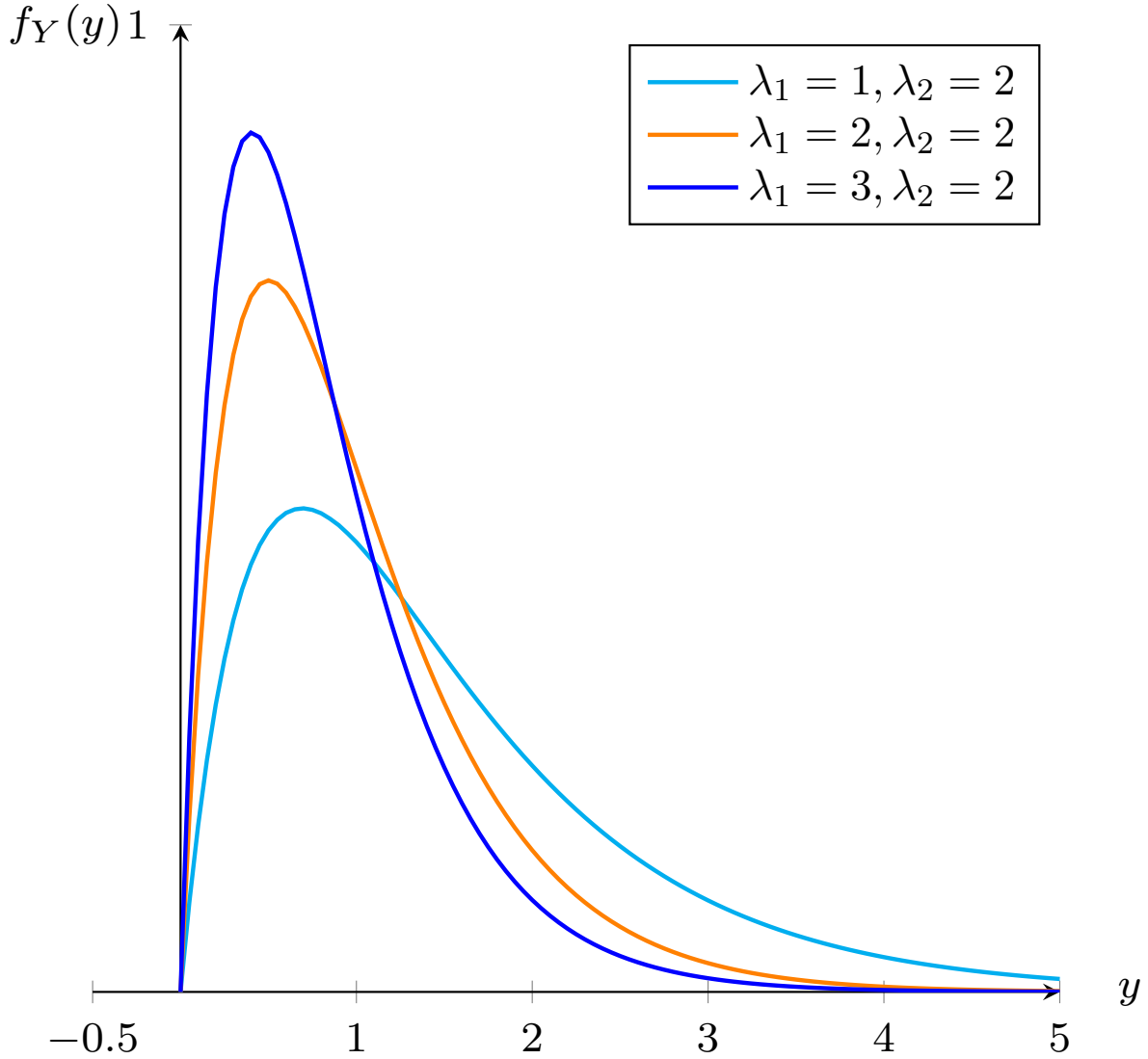
Example (Exponential random variables)

We then have two cases to consider:

$$f_Y(y) = \begin{cases} \dfrac{\lambda_1\lambda_2}{\lambda_1 - \lambda_2} e^{-\lambda_2 y}(1 - e^{-(\lambda_1-\lambda_2)y}) & \lambda_1 \neq \lambda_2 \\ \lambda^2 y e^{-\lambda y} & \lambda_1 = \lambda_2 = \lambda, \text{ say.} \end{cases}$$

In the second case, we see that in fact Y has a gamma distribution with parameters $\alpha = 2$ and $\beta = 1/\lambda$.

Sums of random variables (cont.)



Sums of random variables (cont.)

Note

- We can extend the result to three or more independent random variables

$$Y = Y_1 + Y_2 + Y_3$$

by using the convolution formula recursively; we have

$$Y = (Y_1 + Y_2) + Y_3 = Y_{12} + Y_3$$

say, and so we would compute the pmf or pdf of $Y_{12} = Y_1 + Y_2$ first, and then convolve it with the pmf or pdf for Y_3.

- We can generalize to collections of random variables that are not independent.

Chapter 4: Multivariate distributions

Multivariate probability distributions

We now consider probability specifications for more than one variable at a time.

- We have already used the straightforward concept of *independent* random variables, that is, variables associated with independent events;
- for example, the random variables associated with the outcomes of two independent rolls of dice.

Multivariate probability distributions (cont.)

However, there are many instances where we wish to consider random variables where there is an absence of independence;

- for example where the outcome of one part of the experiment depends on the outcome of the other.

In general we can consider a *vector* of random variables (or *vector random variable*)

$$(Y_1, Y_2, \ldots, Y_n).$$

Multivariate probability distributions (cont.)

Note

The critical thing to remember is that when we distribute probability amongst the possible values of the random variables, the total probability we assign must still equal one.

Discrete bivariate distributions

We consider two variables Y_1 and Y_2 that are both discrete. We can suppose that both variables take values on the integers, $\mathbb{Z}$.

A *discrete bivariate probability mass function* is a function of two arguments

$$p(y_1, y_2)$$

that distributes probability across the possible values of the vector (Y_1, Y_2) so that

$$p(y_1, y_2) = P((Y_1 = y_1) \cap (Y_2 = y_2))$$

for $-\infty < y_1 < \infty$ and $-\infty < y_2 < \infty$.

Discrete bivariate distributions (cont.)

For shorthand, we write

$$P((Y_1 = y_1) \cap (Y_2 = y_2)) \equiv P(Y_1 = y_1, Y_2 = y_2)$$

Discrete bivariate distributions (cont.)

We can think of $p(y_1, y_2)$ as specifying the values in a probability table.

		Y_2		
		1	2	3
	1	$p(1,1)$	$p(1,2)$	$p(1,3)$
	2	$p(2,1)$	$p(2,2)$	$p(2,3)$
Y_1	3	$p(3,1)$	$p(3,2)$	$p(3,3)$
	4	$p(4,1)$	$p(4,2)$	$p(4,3)$
	5	$p(5,1)$	$p(5,2)$	$p(5,3)$

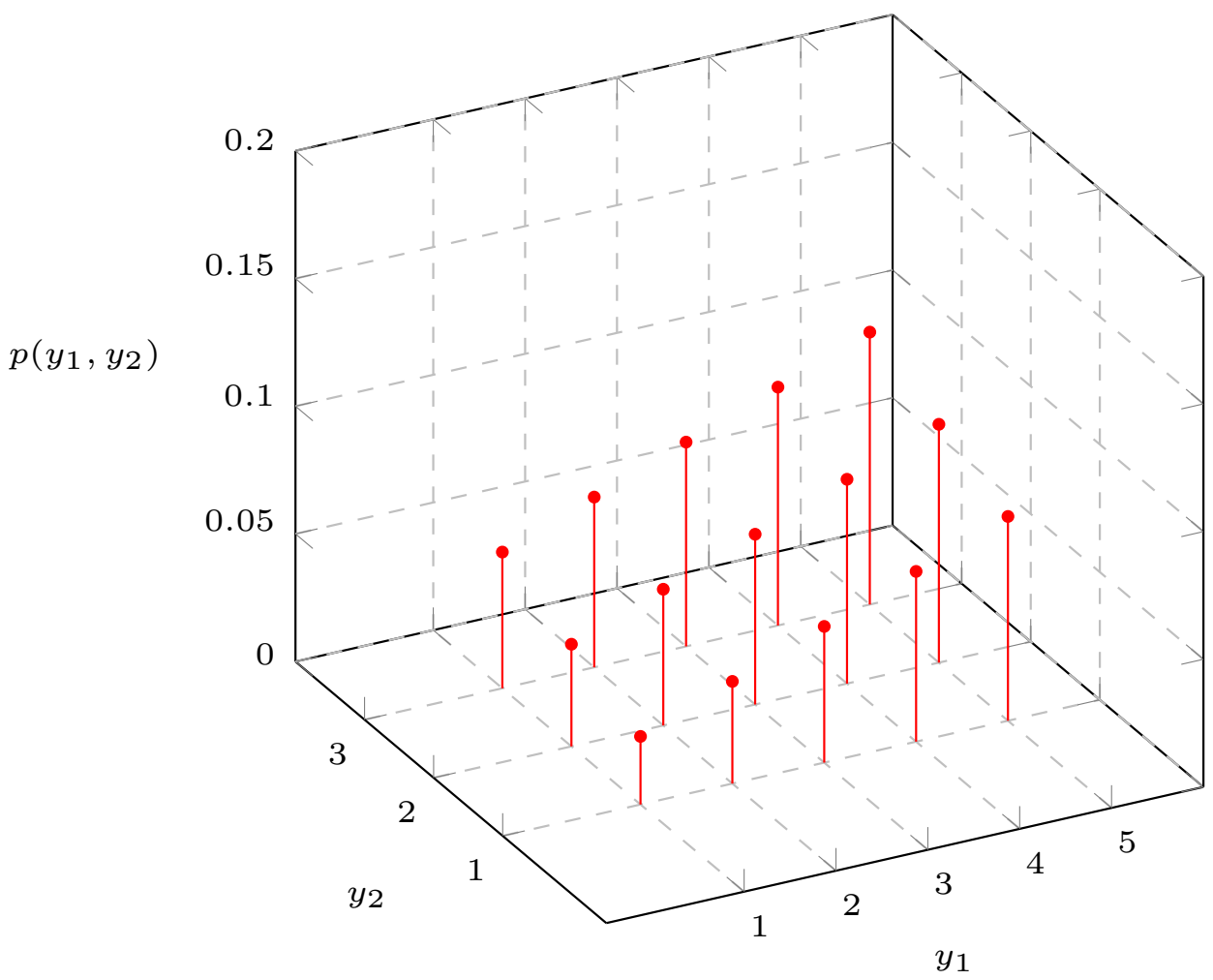
Discrete bivariate distributions (cont.)

For $1 \leq y_1 \leq 5, 1 \leq y_2 \leq 3$

$$p(y_1, y_2) = \frac{(y_1 + y_2)}{75}$$

		Y_2		
		1	2	3
	1	2/75	3/75	4/75
	2	3/75	4/75	5/75
Y_1	3	4/75	5/75	6/75
	4	5/75	6/75	7/75
	5	6/75	7/75	8/75

Discrete bivariate distributions (cont.)



Discrete bivariate distributions (cont.)

The function $p(y_1, y_2)$ is the *joint probability mass function*: it has two basic properties

- "specifies probabilities"

$$0 \leq p(y_1, y_2) \leq 1 \quad \text{for all } y_1, y_2$$

- "sums to one"

$$\sum_{y_1=-\infty}^{\infty} \sum_{y_2=-\infty}^{\infty} p(y_1, y_2) = 1.$$

although $p(y_1, y_2)$ may be zero for some arguments.

Discrete bivariate distributions (cont.)

In the above example,

$$\sum_{y_1=1}^{5}\sum_{y_2=1}^{3}\frac{(y_1+y_2)}{75} = \frac{1}{75}\sum_{y_1=1}^{5}\sum_{y_2=1}^{3}(y_1+y_2)$$

$$= \frac{1}{75}\left[3\sum_{y_1=1}^{5}y_1 + 5\sum_{y_2=1}^{3}y_2\right]$$

$$= \frac{1}{75}\left[3\frac{5\times 6}{2} + 5\frac{3\times 4}{2}\right]$$

$$= \frac{1}{75}\left[45+30\right]$$

$$= 1.$$

Discrete bivariate distributions (cont.)

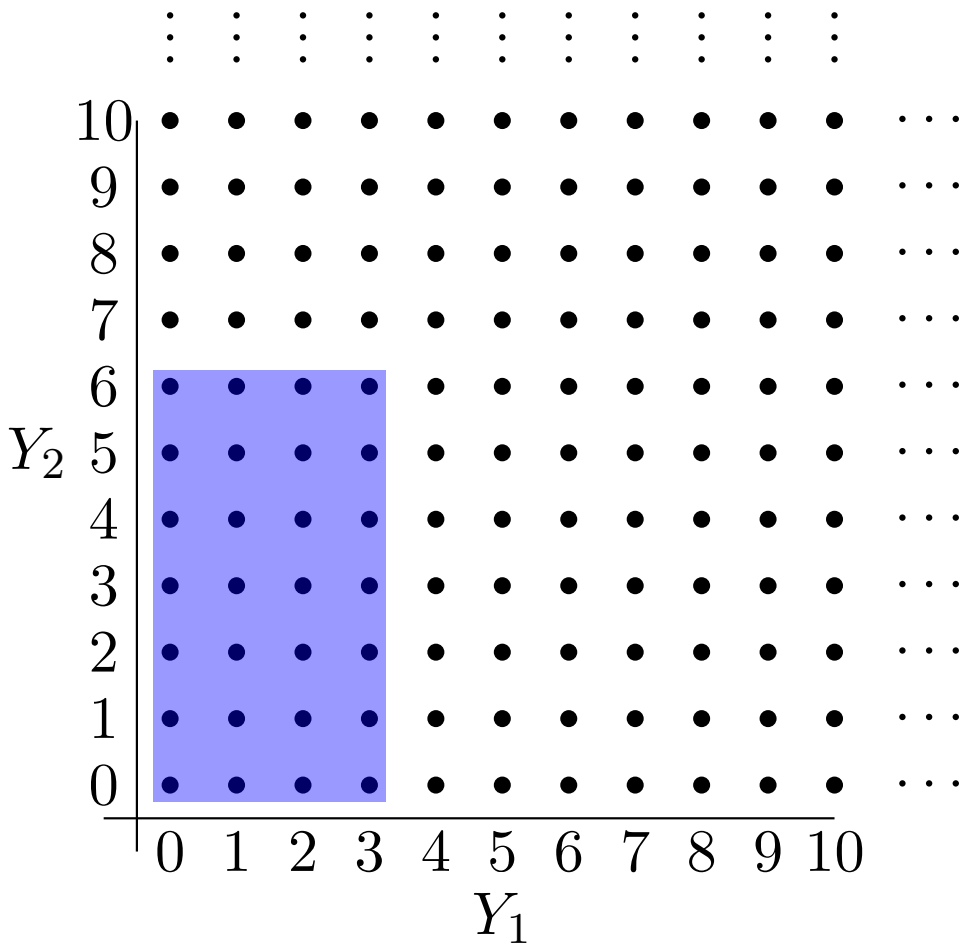
We define the *joint cumulative distribution function* $F(y_1, y_2)$ by

$$F(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2) = \sum_{t_1=-\infty}^{y_1} \sum_{t_2=-\infty}^{y_2} p(t_1, t_2)$$

that is, by accumulating the probabilities in the joint pmf over a range of values up to and including (y_1, y_2)

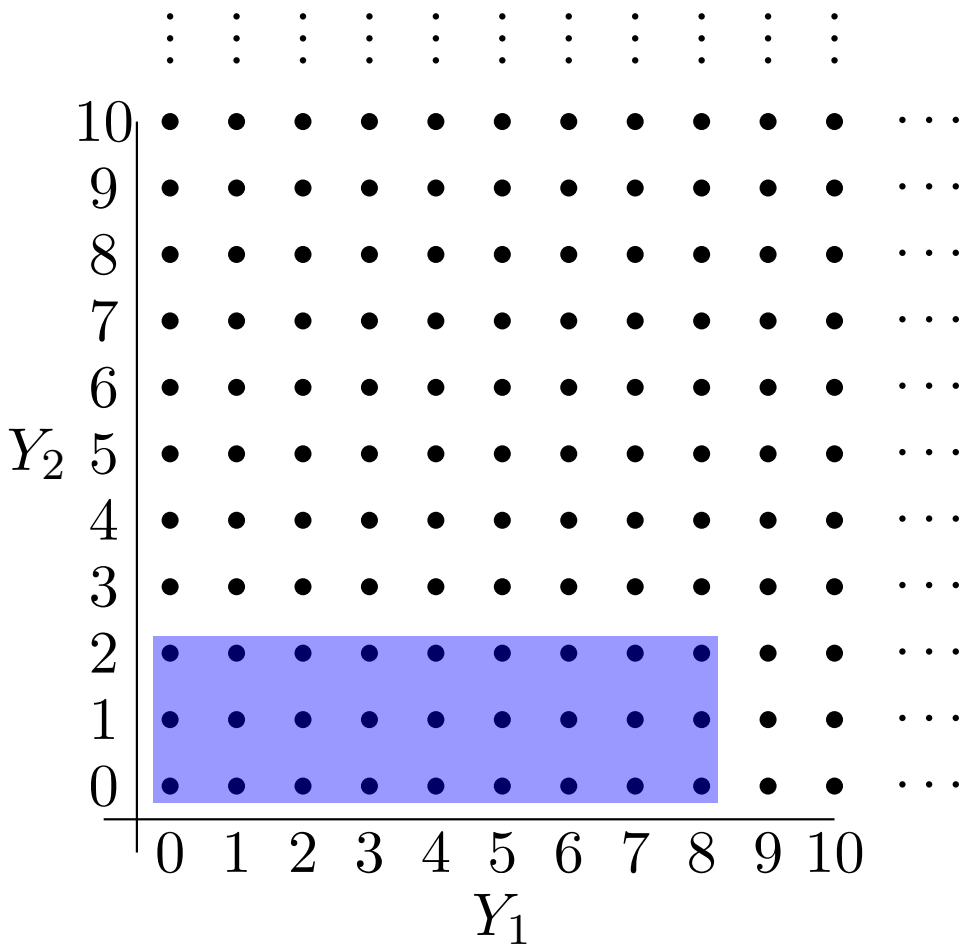
Discrete bivariate distributions (cont.)

$$F(3,6) = P(Y_1 \leq 3, Y_2 \leq 6) = \sum_{t_1=0}^{3} \sum_{t_2=0}^{6} p(t_1, t_2)$$



Discrete bivariate distributions (cont.)

$$F(8,2) = P(Y_1 \leq 8, Y_2 \leq 2) = \sum_{t_1=0}^{8} \sum_{t_2=0}^{2} p(t_1, t_2)$$



Discrete bivariate distributions (cont.)

Example

A bag contains ten balls:

- five red;
- three yellow;
- two white;

Four balls are selected, with all such selections being equally likely.

Discrete bivariate distributions (cont.)

Example

Let

- Y_1 denote the number of red balls selected;
- Y_2 denote the number of yellow balls selected.

Then using combinatorial arguments, we see that the joint pmf of Y_1 and Y_2 is given by

$$p_{Y_1,Y_2}(y_1, y_2) = \frac{\binom{5}{y_1}\binom{3}{y_2}\binom{2}{4-y_1-y_2}}{\binom{10}{4}}$$

for (y_1, y_2) such that the combinatorial terms are defined, and zero when the terms are not.

Discrete bivariate distributions (cont.)

Example

We need (y_1, y_2) simultaneously to satisfy

$$0 \leq y_1 \leq 5 \qquad 0 \leq y_2 \leq 3 \qquad 0 \leq 4 - y_1 - y_2 \leq 2$$

in order to have a non-zero probability.

Total number of selections: $\binom{10}{4} = 210$.

Discrete bivariate distributions (cont.)

Example

		Y_2			
		0	1	2	3
Y_1	0	0	0		
	1	0			
	2				0
	3			0	0
	4		0	0	0
	5	0	0	0	0

Red	Yellow	White
0	2	2
0	3	1
1	1	2
1	2	1
1	3	0
2	0	2
2	1	1
2	2	0
3	0	1
3	1	0
4	0	0

Discrete bivariate distributions (cont.)

Example

Red	Yellow	White	Count
0	2	2	3
0	3	1	2
1	1	2	15
1	2	1	30
1	3	0	5
2	0	2	10
2	1	1	60
2	2	0	30
3	0	1	20
3	1	0	30
4	0	0	5

Discrete bivariate distributions (cont.)

Example

		Y_2			
		0	1	2	3
	0	0.0000	0.0000	0.0143	0.0095
	1	0.0000	0.0714	0.1429	0.0238
	2	0.0476	0.2857	0.1429	0.0000
Y_1	3	0.0952	0.1429	0.0000	0.0000
	4	0.0238	0.0000	0.0000	0.0000
	5	0.0000	0.0000	0.0000	0.0000

Discrete bivariate distributions (cont.)

For more than two variables, we can define a joint pmf in the same way:

$$p(y_1, y_2, \ldots, y_n) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$$

to represent an n-way table.

The marginal mass function

Once we have specified a joint pmf, we can easily compute the probability distribution for one of the variables alone.

The *marginal probability mass function* for Y_1 is merely the pmf of Y_1, which is computed via the Theorem of Total Probability.

The marginal mass function (cont.)

Suppose that the joint pmf for Y_1 and Y_2 is denoted $p_{Y_1,Y_2}(.,.)$. The the marginal pmf for Y_1, $p_{Y_1}(.)$ is given by

$$p_{Y_1}(y_1) = P(Y_1 = y_1) = \sum_{y_2=-\infty}^{\infty} P(Y_1 = y_1, Y_2 = y_2)$$

that is

$$p_{Y_1}(y_1) = \sum_{y_2=-\infty}^{\infty} p_{Y_1,Y_2}(y_1, y_2)$$

The marginal mass function (cont.)

This result uses a partitioning argument:

$$(Y_1 = y_1) = \bigcup_{y_2=-\infty}^{\infty} (Y_1 = y_1) \cap (Y_2 = y_2)$$

The marginal mass function (cont.)

For example

$$P(Y_1 = 2) = P(Y_1 = 2, Y_2 = 1) + P(Y_1 = 2, Y_2 = 2) + P(Y_1 = 2, Y_2 = 3).$$

The marginal mass function (cont.)

If $p_{Y_1,Y_2}(y_1, y_2)$ specifies the values in a probability table, we compute the marginal pmf

- for Y_1 by summing across the rows of the table;
- for Y_2 by summing down the columns of the table.

		Y_2			
		1	2	3	$p_{Y_1}(.)$
Y_1	1	$p_{Y_1,Y_2}(1,1)$	$p_{Y_1,Y_2}(1,2)$	$p_{Y_1,Y_2}(1,3)$	$p_{Y_1}(1)$
	2	$p_{Y_1,Y_2}(2,1)$	$p_{Y_1,Y_2}(2,2)$	$p_{Y_1,Y_2}(2,3)$	$p_{Y_1}(2)$
	3	$p_{Y_1,Y_2}(3,1)$	$p_{Y_1,Y_2}(3,2)$	$p_{Y_1,Y_2}(3,3)$	$p_{Y_1}(3)$
	4	$p_{Y_1,Y_2}(4,1)$	$p_{Y_1,Y_2}(4,2)$	$p_{Y_1,Y_2}(4,3)$	$p_{Y_1}(4)$
	5	$p_{Y_1,Y_2}(5,1)$	$p_{Y_1,Y_2}(5,2)$	$p_{Y_1,Y_2}(5,3)$	$p_{Y_1}(5)$
	$p_{Y_2}(.)$	$p_{Y_2}(1)$	$p_{Y_2}(2)$	$p_{Y_2}(3)$	1

The marginal mass function (cont.)

Example (Previous example)

Four balls selected from 10.

- Y_1 denote the number of red balls selected;
- Y_2 denote the number of yellow balls selected.

The joint pmf of Y_1 and Y_2 is given by

$$p_{Y_1,Y_2}(y_1, y_2) = \frac{\binom{5}{y_1}\binom{3}{y_2}\binom{2}{4-y_1-y_2}}{\binom{10}{4}}$$

for (y_1, y_2) such that the combinatorial terms are defined, and zero when the terms are not.

The marginal mass function (cont.)

Example (Previous example)

We can compute the marginal pmf for Y_1 by summing probabilities in the joint probability table.

$$p_{Y_1}(0) = \frac{\binom{5}{0}\binom{3}{2}\binom{2}{2} + \binom{5}{0}\binom{3}{3}\binom{2}{1}}{\binom{10}{4}} = \frac{3+2}{210} = \frac{5}{210}$$

The marginal mass function (cont.)

Example (Previous example)

$$p_{Y_1}(1) = \frac{\binom{5}{1}\binom{3}{1}\binom{2}{2} + \binom{5}{1}\binom{3}{2}\binom{2}{1} + \binom{5}{1}\binom{3}{3}\binom{2}{0}}{\binom{10}{4}}$$

$$= \frac{15 + 30 + 5}{210} = \frac{50}{210}$$

The marginal mass function (cont.)

Example (Previous example)

$$p_{Y_1}(2) = \frac{\binom{5}{2}\binom{3}{0}\binom{2}{2} + \binom{5}{2}\binom{3}{1}\binom{2}{1} + \binom{5}{2}\binom{3}{0}\binom{2}{2}}{\binom{10}{4}}$$

$$= \frac{10 + 60 + 30}{210} = \frac{100}{210}$$

The marginal mass function (cont.)

Example (Previous example)

$$p_{Y_1}(3) = \frac{\binom{5}{3}\binom{3}{0}\binom{2}{1} + \binom{5}{3}\binom{3}{1}\binom{2}{0}}{\binom{10}{4}}$$

$$= \frac{20 + 30}{210} = \frac{50}{210}$$

The marginal mass function (cont.)

Example (Previous example)

$$p_{Y_1}(4) = \frac{\binom{5}{4}\binom{3}{0}\binom{2}{0}}{\binom{10}{4}} = \frac{5}{210}$$

The marginal mass function (cont.)

Note

In this example we can compute $p_{Y_1}(.)$ directly using the hypergeometric formula

$$p_{Y_1}(y_1) = \frac{\binom{5}{y_1}\binom{5}{4-y_1}}{\binom{10}{4}}$$

for $0 \leq y_1 \leq 5$ and $0 \leq 4 - y_1 \leq 5$.

The marginal mass function (cont.)

Note

y_1	Numerator
0	$\binom{5}{0}\binom{5}{4} = 1 \times 5 = 5$
1	$\binom{5}{1}\binom{5}{2} = 5 \times 10 = 50$
2	$\binom{5}{2}\binom{5}{3} = 10 \times 10 = 100$
3	$\binom{5}{3}\binom{5}{1} = 10 \times 5 = 50$
4	$\binom{5}{4}\binom{5}{0} = 5 \times 1 = 5$

The marginal mass function (cont.)

Note

The marginal pmfs $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$ have all the properties of single variable pmfs; specifically

$$\sum_{y_1=-\infty}^{\infty} p_{Y_1}(y_1) = \sum_{y_1=-\infty}^{\infty} \sum_{y_2=-\infty}^{\infty} p_{Y_1,Y_2}(y_1, y_2) = 1.$$

The conditional mass function

Once we have the joint pmf

$$p_{Y_1,Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2)$$

and the marginal pmf

$$p_{Y_1}(y_1) = P(Y_1 = y_1)$$

we can consider *conditional* pmfs.

The conditional mass function (cont.)

We have that if $P(Y_1 = y_1) > 0$, then the conditional probability that $Y_2 = y_2$, given that $Y_1 = y_1$, is

$$P(Y_2 = y_2 | Y_1 = y_1) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_1 = y_1)}$$

For a **fixed** value of y_1, we can consider how this conditional probability varies as argument y_2 varies.

The conditional mass function (cont.)

The *conditional probability mass function* for Y_2, given that $Y_1 = y_1$, is denoted

$$p_{Y_2|Y_1}(y_2|y_1)$$

and defined by

$$p_{Y_2|Y_1}(y_2|y_1) = P(Y_2 = y_2|Y_1 = y_1)$$

whenever $P(Y_1 = y_1) > 0$.

The conditional mass function (cont.)

The conditional pmfs are obtained by taking 'slices' through the joint pmf, and then standardizing the slice so that the probabilities sum to one.

Recall that

$$p_{Y_2|Y_1}(y_2|y_1) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_1 = y_1)} = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_1}(y_1)}$$

$$= \frac{p_{Y_1,Y_2}(y_1, y_2)}{\sum_{t_2=-\infty}^{\infty} p_{Y_1,Y_2}(y_1, t_2)}.$$

The conditional mass function (cont.)

Given $Y_1 = 2$, we define the conditional pmf $p_{Y_2|Y_1}(y_2|2)$ by examining the second row of the table.

		Y_2			
		1	2	3	$p_{Y_1}(.)$
	1	$p_{Y_1,Y_2}(1,1)$	$p_{Y_1,Y_2}(1,2)$	$p_{Y_1,Y_2}(1,3)$	$p_{Y_1}(1)$
	2	$p_{Y_1,Y_2}(2,1)$	$p_{Y_1,Y_2}(2,2)$	$p_{Y_1,Y_2}(2,3)$	$p_{Y_1}(2)$
Y_1	3	$p_{Y_1,Y_2}(3,1)$	$p_{Y_1,Y_2}(3,2)$	$p_{Y_1,Y_2}(3,3)$	$p_{Y_1}(3)$
	4	$p_{Y_1,Y_2}(4,1)$	$p_{Y_1,Y_2}(4,2)$	$p_{Y_1,Y_2}(4,3)$	$p_{Y_1}(4)$
	5	$p_{Y_1,Y_2}(5,1)$	$p_{Y_1,Y_2}(5,2)$	$p_{Y_1,Y_2}(5,3)$	$p_{Y_1}(5)$
	$p_{Y_2}(.)$	$p_{Y_2}(1)$	$p_{Y_2}(2)$	$p_{Y_2}(3)$	1

The conditional mass function (cont.)

Given $Y_2 = 3$, we define the conditional pmf $p_{Y_1|Y_2}(y_1|3)$ by examining the third column of the table.

		Y_2			
		1	2	3	$p_{Y_1}(.)$
Y_1	1	$p_{Y_1,Y_2}(1,1)$	$p_{Y_1,Y_2}(1,2)$	$p_{Y_1,Y_2}(1,3)$	$p_{Y_1}(1)$
	2	$p_{Y_1,Y_2}(2,1)$	$p_{Y_1,Y_2}(2,2)$	$p_{Y_1,Y_2}(2,3)$	$p_{Y_1}(2)$
	3	$p_{Y_1,Y_2}(3,1)$	$p_{Y_1,Y_2}(3,2)$	$p_{Y_1,Y_2}(3,3)$	$p_{Y_1}(3)$
	4	$p_{Y_1,Y_2}(4,1)$	$p_{Y_1,Y_2}(4,2)$	$p_{Y_1,Y_2}(4,3)$	$p_{Y_1}(4)$
	5	$p_{Y_1,Y_2}(5,1)$	$p_{Y_1,Y_2}(5,2)$	$p_{Y_1,Y_2}(5,3)$	$p_{Y_1}(5)$
	$p_{Y_2}(.)$	$p_{Y_2}(1)$	$p_{Y_2}(2)$	$p_{Y_2}(3)$	1

The conditional mass function (cont.)

Note

We have the fundamental relationship

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1)p_{Y_2|Y_1}(y_2|y_1)$$

whenever $p_{Y_1}(y_1) > 0$.

Independence

Two discrete random variables Y_1 and Y_2 are called *independent* if

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1)p_{Y_2}(y_2)$$

for **all** values y_1, y_2. Equivalently, they are independent if

$$p_{Y_1|Y_2}(y_1|y_2) = p_{Y_1}(y_1) \qquad p_{Y_2|Y_1}(y_2|y_1) = p_{Y_2}(y_2)$$

for **all** values y_1, y_2 where the conditional pmfs are defined.

If Y_1 and Y_2 are not independent, they are *dependent*.

Independence (cont.)

This definition extends to multiple variables: $Y_1, \ldots, Y_n$ are independent if

$$p_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} p_{Y_i}(y_i)$$

for **all** vectors $(y_1, \ldots, y_n)$.

Note that we could make equivalent statements in terms of the joint cdf

$$F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = P\left(\bigcap_{i=1}^{n}(Y_i \leq y_i)\right) = \prod_{i=1}^{n} F_{Y_i}(y_i)$$

The joint density function

If Y_1 and Y_2 are two *continuous* random variables, then we can still consider statements of the form

$$P((Y_1 \leq y_1) \cap (Y_2 \leq y_2))$$

and hence define the *joint cumulative distribution function* cdf

$$F_{Y_1,Y_2}(y_1, y_2) = P((Y_1 \leq y_1) \cap (Y_2 \leq y_2))$$

for any pair of real numbers (y_1, y_2).

The joint density function (cont.)

The joint cdf has the following properties:

- "starts at zero"

$$\lim_{y_1 \longrightarrow -\infty} \lim_{y_2 \longrightarrow -\infty} F_{Y_1,Y_2}(y_1, y_2) = 0$$

- "ends at one"

$$\lim_{y_1 \longrightarrow \infty} \lim_{y_2 \longrightarrow \infty} F_{Y_1,Y_2}(y_1, y_2) = 1$$

- "non-decreasing in y_1 and y_2 in between"

$$F_{Y_1,Y_2}(y_1, y_2) \leq F_{Y_1,Y_2}(y_1 + c, y_2)$$

$$F_{Y_1,Y_2}(y_1, y_2) \leq F_{Y_1,Y_2}(y_1, y_2 + c)$$

for all y_1, y_2, and any $c > 0$.

The joint density function (cont.)

Furthermore, we have that

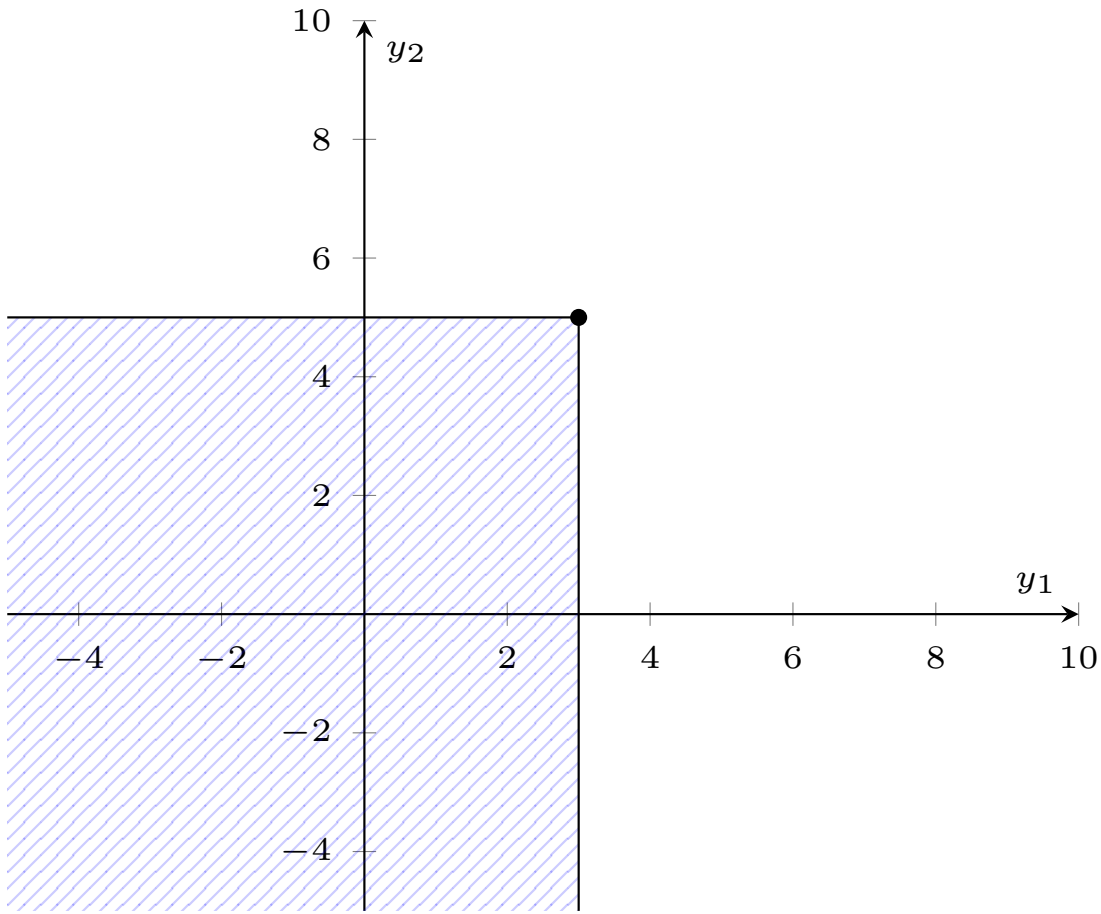
$$\begin{aligned}\lim_{y_1 \longrightarrow \infty} F_{Y_1,Y_2}(y_1, y_2) &= P(Y_1 < \infty, Y_2 \leq y_2) \\ &= P(Y_2 \leq y_2) \\ &= F_{Y_2}(y_2)\end{aligned}$$

and similarly

$$\lim_{y_2 \longrightarrow \infty} F_{Y_1,Y_2}(y_1, y_2) = F_{Y_1}(y_1).$$

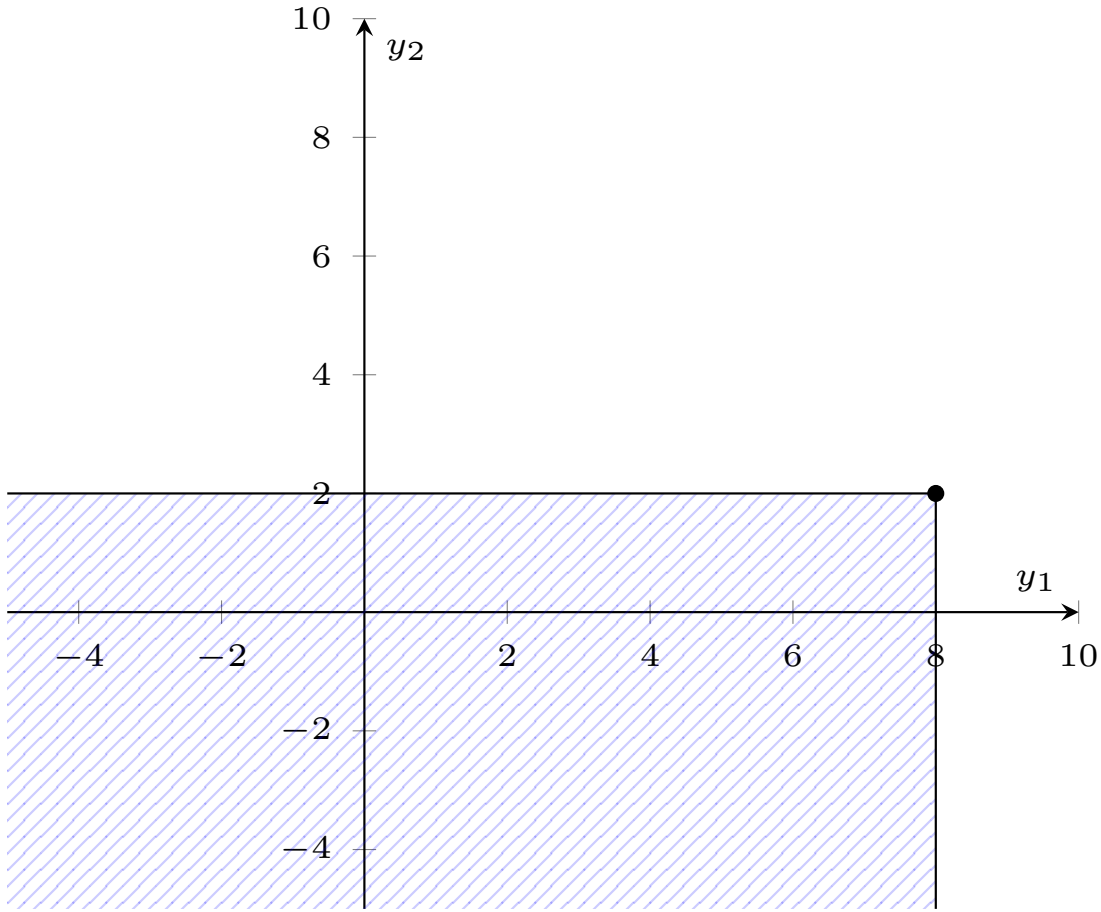
The joint density function (cont.)

$$F_{Y_1,Y_2}(3,5) = P(Y_1 \leq 3, Y_2 \leq 5)$$



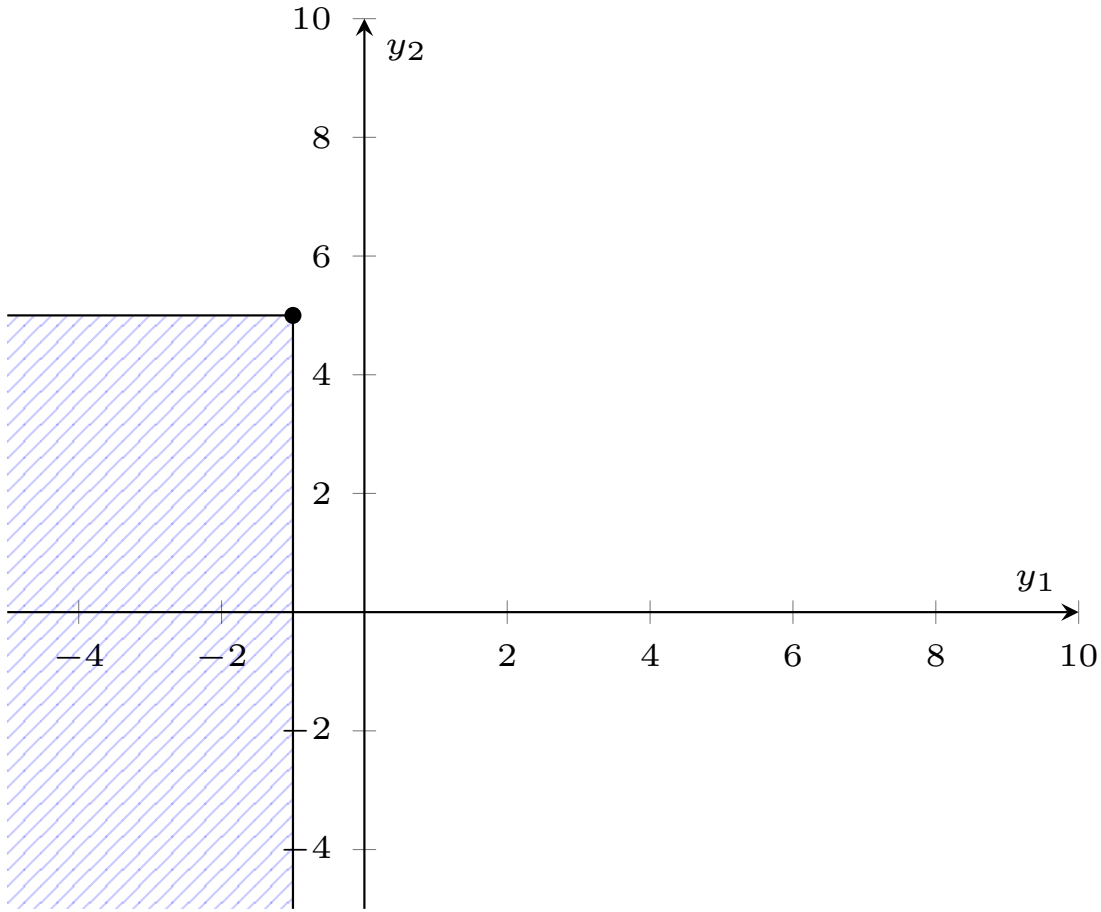
The joint density function (cont.)

$$F_{Y_1,Y_2}(8,2) = P(Y_1 \leq 8, Y_2 \leq 2)$$



The joint density function (cont.)

$$F_{Y_1,Y_2}(-1,5) = P(Y_1 \leq -1, Y_2 \leq 5)$$



The joint density function (cont.)

We accumulate probability over the shaded region, the rectangle

$$(-\infty, y_1] \times (-\infty, y_2],$$

to compute $F_{Y_1,Y_2}(y_1, y_2)$.

However, as in the single variable case, we must have

$$P(Y_1 = y_1, Y_2 = y_2) = 0$$

for all y_1 and y_2.

The joint density function (cont.)

As in the single variable case, we introduce the *joint probability density function* (joint pdf)

$$f_{Y_1,Y_2}(y_1, y_2)$$

to describe how probability is spread around the possible values, where

$$F_{Y_1,Y_2}(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f_{Y_1,Y_2}(t_1, t_2)\, dt_2 dt_1$$

that is, to compute $F_{Y_1,Y_2}(y_1, y_2)$ we integrate $f_{Y_1,Y_2}(y_1, y_2)$ over the rectangle

$$(-\infty, y_1] \times (-\infty, y_2].$$

The joint density function (cont.)

We can compute the double integral as follows: writing

$$\int_{-\infty}^{y_1} \left\{ \int_{-\infty}^{y_2} f_{Y_1,Y_2}(t_1, t_2)\, dt_2 \right\} dt_1$$

- fix t_1, and perform the first (inner) integration

$$\int_{-\infty}^{y_2} f_{Y_1,Y_2}(t_1, t_2)\, dt_2$$

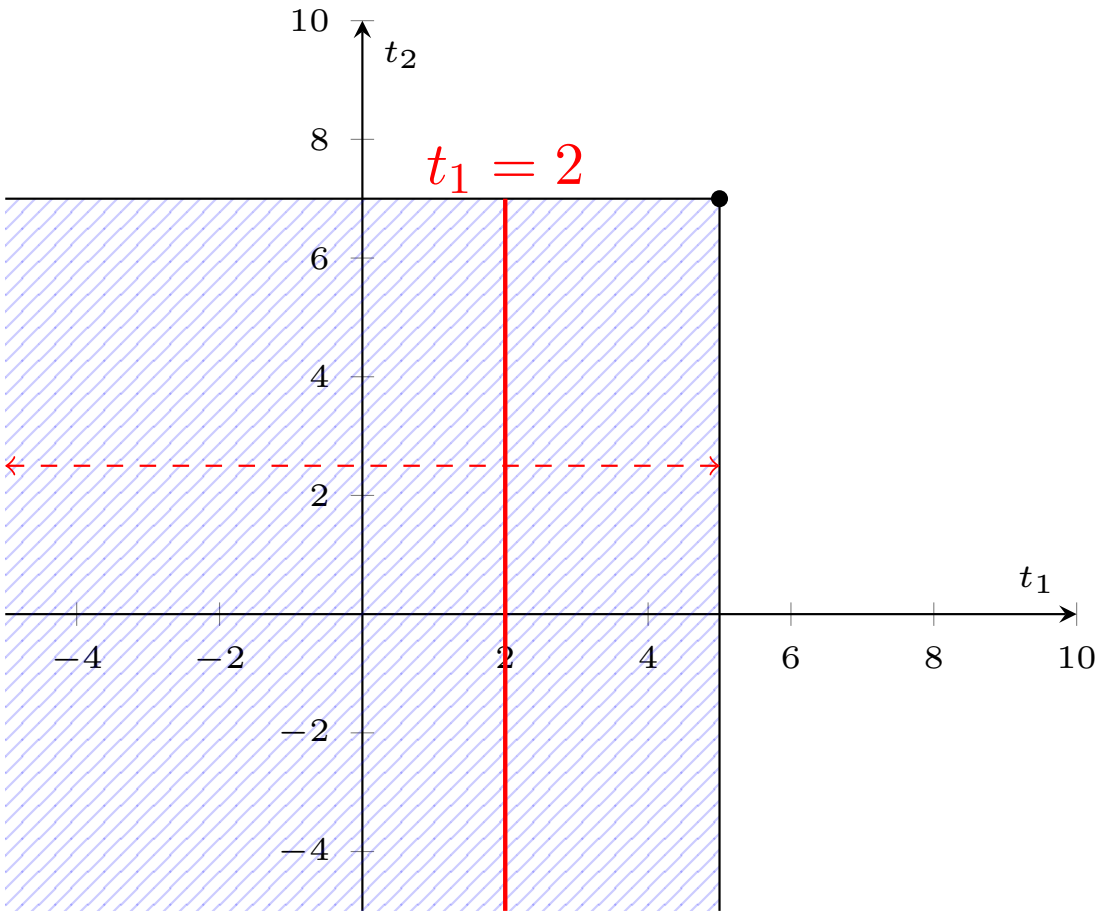
  in the 'strip' at t_1 to obtain a function $g(t_1, y_2)$, say;
- perform the second (outer) integration

$$\int_{-\infty}^{y_1} g(t_1, y_2)\, dt_1.$$

  to obtain the joint cdf.

The joint density function (cont.)

$$F_{Y_1,Y_2}(5,7) = P(Y_1 \leq 5, Y_2 \leq 7)$$



The joint density function (cont.)

The joint pdf describes how the probability is spread 'point-by-point' across the real plane. By the probability axioms, we must have that

- the joint pdf is non-negative

$$f_{Y_1,Y_2}(y_1, y_2) \geq 0 \qquad -\infty < y_1 < \infty, -\infty < y_2 < \infty$$

(as the joint cdf is non-decreasing in both y_1 and y_2);

- the joint pdf integrates to 1

$$\int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 \right\} dy_1 = 1$$

(as the probability must accumulate to 1 over the real plane).

The joint density function (cont.)

Example (Distribution on the unit square)

Suppose Y_1 and Y_2 are continuous with joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = c(y_1 + y_2) \qquad 0 \leq y_1 \leq 1, 0 \leq y_2 \leq 1$$

with $f_{Y_1,Y_2}(y_1, y_2) = 0$ otherwise.

The joint density function (cont.)

Example (Distribution on the unit square)

Then for $0 \leq y_1 \leq 1, 0 \leq y_2 \leq 1$,

$$\begin{aligned} F_{Y_1,Y_2}(y_1, y_2) &= \int_{-\infty}^{y_1} \left\{ \int_{-\infty}^{y_2} f_{Y_1,Y_2}(t_1, t_2)\, dt_2 \right\} dt_1 \\ &= \int_{0}^{y_1} \left\{ \int_{0}^{y_2} c(t_1 + t_2)\, dt_2 \right\} dt_1 \\ &= \int_{0}^{y_1} \left[c\left(t_1 t_2 + \frac{1}{2} t_2^2 \right) \right]_0^{y_2} dt_1 \\ &= \int_{0}^{y_1} c\left(t_1 y_2 + \frac{1}{2} y_2^2 \right) dt_1 \end{aligned}$$

Example (Distribution on the unit square)

$$= \left[c \left(\frac{1}{2} t_1^2 y_2 + \frac{1}{2} y_2^2 t_1 \right) \right]_0^{y_1}$$

$$= \frac{c}{2} \left(y_1^2 y_2 + y_1 y_2^2 \right)$$

We require that $F_{Y_1,Y_2}(1,1) = 1$, so we must have $c = 1$.

The joint density function (cont.)

Example (Distribution on the unit square)

That is

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} (y_1+y_2) & 0 \le y_1 \le 1, 0 \le y_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(no probability outside of the unit square).

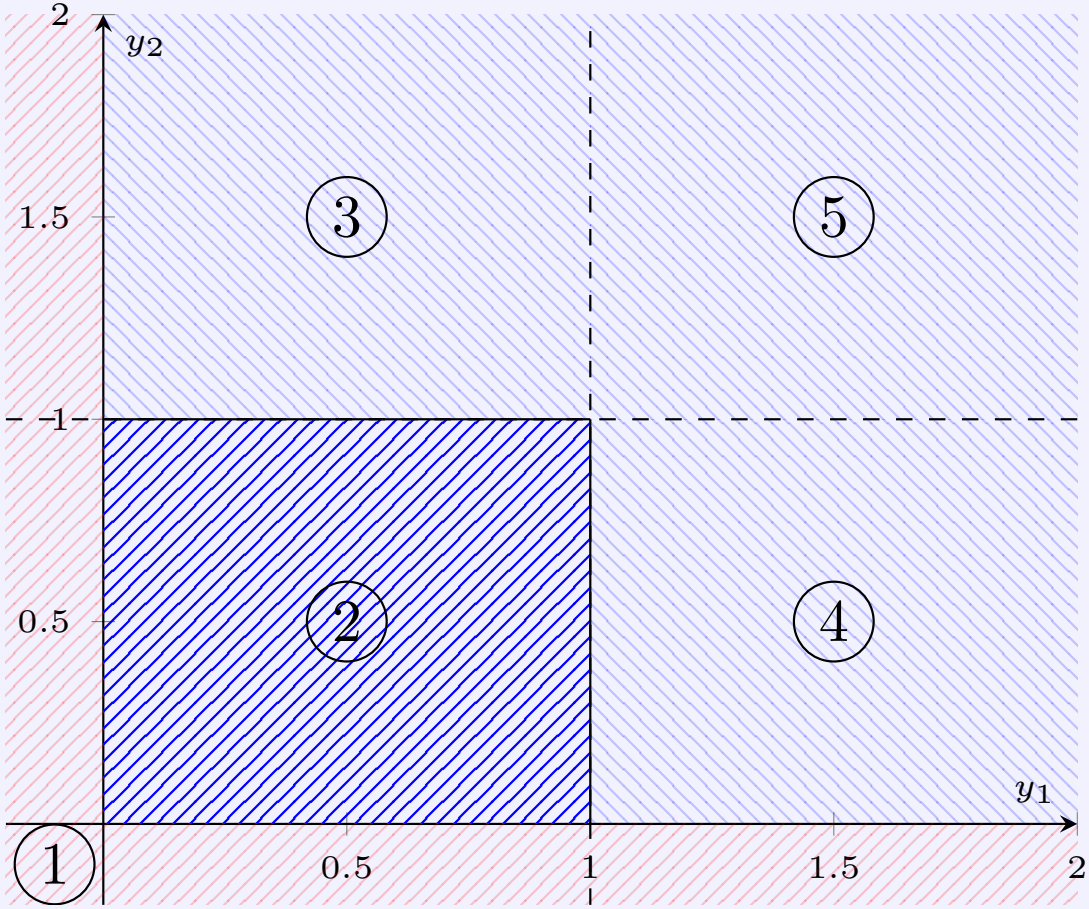
The joint density function (cont.)

Example (Distribution on the unit square)

$$F_{Y_1,Y_2}(y_1,y_2) = \begin{cases} 0 & y_1 < 0 \text{ or } y_2 < 0 \quad (1) \\ \left(y_1^2 y_2 + y_1 y_2^2\right)/2 & 0 \le y_1, y_2 \le 1 \quad (2) \\ \left(y_1^2 + y_1\right)/2 & 0 \le y_1 \le 1, y_2 > 1 \quad (3) \\ \left(y_2 + y_2^2\right)/2 & 0 \le y_2 \le 1, y_1 > 1 \quad (4) \\ 1 & y_1 > 1 \text{ and } y_2 > 1 \quad (5) \end{cases}$$

The joint density function (cont.)

Note: regions for $F_{Y_1,Y_2}(y_1, y_2)$



The joint density function (cont.)

Note

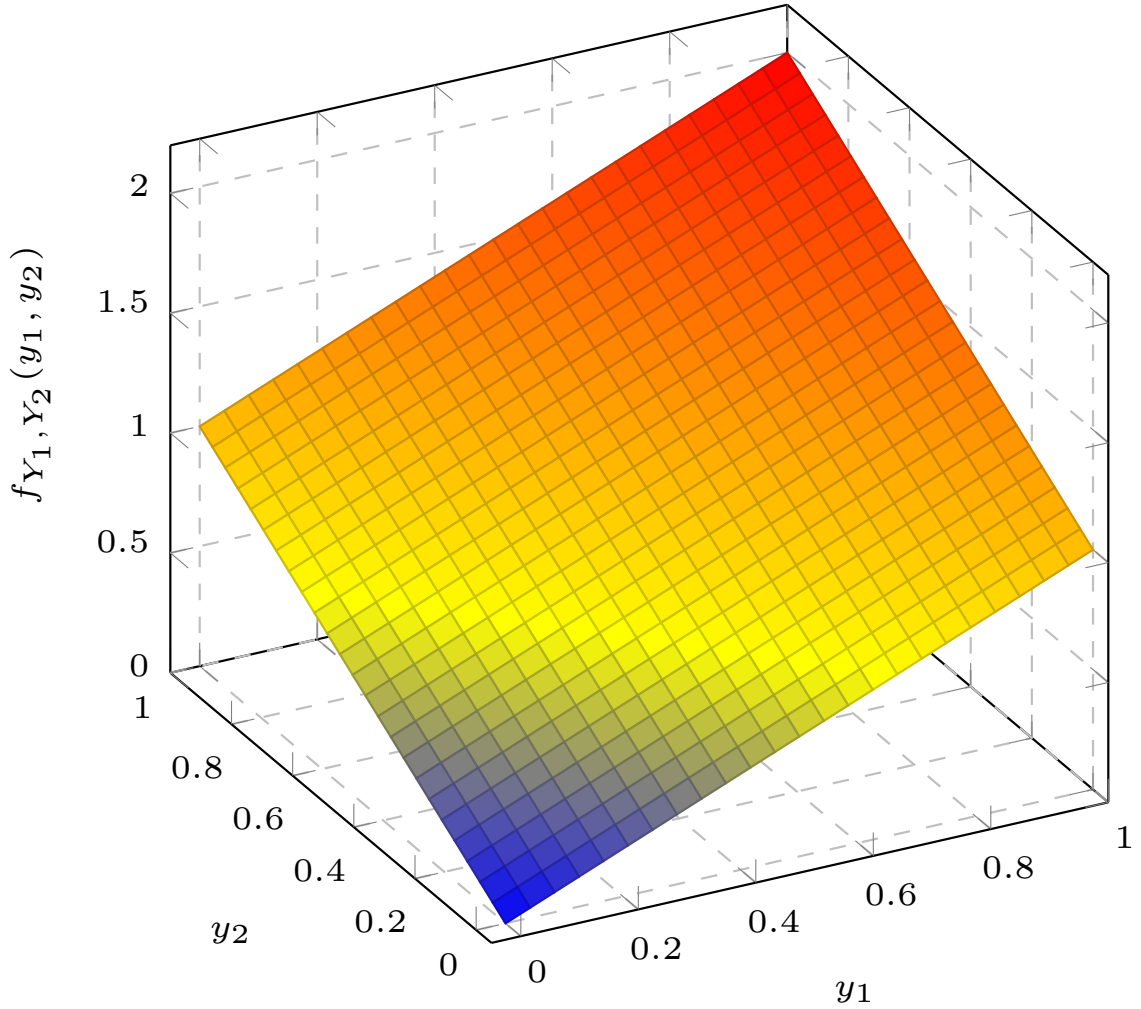
To compute $F_{Y_1,Y_2}(y_1, y_2)$ we always integrate the joint pdf

below and to the left

of (y_1, y_2).

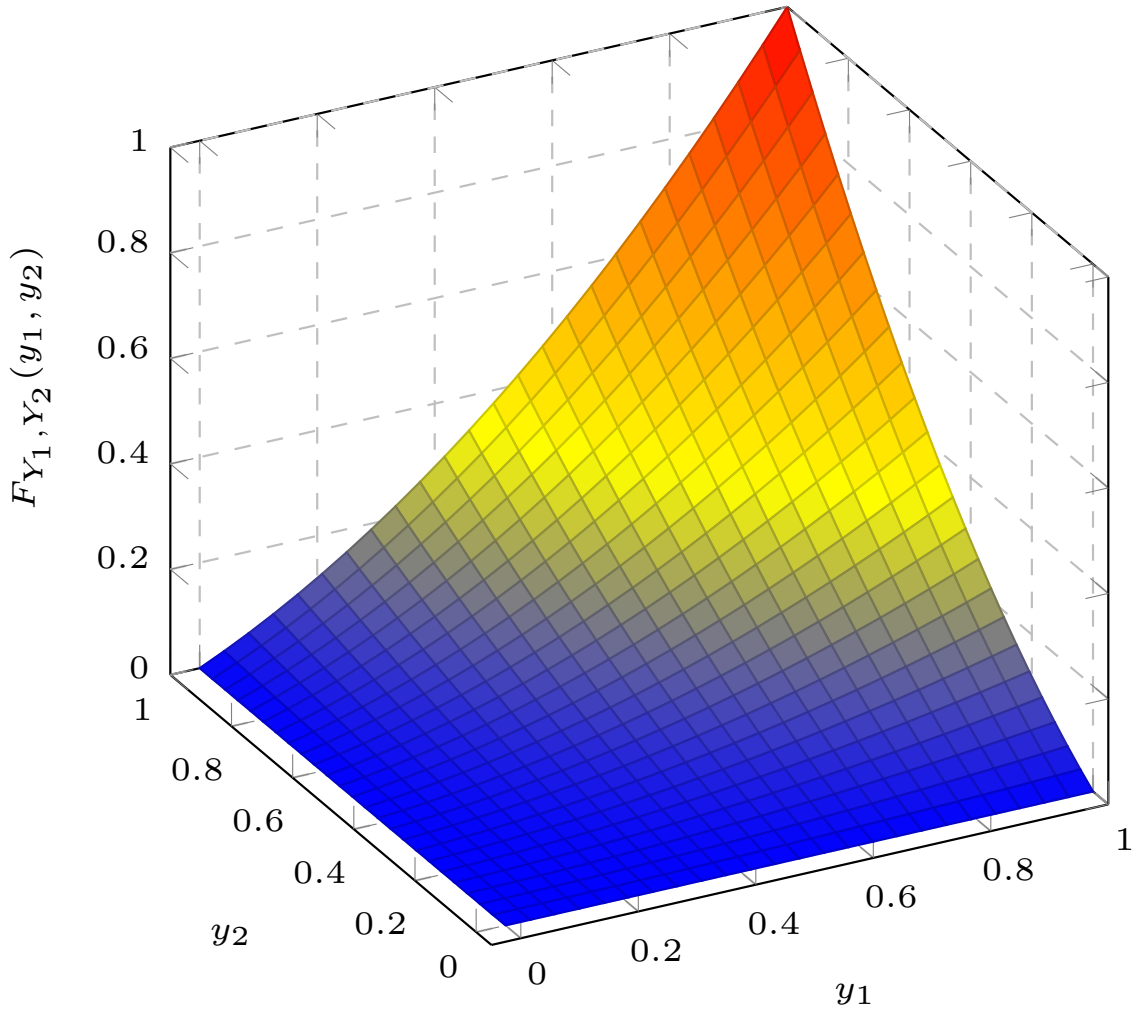
The joint density function (cont.)

$$f_{Y_1,Y_2}(y_1, y_2) = (y_1 + y_2)$$



The joint density function (cont.)

$$F_{Y_1,Y_2}(y_1, y_2) = (y_1^2 y_2 + y_1 y_2^2)/2$$



The joint density function (cont.)

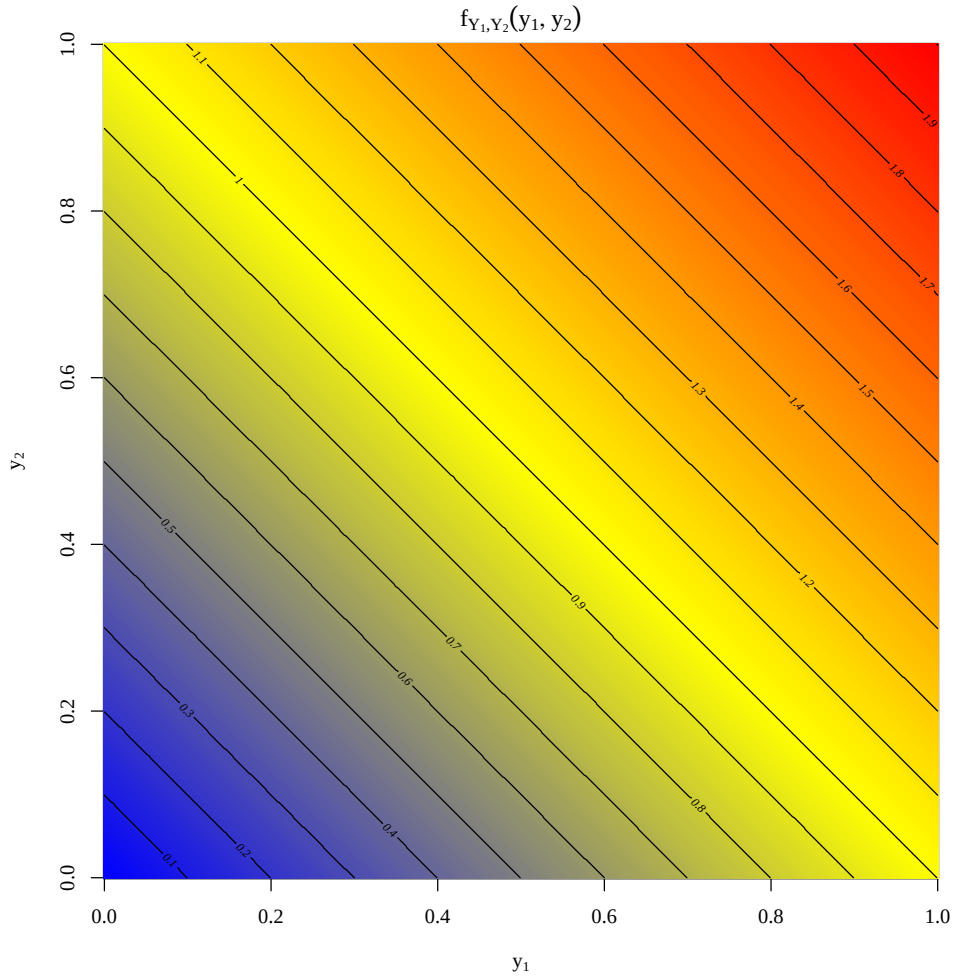


Figure: $f_{Y_1,Y_2}(y_1, y_2)$: image and contour plot

The joint density function (cont.)

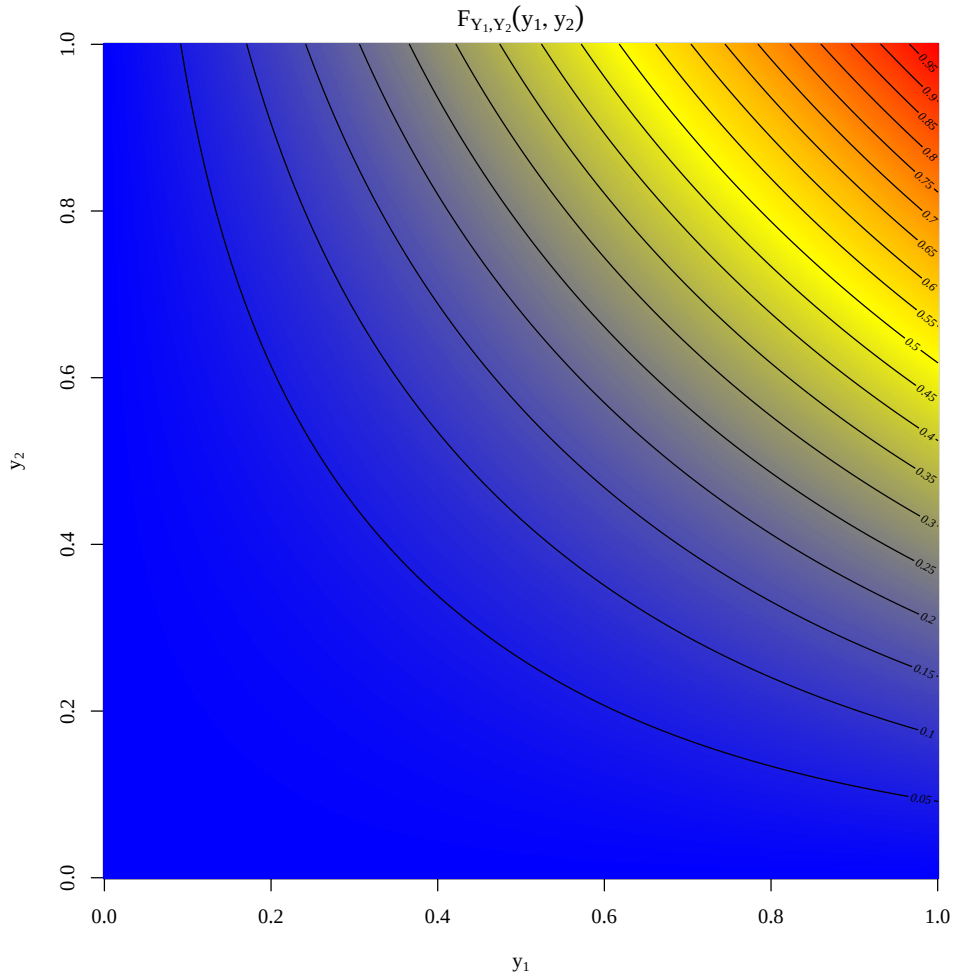


Figure: $F_{Y_1,Y_2}(y_1, y_2)$: image and contour plot

The joint density function (cont.)

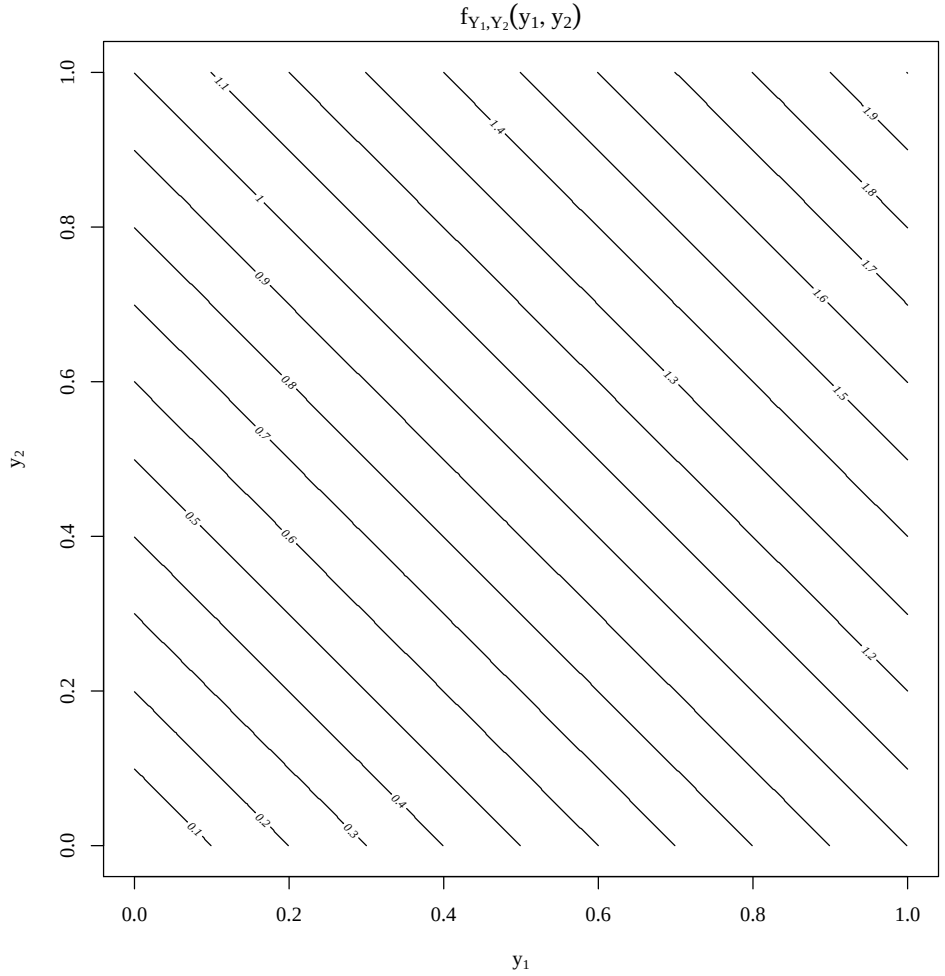


Figure: $f_{Y_1,Y_2}(y_1, y_2)$: contour plot

The joint density function (cont.)

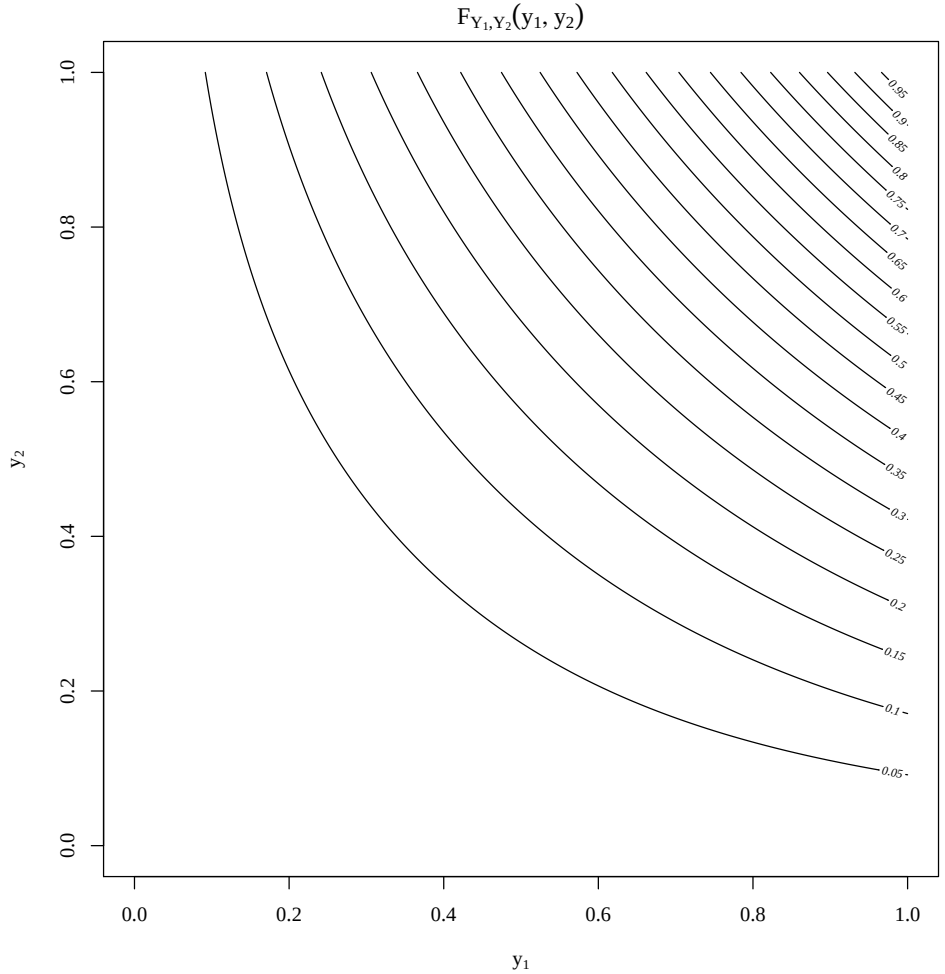


Figure: $F_{Y_1,Y_2}(y_1, y_2)$: contour plot

The joint density function (cont.)

We compute $f_{Y_1,Y_2}(y_1, y_2)$ from $F_{Y_1,Y_2}(y_1, y_2)$ using partial differentiation:

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{\partial^2}{\partial y_1 \partial y_2} \{F_{Y_1,Y_2}(y_1, y_2)\}$$

- **Step 1:** differentiate $F_{Y_1,Y_2}(y_1, y_2)$ with respect to y_2 while holding y_1 constant;
- **Step 2:** take the result of **Step 1**, and differentiate it with respect to y_1.

The joint density function (cont.)

We can regard the calculation as

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{\partial}{\partial y_1}\left\{\frac{\partial F_{Y_1,Y_2}(y_1, y_2)}{\partial y_2}\right\}$$

The joint density function (cont.)

Example

For $0 \leq y_1, y_2 \leq 1$,

$$F_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2}\left(y_1^2 y_2 + y_1 y_2^2\right)$$

- **Step 1:**

$$\frac{\partial F_{Y_1,Y_2}(y_1, y_2)}{\partial y_2} = \frac{1}{2}\left(y_1^2 + 2y_1 y_2\right)$$

- **Step 2:**

$$\frac{\partial\left\{\frac{1}{2}\left(y_1^2 + 2y_1 y_2\right)\right\}}{\partial y_2} = \frac{1}{2}\left(2y_1 + 2y_2\right) = \left(y_1 + y_2\right)$$

The joint density function (cont.)

Note

In the partial differentiation, we could choose to differentiate with respect to y_1 in Step 1, and then differentiate the result with respect to y_2: we will get the same answer

$$\frac{\partial^2}{\partial y_1 \partial y_2}\left\{F_{Y_1,Y_2}(y_1,y_2)\right\} = \frac{\partial^2}{\partial y_2 \partial y_1}\left\{F_{Y_1,Y_2}(y_1,y_2)\right\}.$$

This is a general property of partial derivatives that holds provided the result of each side is continuous.

Constant pdfs

If the joint pdf is constant over a finite region $\mathcal{Y} \in \mathbb{R}^2$

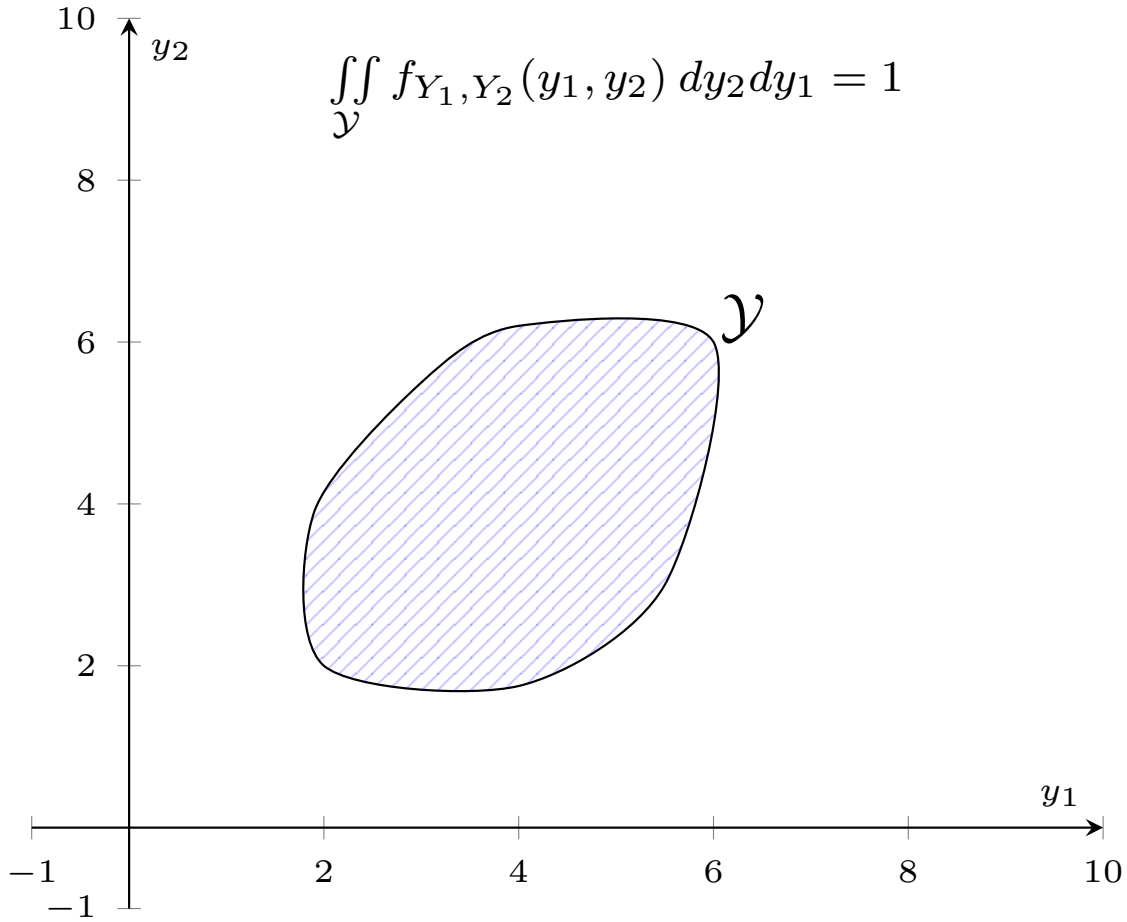
$$f_{Y_1,Y_2}(y_1, y_2) = c \qquad (y_1, y_2) \in \mathcal{Y}$$

with $f_{Y_1,Y_2}(y_1, y_2)$ zero otherwise, then to compute probabilities associated with this pdf, for example

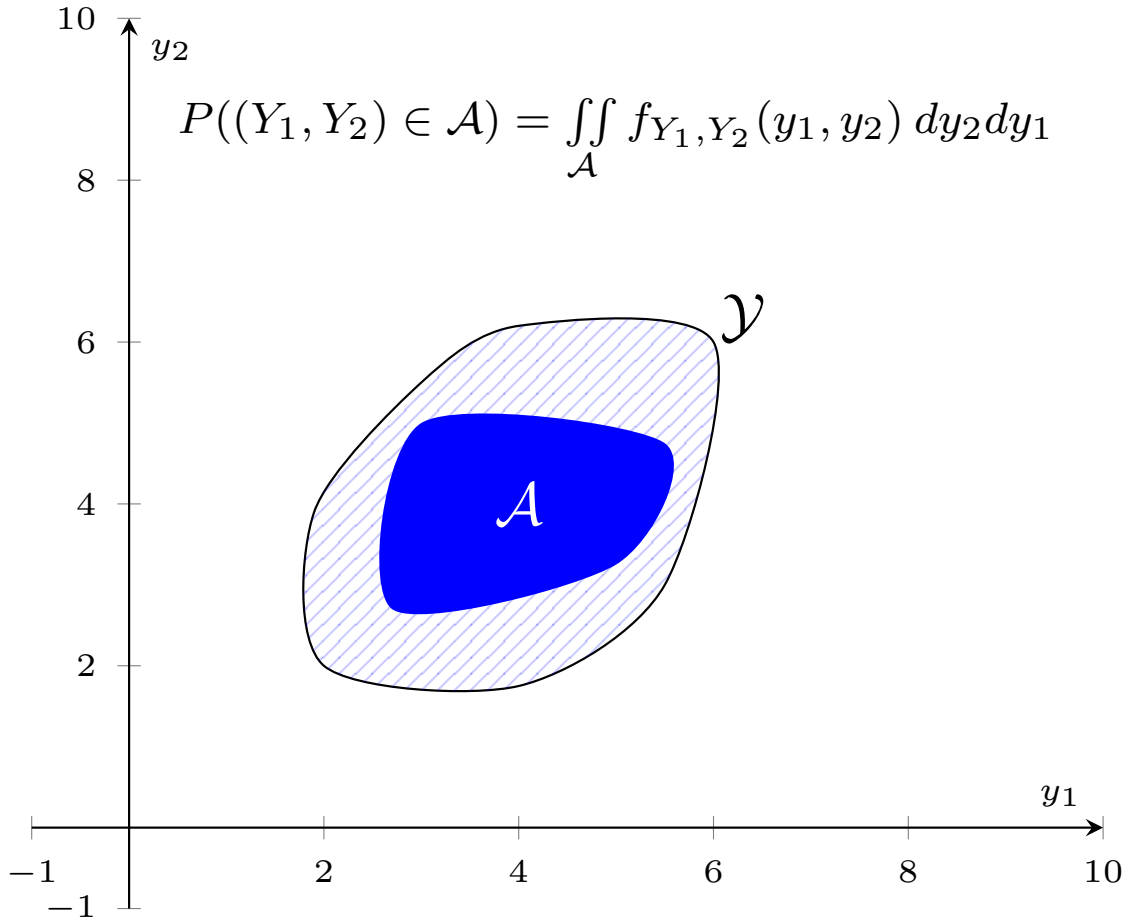
$$P((Y_1, Y_2) \in \mathcal{A}) = \iint_{\mathcal{A}} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 dy_1$$

we must compute the area of $\mathcal{A}$ and divide it by the area of $\mathcal{Y}$.

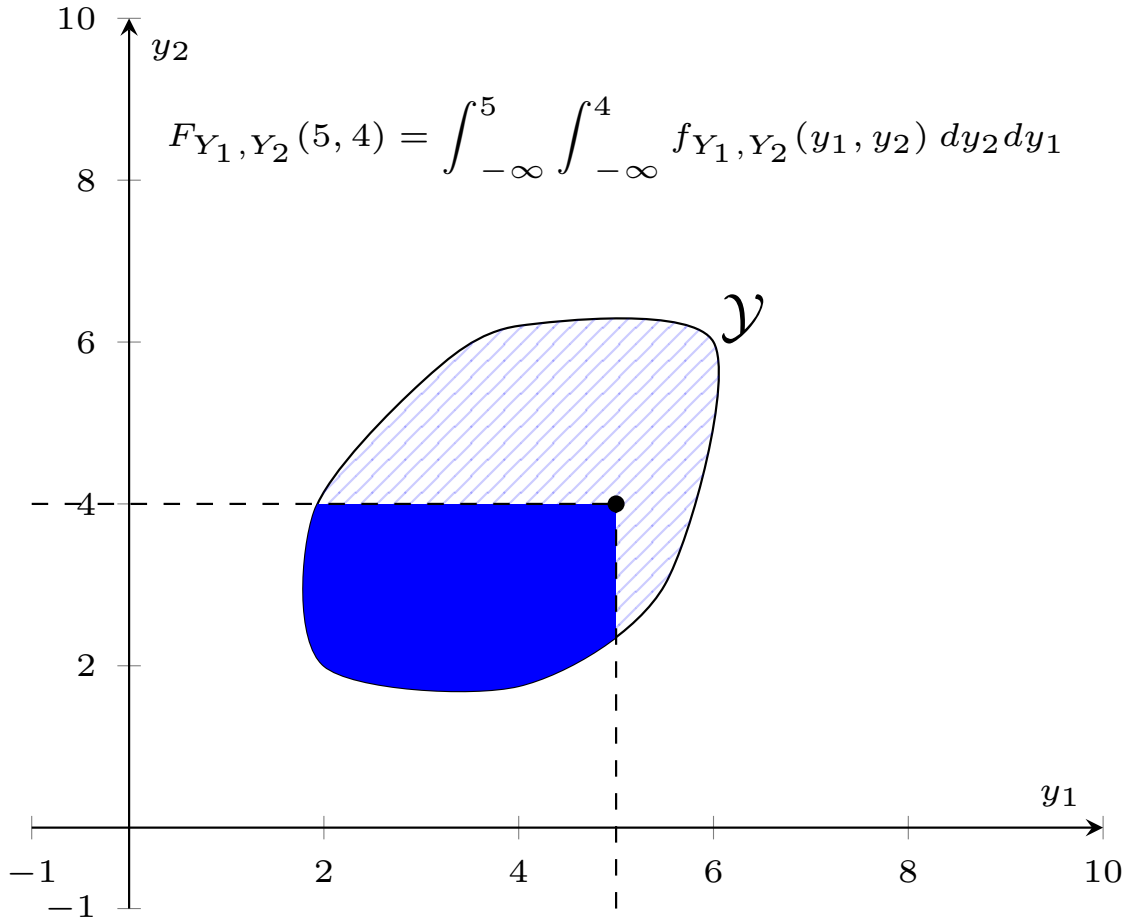
Constant pdfs (cont.)



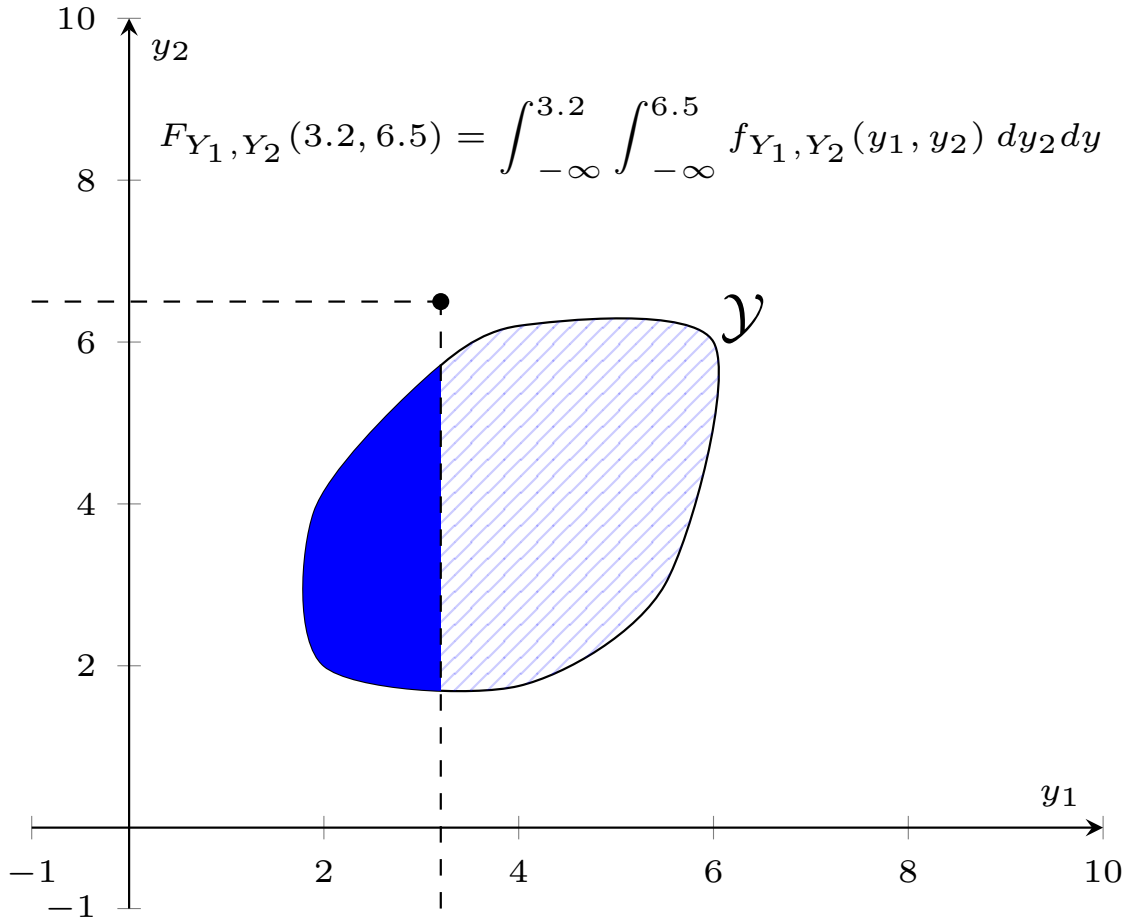
Constant pdfs (cont.)



Constant pdfs (cont.)



Constant pdfs (cont.)



The marginal density function

Once the joint pdf $f_{Y_1,Y_2}(y_1, y_2)$ is specified, we define the *marginal probability density functions* (marginal pdfs) analogously to the discrete case.

We have for Y_1 the marginal pdf $f_{Y_1}(y_1)$ for each fixed y_1 by

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2$$

and for Y_2 the marginal pdf $f_{Y_2}(y_2)$ for each fixed y_2 by

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_1$$

The marginal density function (cont.)

Example (Distribution on the unit square)

For

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} (y_1+y_2) & 0 \le y_1 \le 1, 0 \le y_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

The marginal density function (cont.)

Example (Distribution on the unit square)

We have for $0 \leq y_1 \leq 1$

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 \\ &= \int_0^1 (y_1 + y_2)\, dy_2 \\ &= \left[y_1 y_2 + \frac{y_2^2}{2} \right]_0^1 \\ &= y_1 + \frac{1}{2}. \end{aligned}$$

with $f_{Y_1}(y_1) = 0$ otherwise.

The marginal density function (cont.)

Example (Distribution on the unit square)

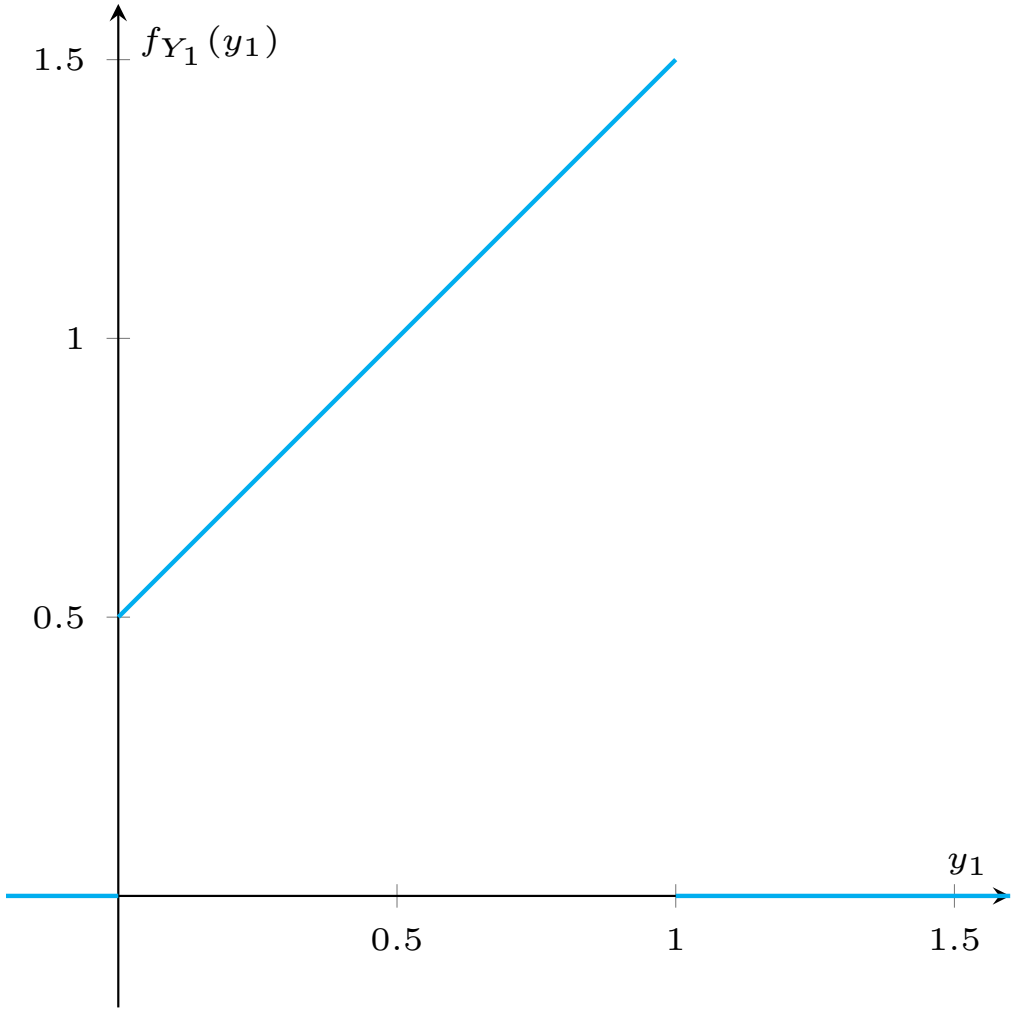
Also for $0 \leq y_1 \leq 1$

$$\begin{aligned} F_{Y_1}(y_1) &= \int_{-\infty}^{y_1} f_{Y_1}(t_1)\, dt_1 \\ &= \int_0^{y_1} \left(t_1 + \frac{1}{2} \right) \, dt_2 \\ &= \left[\frac{t_1^2}{2} + \frac{t_1}{2} \right]_0^{y_1} \\ &= \frac{y_1^2}{2} + \frac{y_1}{2}. \end{aligned}$$

with $F_{Y_1}(y_1) = 0$ for $y_1 < 0$ and $F_{Y_1}(y_1) = 1$ for $y_1 > 1$.

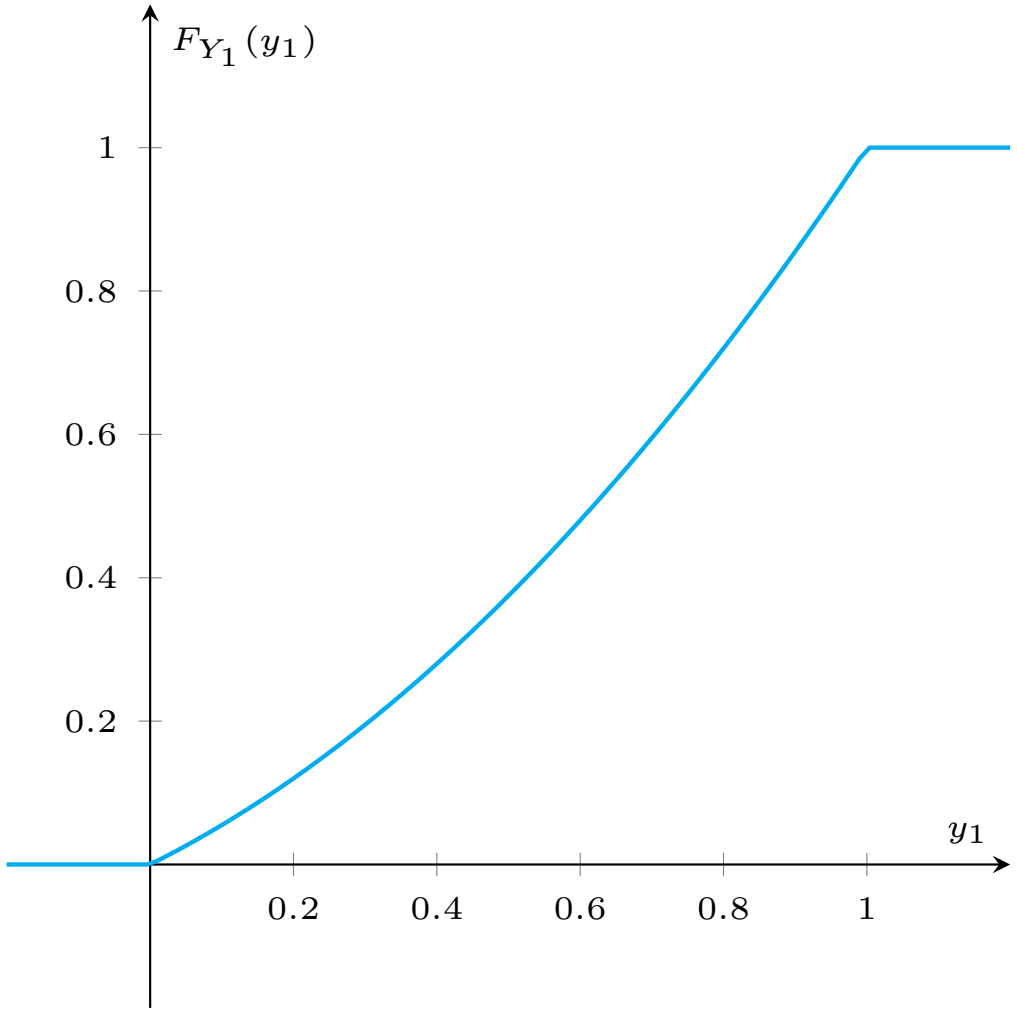
The marginal density function (cont.)

$f_{Y_1}(y_1) = y_1 + 1/2$ for $0 \leq y_1 \leq 1$.



The marginal density function (cont.)

$F_{Y_1}(y_1) = (y_1^2 + y_1)/2$ for $0 \leq y_1 \leq 1$.



The marginal density function (cont.)

Example (Distribution on the unit square)

We can perform a similar calculation and obtain

$$f_{Y_2}(y_2) = y_2 + \frac{1}{2} \qquad 0 \leq y_2 \leq 1$$

with $f_{Y_2}(y_2) = 0$ otherwise, and

$$F_{Y_2}(y_2) = \frac{(y_2^2 + y_2)}{2} \qquad 0 \leq y_2 \leq 1$$

with $F_{Y_2}(y_2) = 0$ for $y_2 < 0$ and $F_{Y_2}(y_2) = 1$ for $y_2 > 1$.

The conditional density function

Once we have the joint pdf

$$f_{Y_1,Y_2}(y_1, y_2)$$

and the marginal pdf

$$f_{Y_1}(y_1)$$

we can consider *conditional* pdfs.

The conditional density function (cont.)

The *conditional probability density function* for Y_2, given that $Y_1 = y_1$, is denoted

$$f_{Y_2|Y_1}(y_2|y_1)$$

and defined by

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_1}(y_1)}$$

whenever $f_{Y_1}(y_1) > 0$.

The conditional density function (cont.)

The function $f_{Y_2|Y_1}(y_2|y_1)$ is a pdf in y_2 for every fixed y_1. That is, for every fixed y_1 where $f_{Y_1}(y_1) > 0$,

- the conditional pdf is non-negative

$$f_{Y_2|Y_1}(y_2|y_1) \geq 0 \qquad -\infty < y_2 < \infty;$$

- the conditional pdf is integrates to 1 over y_2

$$\int_{-\infty}^{\infty} f_{Y_2|Y_1}(y_2|y_1)\, dy_2 = 1.$$

We can define $f_{Y_1|Y_2}(y_1|y_2)$ in an identical fashion.

The conditional density function (cont.)

Conditional pdfs are obtained by taking a 'slice' through the joint pdf at $Y_1 = y_1$, and then standardizing the slice so that the density integrates to one.

The conditional density function (cont.)

Example (Distribution on the unit square)

For

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} (y_1+y_2) & 0 \le y_1 \le 1, 0 \le y_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

we can consider conditioning on values of y_1 or y_2 in the interval $[0,1]$.

The conditional density function (cont.)

Example (Distribution on the unit square)

As

$$f_{Y_1}(y_1) = y_1 + \frac{1}{2} \qquad 0 \leq y_1 \leq 1.$$

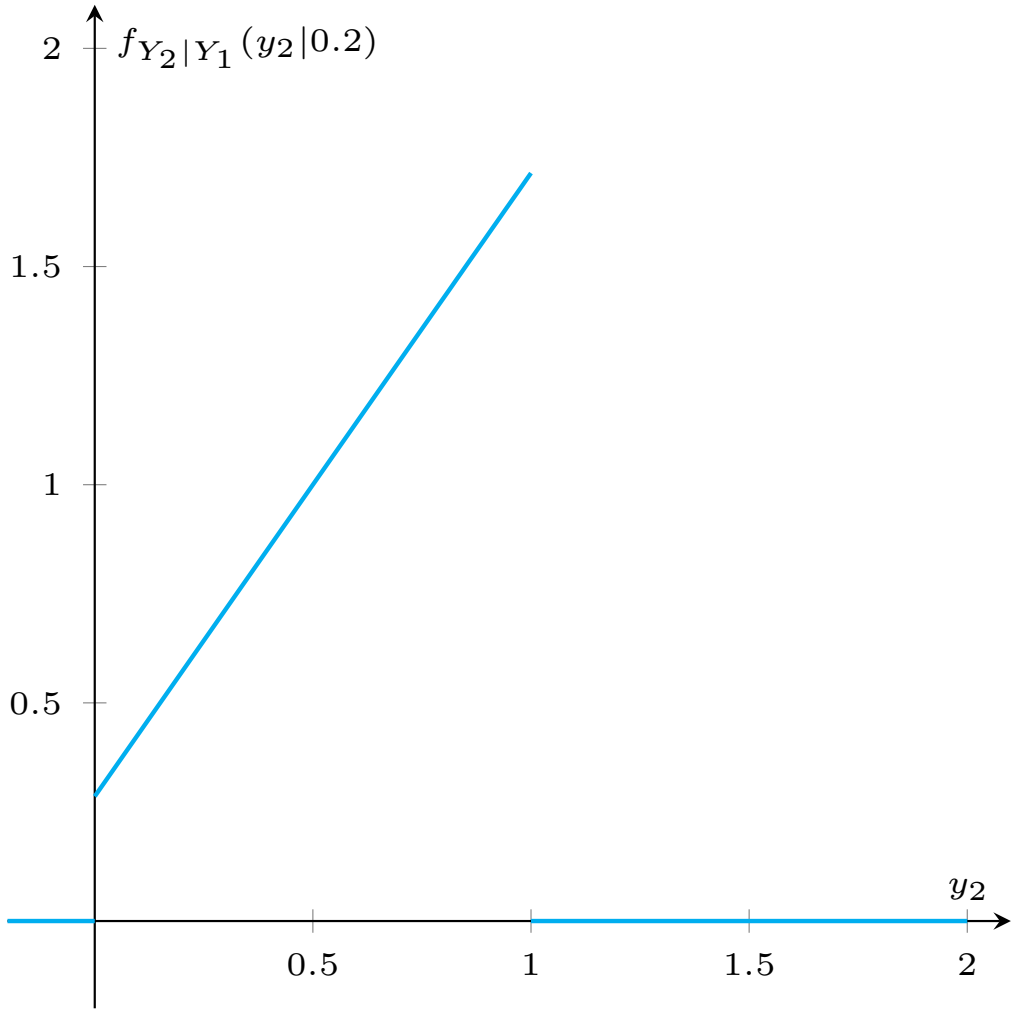
we have for each fixed y_1 in this range

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_1}(y_1)} = \frac{(y_1 + y_2)}{y_1 + 1/2}$$

for $0 \leq y_2 \leq 1$, with the function zero for other values of y_2.

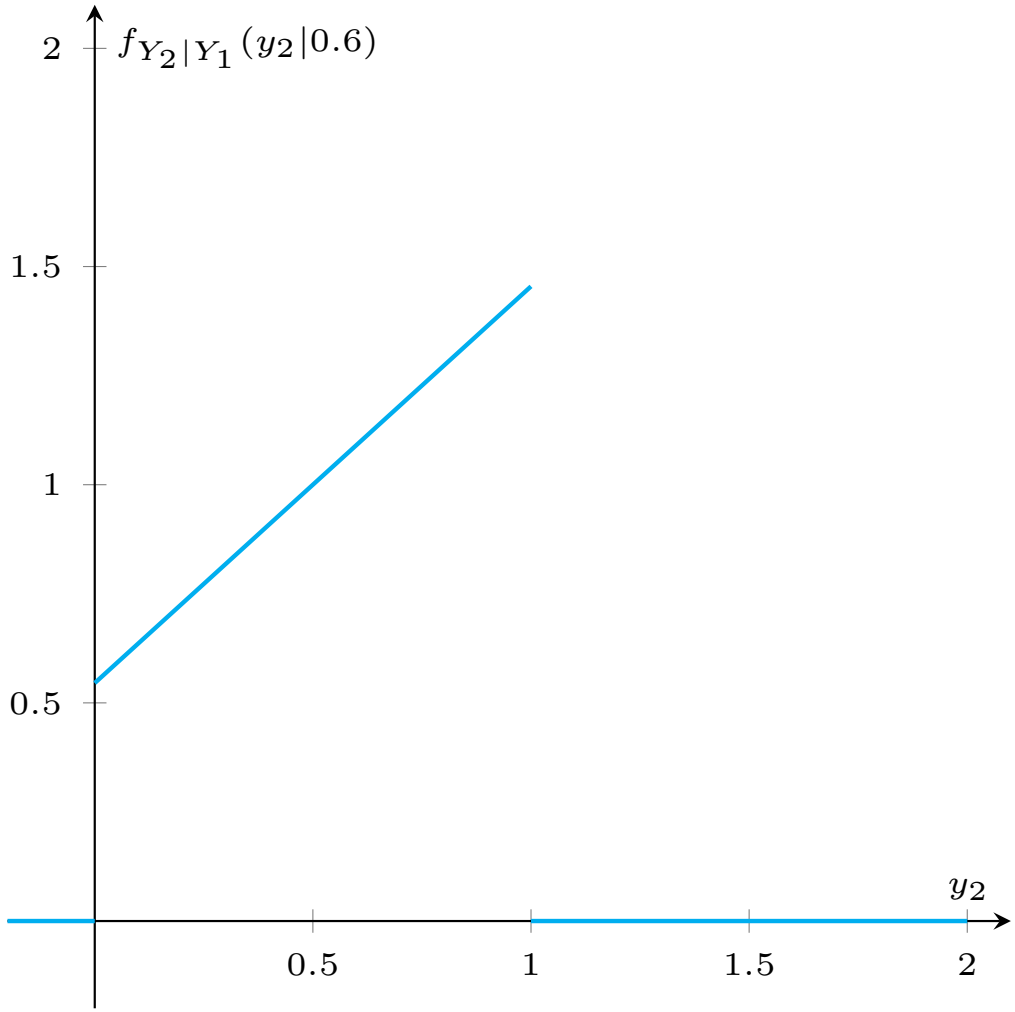
The conditional density function (cont.)

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{(y_1 + y_2)}{y_1 + 1/2} \text{ for } 0 \leq y_2 \leq 1, \text{ with } y_1 = 0.2$$



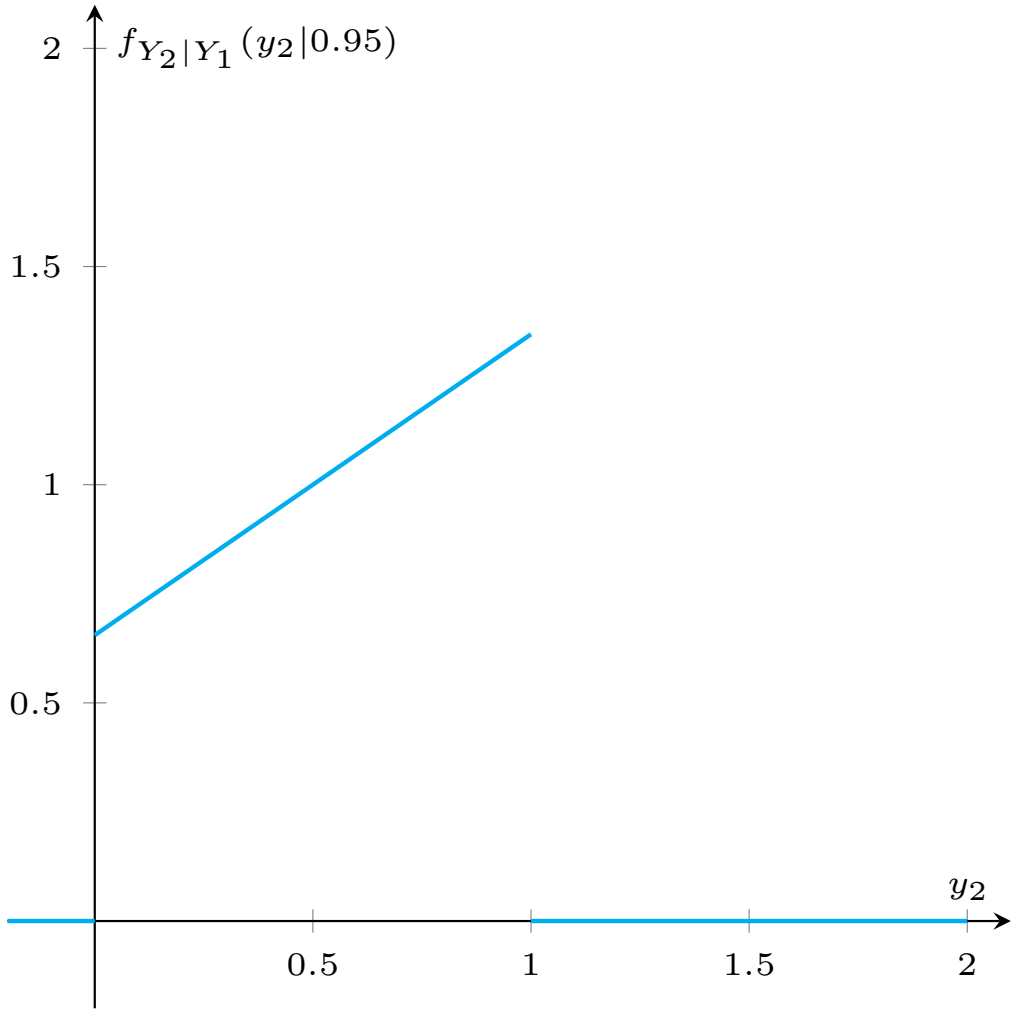
The conditional density function (cont.)

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{(y_1 + y_2)}{y_1 + 1/2} \text{ for } 0 \leq y_2 \leq 1, \text{ with } y_1 = 0.6$$



The conditional density function (cont.)

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{(y_1 + y_2)}{y_1 + 1/2} \text{ for } 0 \leq y_2 \leq 1, \text{ with } y_1 = 0.95$$



The conditional density function (cont.)

As in the discrete case, we have the *chain rule factorization*

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2|Y_1}(y_2|y_1)$$

whenever $f_{Y_1}(y_1) > 0$.

As ever, we can exchange the roles of Y_1 and Y_2 to obtain that

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_2}(y_2) f_{Y_1|Y_2}(y_1|y_2).$$

whenever $f_{Y_1}(y_1) > 0$.

All of the above ideas extend naturally to more than two variables. If $(Y_1, \ldots, Y_n)$ form an n-dimensional random vector, we can consider the *joint pdf*

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)$$

and *joint cdf*

$$F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n).$$

Multivariate distributions (cont.)

We can consider *marginalization* by integrating out $n - d < n$ of the dimensions to leave a d-dimensional probability distribution: for example if $n = 4$ and $d = 2$, we have

$$f_{Y_1,Y_2}(y_1, y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{Y_1,Y_2,Y_3,Y_4}(y_1, y_2, y_3, y_4)\, dy_3 dy_4$$

or

$$f_{Y_2,Y_4}(y_2, y_4) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{Y_1,Y_2,Y_3,Y_4}(y_1, y_2, y_3, y_4)\, dy_1 dy_3$$

and so on. This construction works with any n and any d.

Independence

Continuous random variables $(Y_1, \ldots, Y_n)$ are *independent* if

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i)$$

for **all** vectors $(y_1, \ldots, y_n) \in \mathbb{R}^n$. Equivalently they are independent if

$$F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} F_{Y_i}(y_i)$$

for **all** vectors $(y_1, \ldots, y_n) \in \mathbb{R}^n$.

Random variables that are not independent are termed *dependent*.

Independence (cont.)

In the bivariate case, Y_1 and Y_2 are independent if

$$f_{Y_2|Y_1}(y_2|y_1) = f_{Y_2}(y_2)$$

for all (y_1, y_2) where the conditional pdf is defined, or equivalently if

$$f_{Y_1|Y_2}(y_1|y_2) = f_{Y_1}(y_1)$$

for all (y_1, y_2).

Independence (cont.)

Note

A straightforward way to exclude the possibility of two variables being independent is to assess whether the set of values, $\mathcal{Y}_{12}$ (called the *support* of the joint pdf), for which

$$f_{Y_1,Y_2}(y_1, y_2) > 0$$

is identical to the Cartesian product of the two sets, $\mathcal{Y}_1$ and $\mathcal{Y}_2$ for which the marginal densities satisfy $f_{Y_1}(y_1) > 0$ and $f_{Y_2}(y_2) > 0$.

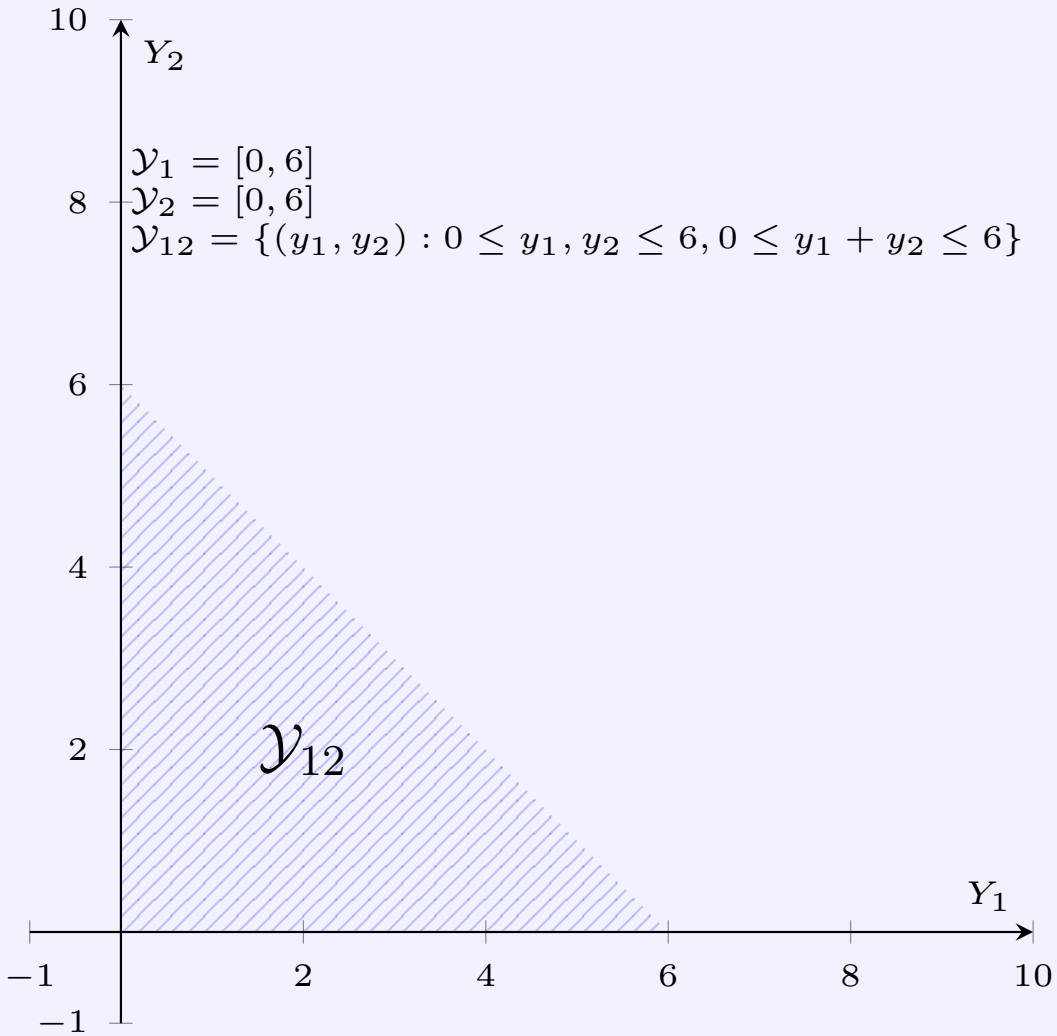
That is, if

$$\mathcal{Y}_{12} \neq \mathcal{Y}_1 \times \mathcal{Y}_2$$

then Y_1 and Y_2 are **not** independent.

Independence (cont.)

Note



Independence (cont.)

Note

We can find points where

- $f_{Y_1}(y_1) > 0$ and $f_{Y_2}(y_2) > 0$, but
- $f_{Y_1,Y_2}(y_1, y_2) = 0$

(for example, the point $(5, 5)$), so we do not meet the requirement for independence that

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2)$$

for **all** $(y_1, y_2) \in \mathbb{R}^2$.

Multivariate expectations

We can extend the idea of expectation to two or more dimensions.

We consider a function $g(y_1, \ldots, y_n)$ of n random variables. The expectation of this function is defined in the discrete case by

$$\mathbb{E}[g(Y_1, \ldots, Y_n)] = \sum_{y_1=-\infty}^{\infty} \cdots \sum_{y_n=-\infty}^{\infty} g(y_1, \ldots, y_n) p_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n).$$

Multivariate expectations (cont.)

In the continuous case we define

$$\mathbb{E}[g(Y_1, \ldots, Y_n)] =$$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, \ldots, y_n) f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)\, dy_1 \ldots dy_n.$$

These definitions hold when the sum/integral is finite.

Multivariate expectations (cont.)

For two variables, this becomes

$$\mathbb{E}[g(Y_1, Y_2)] = \sum_{y_1=-\infty}^{\infty} \sum_{y_2=-\infty}^{\infty} g(y_1, y_2) p_{Y_1,Y_2}(y_1, y_2).$$

or

$$\mathbb{E}[g(Y_1, Y_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y_1, y_2) f_{Y_1,Y_2}(y_1, y_2)\, dy_1 dy_2.$$

Multivariate expectations (cont.)

The properties of expectations for more than one variable match those in the single variable case.

In particular, if a_1, a_2 are constants, and $g_1(.,.)$ and $g_2(.,.)$ are two functions, then

$$\mathbb{E}[a_1 g_1(Y_1, Y_2) + a_2 g_2(Y_1, Y_2)] = a_1 \mathbb{E}[g_1(Y_1, Y_2)] + a_2 \mathbb{E}[g_2(Y_1, Y_2)]$$

provided all the expectations are finite.

Covariance and correlation

For two random variables Y_1 and Y_2, with expectations μ_1 and μ_2 respectively, the *covariance* between Y_1 and Y_2, $\text{Cov}[Y_1, Y_2]$, is defined by

$$\text{Cov}[Y_1, Y_2] = \mathbb{E}[(Y_1 - \mu_1)(Y_2 - \mu_2)]$$

Covariance and correlation (cont.)

Note that

$$\begin{aligned}
\mathrm{Cov}[Y_1, Y_2] &= \mathbb{E}[(Y_1 - \mu_1)(Y_2 - \mu_2)] \\
&= \mathbb{E}[Y_1 Y_2 - \mu_1 Y_2 - Y_1 \mu_2 + \mu_1 \mu_2] \\
&= \mathbb{E}[Y_1 Y_2] - \mu_1 \mathbb{E}[Y_2] - \mu_2 \mathbb{E}[Y_1] + \mu_1 \mu_2 \\
&= \mathbb{E}[Y_1 Y_2] - \mu_1 \mu_2 - \mu_2 \mu_1 + \mu_1 \mu_2 \\
&= \mathbb{E}[Y_1 Y_2] - \mu_1 \mu_2.
\end{aligned}$$

Covariance and correlation (cont.)

The *correlation* between Y_1 and Y_2, $\text{Corr}(Y_1, Y_2)$, is defined by

$$\text{Corr}[Y_1, Y_2] = \frac{\text{Cov}[Y_1, Y_2]}{\sqrt{\mathbb{V}[Y_1]\mathbb{V}[Y_2]}}$$

Covariance and correlation (cont.)

Both covariance and correlation are *measures of dependence*. We have that for any pair of random variables,

$$-1 \leq \text{Corr}[Y_1, Y_2] \leq 1$$

Covariance and correlation (cont.)

If

$$\mathrm{Cov}[Y_1, Y_2] = \mathrm{Corr}[Y_1, Y_2] = 0$$

we say that Y_1 and Y_2 are *uncorrelated*.

Independent random variables

If Y_1 and Y_2 are independent then (in the continuous case – discrete case works in the same way), we have that

$$\begin{aligned}
\mathbb{E}[Y_1 Y_2] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f_{Y_1,Y_2}(y_1, y_2)\, dy_1 dy_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f_{Y_1}(y_1) f_{Y_2}(y_2)\, dy_1 dy_2 \\
&= \int_{-\infty}^{\infty} y_2 f_{Y_2}(y_2) \left\{ \int_{-\infty}^{\infty} y_1 f_{Y_1}(y_1)\, dy_1 \right\} dy_2 \\
&= \int_{-\infty}^{\infty} y_2 f_{Y_2}(y_2) \mathbb{E}[Y_1] dy_2 \\
&= \mathbb{E}[Y_1] \int_{-\infty}^{\infty} y_2 f_{Y_2}(y_2) dy_2 = \mathbb{E}[Y_1]\mathbb{E}[Y_2]
\end{aligned}$$

Independent random variables (cont.)

More generally, if $g_1(Y_1)$ and $g_2(Y_2)$ are two functions of the independent random variables, then

$$\mathbb{E}[g_1(Y_1)g_2(Y_2)] = \mathbb{E}[g_1(Y_1)]\mathbb{E}[g_2(Y_2)]$$

Independent random variables (cont.)

Therefore if two random variables Y_1 and Y_2 are independent, then

$$\begin{aligned}\mathrm{Cov}[Y_1, Y_2] &= \mathbb{E}[Y_1 Y_2] - \mu_1 \mu_2 \\ &= \mathbb{E}[Y_1]\mathbb{E}[Y_2] - \mu_1 \mu_2 \\ &= \mu_1 \mu_2 - \mu_1 \mu_2 = 0\end{aligned}$$

and Y_1 and Y_2 are uncorrelated.

Linear combinations of random variables

Suppose that $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_m$ be random variables with

$$\mathbb{E}[Y_i] = \mu_i \qquad \mathbb{E}[X_j] = \xi_j.$$

Define

$$U_1 = \sum_{i=1}^{n} a_i Y_i \qquad U_2 = \sum_{j=1}^{m} b_j X_j$$

for real constants $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$.

Linear combinations of random variables (cont.)

(a) Expectations:

$$\mathbb{E}[U_1] = \sum_{i=1}^{n} a_i \mathbb{E}[Y_i] = \sum_{i=1}^{n} a_i \mu_i$$

(b) Variances:

$$\mathbb{V}[U_1] = \sum_{i=1}^{n} a_i^2 \mathbb{V}[Y_i] + 2 \sum_{j=2}^{n} \sum_{i=1}^{j-1} a_i a_j \mathrm{Cov}[Y_i, Y_j]$$

(c) Covariances:

$$\mathrm{Cov}[U_1, U_2] = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \mathrm{Cov}[Y_i, X_j]$$

For (b):

$$\mathbb{V}[U_1] = \mathbb{E}[U_1^2] - \{\mathbb{E}[U_1]\}^2.$$

But

$$U_1^2 = \left(\sum_{i=1}^{n} a_i Y_i\right)^2 = \sum_{i=1}^{n} a_i^2 Y_i^2 + 2\sum_{j=2}^{n}\sum_{i=1}^{j-1} a_i a_j Y_i Y_j$$

so therefore

$$\mathbb{E}[U_1^2] = \sum_{i=1}^{n} a_i^2 \mathbb{E}[Y_i^2] + 2\sum_{j=2}^{n}\sum_{i=1}^{j-1} a_i a_j \mathbb{E}[Y_i Y_j]$$

Linear combinations of random variables (cont.)

Similarly

$$\{\mathbb{E}[U_1]\}^2 = \left\{\sum_{i=1}^{n} a_i\mu_i\right\}^2 = \sum_{i=1}^{n} a_i^2\mu_i^2 + 2\sum_{j=2}^{n}\sum_{i=1}^{j-1} a_i a_j \mu_i \mu_j$$

Therefore

$$\mathbb{E}[U_1^2] - \{\mathbb{E}[U_1]\}^2 = \sum_{i=1}^{n} a_i^2(\mathbb{E}[Y_i^2] - \mu_i^2) + 2\sum_{j=2}^{n}\sum_{i=1}^{j-1} a_i a_j(\mathbb{E}[Y_i Y_j] - \mu_i\mu_j)$$

and the result follows.

Linear combinations of random variables (cont.)

For (c): consider the simple case

$$U_1 = a_1 Y_1 + a_2 Y_2 \qquad U_2 = b_1 X_1 + b_2 X_2$$

Then

$$\begin{aligned}
\mathbb{E}[U_1 U_2] &= \mathbb{E}[(a_1 Y_1 + a_2 Y_2)(b_1 X_1 + b_2 X_2)] \\
&= \mathbb{E}[a_1 b_1 Y_1 X_1 + a_1 b_2 Y_1 X_2 + a_2 b_1 Y_2 X_1 + a_2 b_2 Y_2 X_2] \\
&= \sum_{i=1}^{2} \sum_{j=1}^{2} a_i b_j \mathbb{E}[Y_i X_j]
\end{aligned}$$

Linear combinations of random variables (cont.)

Using the same logic

$$\begin{aligned}\mathbb{E}[U_1]\mathbb{E}[U_2] &= \left(\sum_{i=1}^{2} a_i \mathbb{E}[Y_i]\right)\left(\sum_{j=1}^{2} b_j \mathbb{E}[X_j]\right) \\ &= \left(\sum_{i=1}^{2} a_i \mu_i\right)\left(\sum_{j=1}^{2} b_j \xi_j\right) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{2} a_i b_j \mu_i \xi_j\end{aligned}$$

and therefore

$$\mathbb{E}[U_1U_2] - \mathbb{E}[U_1]\mathbb{E}[U_2] = \sum_{i=1}^{2}\sum_{j=1}^{2} a_i b_j (\mathbb{E}[Y_i X_j] - \mu_i \xi_j)$$

and the result follows.

Linear combinations of random variables (cont.)

Example (Sums of independent random variables)

If $Y_1, \ldots, Y_n$ are independent and identically distributed with expectation μ and variance σ^2, then

$$U = \sum_{i=1}^{n} Y_i$$

then

$$\mathbb{E}[U] = \sum_{i=1}^{n} \mathbb{E}[Y_i] = \sum_{i=1}^{n} \mu = n\mu$$

and

$$\mathbb{V}[U] = \sum_{i=1}^{n} \mathbb{V}[Y_i] = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

as the covariance terms are all zero.

Sums of random variables (revisited)

For two random variables Y_1 and Y_2 consider the distribution of

$$Y = Y_1 + Y_2$$

which we studied previously in the context using the convolution calculation in Chapter 3.

Sums of random variables (revisited) (cont.)

We now see that the convolution formula can be viewed in the context of a joint distribution calculation:

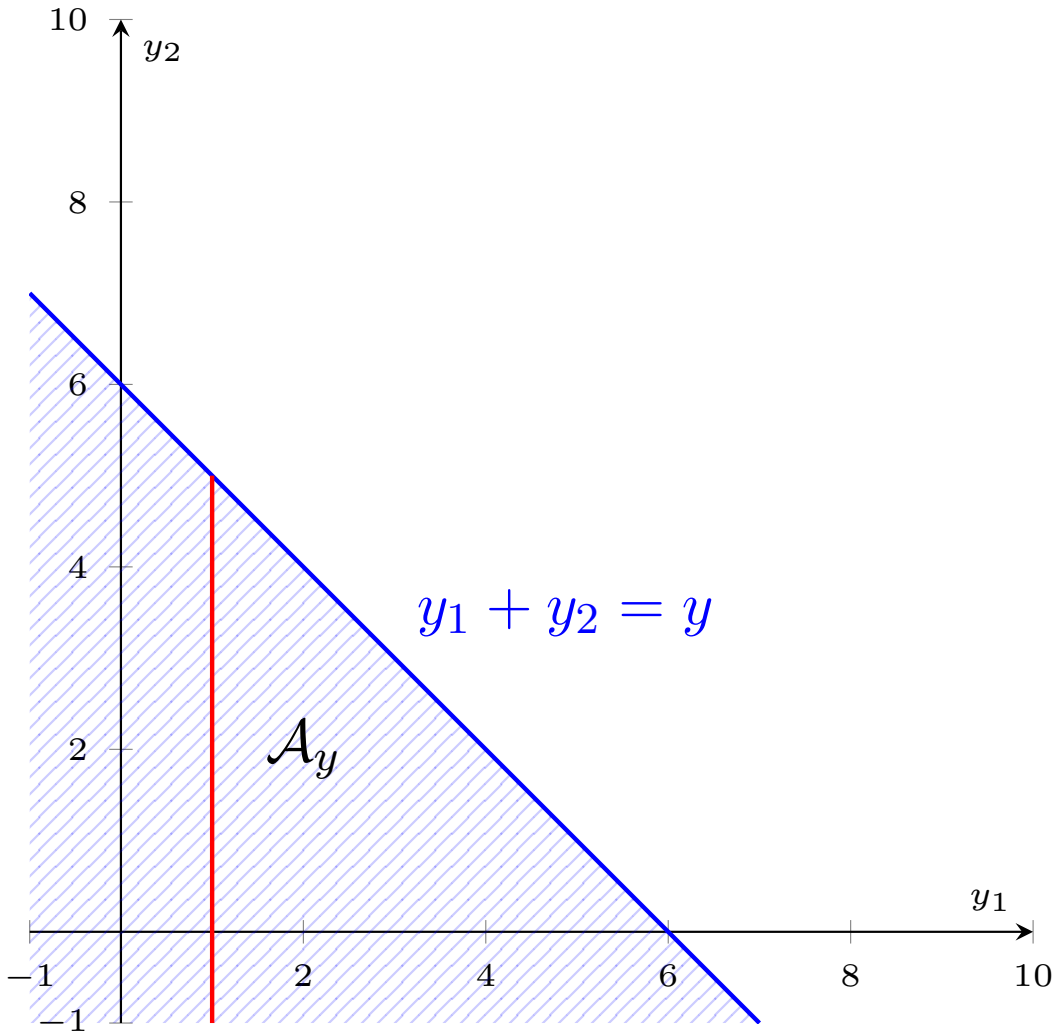
$$F_Y(y) = P(Y \leq y) = P(Y_1 + Y_2 \leq y) = \iint_{\mathcal{A}_y} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 dy_1$$

where

$$\mathcal{A}_y = \left\{(y_1, y_2) \in \mathbb{R}^2 : y_1 + y_2 \leq y\right\}.$$

Sums of random variables (revisited) (cont.)

Need to integrate over the region below the line $y_1 + y_2 = y$.



Therefore we can write

$$F_Y(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y-y_1} f_{Y_1,Y_2}(y_1, y_2)\, dy_2 dy_1.$$

By differentiating with respect to y under the integral sign, using the Fundamental Theorem of Calculus, we see that

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y - y_1)\, dy_1.$$

Sums of random variables (revisited) (cont.)

We now consider an alternate method of calculation: recall the definition of the moment generating function:

$$m_Y(t) = \mathbb{E}\left[e^{tY}\right]$$

for $-b \leq t \leq b$, some $b > 0$.

We can define a similar quantity for the random variable U which is a transformation of Y, $U = g(Y)$ say, as

$$m_U(t) = \mathbb{E}\left[e^{tU}\right] \equiv \mathbb{E}\left[e^{tg(Y)}\right].$$

Sums of random variables (revisited) (cont.)

For two random variables Y_1 and Y_2, with

$$Y = a_1 g_1(Y_1) + a_2 g_2(Y_2),$$

we have

$$m_Y(t) = \mathbb{E}\left[e^{tY}\right] = \mathbb{E}\left[e^{t\{a_1 g_2(Y_1) + a_2 g_2(Y_2)\}}\right]$$

This is the mgf for Y. We can write

$$m_Y(t) = \mathbb{E}\left[e^{tY}\right] = \mathbb{E}\left[e^{ta_1 g_1(Y_1)} e^{ta_2 g_2(Y_2)}\right].$$

Sums of random variables (revisited) (cont.)

If Y_1 and Y_2 are independent we can further write

$$m_Y(t) = \mathbb{E}\left[e^{ta_1 g_1(Y_1)}\right] \mathbb{E}\left[e^{ta_2 g_2(Y_2)}\right].$$

– this follows by the earlier result on expectations.

Sums of random variables (revisited) (cont.)

Now consider the special case $Y = Y_1 + Y_2$:

$$m_Y(t) = \mathbb{E}\left[e^{t\{Y_1+Y_2\}}\right].$$

It follows immediately that if Y_1 and Y_2 are independent,

$$m_Y(t) = \mathbb{E}\left[e^{tY_1}\right]\mathbb{E}\left[e^{tY_2}\right] = m_{Y_1}(t)m_{Y_2}(t).$$

This gives us a way to compute the distribution of sums of independent random variables.

Example (Sum of Binomial random variables)

If Y_1 and Y_2 are independent Binomial random variables with parameters (n_1, p) and (n_2, p), we have

$$m_{Y_1}(t) = (pe^t + q)^{n_1} \qquad m_{Y_2}(t) = (pe^t + q)^{n_2}$$

then if $Y = Y_1 + Y_2$, we have that

$$m_Y(t) = m_{Y_1}(t)m_{Y_2}(t) = (pe^t + q)^{n_1}(pe^t + q)^{n_2} = (pe^t + q)^{n_1+n_2}$$

and we recognize this as the mgf of the Binomial distribution with parameters $(n_1 + n_2, p)$.

Example (Sum of Poisson random variables)

If Y_1 and Y_2 are independent Poisson random variables with parameters λ_1 and λ_2, we have

$$m_{Y_1}(t) = \exp\{\lambda_1(e^t - 1)\} \qquad m_{Y_2}(t) = \exp\{\lambda_2(e^t - 1)\}$$

then if $Y = Y_1 + Y_2$, we have that

$$\begin{aligned} m_Y(t) = m_{Y_1}(t)m_{Y_2}(t) &= \exp\{\lambda_1(e^t - 1)\}\exp\{\lambda_2(e^t - 1)\} \\ &= \exp\{(\lambda_1 + \lambda_2)(e^t - 1)\} \end{aligned}$$

and we recognize this as the mgf of the Poisson distribution with parameter $\lambda_1 + \lambda_2$.

Example (Sum of Normal random variables)

If Y_1 and Y_2 are independent Normal random variables with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2), we have

$$m_{Y_1}(t) = \exp\{\mu_1 t + \sigma_1^2 t^2/2\} \qquad m_{Y_2}(t) = \exp\{\mu_2 t + \sigma_2^2 t^2/2\}$$

then if $Y = Y_1 + Y_2$, we have that

$$\begin{aligned} m_Y(t) = m_{Y_1}(t) m_{Y_2}(t) &= \exp\{\mu_1 t + \sigma_1^2 t^2/2\} \exp\{\mu_2 t + \sigma_2^2 t^2/2\} \\ &= \exp\{(\mu_1 + \mu_2)t + (\sigma_1^2 + \sigma_2^2)t^2/2\} \end{aligned}$$

and we recognize this as the mgf of the Normal distribution with parameters $(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Multivariate transformations

It is possible to consider general multivariate transformations of variables. For example if Y_1 and Y_2 are two variables with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$, we can consider

$$U_1 = h_1(Y_1, Y_2) \qquad U_2 = h_2(Y_1, Y_2)$$

and attempt to compute

$$f_{U_1,U_2}(u_1, u_2).$$

Multivariate transformations (cont.)

For example, we might have

$$U_1 = Y_1 + Y_2 \qquad U_2 = Y_1 - Y_2$$

or

$$U_1 = Y_1 Y_2 \qquad U_2 = Y_1 / Y_2$$

etc.

Multivariate transformations (cont.)

We can approach this from a 'change of variable' perspective. If the transformation

$$(Y_1, Y_2) \longrightarrow (U_1, U_2)$$

is a one-to-one transformation, then the inverse transformation

$$(U_1, U_2) \longrightarrow (Y_1, Y_2)$$

is well-defined. Suppose we write

$$Y_1 = h_1^{-1}(U_1, U_2) \qquad Y_2 = h_2^{-1}(U_1, U_2).$$

Multivariate transformations (cont.)

Then we may derive that for $(u_1, u_2) \in \mathbb{R}^2$,

$$f_{U_1,U_2}(u_1, u_2) = f_{Y_1,Y_2}(h_1^{-1}(u_1, u_2), h_2^{-1}(u_1, u_2))|J(u_1, u_2)|$$

where $J(u_1, u_2)$ is the *Jacobian* of the transformation, that is

$$J(u_1, u_2) = \det \begin{bmatrix} \dfrac{\partial h_1^{-1}(u_1, u_2)}{\partial u_1} & \dfrac{\partial h_1^{-1}(u_1, u_2)}{\partial u_2} \\ \dfrac{\partial h_2^{-1}(u_1, u_2)}{\partial u_1} & \dfrac{\partial h_2^{-1}(u_1, u_2)}{\partial u_2} \end{bmatrix}$$

where det[.] denotes determinant, and |.| denotes absolute value.

Multivariate transformations (cont.)

This result holds provided that all the partial derivatives exist and are continuous, and that

$$|J(u_1, u_2)| > 0.$$

The result extends beyond two dimensions.

Chapter 5: Probability inequalities and theorems

Markov's inequality

Suppose that $g(Y)$ is a function of continuous random variable Y such that

$$\mathbb{E}[|g(Y)|] < \infty.$$

For any $k > 0$, let

$$\mathcal{A}_k = \{y \in \mathbb{R} : |g(y)| \geq k\}.$$

Markov's inequality (cont.)

Then

$$\begin{aligned}
\mathbb{E}[|g(Y)|] &= \int_{-\infty}^{\infty} |g(y)| f_Y(y)\, dy \\
&= \int_{\mathcal{A}_k} |g(y)| f_Y(y)\, dy + \int_{\mathcal{A}_k'} |g(y)| f_Y(y)\, dy \\
&\geq \int_{\mathcal{A}_k} |g(y)| f_Y(y)\, dy \\
&\geq \int_{\mathcal{A}_k} k f_Y(y)\, dy \\
&= k \int_{\mathcal{A}_k} f_Y(y)\, dy \\
&= k P(|g(Y)| \geq k)
\end{aligned}$$

Markov's inequality (cont.)

Thus

$$\mathbb{E}[|g(Y)|] \geq kP(|g(Y)| \geq k)$$

or equivalently

$$P(|g(Y)| \geq k) \leq \frac{\mathbb{E}[|g(Y)|]}{k}.$$

Markov's inequality (cont.)

Thus

$$1 - P(|g(Y)| \geq k) \geq 1 - \frac{\mathbb{E}[|g(Y)|]}{k}$$

or equivalently

$$P(|g(Y)| < k) \geq 1 - \frac{\mathbb{E}[|g(Y)|]}{k}.$$

Note that the result is identical for a discrete random variable.

Chebychev's inequality

aka

- Chebyshev's inequality
- Bienaymé-Chebyshev inequality
- Tchebysheff's Theorem
- Chebysheff, Chebychov, Chebyshov, Tchebychev ...

Chebychev's inequality (cont.)

Suppose that Y has finite mean μ and variance σ^2. In Markov's inequality, consider the function

$$g(t) = (t - \mu)^2.$$

We have for any $k > 0$ that

$$P(|(Y - \mu)^2| < k) \geq 1 - \frac{\mathbb{E}[|(Y - \mu)^2|]}{k}.$$

which can be more simply written as

$$P((Y - \mu)^2 < k) \geq 1 - \frac{\mathbb{E}[(Y - \mu)^2]}{k}.$$

Chebychev's inequality (cont.)

But

$$\mathbb{E}[(Y-\mu)^2] = \sigma^2$$

and if we choose to write $k = c^2\sigma^2$, the inequality becomes

$$P((Y-\mu)^2 < c^2\sigma^2) \geq 1 - \frac{\mathbb{E}[(Y-\mu)^2]}{c^2\sigma^2} = 1 - \frac{1}{c^2}$$

Chebychev's inequality (cont.)

The left hand side of the inequality is equal to

$$P(\sqrt{(Y-\mu)^2} < \sqrt{c^2\sigma^2}) = P(|Y-\mu| < c\sigma)$$

so the inequality becomes

$$P(|Y-\mu| < c\sigma) \geq 1 - \frac{1}{c^2}.$$

Chebychev's inequality (cont.)

An equivalent way to write this is

$$P(|Y - \mu| \geq c\sigma) \leq \frac{1}{c^2}$$

which holds for arbitrary $c > 0$.

Chebychev's inequality (cont.)

This states that for **any** distribution with finite mean μ and variance σ^2, we have that

$$P(\mu - c\sigma < Y < \mu + c\sigma) \geq 1 - \frac{1}{c^2}$$

which bounds how much probability can be present in the tails of the distribution.

The sample mean

Suppose that $Y_1, \ldots, Y_n$ are independent random variables that have the same distribution, and have expectation μ and variance σ^2. Recall that if

$$U = \sum_{i=1}^{n} Y_i$$

then

$$\mathbb{E}[U] = n\mu \qquad\qquad \mathbb{V}[U] = n\sigma^2.$$

The sample mean (cont.)

Let

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

be the *sample mean* random variable. Then

$$\mathbb{E}[\overline{Y}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}n\mu = \mu$$

and

$$\mathbb{V}[\overline{Y}_n] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

The sample mean (cont.)

Now recall Chebychev's inequality:

$$P(|Y - \mu| \geq c\sigma) \leq \frac{1}{c^2}.$$

Applying this to $\overline{Y}_n$, we have that

$$P\left(|\overline{Y}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right) \leq \frac{1}{c^2}.$$

The sample mean (cont.)

We can rewrite this for any n by setting

$$\epsilon = \frac{c\sigma}{\sqrt{n}} \quad \text{so that} \quad c = \frac{\sqrt{n}\epsilon}{\sigma}.$$

Then the inequality becomes

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Convergence in probability

Consider a sequence of random variables $Y_1, Y_2, \ldots, Y_n, \ldots$, and real constant c. We say that the sequence *converges in probability* to c as $n \longrightarrow \infty$ if, for every $\epsilon > 0$,

$$\lim_{n \longrightarrow \infty} P(|Y_n - c| < \epsilon) = 1$$

that is, the probability

$$P(c - \epsilon < Y_n < c + \epsilon)$$

converges to 1 as $n \longrightarrow \infty$. We write

$$Y_n \xrightarrow{p} c$$

as $n \longrightarrow \infty$ to indicate convergence in probability.

Convergence in probability (cont.)

Note

- If, for a fixed $\epsilon > 0$,

$$p_n = P(|Y_n - c| < \epsilon)$$

the definition requires that the real sequence $\{p_n\}$ converges to 1 in the usual sense.

- There is no restriction on the joint distribution of the Y_is in the sequence (that is, they can be dependent).
- We can equivalently write the condition

$$\lim_{n \longrightarrow \infty} P(|Y_n - c| > \epsilon) = 0.$$

The weak law of large numbers

Suppose that $Y_1, Y_2, \ldots, Y_n$, are independent random variables that have the same distribution, and have expectation μ and variance σ^2. For each n, let

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

be the sample mean random variable. Then

$$\overline{Y}_n \xrightarrow{p} \mu$$

as $n \longrightarrow \infty$.

The weak law of large numbers (cont.)

This follows from Chebychev's inequality: recall that for the sample mean

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

or equivalently

$$P(|\overline{Y}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

for any $\epsilon > 0$.

The weak law of large numbers (cont.)

But

$$1 - \frac{\sigma^2}{n\epsilon^2} \longrightarrow 1$$

as $n \longrightarrow \infty$, as σ and ϵ are fixed. Thus

$$\lim_{n \longrightarrow \infty} P(|\overline{Y}_n - \mu| < \epsilon) \geq 1$$

and hence $\overline{Y}_n \xrightarrow{p} \mu$.

The weak law of large numbers (cont.)

What this result means: under the stated conditions, the sample mean random variable converges to the expectation of the distribution of the Ys.

For example, if you repeatedly and independently toss a fair coin, the proportion of heads will converge to 1/2 as the number of tosses increases.

The weak law of large numbers (cont.)

Note

The statement

$$P(|\overline{Y}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

tells us that the *rate of convergence* is (at least) n;

- that is, the probability on the left hand side converges to 1 at rate (at least) n.

Convergence in distribution

Consider the sequence of cdfs, $F_1(y), F_2(y), \ldots, F_n(y), \ldots$, and suppose that

$$\lim_{n \longrightarrow \infty} F_n(y) = F(y)$$

for some other cdf $F(y)$ at all points where $F(y)$ is **continuous**.

Then $F(y)$ is the *limiting distribution* for the sequence.

Convergence in distribution (cont.)

If $Y_1, Y_2, \ldots, Y_n, \ldots$ are a sequence of random variables having cdfs $F_1(y), F_2(y), \ldots, F_n(y), \ldots$ respectively, and Y has cdf $F(y)$, then we say that the sequence $Y_1, Y_2, \ldots, Y_n, \ldots$ *converges in distribution* to Y,

$$Y_n \xrightarrow{d} Y$$

as $n \longrightarrow \infty$.

The central limit theorem

Suppose that $Y_1, Y_2, \ldots, Y_n$, are independent random variables that have the same distribution, and have finite expectation μ and variance σ^2. For each n, let

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

be the sample mean random variable.

The central limit theorem (cont.)

Then the random variable

$$U_n = \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma}$$

converges in distribution to a *standard normal distribution.*

That is,

$$\lim_{n \longrightarrow \infty} F_{U_n}(u) = \lim_{n \longrightarrow \infty} P(U_n \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

This result is the *Central Limit Theorem* (CLT).

The central limit theorem (cont.)

Note that we could write

$$U_n = \frac{(\overline{Y}_n - \mu)}{\sigma/\sqrt{n}}$$

or, multiplying numerator and denominator by n

$$U_n = \frac{\left(\sum_{i=1}^{n} Y_i - n\mu\right)}{\sqrt{n}\sigma}$$

The central limit theorem (cont.)

The result says that, provided the conditions on expectation and variance are met, when n is large, we can approximate the distribution of $\overline{Y}_n$, or of

$$S_n = \sum_{i=1}^{n} Y_i$$

using a Normal distribution, irrespective of the actual distribution of $Y_1, \ldots, Y_n$.

The central limit theorem (cont.)

That is,

- $\overline{Y}_n$ is approximately distributed as

$$Normal(\mu, \sigma^2/n)$$

- S_n is approximately distributed as

$$Normal(n\mu, n\sigma^2)$$

The central limit theorem (cont.)

Note

If $Y_1, \ldots, Y_n$ are independent and identically distributed Normal random variables with expectation μ and variance σ^2, then

- $\overline{Y}_n$ is exactly distributed as

$$Normal(\mu, \sigma^2/n)$$

- S_n is exactly distributed as

$$Normal(n\mu, n\sigma^2)$$

The central limit theorem (cont.)

The result is straightforward to prove using moment-generating functions. For $i = 1, 2, \ldots, n$, let

$$Z_i = \frac{Y_i - \mu}{\sigma}.$$

Then from properties of expectations and variances, we have that

$$\mathbb{E}[Z_i] = 0 \qquad \mathbb{V}[Z_i] = 1.$$

and also

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i$$

The central limit theorem (cont.)

If we assume that $m(t)$ is the mgf of each Y_i, and $m_Z(t)$ is the mgf for each Z_i, then by previous mgf results

$$m_{U_n}(t) = \left\{ m_Z\left(\frac{t}{\sqrt{n}}\right) \right\}^n$$

Now, using a Taylor expansion of $m_Z(t)$ at $t = 0$

$$m_Z(t) = m_Z(0) + t\, m_Z^{(1)}(0) + \frac{t^2}{2}\, m_Z^{(2)}(t_0)$$

where $0 < t_0 < t$ yields the remainder term.

The central limit theorem (cont.)

Now as $\mathbb{E}[Z_i] = 0$, and $\mathbb{V}[Z_i] = 1$ we can deduce that

$$m_Z(0) = 1 \qquad m_Z^{(1)}(0) = \mathbb{E}[Z_i] = 0$$

and hence

$$m_Z(t) = 1 + \frac{t^2}{2} m_Z^{(2)}(t_0).$$

The central limit theorem (cont.)

Therefore

$$m_{U_n}(t) = \left\{1 + \frac{t^2}{2n}\, m_Z^{(2)}(t_0)\right\}^n \qquad 0 < t_0 < \frac{t}{\sqrt{n}}$$

Now, as $n \longrightarrow \infty$

$$\frac{t}{\sqrt{n}} \longrightarrow 0$$

so $t_0 \longrightarrow 0$.

The central limit theorem (cont.)

Hence

$$\begin{aligned}\lim_{n\longrightarrow\infty} m_{U_n}(t) &= \lim_{n\longrightarrow\infty}\left\{1+\frac{t^2}{2n}\,m_Z^{(2)}(t_0)\right\}^n \\ &= \lim_{n\longrightarrow\infty}\left\{1+\frac{t^2}{2n}\,m_Z^{(2)}(0)\right\}^n\end{aligned}$$

The central limit theorem (cont.)

But $m_Z^{(2)}(0) = \mathbb{E}[Z_i^2] = 1$, so therefore

$$\lim_{n \longrightarrow \infty} m_{U_n}(t) = \lim_{n \longrightarrow \infty} \left\{ 1 + \frac{t^2}{2n} \right\}^n = e^{t^2/2}$$

by definition of the exponential function. Hence the limiting mgf is that of a standard Normal distribution, and so

$$U_n \xrightarrow{d} Z$$

where Z has a standard Normal distribution.

Applications of the central limit theorem

We can use the central limit theorem to approximate some standard distributions for which it is otherwise awkward to compute probabilities.

Suppose we wish to approximate the distribution of random variable Y. We must first attempt to decompose Y as

$$Y = \sum_{i=1}^{n} Y_i$$

where $Y_1, \ldots, Y_n$ are independent and identically distributed, and have expectation μ and variance σ^2 that can be easily computed.

Applications of the central limit theorem (cont.)

We then can say that Y is approximately distributed as

$$Normal(n\mu, n\sigma^2).$$

and compute, for example

$$F_Y(y) = \Pr(Y \leq y) = P\left(\frac{Y - n\mu}{\sqrt{n}\sigma} \leq \frac{y - n\mu}{\sqrt{n}\sigma}\right) \doteq \Phi\left(\frac{y - n\mu}{\sqrt{n}\sigma}\right)$$

where $\Phi(.)$ is the standard Normal cdf.

Applications of the central limit theorem (cont.)

Example (Binomial case)

If $Y \sim Binomial(n, p)$, the mgf of Y takes the form

$$m_Y(t) = (pe^t + q)^n$$

By the uniqueness of mgfs, we recognize that Y has the representation

$$Y = \sum_{i=1}^{n} Y_i$$

where $Y_i \sim Binomial(1, p) \equiv Bernoulli(p)$ are independent. We have

$$\mathbb{E}[Y_i] = p \qquad \mathbb{V}[Y_i] = pq.$$

Applications of the central limit theorem (cont.)

Example (Binomial case)

Therefore for large n, Y is approximately distributed as

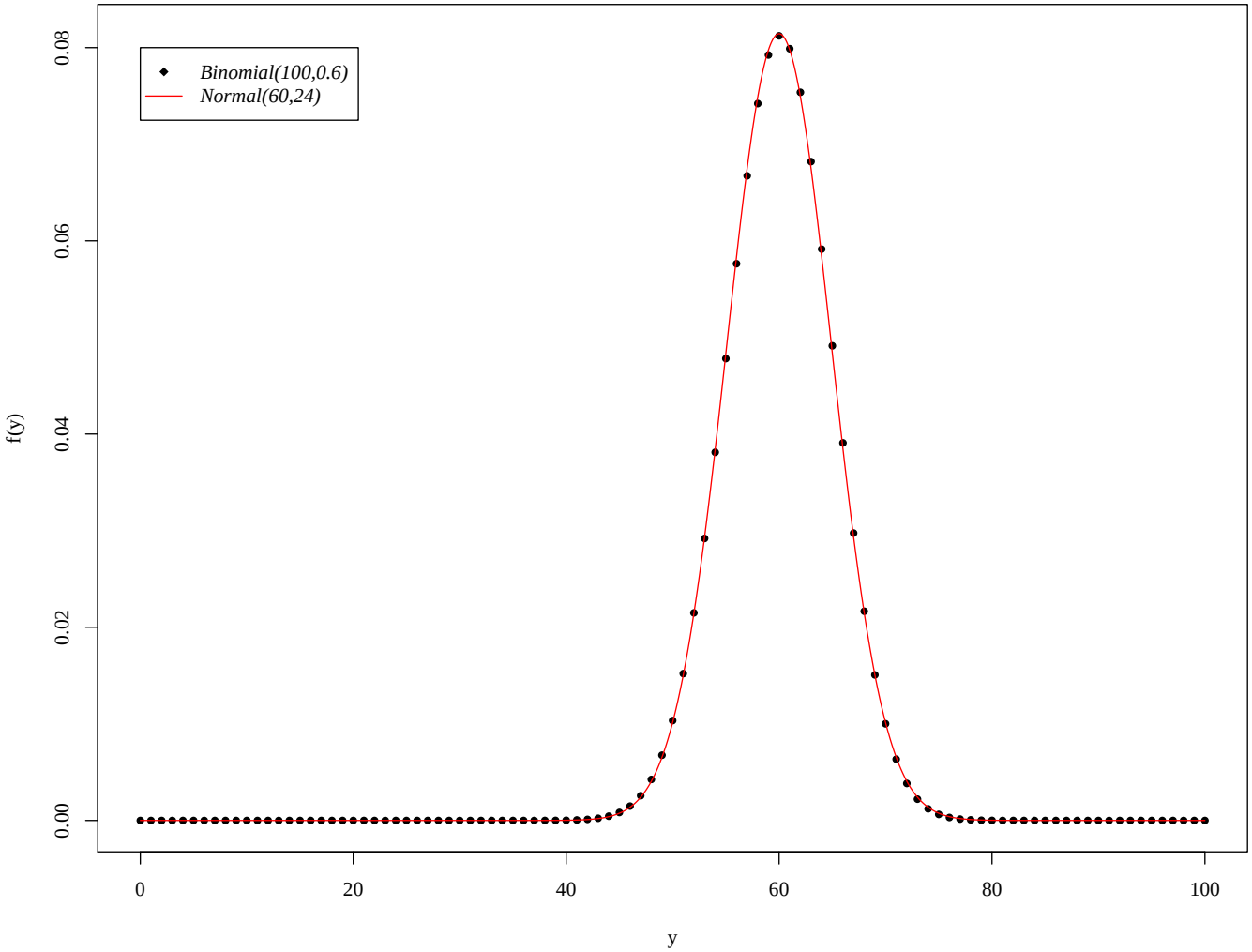
$$Normal(np, npq)$$

and

$$F_Y(y) = P(Y \leq y) \doteq \Phi\left(\frac{y - np}{\sqrt{npq}}\right)$$

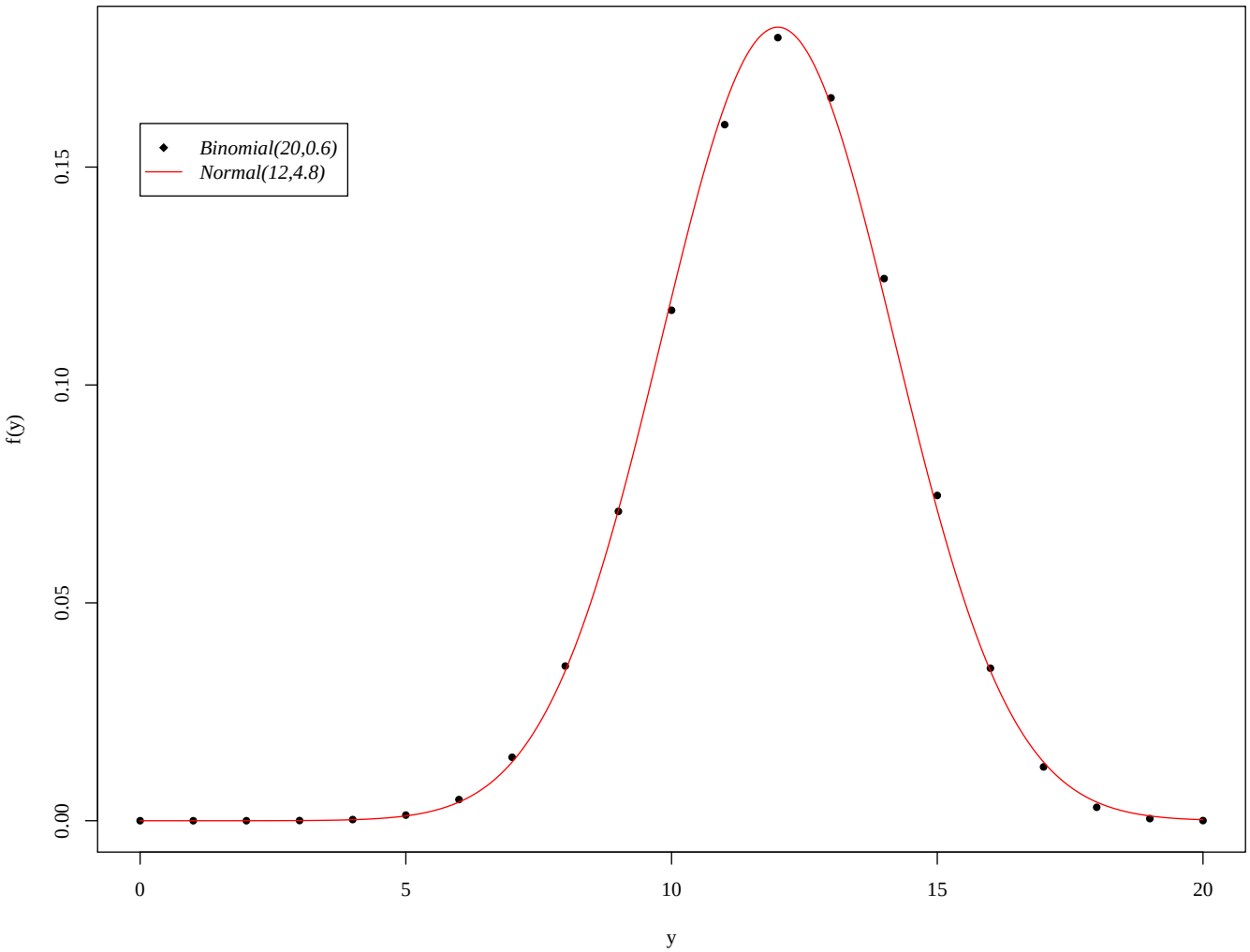
Applications of the central limit theorem (cont.)

Binomial: $n = 100$, with $p = 0.6$.



Applications of the central limit theorem (cont.)

Binomial: $n = 20$, with $p = 0.6$.



Applications of the central limit theorem (cont.)

Example (Poisson case)

If $Y \sim Poisson(\lambda)$, the mgf of Y takes the form

$$m_Y(t) = \exp\{\lambda(e^t - 1)\}.$$

By the uniqueness of mgfs, we recognize that Y has the representation

$$Y = \sum_{i=1}^{n} Y_i$$

where the $Y_i \sim Poisson(\lambda/n)$ are independent. We have

$$\mathbb{E}[Y_i] = \frac{\lambda}{n} \qquad \mathbb{V}[Y_i] = \frac{\lambda}{n}.$$

Applications of the central limit theorem (cont.)

Example (Poisson case)

Therefore Y is approximately distributed as

$$Normal(\lambda, \lambda)$$

and

$$F_Y(y) = P(Y \leq y) \simeq \Phi\left(\frac{y-\lambda}{\sqrt{\lambda}}\right)$$

Note that for this approximation to work, we need to have λ large.

Applications of the central limit theorem (cont.)

Continuity correction

When we use the central limit theorem to approximate a discrete distribution, a *continuity correction* is sometimes applied: that is, if Y is discrete we compute

$$F_Y(y) = P(Y \leq y) \simeq \Phi\left(\frac{y + \frac{1}{2} - n\mu}{\sqrt{n}\sigma}\right)$$

Applications of the central limit theorem (cont.)

Example (Gamma case)

If $Y \sim Gamma(\alpha, \beta)$, the mgf of Y takes the form

$$m_Y(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}$$

By the uniqueness of mgfs, we recognize that Y has the representation

$$Y = \sum_{i=1}^{n} Y_i$$

where the $Y_i \sim Gamma(\alpha/n, \beta)$ are independent. We have

$$\mathbb{E}[Y_i] = \frac{\alpha\beta}{n} \qquad \mathbb{V}[Y_i] = \frac{\alpha\beta^2}{n}.$$

Applications of the central limit theorem (cont.)

Example (Gamma case)

Therefore Y is approximately distributed as

$$Normal(\alpha\beta, \alpha\beta^2)$$

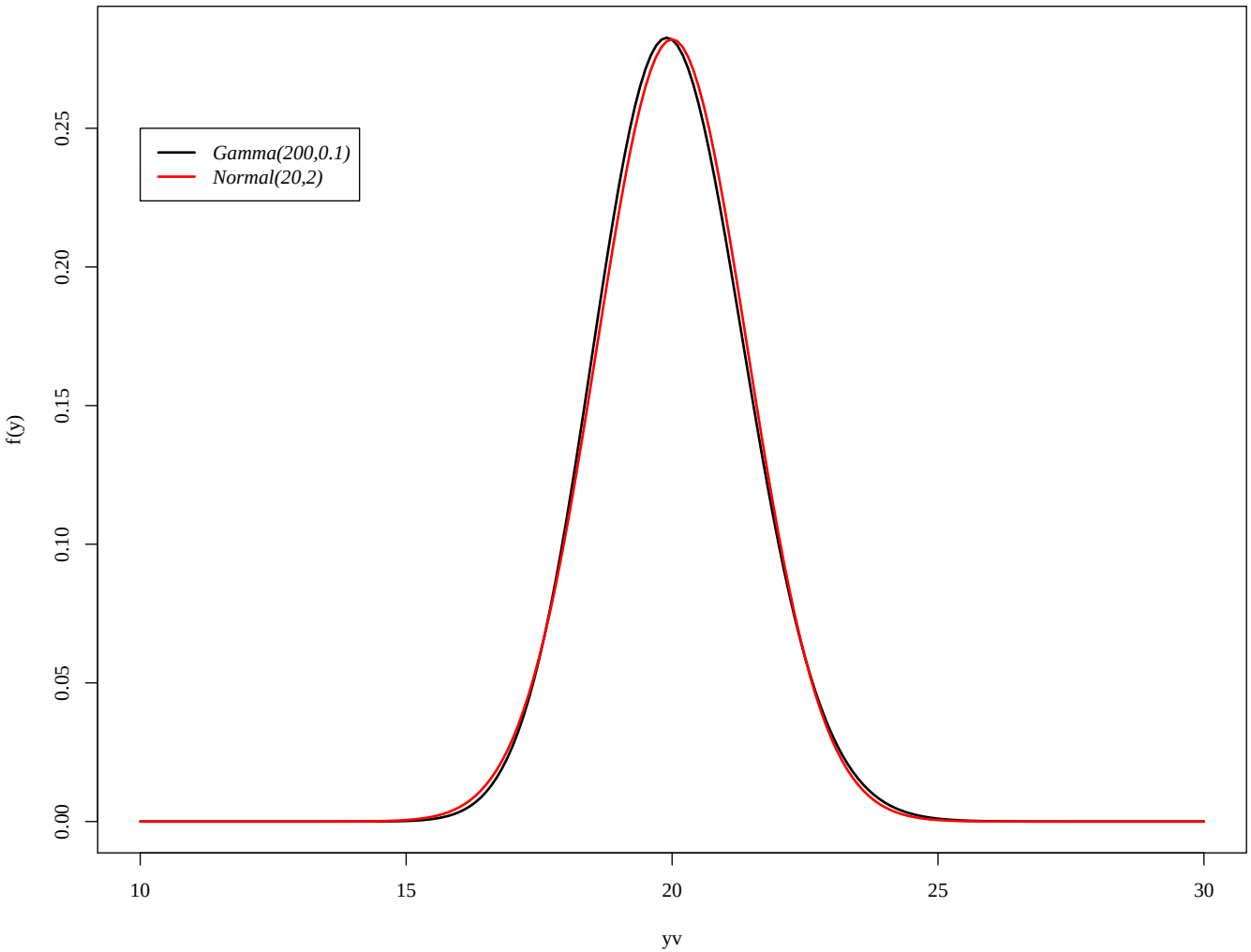
and

$$F_Y(y) = P(Y \leq y) \doteq \Phi\left(\frac{y - \alpha\beta}{\sqrt{\alpha\beta^2}}\right)$$

Note that for this approximation to work, we need to have α large.

Applications of the central limit theorem (cont.)

Gamma: $\alpha = 200$, with $\beta = 0.1$.



Applications of the central limit theorem (cont.)

Gamma: $\alpha = 50$, with $\beta = 0.4$.

