



Probability 1 (MATH 11300) lecture slides

Márton Balázs

School of Mathematics
University of Bristol
Autumn, 2016

December 1, 2016

To know...

- ▶ <http://www.maths.bris.ac.uk/~mb13434/probl/>
- ▶ m.balazs@bristol.ac.uk
- ▶ Drop in Sessions: Tuesdays 5pm, office 3.7 in the Maths bld.
- ▶ 22+2 lectures, 12 exercise classes, 11 mandatory HW sets.
- ▶ HW on two weeks is assessed and counts 10% towards final mark.
- ▶ These slides have gaps, **come to lectures.**  
- ▶ Core reading: *Probability 1*, Pearson Custom Publishing.
Compiled from A First Course in Probability by S. Ross.
- ▶ **Probability is difficult, but interesting, useful, and fun.**
- ▶ **Not-to-hand-in extra problem sheets for those interested. They won't affect your marks in any way.**
- ▶ This material is copyright of the University unless explicitly stated otherwise. It is provided exclusively for educational purposes at the University and is to be downloaded or copied for your private study only.

1. Elementary probability
2. Conditional probability
3. Discrete random variables
4. Continuous random variables
5. Joint distributions
6. Expectation, covariance
7. Law of Large Numbers, Central Limit Theorem

1. Elementary probability

Combinatorics

Sample space

Probability

Equally likely outcomes

Objectives:

- ▶ To define events and sample spaces, describe them in simple examples
- ▶ To list the axioms of probability, and use them to prove simple results
- ▶ To use counting arguments to calculate probabilities when there are equally likely outcomes

Combinatorics / the basic principle of counting

In this part we'll learn how to count in some typical scenarios. The starting point is the following:

- ▶ Suppose an experiment has n outcomes; and another experiment has m outcomes.
- ▶ Then the two experiments jointly have $n \cdot m$ outcomes.

Example

Rolling a die and flipping a coin can have a total of $6 \cdot 2 = 12$ different outcomes, combined.

1. Permutations

Definition

Let $H = \{h_1, h_2, \dots, h_n\}$ be a set of n different objects. The permutations of H are the different orders in which one can write all of the elements of H . There are $n! = 1 \cdot 2 \cdot 3 \cdots n$ of them. We set $0! = 1$.



1. Permutations

Definition

Let $H = \{h_1, h_2, \dots, h_n\}$ be a set of n different objects. The permutations of H are the different orders in which one can write all of the elements of H . There are $n! = 1 \cdot 2 \cdot 3 \cdots n$ of them. We set $0! = 1$.



Example

The results of a horse race with horses $H = \{A, B, C, D, E, F, G\}$ are permutations of H . A possible outcome is (E, G, A, C, B, D, F) (E is the winner, G is second, etc.). There are $7! = 5\,040$ possible outcomes.

2. Permutations with repetitions

Definition

Let $H = \{h_1 \dots h_1, h_2 \dots h_2, \dots, h_r \dots h_r\}$ be a set of r different types of **repeated** objects: n_1 many of h_1 , n_2 of h_2 , \dots n_r of h_r . The permutations with repetitions of H are the different orders in which one can write all of the elements of H . There are

$$\binom{n}{n_1, n_2, \dots, n_r} := \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_r!}$$

of them, where $n = n_1 + \dots + n_r$ is the total number of objects. This formula is also known as the *multinomial coefficient*.



2. Permutations with repetitions

Example

We can make

$$\binom{11}{5, 2, 2, 1, 1} = \frac{11!}{5! \cdot 2! \cdot 2! \cdot 1! \cdot 1!} = 83\,160$$

different words out of the letters A, B, R, A, C, A, D, A, B, R, A.

3. k -permutations

Definition

Let $H = \{h_1, h_2, \dots, h_n\}$ be a set of n different objects. The k -permutations of H are the different ways in which one can pick and write k of the elements of H in order. There are $\frac{n!}{(n-k)!}$ of these k -permutations.



3. k -permutations

Definition

Let $H = \{h_1, h_2, \dots, h_n\}$ be a set of n different objects. The k -permutations of H are the different ways in which one can pick and write k of the elements of H **in order**. There are $\frac{n!}{(n-k)!}$ of these k -permutations.



Example

The first three places of a horse race with horses $H = \{A, B, C, D, E, F, G\}$ form a 3-permutation of H . A possible outcome is (E, G, A) (E is the winner, G is second, A is third.). There are $\frac{7!}{(7-3)!} = 210$ possible outcomes for the first three places.

4. k -permutations with repetitions

Definition

Let $H = \{h_1 \dots, h_2 \dots, \dots, h_r \dots\}$ be a set of r different types of **repeated** objects, **each of infinite supply**. The k -permutations with repetitions of H are the different orders in which one can write an ordered sequence of length k using the elements of H . There are r^k such sequences.



4. k -permutations with repetitions

Definition

Let $H = \{h_1 \dots, h_2 \dots, \dots, h_r \dots\}$ be a set of r different types of **repeated** objects, **each of infinite supply**. The k -permutations with repetitions of H are the different orders in which one can write an ordered sequence of length k using the elements of H . There are r^k such sequences.



(The case when the elements of H are only of finite supply is much more complicated.)

4. k -permutations with repetitions

Definition

Let $H = \{h_1 \dots, h_2 \dots, \dots, h_r \dots\}$ be a set of r different types of **repeated** objects, **each of infinite supply**. The k -permutations with repetitions of H are the different orders in which one can write an ordered sequence of length k using the elements of H . There are r^k such sequences.



(The case when the elements of H are only of finite supply is much more complicated.)

Example

There are $26^3 = 17576$ possible $k = 3$ -letter words using the $r = 26$ letters of the English alphabet.

5. k -combinations

Definition

Let $H = \{h_1, h_2, \dots, h_n\}$ be a set of n different objects. The k -combinations of H are the different ways in which one can pick k of the elements of H **without order**. There are

$$\binom{n}{k} := \frac{n!}{k! \cdot (n-k)!}$$

of these k -combinations. This formula is also known as the *binomial coefficient* (" n choose k ").



5. k -combinations

Example

There are

$$\binom{30}{5} = \frac{30!}{5! \cdot (30 - 5)!} = 142\,506$$

possible ways to form a committee of 5 students out of a class of 30 students.

5. k -combinations

Example

There are

$$\binom{30}{5} = \frac{30!}{5! \cdot (30 - 5)!} = 142\,506$$

possible ways to form a committee of 5 students out of a class of 30 students.

Remark

In a similar way, there are

$$\binom{n}{k_1, k_2, \dots, k_r} := \frac{n!}{k_1! \cdot k_2! \cdot \dots \cdot k_r!}$$

many ways to form **unordered** groups of sizes k_1, k_2, \dots, k_r of n objects ($n = k_1 + \dots + k_r$). Thus, the multinomial coefficient generalizes the binomial coefficient.

6. The Binomial coefficient

- ▶ Recall the definition, for n, k non-negative integers,

$$\binom{n}{k} := \frac{n!}{k! \cdot (n-k)!} = \binom{n}{n-k} \quad \text{for } 0 \leq k \leq n.$$

We extend this by $\binom{n}{k} \equiv 0$ in all other cases. (It is possible to define these coefficients for any real n , but we won't need that.)

6. The Binomial coefficient

Theorem (Pascal's Identity)

For any k and $1 \leq n$ integers,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

6. The Binomial coefficient

Theorem (Pascal's Identity)

For any k and $1 \leq n$ integers,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Proof.

Either write out the factorials, or count the number of k -combinations of n objects in two ways:

- ▶ the first object is chosen, and the remaining $k - 1$ objects need to be picked out of $n - 1$, or
- ▶ the first object is *not* chosen, and all k objects need to be picked out of $n - 1$.



7. The Binomial Theorem

Theorem (Newton's Binomial Theorem)

For any real numbers x and y , and $n \geq 1$, we have

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} \cdot x^k \cdot y^{n-k}.$$

7. The Binomial Theorem

Theorem (Newton's Binomial Theorem)

For any real numbers x and y , and $n \geq 1$, we have

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} \cdot x^k \cdot y^{n-k}.$$

Proof.

In the product $(x + y) \cdots (x + y)$ on the left hand-side we need to pick x or y from each parenthesis in all possible ways, and multiply them. Picking x from k of these parentheses and y from the remaining $n - k$ can be done in $\binom{n}{k}$ ways, each of which contributes $x^k \cdot y^{n-k}$ to the final sum. □

8. The Multinomial Theorem

The Binomial Theorem generalises to

Theorem (Multinomial Theorem)

Let x_1, x_2, \dots, x_r be real numbers, $n \geq 1$. Then

$$(x_1 + x_2 + \dots + x_r)^n = \sum_{\substack{n_1, \dots, n_r \geq 0 \\ n_1 + \dots + n_r = n}} \binom{n}{n_1 \ n_2 \ \dots \ n_r} \cdot x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_r^{n_r}.$$

Sample space

Here we are (almost) going to define a mathematical model for various experiments. To do it properly, we would need some tools from *measure theory*. This will be skipped for now, but you are welcome to revisit this point some time later during your studies!

Sample space

Here we are (almost) going to define a mathematical model for various experiments. To do it properly, we would need some tools from *measure theory*. This will be skipped for now, but you are welcome to revisit this point some time later during your studies!

- ▶ We always consider an *experiment*. Ω will denote the set of all possible outcomes of this experiment.

Sample space

Here we are (almost) going to define a mathematical model for various experiments. To do it properly, we would need some tools from *measure theory*. This will be skipped for now, but you are welcome to revisit this point some time later during your studies!

- ▶ We always consider an *experiment*. Ω will denote the set of all possible outcomes of this experiment.
- ▶ An event will be a collection of possible outcomes. Therefore, an event E will be considered a subset of Ω :
 $E \subseteq \Omega$.

Sample space

Here we are (almost) going to define a mathematical model for various experiments. To do it properly, we would need some tools from *measure theory*. This will be skipped for now, but you are welcome to revisit this point some time later during your studies!

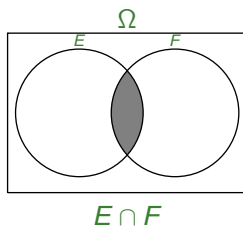
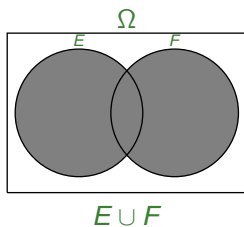
- ▶ We always consider an *experiment*. Ω will denote the set of all possible outcomes of this experiment.
- ▶ An event will be a collection of possible outcomes. Therefore, an event E will be considered a subset of Ω :
 $E \subseteq \Omega$.
- ▶ Sometimes Ω is too large, and not all its subsets can be defined as events. This is where measure theory helps...

Sample space

Here we are (almost) going to define a mathematical model for various experiments. To do it properly, we would need some tools from *measure theory*. This will be skipped for now, but you are welcome to revisit this point some time later during your studies!

- ▶ We always consider an *experiment*. Ω will denote the set of all possible outcomes of this experiment.
- ▶ An event will be a collection of possible outcomes. Therefore, an event E will be considered a subset of Ω :
 $E \subseteq \Omega$.
- ▶ Sometimes Ω is too large, and not all its subsets can be defined as events. This is where measure theory helps...
- ▶ It makes perfect sense to define the union $E \cup F$ and the intersection $E \cap F$ of two events, E and F .

Sample space



Notation: sometimes $E \cup F = E + F$, $E \cap F = EF$.

1. Examples

Example

Experiment: Is it going to rain today?

Sample space: $\Omega = \{r, n\}$.

$$|\Omega| = 2.$$

An event: $E = \{r\}$.

$$|E| = 1.$$

1. Examples

Example

Experiment: Finishing order of a race of 7 horses.

Sample space: $\Omega = \{\text{permutations of } \mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}, \mathcal{F}, \mathcal{G}\}.$
 $|\Omega| = 7!.$

An event: $E = \{\text{horse } \mathcal{B} \text{ wins}\}$
 $= \{\text{permutations that start with } \mathcal{B}\}.$
 $|E| = 6!.$

Another event: $F = \{\mathcal{G} \text{ wins, } \mathcal{D} \text{ is second}\}.$
 $= \{\text{permutations starting as } (\mathcal{G}, \mathcal{D}, \dots)\}.$
 $|F| = 5!.$

Notice $E \cap F = \emptyset$ in this example. We call \emptyset the null event. This is the event that never happens.

1. Examples

Example

Experiment: Flipping two coins.

Sample space: $\Omega = \{\text{ordered pairs of the two outcomes}\}.$
 $= \{(H, H), (H, T), (T, H), (T, T)\}.$
 $|\Omega| = 4.$

An event: $E = \{\text{the two coins come up different}\}$
 $= \{(H, T), (T, H)\}.$
 $|E| = 2.$

Another event: $F = \{\text{both flips come up heads}\}.$
 $= \{(H, H)\}.$
 $|F| = 1.$

Notice: $E \cup F = \{(H, T), (T, H), (H, H)\}$
 $= \{\text{at least one } H\}.$

1. Examples

Example

Experiment: Rolling two dice.

Sample space: $\Omega = \{\text{ordered pairs of the two outcomes}\}$
 $= \{(i, j) : i, j = 1 \dots 6\}.$

$$|\Omega| = 36.$$

An event: $E = \{\text{the sum of the rolls is 4}\}$
 $= \{(1, 3), (2, 2), (3, 1)\}.$

$$|E| = 3.$$

Another event: $F = \{\text{the two rolls are the same}\}.$
 $= \{(i, i) : i = 1 \dots 6\}.$


$$|F| = 6.$$

Notice: $E \cap F = \{(2, 2)\}.$

1. Examples

Example

Experiment: Repeatedly rolling a die until we first see 6.

Sample space: $\Omega = \{\text{sequences of numbers between 1}$
and 5, and then a 6\}. \rightsquigarrow 

$$|\Omega| = \infty.$$

An event: $E = \{\text{roll 4 first, get 6 on the third roll}\}$
 $= \{(4, k, 6) : k = 1 \dots 5\}.$

$$|E| = 5.$$

1. Examples

Example

Experiment: Lifetime of a device (measured in years).

Sample space: $\Omega = [0, \infty)$

$|\Omega| = \infty$ (uncountable).

An event: $E = \{\text{shouldn't have bought it}\} = \{0\}$

$|E| = 1$.

Another event: $F = \{\text{device lasts for at least 5 years}\}$
 $= [5, \infty)$.

$|F| = \infty$.

Another event: $G = \{\text{device is dead by its 6th birthday}\}$
 $= [0, 6)$.

$|G| = \infty$.

Notice: $F \cap G = [5, 6)$, $F \cup G = [0, \infty) = \Omega$.

2. The union and the intersection

Inspired by the above:

Remark

The **union** $E \cup F$ of events E and F always means **E OR F** .
The **intersection** $E \cap F$ of events E and F always means **E AND F** .

Similarly:

Remark

The **union** $\bigcup_i E_i$ of events E_i always means **at least one of the E_i 's**.
The **intersection** $\bigcap_i E_i$ of events E_i always means **each of the E_i 's**.

2. The union and the intersection

Definition

If $E \cap F = \emptyset$, then we say that the events E and F are mutually exclusive events.

If the events E_1, E_2, \dots satisfy $E_i \cap E_j = \emptyset$ whenever $i \neq j$, then we say that the E_i 's are mutually exclusive events.

Mutually exclusive events cannot happen at the same time.

3. Inclusion and implication

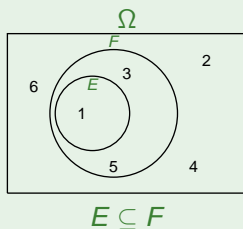
Remark

If the event E is a subset of the event F , $E \subseteq F$, then the occurrence of E implies that of F .

Example

The experiment is rolling a die.

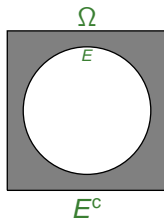
$E = \{\text{rolling 1 on a die}\} \subseteq \{\text{rolling an odd no. on a die}\} = F$.



4. Complementary events

Definition

The complement of an event E is $E^c := \Omega - E$. This is the event that E does *not* occur.



Notice: $E \cap E^c = \emptyset$, $E \cup E^c = \Omega$.

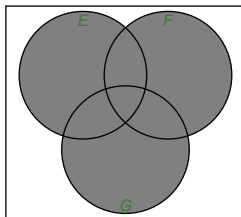
Notation: sometimes $E^c = \bar{E} = E^*$.

5. Simple properties of events

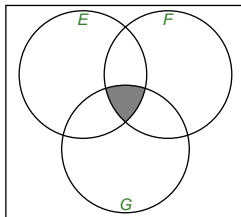
Commutativity: $E \cup F = F \cup E,$
 $E \cap F = F \cap E.$

5. Simple properties of events

Associativity: $E \cup (F \cup G) = (E \cup F) \cup G = E \cup F \cup G,$



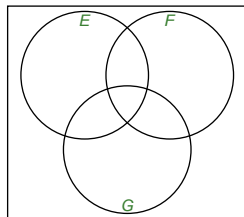
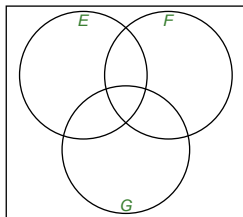
$E \cap (F \cap G) = (E \cap F) \cap G = E \cap F \cap G.$



5. Simple properties of events

Distributivity:

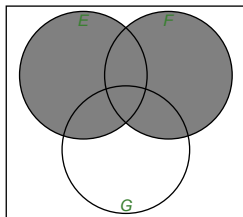
$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$



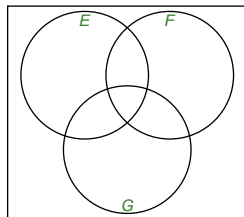
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$



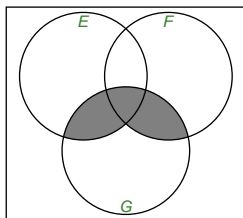
$E \cup F$



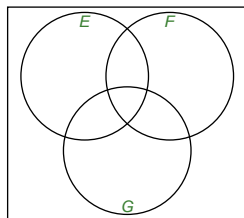
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$



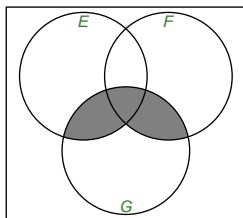
$$(E \cup F) \cap G$$



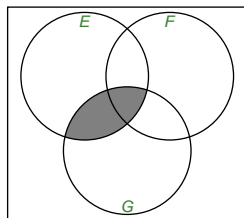
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$



$$(E \cup F) \cap G$$

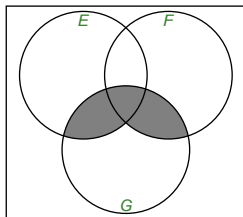


$$(E \cap G) \cup (F \cap G)$$

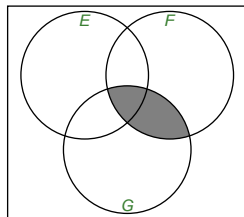
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$



$$(E \cup F) \cap G$$

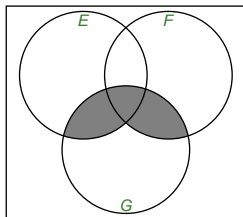


$$E \cap G \cup F \cap G$$

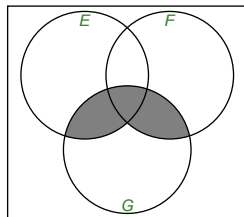
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$



$$(E \cup F) \cap G$$

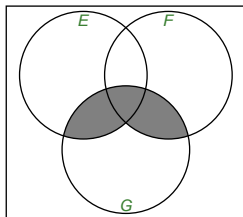


$$(E \cap G) \cup (F \cap G)$$

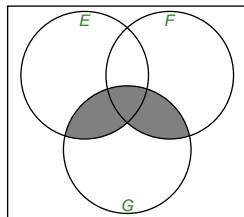
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$

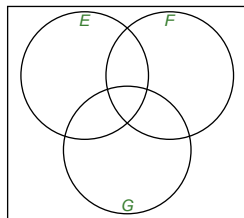
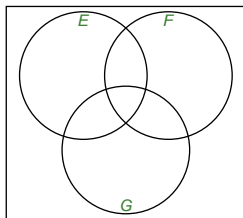


$$(E \cup F) \cap G$$



$$(E \cap G) \cup (F \cap G)$$

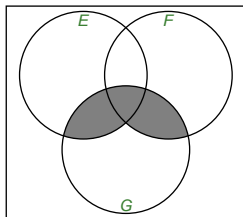
$$(E \cap F) \cup G = (E \cup G) \cap (F \cup G).$$



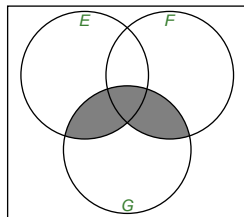
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$

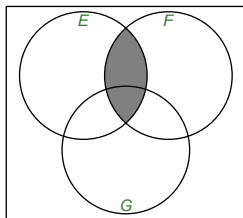


$$(E \cup F) \cap G$$

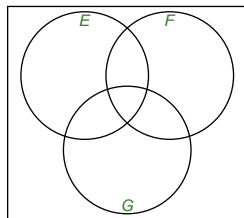


$$(E \cap G) \cup (F \cap G)$$

$$(E \cap F) \cup G = (E \cup G) \cap (F \cup G).$$



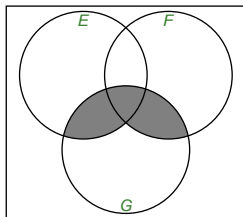
$$(E \cap F) \cup G$$



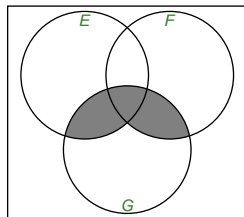
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$

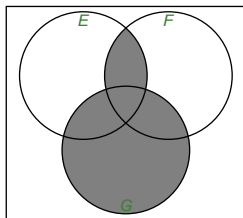


$$(E \cup F) \cap G$$

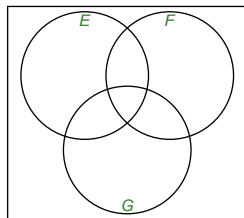


$$(E \cap G) \cup (F \cap G)$$

$$(E \cap F) \cup G = (E \cup G) \cap (F \cup G).$$



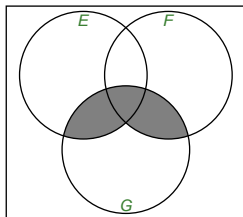
$$(E \cap F) \cup G$$



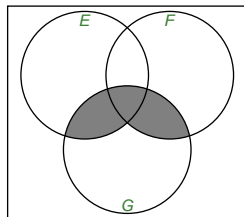
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$

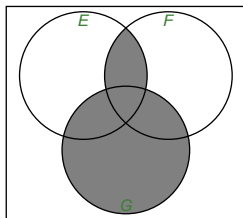


$$(E \cup F) \cap G$$

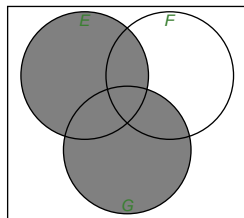


$$(E \cap G) \cup (F \cap G)$$

$$(E \cap F) \cup G = (E \cup G) \cap (F \cup G).$$



$$(E \cap F) \cup G$$

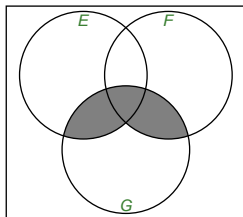


$$E \cup G$$

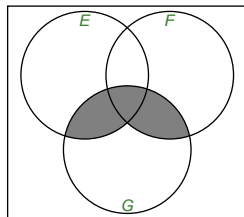
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$

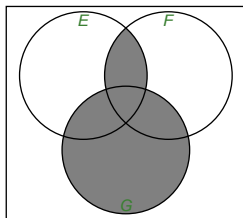


$$(E \cup F) \cap G$$

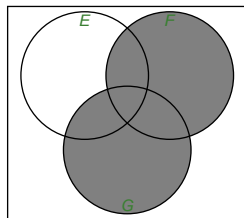


$$(E \cap G) \cup (F \cap G)$$

$$(E \cap F) \cup G = (E \cup G) \cap (F \cup G).$$



$$(E \cap F) \cup G$$

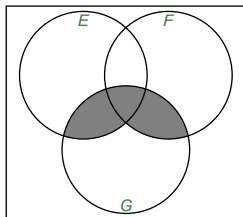


$$(E \cup G) \cap (F \cup G)$$

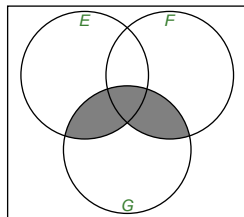
5. Simple properties of events

Distributivity:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G),$$

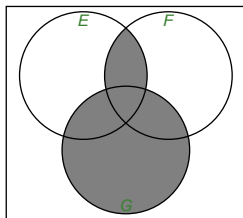


$$(E \cup F) \cap G$$

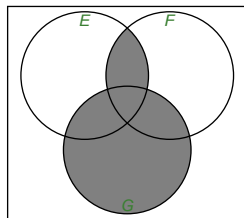


$$(E \cap G) \cup (F \cap G)$$

$$(E \cap F) \cup G = (E \cup G) \cap (F \cup G).$$



$$(E \cap F) \cup G$$

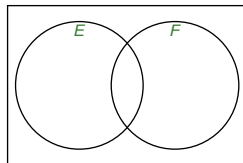
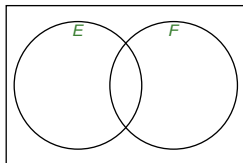


$$(E \cup G) \cap (F \cup G)$$

5. Simple properties of events

De Morgan's Law:

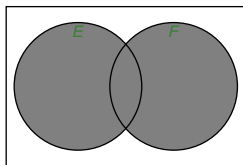
$$(E \cup F)^c = E^c \cap F^c.$$



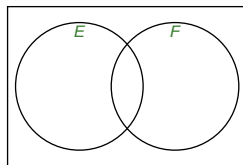
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$



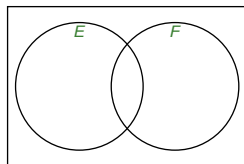
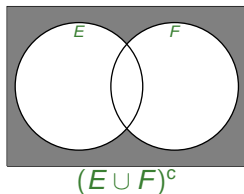
$E \cup F$



5. Simple properties of events

De Morgan's Law:

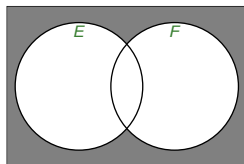
$$(E \cup F)^c = E^c \cap F^c.$$



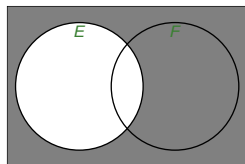
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$



$$(E \cup F)^c$$

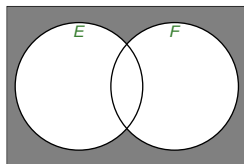


$$E^c$$

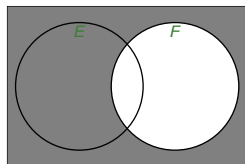
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$



$$(E \cup F)^c$$

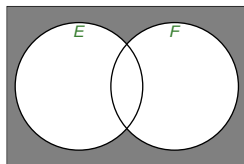


$$F^c$$

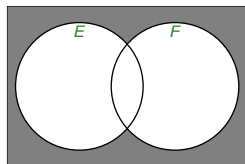
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$



$$(E \cup F)^c$$

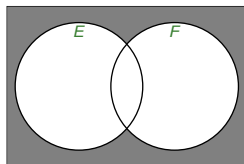


$$E^c \cap F^c$$

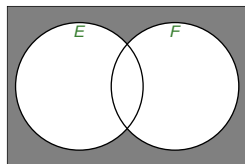
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$

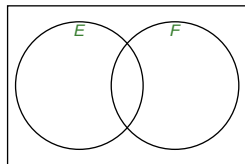
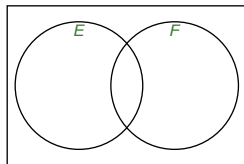


$$(E \cup F)^c$$



$$E^c \cap F^c$$

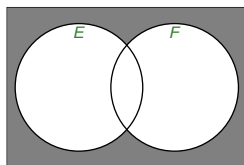
$$(E \cap F)^c = E^c \cup F^c.$$



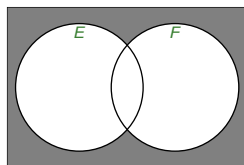
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$

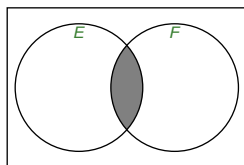


$$(E \cup F)^c$$

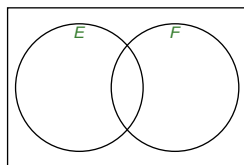


$$E^c \cap F^c$$

$$(E \cap F)^c = E^c \cup F^c.$$



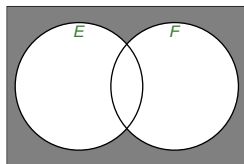
$$E \cap F$$



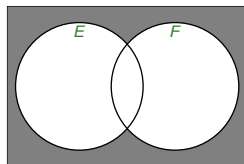
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$

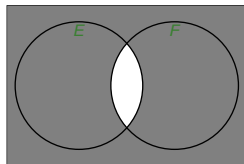


$$(E \cup F)^c$$

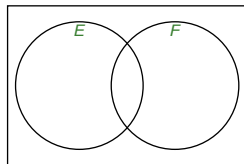


$$E^c \cap F^c$$

$$(E \cap F)^c = E^c \cup F^c.$$



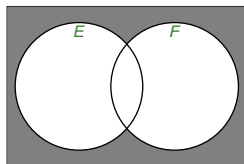
$$(E \cap F)^c$$



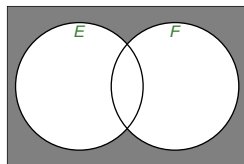
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$

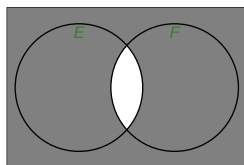


$$(E \cup F)^c$$

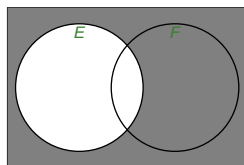


$$E^c \cap F^c$$

$$(E \cap F)^c = E^c \cup F^c.$$



$$(E \cap F)^c$$

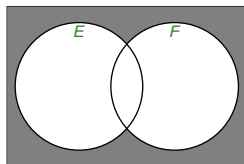


$$E^c \cup F^c$$

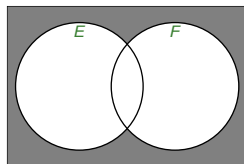
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$

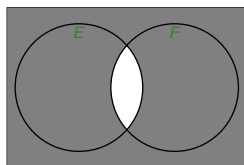


$$(E \cup F)^c$$

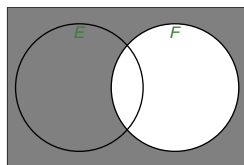


$$E^c \cap F^c$$

$$(E \cap F)^c = E^c \cup F^c.$$



$$(E \cap F)^c$$

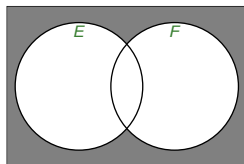


$$F^c$$

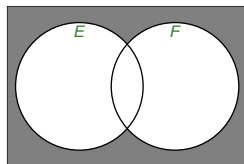
5. Simple properties of events

De Morgan's Law:

$$(E \cup F)^c = E^c \cap F^c.$$

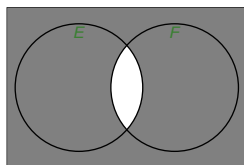


$$(E \cup F)^c$$

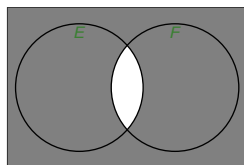


$$E^c \cap F^c$$

$$(E \cap F)^c = E^c \cup F^c.$$



$$(E \cap F)^c$$



$$E^c \cup F^c$$

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that **someone** has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= **Everyone** doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= **Someone** doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone **doesn't** have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone **doesn't** have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that **everyone** has an umbrella.
= Someone doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone doesn't have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone **doesn't** have an umbrella.

5. Simple properties of events

Let E_i 's be events (e.g., Person i has an umbrella). Then

De Morgan's Law:
$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c.$$

Not true that someone has an umbrella.
= Everyone doesn't have an umbrella.

$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c.$$

Not true that everyone has an umbrella.
= Someone doesn't have an umbrella.

Probability

Finally, we can now define what probability is.

Definition (axioms of probability)

The probability \mathbf{P} on a sample space Ω assigns *numbers* to *events* of Ω in such a way, that:

1. the probability of any event is non-negative: $\mathbf{P}\{E\} \geq 0$;
2. the probability of the sample space is one: $\mathbf{P}\{\Omega\} = 1$;
3. for any finitely or countably infinitely many *mutually exclusive* events E_1, E_2, \dots ,

$$\mathbf{P}\left\{\bigcup_i E_i\right\} = \sum_i \mathbf{P}\{E_i\}.$$

Probability

Notation:

$$\bigcup_{i=1}^n E_i = E_1 \cup E_2 \cup \dots \cup E_n \quad , \text{ or}$$

$$\bigcup_{i=1}^{\infty} E_i = E_1 \cup E_2 \cup \dots \quad ,$$

$$\sum_{i=1}^n \mathbf{P}\{E_i\} = \mathbf{P}\{E_1\} + \mathbf{P}\{E_2\} + \dots + \mathbf{P}\{E_n\} \quad , \text{ or}$$

$$\sum_{i=1}^{\infty} \mathbf{P}\{E_i\} = \mathbf{P}\{E_1\} + \mathbf{P}\{E_2\} + \dots \quad .$$

A few simple facts

Proposition

For any event, $\mathbf{P}\{E^c\} = 1 - \mathbf{P}\{E\}$.

Proof.

We know that E and E^c are mutually exclusive, and $E \cup E^c = \Omega$. Therefore by Axiom 3, and then 2,

$$\mathbf{P}\{E\} + \mathbf{P}\{E^c\} = \mathbf{P}\{E \cup E^c\} = \mathbf{P}\{\Omega\} = 1.$$



Corollary

We have $\mathbf{P}\{\emptyset\} = \mathbf{P}\{\Omega^c\} = 1 - \mathbf{P}\{\Omega\} = 1 - 1 = 0$.

For any event E , $\mathbf{P}\{E\} = 1 - \mathbf{P}\{E^c\} \leq 1$.

A few simple facts

Proposition (Inclusion-exclusion principle)

For any events E and F , $\mathbf{P}\{E \cup F\} = \mathbf{P}\{E\} + \mathbf{P}\{F\} - \mathbf{P}\{E \cap F\}$.



Proposition (Boole's inequality)

For any events E_1, E_2, \dots, E_n ,

$$\mathbf{P}\left\{\bigcup_{i=1}^n E_i\right\} \leq \sum_{i=1}^n \mathbf{P}\{E_i\}.$$

A few simple facts

Proof by induction.

When $n = 2$,

$$\mathbf{P}\{E_1 \cup E_2\} = \mathbf{P}\{E_1\} + \mathbf{P}\{E_2\} - \mathbf{P}\{E_1 \cap E_2\} \leq \mathbf{P}\{E_1\} + \mathbf{P}\{E_2\}.$$

Now suppose true for n . Then

$$\begin{aligned} \mathbf{P}\left\{\bigcup_{i=1}^{n+1} E_i\right\} &= \mathbf{P}\left\{\left(\bigcup_{i=1}^n E_i\right) \cup E_{n+1}\right\} \leq \mathbf{P}\left\{\bigcup_{i=1}^n E_i\right\} + \mathbf{P}\{E_{n+1}\} \\ &\leq \sum_{i=1}^n \mathbf{P}\{E_i\} + \mathbf{P}\{E_{n+1}\} = \sum_{i=1}^{n+1} \mathbf{P}\{E_i\}. \end{aligned}$$



A few simple facts

Proposition (Inclusion-exclusion principle)

For any events E, F, G ,

$$\begin{aligned} \mathbf{P}\{E \cup F \cup G\} &= \mathbf{P}\{E\} + \mathbf{P}\{F\} + \mathbf{P}\{G\} \\ &\quad - \mathbf{P}\{E \cap F\} - \mathbf{P}\{E \cap G\} - \mathbf{P}\{F \cap G\} \\ &\quad + \mathbf{P}\{E \cap F \cap G\}. \end{aligned}$$

A few simple facts

Example

In the sports club,

36 members play tennis,	22 play tennis and squash,
28 play squash,	12 play tennis and badminton,
18 play badminton,	9 play squash and badminton,
4 play tennis, squash and badminton.	

How many play at least one of these games?

A few simple facts

Example

36 members play tennis, 22 play tennis and squash,
28 play squash, 12 play tennis and badminton,
18 play badminton, 9 play squash and badminton,
 4 play tennis, squash and badminton.

Solution

Introduce probability by picking a random member out of those N enrolled to the club. Then

$$T := \{\text{that person plays tennis}\},$$
$$S := \{\text{that person plays squash}\},$$
$$B := \{\text{that person plays badminton}\}.$$

A few simple facts

Example

36 members play tennis, 22 play tennis and squash,
 28 play squash, 12 play tennis and badminton,
 18 play badminton, 9 play squash and badminton,
 4 play tennis, squash and badminton.

Solution (... cont'd)

$$\begin{aligned}
 \mathbf{P}\{T \cup S \cup B\} &= \mathbf{P}\{T\} + \mathbf{P}\{S\} + \mathbf{P}\{B\} \\
 &\quad - \mathbf{P}\{T \cap S\} - \mathbf{P}\{T \cap B\} - \mathbf{P}\{S \cap B\} \\
 &\quad + \mathbf{P}\{T \cap S \cap B\} \\
 &= \frac{36}{N} + \frac{28}{N} + \frac{18}{N} - \frac{22}{N} - \frac{12}{N} - \frac{9}{N} + \frac{4}{N} = \frac{43}{N}.
 \end{aligned}$$

Our answer is therefore 43 members.

A few simple facts

Proposition (Inclusion-exclusion principle)

For any events E_1, E_2, \dots, E_n ,

$$\begin{aligned} \mathbf{P}\{E_1 \cup E_2 \cup \dots \cup E_n\} &= \sum_{1 \leq i \leq n} \mathbf{P}\{E_i\} \\ &- \sum_{1 \leq i_1 < i_2 \leq n} \mathbf{P}\{E_{i_1} \cap E_{i_2}\} \\ &+ \sum_{1 \leq i_1 < i_2 < i_3 \leq n} \mathbf{P}\{E_{i_1} \cap E_{i_2} \cap E_{i_3}\} \\ &- \dots \\ &+ (-1)^{n+1} \mathbf{P}\{E_1 \cap E_2 \cap \dots \cap E_n\}. \end{aligned}$$



A few simple facts

Proposition

If $E \subseteq F$, then $\mathbf{P}\{F - E\} = \mathbf{P}\{F\} - \mathbf{P}\{E\}$.



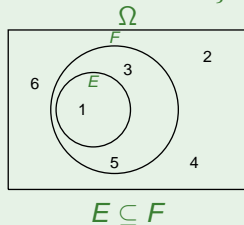
Corollary

If $E \subseteq F$, then $\mathbf{P}\{E\} \leq \mathbf{P}\{F\}$.

Example

$E = \{\text{rolling 1 on a die}\} \subseteq \{\text{rolling an odd no. on a die}\} = F$.

$$\frac{1}{6} = \mathbf{P}\{E\} \leq \mathbf{P}\{F\} = \frac{1}{2}.$$



Equally likely outcomes

A very special but important case is when the sample space is finite: $|\Omega| = N < \infty$, and each outcome of our experiment has equal probability. Then necessarily this probability equals $\frac{1}{N}$:

$$\mathbf{P}\{\omega\} = \frac{1}{N} \quad \forall \omega \in \Omega.$$

Definition

These outcomes $\omega \in \Omega$ are also called elementary events.

Let $E \subseteq \Omega$ be an event that consists of k elementary events: $|E| = k$. Then

$$\mathbf{P}\{E\} = \frac{|E|}{|\Omega|} = \frac{k}{N}.$$

We thus see why *counting* will be important.

Equally likely outcomes

Example

Rolling two dice, what is the probability that the sum of the numbers shown is 7?

Defining $E = \{\text{sum is 7}\} = \{(1, 6), (2, 5), \dots, (6, 1)\}$ in the sample space $\Omega = \{(i, j) : i, j = 1 \dots 6\}$, and noting that each pair of numbers is equally likely, we have

$$\mathbf{P}\{E\} = \frac{|E|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}.$$

A **wrong** solution would be to say that 7 is one out of the possible values 2, 3, ..., 12 for the sum, and the answer is $\frac{1}{11}$. These sums are **not equally likely**, e.g., 12 only occurs as one case out of 36. The pairs of numbers above are equally likely.

Equally likely outcomes

Often there are more than one ways of solving a probability problem. Sometimes we can count in order, or without order.

Example

An urn contains 6 red and 5 blue balls. We draw three balls at random, at once (that is, without replacement). What is the chance of drawing one red and two blue balls?

Equally likely outcomes

Solution (with order)

Define the sample space Ω as ordered choices of 3 out of the 11 balls (3-permutations of them). Then each choice is equally likely, and $|\Omega| = \frac{11!}{(11-3)!} = 11 \cdot 10 \cdot 9$. Now, our event E consists of:

- ▶ drawing red-blue-blue, in $6 \cdot 5 \cdot 4$ many ways, or
- ▶ drawing blue-red-blue, in $5 \cdot 6 \cdot 4$ many ways, or
- ▶ drawing blue-blue-red, in $5 \cdot 4 \cdot 6$ many ways.

Thus, $|E| = 6 \cdot 5 \cdot 4 + 5 \cdot 6 \cdot 4 + 5 \cdot 4 \cdot 6 = 3 \cdot 6 \cdot 5 \cdot 4$, and the answer is

$$\mathbf{P}\{E\} = \frac{|E|}{|\Omega|} = \frac{3 \cdot 6 \cdot 5 \cdot 4}{11 \cdot 10 \cdot 9} = \frac{4}{11}.$$

Equally likely outcomes

Solution (without order)

Define the sample space Ω as unordered choices of 3 out of the 11 balls (3-combinations of them). Then each choice is equally likely, and $|\Omega| = \binom{11}{3} = \frac{11 \cdot 10 \cdot 9}{6}$. Now, our event E consists of picking 1 out of the 6 red balls and 2 out of the 5 blue balls, with no respect to order. Thus, $|E| = \binom{6}{1} \cdot \binom{5}{2} = 6 \cdot 10$, and the answer is

$$\mathbf{P}\{E\} = \frac{|E|}{|\Omega|} = \frac{6 \cdot 10}{11 \cdot 10 \cdot 9 / 6} = \frac{4}{11}.$$

Both solutions are fine. Sometimes *in order* is easier, sometimes harder than *without order*. But never mix the two.

Equally likely outcomes

Example


Out of n people, what is the probability that there are no coinciding birthdays?

Solution

Assume 365 days in the year. The answer is of course zero, if $n > 365$ (this is called the pigeonhole principle).

Otherwise, our sample space Ω is a possible birthday for all n people, $|\Omega| = 365^n$. The event E of no coinciding birthdays can occur in

$$|E| = 365 \cdot 364 \cdots (365 - n + 1) = \frac{365!}{(365 - n)!}$$

many ways \rightsquigarrow  .

Equally likely outcomes

Solution (. . . cont'd)

The answer is

$$\mathbf{P}\{E\} = \frac{|E|}{|\Omega|} = \frac{365!}{(365 - n)! \cdot 365^n}.$$

This is

88%	for $n = 10$,
59%	for $n = 20$,
29%	for $n = 30$,
11%	for $n = 40$,
3%	for $n = 50$,
0.00003%	for $n = 100$.

Equally likely outcomes

Example

Flipping two fair coins, what is the probability of getting at least one Head?

Solution (straightforward)

As we have seen, the sample space is $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$ of equally likely outcomes (*coin is fair*), and we have at least one Head in 3 out of the 4 cases so the answer is $\frac{3}{4}$.

Equally likely outcomes

Solution (using the complement)

The complement of our event G in question is {no Heads}, which happens with probability $\frac{1}{4}$. Therefore

$$\mathbf{P}\{G\} = 1 - \mathbf{P}\{G^c\} = 1 - \frac{1}{4} = \frac{3}{4}.$$

Solution (using inclusion-exclusion)

Define E as the event that the first coin comes up Head, F that the second coin comes up Head. We are looking for the union of these events:

$$\mathbf{P}\{E \cup F\} = \mathbf{P}\{E\} + \mathbf{P}\{F\} - \mathbf{P}\{E \cap F\} = \frac{2}{4} + \frac{2}{4} - \frac{1}{4} = \frac{3}{4}.$$

2. Conditional probability

Conditional probability

Bayes' Theorem

Independence

Objectives:

- ▶ To understand what conditioning means, reduce the sample space
- ▶ To use the conditional probability, Law of Total Probability and Bayes' Theorem
- ▶ To understand and use independence and conditional independence

Conditional probability

Often one is given *partial information* on the outcome of an experiment. This changes our view of likelihoods for various outcomes. We shall build a mathematical model to handle this issue.

Example

We roll two dice. What is the probability that the sum of the numbers is 8? And if we know that the first die shows a 5?

The first question is by now easy: 5 cases out of 36, so the answer is $\frac{5}{36}$. Now, given the fact that the first die shows 5, we get a sum of 8 if and only if the second die shows 3. This has probability $\frac{1}{6}$, being the answer to the second question.

We see that partial information can change the probability of the outcomes of our experiment.

1. The reduced sample space

What happened was that we have reduced our world to the event $F = \{\text{first die shows 5}\} = \{(5, 1), (5, 2), \dots, (5, 6)\}$ that was given to us.

Definition

The event that is given to us is also called a reduced sample space. We can simply work in this set to figure out the conditional probabilities given this event.

The event F has 6 equally likely outcomes. Only one of them, $(5, 3)$, provides a sum of 8. Therefore, the conditional probability is $\frac{1}{6}$.

2. The formal definition

Let us also name the event

$$E = \{\text{the sum is 8}\} = \{(2, 6), (3, 5), \dots, (6, 2)\}.$$

The above question can be reformulated as: “In what proportion of cases in F will also E occur?” or, equivalently, “How does the probability of both E and F compare to the probability of F only?”

Definition

Let F be an event with $\mathbf{P}\{F\} > 0$ (we'll assume this from now on). Then the conditional probability of E , given F is defined as

$$\mathbf{P}\{E|F\} := \frac{\mathbf{P}\{E \cap F\}}{\mathbf{P}\{F\}}.$$

2. The formal definition

To answer the question we began with, with the formal definition we can now write $E \cap F = \{(5, 3)\}$ (the sum is 8 and the first die shows 5), and

$$\mathbf{P}\{E|F\} := \frac{\mathbf{P}\{E \cap F\}}{\mathbf{P}\{F\}} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}.$$

3. It's well-behaved

Proposition

The conditional probability $\mathbf{P}\{\cdot | F\}$ is a proper probability (it satisfies the axioms):

1. the conditional probability of any event is non-negative:
 $\mathbf{P}\{E | F\} \geq 0$;
2. the conditional probability of the sample space is one:
 $\mathbf{P}\{\Omega | F\} = 1$;
3. for any finitely or countably infinitely many *mutually exclusive* events E_1, E_2, \dots ,

$$\mathbf{P}\left\{\bigcup_i E_i | F\right\} = \sum_i \mathbf{P}\{E_i | F\}.$$



3. It's well-behaved

Corollary

All statements remain valid for $\mathbf{P}\{\cdot | F\}$. E.g.

- ▶ $\mathbf{P}\{E^c | F\} = 1 - \mathbf{P}\{E | F\}$.
- ▶ $\mathbf{P}\{\emptyset | F\} = 0$.
- ▶ $\mathbf{P}\{E | F\} = 1 - \mathbf{P}\{E^c | F\} \leq 1$.
- ▶ $\mathbf{P}\{E \cup G | F\} = \mathbf{P}\{E | F\} + \mathbf{P}\{G | F\} - \mathbf{P}\{E \cap G | F\}$.
- ▶ If $E \subseteq G$, then $\mathbf{P}\{G - E | F\} = \mathbf{P}\{G | F\} - \mathbf{P}\{E | F\}$.
- ▶ If $E \subseteq G$, then $\mathbf{P}\{E | F\} \leq \mathbf{P}\{G | F\}$.

Remark

BUT: Don't change the condition! E.g., $\mathbf{P}\{E | F\}$ and $\mathbf{P}\{E | F^c\}$ have nothing to do with each other.

4. Multiplication rule

Proposition (Multiplication rule)

For E_1, E_2, \dots, E_n events,

$$\mathbf{P}\{E_1 \cap \dots \cap E_n\} = \mathbf{P}\{E_1\} \cdot \mathbf{P}\{E_2 | E_1\} \cdot \mathbf{P}\{E_3 | E_1 \cap E_2\} \\ \dots \mathbf{P}\{E_n | E_1 \cap \dots \cap E_{n-1}\}.$$

Proof.

Just write out the conditionals. \rightsquigarrow



Example

An urn contains 6 red and 5 blue balls. We draw three balls at random, at once (that is, without replacement). What is the chance of drawing one red and two blue balls?

4. Multiplication rule

Solution (with order)

Define $R_1, R_2, R_3, B_1, B_2, B_3$ for the colors of the respective draws. We need

$$\begin{aligned} & \mathbf{P}\{R_1 \cap B_2 \cap B_3\} + \mathbf{P}\{B_1 \cap R_2 \cap B_3\} + \mathbf{P}\{B_1 \cap B_2 \cap R_3\} \\ &= \mathbf{P}\{R_1\} \cdot \mathbf{P}\{B_2 \mid R_1\} \cdot \mathbf{P}\{B_3 \mid R_1 \cap B_2\} \\ & \quad + \mathbf{P}\{B_1\} \cdot \mathbf{P}\{R_2 \mid B_1\} \cdot \mathbf{P}\{B_3 \mid B_1 \cap R_2\} \\ & \quad + \mathbf{P}\{B_1\} \cdot \mathbf{P}\{B_2 \mid B_1\} \cdot \mathbf{P}\{R_3 \mid B_1 \cap B_2\} \\ &= \frac{6}{11} \cdot \frac{5}{10} \cdot \frac{4}{9} + \frac{5}{11} \cdot \frac{6}{10} \cdot \frac{4}{9} + \frac{5}{11} \cdot \frac{4}{10} \cdot \frac{6}{9} = \frac{4}{11}. \end{aligned}$$

Bayes' Theorem

The aim here is to say something about $\mathbf{P}\{F | E\}$, once we know $\mathbf{P}\{E | F\}$ (and other things. . .). This will be very useful, and serve as a fundamental tool in probability and statistics.

1. The Law of Total Probability

Theorem (Law of Total Probability; aka. Partition Thm.)

For any events E and F ,

$$\mathbf{P}\{E\} = \mathbf{P}\{E | F\} \cdot \mathbf{P}\{F\} + \mathbf{P}\{E | F^c\} \cdot \mathbf{P}\{F^c\}.$$

(As usual, we assume that the conditionals exist.)

Proof.



1. The Law of Total Probability

Example

According to an insurance company,

- ▶ 30% of population are *accident-prone*, they will have an accident in any given year with 0.4 chance;
- ▶ the remaining 70% of population will have an accident in any given year with 0.2 chance.

Accepting this model, what is the probability that a new customer will have an accident in 2016?

1. The Law of Total Probability

Solution

Define the following events:

- ▶ $F := \{\text{new customer is accident-prone}\};$
- ▶ $A_{2016} := \{\text{new customer has accident in 2016}\}.$

Given are: $\mathbf{P}\{F\} = 0.3$, $\mathbf{P}\{A_{2016} | F\} = 0.4$, $\mathbf{P}\{A_{2016} | F^c\} = 0.2$.

Therefore,

$$\begin{aligned}\mathbf{P}\{A_{2016}\} &= \mathbf{P}\{A_{2016} | F\} \cdot \mathbf{P}\{F\} + \mathbf{P}\{A_{2016} | F^c\} \cdot \mathbf{P}\{F^c\} \\ &= 0.4 \cdot 0.3 + 0.2 \cdot 0.7 = 0.26.\end{aligned}$$

Notice a *weighted average* of 0.4 and 0.2 with weights 30% and 70%.

1. The Law of Total Probability

Solution

Define the following events:

- ▶ $F := \{\text{new customer is accident-prone}\};$
- ▶ $A_{2016} := \{\text{new customer has accident in 2016}\}.$

Given are: $\mathbf{P}\{F\} = 0.3$, $\mathbf{P}\{A_{2016} | F\} = 0.4$, $\mathbf{P}\{A_{2016} | F^c\} = 0.2$.

Therefore,

$$\begin{aligned}\mathbf{P}\{A_{2016}\} &= \mathbf{P}\{A_{2016} | F\} \cdot \mathbf{P}\{F\} + \mathbf{P}\{A_{2016} | F^c\} \cdot \mathbf{P}\{F^c\} \\ &= 0.4 \cdot 0.3 + 0.2 \cdot 0.7 = 0.26.\end{aligned}$$

Notice a *weighted average* of 0.4 and 0.2 with weights 30% and 70%.

1. The Law of Total Probability

Definition

Finitely or countably infinitely many events F_1, F_2, \dots form a complete system of events, or a partition of Ω , if $F_i \cap F_j = \emptyset$ and $\bigcup_j F_j = \Omega$.

Notice that exactly one of the F_i 's occurs.

Theorem (Law of Total Probability; aka. Partition Thm.)

For any event E and a complete system F_1, F_2, \dots , we have

$$\mathbf{P}\{E\} = \sum_i \mathbf{P}\{E | F_i\} \cdot \mathbf{P}\{F_i\}.$$

For any event F , the pair $F_1 := F$ and $F_2 := F^c$ form a complete system, and we are back to the previous version of the Theorem.

2. Bayes' Theorem

Theorem (Bayes' Theorem)


For any events E, F ,

$$P\{F|E\} = \frac{P\{E|F\} \cdot P\{F\}}{P\{E|F\} \cdot P\{F\} + P\{E|F^c\} \cdot P\{F^c\}}.$$

If $\{F_i\}_i$ is a complete system of events, then

$$P\{F_i|E\} = \frac{P\{E|F_i\} \cdot P\{F_i\}}{\sum_j P\{E|F_j\} \cdot P\{F_j\}}.$$

Proof.

Combine the definition of the conditional with the Law of Total Probability.  □

2. Bayes' Theorem

Let us go back to the insurance company. Imagine it's the 1st January 2017.

Example

We learn that the new customer did have an accident in 2016. Now what is the chance that (s)he is accident-prone?

According to Bayes' Theorem,

$$\begin{aligned}\mathbf{P}\{F | A_{2016}\} &= \frac{\mathbf{P}\{A_{2016} | F\} \cdot \mathbf{P}\{F\}}{\mathbf{P}\{A_{2016} | F\} \cdot \mathbf{P}\{F\} + \mathbf{P}\{A_{2016} | F^c\} \cdot \mathbf{P}\{F^c\}} \\ &= \frac{0.4 \cdot 0.3}{0.4 \cdot 0.3 + 0.2 \cdot 0.7} = \frac{6}{13} \simeq 0.46.\end{aligned}$$

C.f. the unconditioned probability $\mathbf{P}\{F\} = 0.3$.

Independence

In some special cases partial information on an experiment does not change the likelihood of an event. In this case we talk about *independence*.

Definition

Events E and F are independent, if $\mathbf{P}\{E | F\} = \mathbf{P}\{E\}$.

Notice that, **except for some degenerate cases**, this is equivalent to $\mathbf{P}\{E \cap F\} = \mathbf{P}\{E\} \cdot \mathbf{P}\{F\}$, and to $\mathbf{P}\{F | E\} = \mathbf{P}\{F\}$.

Don't mix independence with mutually exclusive events.

Independence is usually trivial, or rather tricky.

Independence

Proposition

Let E and F be independent events. Then E and F^c are also independent.

Proof.



Example

Rolling two dice, Let E be the event that the sum of the numbers is 6, F the event that the first die shows 3. These events are not independent:

$$\frac{1}{36} = \mathbf{P}\{E \cap F\} \neq \mathbf{P}\{E\} \cdot \mathbf{P}\{F\} = \frac{5}{36} \cdot \frac{1}{6}.$$

Independence

Example

Rolling two dice, Let E be the event that the sum of the numbers is 7, F the event that the first die shows 3. These events are independent (!):

$$\frac{1}{36} = \mathbf{P}\{E \cap F\} = \mathbf{P}\{E\} \cdot \mathbf{P}\{F\} = \frac{6}{36} \cdot \frac{1}{6} = \frac{1}{36}.$$

Equivalently,

$$\frac{1}{6} = \mathbf{P}\{E | F\} = \mathbf{P}\{E\}.$$

Or,

$$\frac{1}{6} = \mathbf{P}\{F | E\} = \mathbf{P}\{F\}.$$

Independence

Example

Rolling two dice, Let

- ▶ E be the event that the sum of the numbers is 7,
- ▶ F the event that the first die shows 3,
- ▶ G the event that the second die shows 4.

$$\frac{1}{36} = \mathbf{P}\{E \cap F\} = \mathbf{P}\{E\} \cdot \mathbf{P}\{F\} = \frac{6}{36} \cdot \frac{1}{6} = \frac{1}{36}.$$

$$\frac{1}{36} = \mathbf{P}\{E \cap G\} = \mathbf{P}\{E\} \cdot \mathbf{P}\{G\} = \frac{6}{36} \cdot \frac{1}{6} = \frac{1}{36}.$$

$$\frac{1}{36} = \mathbf{P}\{F \cap G\} = \mathbf{P}\{F\} \cdot \mathbf{P}\{G\} = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

E, F, G are pairwise independent.

But we have a bad feeling about this...

Independence

Example

- ▶ E is the event that the sum of the numbers is 7,
- ▶ F the event that the first die shows 3,
- ▶ G the event that the second die shows 4.

Are these events independent?

$$1 = \mathbf{P}\{E | F \cap G\} \neq \mathbf{P}\{E\} = \frac{1}{6}!$$

Or, equivalently,

$$\frac{1}{36} = \mathbf{P}\{E \cap F \cap G\} \neq \mathbf{P}\{E\} \cdot \mathbf{P}\{F \cap G\} = \frac{1}{6} \cdot \frac{1}{36}!$$

Recall $\mathbf{P}\{F \cap G\} = \mathbf{P}\{F\} \cdot \mathbf{P}\{G\}$ from the previous page.

Independence

Definition


Three events E , F , G are (mutually) independent, if

$$\mathbf{P}\{E \cap F\} = \mathbf{P}\{E\} \cdot \mathbf{P}\{F\},$$

$$\mathbf{P}\{E \cap G\} = \mathbf{P}\{E\} \cdot \mathbf{P}\{G\},$$

$$\mathbf{P}\{F \cap G\} = \mathbf{P}\{F\} \cdot \mathbf{P}\{G\},$$

$$\mathbf{P}\{E \cap F \cap G\} = \mathbf{P}\{E\} \cdot \mathbf{P}\{F\} \cdot \mathbf{P}\{G\}.$$

And, for more events the definition is that any (finite) collection of events have this factorisation property. \rightsquigarrow 

Independence

Below, $0 < p < 1$ is a probability parameter.

Example

n independent experiments are performed, each of which succeeds with probability p . What is the probability that every single experiment succeeds?

Easy: p^n .

Example (...)

... What is the probability that at least one experiment succeeds?

Looking at the complement, $1 - \mathbf{P}\{\text{each fails}\} = 1 - (1 - p)^n$.

Independence

Below, $0 < p < 1$ is a probability parameter.

Example

n independent experiments are performed, each of which succeeds with probability p . What is the probability that every single experiment succeeds?

Easy: p^n . $\xrightarrow{n \rightarrow \infty} 0$

Example (...)

... What is the probability that at least one experiment succeeds?

Looking at the complement, $1 - \mathbf{P}\{\text{each fails}\} = 1 - (1 - p)^n$.

Independence

Below, $0 < p < 1$ is a probability parameter.

Example

n independent experiments are performed, each of which succeeds with probability p . What is the probability that every single experiment succeeds?

Easy: p^n . $\xrightarrow{n \rightarrow \infty} 0$

Example (...)

... What is the probability that at least one experiment succeeds?

Looking at the complement, $1 - \mathbf{P}\{\text{each fails}\} = 1 - (1 - p)^n$.

$\xrightarrow{n \rightarrow \infty} 1$

Independence

Below, $0 < p < 1$ is a probability parameter.

Example

n independent experiments are performed, each of which succeeds with probability p . What is the probability that every single experiment succeeds?

Easy: p^n . $\xrightarrow{n \rightarrow \infty} 0$

Example (Murphy's Law)

... What is the probability that at least one experiment succeeds?

Looking at the complement, $1 - \mathbf{P}\{\text{each fails}\} = 1 - (1 - p)^n$.

$\xrightarrow{n \rightarrow \infty} 1$

Independence

Example

n independent experiments are performed, each of which succeeds with probability p . What is the probability that exactly k of them succeed?

$$\binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}.$$



Conditional independence

Back to the insurance company and to the 1st January 2017 again.

Example

We learn that the new customer did have an accident in 2016. Now what is the chance that (s)he will have one in 2017?

The question is $\mathbf{P}\{A_{2017} \mid A_{2016}\}$. We again consider F (being accident-prone):

Conditional independence

Example (... cont'd)

$$\begin{aligned}
 \mathbf{P}\{A_{2017} \mid A_{2016}\} &= \frac{\mathbf{P}\{A_{2017} \cap A_{2016}\}}{\mathbf{P}\{A_{2016}\}} \\
 &= \frac{\mathbf{P}\{A_{2017} \cap A_{2016} \cap F\}}{\mathbf{P}\{A_{2016}\}} + \frac{\mathbf{P}\{A_{2017} \cap A_{2016} \cap F^c\}}{\mathbf{P}\{A_{2016}\}} \\
 &= \frac{\mathbf{P}\{A_{2017} \cap A_{2016} \cap F\}}{\mathbf{P}\{A_{2016} \cap F\}} \cdot \frac{\mathbf{P}\{A_{2016} \cap F\}}{\mathbf{P}\{A_{2016}\}} \\
 &\quad + \frac{\mathbf{P}\{A_{2017} \cap A_{2016} \cap F^c\}}{\mathbf{P}\{A_{2016} \cap F^c\}} \cdot \frac{\mathbf{P}\{A_{2016} \cap F^c\}}{\mathbf{P}\{A_{2016}\}} \\
 &= \mathbf{P}\{A_{2017} \mid A_{2016} \cap F\} \cdot \mathbf{P}\{F \mid A_{2016}\} \\
 &\quad + \mathbf{P}\{A_{2017} \mid A_{2016} \cap F^c\} \cdot \mathbf{P}\{F^c \mid A_{2016}\}.
 \end{aligned}$$

Conditional independence

Example (... cont'd)

$$\begin{aligned}
 \mathbf{P}\{A_{2017} | A_{2016}\} &= \frac{\mathbf{P}\{A_{2017} \cap A_{2016}\}}{\mathbf{P}\{A_{2016}\}} \\
 &= \frac{\mathbf{P}\{A_{2017} \cap A_{2016} \cap F\}}{\mathbf{P}\{A_{2016}\}} + \frac{\mathbf{P}\{A_{2017} \cap A_{2016} \cap F^c\}}{\mathbf{P}\{A_{2016}\}} \\
 &= \frac{\mathbf{P}\{A_{2017} \cap A_{2016} \cap F\}}{\mathbf{P}\{A_{2016} \cap F\}} \cdot \frac{\mathbf{P}\{A_{2016} \cap F\}}{\mathbf{P}\{A_{2016}\}} \\
 &\quad + \frac{\mathbf{P}\{A_{2017} \cap A_{2016} \cap F^c\}}{\mathbf{P}\{A_{2016} \cap F^c\}} \cdot \frac{\mathbf{P}\{A_{2016} \cap F^c\}}{\mathbf{P}\{A_{2016}\}} \\
 &= \mathbf{P}\{A_{2017} | A_{2016} \cap F\} \cdot \mathbf{P}\{F | A_{2016}\} \\
 &\quad + \mathbf{P}\{A_{2017} | A_{2016} \cap F^c\} \cdot \mathbf{P}\{F^c | A_{2016}\}.
 \end{aligned}$$

Conditional Law of Total Probability, very useful.

Conditional independence

Example (... cont'd)

Now, we have conditional independence:

$$\begin{aligned} \mathbf{P}\{A_{2017} \mid A_{2016} \cap F\} &= \mathbf{P}\{A_{2017} \mid F\} = 0.4 && \text{and} \\ \mathbf{P}\{A_{2017} \mid A_{2016} \cap F^c\} &= \mathbf{P}\{A_{2017} \mid F^c\} = 0.2. \end{aligned}$$

Thus,

$$\begin{aligned} &\mathbf{P}\{A_{2017} \mid A_{2016}\} \\ &= \mathbf{P}\{A_{2017} \mid F\} \cdot \mathbf{P}\{F \mid A_{2016}\} + \mathbf{P}\{A_{2017} \mid F^c\} \cdot \mathbf{P}\{F^c \mid A_{2016}\} \\ &= 0.4 \cdot \frac{6}{13} + 0.2 \cdot \frac{7}{13} \simeq 0.29. \end{aligned}$$

C.f. $\mathbf{P}\{A_{2017}\} = \mathbf{P}\{A_{2016}\} = 0.26$ from before.

A_{2016} and A_{2017} are dependent!

3. Discrete random variables

Mass function

Expectation, variance

Bernoulli, Binomial

Poisson

Geometric

Objectives:

- ▶ To build a mathematical model for discrete random variables
- ▶ To define and get familiar with the probability mass function, expectation and variance of such variables
- ▶ To get experience in working with some of the basic distributions (Bernoulli, Binomial, Poisson, Geometric)

Random variables

The best way of thinking about random variables is just to consider them as random numbers.

But *random* means that there must be some kind of experiment behind these numbers. They actually fit well in our framework:

Definition

A random variable is a function from the sample space Ω to the real numbers \mathbb{R} .

The usual notation for random variables is X, Y, Z , etc., we often don't mark them as functions: $X(\omega), Y(\omega), Z(\omega)$, etc.

Random variables

Example

Flipping three coins, let X count the number of Heads obtained. Then, as a function on Ω ,

$$X(T, T, T) = 0;$$

$$X(T, T, H) = X(T, H, T) = X(H, T, T) = 1;$$

$$X(T, H, H) = X(H, T, H) = X(H, H, T) = 2;$$

$$X(H, H, H) = 3.$$

Instead, we'll just say that X can take on values 0, 1, 2, 3 with respective probabilities $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$.

How did we get $\frac{3}{8}$? Well,

$$\mathbf{P}\{X = 1\} = \mathbf{P}\{(T, T, H), (T, H, T), (H, T, T)\} = \frac{3}{8}.$$

Discrete random variables

Definition

A random variable X that can take on finitely or countably infinitely many possible values is called discrete.

Example

The number of Heads in three coinflips is discrete.

Example

The number of coinflips needed to first see a Head is discrete: it can be $1, 2, 3, \dots$

Example

The lifetime of a device is *not discrete*, it can be anything in the real interval $[0, \infty)$.

Mass function

The *distribution* of a random variable will be the object of central importance to us.

Definition

Let X be a discrete random variable with possible values x_1, x_2, \dots . The probability mass function (pmf), or distribution of a random variable tells us the probabilities of these possible values:

$$p_X(x_j) = \mathbf{P}\{X = x_j\},$$

for all possible x_j 's.

Often the possible values are just integers, $x_j = i$, and we can just write $p_X(i)$ for the mass function.

We also omit the subscript X if it's clear which random variable we are considering and simply put $p(i)$.

Mass function

Proposition

For any discrete random variable X ,

$$p(x_j) \geq 0, \quad \text{and} \quad \sum_i p(x_i) = 1.$$

Proof.



Remark

Vice versa: any function p which is only non-zero in countably many x_j values, and which has the above properties, is a probability mass function. There is a sample space and a random variable that realises this mass function.

Mass function

Example

We have seen X , the number of Heads in three coinflips. Its possible values are $X = 0, 1, 2, 3$, and its mass function is given by

$$p(0) = p(3) = \frac{1}{8}; \quad p(1) = p(2) = \frac{3}{8}.$$

Indeed,

$$\sum_{i=0}^3 p(i) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1.$$

Mass function

Example

Fix a positive parameter $\lambda > 0$, and define

$$p(i) = c \cdot \frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \dots$$

How should we choose c to make this into a mass function? In that case, what are $\mathbf{P}\{X = 0\}$ and $\mathbf{P}\{X > 2\}$ for the random variable X having this mass function?

Mass function

Solution

First, $p(i) \geq 0$ iff $c \geq 0$. Second, we need

$$\sum_{i=0}^{\infty} p(i) = c \cdot \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = c \cdot e^{\lambda} = 1,$$

from which $c = e^{-\lambda}$. To answer the probabilities,

$$\mathbf{P}\{X = 0\} = p(0) = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-\lambda};$$

Mass function

Solution (. . . cont'd)

$$\begin{aligned}\mathbf{P}\{X > 2\} &= 1 - \mathbf{P}\{X \leq 2\} \\ &= 1 - \mathbf{P}\{X = 0\} - \mathbf{P}\{X = 1\} - \mathbf{P}\{X = 2\} \\ &= 1 - e^{-\lambda} \cdot \frac{\lambda^0}{0!} - e^{-\lambda} \cdot \frac{\lambda^1}{1!} - e^{-\lambda} \cdot \frac{\lambda^2}{2!} \\ &= 1 - e^{-\lambda} - e^{-\lambda} \cdot \lambda - e^{-\lambda} \cdot \frac{\lambda^2}{2}.\end{aligned}$$

Expectation, variance

Once we have a random variable, we would like to quantify its *typical* behaviour in some sense. Two of the most often used quantities for this are the *expectation* and the *variance*.

1. Expectation

Definition

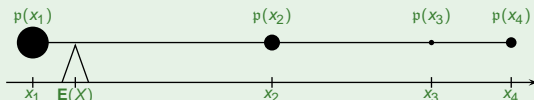
The expectation, or mean, or expected value of a discrete random variable X is defined as

$$EX := \sum_i x_i \cdot p(x_i),$$

provided that this sum exists.

Remark

The expectation is nothing else than a weighted average of the possible values x_i with weights $p(x_i)$. **A center of mass, in other words.**



1. Expectation

Remark

Why is this definition natural? \rightsquigarrow



Example (an important one...)

Let X be an indicator variable:

$$X = \begin{cases} 1, & \text{if event } E \text{ occurs,} \\ 0, & \text{if event } E^c \text{ occurs.} \end{cases}$$

Its mass function is $p(1) = \mathbf{P}\{E\}$ and $p(0) = 1 - \mathbf{P}\{E\}$. Its expectation is

$$\mathbf{E}X = 0 \cdot p(0) + 1 \cdot p(1) = \mathbf{P}\{E\}.$$

1. Expectation

Example (fair die)

Let X be the number shown after rolling a fair die. Then $X = 1, 2, \dots, 6$, each with probability $\frac{1}{6}$. The expectation is

$$EX = \sum_{i=1}^6 i \cdot p(i) = \sum_{i=1}^6 i \cdot \frac{1}{6} = \frac{1+6}{2} \cdot 6 \cdot \frac{1}{6} = \frac{7}{2}.$$

The expected value is not necessarily a possible value. Have you ever seen a die showing 3.5...?

2. A few properties of expectation

Proposition (expectation of a function of a r.v.)

Let X be a discrete random variable, and $g : \mathbb{R} \rightarrow \mathbb{R}$ function.
Then

$$\mathbf{E}g(X) = \sum_i g(x_i) \cdot p(x_i),$$

if exists...

This formula is rather natural.

Proof.



2. A few properties of expectation

Corollary (linearity of expectations, first version)

Let X be a discrete random variable, a and b fixed real numbers. Then

$$\mathbf{E}(aX + b) = a \cdot \mathbf{E}X + b.$$

Proof.

According to the above, (with $g(x) = ax + b$),

$$\begin{aligned}\mathbf{E}(aX + b) &= \sum_i (ax_i + b) \cdot p(x_i) = a \cdot \sum_i x_i p(x_i) + b \cdot \sum_i p(x_i) \\ &= a \cdot \mathbf{E}X + b.\end{aligned}$$



2. A few properties of expectation

Corollary (linearity of expectations, first version)

Let X be a discrete random variable, a and b fixed real numbers. Then

$$\mathbf{E}(aX + b) = a \cdot \mathbf{E}X + b.$$

Proof.

According to the above, (with $g(x) = ax + b$),

$$\begin{aligned}\mathbf{E}(aX + b) &= \sum_i (ax_i + b) \cdot p(x_i) = a \cdot \sum_i x_i p(x_i) + b \cdot \sum_i p(x_i) \\ &= a \cdot \mathbf{E}X + b.\end{aligned}$$



2. A few properties of expectation

Corollary (linearity of expectations, first version)

Let X be a discrete random variable, a and b fixed real numbers. Then

$$\mathbf{E}(aX + b) = a \cdot \mathbf{E}X + b.$$

Proof.

According to the above, (with $g(x) = ax + b$),

$$\begin{aligned}\mathbf{E}(aX + b) &= \sum_i (ax_i + b) \cdot p(x_i) = a \cdot \sum_i x_i p(x_i) + b \cdot \sum_i p(x_i) \\ &= a \cdot \mathbf{E}X + b \cdot 1.\end{aligned}$$



2. A few properties of expectation

Definition (moments)

Let n be a positive integer. The n^{th} moment of a random variable X is defined as

$$\mathbf{E}X^n.$$

The n^{th} absolute moment of X is

$$\mathbf{E}|X|^n.$$

Remark

Our notation in this definition and in the future will be

$$\mathbf{E}X^n := \mathbf{E}(X^n) \neq (\mathbf{E}X)^n !!$$

3. Variance

Example

Define $X \equiv 0$,

$$Y = \begin{cases} 1, \text{ wp. } \frac{1}{2}, \\ -1, \text{ wp. } \frac{1}{2}, \end{cases} \quad Z = \begin{cases} 2, \text{ wp. } \frac{1}{5}, \\ -\frac{1}{2}, \text{ wp. } \frac{4}{5}, \end{cases} \quad U = \begin{cases} 10, \text{ wp. } \frac{1}{2}, \\ -10, \text{ wp. } \frac{1}{2}, \end{cases}$$

Notice $\mathbf{E}X = \mathbf{E}Y = \mathbf{E}Z = \mathbf{E}U = 0$, the expectation does not distinguish between these rv.'s. Yet they are clearly different.

Definition (variance, standard deviation)

The variance and the standard deviation of a random variable are defined as $\mathbf{Var}X := \mathbf{E}(X - \mathbf{E}X)^2$ and $\mathbf{SD}X := \sqrt{\mathbf{Var}X}$.

Why this definition? \rightsquigarrow 

3. Variance

Example (... cont'd)

$$\mathbf{Var}X = \mathbf{E}(X - 0)^2 = 0^2 = 0,$$

$$\mathbf{SD}X = \sqrt{0} = 0.$$

$$\mathbf{Var}Y = \mathbf{E}(Y - 0)^2 = 1^2 \cdot \frac{1}{2} + (-1)^2 \cdot \frac{1}{2} = 1,$$

$$\mathbf{SD}Y = \sqrt{1} = 1.$$

$$\mathbf{Var}Z = \mathbf{E}(Z - 0)^2 = 2^2 \cdot \frac{1}{5} + \left(-\frac{1}{2}\right)^2 \cdot \frac{4}{5} = 1,$$

$$\mathbf{SD}Z = \sqrt{1} = 1.$$

$$\mathbf{Var}U = \mathbf{E}(U - 0)^2 = 10^2 \cdot \frac{1}{2} + (-10)^2 \cdot \frac{1}{2} = 100,$$

$$\mathbf{SD}U = \sqrt{100} = 10.$$

3. Variance

These numbers do distinguish between most of our variables
though finer information would be needed to make a difference
between Y and Z .

4. A few properties of the variance

Proposition (an equivalent form of the variance)

For any X , $\text{Var}X = \mathbf{E}X^2 - (\mathbf{E}X)^2$.

Proof.



Corollary

For any X , $\mathbf{E}X^2 \geq (\mathbf{E}X)^2$, with equality only if $X = \text{const. a.s.}$



New notation a.s. (almost surely) means *with probability one*.

4. A few properties of the variance

Example

The variance of the number X shown after rolling a fair die is

$$\mathbf{Var}X = \mathbf{E}X^2 - (\mathbf{E}X)^2 = (1^2 + 2^2 + \dots + 6^2) \cdot \frac{1}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

and its standard deviation is $\sqrt{35/12} \simeq 1.71$.

The two most important numbers we can say about a fair die are the average of **3.5** and typical deviations of **1.71** around this average.

4. A few properties of the variance

Example (an important one)

The variance of the indicator variable X of the event E is

$$\mathbf{Var}X = \mathbf{E}X^2 - (\mathbf{E}X)^2 = 1^2 \cdot \mathbf{P}\{E\} - (\mathbf{P}\{E\})^2 = \mathbf{P}\{E\} \cdot (1 - \mathbf{P}\{E\})$$

and the standard deviation is $\mathbf{SD}X = \sqrt{\mathbf{P}\{E\} \cdot (1 - \mathbf{P}\{E\})}$.

4. A few properties of the variance

Proposition (*nonlinearity of the variance*)

Let X be a random variable, a and b fixed real numbers. Then

$$\mathbf{Var}(aX + b) = a^2 \cdot \mathbf{Var}X.$$

Proof.



Notice the square on a^2 and also that, in particular, $\mathbf{Var}(X + b) = \mathbf{Var}X = \mathbf{Var}(-X)$: the variance is invariant to shifting the random variable by a constant b or to reflecting it.

Bernoulli, Binomial

In this part we'll get to know the Bernoulli and the Binomial distributions.

The setting will be that a fixed number of *independent* trials will be made, each succeeding with probability p . We will be counting the number of successes.

1. Definition

Definition

Suppose that n independent trials are performed, each succeeding with probability p . Let X count the number of successes within the n trials. Then X has the Binomial distribution with parameters n and p or, in short, $X \sim \text{Binom}(n, p)$.

The special case of $n = 1$ is called the Bernoulli distribution with parameter p .

Notice that the Bernoulli distribution is just another name for the *indicator variable* from before.

2. Mass function

Proposition

Let $X \sim \text{Binom}(n, p)$. Then $X = 0, 1, \dots, n$, and its mass function is

$$p(i) = \mathbf{P}\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \dots, n.$$

In particular, the *Bernoulli*(p) variable can take on values 0 or 1, with respective probabilities

$$p(0) = 1 - p, \quad p(1) = p.$$



2. Mass function

Remark

That the above is indeed a mass function we verify via the Binomial Theorem ($p(i) \geq 0$ is clear):

$$\sum_{i=0}^n p(i) = \sum_{i=0}^n \binom{n}{i} p^i (1-p)^{n-i} = [p + (1-p)]^n = 1.$$

2. Mass function

Example

Screws are sold in packages of 10. Due to a manufacturing error, each screw today is independently defective with probability 0.1. If there is money-back guarantee that at most one screw is defective in a package, what percentage of packages is returned?

Define X to be the number of defective screws in a package. Then $X \sim \text{Binom}(10, 0.1)$, and the answer is the chance that a given package has 2 or more faulty screws:

$$\begin{aligned}\mathbf{P}\{X \geq 2\} &= 1 - \mathbf{P}\{X = 0\} - \mathbf{P}\{X = 1\} \\ &= 1 - \binom{10}{0} 0.1^0 0.9^{10} - \binom{10}{1} 0.1^1 0.9^9 \simeq 0.2639.\end{aligned}$$

3. Expectation, variance

Proposition

Let $X \sim \text{Binom}(n, p)$. Then

$$\mathbf{E}X = np, \quad \text{and} \quad \mathbf{Var}X = np(1 - p).$$

Proof.

We first need to calculate

$$\mathbf{E}X = \sum_i i \cdot p(i) = \sum_{i=0}^n i \cdot \binom{n}{i} p^i (1-p)^{n-i}.$$

To handle this, here is a cute trick: $i = \frac{d}{dt} t^i \Big|_{t=1}$. 

3. Expectation, variance

Proof.

$$\begin{aligned} \mathbf{EX} &= \sum_{i=0}^n \binom{n}{i} i \cdot p^i (1-p)^{n-i} \\ &= \sum_{i=0}^n \binom{n}{i} \frac{d}{dt} t^i \Big|_{t=1} \cdot p^i (1-p)^{n-i} \\ &= \frac{d}{dt} \left(\sum_{i=0}^n \binom{n}{i} (tp)^i (1-p)^{n-i} \right) \Big|_{t=1} \\ &= \frac{d}{dt} (tp + 1 - p)^n \Big|_{t=1} = n(tp + 1 - p)^{n-1} \cdot p \Big|_{t=1} = np. \end{aligned}$$

3. Expectation, variance

Proof.

$$\begin{aligned} \mathbf{EX} &= \sum_{i=0}^n \binom{n}{i} i \cdot p^i (1-p)^{n-i} \\ &= \sum_{i=0}^n \binom{n}{i} \frac{d}{dt} t^i \Big|_{t=1} \cdot p^i (1-p)^{n-i} \\ &= \frac{d}{dt} \left(\sum_{i=0}^n \binom{n}{i} (tp)^i (1-p)^{n-i} \right) \Big|_{t=1} \\ &= \frac{d}{dt} (tp + 1 - p)^n \Big|_{t=1} = n(tp + 1 - p)^{n-1} \cdot p \Big|_{t=1} = np. \end{aligned}$$

Poisson

The Poisson distribution is of central importance in Probability. We won't see immediately why, we'll just start with defining its distribution. Later we'll see how it comes from the Binomial.

1. Mass function

Definition

Fix a positive real number λ . The random variable X is Poisson distributed with parameter λ , in short $X \sim \text{Poi}(\lambda)$, if it is non-negative integer valued, and its mass function is

$$p(i) = \mathbf{P}\{X = i\} = e^{-\lambda} \cdot \frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \dots$$

We have already seen in an example that this is indeed a mass function.

Ok, nice, but why this distribution?

2. Poisson approximation of Binomial

Proposition

Fix $\lambda > 0$, and suppose that $Y_n \sim \text{Binom}(n, p)$ with $p = p(n)$ in such a way that $n \cdot p \rightarrow \lambda$. Then the distribution of Y_n converges to $\text{Poisson}(\lambda)$:

$$\forall i \geq 0 \quad \mathbf{P}\{Y_n = i\} \xrightarrow{n \rightarrow \infty} e^{-\lambda} \frac{\lambda^i}{i!}.$$

That is, take $Y \sim \text{Binom}(n, p)$ with large n , small p , such that $np \simeq \lambda$. Then Y is approximately $\text{Poisson}(\lambda)$ distributed.

2. Poisson approximation of Binomial

Proof.

$$\begin{aligned}\mathbf{P}\{Y_n = i\} &= \binom{n}{i} \cdot p^i (1-p)^{n-i} \\ &= \frac{1}{i!} \cdot [np] \cdot [(n-1)p] \cdots [(n-i+1)p] \cdot \frac{(1-p)^n}{(1-p)^i}.\end{aligned}$$

Now, $np \rightarrow \lambda$, $(n-1)p \rightarrow \lambda$, \dots , $(n-i+1)p \rightarrow \lambda$.

$$(1-p)^n = \left(1 - \frac{1}{1/p}\right)^n \xrightarrow{n \rightarrow \infty} e^{-\lambda}. \quad \rightsquigarrow \img alt="whiteboard icon" data-bbox="744 658 790 712"/>$$

$(1-p)^i \rightarrow 1$. Therefore, $\mathbf{P}\{Y_n = i\} \rightarrow \frac{1}{i!} \lambda^i e^{-\lambda}$. □

2. Poisson approximation of Binomial

Proof.

$$\begin{aligned} \mathbf{P}\{Y_n = i\} &= \binom{n}{i} \cdot p^i (1-p)^{n-i} \\ &= \frac{1}{i!} \cdot [np] \cdot [(n-1)p] \cdots [(n-i+1)p] \cdot \frac{(1-p)^n}{(1-p)^i}. \end{aligned}$$

Now, $np \rightarrow \lambda$, $(n-1)p \rightarrow \lambda$, \dots , $(n-i+1)p \rightarrow \lambda$.

$$(1-p)^n = \left(1 - \frac{1}{1/p}\right)^n \xrightarrow[n \rightarrow \infty]{} e^{-\lambda}. \quad \rightsquigarrow \img alt="whiteboard icon" data-bbox="744 658 791 711"/>$$

$(1-p)^i \rightarrow 1$. Therefore, $\mathbf{P}\{Y_n = i\} \rightarrow \frac{1}{i!} \lambda^i e^{-\lambda}$. □

2. Poisson approximation of Binomial

Proof.

$$\begin{aligned} \mathbf{P}\{Y_n = i\} &= \binom{n}{i} \cdot p^i (1-p)^{n-i} \\ &= \frac{1}{i!} \cdot [np] \cdot [(n-1)p] \cdots [(n-i+1)p] \cdot \frac{(1-p)^n}{(1-p)^i}. \end{aligned}$$

Now, $np \rightarrow \lambda$, $(n-1)p \rightarrow \lambda$, \dots , $(n-i+1)p \rightarrow \lambda$.

$$(1-p)^n = \left(1 - \frac{1}{1/p}\right)^n \xrightarrow{n \rightarrow \infty} e^{-\lambda}. \quad \rightsquigarrow \img alt="whiteboard icon" data-bbox="744 658 791 711"/>$$

$(1-p)^i \rightarrow 1$. Therefore, $\mathbf{P}\{Y_n = i\} \rightarrow \frac{1}{i!} \lambda^i e^{-\lambda}$. □

2. Poisson approximation of Binomial

Proof.

$$\begin{aligned} \mathbf{P}\{Y_n = i\} &= \binom{n}{i} \cdot p^i (1-p)^{n-i} \\ &= \frac{1}{i!} \cdot [np] \cdot [(n-1)p] \cdots [(n-i+1)p] \cdot \frac{(1-p)^n}{(1-p)^i}. \end{aligned}$$

Now, $np \rightarrow \lambda$, $(n-1)p \rightarrow \lambda$, \dots , $(n-i+1)p \rightarrow \lambda$.

$$(1-p)^n = \left(1 - \frac{1}{1/p}\right)^n \xrightarrow{n \rightarrow \infty} e^{-\lambda}. \quad \rightsquigarrow \text{graph icon}$$

$(1-p)^i \rightarrow 1$. Therefore, $\mathbf{P}\{Y_n = i\} \rightarrow \frac{1}{i!} \lambda^i e^{-\lambda}$. □

2. Poisson approximation of Binomial

Proof.

$$\begin{aligned} \mathbf{P}\{Y_n = i\} &= \binom{n}{i} \cdot p^i (1-p)^{n-i} \\ &= \frac{1}{i!} \cdot [np] \cdot [(n-1)p] \cdots [(n-i+1)p] \cdot \frac{(1-p)^n}{(1-p)^i}. \end{aligned}$$

Now, $np \rightarrow \lambda$, $(n-1)p \rightarrow \lambda$, \dots , $(n-i+1)p \rightarrow \lambda$.

$$(1-p)^n = \left(1 - \frac{1}{1/p}\right)^n \xrightarrow{n \rightarrow \infty} e^{-\lambda}. \quad \rightsquigarrow \text{📌}$$

$(1-p)^i \rightarrow 1$. Therefore, $\mathbf{P}\{Y_n = i\} \rightarrow \frac{1}{i!} \lambda^i e^{-\lambda}$. □

2. Poisson approximation of Binomial

Proof.

$$\begin{aligned} \mathbf{P}\{Y_n = i\} &= \binom{n}{i} \cdot p^i (1-p)^{n-i} \\ &= \frac{1}{i!} \cdot [np] \cdot [(n-1)p] \cdots [(n-i+1)p] \cdot \frac{(1-p)^n}{(1-p)^i}. \end{aligned}$$

Now, $np \rightarrow \lambda$, $(n-1)p \rightarrow \lambda$, \dots , $(n-i+1)p \rightarrow \lambda$.

$$(1-p)^n = \left(1 - \frac{1}{1/p}\right)^n \xrightarrow{n \rightarrow \infty} e^{-\lambda}. \quad \rightsquigarrow \img alt="whiteboard icon" data-bbox="744 658 790 712"/>$$

$(1-p)^i \rightarrow 1$. Therefore, $\mathbf{P}\{Y_n = i\} \rightarrow \frac{1}{i!} \lambda^i e^{-\lambda}$. □

2. Poisson approximation of Binomial

Proof.

$$\begin{aligned} \mathbf{P}\{Y_n = i\} &= \binom{n}{i} \cdot p^i (1-p)^{n-i} \\ &= \frac{1}{i!} \cdot [np] \cdot [(n-1)p] \cdots [(n-i+1)p] \cdot \frac{(1-p)^n}{(1-p)^i}. \end{aligned}$$

Now, $np \rightarrow \lambda$, $(n-1)p \rightarrow \lambda$, \dots , $(n-i+1)p \rightarrow \lambda$.

$$(1-p)^n = \left(1 - \frac{1}{1/p}\right)^n \xrightarrow{n \rightarrow \infty} e^{-\lambda}. \quad \rightsquigarrow \text{📌}$$

$(1-p)^i \rightarrow 1$. Therefore, $\mathbf{P}\{Y_n = i\} \rightarrow \frac{1}{i!} \lambda^i e^{-\lambda}$. □

2. Poisson approximation of Binomial

Proof.

$$\begin{aligned} \mathbf{P}\{Y_n = i\} &= \binom{n}{i} \cdot p^i (1-p)^{n-i} \\ &= \frac{1}{i!} \cdot [np] \cdot [(n-1)p] \cdots [(n-i+1)p] \cdot \frac{(1-p)^n}{(1-p)^i}. \end{aligned}$$

Now, $np \rightarrow \lambda$, $(n-1)p \rightarrow \lambda$, \dots , $(n-i+1)p \rightarrow \lambda$.

$$(1-p)^n = \left(1 - \frac{1}{1/p}\right)^n \xrightarrow[n \rightarrow \infty]{} e^{-\lambda}. \quad \rightsquigarrow \text{graph icon}$$

$(1-p)^i \rightarrow 1$. Therefore, $\mathbf{P}\{Y_n = i\} \rightarrow \frac{1}{i!} \lambda^i e^{-\lambda}$. □

2. Poisson approximation of Binomial

Proof.

$$\begin{aligned} \mathbf{P}\{Y_n = i\} &= \binom{n}{i} \cdot p^i (1-p)^{n-i} \\ &= \frac{1}{i!} \cdot [np] \cdot [(n-1)p] \cdots [(n-i+1)p] \cdot \frac{(1-p)^n}{(1-p)^i}. \end{aligned}$$

Now, $np \rightarrow \lambda$, $(n-1)p \rightarrow \lambda$, \dots , $(n-i+1)p \rightarrow \lambda$.

$$(1-p)^n = \left(1 - \frac{1}{1/p}\right)^n \xrightarrow[n \rightarrow \infty]{} e^{-\lambda}. \quad \rightsquigarrow \img alt="whiteboard icon" data-bbox="743 658 790 712"/>$$

$(1-p)^i \rightarrow 1$. Therefore, $\mathbf{P}\{Y_n = i\} \rightarrow \frac{1}{i!} \lambda^i e^{-\lambda}$. □

3. Expectation, variance

Proposition

For $X \sim \text{Poi}(\lambda)$, $\mathbf{E}X = \mathbf{Var}X = \lambda$.

Recall np and $np(1 - p)$ for the Binomial...

Proof.

$$\begin{aligned}\mathbf{E}X &= \sum_{i=0}^{\infty} i p(i) = \sum_{i=1}^{\infty} i \cdot e^{-\lambda} \frac{\lambda^i}{i!} \\ &= \lambda \sum_{i=1}^{\infty} e^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} = \lambda \sum_{j=0}^{\infty} e^{-\lambda} \frac{\lambda^j}{j!} = \lambda.\end{aligned}$$



4. Examples

Example

Because of the approximation of Binomial, the

- ▶ number of typos on a page of a book;
- ▶ number of citizens over 100 years of age in a city;
- ▶ number of incoming calls per hour in a customer centre;
- ▶ number of customers in a post office today

are each well approximated by the Poisson distribution.

Many independent small probability events, summing up to “a few” in expectation.  

4. Examples

Example

A book on average has $1/2$ typos per page. What is the probability that the next page has at least three of them?

The number X of typos on a page follows a **Poisson**(λ) distribution, where λ can be determined from $\frac{1}{2} = \mathbf{E}X = \lambda$. To answer the question,

$$\begin{aligned} \mathbf{P}\{X \geq 3\} &= 1 - \mathbf{P}\{X \leq 2\} \\ &= 1 - \mathbf{P}\{X = 0\} - \mathbf{P}\{X = 1\} - \mathbf{P}\{X = 2\} \\ &= 1 - \frac{(1/2)^0}{0!} \cdot e^{-1/2} - \frac{(1/2)^1}{1!} \cdot e^{-1/2} - \frac{(1/2)^2}{2!} \cdot e^{-1/2} \\ &\simeq 0.014. \end{aligned}$$

4. Examples

Example

Screws are sold in packages of 10. Due to a manufacturing error, each screw today is independently defective with probability 0.1. If there is money-back guarantee that at most one screw is defective in a package, what percentage of packages is returned?

Define X as before; $X \sim \text{Binom}(10, 0.1)$. However, it can already well be approximated by a $\text{Poi}(1)$ distribution ($\lambda = 1 = 10 \cdot 0.1 = np$). Thus,

$$\begin{aligned} \mathbf{P}\{X \geq 2\} &= 1 - \mathbf{P}\{X = 0\} - \mathbf{P}\{X = 1\} \\ &\simeq 1 - e^{-1} \frac{1^0}{0!} - e^{-1} \frac{1^1}{1!} \simeq 0.2642. \end{aligned}$$

Compare this with the exact value 0.2639 from the Binomial.

Geometric

In this setting we again perform independent trials. However, the question we ask is now different: we'll be waiting for the first success.

1. Mass function

Definition

Suppose that independent trials, each succeeding with probability p , are repeated until the first success. The total number X of trials made has the Geometric(p) distribution (in short, $X \sim \text{Geom}(p)$).

Proposition

X can take on positive integers, with probabilities
 $p(i) = (1 - p)^{i-1} \cdot p, i = 1, 2, \dots$

That this is a mass function, we verify by $p(i) \geq 0$ and

$$\sum_{i=1}^{\infty} p(i) = \sum_{i=1}^{\infty} (1 - p)^{i-1} \cdot p = \frac{p}{1 - (1 - p)} = 1.$$

1. Mass function

Remark

For a $\text{Geometric}(p)$ random variable and any $k \geq 1$ we have $\mathbf{P}\{X \geq k\} = (1 - p)^{k-1}$ (we have at least $k - 1$ failures).

Corollary

The Geometric random variable is (discrete) memoryless: for every $k \geq 1, n \geq 0$

$$\mathbf{P}\{X \geq n + k \mid X > n\} = \mathbf{P}\{X \geq k\}.$$



2. Expectation, variance

Proposition

For a *Geometric*(p) random variable X ,

$$\mathbf{E}X = \frac{1}{p}, \quad \mathbf{Var}X = \frac{1-p}{p^2}.$$

2. Expectation, variance

Proof.

$$\begin{aligned} \mathbf{E}X &= \sum_{i=1}^{\infty} i \cdot (1-p)^{i-1} p = \sum_{i=0}^{\infty} i \cdot (1-p)^{i-1} p \\ &= \sum_{i=0}^{\infty} \frac{d}{dt} t^i \Big|_{t=1} \cdot (1-p)^{i-1} p = \frac{d}{dt} \left(\sum_{i=0}^{\infty} t^i \cdot (1-p)^{i-1} p \right) \Big|_{t=1} \\ &= \frac{p}{1-p} \cdot \frac{d}{dt} \frac{1}{1-(1-p)t} \Big|_{t=1} \\ &= \frac{p}{1-p} \cdot \frac{1-p}{(1-(1-p))^2} = \frac{1}{p}. \end{aligned}$$



2. Expectation, variance

Proof.

$$\begin{aligned} \mathbf{E}X &= \sum_{i=1}^{\infty} i \cdot (1-p)^{i-1} p = \sum_{i=0}^{\infty} i \cdot (1-p)^{i-1} p \\ &= \sum_{i=0}^{\infty} \frac{d}{dt} t^i \Big|_{t=1} \cdot (1-p)^{i-1} p = \frac{d}{dt} \left(\sum_{i=0}^{\infty} t^i \cdot (1-p)^{i-1} p \right) \Big|_{t=1} \\ &= \frac{p}{1-p} \cdot \frac{d}{dt} \frac{1}{1-(1-p)t} \Big|_{t=1} \\ &= \frac{p}{1-p} \cdot \frac{1-p}{(1-(1-p))^2} = \frac{1}{p}. \end{aligned}$$



3. Example

Example

To first see 3 appearing on a fair die, we wait $X \sim \text{Geom}(\frac{1}{6})$ many rolls. Our average waiting time is $\mathbf{E}X = \frac{1}{1/6} = 6$ rolls, and the standard deviation is

$$\text{SD } X = \sqrt{\mathbf{Var}X} = \sqrt{\frac{1 - \frac{1}{6}}{(\frac{1}{6})^2}} = \sqrt{30} \simeq 5.48.$$

3. Example

Example (... cont'd)

The chance that 3 first comes on the 7th roll is

$$p(7) = \mathbf{P}\{X = 7\} = \left(1 - \frac{1}{6}\right)^6 \cdot \frac{1}{6} \simeq 0.056,$$

while the chance that 3 first comes on the 7th or later rolls is

$$\mathbf{P}\{X \geq 7\} = \left(1 - \frac{1}{6}\right)^6 \simeq 0.335.$$

4. Continuous random variables

Distribution, density

Uniform

Exponential

Normal

Transformations

Objectives:

- ▶ To build a mathematical model of continuous random variables
- ▶ To define and get familiar with the cumulative distribution function, probability density function, expectation and variance of such variables
- ▶ To get experience in working with some of the basic distributions (Uniform, Exponential, Normal)
- ▶ To find the distribution of a function of a random variable

Non-discrete random variables

Spin a pencil on the table. Let X be the angle it points to after it has stopped.

- ▶ What is the probability that $X = 0^\circ$?
- ▶ What is the probability that $X = 90^\circ$?
- ▶ What is the probability that $X = 258.4562^\circ$?

These are all zero. $\mathbf{P}\{X = x\} = 0$ for any $x \in [0, 360^\circ)$. This random variable has no mass function. It is *not* a discrete random variable, it can take on uncountably many values.

We need a new framework to handle these kind of phenomena.

Distribution function

Definition

The cumulative distribution function (cdf) of a random variable X is given by

$$F : \mathbb{R} \rightarrow [0, 1], \quad x \mapsto F(x) = \mathbf{P}\{X \leq x\}.$$

Notice that this function is well defined for *any* random variable.

Remark

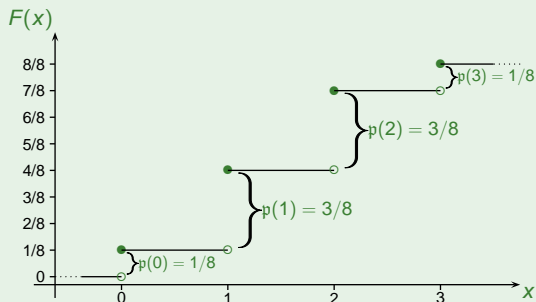
The distribution function contains all relevant information about the distribution of our random variable. E.g., for any fixed $a < b$,

$$\mathbf{P}\{a < X \leq b\} = \mathbf{P}\{X \leq b\} - \mathbf{P}\{X \leq a\} = F(b) - F(a).$$

Distribution function

Example

Flip a coin three times, and let X be the number of Heads obtained. Its distribution function is given by



Distribution function

Definition

A random variable with piecewise constant distribution function is called *discrete*. Its mass function values equal to the jump sizes in the distribution function.

And this is equivalent to our earlier definition (taking on countably many possible values).

Distribution function

Proposition

A cumulative distribution function F

- ▶ is non-decreasing;
- ▶ has limit $\lim_{x \rightarrow -\infty} F(x) = 0$ on the left;
- ▶ has limit $\lim_{x \rightarrow \infty} F(x) = 1$ on the right;
- ▶ is continuous from the right.



Vice versa: any function F with the above properties is a cumulative distribution function. There is a sample space and a random variable on it that realises this distribution function.

Distribution function

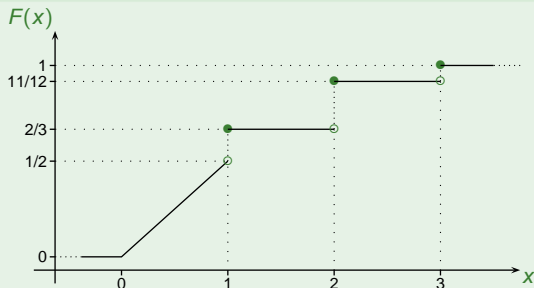
Example

Let F be the function given by

$$F(x) = \begin{cases} 0, & x < 0, \\ x/2, & 0 \leq x < 1, \\ 2/3, & 1 \leq x < 2, \\ 11/12, & 2 \leq x < 3, \\ 1, & 3 \leq x. \end{cases}$$

Distribution function

Example (... cont'd)



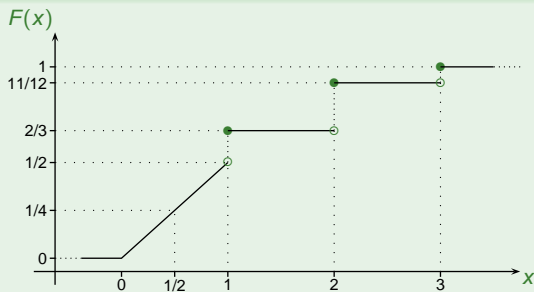
$$\mathbf{P}\{X \leq 3\} = F(3) = 1,$$

$$\mathbf{P}\{X < 3\} = \lim_{x \nearrow 3} F(x) = \frac{11}{12},$$

$$\mathbf{P}\{X = 3\} = F(3) - \lim_{x \nearrow 3} F(x) = \frac{1}{12}.$$

Distribution function

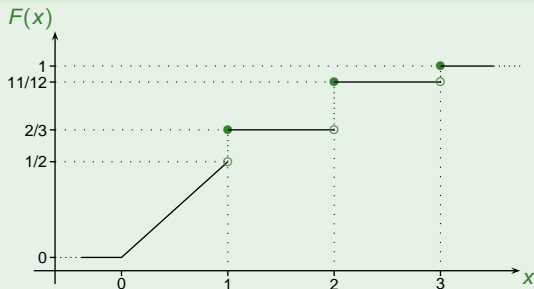
Example (... cont'd)



$$\mathbf{P}\left\{X \geq \frac{1}{2}\right\} = \mathbf{P}\left\{X > \frac{1}{2}\right\} = \frac{3}{4}.$$

Distribution function

Example (... cont'd)



$$\mathbf{P}\{2 < X \leq 4\} = F(4) - F(2) = 1 - \frac{11}{12} = \frac{1}{12},$$

$$\begin{aligned} \mathbf{P}\{2 \leq X < 4\} &= \mathbf{P}\{2 < X \leq 4\} - \mathbf{P}\{X = 4\} + \mathbf{P}\{X = 2\} \\ &= \frac{1}{12} - 0 + \left(\frac{11}{12} - \frac{2}{3}\right) = \frac{1}{3}. \end{aligned}$$

1. Density function

Definition

Suppose that a random variable has its distribution function in the form of

$$F(a) = \int_{-\infty}^a f(x) dx, \quad (\forall a \in \mathbb{R})$$

with a function $f \geq 0$. Then the distribution is called (absolutely) continuous, and f is the probability density function (pdf).

We'll assume that X is continuous for the rest of this chapter.

1. Density function

Proposition

A probability density function f

- ▶ is non-negative;
- ▶ has total integral $\int_{-\infty}^{\infty} f(x) dx = 1$.



Vice versa: any function f with the above properties is a probability density function. There is a sample space and a continuous random variable on it that realises this density.

2. Properties of the density function

Proposition

For any* subset $B \subseteq \mathbb{R}$,

$$\mathbf{P}\{X \in B\} = \int_B f(x) \, dx.$$



Corollary

Indeed, for a continuous random variable X ,

$$\mathbf{P}\{X = a\} = \int_{\{a\}} f(x) \, dx = 0 \quad (\forall a \in \mathbb{R}).$$

2. Properties of the density function

Corollary

For a small ε ,

$$\mathbf{P}\{X \in (a, a + \varepsilon]\} = \int_a^{a+\varepsilon} f(x) dx \simeq f(a) \cdot \varepsilon.$$

There is no particular value that X can take on with positive chance. We can only talk about intervals, and the density tells us the likelihood that X is *around a point* a .

2. Properties of the density function

Corollary

To get to the density from a(n absolutely continuous!) distribution function,

$$f(a) = \frac{dF(a)}{da} \quad (\text{a.e. } a \in \mathbb{R}).$$

New notation a.e. (almost every): for all but a zero-measure set of numbers, so it's no problem for any integrals.

3. Expectation, variance

The way of defining the expectation will be no surprise for anyone (c.f. the discrete case):

Definition

The expected value of a continuous random variable X is defined by

$$EX = \int_{-\infty}^{\infty} x \cdot f(x) dx,$$

if the integral exists.

3. Expectation, variance

In a way similar to the discrete case,

Proposition

Let X be a continuous random variable, and g an $\mathbb{R} \rightarrow \mathbb{R}$ function. Then

$$\mathbf{E}g(X) = \int_{-\infty}^{\infty} g(x) \cdot f(x) \, dx$$

if exists.

From here we can define moments, absolute moments

$$\mathbf{E}X^n = \int_{-\infty}^{\infty} x^n \cdot f(x) \, dx, \quad \mathbf{E}|X|^n = \int_{-\infty}^{\infty} |x|^n \cdot f(x) \, dx,$$

variance $\mathbf{Var}X = \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2$ and standard deviation $\mathbf{SD}X = \sqrt{\mathbf{Var}X}$ as in the discrete case. These enjoy the same properties as before.

Uniform

We are given real numbers $\alpha < \beta$, and wish to define a random variable X that's *equally likely to fall anywhere* in this interval. Thinking about the definitions, we can do that by assuming a constant density on this interval.

1. Density, distribution function

Definition

Fix $\alpha < \beta$ reals. We say that X has the uniform distribution over the interval (α, β) , in short, $X \sim U(\alpha, \beta)$, if its density is given by

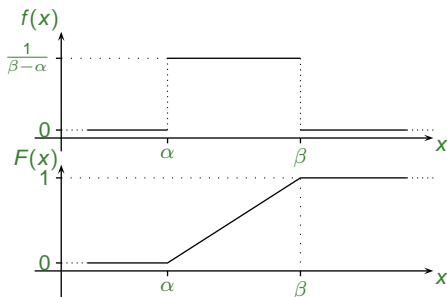
$$f(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \text{if } x \in (\alpha, \beta), \\ 0, & \text{otherwise.} \end{cases}$$

Notice that this is exactly the value of the constant that makes this a density.

1. Density, distribution function

Integrating this density,

$$F(x) = \begin{cases} 0, & \text{if } x \leq \alpha, \\ \frac{x - \alpha}{\beta - \alpha}, & \text{if } \alpha \leq x \leq \beta, \\ 1, & \text{if } \beta \leq x. \end{cases}$$



1. Density, distribution function

Remark

If $X \sim U(\alpha, \beta)$, and $\alpha < a < b < \beta$, then

$$\mathbf{P}\{a < X \leq b\} = \int_a^b f(x) \, dx = \frac{b - a}{\beta - \alpha}.$$

Probabilities are computed by proportions of lengths.

2. Expectation, variance

Proposition

For $X \sim U(\alpha, \beta)$,

$$\mathbf{E}X = \frac{\alpha + \beta}{2}, \quad \mathbf{Var}X = \frac{(\beta - \alpha)^2}{12}.$$

Proof.

$$\mathbf{E}X = \int_{-\infty}^{\infty} xf(x) dx = \int_{\alpha}^{\beta} \frac{x}{\beta - \alpha} dx = \frac{\frac{\beta^2}{2} - \frac{\alpha^2}{2}}{\beta - \alpha} = \frac{\alpha + \beta}{2}.$$

Exponential

The Exponential is a very special distribution because of its *memoryless* property. It is often considered as a waiting time, and is widely used in the theory of stochastic processes.

1. Density, distribution function

Definition

Fix a positive parameter λ . X is said to have the Exponential distribution with parameter λ or, in short, $X \sim \text{Exp}(\lambda)$, if its density is given by

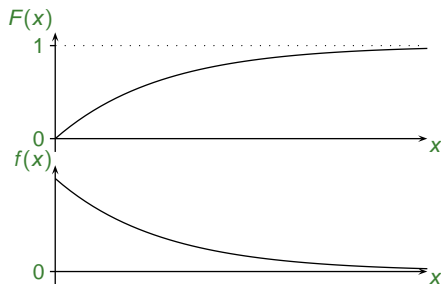
$$f(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ \lambda e^{-\lambda x}, & \text{if } x \geq 0. \end{cases}$$

Remark

Its distribution function can easily be integrated from the density:

$$F(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1 - e^{-\lambda x}, & \text{if } x \geq 0. \end{cases}$$

1. Density, distribution function



$$F(x) = 1 - e^{-\lambda x}; \quad f(x) = \lambda e^{-\lambda x}.$$

2. Expectation, variance

Proposition

For $X \sim \text{Exp}(\lambda)$,

$$\mathbf{E}X = \frac{1}{\lambda}; \quad \mathbf{Var}X = \frac{1}{\lambda^2}.$$



Thinking about X as a waiting time, we now see that λ describes how fast the event we wait for happens. Therefore λ is also called the rate of the exponential waiting time.

2. Expectation, variance

Proof.

We need to compute

$$\mathbf{E}X = \int_0^{\infty} x \lambda e^{-\lambda x} dx \quad \text{and} \quad \mathbf{E}X^2 = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx$$

using integration by parts.  



3. The memoryless property

Proposition

The exponential is the only continuous non-negative memoryless distribution. That is, the only distribution with $X \geq 0$ and

$$\mathbf{P}\{X > t + s \mid X > t\} = \mathbf{P}\{X > s\} \quad (\forall t, s \geq 0).$$

Suppose we have waited for time t . The chance of waiting an additional time s is the same as if we would start waiting anew. The distribution does not remember its past.

3. The memoryless property

Proof.

To prove that the Exponential distribution is memoryless,

$$\begin{aligned}\mathbf{P}\{X > t + s | X > t\} &= \frac{\mathbf{P}\{\{X > t + s\} \cap \{X > t\}\}}{\mathbf{P}\{X > t\}} \\ &= \frac{\mathbf{P}\{X > t + s\}}{\mathbf{P}\{X > t\}} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbf{P}\{X > s\}.\end{aligned}$$

To prove that the Exponential is the only one, \rightsquigarrow  .



3. The memoryless property

Example

Suppose that the length of a phone call in a phone booth is exponentially distributed with mean 10 minutes. If we arrive to an already occupied booth (but there is no one else queuing), what is the probability that we'll wait between 10 and 20 minutes for the booth to free up?

Notice that by the memoryless property we don't care about how long the phone booth has been occupied for. (For all other distributions this would matter!). The remaining time of that person's phone call is $X \sim \text{Exp}(\lambda)$ with $\lambda = 1/\mathbf{E}X = 1/10$, and we calculate

$$\begin{aligned}\mathbf{P}\{10 < X \leq 20\} &= F(20) - F(10) \\ &= 1 - e^{-20/10} - (1 - e^{-10/10}) = e^{-1} - e^{-2} \simeq 0.233.\end{aligned}$$

Normal

The Normal, or Gaussian, is a very nice distribution on its own, but we won't see why it is useful until a bit later.

1. Density, distribution function

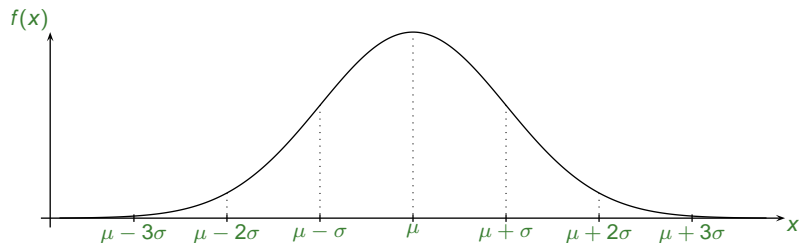
Definition

Let $\mu \in \mathbb{R}$, $\sigma > 0$ be real parameters. X has the Normal distribution with parameters μ and σ^2 or, in short $X \sim \mathcal{N}(\mu, \sigma^2)$, if its density is given by

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (x \in \mathbb{R}).$$

To prove that this is a density, 2-dim. polar coordinates are needed, anyone interested come and see me after class.

1. Density, distribution function



Definition

The case $\mu = 0$, $\sigma^2 = 1$ is called standard normal distribution ($\mathcal{N}(0, 1)$). Its density is denoted by φ , and its distribution function by Φ :

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}, \quad \Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \cdot e^{-y^2/2} dy \quad (x \in \mathbb{R}).$$

1. Density, distribution function

Remark

The standard normal distribution function

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \cdot e^{-y^2/2} dy$$

has *no closed form*, its values will be looked up in tables.

Next we'll establish some tools that will enable us to use such tables to find probabilities of normal random variables.

2. Symmetry

Proposition

For any $z \in \mathbb{R}$,

$$\Phi(-z) = 1 - \Phi(z).$$

Proof.

The standard normal distribution is symmetric: if $X \sim \mathcal{N}(0, 1)$, $-X \sim \mathcal{N}(0, 1)$ as well. Therefore

$$\Phi(-z) = \mathbf{P}\{X < -z\} = \mathbf{P}\{-X > z\} = \mathbf{P}\{X > z\} = 1 - \Phi(z).$$



That's why most tables only have entries for positive values of z .

3. Linear transformations

Proposition

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, and $\alpha, \beta \in \mathbb{R}$ fixed numbers. Then $\alpha X + \beta \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$.

Proof.

We prove for positive α , for negatives it's similar. Start with the distribution function of $Y = \alpha X + \beta$:

$$\begin{aligned} F_Y(y) &= \mathbf{P}\{Y < y\} = \mathbf{P}\{\alpha X + \beta < y\} \\ &= \mathbf{P}\left\{X < \frac{y - \beta}{\alpha}\right\} = F_X\left(\frac{y - \beta}{\alpha}\right). \end{aligned}$$

3. Linear transformations

Proof.

Differentiate this to get

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-\beta}{\alpha}\right) = f_X\left(\frac{y-\beta}{\alpha}\right) \cdot \frac{1}{\alpha} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\frac{y-\beta}{\alpha} - \mu\right)^2}{2\sigma^2}\right) \cdot \frac{1}{\alpha} = \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(y-(\alpha\mu+\beta))^2}{2(\alpha\sigma)^2}}, \end{aligned}$$

which implies the statement: $Y \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$. □

Corollary

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then its standardised version $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
Just use $\alpha = \frac{1}{\sigma}$ *and* $\beta = -\frac{\mu}{\sigma}$.

3. Linear transformations

Proof.

Differentiate this to get

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-\beta}{\alpha}\right) = f_X\left(\frac{y-\beta}{\alpha}\right) \cdot \frac{1}{\alpha} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\frac{y-\beta}{\alpha} - \mu\right)^2}{2\sigma^2}\right) \cdot \frac{1}{\alpha} = \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(y-(\alpha\mu+\beta))^2}{2(\alpha\sigma)^2}}, \end{aligned}$$

which implies the statement: $Y \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$. □

Corollary

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then its standardised version $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
Just use $\alpha = \frac{1}{\sigma}$ *and* $\beta = -\frac{\mu}{\sigma}$.

3. Linear transformations

Proof.

Differentiate this to get

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-\beta}{\alpha}\right) = f_X\left(\frac{y-\beta}{\alpha}\right) \cdot \frac{1}{\alpha} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\frac{y-\beta}{\alpha} - \mu\right)^2}{2\sigma^2}\right) \cdot \frac{1}{\alpha} = \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(y-(\alpha\mu+\beta))^2}{2(\alpha\sigma)^2}}, \end{aligned}$$

which implies the statement: $Y \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$. □

Corollary

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then its standardised version $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
Just use $\alpha = \frac{1}{\sigma}$ *and* $\beta = -\frac{\mu}{\sigma}$.

3. Linear transformations

Proof.

Differentiate this to get

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-\beta}{\alpha}\right) = f_X\left(\frac{y-\beta}{\alpha}\right) \cdot \frac{1}{\alpha} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\frac{y-\beta}{\alpha} - \mu\right)^2}{2\sigma^2}\right) \cdot \frac{1}{\alpha} = \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(y-(\alpha\mu+\beta))^2}{2(\alpha\sigma)^2}}, \end{aligned}$$

which implies the statement: $Y \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$. □

Corollary

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then its standardised version $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
Just use $\alpha = \frac{1}{\sigma}$ *and* $\beta = -\frac{\mu}{\sigma}$.

3. Linear transformations

Proof.

Differentiate this to get

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-\beta}{\alpha}\right) = f_X\left(\frac{y-\beta}{\alpha}\right) \cdot \frac{1}{\alpha} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\frac{y-\beta}{\alpha} - \mu\right)^2}{2\sigma^2}\right) \cdot \frac{1}{\alpha} = \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(y-(\alpha\mu+\beta))^2}{2(\alpha\sigma)^2}}, \end{aligned}$$

which implies the statement: $Y \sim \mathcal{N}(\alpha\mu + \beta, \alpha^2\sigma^2)$. □

Corollary

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then its standardised version $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.
 Just use $\alpha = \frac{1}{\sigma}$ and $\beta = -\frac{\mu}{\sigma}$.

4. Expectation and variance

Proposition

If $X \sim \mathcal{N}(0, 1)$ is standard normal, then its mean is 0 and its variance is 1.

That the mean is zero follows from symmetry.

4. Expectation and variance

Corollary

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then its mean is μ and its variance is σ^2 .

Proof.

$$\mathbf{E}X = \sigma \cdot \mathbf{E}\left(\frac{X - \mu}{\sigma}\right) + \mu = 0 + \mu = \mu,$$

$$\mathbf{Var}X = \sigma^2 \cdot \mathbf{Var}\left(\frac{X - \mu}{\sigma}\right) = \sigma^2 \cdot 1 = \sigma^2$$

as $\frac{X - \mu}{\sigma}$ is standard. □

$\mathcal{N}(\mu, \sigma^2)$ is also said to be the normal distribution with mean μ , variance σ^2 .

5. Example

Example

Let X be normally distributed with mean 3 and variance 4. What is the chance that X is positive?

We have $X \sim \mathcal{N}(3, 4)$, hence $\frac{X-3}{2} \sim \mathcal{N}(0, 1)$:

$$\begin{aligned}\mathbf{P}\{X > 0\} &= \mathbf{P}\left\{\frac{X-3}{2} > \frac{0-3}{2}\right\} = 1 - \Phi(-1.5) \\ &= 1 - [1 - \Phi(1.5)] = \Phi(1.5) \simeq 0.9332.\end{aligned}$$

Make sure you know how to use the table...

5. Example

z	$\Phi(z)$									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936

6. Why Normal?

Theorem (DeMoivre-Laplace)

Fix p , and let $X_n \sim \text{Binom}(n, p)$. Then for every fixed $a < b$ reals,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ a < \frac{X_n - np}{\sqrt{np(1-p)}} \leq b \right\} = \Phi(b) - \Phi(a).$$

That is, take $X \sim \text{Binom}(n, p)$ with large n , fixed (not small) p . Then $\frac{X - np}{\sqrt{np(1-p)}}$ is approximately $\mathcal{N}(0, 1)$ distributed.

This will be a special case of the *Central Limit Theorem*, no proof here. In fact, Normal will appear in many similar scenarios. Measured quantities, heights of people, length of these lectures, etc.

6. Why Normal?

Example

The ideal size of a course is 150 students. On average, 30% of those accepted will enroll, thus the organisers accept 450 students. What is the chance that more than 150 students will enroll?

The number of enrolling students is $X \sim \text{Binom}(450, 0.3)$. With the DeMoivre-Laplace Thm ($np = 135$, $\sqrt{np(1-p)} \simeq 9.72$)

$$\begin{aligned} \mathbf{P}\{X > 150\} &= \mathbf{P}\{X > 150.5\} \simeq \mathbf{P}\left\{\frac{X - 135}{9.72} > \frac{150.5 - 135}{9.72}\right\} \\ &\simeq 1 - \Phi\left(\frac{150.5 - 135}{9.72}\right) \simeq 1 - \Phi(1.59) \simeq 0.0559. \end{aligned}$$

6. Why Normal?

Example

The ideal size of a course is 150 students. On average, 30% of those accepted will enroll, thus the organisers accept 450 students. What is the chance that more than 150 students will enroll?

The number of enrolling students is $X \sim \text{Binom}(450, 0.3)$. With the DeMoivre-Laplace Thm ($np = 135$, $\sqrt{np(1-p)} \simeq 9.72$)

$$\begin{aligned} \mathbf{P}\{X > 150\} &= \mathbf{P}\{X > 150.5\} \simeq \mathbf{P}\left\{\frac{X - 135}{9.72} > \frac{150.5 - 135}{9.72}\right\} \\ &\simeq 1 - \Phi\left(\frac{150.5 - 135}{9.72}\right) \simeq 1 - \Phi(1.59) \simeq 0.0559. \end{aligned}$$



6. Why Normal?

Example

The ideal size of a course is 150 students. On average, 30% of those accepted will enroll, thus the organisers accept 450 students. What is the chance that more than 150 students will enroll?

The number of enrolling students is $X \sim \text{Binom}(450, 0.3)$. With the DeMoivre-Laplace Thm ($np = 135$, $\sqrt{np(1-p)} \simeq 9.72$)

$$\begin{aligned} \mathbf{P}\{X > 150\} &= \mathbf{P}\{X > 150.5\} \simeq \mathbf{P}\left\{\frac{X - 135}{9.72} > \frac{150.5 - 135}{9.72}\right\} \\ &\simeq 1 - \Phi\left(\frac{150.5 - 135}{9.72}\right) \simeq 1 - \Phi(1.59) \simeq 0.0559. \end{aligned}$$

Transformations

Let X be a random variable, and $g(X)$ a function of it. If X is discrete then the distribution of $g(X)$ is rather straightforward. In the continuous case the question is more interesting.

We have in fact seen an example before: an *affine transformation* $g(x) = ax + b$ of Normal keeps it Normal. We'll see more examples, and then a general statement about this phenomenon.

Transformations

Example

Let X be a continuous random variable with density f_X .
Determine the density of $Y := X^2$.

Transformations

Solution

Fix $y > 0$ (for $y \leq 0$ the density is trivially zero), and start with the distribution function:

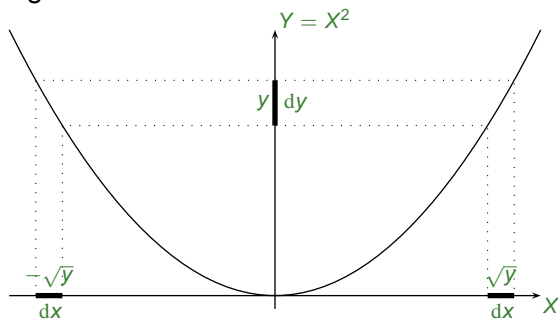
$$\begin{aligned}F_Y(y) &= \mathbf{P}\{Y < y\} = \mathbf{P}\{X^2 < y\} = \mathbf{P}\{-\sqrt{y} < X < \sqrt{y}\} \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}).\end{aligned}$$

Differentiate this to get

$$\begin{aligned}f_Y(y) &= F'_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + F'_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \\ &= f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}}.\end{aligned}$$

Transformations

What is going on?

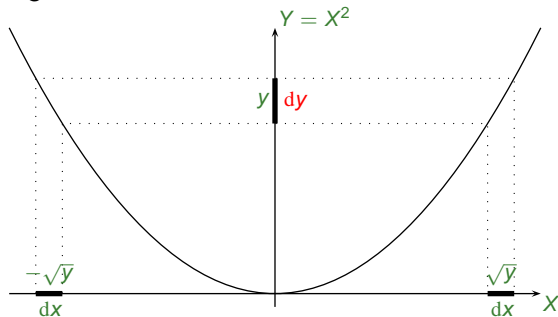


$$f_Y(y) dy = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} dy + f_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} dy.$$



Transformations

What is going on?

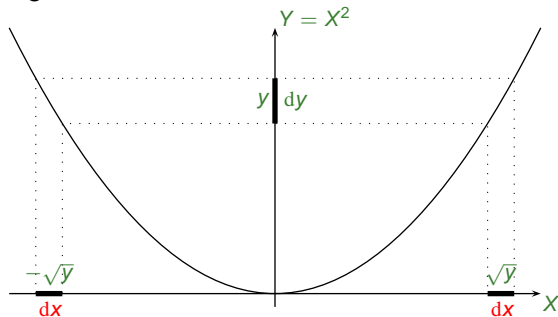


$$f_Y(y) dy = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} dy + f_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} dy.$$



Transformations

What is going on?



$$f_Y(y) dy = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} dy + f_X(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} dy.$$



Transformations

There is a general formula along the same lines:

Proposition

Let X be a continuous random variable with density f_X , and g a continuously differentiable function with nonzero derivative. Then the density of $Y = g(X)$ is given by

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}.$$

Proving this can be done in a way very similar to the scheme above.

The scheme is important.

5. Joint distributions

Joint distributions

Independence, convolutions

Objectives:

- ▶ To build a mathematical model for several random variables on a common probability space
- ▶ To get familiar with joint, marginal and conditional discrete distributions
- ▶ To understand discrete convolutions
- ▶ To get familiar with the Gamma distribution (via a continuous convolution)

Joint distributions

Often an experiment can result in several random quantities at the same time. In this case we have several random variables defined on a common probability space. Their relations can be far from trivial, and are described by *joint distributions*. Here we'll familiarise ourselves with the basics of joint distributions.

For most part we restrict our attention to the discrete case, as the jointly continuous case would require multivariable calculus and more time.

At the end of this chapter we introduce the Gamma distribution, purely motivated by joint distributions.

1. Joint mass function

Most examples will involve two random variables, but everything can be generalised for more of them.

Definition

Suppose two discrete random variables X and Y are defined on a common probability space, and can take on values x_1, x_2, \dots and y_1, y_2, \dots , respectively. The joint probability mass function of them is defined as

$$p(x_i, y_j) = \mathbf{P}\{X = x_i, Y = y_j\}, \quad i = 1, 2, \dots, j = 1, 2, \dots$$

This function contains all information about the joint distribution of X and Y .

1. Joint mass function

Definition

The marginal mass functions are

$$p_X(x_i) := \mathbf{P}\{X = x_i\}, \quad \text{and} \quad p_Y(y_j) := \mathbf{P}\{Y = y_j\}.$$

It is clear from the Law of Total Probability that

Proposition

$$p_X(x_i) = \sum_j p(x_i, y_j), \quad \text{and} \quad p_Y(y_j) = \sum_i p(x_i, y_j).$$



1. Joint mass function

Proposition

Any joint mass function satisfies

- ▶ $p(\mathbf{x}, \mathbf{y}) \geq 0, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n;$
- ▶ $\sum_{i,j} p(x_i, y_j) = 1.$

Vice versa: any function p which is only non-zero in countably many (x_i, y_j) values, and which has the above properties, is a joint probability mass function. There is a sample space and random variables that realise this joint mass function.

1. Joint mass function

Proof.

Non-negativity is clear. For the double sum,

$$\sum_{i,j} p(x_i, y_j) = \sum_i \sum_j p(x_i, y_j) = \sum_i p_X(x_i) = 1.$$



1. Joint mass function

Example

An urn has 3 red, 4 white, 5 black balls. Drawing 3 at once, let X be the number of red, Y the number of white balls drawn. The joint mass function is:

$Y \setminus X$	0	1	2	3	$p_Y(\cdot)$
0	$\frac{\binom{5}{3}}{\binom{12}{3}}$	$\frac{\binom{3}{1} \cdot \binom{5}{2}}{\binom{12}{3}}$	$\frac{\binom{3}{2} \cdot \binom{5}{1}}{\binom{12}{3}}$	$\frac{1}{\binom{12}{3}}$	$\frac{\binom{8}{3}}{\binom{12}{3}}$
1	$\frac{\binom{4}{1} \cdot \binom{5}{2}}{\binom{12}{3}}$	$\frac{\binom{4}{1} \cdot \binom{3}{1} \cdot \binom{5}{1}}{\binom{12}{3}}$	$\frac{\binom{4}{1} \cdot \binom{3}{2}}{\binom{12}{3}}$	0	$\frac{\binom{4}{1} \cdot \binom{8}{2}}{\binom{12}{3}}$
2	$\frac{\binom{4}{2} \cdot \binom{5}{1}}{\binom{12}{3}}$	$\frac{\binom{4}{2} \cdot \binom{3}{1}}{\binom{12}{3}}$	0	0	$\frac{\binom{4}{2} \cdot \binom{8}{1}}{\binom{12}{3}}$
3	$\frac{\binom{3}{3}}{\binom{12}{3}}$	0	0	0	$\frac{\binom{3}{3}}{\binom{12}{3}}$
$p_X(\cdot)$	$\frac{\binom{9}{3}}{\binom{12}{3}}$	$\frac{\binom{3}{1} \cdot \binom{9}{2}}{\binom{12}{3}}$	$\frac{\binom{3}{2} \cdot \binom{9}{1}}{\binom{12}{3}}$	$\frac{1}{\binom{12}{3}}$	1

2. Conditional mass function

Definition

Suppose $p_Y(y_j) > 0$. The conditional mass function of X , given $Y = y_j$ is defined by

$$p_{X|Y}(x | y_j) := \mathbf{P}\{X = x | Y = y_j\} = \frac{p(x, y_j)}{p_Y(y_j)}.$$

As the conditional probability was a proper probability, this is a proper mass function: $\forall x, y_j,$

$$p_{X|Y}(x | y_j) \geq 0, \quad \sum_i p_{X|Y}(x_i | y_j) = 1.$$

2. Conditional mass function

Example

Let X and Y have joint mass function

$X \backslash Y$	0	1
0	0.4	0.2
1	0.1	0.3

The conditional distribution of X given $Y = 0$ is

$$p_{X|Y}(0|0) = \frac{p(0,0)}{p_Y(0)} = \frac{p(0,0)}{p(0,0) + p(1,0)} = \frac{0.4}{0.4 + 0.1} = \frac{4}{5},$$

$$p_{X|Y}(1|0) = \frac{p(1,0)}{p_Y(0)} = \frac{p(1,0)}{p(0,0) + p(1,0)} = \frac{0.1}{0.4 + 0.1} = \frac{1}{5}.$$

Independence, convolutions

An important special case of joint distributions is the one of *independent* variables: whatever the value of some of them is, it does not influence the distribution of the others. We'll make this precise in this part, and then use it to determine the distribution of the sum of independent variables.

As a slight generalisation and application, we'll also introduce the *Gamma* distribution.

1. Independent r.v.'s

Definition

Random variables X and Y are independent, if events formulated with them are so. That is, if for every $A, B \subseteq \mathbb{R}$

$$\mathbf{P}\{X \in A, Y \in B\} = \mathbf{P}\{X \in A\} \cdot \mathbf{P}\{Y \in B\}.$$

Similarly, random variables X_1, X_2, \dots are independent, if events formulated with them are so. That is, if for every $A_{i_1}, A_{i_2}, \dots, A_{i_n} \subseteq \mathbb{R}$

$$\begin{aligned} \mathbf{P}\{X_{i_1} \in A_{i_1}, X_{i_2} \in A_{i_2}, \dots, X_{i_n} \in A_{i_n}\} \\ = \mathbf{P}\{X_{i_1} \in A_{i_1}\} \cdot \mathbf{P}\{X_{i_2} \in A_{i_2}\} \cdots \mathbf{P}\{X_{i_n} \in A_{i_n}\}. \end{aligned}$$

Recall *mutual independence* for events...

1. Independent r.v.'s

Remark

People use the abbreviation i.i.d. for **i**ndependent and **i**dentically **d**istributed random variables.

Proposition

Two random variables X and Y are independent if and only if their joint mass function factorises into the product of the marginals:

$$p(x_i, y_j) = p_X(x_i) \cdot p_Y(y_j), \quad (\forall x_i, y_j).$$



1. Independent r.v.'s

Example (a trivial one)

Rolling two dice, let X and Y be the two numbers shown on them. Then every pair of numbers have equal probability:

$$p(i, j) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = p_X(i) \cdot p_Y(j) \quad (\forall i, j = 1, \dots, 6),$$

and we see that these variables are independent (in fact i.i.d. as well).

2. Discrete convolution

We restrict ourselves now to integer valued random variables. Let X and Y be such, and also independent. What is the distribution of their sum?

Proposition

Let X and Y be independent, integer valued random variables with respective mass functions p_X and p_Y . Then

$$p_{X+Y}(k) = \sum_{i=-\infty}^{\infty} p_X(k-i) \cdot p_Y(i), \quad (\forall k \in \mathbb{Z}).$$

This formula is called the discrete convolution of the mass functions p_X and p_Y .

2. Discrete convolution

Proof.

Using the Law of Total Probability, and independence,

$$\begin{aligned} p_{X+Y}(k) &= \mathbf{P}\{X + Y = k\} = \sum_{i=-\infty}^{\infty} \mathbf{P}\{X + Y = k, Y = i\} \\ &= \sum_{i=-\infty}^{\infty} \mathbf{P}\{X = k - i, Y = i\} = \sum_{i=-\infty}^{\infty} p_X(k - i) \cdot p_Y(i). \end{aligned}$$

□

2. Discrete convolution

Proof.

Using the **Law of Total Probability**, and independence,

$$\begin{aligned} p_{X+Y}(k) &= \mathbf{P}\{X + Y = k\} = \sum_{i=-\infty}^{\infty} \mathbf{P}\{X + Y = k, Y = i\} \\ &= \sum_{i=-\infty}^{\infty} \mathbf{P}\{X = k - i, Y = i\} = \sum_{i=-\infty}^{\infty} p_X(k - i) \cdot p_Y(i). \end{aligned}$$

□

2. Discrete convolution

Proof.

Using the Law of Total Probability, and **independence**,

$$\begin{aligned} p_{X+Y}(k) &= \mathbf{P}\{X + Y = k\} = \sum_{i=-\infty}^{\infty} \mathbf{P}\{X + Y = k, Y = i\} \\ &= \sum_{i=-\infty}^{\infty} \mathbf{P}\{X = k - i, Y = i\} = \sum_{i=-\infty}^{\infty} p_X(k - i) \cdot p_Y(i). \end{aligned}$$

□

2. Discrete convolution

Example

Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ be independent. Then $X + Y \sim \text{Poi}(\lambda + \mu)$.

Example (of course...)

Let $X \sim \text{Binom}(n, p)$ and $Y \sim \text{Binom}(m, p)$ be independent (notice the same $p!$). Then $X + Y \sim \text{Binom}(n + m, p)$.

3. Continuous convolution

Recall the discrete convolution formula

$$p_{X+Y}(k) = \sum_{i=-\infty}^{\infty} p_X(k-i) \cdot p_Y(i), \quad (\forall k \in \mathbb{Z}).$$

In a very similar way we state without proof the continuous convolution formula for densities:

Proposition

Suppose X and Y are independent continuous random variables with respective densities f_X and f_Y . Then their sum is a continuous random variable with density

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy, \quad (\forall a \in \mathbb{R}).$$

4. Gamma distribution

We'll only use the continuous convolution formula on a particular case:

Let X and Y be i.i.d. $\text{Exp}(\lambda)$, and see the density of their sum ($a \geq 0$):

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy = \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda e^{-\lambda y} dy \\ &= \lambda^2 a \cdot e^{-\lambda a}. \end{aligned}$$

This density is called the $\text{Gamma}(2, \lambda)$ density.

4. Gamma distribution

We'll only use the continuous convolution formula on a particular case:

Let X and Y be i.i.d. $\text{Exp}(\lambda)$, and see the density of their sum ($a \geq 0$):

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy = \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda e^{-\lambda y} dy \\ &= \lambda^2 a \cdot e^{-\lambda a}. \end{aligned}$$

This density is called the $\text{Gamma}(2, \lambda)$ density.

4. Gamma distribution

We'll only use the continuous convolution formula on a particular case:

Let X and Y be i.i.d. $\text{Exp}(\lambda)$, and see the density of their sum ($a \geq 0$):

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy = \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda e^{-\lambda y} dy \\ &= \lambda^2 a \cdot e^{-\lambda a}. \end{aligned}$$

This density is called the $\text{Gamma}(2, \lambda)$ density.

4. Gamma distribution

We'll only use the continuous convolution formula on a particular case:

Let X and Y be i.i.d. $\text{Exp}(\lambda)$, and see the density of their sum ($a \geq 0$):

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy = \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda e^{-\lambda y} dy \\ &= \lambda^2 a \cdot e^{-\lambda a}. \end{aligned}$$

This density is called the $\text{Gamma}(2, \lambda)$ density.

4. Gamma distribution

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(2, \lambda)$ be independent.
Again,

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda^2 y e^{-\lambda y} dy = \frac{\lambda^3 a^2 e^{-\lambda a}}{2}. \end{aligned}$$

This is the $\text{Gamma}(3, \lambda)$ density.

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(3, \lambda)$ be independent.

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \frac{\lambda^3 y^2 e^{-\lambda y}}{2} dy = \frac{\lambda^4 a^3 e^{-\lambda a}}{2 \cdot 3}. \end{aligned}$$

This is the $\text{Gamma}(4, \lambda)$ density.

4. Gamma distribution

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(2, \lambda)$ be independent.
Again,

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda^2 y e^{-\lambda y} dy = \frac{\lambda^3 a^2 e^{-\lambda a}}{2}. \end{aligned}$$

This is the $\text{Gamma}(3, \lambda)$ density.

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(3, \lambda)$ be independent.

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \frac{\lambda^3 y^2 e^{-\lambda y}}{2} dy = \frac{\lambda^4 a^3 e^{-\lambda a}}{2 \cdot 3}. \end{aligned}$$

This is the $\text{Gamma}(4, \lambda)$ density.

4. Gamma distribution

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(2, \lambda)$ be independent.
Again,

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda^2 y e^{-\lambda y} dy = \frac{\lambda^3 a^2 e^{-\lambda a}}{2}. \end{aligned}$$

This is the $\text{Gamma}(3, \lambda)$ density.

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(3, \lambda)$ be independent.

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \frac{\lambda^3 y^2 e^{-\lambda y}}{2} dy = \frac{\lambda^4 a^3 e^{-\lambda a}}{2 \cdot 3}. \end{aligned}$$

This is the $\text{Gamma}(4, \lambda)$ density.

4. Gamma distribution

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(2, \lambda)$ be independent.
Again,

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda^2 y e^{-\lambda y} dy = \frac{\lambda^3 a^2 e^{-\lambda a}}{2}. \end{aligned}$$

This is the $\text{Gamma}(3, \lambda)$ density.

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(3, \lambda)$ be independent.

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \frac{\lambda^3 y^2 e^{-\lambda y}}{2} dy = \frac{\lambda^4 a^3 e^{-\lambda a}}{2 \cdot 3}. \end{aligned}$$

This is the $\text{Gamma}(4, \lambda)$ density.

4. Gamma distribution

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(2, \lambda)$ be independent.
Again,

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda^2 y e^{-\lambda y} dy = \frac{\lambda^3 a^2 e^{-\lambda a}}{2}. \end{aligned}$$

This is the $\text{Gamma}(3, \lambda)$ density.

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(3, \lambda)$ be independent.

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \frac{\lambda^3 y^2 e^{-\lambda y}}{2} dy = \frac{\lambda^4 a^3 e^{-\lambda a}}{2 \cdot 3}. \end{aligned}$$

This is the $\text{Gamma}(4, \lambda)$ density.

4. Gamma distribution

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(2, \lambda)$ be independent.
Again,

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda^2 y e^{-\lambda y} dy = \frac{\lambda^3 a^2 e^{-\lambda a}}{2}. \end{aligned}$$

This is the $\text{Gamma}(3, \lambda)$ density.

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(3, \lambda)$ be independent.

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \frac{\lambda^3 y^2 e^{-\lambda y}}{2} dy = \frac{\lambda^4 a^3 e^{-\lambda a}}{2 \cdot 3}. \end{aligned}$$

This is the $\text{Gamma}(4, \lambda)$ density.

4. Gamma distribution

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(2, \lambda)$ be independent.
Again,

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda^2 y e^{-\lambda y} dy = \frac{\lambda^3 a^2 e^{-\lambda a}}{2}. \end{aligned}$$

This is the $\text{Gamma}(3, \lambda)$ density.

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(3, \lambda)$ be independent.

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \frac{\lambda^3 y^2 e^{-\lambda y}}{2} dy = \frac{\lambda^4 a^3 e^{-\lambda a}}{2 \cdot 3}. \end{aligned}$$

This is the $\text{Gamma}(4, \lambda)$ density.

4. Gamma distribution

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(2, \lambda)$ be independent.
Again,

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \lambda^2 y e^{-\lambda y} dy = \frac{\lambda^3 a^2 e^{-\lambda a}}{2}. \end{aligned}$$

This is the $\text{Gamma}(3, \lambda)$ density.

Now, let $X \sim \text{Exp}(\lambda)$, and $Y \sim \text{Gamma}(3, \lambda)$ be independent.

$$\begin{aligned} f_{X+Y}(a) &= \int_{-\infty}^{\infty} f_X(a-y) \cdot f_Y(y) dy \\ &= \int_0^a \lambda e^{-\lambda(a-y)} \cdot \frac{\lambda^3 y^2 e^{-\lambda y}}{2} dy = \frac{\lambda^4 a^3 e^{-\lambda a}}{2 \cdot 3}. \end{aligned}$$

This is the $\text{Gamma}(4, \lambda)$ density.

4. Gamma distribution

Inductively,

Proposition

The convolution of n i.i.d. $\text{Exp}(\lambda)$ distributions results in the Gamma(n, λ) density:

$$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}, \quad \forall x \geq 0$$

and zero otherwise.

This is the density of the sum of n i.i.d. $\text{Exp}(\lambda)$ random variables. In particular, $\text{Gamma}(1, \lambda) \equiv \text{Exp}(\lambda)$.

4. Gamma distribution

Corollary

Make a change $z = \lambda x$ of the integration variable, and write

$$\int_0^{\infty} f(x) dx = \int_0^{\infty} \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} dx = \int_0^{\infty} \frac{(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!} \lambda dx = 1:$$

$$(n-1)! = \int_0^{\infty} z^{n-1} e^{-z} dz.$$

Definition

The gamma function is defined, for every $\alpha > 0$ real numbers, by

$$\Gamma(\alpha) := \int_0^{\infty} z^{\alpha-1} e^{-z} dz.$$

In particular, $\Gamma(n) = (n-1)!$ for positive integer n 's.

4. Gamma distribution

Corollary

Make a change $z = \lambda x$ of the integration variable, and write

$$\int_0^{\infty} f(x) dx = \int_0^{\infty} \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} dx = \int_0^{\infty} \frac{(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!} \lambda dx = 1:$$

$$(n-1)! = \int_0^{\infty} z^{n-1} e^{-z} dz.$$

Definition

The gamma function is defined, for every $\alpha > 0$ real numbers, by

$$\Gamma(\alpha) := \int_0^{\infty} z^{\alpha-1} e^{-z} dz.$$

In particular, $\Gamma(n) = (n-1)!$ for positive integer n 's.

4. Gamma distribution

Corollary

Make a change $z = \lambda x$ of the integration variable, and write

$$\int_0^{\infty} f(x) dx = \int_0^{\infty} \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} dx = \int_0^{\infty} \frac{(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!} \lambda dx = 1:$$

$$(n-1)! = \int_0^{\infty} z^{n-1} e^{-z} dz.$$

Definition

The gamma function is defined, for every $\alpha > 0$ real numbers, by

$$\Gamma(\alpha) := \int_0^{\infty} z^{\alpha-1} e^{-z} dz.$$

In particular, $\Gamma(n) = (n-1)!$ for positive integer n 's.

4. Gamma distribution

Corollary

Make a change $z = \lambda x$ of the integration variable, and write

$$\int_0^{\infty} f(x) dx = \int_0^{\infty} \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} dx = \int_0^{\infty} \frac{(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!} \lambda dx = 1:$$

$$(n-1)! = \int_0^{\infty} z^{n-1} e^{-z} dz.$$

Definition

The gamma function is defined, for every $\alpha > 0$ real numbers, by

$$\Gamma(\alpha) := \int_0^{\infty} z^{\alpha-1} e^{-z} dz.$$

In particular, $\Gamma(n) = (n-1)!$ for positive integer n 's.

4. Gamma distribution

Corollary

Make a change $z = \lambda x$ of the integration variable, and write

$$\int_0^{\infty} f(x) dx = \int_0^{\infty} \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} dx = \int_0^{\infty} \frac{(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!} \lambda dx = 1:$$

$$(n-1)! = \int_0^{\infty} z^{n-1} e^{-z} dz.$$

Definition

The gamma function is defined, for every $\alpha > 0$ real numbers, by

$$\Gamma(\alpha) := \int_0^{\infty} z^{\alpha-1} e^{-z} dz.$$

In particular, $\Gamma(n) = (n-1)!$ for positive integer n 's.

4. Gamma distribution

Integration by parts yields

Proposition

$$\Gamma(1) = 1, \quad \text{and} \quad \Gamma(\alpha + 1) = \alpha \cdot \Gamma(\alpha), \quad (\forall \alpha > 0).$$



With this tool in hand, we can generalise to

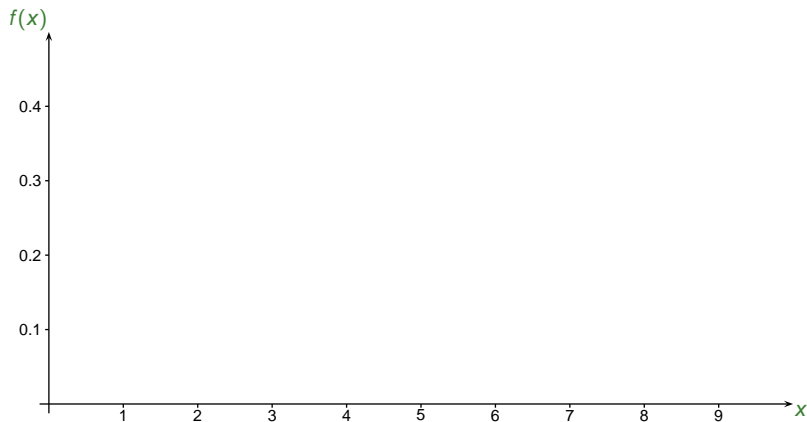
Definition

The Gamma(α , λ) distribution of positive real shape and rate parameters α and λ is the one with density

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad \forall x \geq 0$$

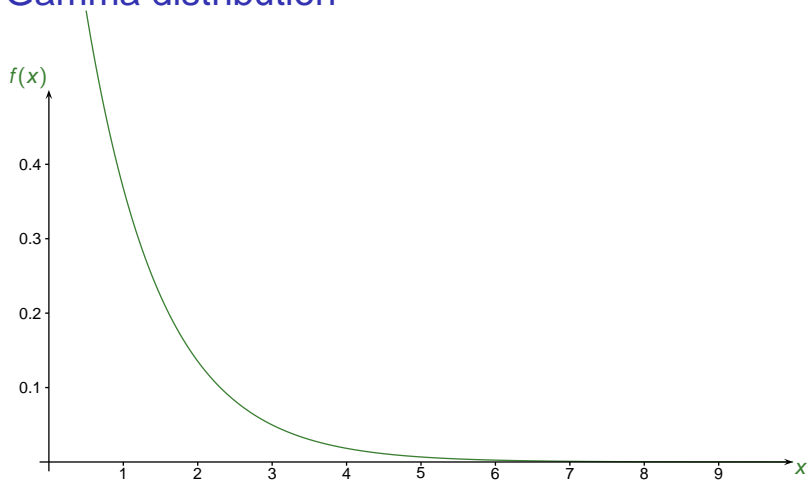
and zero otherwise.

4. Gamma distribution



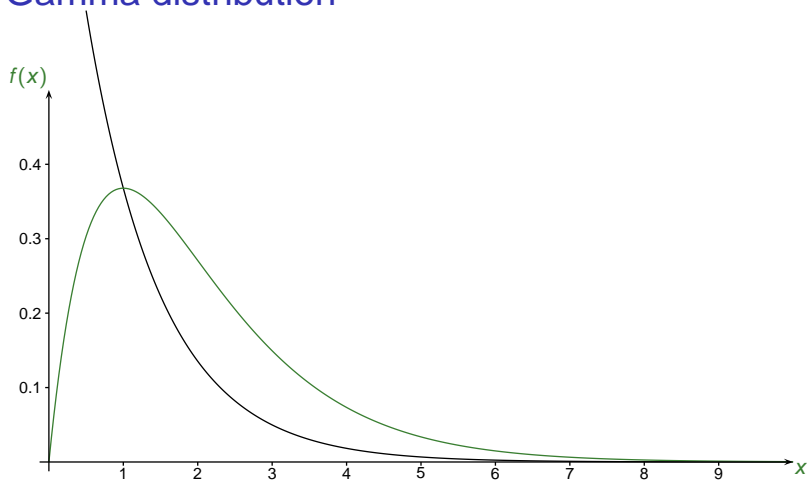
$$\text{Gamma}(\alpha, \lambda) : \quad f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$$

4. Gamma distribution



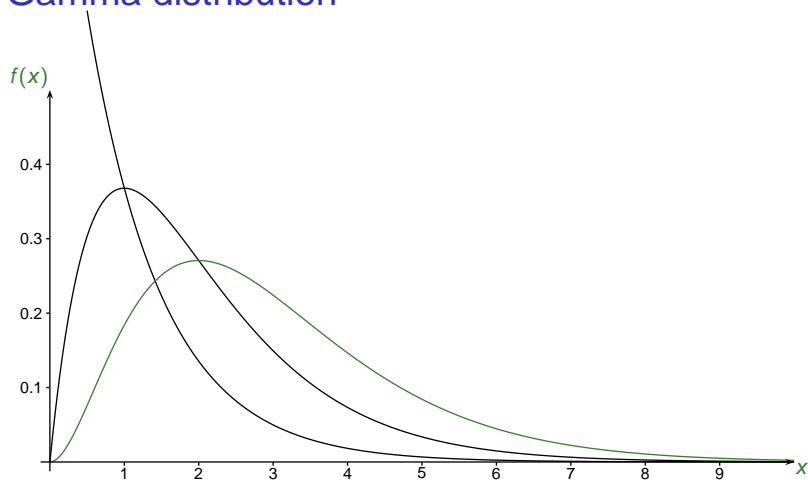
$$\text{Gamma}(1, 1) : \quad f(x) = \frac{1^1 x^{1-1} e^{-1x}}{\Gamma(1)} = e^{-x} \quad \sim \text{Exp}(1)$$

4. Gamma distribution



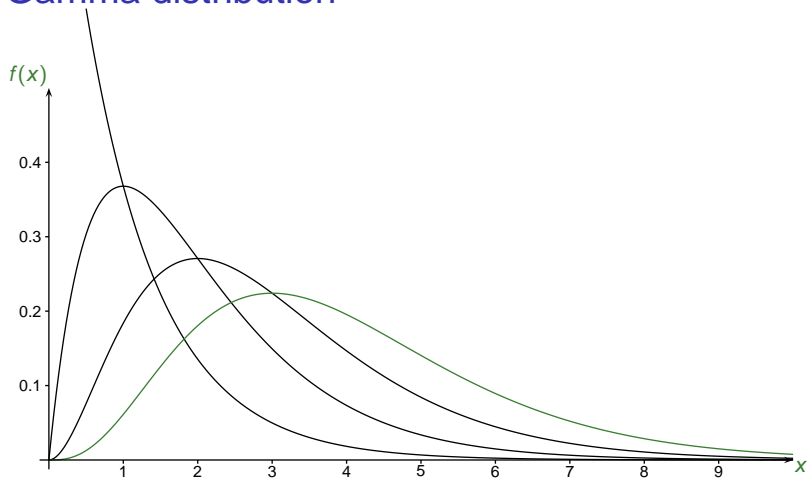
Gamma(2, 1) :
$$f(x) = \frac{1^2 x^{2-1} e^{-1x}}{\Gamma(2)} = x e^{-x}$$

4. Gamma distribution



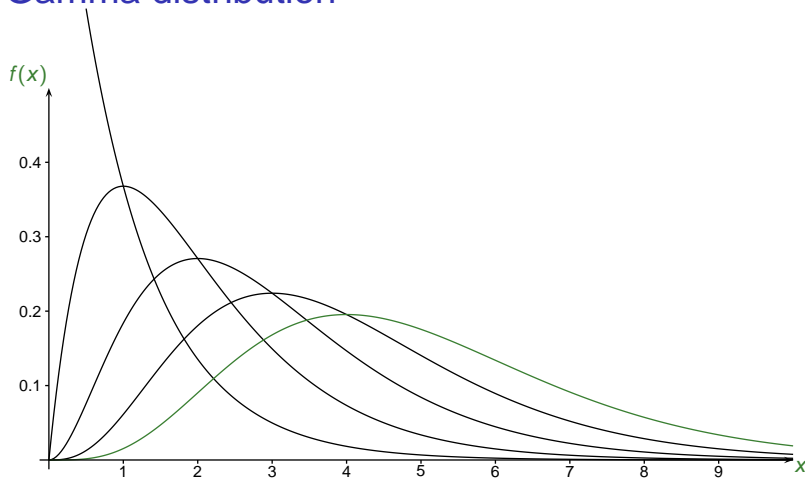
Gamma(3, 1) :
$$f(x) = \frac{1^3 x^{3-1} e^{-1x}}{\Gamma(3)} = x^2 e^{-x} / 2$$

4. Gamma distribution



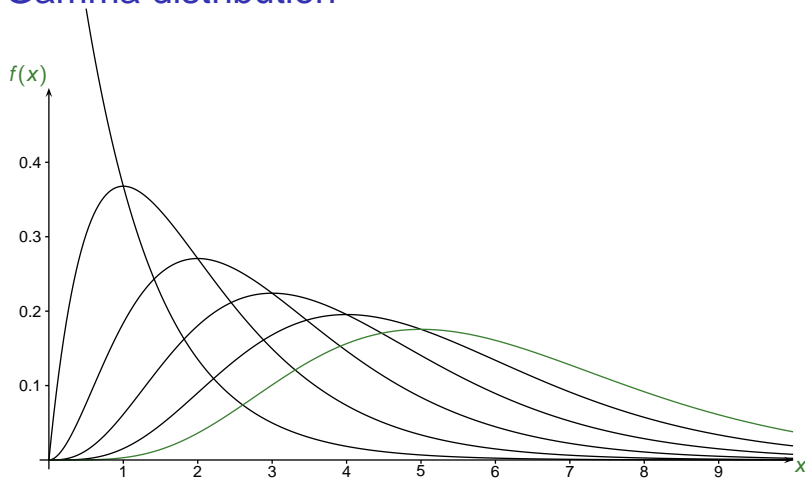
Gamma(4, 1) :
$$f(x) = \frac{1^4 x^{4-1} e^{-1x}}{\Gamma(4)} = x^3 e^{-x} / 6$$

4. Gamma distribution



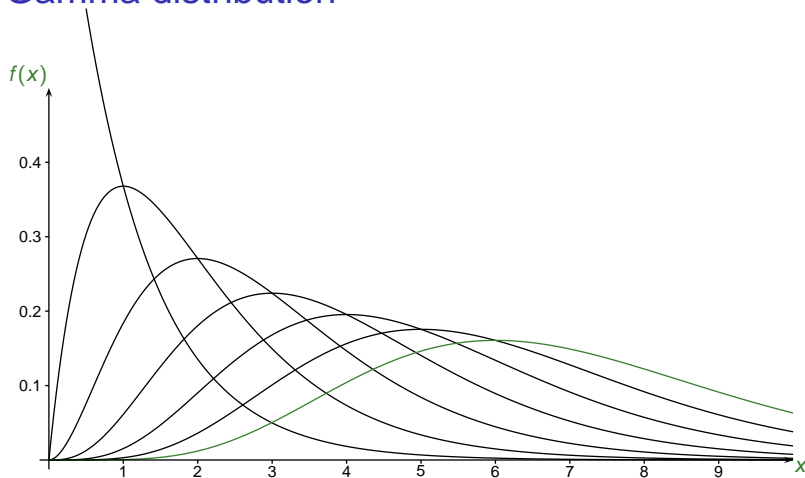
Gamma(5, 1) :
$$f(x) = \frac{1^5 x^{5-1} e^{-1x}}{\Gamma(5)} = x^4 e^{-x} / 24$$

4. Gamma distribution



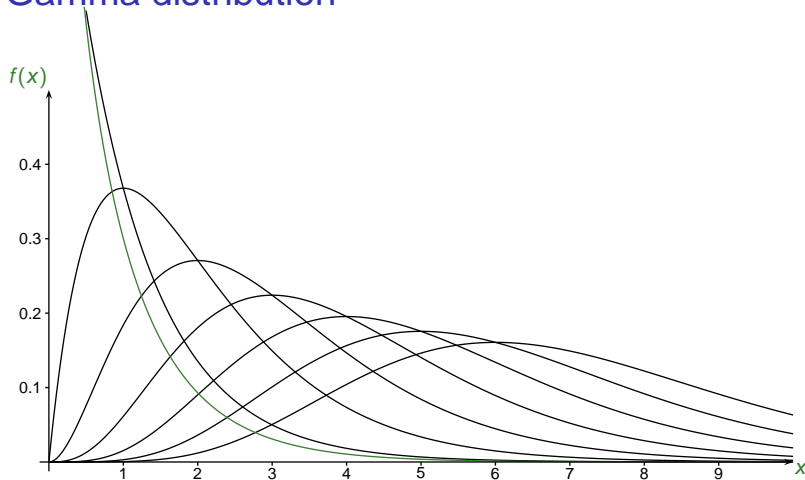
Gamma(6, 1) :
$$f(x) = \frac{1^6 x^{6-1} e^{-1x}}{\Gamma(6)} = x^5 e^{-x} / 120$$

4. Gamma distribution



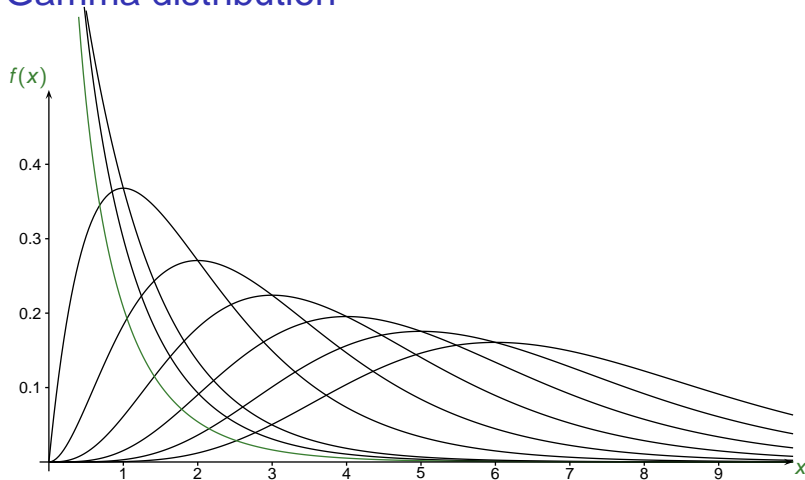
Gamma(7, 1) :
$$f(x) = \frac{1^7 x^{7-1} e^{-1x}}{\Gamma(7)} = x^6 e^{-x} / 720$$

4. Gamma distribution



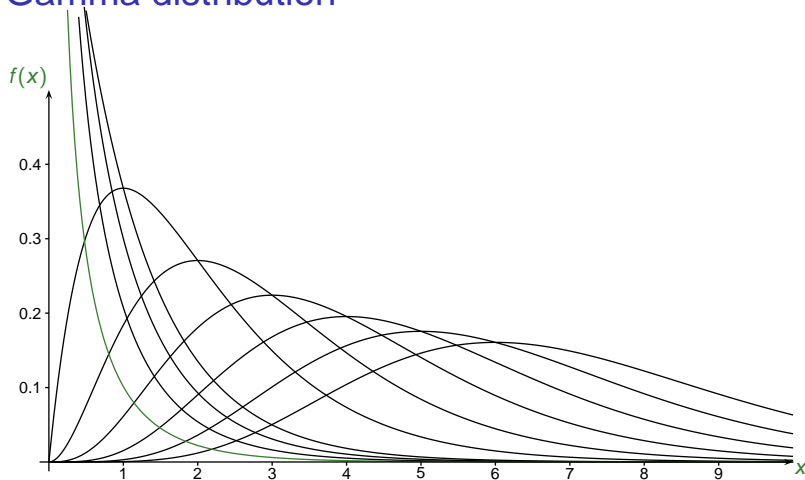
Gamma(0.75, 1) :
$$f(x) = \frac{10.75 x^{0.75-1} e^{-1x}}{\Gamma(0.75)} \simeq x^{-0.25} e^{-x} / 1.23$$

4. Gamma distribution



Gamma(0.5, 1) :
$$f(x) = \frac{1^{0.5} x^{0.5-1} e^{-1x}}{\Gamma(0.5)} \simeq x^{-0.5} e^{-x} / 1.72$$

4. Gamma distribution



Gamma(0.25, 1) : $f(x) = \frac{1^{0.25} x^{0.25-1} e^{-1x}}{\Gamma(0.25)} \simeq x^{-0.75} e^{-x} / 3.63$

4. Gamma distribution

Proposition

If $X \sim \text{Gamma}(\alpha, \lambda)$, then

$$\mathbf{E}X = \frac{\alpha}{\lambda}, \quad \mathbf{Var}X = \frac{\alpha}{\lambda^2}.$$

This is rather natural from the sum of i.i.d. Exponentials when α is integer, and can easily be integrated out in all cases.

6. Expectation, covariance

Properties of expectations

Covariance

Conditional expectation

Moment generating functions

Objectives:

- ▶ To explore further properties of expectations of a single and multiple variables
- ▶ To define covariance, and use it for computing variances of sums
- ▶ To explore and use conditional expectations
- ▶ To define and use moment generating functions

Properties of expectations

Recall the respective definitions

$$\mathbf{E}X = \sum_i x_i p(x_i) \quad \text{or} \quad \mathbf{E}X = \int_{-\infty}^{\infty} xf(x) dx$$

for the discrete and continuous cases. In this chapter we'll explore properties of expected values. We'll always assume that the expectations we talk about exist. Most proofs will be done for the discrete case, but everything in this chapter is very general **even beyond discrete and continuous...**

1. A simple monotonicity property

Proposition

Suppose that $a \leq X \leq b$ a.s. Then $a \leq \mathbf{E}X \leq b$.

Recall: a.s. means *with probability one*.

Proof.

Due to the assumption, all possible values satisfy $a \leq x_i \leq b$.

Therefore

$$a = a \cdot 1 = \sum_i a p(x_i) \leq \sum_i x_i p(x_i) \leq \sum_i b p(x_i) = b \cdot 1 = b.$$



1. A simple monotonicity property

Proposition

Suppose that $a \leq X \leq b$ a.s. Then $a \leq \mathbf{E}X \leq b$.

Recall: a.s. means *with probability one*.

Proof.

Due to the assumption, all possible values satisfy $a \leq x_i \leq b$.

Therefore

$$a = a \cdot 1 = \sum_i a p(x_i) \leq \sum_i x_i p(x_i) \leq \sum_i b p(x_i) = b \cdot 1 = b.$$



1. A simple monotonicity property

Proposition

Suppose that $a \leq X \leq b$ a.s. Then $a \leq \mathbf{E}X \leq b$.

Recall: a.s. means *with probability one*.

Proof.

Due to the assumption, all possible values satisfy $a \leq x_i \leq b$.

Therefore

$$a = a \cdot 1 = \sum_i a p(x_i) \leq \sum_i x_i p(x_i) \leq \sum_i b p(x_i) = b \cdot 1 = b.$$



1. A simple monotonicity property

Proposition

Suppose that $a \leq X \leq b$ a.s. Then $a \leq \mathbf{EX} \leq b$.

Recall: a.s. means *with probability one*.

Proof.

Due to the assumption, all possible values satisfy $a \leq x_i \leq b$.

Therefore

$$a = a \cdot 1 = \sum_i a p(x_i) \leq \sum_i x_i p(x_i) \leq \sum_i b p(x_i) = b \cdot 1 = b.$$



2. Expectation of functions of variables

Proposition


Suppose that X and Y are discrete random variables, and $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ function. Then

$$\mathbf{E}g(X, Y) = \sum_{i,j} g(x_i, y_j) \cdot p(x_i, y_j).$$

There is a very analogous formula for continuous random variables, using *joint densities*, beyond the scope of this unit.

A similar formula holds for functions of 3, 4, etc. random variables.

Proof.

The proof goes as in the one-variable case \rightsquigarrow  .



3. Expectation of sums and differences

Corollary (a very important one)

Let X and Y be any random variables. Then

$$\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y \quad \text{and} \quad \mathbf{E}(X - Y) = \mathbf{E}X - \mathbf{E}Y.$$

Proof.

$$\begin{aligned}\mathbf{E}(X \pm Y) &= \sum_{i,j} (x_i \pm y_j) p(x_i, y_j) \\ &= \sum_i \sum_j x_i p(x_i, y_j) \pm \sum_j \sum_i y_j p(x_i, y_j) \\ &= \sum_i x_i p_X(x_i) \pm \sum_j y_j p_Y(y_j) = \mathbf{E}X \pm \mathbf{E}Y.\end{aligned}$$



3. Expectation of sums and differences

Corollary (a very important one)

Let X and Y be any random variables. Then

$$\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y \quad \text{and} \quad \mathbf{E}(X - Y) = \mathbf{E}X - \mathbf{E}Y.$$

Proof.

$$\begin{aligned}\mathbf{E}(X \pm Y) &= \sum_{i,j} (x_i \pm y_j) p(x_i, y_j) \\ &= \sum_i \sum_j x_i p(x_i, y_j) \pm \sum_j \sum_i y_j p(x_i, y_j) \\ &= \sum_i x_i p_X(x_i) \pm \sum_j y_j p_Y(y_j) = \mathbf{E}X \pm \mathbf{E}Y.\end{aligned}$$



3. Expectation of sums and differences

Corollary (a very important one)

Let X and Y be any random variables. Then

$$\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y \quad \text{and} \quad \mathbf{E}(X - Y) = \mathbf{E}X - \mathbf{E}Y.$$

Proof.

$$\begin{aligned} \mathbf{E}(X \pm Y) &= \sum_{i,j} (x_i \pm y_j) p(x_i, y_j) \\ &= \sum_i \sum_j x_i p(x_i, y_j) \pm \sum_j \sum_i y_j p(x_i, y_j) \\ &= \sum_i x_i p_X(x_i) \pm \sum_j y_j p_Y(y_j) = \mathbf{E}X \pm \mathbf{E}Y. \end{aligned}$$



3. Expectation of sums and differences

Corollary (a very important one)

Let X and Y be any random variables. Then

$$\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y \quad \text{and} \quad \mathbf{E}(X - Y) = \mathbf{E}X - \mathbf{E}Y.$$

Proof.

$$\begin{aligned} \mathbf{E}(X \pm Y) &= \sum_{i,j} (x_i \pm y_j) p(x_i, y_j) \\ &= \sum_i \sum_j x_i p(x_i, y_j) \pm \sum_j \sum_i y_j p(x_i, y_j) \\ &= \sum_i x_i p_X(x_i) \pm \sum_j y_j p_Y(y_j) = \mathbf{E}X \pm \mathbf{E}Y. \end{aligned}$$



3. Expectation of sums and differences

Corollary (a very important one)

Let X and Y be any random variables. Then

$$\mathbf{E}(X + Y) = \mathbf{E}X + \mathbf{E}Y \quad \text{and} \quad \mathbf{E}(X - Y) = \mathbf{E}X - \mathbf{E}Y.$$

Proof.

$$\begin{aligned} \mathbf{E}(X \pm Y) &= \sum_{i,j} (x_i \pm y_j) p(x_i, y_j) \\ &= \sum_i \sum_j x_i p(x_i, y_j) \pm \sum_j \sum_i y_j p(x_i, y_j) \\ &= \sum_i x_i p_X(x_i) \pm \sum_j y_j p_Y(y_j) = \mathbf{E}X \pm \mathbf{E}Y. \end{aligned}$$



3. Expectation of sums and differences

Corollary

Let X and Y be such that $X \leq Y$ a.s. Then $\mathbf{E}X \leq \mathbf{E}Y$.

Proof.

Just look at the difference $Y - X$. This is a.s. non-negative, hence its expectation is non-negative as well:

$$0 \leq \mathbf{E}(Y - X) = \mathbf{E}Y - \mathbf{E}X.$$



3. Expectation of sums and differences

Example (sample mean)

Let X_1, X_2, \dots, X_n be identically distributed random variables with mean μ . Their sample mean is

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i.$$

Its expectation is

$$\mathbf{E}\bar{X} = \mathbf{E}\frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} \mathbf{E} \sum_{i=1}^n X_i = \frac{1}{n} \sum_{i=1}^n \mathbf{E}X_i = \frac{1}{n} \sum_{i=1}^n \mu = \mu.$$

3. Expectation of sums and differences

Example

Let A_1, A_2, \dots, A_n be events, and X_1, X_2, \dots, X_n their respective indicator variables. Then

$$X := \sum_{i=1}^n X_i$$

counts the number of these events that occur. The *expected* number of them is

$$\mathbf{E}X = \mathbf{E} \sum_{i=1}^n X_i = \sum_{i=1}^n \mathbf{E}X_i = \sum_{i=1}^n \mathbf{P}A_i.$$

3. Expectation of sums and differences

Example (Boole's inequality)

Let A_1, A_2, \dots, A_n be events, and X_1, X_2, \dots, X_n their respective indicator variables. Define further

$$X := \sum_{i=1}^n X_i \quad \text{and} \quad Y := \begin{cases} 1, & \text{if } X \geq 1, \\ 0, & \text{if } X = 0. \end{cases}$$

Notice that $Y = \mathbf{1}\left\{\bigcup_{i=1}^n A_i\right\}$ and that $Y \leq X \rightsquigarrow$ , thus

$$\mathbf{P}\left\{\bigcup_{i=1}^n A_i\right\} = \mathbf{E}Y \leq \mathbf{E}X = \mathbf{E}\sum_{i=1}^n X_i = \sum_{i=1}^n \mathbf{E}X_i = \sum_{i=1}^n \mathbf{P}\{A_i\}$$

known as Boole's inequality.

3. Expectation of sums and differences

Example (Binomial distribution)

Suppose that n independent trials are made, each succeeding with probability p . Define X_i as the indicator of success in the i^{th} trial, $i = 1, 2, \dots, n$. Then

$$X := \sum_{i=1}^n X_i$$

counts the total number of successes, therefore $X \sim \text{Binom}(n, p)$. Its expectation is

$$\mathbf{E}X = \mathbf{E} \sum_{i=1}^n X_i = \sum_{i=1}^n \mathbf{E}X_i = \sum_{i=1}^n p = np.$$

Simpler than before, is it?

3. Expectation of sums and differences

Example (Gamma distribution)

Let n be a positive integer, $\lambda > 0$ real, and $X \sim \text{Gamma}(n, \lambda)$. Then we know

$$X \stackrel{d}{=} \sum_{i=1}^n X_i,$$

where X_1, X_2, \dots, X_n are i.i.d. $\text{Exp}(\lambda)$. Therefore

$$\mathbf{E}X = \mathbf{E} \sum_{i=1}^n X_i = \sum_{i=1}^n \mathbf{E}X_i = \sum_{i=1}^n \frac{1}{\lambda} = \frac{n}{\lambda}.$$

Here $\stackrel{d}{=}$ means “equal in distribution”.

We have seen the above formula $\mathbf{E}X = \frac{\alpha}{\lambda}$ in more generality for $X \sim \text{Gamma}(\alpha, \lambda)$ with any real $\alpha > 0$.

Covariance

In this part we investigate the relation of independence to expected values. It will give us some (not perfect) way of measuring independence.

Again, we assume that all the expectations we talk about exist.

1. Independence




We start with a simple observation:

Proposition

Let X and Y be *independent* random variables, and g, h functions. Then

$$\mathbf{E}(g(X) \cdot h(Y)) = \mathbf{E}g(X) \cdot \mathbf{E}h(Y).$$

Proof.

This is true for any random variables. For the discrete case, the proof uses the factorisation of the mass functions:   

2. Covariance

Then, the following is a natural object to measure independence:

Definition

The covariance of the random variables X and Y is

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)].$$

Before exploring its properties, notice

$$\begin{aligned}\mathbf{Cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}[X \cdot \mathbf{E}Y] - \mathbf{E}[(\mathbf{E}X) \cdot Y] + \mathbf{E}[\mathbf{E}X \cdot \mathbf{E}Y] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}X \cdot \mathbf{E}Y - \mathbf{E}X \cdot \mathbf{E}Y + \mathbf{E}X \cdot \mathbf{E}Y \\ &= \mathbf{E}XY - \mathbf{E}X \cdot \mathbf{E}Y.\end{aligned}$$

2. Covariance

Then, the following is a natural object to measure independence:

Definition

The covariance of the random variables X and Y is

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)].$$

Before exploring its properties, notice

$$\begin{aligned} \mathbf{Cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}[X \cdot \mathbf{E}Y] - \mathbf{E}[(\mathbf{E}X) \cdot Y] + \mathbf{E}[\mathbf{E}X \cdot \mathbf{E}Y] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}X \cdot \mathbf{E}Y - \mathbf{E}X \cdot \mathbf{E}Y + \mathbf{E}X \cdot \mathbf{E}Y \\ &= \mathbf{E}XY - \mathbf{E}X \cdot \mathbf{E}Y. \end{aligned}$$

2. Covariance

Then, the following is a natural object to measure independence:

Definition

The covariance of the random variables X and Y is

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)].$$

Before exploring its properties, notice

$$\begin{aligned}\mathbf{Cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}[X \cdot \mathbf{E}Y] - \mathbf{E}[(\mathbf{E}X) \cdot Y] + \mathbf{E}[\mathbf{E}X \cdot \mathbf{E}Y] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}X \cdot \mathbf{E}Y - \mathbf{E}X \cdot \mathbf{E}Y + \mathbf{E}X \cdot \mathbf{E}Y \\ &= \mathbf{E}XY - \mathbf{E}X \cdot \mathbf{E}Y.\end{aligned}$$

2. Covariance

Then, the following is a natural object to measure independence:

Definition

The covariance of the random variables X and Y is

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)].$$

Before exploring its properties, notice

$$\begin{aligned} \mathbf{Cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}[X \cdot \mathbf{E}Y] - \mathbf{E}[(\mathbf{E}X) \cdot Y] + \mathbf{E}[\mathbf{E}X \cdot \mathbf{E}Y] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}X \cdot \mathbf{E}Y - \mathbf{E}X \cdot \mathbf{E}Y + \mathbf{E}X \cdot \mathbf{E}Y \\ &= \mathbf{E}XY - \mathbf{E}X \cdot \mathbf{E}Y. \end{aligned}$$

2. Covariance

Then, the following is a natural object to measure independence:

Definition

The covariance of the random variables X and Y is

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)].$$

Before exploring its properties, **notice**

$$\begin{aligned} \mathbf{Cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}[X \cdot \mathbf{E}Y] - \mathbf{E}[(\mathbf{E}X) \cdot Y] + \mathbf{E}[\mathbf{E}X \cdot \mathbf{E}Y] \\ &= \mathbf{E}[X \cdot Y] - \mathbf{E}X \cdot \mathbf{E}Y - \mathbf{E}X \cdot \mathbf{E}Y + \mathbf{E}X \cdot \mathbf{E}Y \\ &= \mathbf{E}XY - \mathbf{E}X \cdot \mathbf{E}Y. \end{aligned}$$

2. Covariance

Remark

From either forms

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)] = \mathbf{E}XY - \mathbf{E}X \cdot \mathbf{E}Y$$

it is clear that for **independent** random variables,

$$\mathbf{Cov}(X, Y) = 0.$$

2. Covariance

Example

This is not true the other way around: let

$$X := \begin{cases} -1, & \text{with prob. } \frac{1}{3}, \\ 0, & \text{with prob. } \frac{1}{3}, \\ 1, & \text{with prob. } \frac{1}{3}, \end{cases} \quad Y := \begin{cases} 0, & \text{if } X \neq 0, \\ 1, & \text{if } X = 0. \end{cases}$$

Then $X \cdot Y = 0$ and $\mathbf{E}X = 0$, thus $\mathbf{Cov}(X, Y) = 0$, but these variables are clearly **not independent**.

2. Covariance

Proposition (properties of covariance)

Fix a_i, b, c_j, d real numbers. Covariance is

- ▶ *positive semidefinite*: $\mathbf{Cov}(X, X) = \mathbf{Var}X \geq 0$,
- ▶ *symmetric*: $\mathbf{Cov}(X, Y) = \mathbf{Cov}(Y, X)$,
- ▶ *almost bilinear*:

$$\mathbf{Cov}\left(\sum_i a_i X_i + b, \sum_j c_j Y_j + d\right) = \sum_{i,j} a_i c_j \mathbf{Cov}(X_i, Y_j).$$



3. Variance

Now we can answer a long overdue question: what happens to the variance of sums or random variables?

Proposition (variance of sums)

Let X_1, X_2, \dots, X_n be random variables. Then

$$\mathbf{Var} \sum_{i=1}^n X_i = \sum_{i=1}^n \mathbf{Var} X_i + 2 \sum_{1 \leq i < j \leq n} \mathbf{Cov}(X_i, X_j).$$

In particular, variances of *independent* random variables are additive.

No additivity, however, of variances in general.

3. Variance

Proof.

Just using properties of the covariance,

$$\begin{aligned}\mathbf{Var} \sum_{i=1}^n X_i &= \mathbf{Cov} \left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j \right) = \sum_{i,j=1}^n \mathbf{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \mathbf{Cov}(X_i, X_i) + \sum_{i \neq j} \mathbf{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \mathbf{Var} X_i + \sum_{i < j} \mathbf{Cov}(X_i, X_j) + \sum_{i > j} \mathbf{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \mathbf{Var} X_i + 2 \sum_{1 \leq i < j \leq n} \mathbf{Cov}(X_i, X_j).\end{aligned}$$



3. Variance

Remark

Notice that for **independent** variables,

$$\begin{aligned}\mathbf{Var}(X - Y) &= \mathbf{Var}(X + (-Y)) \\ &= \mathbf{Var}X + \mathbf{Var}(-Y) + 2\mathbf{Cov}(X, -Y) \\ &= \mathbf{Var}X + \mathbf{Var}Y - 2\mathbf{Cov}(X, Y) \\ &= \mathbf{Var}X + \mathbf{Var}Y.\end{aligned}$$

3. Variance

Example (variance of the sample mean)

Suppose that X_i 's are **i.i.d.**, each of variance σ^2 . Recall the definition

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i$$

of the sample mean. Its variance is

$$\begin{aligned} \text{Var} \bar{X} &= \text{Var} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \text{Var} \left(\sum_{i=1}^n X_i \right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var} X_i = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Decreases with n , that's why we like sample averages.

3. Variance

Example (Binomial distribution)

Suppose that n independent trials are made, each succeeding with probability p . Define X_i as the indicator of success in the i^{th} trial, $i = 1, 2, \dots, n$. Then

$$X := \sum_{i=1}^n X_i$$

counts the total number of successes, therefore $X \sim \text{Binom}(n, p)$. Its variance is

$$\text{Var}X = \text{Var} \sum_{i=1}^n X_i = \sum_{i=1}^n \text{Var}X_i = \sum_{i=1}^n p(1-p) = np(1-p).$$

Simpler than before, is it?

3. Variance

Example (Gamma distribution)

Let n be a positive integer, $\lambda > 0$ real, and $X \sim \text{Gamma}(n, \lambda)$. Then we know

$$X \stackrel{d}{=} \sum_{i=1}^n X_i,$$

where X_1, X_2, \dots, X_n are **i.i.d.** $\text{Exp}(\lambda)$. Therefore

$$\text{Var}X = \text{Var} \sum_{i=1}^n X_i = \sum_{i=1}^n \text{Var}X_i = \sum_{i=1}^n \frac{1}{\lambda^2} = \frac{n}{\lambda^2}.$$

Here $\stackrel{d}{=}$ means “equal in distribution”.

We have seen the above formula $\text{Var}X = \frac{\alpha}{\lambda^2}$ in more generality for $X \sim \text{Gamma}(\alpha, \lambda)$ with any real $\alpha > 0$.

4. Cauchy-Schwarz inequality

Theorem (Cauchy-Schwarz inequality)

For every X and Y ,

$$|\mathbf{E}XY| \leq \sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2},$$

with equality iff $Y = \text{const.} \cdot X$ a.s.

This is essentially the same statement as Cauchy-Schwarz in linear algebra.

4. Cauchy-Schwarz inequality

Proof.

$$\begin{aligned} 0 &\leq \mathbf{E}\left(\frac{X}{\sqrt{\mathbf{E}X^2}} \pm \frac{Y}{\sqrt{\mathbf{E}Y^2}}\right)^2 \\ &= \mathbf{E}\left(\frac{X^2}{\mathbf{E}X^2}\right) + \mathbf{E}\left(\frac{Y^2}{\mathbf{E}Y^2}\right) \pm 2\mathbf{E}\frac{XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}} \\ &= 2 \pm 2\frac{\mathbf{E}XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}}. \end{aligned}$$



4. Cauchy-Schwarz inequality

Proof.

$$\begin{aligned}
 0 &\leq \mathbf{E} \left(\frac{X}{\sqrt{\mathbf{E}X^2}} - \frac{Y}{\sqrt{\mathbf{E}Y^2}} \right)^2 \\
 &= \mathbf{E} \left(\frac{X^2}{\mathbf{E}X^2} \right) + \mathbf{E} \left(\frac{Y^2}{\mathbf{E}Y^2} \right) - 2\mathbf{E} \frac{XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}} \\
 &= 2 - 2 \frac{\mathbf{E}XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}}.
 \end{aligned}$$

When $-$,

$$\begin{aligned}
 2 \frac{\mathbf{E}XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}} &\leq 2, \\
 \mathbf{E}XY &\leq \sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}.
 \end{aligned}$$



4. Cauchy-Schwarz inequality

Proof.

$$\begin{aligned}
 0 &\leq \mathbf{E} \left(\frac{X}{\sqrt{\mathbf{E}X^2}} + \frac{Y}{\sqrt{\mathbf{E}Y^2}} \right)^2 \\
 &= \mathbf{E} \left(\frac{X^2}{\mathbf{E}X^2} \right) + \mathbf{E} \left(\frac{Y^2}{\mathbf{E}Y^2} \right) + 2\mathbf{E} \frac{XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}} \\
 &= 2 + 2 \frac{\mathbf{E}XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}}.
 \end{aligned}$$

When +,

$$\begin{aligned}
 -2 \frac{\mathbf{E}XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}} &\leq 2, \\
 \mathbf{E}XY &\geq -\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}.
 \end{aligned}$$





4. Cauchy-Schwarz inequality

Proof.

$$\begin{aligned}
 0 &\leq \mathbf{E} \left(\frac{X}{\sqrt{\mathbf{E}X^2}} + \frac{Y}{\sqrt{\mathbf{E}Y^2}} \right)^2 \\
 &= \mathbf{E} \left(\frac{X^2}{\mathbf{E}X^2} \right) + \mathbf{E} \left(\frac{Y^2}{\mathbf{E}Y^2} \right) + 2\mathbf{E} \frac{XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}} \\
 &= 2 + 2 \frac{\mathbf{E}XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}}.
 \end{aligned}$$

When +,

$$\begin{aligned}
 -2 \frac{\mathbf{E}XY}{\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}} &\leq 2, \\
 \mathbf{E}XY &\geq -\sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}.
 \end{aligned}$$

For the “equality iff” part,   .



4. Cauchy-Schwarz inequality

Corollary

Apply Cauchy-Schwarz on $|X|$ and $|Y|$ to get

$$\mathbf{E}|XY| = \mathbf{E}(|X| \cdot |Y|) \leq \sqrt{\mathbf{E}|X|^2} \cdot \sqrt{\mathbf{E}|Y|^2} = \sqrt{\mathbf{E}X^2} \cdot \sqrt{\mathbf{E}Y^2}.$$

Corollary

Apply Cauchy-Schwarz on $\tilde{X} := X - \mathbf{E}X$ and $\tilde{Y} := Y - \mathbf{E}Y$:

$$\begin{aligned} |\mathbf{Cov}(X, Y)| &= |\mathbf{E}[(X - \mathbf{E}X) \cdot (Y - \mathbf{E}Y)]| = |\mathbf{E}[\tilde{X} \cdot \tilde{Y}]| \\ &\leq \sqrt{\mathbf{E}\tilde{X}^2} \cdot \sqrt{\mathbf{E}\tilde{Y}^2} \\ &= \sqrt{\mathbf{E}(X - \mathbf{E}X)^2} \cdot \sqrt{\mathbf{E}(Y - \mathbf{E}Y)^2} = \text{SD } X \cdot \text{SD } Y. \end{aligned}$$

5. Correlation

Definition

The correlation coefficient of random variables X and Y is

$$\rho(X, Y) := \frac{\mathbf{Cov}(X, Y)}{\mathbf{SD} X \cdot \mathbf{SD} Y}.$$


Remark

The previous corollary precisely states that

$$-1 \leq \rho(X, Y) \leq 1,$$

and the “equality iff” part of Cauchy-Schwarz implies that we have equality iff $\tilde{Y} = a\tilde{X}$, that is, $Y = aX + b$ for some fixed a, b .

5. Correlation

Positive correlation means that typical values of Y tend to be larger when those of X are larger. Negative correlation means that typical values of Y tend to be smaller when those of X are larger. \rightsquigarrow 

5. Correlation

Example

Rolling two dice, let X be the number shown on the first die, Y the one shown on the second die, $Z = X + Y$ the sum of the two numbers. Clearly, X and Y are independent, $\mathbf{Cov}(X, Y) = 0$, $\rho(X, Y) = 0$. For X and Z ,

$$\begin{aligned}\mathbf{Cov}(X, Z) &= \mathbf{Cov}(X, X + Y) = \mathbf{Cov}(X, X) + \mathbf{Cov}(X, Y) \\ &= \mathbf{Var}X + 0 = \mathbf{Var}X;\end{aligned}$$

$$\begin{aligned}\mathbf{Var}(Z) &= \mathbf{Var}(X + Y) = \mathbf{Var}X + \mathbf{Var}Y \quad \text{indep.}! \\ &= \mathbf{Var}X + \mathbf{Var}X = 2\mathbf{Var}X;\end{aligned}$$

$$\rho(X, Z) = \frac{\mathbf{Cov}(X, Z)}{\mathbf{SD} X \cdot \mathbf{SD} Z} = \frac{\mathbf{Var}X}{\sqrt{\mathbf{Var}X} \cdot \sqrt{2\mathbf{Var}X}} = \frac{1}{\sqrt{2}}.$$

Conditional expectation

Recall that the *conditional mass function* $p_{X|Y}(x|y_j)$ is a proper mass function. This allows in particular to build an expectation out of it, a very useful tool.

As usual, we assume that all expectations exist. Everything here is completely general, but most proofs are only shown for the discrete case.

1. Conditional expectation

We can therefore define

Definition

The conditional expectation of X , given $Y = y_j$ is

$$\mathbf{E}(X | Y = y_j) := \sum_i x_i \cdot p_{X|Y}(x_i | y_j).$$

Example

Let X and Y be independent $\text{Poi}(\lambda)$ and $\text{Poi}(\mu)$ variables, and $Z = X + Y$. Find the conditional expectation $\mathbf{E}(X | Z = k)$.

1. Conditional expectation

Solution

Start with the conditional mass function ($0 \leq i \leq k$):

$$p_{X|Z}(i|k) = \frac{p(i, k)}{p_Z(k)},$$

where $p(i, k)$ is the joint mass function of X and Z at (i, k) . To find this latter one, write

$$\begin{aligned} p(i, k) &= \mathbf{P}\{X = i, Z = k\} = \mathbf{P}\{X = i, X + Y = k\} \\ &= \mathbf{P}\{X = i, Y = k - i\} = e^{-\lambda} \frac{\lambda^i}{i!} \cdot e^{-\mu} \frac{\mu^{k-i}}{(k-i)!}. \end{aligned}$$

Recall also that $Z = X + Y \sim \text{Poi}(\lambda + \mu)$, $p_Z(k) = e^{-\lambda - \mu} \frac{(\lambda + \mu)^k}{k!}$.

1. Conditional expectation

Solution

Combining,

$$\begin{aligned} p_{X|Z}(i|k) &= \frac{p(i, k)}{p_Z(k)} = \frac{e^{-\lambda} \frac{\lambda^i}{i!} \cdot e^{-\mu} \frac{\mu^{k-i}}{(k-i)!}}{e^{-\lambda-\mu} \frac{(\lambda+\mu)^k}{k!}} \\ &= \binom{k}{i} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^i \left(\frac{\mu}{\lambda+\mu}\right)^{k-i} = \binom{k}{i} \cdot p^i (1-p)^{k-i} \end{aligned}$$

with $p := \frac{\lambda}{\lambda+\mu}$. We conclude $(X|Z=k) \sim \text{Binom}(k, p)$,
therefore

$$\mathbf{E}(X|Z=k) = kp = k \cdot \frac{\lambda}{\lambda+\mu}.$$

1. Conditional expectation

Solution

Combining,

$$\begin{aligned} p_{X|Z}(i|k) &= \frac{p(i, k)}{p_Z(k)} = \frac{e^{-\lambda} \frac{\lambda^i}{i!} \cdot e^{-\mu} \frac{\mu^{k-i}}{(k-i)!}}{e^{-\lambda-\mu} \frac{(\lambda+\mu)^k}{k!}} \\ &= \binom{k}{i} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^i \left(\frac{\mu}{\lambda+\mu}\right)^{k-i} = \binom{k}{i} \cdot p^i (1-p)^{k-i} \end{aligned}$$

with $p := \frac{\lambda}{\lambda+\mu}$. We conclude $(X|Z=k) \sim \text{Binom}(k, p)$,
therefore

$$\mathbf{E}(X|Z=k) = kp = k \cdot \frac{\lambda}{\lambda+\mu}.$$

1. Conditional expectation

Solution

Combining,

$$\begin{aligned} p_{X|Z}(i|k) &= \frac{p(i, k)}{p_Z(k)} = \frac{e^{-\lambda} \frac{\lambda^i}{i!} \cdot e^{-\mu} \frac{\mu^{k-i}}{(k-i)!}}{e^{-\lambda-\mu} \frac{(\lambda+\mu)^k}{k!}} \\ &= \binom{k}{i} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^i \left(\frac{\mu}{\lambda+\mu}\right)^{k-i} = \binom{k}{i} \cdot p^i (1-p)^{k-i} \end{aligned}$$

with $p := \frac{\lambda}{\lambda+\mu}$. We conclude $(X|Z=k) \sim \text{Binom}(k, p)$,
therefore

$$\mathbf{E}(X|Z=k) = kp = k \cdot \frac{\lambda}{\lambda+\mu}.$$

1. Conditional expectation

Solution

Combining,

$$\begin{aligned} p_{X|Z}(i|k) &= \frac{p(i, k)}{p_Z(k)} = \frac{e^{-\lambda} \frac{\lambda^i}{i!} \cdot e^{-\mu} \frac{\mu^{k-i}}{(k-i)!}}{e^{-\lambda-\mu} \frac{(\lambda+\mu)^k}{k!}} \\ &= \binom{k}{i} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^i \left(\frac{\mu}{\lambda+\mu}\right)^{k-i} = \binom{k}{i} \cdot p^i (1-p)^{k-i} \end{aligned}$$

with $p := \frac{\lambda}{\lambda+\mu}$. We conclude $(X|Z=k) \sim \text{Binom}(k, p)$,
therefore

$$\mathbf{E}(X|Z=k) = kp = k \cdot \frac{\lambda}{\lambda+\mu}.$$

1. Conditional expectation

Solution

Combining,

$$\begin{aligned} p_{X|Z}(i|k) &= \frac{p(i, k)}{p_Z(k)} = \frac{e^{-\lambda} \frac{\lambda^i}{i!} \cdot e^{-\mu} \frac{\mu^{k-i}}{(k-i)!}}{e^{-\lambda-\mu} \frac{(\lambda+\mu)^k}{k!}} \\ &= \binom{k}{i} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^i \left(\frac{\mu}{\lambda+\mu}\right)^{k-i} = \binom{k}{i} \cdot p^i (1-p)^{k-i} \end{aligned}$$

with $p := \frac{\lambda}{\lambda+\mu}$. We conclude $(X|Z=k) \sim \text{Binom}(k, p)$,
therefore

$$\mathbf{E}(X|Z=k) = kp = k \cdot \frac{\lambda}{\lambda+\mu}.$$

1. Conditional expectation

Solution

Combining,

$$\begin{aligned} p_{X|Z}(i|k) &= \frac{p(i, k)}{p_Z(k)} = \frac{e^{-\lambda} \frac{\lambda^i}{i!} \cdot e^{-\mu} \frac{\mu^{k-i}}{(k-i)!}}{e^{-\lambda-\mu} \frac{(\lambda+\mu)^k}{k!}} \\ &= \binom{k}{i} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^i \left(\frac{\mu}{\lambda+\mu}\right)^{k-i} = \binom{k}{i} \cdot p^i (1-p)^{k-i} \end{aligned}$$

with $p := \frac{\lambda}{\lambda+\mu}$. We conclude $(X|Z=k) \sim \text{Binom}(k, p)$,
therefore

$$\mathbf{E}(X|Z=k) = kp = k \cdot \frac{\lambda}{\lambda+\mu}.$$

1. Conditional expectation

Solution

Combining,

$$\begin{aligned} p_{X|Z}(i|k) &= \frac{p(i, k)}{p_Z(k)} = \frac{e^{-\lambda} \frac{\lambda^i}{i!} \cdot e^{-\mu} \frac{\mu^{k-i}}{(k-i)!}}{e^{-\lambda-\mu} \frac{(\lambda+\mu)^k}{k!}} \\ &= \binom{k}{i} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^i \left(\frac{\mu}{\lambda+\mu}\right)^{k-i} = \binom{k}{i} \cdot p^i (1-p)^{k-i} \end{aligned}$$

with $p := \frac{\lambda}{\lambda+\mu}$. We conclude $(X|Z=k) \sim \text{Binom}(k, p)$,
therefore

$$\mathbf{E}(X|Z=k) = kp = k \cdot \frac{\lambda}{\lambda+\mu}.$$

1. Conditional expectation

Remark

We abbreviate our assertion

$$\mathbf{E}(X | Z = k) = k \cdot \frac{\lambda}{\lambda + \mu} \quad \text{as} \quad \mathbf{E}(X | Z) = Z \cdot \frac{\lambda}{\lambda + \mu}.$$

Notice, however, that something deeper actually happened.

- ▶ $\mathbf{E}(X | Z = k)$ is the expected value of X if I know that $Z = k$. It is a function **of k** .
- ▶ $\mathbf{E}(X | Z)$ is the expected value of X if I know the value of Z , *but I won't tell you*. It is the same function **of Z** .

As such, $\mathbf{E}(X | Z)$ **itself is a random variable** (namely, a function of Z).

2. Tower rule

It therefore makes sense to talk about the expectation of the random variable $\mathbf{E}(X | Y)$.

Proposition (Tower rule; aka. Law of total expectation)

For any random variables, $\mathbf{E}\mathbf{E}(X | Y) = \mathbf{E}X$.

Proof.

$$\begin{aligned}\mathbf{E}\mathbf{E}(X | Y) &= \sum_j \mathbf{E}(X | Y = y_j) \cdot p_Y(y_j) \\ &= \sum_j \sum_i x_i p_{X|Y}(x_i | y_j) \cdot p_Y(y_j) \\ &= \sum_i \sum_j x_i p(x_i, y_j) = \sum_i x_i p_X(x_i) = \mathbf{E}X.\end{aligned}$$



2. Tower rule

It therefore makes sense to talk about the expectation of the random variable $\mathbf{E}(X | Y)$.

Proposition (Tower rule; aka. Law of total expectation)

For any random variables, $\mathbf{E}\mathbf{E}(X | Y) = \mathbf{E}X$.

Proof.

$$\begin{aligned}\mathbf{E}\mathbf{E}(X | Y) &= \sum_j \mathbf{E}(X | Y = y_j) \cdot p_Y(y_j) \\ &= \sum_j \sum_i x_i p_{X|Y}(x_i | y_j) \cdot p_Y(y_j) \\ &= \sum_i \sum_j x_i p(x_i, y_j) = \sum_i x_i p_X(x_i) = \mathbf{E}X.\end{aligned}$$



2. Tower rule

It therefore makes sense to talk about the expectation of the random variable $\mathbf{E}(X | Y)$.

Proposition (Tower rule; aka. Law of total expectation)

For any random variables, $\mathbf{E}\mathbf{E}(X | Y) = \mathbf{E}X$.

Proof.

$$\begin{aligned}\mathbf{E}\mathbf{E}(X | Y) &= \sum_j \mathbf{E}(X | Y = y_j) \cdot p_Y(y_j) \\ &= \sum_j \sum_i x_i p_{X|Y}(x_i | y_j) \cdot p_Y(y_j) \\ &= \sum_i \sum_j x_i p(x_i, y_j) = \sum_i x_i p_X(x_i) = \mathbf{E}X.\end{aligned}$$



2. Tower rule

It therefore makes sense to talk about the expectation of the random variable $\mathbf{E}(X | Y)$.

Proposition (Tower rule; aka. Law of total expectation)

For any random variables, $\mathbf{E}\mathbf{E}(X | Y) = \mathbf{E}X$.

Proof.

$$\begin{aligned}\mathbf{E}\mathbf{E}(X | Y) &= \sum_j \mathbf{E}(X | Y = y_j) \cdot p_Y(y_j) \\ &= \sum_j \sum_i x_i p_{X|Y}(x_i | y_j) \cdot p_Y(y_j) \\ &= \sum_i \sum_j x_i p(x_i, y_j) = \sum_i x_i p_X(x_i) = \mathbf{E}X.\end{aligned}$$



2. Tower rule

It therefore makes sense to talk about the expectation of the random variable $\mathbf{E}(X | Y)$.

Proposition (Tower rule; aka. Law of total expectation)

For any random variables, $\mathbf{E}\mathbf{E}(X | Y) = \mathbf{E}X$.

Proof.

$$\begin{aligned}
 \mathbf{E}\mathbf{E}(X | Y) &= \sum_j \mathbf{E}(X | Y = y_j) \cdot p_Y(y_j) \\
 &= \sum_j \sum_i x_i p_{X|Y}(x_i | y_j) \cdot p_Y(y_j) \\
 &= \sum_i \sum_j x_i p(x_i, y_j) = \sum_i x_i p_X(x_i) = \mathbf{E}X.
 \end{aligned}$$



2. Tower rule

It therefore makes sense to talk about the expectation of the random variable $\mathbf{E}(X | Y)$.

Proposition (Tower rule; aka. Law of total expectation)

For any random variables, $\mathbf{E}\mathbf{E}(X | Y) = \mathbf{E}X$.

Proof.

$$\begin{aligned}
 \mathbf{E}\mathbf{E}(X | Y) &= \sum_j \mathbf{E}(X | Y = y_j) \cdot p_Y(y_j) \\
 &= \sum_j \sum_i x_i p_{X|Y}(x_i | y_j) \cdot p_Y(y_j) \\
 &= \sum_i \sum_j x_i p(x_i, y_j) = \sum_i x_i p_X(x_i) = \mathbf{E}X.
 \end{aligned}$$



2. Tower rule

Example

A disoriented miner finds himself in a room of the mine with three doors:

- ▶ The first door brings him to safety after a 3 hours long hike.
- ▶ The second door takes him back to the same room after 5 hours of climbing.
- ▶ The third door takes him again back to the same room after 7 hours of exhausting climbing.

The disoriented miner chooses one of the three doors with equal chance independently each time he is in that room. What is the expected time after which the miner is safe?

2. Tower rule

Solution

Let X be the time to reach safety, and Y the initial choice of a door ($= 1, 2, 3$). Then

$$\begin{aligned} \mathbf{E}X &= \mathbf{E}\mathbf{E}(X | Y) \\ &= \mathbf{E}(X | Y = 1) \cdot \mathbf{P}\{Y = 1\} + \mathbf{E}(X | Y = 2) \cdot \mathbf{P}\{Y = 2\} \\ &\quad + \mathbf{E}(X | Y = 3) \cdot \mathbf{P}\{Y = 3\} \\ &= 3 \cdot \frac{1}{3} + (\mathbf{E}X + 5) \cdot \frac{1}{3} + (\mathbf{E}X + 7) \cdot \frac{1}{3}, \end{aligned}$$

which we rearrange as

$$3\mathbf{E}X = 15 + 2\mathbf{E}X; \quad \mathbf{E}X = 15.$$

3. Conditional variance

A Tower rule-like formula also exists for variances. We start with defining

Definition

The conditional variance of X , given Y is

$$\mathbf{Var}(X | Y) = \mathbf{E}[(X - \mathbf{E}(X | Y))^2 | Y] = \mathbf{E}(X^2 | Y) - [\mathbf{E}(X | Y)]^2.$$

No surprise here, just use conditionals everywhere in the definition of variance.

Notice that $\mathbf{Var}(X | Y)$ is again a function of Y .

3. Conditional variance

Proposition

The conditional variance formula holds:

$$\mathbf{Var}X = \mathbf{EVar}(X | Y) + \mathbf{VarE}(X | Y).$$

In words: the variance is the expectation of the conditional variance plus the variance of the conditional expectation.

3. Conditional variance

Proof.

$$\begin{aligned}\mathbf{Var}X &= \mathbf{E}X^2 - [\mathbf{E}X]^2 = \mathbf{E}\mathbf{E}(X^2 | Y) - [\mathbf{E}\mathbf{E}(X | Y)]^2 = \\ &= \mathbf{E}(\mathbf{E}(X^2 | Y) - [\mathbf{E}(X | Y)]^2) \\ &\quad + \mathbf{E}[\mathbf{E}(X | Y)]^2 - [\mathbf{E}\mathbf{E}(X | Y)]^2 \\ &= \mathbf{E}\mathbf{Var}(X | Y) + \mathbf{Var}\mathbf{E}(X | Y).\end{aligned}$$



3. Conditional variance

Proof.

$$\begin{aligned}\text{Var}X &= \mathbf{E}X^2 - [\mathbf{E}X]^2 = \mathbf{E}\mathbf{E}(X^2 | Y) - [\mathbf{E}\mathbf{E}(X | Y)]^2 = \\ &= \mathbf{E}(\mathbf{E}(X^2 | Y) - [\mathbf{E}(X | Y)]^2) \\ &\quad + \mathbf{E}[\mathbf{E}(X | Y)]^2 - [\mathbf{E}\mathbf{E}(X | Y)]^2 \\ &= \mathbf{E}\text{Var}(X | Y) + \text{Var}\mathbf{E}(X | Y).\end{aligned}$$



3. Conditional variance

Proof.

$$\begin{aligned}\text{Var}X &= \mathbf{E}X^2 - [\mathbf{E}X]^2 = \mathbf{E}\mathbf{E}(X^2 | Y) - [\mathbf{E}\mathbf{E}(X | Y)]^2 = \\ &= \mathbf{E}(\mathbf{E}(X^2 | Y) - [\mathbf{E}(X | Y)]^2) \\ &\quad + \mathbf{E}[\mathbf{E}(X | Y)]^2 - [\mathbf{E}\mathbf{E}(X | Y)]^2 \\ &= \mathbf{E}\text{Var}(X | Y) + \text{Var}\mathbf{E}(X | Y).\end{aligned}$$



3. Conditional variance

Proof.

$$\begin{aligned}\text{Var}X &= \mathbf{E}X^2 - [\mathbf{E}X]^2 = \mathbf{E}\mathbf{E}(X^2 | Y) - [\mathbf{E}\mathbf{E}(X | Y)]^2 = \\ &= \mathbf{E}(\mathbf{E}(X^2 | Y) - [\mathbf{E}(X | Y)]^2) \\ &\quad + \mathbf{E}[\mathbf{E}(X | Y)]^2 - [\mathbf{E}\mathbf{E}(X | Y)]^2 \\ &= \mathbf{E}\text{Var}(X | Y) + \text{Var}\mathbf{E}(X | Y).\end{aligned}$$



3. Conditional variance

Proof.

$$\begin{aligned}\text{Var}X &= \mathbf{E}X^2 - [\mathbf{E}X]^2 = \mathbf{E}\mathbf{E}(X^2 | Y) - [\mathbf{E}\mathbf{E}(X | Y)]^2 = \\ &= \mathbf{E}(\mathbf{E}(X^2 | Y) - [\mathbf{E}(X | Y)]^2) \\ &\quad + \mathbf{E}[\mathbf{E}(X | Y)]^2 - [\mathbf{E}\mathbf{E}(X | Y)]^2 \\ &= \mathbf{E}\text{Var}(X | Y) + \text{Var}\mathbf{E}(X | Y).\end{aligned}$$



3. Conditional variance

Proof.

$$\begin{aligned}\text{Var}X &= \mathbf{E}X^2 - [\mathbf{E}X]^2 = \mathbf{E}\mathbf{E}(X^2 | Y) - [\mathbf{E}\mathbf{E}(X | Y)]^2 = \\ &= \mathbf{E}(\mathbf{E}(X^2 | Y) - [\mathbf{E}(X | Y)]^2) \\ &\quad + \mathbf{E}[\mathbf{E}(X | Y)]^2 - [\mathbf{E}\mathbf{E}(X | Y)]^2 \\ &= \mathbf{E}\text{Var}(X | Y) + \text{Var}\mathbf{E}(X | Y).\end{aligned}$$



3. Conditional variance

Example

An initially empty train departs from the station at a $T \sim U(0, t)$ time. While in the station, it collects passengers, and given the time T of departure it has Poisson many passengers with mean λT . What is the overall mean and variance of the number of passengers who take this train from the station?

3. Conditional variance

Solution

Let us start with translating the problem. Call N the number of passengers on the train. Then $T \sim U(0, t)$ and $(N | T) \sim \text{Poi}(\lambda T)$. Thus,

$$\begin{aligned} \mathbf{E}N &= \mathbf{E}\mathbf{E}(N | T) = \mathbf{E}(\lambda T) = \lambda \mathbf{E}T = \lambda \frac{t}{2}, \\ \mathbf{Var}N &= \mathbf{E}\mathbf{Var}(N | T) + \mathbf{Var}\mathbf{E}(N | T) = \mathbf{E}(\lambda T) + \mathbf{Var}(\lambda T) \\ &= \lambda \mathbf{E}T + \lambda^2 \mathbf{Var}T = \lambda \frac{t}{2} + \lambda^2 \frac{t^2}{12}. \end{aligned}$$

Notice how N fluctuates for two reasons \rightsquigarrow  .

4. Random sums

Generalise slightly from before:

- ▶ $\mathbf{E}(Z \cdot X | Z = k)$ is the expected value of $Z \cdot X$ if I know that $Z = k$. It is a function of k .
- ▶ $\mathbf{E}(Z \cdot X | Z)$ is the expected value of $Z \cdot X$ if I know the value of Z , *but I won't tell you*. It is the same function of Z .

It is clear that

$\mathbf{E}(Z \cdot X | Z = k) = \mathbf{E}(k \cdot X | Z = k) = k \cdot \mathbf{E}(X | Z = k)$, and by the analogy we conclude

$$\mathbf{E}(Z \cdot X | Z) = Z \cdot \mathbf{E}(X | Z).$$

The message of this formula is the fact that **given Z , Z is not random**, it goes in and out the conditional expectation. This is the second very important property of conditional expectations after the tower rule.

4. Random sums

Example (random sums)

A local store sees a random N number of customers a day, the i^{th} of whom independently of everything spends an amount X_i with mean μ and variance σ^2 . Let's find the mean and variance of the total amount spent in the store in a day.

Solution

Notice that we are after the mean and variance of

$$\sum_{i=1}^N X_i,$$

and the summation boundary is random too.

4. Random sums

Solution (. . . cont'd)

It is completely **WRONG** to write

$$\mathbf{E} \sum_{i=1}^N X_i = \sum_{i=1}^N \mathbf{E} X_i$$

as we have no right to commute the **E** over the random N .

4. Random sums

Solution (. . . cont'd)

However, conditioning on N makes N non-random for the conditional expectation, and that's the way to proceed:

$$\begin{aligned}\mathbf{E} \sum_{i=1}^N X_i &= \mathbf{E} \mathbf{E} \left(\sum_{i=1}^N X_i \mid N \right) = \mathbf{E} \sum_{i=1}^N \mathbf{E}(X_i \mid N) \\ &= \mathbf{E} \sum_{i=1}^N \mathbf{E} X_i = \mathbf{E} \sum_{i=1}^N \mu = \mathbf{E}(N\mu) = \mu \cdot \mathbf{E} N.\end{aligned}$$

No surprise here. But how about the variance?

4. Random sums

Solution (. . . cont'd)

$$\begin{aligned}
 \mathbf{Var} \sum_{i=1}^N X_i &= \mathbf{EVar} \left(\sum_{i=1}^N X_i \mid N \right) + \mathbf{VarE} \left(\sum_{i=1}^N X_i \mid N \right) \\
 &= \mathbf{E} \sum_{i=1}^N \mathbf{Var}(X_i \mid N) + \mathbf{Var} \sum_{i=1}^N \mathbf{E}(X_i \mid N) \\
 &= \mathbf{E} \sum_{i=1}^N \mathbf{Var} X_i + \mathbf{Var} \sum_{i=1}^N \mathbf{E} X_i = \mathbf{E} \sum_{i=1}^N \sigma^2 + \mathbf{Var} \sum_{i=1}^N \mu \\
 &= \mathbf{E}(N\sigma^2) + \mathbf{Var}(N\mu) = \sigma^2 \cdot \mathbf{E}N + \mu^2 \cdot \mathbf{Var}N.
 \end{aligned}$$

Notice again the two sources of fluctuations \rightsquigarrow



Moment generating functions

Moment generating functions will be an exciting new tool to make probabilistic computations very efficient. Moreover, they will serve as the fundamental tool to prove the Central Limit Theorem.

People working in probability often use the *characteristic function* instead, the complex big brother of the moment generating function.

1. Definition

Definition

The moment generating function of the random variable X is

$$M(t) := \mathbf{E}e^{tX}, \quad (\forall t \in \mathbb{R}).$$

This always exists due to $e^{tX} > 0$, but is not always finite. For most cases, it will be finite at least for t 's sufficiently close to 0. We'll assume this in this chapter.

Proposition (how it generates moments)

$$M(0) = 1, \quad \text{and} \quad M^{[n]}(0) = \mathbf{E}X^n,$$

where $M^{[n]}(0)$ denotes the n^{th} derivative at zero. \rightsquigarrow



2. Examples

Example (Binomial distribution)

Let $X \sim \text{Binom}(n, p)$.

$$M(t) = \mathbf{E}e^{tX} = \sum_{k=0}^n \binom{n}{k} e^{tk} p^k (1-p)^{n-k} = (e^t p + 1 - p)^n,$$

$$M(0) = (p + 1 - p)^n = 1,$$

$$\mathbf{E}X = M'(0) = npe^t (e^t p + 1 - p)^{n-1} \Big|_{t=0} = np,$$

$$\begin{aligned} \mathbf{E}X^2 &= M''(0) = npe^t (e^t p + 1 - p)^{n-1} \\ &\quad + n(n-1)p^2 e^{2t} (e^t p + 1 - p)^{n-2} \Big|_{t=0} \\ &= np + n(n-1)p^2, \end{aligned}$$

and that gives, as before, $\mathbf{Var}X = np(1-p)$.

2. Examples

Example (Poisson distribution)

Let $X \sim \text{Poi}(\lambda)$.

$$\begin{aligned}M(t) &= \mathbf{E}e^{tX} = \sum_{i=0}^{\infty} e^{ti} \frac{\lambda^i}{i!} \cdot e^{-\lambda} = \sum_{i=0}^{\infty} \frac{(\lambda e^t)^i}{i!} \cdot e^{-\lambda} \\ &= e^{\lambda e^t} \cdot e^{-\lambda} = e^{\lambda(e^t-1)},\end{aligned}$$

$$M(0) = e^{\lambda(e^0-1)} = 1,$$

$$\mathbf{E}X = M'(0) = \lambda e^t \cdot e^{\lambda(e^t-1)} \Big|_{t=0} = \lambda,$$

$$\mathbf{E}X^2 = M''(0) = \lambda e^t \cdot e^{\lambda(e^t-1)} + \lambda^2 e^{2t} \cdot e^{\lambda(e^t-1)} \Big|_{t=0} = \lambda + \lambda^2,$$

and $\mathbf{Var}X = \lambda$ as before.

2. Examples

Example (Exponential distribution)

Let $X \sim \text{Exp}(\lambda)$. The moment generating function is only finite for $t < \lambda$:

$$M(t) = \mathbf{E}e^{tX} = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t},$$

$$M(0) = \frac{\lambda}{\lambda - 0} = 1,$$

$$\mathbf{E}X = M'(0) = \frac{\lambda}{(\lambda - t)^2} \Big|_{t=0} = \frac{1}{\lambda},$$

$$\mathbf{E}X^2 = M''(0) = \frac{2\lambda}{(\lambda - t)^3} \Big|_{t=0} = \frac{2}{\lambda^2},$$

and $\mathbf{Var}X = \frac{1}{\lambda^2}$.

2. Examples

Example (Normal distribution)

First take $Z \sim \mathcal{N}(0, 1)$.

$$\begin{aligned}M_{\mathcal{N}(0,1)}(t) &= \mathbf{E}e^{tZ} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz} e^{-z^2/2} dz \\ &= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-t)^2/2} dz = e^{t^2/2}.\end{aligned}$$



2. Examples

Example (Normal distribution)

Next, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$,

$$\begin{aligned} M_{\mathcal{N}(\mu, \sigma^2)}(t) &= \mathbf{E}e^{tX} = \mathbf{E}\left(e^{\sigma t \cdot \frac{X-\mu}{\sigma}}\right) \cdot e^{\mu t} \\ &= M_{\mathcal{N}(0, 1)}(\sigma t) \cdot e^{\mu t} = e^{\sigma^2 t^2/2 + \mu t}, \end{aligned}$$

$$M_{\mathcal{N}(\mu, \sigma^2)}(0) = e^{\sigma^2 0^2/2 + \mu 0} = 1,$$

$$\mathbf{E}X = M'_{\mathcal{N}(\mu, \sigma^2)}(0) = (\sigma^2 t + \mu)e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} = \mu,$$

$$\begin{aligned} \mathbf{E}X^2 &= M''_{\mathcal{N}(\mu, \sigma^2)}(0) = \sigma^2 e^{\sigma^2 t^2/2 + \mu t} + (\sigma^2 t + \mu)^2 e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} \\ &= \sigma^2 + \mu^2, \end{aligned}$$

from which $\mathbf{Var}X = \sigma^2$.

2. Examples

Example (Normal distribution)

Next, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$,

$$\begin{aligned}M_{\mathcal{N}(\mu, \sigma^2)}(t) &= \mathbf{E}e^{tX} = \mathbf{E}\left(e^{\sigma t \cdot \frac{X-\mu}{\sigma}}\right) \cdot e^{\mu t} \\ &= M_{\mathcal{N}(0, 1)}(\sigma t) \cdot e^{\mu t} = e^{\sigma^2 t^2/2 + \mu t},\end{aligned}$$

$$M_{\mathcal{N}(\mu, \sigma^2)}(0) = e^{\sigma^2 0^2/2 + \mu 0} = 1,$$

$$\mathbf{E}X = M'_{\mathcal{N}(\mu, \sigma^2)}(0) = (\sigma^2 t + \mu)e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} = \mu,$$

$$\begin{aligned}\mathbf{E}X^2 &= M''_{\mathcal{N}(\mu, \sigma^2)}(0) = \sigma^2 e^{\sigma^2 t^2/2 + \mu t} + (\sigma^2 t + \mu)^2 e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} \\ &= \sigma^2 + \mu^2,\end{aligned}$$

from which $\mathbf{Var}X = \sigma^2$.

2. Examples

Example (Normal distribution)

Next, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$,

$$\begin{aligned} M_{\mathcal{N}(\mu, \sigma^2)}(t) &= \mathbf{E}e^{tX} = \mathbf{E}\left(e^{\sigma t \cdot \frac{X-\mu}{\sigma}}\right) \cdot e^{\mu t} \\ &= M_{\mathcal{N}(0, 1)}(\sigma t) \cdot e^{\mu t} = e^{\sigma^2 t^2/2 + \mu t}, \end{aligned}$$

$$M_{\mathcal{N}(\mu, \sigma^2)}(0) = e^{\sigma^2 0^2/2 + \mu 0} = 1,$$

$$\mathbf{E}X = M'_{\mathcal{N}(\mu, \sigma^2)}(0) = (\sigma^2 t + \mu)e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} = \mu,$$

$$\begin{aligned} \mathbf{E}X^2 &= M''_{\mathcal{N}(\mu, \sigma^2)}(0) = \sigma^2 e^{\sigma^2 t^2/2 + \mu t} + (\sigma^2 t + \mu)^2 e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} \\ &= \sigma^2 + \mu^2, \end{aligned}$$

from which $\mathbf{Var}X = \sigma^2$.

2. Examples

Example (Normal distribution)

Next, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$,

$$\begin{aligned} M_{\mathcal{N}(\mu, \sigma^2)}(t) &= \mathbf{E}e^{tX} = \mathbf{E}\left(e^{\sigma t \cdot \frac{X-\mu}{\sigma}}\right) \cdot e^{\mu t} \\ &= M_{\mathcal{N}(0, 1)}(\sigma t) \cdot e^{\mu t} = e^{\sigma^2 t^2/2 + \mu t}, \end{aligned}$$

$$M_{\mathcal{N}(\mu, \sigma^2)}(0) = e^{\sigma^2 0^2/2 + \mu 0} = 1,$$

$$\mathbf{E}X = M'_{\mathcal{N}(\mu, \sigma^2)}(0) = (\sigma^2 t + \mu)e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} = \mu,$$

$$\begin{aligned} \mathbf{E}X^2 &= M''_{\mathcal{N}(\mu, \sigma^2)}(0) = \sigma^2 e^{\sigma^2 t^2/2 + \mu t} + (\sigma^2 t + \mu)^2 e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} \\ &= \sigma^2 + \mu^2, \end{aligned}$$

from which $\mathbf{Var}X = \sigma^2$.

2. Examples

Example (Normal distribution)

Next, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$,

$$\begin{aligned} M_{\mathcal{N}(\mu, \sigma^2)}(t) &= \mathbf{E}e^{tX} = \mathbf{E}\left(e^{\sigma t \cdot \frac{X-\mu}{\sigma}}\right) \cdot e^{\mu t} \\ &= M_{\mathcal{N}(0, 1)}(\sigma t) \cdot e^{\mu t} = e^{\sigma^2 t^2/2 + \mu t}, \end{aligned}$$

$$M_{\mathcal{N}(\mu, \sigma^2)}(0) = e^{\sigma^2 0^2/2 + \mu 0} = 1,$$

$$\mathbf{E}X = M'_{\mathcal{N}(\mu, \sigma^2)}(0) = (\sigma^2 t + \mu)e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} = \mu,$$

$$\begin{aligned} \mathbf{E}X^2 &= M''_{\mathcal{N}(\mu, \sigma^2)}(0) = \sigma^2 e^{\sigma^2 t^2/2 + \mu t} + (\sigma^2 t + \mu)^2 e^{\sigma^2 t^2/2 + \mu t} \Big|_{t=0} \\ &= \sigma^2 + \mu^2, \end{aligned}$$

from which $\mathbf{Var}X = \sigma^2$.

2. Examples

Example (Gamma distribution)

Let $X \sim \text{Gamma}(\alpha, \lambda)$, and $t < \lambda$.

$$\begin{aligned}M(t) &= \mathbf{E}e^{tX} = \int_0^{\infty} e^{tx} \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} dx \\&= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha} \int_0^{\infty} \frac{(\lambda-t)^{\alpha} x^{\alpha-1} e^{-(\lambda-t)x}}{\Gamma(\alpha)} dx \\&= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha} \int_0^{\infty} f_{\text{Gamma}(\alpha, \lambda-t)}(x) dx = \left(\frac{\lambda}{\lambda-t}\right)^{\alpha},\end{aligned}$$

2. Examples

Example (Gamma distribution)

Let $X \sim \text{Gamma}(\alpha, \lambda)$, and $t < \lambda$.

$$\begin{aligned}M(t) &= \mathbf{E}e^{tX} = \int_0^{\infty} e^{tx} \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} dx \\&= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha} \int_0^{\infty} \frac{(\lambda-t)^{\alpha} x^{\alpha-1} e^{-(\lambda-t)x}}{\Gamma(\alpha)} dx \\&= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha} \int_0^{\infty} f_{\text{Gamma}(\alpha, \lambda-t)}(x) dx = \left(\frac{\lambda}{\lambda-t}\right)^{\alpha},\end{aligned}$$

2. Examples

Example (Gamma distribution)

Let $X \sim \text{Gamma}(\alpha, \lambda)$, and $t < \lambda$.

$$\begin{aligned}M(t) &= \mathbf{E}e^{tX} = \int_0^{\infty} e^{tx} \frac{\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} dx \\&= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha} \int_0^{\infty} \frac{(\lambda-t)^{\alpha} x^{\alpha-1} e^{-(\lambda-t)x}}{\Gamma(\alpha)} dx \\&= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha} \int_0^{\infty} f_{\text{Gamma}(\alpha, \lambda-t)}(x) dx = \left(\frac{\lambda}{\lambda-t}\right)^{\alpha},\end{aligned}$$

2. Examples

Example (Gamma distribution)

$$M(t) = \left(\frac{\lambda}{\lambda - t} \right)^\alpha,$$

$$M(0) = \left(\frac{\lambda}{\lambda - 0} \right)^\alpha = 1,$$

$$\mathbf{E}X = M'(0) = \alpha \frac{\lambda^\alpha}{(\lambda - t)^{\alpha+1}} \Big|_{t=0} = \frac{\alpha}{\lambda},$$

$$\mathbf{E}X^2 = M''(0) = \alpha(\alpha + 1) \frac{\lambda^\alpha}{(\lambda - t)^{\alpha+2}} \Big|_{t=0} = \frac{\alpha(\alpha + 1)}{\lambda^2},$$

and the variance is

$$\mathbf{Var}X = \frac{\alpha(\alpha + 1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.$$

3. Independent sums


The following two statements offer a very elegant and often used alternative to convolutions.

Proposition

Let X and Y be *independent* random variables. Then the moment generating function M_{X+Y} of their sum is of product form:

$$M_{X+Y}(t) = \mathbf{E}e^{t(X+Y)} = \mathbf{E}(e^{tX}e^{tY}) = \mathbf{E}e^{tX} \cdot \mathbf{E}e^{tY} = M_X(t) \cdot M_Y(t).$$

Theorem

The moment generating function on an open interval around zero uniquely determines the distribution. \rightsquigarrow 

3. Independent sums

Example (Binomials)

Let $X \sim \text{Binom}(n, p)$, $Y \sim \text{Binom}(m, p)$ be **independent**. Then $X + Y \sim \text{Binom}(n + m, p)$:

$$\begin{aligned}M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \\ &= (e^t p + 1 - p)^n \cdot (e^t p + 1 - p)^m = (e^t p + 1 - p)^{n+m},\end{aligned}$$

which is the $\text{Binom}(n + m, p)$ moment generating function, therefore $X + Y \sim \text{Binom}(n + m, p)$.

3. Independent sums

Example (Binomials)

Let $X \sim \text{Binom}(n, p)$, $Y \sim \text{Binom}(m, p)$ be **independent**. Then $X + Y \sim \text{Binom}(n + m, p)$:

$$\begin{aligned}M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \\ &= (e^t p + 1 - p)^n \cdot (e^t p + 1 - p)^m = (e^t p + 1 - p)^{n+m},\end{aligned}$$

which is the $\text{Binom}(n + m, p)$ moment generating function, therefore $X + Y \sim \text{Binom}(n + m, p)$.

3. Independent sums

Example (Binomials)

Let $X \sim \text{Binom}(n, p)$, $Y \sim \text{Binom}(m, p)$ be **independent**. Then $X + Y \sim \text{Binom}(n + m, p)$:

$$\begin{aligned}M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \\ &= (e^t p + 1 - p)^n \cdot (e^t p + 1 - p)^m = (e^t p + 1 - p)^{n+m},\end{aligned}$$

which is the $\text{Binom}(n + m, p)$ moment generating function, therefore $X + Y \sim \text{Binom}(n + m, p)$.

3. Independent sums

Example (Poissons)

Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ be independent. Then $X + Y \sim \text{Poi}(\lambda + \mu)$:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = e^{\lambda(e^t-1)} \cdot e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}.$$

Example (Normals)

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent. Then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$:

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) = e^{\sigma_1^2 t^2 / 2 + \mu_1 t} \cdot e^{\sigma_2^2 t^2 / 2 + \mu_2 t} \\ &= e^{(\sigma_1^2 + \sigma_2^2) t^2 / 2 + (\mu_1 + \mu_2) t}. \end{aligned}$$

3. Independent sums

Example (Poissons)

Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ be independent. Then $X + Y \sim \text{Poi}(\lambda + \mu)$:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = e^{\lambda(e^t-1)} \cdot e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}.$$

Example (Normals)

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent. Then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$:

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) = e^{\sigma_1^2 t^2 / 2 + \mu_1 t} \cdot e^{\sigma_2^2 t^2 / 2 + \mu_2 t} \\ &= e^{(\sigma_1^2 + \sigma_2^2) t^2 / 2 + (\mu_1 + \mu_2) t}. \end{aligned}$$

3. Independent sums

Example (Poissons)

Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ be independent. Then $X + Y \sim \text{Poi}(\lambda + \mu)$:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = e^{\lambda(e^t-1)} \cdot e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}.$$

Example (Normals)

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent. Then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$:

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) = e^{\sigma_1^2 t^2 / 2 + \mu_1 t} \cdot e^{\sigma_2^2 t^2 / 2 + \mu_2 t} \\ &= e^{(\sigma_1^2 + \sigma_2^2) t^2 / 2 + (\mu_1 + \mu_2) t}. \end{aligned}$$

3. Independent sums

Example (Poissons)

Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ be independent. Then $X + Y \sim \text{Poi}(\lambda + \mu)$:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = e^{\lambda(e^t-1)} \cdot e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}.$$

Example (Normals)

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent. Then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$:

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) = e^{\sigma_1^2 t^2 / 2 + \mu_1 t} \cdot e^{\sigma_2^2 t^2 / 2 + \mu_2 t} \\ &= e^{(\sigma_1^2 + \sigma_2^2) t^2 / 2 + (\mu_1 + \mu_2) t}. \end{aligned}$$

3. Independent sums

Example (Poissons)

Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$ be independent. Then $X + Y \sim \text{Poi}(\lambda + \mu)$:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = e^{\lambda(e^t-1)} \cdot e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}.$$

Example (Normals)

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent. Then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$:

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) = e^{\sigma_1^2 t^2 / 2 + \mu_1 t} \cdot e^{\sigma_2^2 t^2 / 2 + \mu_2 t} \\ &= e^{(\sigma_1^2 + \sigma_2^2) t^2 / 2 + (\mu_1 + \mu_2) t}. \end{aligned}$$

3. Independent sums

Example (Gammas)

Let $X \sim \text{Gamma}(\alpha, \lambda)$, $Y \sim \text{Gamma}(\beta, \lambda)$ be independent.
Then $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$:

$$\begin{aligned}M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \\ &= \left(\frac{\lambda}{\lambda - t}\right)^\alpha \cdot \left(\frac{\lambda}{\lambda - t}\right)^\beta = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha+\beta}.\end{aligned}$$

In particular, the convolution of n many
 $\text{Exp}(\lambda) = \text{Gamma}(1, \lambda)$'s is $\text{Gamma}(n, \lambda)$.

3. Independent sums

Example (Gammas)

Let $X \sim \text{Gamma}(\alpha, \lambda)$, $Y \sim \text{Gamma}(\beta, \lambda)$ be independent.
Then $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$:

$$\begin{aligned}M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \\ &= \left(\frac{\lambda}{\lambda - t}\right)^\alpha \cdot \left(\frac{\lambda}{\lambda - t}\right)^\beta = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha+\beta}.\end{aligned}$$

In particular, the convolution of n many
 $\text{Exp}(\lambda) = \text{Gamma}(1, \lambda)$'s is $\text{Gamma}(n, \lambda)$.

3. Independent sums

Example (Gammas)

Let $X \sim \text{Gamma}(\alpha, \lambda)$, $Y \sim \text{Gamma}(\beta, \lambda)$ be independent.
Then $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$:

$$\begin{aligned}M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \\ &= \left(\frac{\lambda}{\lambda - t}\right)^\alpha \cdot \left(\frac{\lambda}{\lambda - t}\right)^\beta = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha+\beta}.\end{aligned}$$

In particular, the convolution of n many
 $\text{Exp}(\lambda) = \text{Gamma}(1, \lambda)$'s is $\text{Gamma}(n, \lambda)$.

3. Independent sums

Example (Gammas)

Let $X \sim \text{Gamma}(\alpha, \lambda)$, $Y \sim \text{Gamma}(\beta, \lambda)$ be independent.
Then $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$:

$$\begin{aligned}M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \\ &= \left(\frac{\lambda}{\lambda - t}\right)^\alpha \cdot \left(\frac{\lambda}{\lambda - t}\right)^\beta = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha+\beta}.\end{aligned}$$

In particular, the convolution of n many
 $\text{Exp}(\lambda) = \text{Gamma}(1, \lambda)$'s is $\text{Gamma}(n, \lambda)$.

3. Independent sums

Example (Gammas)

Let $X \sim \text{Gamma}(\alpha, \lambda)$, $Y \sim \text{Gamma}(\beta, \lambda)$ be independent.
Then $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$:

$$\begin{aligned}M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) \\ &= \left(\frac{\lambda}{\lambda - t}\right)^\alpha \cdot \left(\frac{\lambda}{\lambda - t}\right)^\beta = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha+\beta}.\end{aligned}$$

In particular, the convolution of n many
 $\text{Exp}(\lambda) = \text{Gamma}(1, \lambda)$'s is $\text{Gamma}(n, \lambda)$.

7. Law of Large Numbers, Central Limit Theorem

Markov's, Chebyshev's inequality

Weak Law of Large Numbers

Central Limit Theorem

Objectives:

- ▶ To get familiar with general inequalities like Markov's and Chebyshev's
- ▶ To (almost) prove and use the Weak Law of Large Numbers
- ▶ To (almost) prove and use the Central Limit Theorem

Markov's, Chebyshev's inequality

Knowing a distribution of a random variable makes it possible to compute its moments. Vice-versa, knowing a few moments gives some bounds on certain probabilities. We'll explore such bounds in this part.

Our bounds here will be very general, and that makes them very useful in theoretical considerations. The price to pay is that they are often not sharp enough for practical applications.

1. Markov's inequality


Theorem (Markov's inequality)

Let X be a *non-negative* random variable. Then for all $a > 0$ reals,

$$\mathbf{P}\{X \geq a\} \leq \frac{\mathbf{E}X}{a}.$$

Of course this inequality is useless for $a \leq \mathbf{E}X$.

Proof.

Let $Y = \begin{cases} 1, & \text{if } X \geq a, \\ 0, & \text{if } X < a. \end{cases}$ Then $X \geq aY \rightsquigarrow$  , thus

$$\mathbf{E}X \geq \mathbf{E}aY = a\mathbf{E}Y = a\mathbf{P}\{X \geq a\}.$$



2. Chebyshev's inequality

Theorem (Chebyshev's inequality)

Let X be a random variable with mean μ and variance σ^2 both finite. Then for all $b > 0$ reals,

$$\mathbf{P}\{|X - \mu| \geq b\} \leq \frac{\mathbf{Var}X}{b^2}.$$

Of course this inequality is useless for $b \leq \text{SD } X$.

Proof.

Apply Markov's inequality on the random variable $(X - \mu)^2 \geq 0$:

$$\mathbf{P}\{|X - \mu| \geq b\} = \mathbf{P}\{(X - \mu)^2 \geq b^2\} \leq \frac{\mathbf{E}(X - \mu)^2}{b^2} = \frac{\mathbf{Var}X}{b^2}.$$



3. Examples

Example

On average, 50 gadgets per day are manufactured in a factory. What can be said on the probability that at least 75 will be produced tomorrow?

Let X be the number of gadgets produced tomorrow. Clearly $X \geq 0$, thus using Markov's inequality,

$$\mathbf{P}\{X \geq 75\} \leq \frac{\mathbf{E}X}{75} = \frac{50}{75} = \frac{2}{3}.$$

Notice that *nothing* was assumed on the distribution of X except its mean!

3. Examples

Example

On average, 50 gadgets per day are manufactured in a factory and the standard deviation of this number is 5. What can be said on the probability that more than 40 but less than 60 will be produced tomorrow?

Let X be the number of gadgets produced tomorrow, $\mu = 50$, $\text{Var}X = 25$. By Chebyshev's inequality,

$$\begin{aligned}\mathbf{P}\{40 < X < 60\} &= \mathbf{P}\{-10 < X - \mu < 10\} = \mathbf{P}\{|X - \mu| < 10\} \\ &= 1 - \mathbf{P}\{|X - \mu| \geq 10\} \\ &\geq 1 - \frac{\text{Var}X}{10^2} = 1 - \frac{25}{10^2} = \frac{3}{4}.\end{aligned}$$

Again, only μ and σ^2 were needed.

Weak Law of Large Numbers

Finally, we arrived to the first of two very important theorems we cover. The *Law of Large Numbers* will tell us that the sample mean of an i.i.d. sample converges to the expectation of the individual variables.

A suitable application will make the connection between our abstract definition of probability and relative frequencies observed in a large number of independent trials.

Weak Law of Large Numbers

Theorem (Weak Law of Large Numbers (WLLN))

Let X_1, X_2, \dots be a sequence of i.i.d. random variables with finite mean μ . Then for every $\varepsilon > 0$,

$$\mathbf{P}\left\{\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \mu\right| > \varepsilon\right\} \xrightarrow{n \rightarrow \infty} 0.$$

Weak Law of Large Numbers

Proof.

We prove a bit less: we'll assume a finite variance σ^2 for our variables. The WLLN is true without this assumption.

$$\begin{aligned} \mathbf{P}\left\{\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right\} &= \mathbf{P}\{|\bar{X} - \mu| > \varepsilon\} \\ &\leq \frac{\mathbf{Var}\bar{X}}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$



Weak Law of Large Numbers

Proof.

We prove a bit less: we'll assume a finite variance σ^2 for our variables. The WLLN is true without this assumption.

$$\begin{aligned} \mathbf{P}\left\{\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right\} &= \mathbf{P}\left\{|\bar{X} - \mu| > \varepsilon\right\} \\ &\leq \frac{\mathbf{Var}\bar{X}}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$



Weak Law of Large Numbers

Proof.

We prove a bit less: we'll assume a finite variance σ^2 for our variables. The WLLN is true without this assumption.

$$\begin{aligned} \mathbf{P}\left\{\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right\} &= \mathbf{P}\{|\bar{X} - \mu| > \varepsilon\} \\ &\leq \frac{\mathbf{Var}\bar{X}}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$



Weak Law of Large Numbers

Corollary

Let $Y_n \sim \text{Binom}(n, p)$. Then for all $\varepsilon > 0$,

$$\mathbf{P}\left\{\left|\frac{Y_n}{n} - p\right| \geq \varepsilon\right\} \xrightarrow{n \rightarrow \infty} 0.$$

Notice that $\frac{Y_n}{n}$ is precisely the relative frequency of successes in n independent trials, and p is the probability of success in a particular trial.

Weak Law of Large Numbers

Proof.

Let $X_i, i = 1 \dots n$ be i.i.d. **Bernoulli(p)** variables. Then we know

$$Y_n \stackrel{d}{=} \sum_{i=1}^n X_i \quad \text{and} \quad \mathbf{E}X_i = \mu = p,$$

hence by the WLLN

$$\mathbf{P}\left\{\left|\frac{Y_n}{n} - p\right| \geq \varepsilon\right\} = \mathbf{P}\left\{\left|\frac{\sum_{i=1}^n X_i}{n} - \mu\right| \geq \varepsilon\right\} \xrightarrow[n \rightarrow \infty]{} 0.$$



Central Limit Theorem

The WLLN tells us that the sample mean of an i.i.d. sequence is close to the expectation of the variables. A second, finer approach will be the Central Limit Theorem. It will tell us the order of magnitude of the distance between the sample mean and the true mean of our random variables.

Central Limit Theorem

We start with an auxiliary proposition without proof.

Proposition

Suppose that the moment generating function values $M_{Z_n}(t)$ of a sequence Z_n of random variables converge to the moment generating function $M_Z(t)$ of a random variable Z at every t of an open interval that contains 0. Then the distribution function values $F_{Z_n}(z)$ of Z_n converge to $F_Z(z)$ at every point z where this latter is continuous.

Central Limit Theorem

Theorem (Central Limit Theorem (CLT))

Let X_1, X_2, \dots be i.i.d. random variables with both their mean μ and variance σ^2 finite. Then for every real $a < b$,

$$\mathbf{P}\left\{a < \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \leq b\right\} \xrightarrow{n \rightarrow \infty} \Phi(b) - \Phi(a).$$

Remark

Notice the mean $n\mu$ and standard deviation $\sqrt{n}\sigma$ of the sum $X_1 + X_2 + \dots + X_n$.

Central Limit Theorem

Theorem (Central Limit Theorem (CLT))

Let X_1, X_2, \dots be i.i.d. random variables with both their mean μ and variance σ^2 finite. Then for every real $a < b$,

$$\mathbf{P}\left\{a < \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \leq b\right\} \xrightarrow{n \rightarrow \infty} \Phi(b) - \Phi(a).$$

Remark

Notice the mean $n\mu$ and standard deviation $\sqrt{n}\sigma$ of the sum $X_1 + X_2 + \dots + X_n$.

Central Limit Theorem

Theorem (Central Limit Theorem (CLT))

Let X_1, X_2, \dots be i.i.d. random variables with both their mean μ and variance σ^2 finite. Then for every real $a < b$,

$$\mathbf{P}\left\{a < \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \leq b\right\} \xrightarrow{n \rightarrow \infty} \Phi(b) - \Phi(a).$$

Remark

Notice the mean $n\mu$ and standard deviation $\sqrt{n}\sigma$ of the sum $X_1 + X_2 + \dots + X_n$.

Central Limit Theorem

Remark

When X_i are i.i.d. Bernoulli(p), $\mu = p$, $\sigma^2 = p(1 - p)$,
 $Y_n := X_1 + \cdots + X_n \sim \text{Binom}(n, p)$, and the CLT becomes the
DeMoivre-Laplace Theorem:

$$\begin{aligned} \mathbf{P}\left\{a < \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \leq b\right\} \\ = \mathbf{P}\left\{a < \frac{Y_n - np}{\sqrt{np(1-p)}} \leq b\right\} \xrightarrow{n \rightarrow \infty} \Phi(b) - \Phi(a). \end{aligned}$$

Central Limit Theorem

Proof.

Again, we prove a little bit less: we assume that the moment generating function $M(t)$ of the X_i 's is finite for small t 's.

Assume first $\mu = 0$, $\sigma = 1$. Then we look at $\sum_{i=1}^n \frac{X_i}{\sqrt{n}} \rightsquigarrow \text{[drawing icon]}$. Its moment generating function is

$$\begin{aligned} M_{\sum_{i=1}^n \frac{X_i}{\sqrt{n}}}(t) &= \mathbf{E} e^{t \sum_{i=1}^n \frac{X_i}{\sqrt{n}}} = \mathbf{E} \prod_{i=1}^n e^{t \frac{X_i}{\sqrt{n}}} = \prod_{i=1}^n \mathbf{E} e^{(\frac{t}{\sqrt{n}}) X_i} \\ &= \prod_{i=1}^n M\left(\frac{t}{\sqrt{n}}\right) = \left[M\left(\frac{t}{\sqrt{n}}\right)\right]^n. \end{aligned}$$

We'll take log of this function, therefore define $\Psi(x) := \ln M(x)$.

Central Limit Theorem

Proof.

For small x 's,

$$\begin{aligned}\Psi(x) &= \Psi(0) + \Psi'(0) \cdot x + \Psi''(0) \cdot \frac{x^2}{2} + \mathcal{O}(x^3) \\ &= 0 + \mathbf{EX} \cdot x + \mathbf{Var}X \cdot \frac{x^2}{2} + \mathcal{O}(x^3) \\ &= 0 + 0 \cdot x + 1 \cdot \frac{x^2}{2} + \mathcal{O}(x^3). \quad \rightsquigarrow \text{📝}\end{aligned}$$

Central Limit Theorem

Proof.

$\Psi(x) = \frac{x^2}{2} + \mathcal{O}(x^3)$. Therefore,

$$\begin{aligned}\ln M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) &= \ln \left[M\left(\frac{t}{\sqrt{n}}\right) \right]^n = n \ln M\left(\frac{t}{\sqrt{n}}\right) = n\Psi\left(\frac{t}{\sqrt{n}}\right) \\ &= n \cdot \frac{t^2}{2n} + n\mathcal{O}\left(\frac{t^3}{n^{3/2}}\right) \xrightarrow{n \rightarrow \infty} \frac{t^2}{2},\end{aligned}$$

$$M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) \xrightarrow{n \rightarrow \infty} e^{t^2/2} = M_{\mathcal{N}(0,1)}(t),$$

and we are done in the case $\mu = 0, \sigma = 1$.

Central Limit Theorem

Proof.

$\Psi(\mathbf{x}) = \frac{x^2}{2} + \mathcal{O}(x^3)$. Therefore,

$$\begin{aligned}\ln M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) &= \ln \left[M\left(\frac{t}{\sqrt{n}}\right) \right]^n = n \ln M\left(\frac{t}{\sqrt{n}}\right) = n\Psi\left(\frac{t}{\sqrt{n}}\right) \\ &= n \cdot \frac{t^2}{2n} + n\mathcal{O}\left(\frac{t^3}{n^{3/2}}\right) \xrightarrow{n \rightarrow \infty} \frac{t^2}{2},\end{aligned}$$

$$M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) \xrightarrow{n \rightarrow \infty} e^{t^2/2} = M_{\mathcal{N}(0,1)}(t),$$

and we are done in the case $\mu = 0, \sigma = 1$.

Central Limit Theorem

Proof.

$\Psi(x) = \frac{x^2}{2} + \mathcal{O}(x^3)$. Therefore,

$$\begin{aligned}\ln M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) &= \ln \left[M\left(\frac{t}{\sqrt{n}}\right) \right]^n = n \ln M\left(\frac{t}{\sqrt{n}}\right) = n\Psi\left(\frac{t}{\sqrt{n}}\right) \\ &= n \cdot \frac{t^2}{2n} + n\mathcal{O}\left(\frac{t^3}{n^{3/2}}\right) \xrightarrow{n \rightarrow \infty} \frac{t^2}{2},\end{aligned}$$

$$M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) \xrightarrow{n \rightarrow \infty} e^{t^2/2} = M_{\mathcal{N}(0,1)}(t),$$

and we are done in the case $\mu = 0, \sigma = 1$.

Central Limit Theorem

Proof.

$\Psi(x) = \frac{x^2}{2} + \mathcal{O}(x^3)$. Therefore,

$$\begin{aligned}\ln M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) &= \ln \left[M\left(\frac{t}{\sqrt{n}}\right) \right]^n = n \ln M\left(\frac{t}{\sqrt{n}}\right) = n\Psi\left(\frac{t}{\sqrt{n}}\right) \\ &= n \cdot \frac{t^2}{2n} + n\mathcal{O}\left(\frac{t^3}{n^{3/2}}\right) \xrightarrow{n \rightarrow \infty} \frac{t^2}{2},\end{aligned}$$

$$M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) \xrightarrow{n \rightarrow \infty} e^{t^2/2} = M_{\mathcal{N}(0,1)}(t),$$

and we are done in the case $\mu = 0, \sigma = 1$.

Central Limit Theorem

Proof.

$\Psi(x) = \frac{x^2}{2} + \mathcal{O}(x^3)$. Therefore,

$$\begin{aligned}\ln M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) &= \ln \left[M\left(\frac{t}{\sqrt{n}}\right) \right]^n = n \ln M\left(\frac{t}{\sqrt{n}}\right) = n\Psi\left(\frac{t}{\sqrt{n}}\right) \\ &= n \cdot \frac{t^2}{2n} + n\mathcal{O}\left(\frac{t^3}{n^{3/2}}\right) \xrightarrow{n \rightarrow \infty} \frac{t^2}{2},\end{aligned}$$

$$M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) \xrightarrow{n \rightarrow \infty} e^{t^2/2} = M_{\mathcal{N}(0,1)}(t),$$

and we are done in the case $\mu = 0, \sigma = 1$.

Central Limit Theorem

Proof.

$\Psi(x) = \frac{x^2}{2} + \mathcal{O}(x^3)$. Therefore,

$$\begin{aligned} \ln M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) &= \ln \left[M\left(\frac{t}{\sqrt{n}}\right) \right]^n = n \ln M\left(\frac{t}{\sqrt{n}}\right) = n\Psi\left(\frac{t}{\sqrt{n}}\right) \\ &= n \cdot \frac{t^2}{2n} + n\mathcal{O}\left(\frac{t^3}{n^{3/2}}\right) \xrightarrow{n \rightarrow \infty} \frac{t^2}{2}, \end{aligned}$$

$$M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) \xrightarrow{n \rightarrow \infty} e^{t^2/2} = M_{\mathcal{N}(0,1)}(t),$$

and we are done in the case $\mu = 0, \sigma = 1$.

Central Limit Theorem

Proof.

$\Psi(x) = \frac{x^2}{2} + \mathcal{O}(x^3)$. Therefore,

$$\begin{aligned}\ln M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) &= \ln \left[M\left(\frac{t}{\sqrt{n}}\right) \right]^n = n \ln M\left(\frac{t}{\sqrt{n}}\right) = n\Psi\left(\frac{t}{\sqrt{n}}\right) \\ &= n \cdot \frac{t^2}{2n} + n\mathcal{O}\left(\frac{t^3}{n^{3/2}}\right) \xrightarrow{n \rightarrow \infty} \frac{t^2}{2},\end{aligned}$$

$$M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) \xrightarrow{n \rightarrow \infty} e^{t^2/2} = M_{\mathcal{N}(0,1)}(t),$$

and **we are done** in the case $\mu = 0, \sigma = 1$.

Central Limit Theorem

Proof.

For general μ and σ^2 ,

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{\frac{X_1 - \mu}{\sigma} + \frac{X_2 - \mu}{\sigma} + \cdots + \frac{X_n - \mu}{\sigma}}{\sqrt{n}},$$

and $\frac{X_i - \mu}{\sigma}$ are i.i.d. with mean zero and variance 1, thus we can apply the previous case to conclude the proof. \square

Remark

A natural question is where the CLT starts giving a useful approximation for practical purposes. People usually agree that, depending on the required level of accuracy, one or two dozens of random variables are enough in most cases. See the *Berry-Esseen theorems* for a theoretical bound on this.

Central Limit Theorem

Proof.

For general μ and σ^2 ,

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{\frac{X_1 - \mu}{\sigma} + \frac{X_2 - \mu}{\sigma} + \cdots + \frac{X_n - \mu}{\sigma}}{\sqrt{n}},$$

and $\frac{X_i - \mu}{\sigma}$ are i.i.d. with mean zero and variance 1, thus we can apply the previous case to conclude the proof. \square

Remark

A natural question is where the CLT starts giving a useful approximation for practical purposes. People usually agree that, depending on the required level of accuracy, one or two dozens of random variables are enough in most cases. **See the *Berry-Esseen theorems* for a theoretical bound on this.**

Central Limit Theorem

Proof.

For general μ and σ^2 ,

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{\frac{X_1 - \mu}{\sigma} + \frac{X_2 - \mu}{\sigma} + \cdots + \frac{X_n - \mu}{\sigma}}{\sqrt{n}},$$

and $\frac{X_i - \mu}{\sigma}$ are i.i.d. with mean zero and variance 1, thus we can apply the previous case to conclude the proof. \square

Remark

A natural question is where the CLT starts giving a useful approximation for practical purposes. People usually agree that, depending on the required level of accuracy, one or two dozens of random variables are enough in most cases. See the *Berry-Esseen theorems* for a theoretical bound on this.

Central Limit Theorem

Example

An astronomer measures the unknown distance μ of an astronomical object. He performs n i.i.d. measurements each with mean μ and standard deviation 2 lightyears. How large should n be to have ± 0.5 lightyears accuracy with at least 95% probability?

Central Limit Theorem

Solution

Clearly, the outcome of the n measurements will be the sample mean \bar{X} . We wish to bring this closer to μ than 0.5 with high probability:

$$\begin{aligned} 0.95 &\leq \mathbf{P}\left\{\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \leq 0.5\right\} \\ &= \mathbf{P}\left\{\left|\frac{X_1 + X_2 + \cdots + X_n - n\mu}{2\sqrt{n}}\right| \leq \frac{0.5\sqrt{n}}{2}\right\} \\ &= \mathbf{P}\left\{-\frac{\sqrt{n}}{4} \leq \frac{X_1 + X_2 + \cdots + X_n - n\mu}{2\sqrt{n}} \leq \frac{\sqrt{n}}{4}\right\}. \end{aligned}$$

We now assume that we need enough measurements for the CLT to kick in:

Central Limit Theorem

Solution

Clearly, the outcome of the n measurements will be the sample mean \bar{X} . We wish to bring this closer to μ than 0.5 with high probability:

$$\begin{aligned}
 0.95 &\leq \mathbf{P}\left\{\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \leq 0.5\right\} \\
 &= \mathbf{P}\left\{\left|\frac{X_1 + X_2 + \cdots + X_n - n\mu}{2\sqrt{n}}\right| \leq \frac{0.5\sqrt{n}}{2}\right\} \\
 &= \mathbf{P}\left\{-\frac{\sqrt{n}}{4} \leq \frac{X_1 + X_2 + \cdots + X_n - n\mu}{2\sqrt{n}} \leq \frac{\sqrt{n}}{4}\right\}.
 \end{aligned}$$

We now assume that we need enough measurements for the CLT to kick in:

Central Limit Theorem

Solution (... cont'd)

$$0.95 \leq \mathbf{P} \left\{ -\frac{\sqrt{n}}{4} \leq \frac{X_1 + X_2 + \cdots + X_n - n\mu}{2\sqrt{n}} \leq \frac{\sqrt{n}}{4} \right\}$$

$$0.95 \lesssim \Phi\left(\frac{\sqrt{n}}{4}\right) - \Phi\left(-\frac{\sqrt{n}}{4}\right) = 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1$$

$$0.975 \lesssim \Phi\left(\frac{\sqrt{n}}{4}\right)$$

$$1.96 \lesssim \frac{\sqrt{n}}{4}$$

$$61.5 \lesssim n,$$

the astronomer needs at least 62 measurements.

Closing remarks

This unit was the first step in your probability and statistics studies at the University. Probability has a broad range of applications in e.g.

- ▶ statistics
- ▶ financial mathematics
- ▶ actuarial sciences
- ▶ physics
- ▶ computer sciences and electrical engineering
- ▶ social sciences
- ▶ traffic engineering
- ▶ biology

Closing remarks

as well as beautiful pure mathematical research questions on its own connecting to, among others,

- ▶ analysis (potential theory, asymptotic analysis, complex function theory, functional analysis, dynamical systems, fractals)
- ▶ combinatorics (counting problems, coloring, combinatorial processes, graph theory)
- ▶ mathematical physics (statistical physics problems, quantum systems)
- ▶ number theory (Riemann Zeta function, distribution of primes)
- ▶ algebra (group theory, random matrix theory).

Closing remarks

Your next unit related to probability will be *Statistics 1*. For further studies in probability, please consult the aids for unit choices, available on Intranet.

Thank you.