1 Introduction

Looking for a match

The problem. Items are inspected until a pair of matching types is found. Assuming that there are \( r \) types and each item is equally likely to be any of the types, how many would you expect to look at in order to find a type match with an item already inspected?

Example. Ask people in the class until a matching birthday is found (\( r = 365 \)). How many would you expect to ask?

A related problem. How many would you need to inspect in order to be at least 50% certain of getting at least one match?

Let’s do it for birthdays.

\[ p_n := \Pr(\text{At least one match in a group of } n). \]

\[ 1 - p_n = \Pr(\text{All } n \text{ have different birthdays}) \]

\[ = 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365 - n + 1}{365} \]

\[ = \prod_{i=1}^{n-1} \left(1 - \frac{i}{365}\right). \]

Use an approximation. We know that, for \( x \) such that \(-1 < x \leq 1\),

\[ \log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \]

\[ \simeq x \text{ when } |x| \text{ is small.} \]

Thus, for \( n << 365 \),

\[ \log(1 - p_n) = \sum_{i=1}^{n-1} \log \left(1 - \frac{i}{365}\right) \simeq -\sum_{i=1}^{n-1} \frac{i}{365} \]

\[ = -\frac{1}{365} \cdot \frac{1}{2} n(n - 1). \]

and so \( p_n \simeq 1 - \exp(-n(n - 1)/730) \). Hence,

\[ p_{23} \simeq 1 - \exp(-23 \cdot 22/730) \simeq 0.500002. \]

The answer to the second problem is 23.
What about the original problem?

Let $N$ be the number of people asked in order to get the first match (this must be at least 2, but a match is assured once we ask 366 people).

The expected value of $N$ is given by

$$E(N) = \sum_{m=1}^{366} m \Pr(N = m) = \sum_{m=1}^{366} \sum_{n=1}^{m} \Pr(N = m)$$

$$= \sum_{n=1}^{366} \sum_{m=n}^{366} \Pr(N = m) = \sum_{n=1}^{366} \Pr(N \geq n)$$

$$= \sum_{n=0}^{365} \Pr(N > n) = \sum_{n=0}^{365} (1 - \Pr(N \leq n))$$

But, $N \leq n$ if and only if, in a group of $n$, there is at least one match. Hence,

$$E(N) = \sum_{n=0}^{365} (1 - p_n) = 2 + \sum_{n=2}^{365} \prod_{i=1}^{n-1} \left(1 - \frac{i}{365}\right).$$

This gives the result $E(N) \simeq 23.6139$. 
2 Events

A random experiment or trial is an experiment whose outcome is not known in advance.

Sample space. The set $S$ of all possible outcomes is called the sample space. Outcomes are sometimes called sample points.

Example. The lifetime of a component is measured: $S = [0, \infty)$. This is an example of a continuous sample space.

Example. Count the number of packets to arrive at a node in a communications network during a one-minute period: $S = \{0, 1, \ldots\}$. This is an example of a discrete sample space.

Events. An event is any set of outcomes (any subset of $S$). Events are usually denoted by upper case letters, $A$, $B$, $C$, et cetera.

Example. Let $A$ be the event that the lifetime is at least 400 hours: $A = [400, \infty)$.

Example. Let $A$ be the event that between 1000 and 2000 packets arrive in a one-minute period: $A = \{1000, 1001, \ldots, 2000\}$.

Combining events. If $A$ and $B$ are two events, then the event that $A$ or $B$ occurs is represented by the union of $A$ and $B$, denoted $A \cup B$, and consists of all outcomes in $A$ or in $B$ (or in both).

Example. Let $A$ be the event that a measured voltage is less than $-5$ volts: $A = \{v : v < -5\}$. Let $B$ be the event that the voltage is greater than 5 volts: $B = \{v : v > 5\}$. Then, $A \cup B = \{v : |v| > 5\}$ is the event that the voltage is less than $-5$ volts or greater than 5 volts.

If $A$ and $B$ are two events, then the event that $A$ and $B$ occur is represented by the intersection of $A$ and $B$, denoted $A \cap B$, and consists of all outcomes in both $A$ and $B$.

Example. In the setup of the previous example, let $C = \{v : v > -5\}$ and $D = \{v : v < 5\}$. Then, $C \cap D = \{v : |v| < 5\}$ is the event that the voltage is between $-5$ volts and 5 volts.
More generally, if \( \{ A_i \} \) is a collection of events, then \( \bigcup_i A_i = A_1 \cup A_2 \cup \ldots \) is the event that \( A_i \) occurs for at least one value of \( i \), while \( \bigcap_i A_i = A_1 \cap A_2 \cap \ldots \) is the event that \( A_i \) occurs for all values of \( i \).

**Example.** A system consists of several components. Let \( A_i \) be the event that component \( i \) fails. If all the components are in series, then \( \bigcup_i A_i \) is the event that the system fails. If all the components are in parallel, then \( \bigcap_i A_i \) is the event that the system fails.

If \( A \) is an event, then \( A^c \) (pronounced \( A \) complement), the event that \( A \) does not occur, consists of all outcomes in \( S \) which are not in \( A \).

**Example.** In the setup of the previous example, if all the components are in series, then \( (\bigcup_i A_i)^c \) is the event that the system functions. If all the components are in parallel, then \( (\bigcap_i A_i)^c \) is the event that the system functions.

Note (De Morgan’s Laws):

\[
(\bigcup_i A_i)^c = \bigcap_i A_i^c \quad \text{and} \quad (\bigcap_i A_i)^c = \bigcup_i A_i^c.
\]

If \( A \) and \( B \) are two events, and all the outcomes in \( A \) are also in \( B \), then \( A \) is a subset of \( B \), written \( A \subset B \). Clearly this means that if \( A \) occurs then \( B \) must occur (\( A \) implies \( B \)).

The empty set \( \emptyset \) is called the impossible event because it never occurs, while its complement \( S \) is called the certain event because it always occurs.

Two events, \( A \) and \( B \), are said to be disjoint if they have no outcomes in common (they cannot occur together), that is, \( A \cap B = \emptyset \). More generally, \( A_1, A_2, \ldots \) are said to be mutually exclusive if each pair of events \( (A_i, A_j), i \neq j \), is disjoint.

\( A_1, A_2, \ldots \) are said to be exhaustive if \( A_1 \cup A_2 \cup \cdots = S \), that is, at least one of these events must occur. A collection \( \mathcal{P} = \{ A_1, A_2, \ldots \} \) of mutually exclusive and exhaustive events is called a partition of \( S \).
Venn Diagrams

Venn diagrams are used to illustrate events and the relationship between them (complement, union, intersection, partition).

They were popularized by John Venn (1834-1923)\textsuperscript{1}.

He was an Anglican cleric and taught at Gonville and Caius College, Cambridge.

\textsuperscript{1}Venn J., Symbolic Logic, MacMillan, London 1881, 2nd Ed., 1894.

Fig 1. Venn diagram illustrating three events, $F$, $G$ and $C$.

Fig 2. The event $C$ is shaded...
Fig 3. The event $C^c$ is shaded

Fig 4. The event $F \cap G$ is shaded

Fig 5. The event $F \cup G$ is shaded

Fig 6. The event $F \cap G^c$ is shaded
Let $A_i$ be the event that the $i$th component is functioning, $i = 1, 2, 3$.

Let $D_a$, $D_b$, $D_c$ be the events that resp. the series, parallel and 2-out-of-3 system is functioning.

Then,

$$D_a = A_1 \cap A_2 \cap A_3.$$  
$$D_b = A_1 \cup A_2 \cup A_3.$$  
$$D_c = (A_1 \cap A_2 \cap A_3) \cup (A_1 \cap A_2 \cap A_3^c) \cup (A_1 \cap A_2^c \cap A_3) \cup (A_1^c \cap A_2 \cap A_3^c) \cup (A_1^c \cap A_2^c \cap A_3) \cup (A_1 \cap A_2^c \cap A_3^c).$$
3 The Rules of Probability

The third and final ingredient in the model for a random experiment is to seek a “measure” which tells us how likely it is that a particular event will occur.

Definition. A probability \( P \) is a rule (or function) which assigns a number between 0 and 1 to each event, and which satisfies the following axioms:

Axiom 1: \( 0 \leq P(A) \leq 1 \).

Axiom 2: \( P(S) = 1 \).

Axiom 3: For any disjoint \( A_1, A_2, \ldots \)

\[
P(\bigcup_i A_i) = \sum_i P(A_i).
\]

Note that the probability (measure) has similar properties to the measure of length, area, volume, weight, etc.

Example. Random experiment: throw a die. Here, the sample space is \( S = \{1, 2, \ldots, 6\} \). Let all subsets of \( S \) be events, and define \( P \) by

\[
P(A) = \frac{|A|}{6}, \ A \subset S.
\]

This completely specifies/models the experiment. For example, the probability of getting an even number is \( P(\{2, 4, 6\}) = 3/6 = 1/2 \).

Some consequences of the Axioms:

- If \( A \subset B \) then \( P(A) \leq P(B) \).
- \( P(\emptyset) = 0 \).
- \( P(A^c) = 1 - P(A) \)
- \( P(A \cup B) = P(A) + P(B) - P(A \cap B) \)
Discrete Sample Spaces

Let $S$ be a discrete sample space, e.g. $S = \{a_1, a_2, \ldots, a_n\}$. Let $P(\{a_i\}) = p_i$, for $i = 1, \ldots, n$, and define

$$P(A) = \sum_{i:a_i \in A} p_i, \text{ for all } A \subset S.$$ 

Then $P$ is a probability.

Thus we can specify $P$ by specifying only the probabilities of the elementary events $\{a_i\}$.

Example. We draw two cards from a full deck of 52 cards (no jokers). What is the probability of drawing at least one Ace?

Give all the cards a number from 1 to 52. Draw the cards one-by-one. Write a possible outcome as $(x, y)$, $x \neq y$.

Each elementary event $\{(x, y)\}$ has the same probability $1/(52 \times 51)$.

Let $A$ be the event: “at least one Ace”. Then,

$$P(A) = \sum_{(x,y) \in A} \frac{1}{52 \times 51} = \frac{|A|}{52 \times 51}.$$ 

We need to count how many elements are in $A$. Easier: $|A^c| = 48 \times 47$. Hence,

$$P(A) = 1 - P(A^c) = 1 - \frac{48 \times 47}{52 \times 51} \approx 0.15.$$
Remark. In many cases, as above, we can choose $S$ such that each elementary event is equally likely, i.e. $\mathbb{P}(\{a_i\}) = 1/n$.

Question. What if we choose the two cards at the same time (no order). Does that change the model? Does it change the probability?

Exercises.

1. We randomly take a ball from an urn containing 10 balls, numbered 0,1, $\ldots$ ,9. Find the probability of the event “the number selected is a multiple of 2 or 5”.

2. We randomly take two balls from the above urn (without putting them back). Find the probability of the event “the sum of the numbers is less than 3”.

3. Give a model for the experiment in which we toss with a fair coin and note (only) the number of tosses required until “heads”.

Remark. Problems of the type above often involve counting techniques (combinatorics).

Remark. Note that Exercise 3 above is an example showing how also for infinite discrete sample spaces we can specify $\mathbb{P}$ via the elementary events.

Continuous Sample Spaces

For continuous sample spaces, we can no longer specify $\mathbb{P}$ via the elementary events. Why not?

Example. We select randomly a point in the unit square.
Here, $S = \{(x, y), 0 \leq x \leq 1, 0 \leq y \leq 1\}$.

We can define $\mathbb{P}$ by specifying the probability of rectangles $[a, b] \times [c, d]$ in the unit square. For example, defining

$$\mathbb{P}([a, b] \times [c, d]) = (b - a)(d - c)$$

assigns a “uniform” probability. Each point is equally likely. For a set $A$, $\mathbb{P}(A)$ is thus equal to the area of $A$.

Exercise. For the above experiment, with the uniform probability measure, calculate the probability that the sum of the $x$ and $y$ coordinate of the selected point is larger than 1.

4 Conditional Probability

How do probabilities change when we know some event $B \subset S$ has occurred?

Suppose $B$ has occurred. Thus, we know that the outcome lies in $B$. Then $A$ will occur if and only if $A \cap B$ occurs, and the relative chance of $A$ occurring is therefore $\mathbb{P}(A \cap B)/\mathbb{P}(B)$.
This leads to the definition of the conditional probability of $A$ given $B$:

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

**Example.** We throw two dice. Given that the sum of the eyes is 10, what is the probability that one 6 is cast.

Let $B$ be the event that the sum is 10, 

$$B = \{(4, 6), (5, 5), (6, 4)\}.$$ 

Let $A$ be the event that one 6 is cast, 

$$A = \{(1, 6), \ldots, (5, 6), (6, 1), \ldots, (6, 5)\}.$$ 

Then, $A \cap B = \{(4, 6), (6, 4)\}$. And, since all elementary events are equally likely, we have 

$$P(A \mid B) = \frac{2/36}{3/36} = \frac{2}{3}.$$ 

**Chain rule**

By the definition of conditional probability we have 

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

We can generalize this to $n$ intersections $A_1 \cap A_2 \cap \cdots \cap A_n$ (abbreviated as $A_1 A_2 \cdots A_n$).

This gives the chain rule of probability:

$$P(A_1 A_2 \cdots A_n) = P(A_1)P(A_2 \mid A_1) \cdots P(A_n \mid A_1 A_2 \cdots A_{n-1}).$$
**Example.** Draw five cards from a full deck of 52 cards. What is the probability of no Ace?

Let $A_i$ be event that the $i$th draw is no ace, $i = 1, 2, 3, 4, 5$. We are interested in $A := A_1 \cdots A_5$. Using the chain rule, we see that

$$
\mathbb{P}(A) = \frac{48}{52} \times \frac{47}{51} \times \frac{46}{50} \times \frac{45}{49} \times \frac{44}{48} \approx 0.66. 
$$

**Exercise.** (Birthday problem) What is the probability that no one in a randomly selected group of $n < 365$ persons shares a birthday with someone else?

---

**Law of Total Probability and Bayes’s rule**

Suppose $B_1, B_2, \ldots, B_n$ is a partition of $S$.

![Partition Diagram]

Then, by the third Axiom,

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

Hence,

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

This is called the **Law (or Theorem) of Total Probability**.
Combining the Law of Total Probability with the definition of conditional probability gives the famous **Bayes’ Rule**:

\[
P(B_j | A) = \frac{P(A | B_j) P(B_j)}{\sum_{i=1}^{n} P(A | B_i) P(B_i)}
\]

**Example.** A company has three factories (1, 2 and 3) that produce the same chip, each producing 15%, 35% and 50% of the total production. The probability of a defective chip at 1, 2, 3 is 0.01, 0.05, 0.02, resp. Suppose someone shows us a defective chip. What is the probability that this chip comes from factory 1?

Let \( B_i \) denote the event that the chip is produced by \( i \). The \( B_i \)'s form a partition of \( S \). Let \( A \) denote the event that the chip is faulty.

By Bayes’ rule,

\[
P(B_1 | A) = \frac{0.15 \times 0.01}{0.15 \times 0.01 + 0.35 \times 0.05 + 0.5 \times 0.02} = 0.052 .
\]

**Independence**

We say \( A \) and \( B \) are **independent** if the knowledge that \( A \) has occurred does not change the probability that \( B \) occurs. That is

\[
A, B \text{ independent } \iff P(A | B) = P(A)
\]

Since \( P(A | B) = P(A \cap B)/P(B) \) an alternative definition is

\[
A, B \text{ independent } \iff P(A \cap B) = P(A)P(B)
\]

This definition covers the case \( B = \emptyset \), which is always independent of every event.

We can extend this to arbitrary many events:

The events \( A_1, A_2, \ldots \), are said to be **independent** if for any \( n \) and any choice of distinct indices \( i_1, \ldots, i_k \),

\[
P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}) .
\]
Exercise. We randomly select a point in the unit square. Show that the events \( \{(x, y) : x + y \leq 1\} \) and \( \{(x, y) : x - y \leq 0\} \) are independent under the “uniform” probability measure discussed before.

Remark. In most cases independence of events is a model assumption. That is, we assume that there exists a \( \mathbb{P} \) such that certain events are independent. We will see an example of this in the next lecture.

5 Bernoulli Trials

A Bernoulli trial (or Bernoulli experiment) involves performing an experiment once and noting whether a particular event occurs ("success") or not ("failure"). The probability of success is \( p \) \((0 \leq p \leq 1)\).

Often we have a sequence of independent Bernoulli trials. That is, we sequentially perform Bernoulli experiments, such that the outcome (success or failure) of each experiment does not depend on the other experiments.

Examples.

1. We flip a coin a number of times.

2. We randomly select \( n \) people from a large population and ask if they vote for a certain political party.

Here is a way to depict the outcomes:
Let the sample space for each trial be \{0, 1\}. (1 for a success, 0 for a failure).

The sample space \( S \) of a sequence of \( n \) trials is
\[
S = \{(0, \ldots, 0), \ldots, (1, \ldots, 1)\}.
\]

How can we specify \( \mathbb{P} \)?

Let \( A_i \) denote the event of “success” during the \( i \)th trial.

By definition
\[
\mathbb{P}(A_i) = p, \quad i = 1, 2, \ldots.
\]

Moreover, \( \mathbb{P} \) must be such that \( A_1, A_2, \ldots \) are independent.

These two rules completely specify \( \mathbb{P} \).

**Exercise.** Since \( A_i \) is an event, it must be a subset of \( S \). Which one?

**Exercise.** If we perform \( n \) trials, then \( S \) is a discrete sample space. Identify the probabilities of the elementary sets. What changes when we continue the trials indefinitely?
Binomial Law

Let $p_n(k)$ be the probability of $k$ successes in $n$ trials, then

$$p_n(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n,$$

where

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

The probabilities form a so-called binomial distribution (or law).

Exercise. Prove this.

Example. In a large country 51% favours party A and 49% favours party B. We randomly select 200 people from this population. What is the probability that of this group more people vote for B than for A.

Let a vote for A be a “success”. Selecting the 200 people is equivalent to performing a sequence of 200 independent Bernoulli Trials with success probability 0.51.

We are looking for the probability that we have less than 100 successes; which is (use computer)

$$\sum_{k=0}^{99} \binom{200}{k} (0.51)^k (0.49)^{200-k} \approx 0.36.$$
Geometric Law

Let \( p(m) \) be the probability that we require \( m \) trials until a success is reached, then

\[
p(m) = (1 - p)^{m-1}p, \quad m = 0, 1, \ldots, n.
\]

The probabilities form a so-called geometric distribution (or law).

Exercise. Prove this.

Example. In the game Ludo you have to throw a six before you can put your token onto the board. What is the probability that you need more than 6 throws of the die before this happens?

We are again dealing with a sequence of independent Bernoulli trials, with success probability \( 1/6 \) of throwing a 6. Hence the required probability is

\[
\sum_{m=7}^{\infty} \frac{1}{6} \left( \frac{5}{6} \right)^{m-1} = \left( \frac{5}{6} \right)^6 \frac{1}{6} \sum_{k=0}^{\infty} \left( \frac{5}{6} \right)^k
\]

\[
= \left( \frac{5}{6} \right)^6 \frac{1}{6} \left( \frac{1}{1 - 5/6} \right)
\]

\[
= \left( \frac{5}{6} \right)^6 \approx 0.33.
\]
Multinomial Law

The Binomial Law can be generalized to the Multinomial Law in the following way:

Suppose we have a sequence of \( n \) independent trials with possible outcomes \( \{1, 2, \ldots, m\} \), instead of just two (failure and success).
Let \( p_1, \ldots, p_m \) be the corresponding probabilities.

Suppose \( k_1 + k_2 + \cdots + k_m = n \).

The probability of \( k_1 \) occurrences of 1, \( k_2 \) occurrences of 2, \ldots, \( k_m \) occurrences of \( m \), in \( n \) trials is

\[
p_n(k_1, \ldots, k_m) := \frac{n!}{k_1! k_2! \cdots k_m!} p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}.
\]

The probabilities form a so-called multinomial distribution (or law).

6 Random Variables

The outcome of a random experiment is often expressed as a number or measurement.

Definition. A function \( X \) assigning to every outcome \( s \in S \) a real number is called a random variable.

Events such as \( \{ s \in S : X(s) \leq x \} \) and \( \{ s \in S : X(s) = x \} \), for some real number \( x \), are abbreviated to \( \{ X \leq x \} \) and \( \{ X = x \} \), respectively.

The corresponding probabilities of these events, \( \mathbb{P}(\{ X \leq x \}) \) and \( \mathbb{P}(\{ X = x \}) \), are abbreviated further to \( \mathbb{P}(X \leq x) \) and \( \mathbb{P}(X = x) \), respectively.
Example. We throw a coin \( n \) times. The sample space is \( S = \{0, 1\}^n \), i.e. sequences of length \( n \) of 0’s (failures) and 1’s (successes).

Consider the function \( X : S \rightarrow \{0, \ldots, n\} \) which maps \( s = (s_1, \ldots, s_n) \) to
\[
X(s) := s_1 + s_2 + \cdots + s_n .
\]

\( X \) is a random variable. The set \( \{X = k\} \) corresponds to the set of outcomes with exactly \( k \) successes. Hence, we can interpret \( X \) as the total number of successes in \( n \) Bernoulli trials.

If we have independent Bernoulli trials with success parameter \( p \), then
\[
P(X = k) = P(\{X = k\}) = P(\{s \in S : X(s) = k\}) = \binom{n}{k} p^k (1 - p)^{n-k} .
\]

Important remarks.

- Random variables are usually the most convenient way to describe random experiments; they allow us to use intuitive notations for certain events, such as \( \{X > 1000\} \), \( \{\max(X, Y) \leq Z\} \), etc.

- Although mathematically a random variable is neither random nor a variable (it is a function), in practice we may interpret a random variable as the measurement on a random experiment which we will carry out “tomorrow”. However, all the thinking about the experiment is done “today”.

- We denote random variables by upper case Roman letters, \( X, Y, \ldots \).

- Numbers we get when we make the measurement (the outcomes of the random variables) are denoted by the lower case letter, such as \( x_1, x_2, x_3 \) for three values for \( X \).
Range

The set of all possible values a random variable, \( X \), can take is called the range of \( X \), often denoted by \( S_X \).

**Discrete** random variables can only take isolated values.

For example: a count can only take non-negative integer values.

**Continuous** random variables can take values in an interval.

For example: rainfall measurements, lifetimes of components, lengths, . . . are (at least in principle) continuous.

If we know, or can find the probabilities for all events defined by a random variable \( X \), we know the (probability) distribution of \( X \).

---

Probability mass function

**Definition.** For a *discrete* random variable \( X \), the function \( x \mapsto \mathbb{P}(X = x) \) is called the **probability mass function** (pmf) of \( X \).

We have for any \( B \subset S_X \),

\[
\mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x).
\]

**Example:** Toss a die and let \( X \) be its face value. \( X \) is discrete with range \( \{1, 2, 3, 4, 5, 6\} \). If the die is fair the probability mass function is given by

\[
\begin{array}{cccccc|c}
  x & 1 & 2 & 3 & 4 & 5 & 6 & \sum \\
\hline
  \mathbb{P}(X = x) & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1 \\
\end{array}
\]
Example: Toss two dice and let $M$ be the largest face value showing. The distribution of $M$ can be found to be

<table>
<thead>
<tr>
<th>$m$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>$\sum$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{P}(M = m)$</td>
<td>$\frac{1}{36}$</td>
<td>$\frac{3}{36}$</td>
<td>$\frac{5}{36}$</td>
<td>$\frac{7}{36}$</td>
<td>$\frac{9}{36}$</td>
<td>$\frac{11}{36}$</td>
<td>1</td>
</tr>
</tbody>
</table>

or, as a formula:

$$\mathbb{P}(M = m) = \frac{2m - 1}{36}, \text{ for } m = 1, 2, \ldots, 6.$$  

We can now work out the probability of any event defined by $M$ so we know the distribution of $M$.

For example:

$$\mathbb{P}(M > 4) = \mathbb{P}(M = 5) + \mathbb{P}(M = 6) = \frac{9 + 11}{36} = \frac{5}{9}.$$  

Probability Density Function

For a continuous random variable $X$ the probability $\mathbb{P}(X = x)$ is always 0. Hence, we cannot characterize the distribution of $X$ via the probability mass function.

Instead, we have:

**Definition.** We say that a continuous random variable $X$ has a **probability density function** (pdf) $f$ if for all $a, b$

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x) \, dx.$$  

----

© STAT2202

43

© STAT2202

44
• $f(x)$ can be interpreted as the “infinitesimal” probability that $X = x$. More precisely,

$$
\mathbb{P}(x \leq X \leq x + h) = \int_{x}^{x+h} f(u) \, du \approx h \, f(x).
$$

Properties of $f$:

1. $f(x) \geq 0$, for all $x$.
2. $\int_{-\infty}^{\infty} f(x) \, dx = 1$.

- Any function $f$ satisfying these conditions, and for which $\int_{a}^{b} f(x) \, dx$ is well-defined, can be a pdf.

**Remark.** Not all continuous random variables have pdf’s. (Cantor example)

**Cumulative Distribution Function**

Although we will usually work with pmf’s for discrete and pdf’s for continuous random variables, the following function is defined for both continuous and discrete random variables.

**Definition.** The cumulative distribution function (cdf) of $X$ is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$
F(x) = \mathbb{P}(X \leq x).
$$
Properties of $F$:

1. $0 \leq F(x) \leq 1$.
2. $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$.
3. $F$ is non-decreasing:
   If $x < y$, then $F(x) \leq F(y)$.
4. $F$ is right-continuous: If $x_n \downarrow x$, then $\lim_{n \to \infty} F(x_n) = F(x)$.

- For a discrete random variable $X$ the cdf $F$ is a step function with jumps of size $\mathbb{P}(X = x)$ at all the points $x \in S_X$.

- For a continuous random variable $X$ with pdf $f$, the cdf $F$ is continuous, and satisfies

$$F(x) = \int_{-\infty}^{x} f(u) \, du.$$ 

Consequently, $f(x) = \frac{dF(x)}{dx}$.

### 7 Important Discrete Distributions

A random variable is said the have a discrete distribution if $S_X$ is countable, and for any subset $B \subset S_X$,

$$\mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x).$$

Think of $X$ as the measurement of a random experiment that will be carried out tomorrow.

However, all the “thinking” is done today. The behaviour of the experiment is summarized by the probability mass function.
Bernoulli distribution

We say that a random variable $X$ has a **Bernoulli distribution** with success parameter $p$ if $S_X = \{0, 1\}$, and

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p.$$  

A Bernoulli random variable describes the outcome of a *Bernoulli trial*.

Binomial distribution

We say that a random variable $X$ has a **Binomial distribution** with parameters $n$ and $p$ if

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n.$$  

We write $X \sim \text{Bin}(n, p)$.
A Binomial random variable is used to describe the total number of successes in a sequence of \( n \) independent Bernoulli trials with success probability \( p \).

**Example:** A population of 12 items contains 5 defectives.

A sample of size 4 is taken with replacement. Let \( X \) be the number of defectives. Then \( X \sim \text{Bin}(4, 5/12) \).

\[
\mathbb{P}(X = k) = \binom{4}{k} \left(\frac{5}{12}\right)^k \left(\frac{7}{12}\right)^{4-k}, \quad k = 0, 1, \ldots, 4
\]

<table>
<thead>
<tr>
<th>( k )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>( \sum )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbb{P}(X = k) )</td>
<td>.116</td>
<td>.331</td>
<td>.354</td>
<td>.169</td>
<td>.030</td>
<td>1</td>
</tr>
</tbody>
</table>

**Geometric distribution**

We say that a random variable \( X \) has a **Geometric distribution** with parameter \( p \) if

\[
\mathbb{P}(X = k) = p(1 - p)^{k-1}, \quad k = 1, 2, \ldots
\]

We write \( X \sim G(p) \).

![Geometric distribution graph](image)
A Geometric random variable is used to describe the “time” of first success in a sequence of independent Bernoulli trials with success probability $p$.

**Example:**

Let $X$ be the number of times that you have to throw a (fair) die to get a six.

Then $X \sim G(1/6)$.

---

**Poisson distribution**

We say that a random variable $X$ has a Poisson distribution with parameter $\lambda$ if

$$
\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2 \ldots
$$

We write $X \sim \text{Po}(\lambda)$. 

---

[Diagram of Geometric distribution with $p=0.8$]

[Histogram of Poisson distribution with $\lambda = 10$]
A Poisson distribution is the limit of Binomial distributions in the following sense:

Let $X_n \sim \text{Bin}(n, \lambda/n)$ and $X \sim \text{Po}(\lambda)$, then

$$\lim_{n \to \infty} \Pr(X_n = k) = \Pr(X = k),$$

for all $k$.

The Poisson distribution is used in many probability models in which there are many possible events which may occur, each having a very small probability.

Examples:

- Deaths per year due to a horse kick in the Prussian cavalry.
- Number of telephone calls arriving at a telephone exchange in a given time.
- Number of accidents on a stretch of road per unit time.
- Number of surviving carp in a given area.
The Hypergeometric distribution

We say that a random variable $X$ has a Hypergeometric distribution with parameters $N$, $n$ and $r$ if

$$\mathbb{P}(X = k) = \frac{\binom{r}{k} \binom{N-r}{n-k}}{\binom{N}{n}},$$

for $\max\{0, r + n - N\} \leq k \leq \min\{n, r\}$.

We write $X \sim \text{Hyp}(n, r, N)$.

The hypergeometric distribution is used in the following situation.

Consider an urn with $N$ balls, $r$ of which are red. We draw at random $n$ balls from the urn without replacement.

The number of red balls amongst the $n$ chosen balls has a $\text{Hyp}(n, r, N)$ distribution.

In table form:

<table>
<thead>
<tr>
<th></th>
<th>Red</th>
<th>Not Red</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected</td>
<td>$k$</td>
<td>$n-k$</td>
<td>$n$</td>
</tr>
<tr>
<td>Not Selected</td>
<td>$r-k$</td>
<td>$N-n-r+k$</td>
<td>$N-n$</td>
</tr>
<tr>
<td>Total</td>
<td>$r$</td>
<td>$N-r$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Example: Five cards are selected from a full deck of 52 cards. Let $X$ be the number of Aces. Then $X \sim \text{Hyp}(n = 5, r = 4, N = 52)$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>$\sum$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{P}(X = k)$</td>
<td>0.659</td>
<td>0.299</td>
<td>0.040</td>
<td>0.002</td>
<td>0.000</td>
<td>1</td>
</tr>
</tbody>
</table>
8 Important Continuous Distributions

A random variable is said to have a continuous distribution with probability density function (pdf) $f$ if for all $a, b$

$$P(a \leq X \leq b) = \int_{a}^{b} f(u) \, du.$$  

The corresponding cumulative distribution function (cdf) $F$ is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u) \, du.$$  

Think of $X$ as the result of a random experiment that will be carried out tomorrow. However, all the “thinking” is done today. The behaviour of the experiment is summarized by $f$ or $F$.

Uniform distribution

We say that a random variable $X$ has a uniform distribution on the interval $[a, b]$, if it has density function $f$, given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim U[a, b]$. 

---

© STAT2202 59

© STAT2202 60
Examples

- Draw “at random” a number in an interval. Each number is equally likely.

- Spin-a-dial

- Round-off errors

- (Pseudo-) random number generator

Exercise

Let $X$ have a uniform distribution on $[0,1]$. 

1. Determine $\mathbb{P}(-2 \leq X \leq 1/2)$.

2. What is the cdf of $X$?
Exponential distribution

We say that a random variable $X$ has an exponential distribution with parameter $\lambda$, if it has density function $f$, given by

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

We write $X \sim \text{Exp}(\lambda)$.

Examples

- Lifetime of an electrical component
- Time between arrivals of calls at a telephone exchange
- Time elapsed until a Geiger counter registers a radio-active particle.

Properties

- The cdf of $X$ is given by
  $$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$
- $\mathbb{P}(X > x) = e^{-\lambda x}, \ x \geq 0.$
- The exponential distribution is memoryless: for any $s, t > 0$
  $$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t).$$
• It is the only memoryless continuous distribution.

• It may be viewed as a continuous version of the geometric distribution.
  Compare: Number of tosses with a die until a 6 is thrown.

**Exercise.** Prove the memoryless property.

---

**Gamma distribution**

We say that a random variable $X$ has a **gamma distribution** with **shape** parameter $\alpha$ and **scale** parameter $\lambda$ if it has pdf $f$, given by

\[
 f(x) = \begin{cases} 
 \frac{\lambda^\alpha x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}, & x \geq 0 \\
 0, & x < 0 
\end{cases}
\]

We write $X \sim \text{Gam}(\alpha, \lambda)$.
Note: $\Gamma$ is the Gamma-function defined as

$$\Gamma(\alpha) = \int_{0}^{\infty} u^{\alpha-1} e^{-u} \, du, \quad \alpha > 0.$$ 

- Notice that the exponential distribution is a member of this family.

- A $\text{Gam}(n/2, 1/2)$-distribution (with $n = 1, 2, \ldots$) is called a \textbf{chi-square}-distribution with $n$ “degrees of freedom”. We write $X \sim \chi^2_n$.

- A $\text{Gam}(n, c)$-distribution (with $n = 1, 2, \ldots$) is called an \textbf{Erlang}-distribution (with parameters $n$ and $c$).

We finally mention a few properties of the $\Gamma$-function.

1. $\Gamma(a + 1) = a \Gamma(a)$, for $a \in \mathbb{R}_+$.
2. $\Gamma(n) = (n - 1)!$ for $n = 1, 2, \ldots$.
3. $\Gamma(1/2) = \sqrt{\pi}$.

\textbf{Normal, or Gaussian, distribution}

We say that a random variable $X$ has an \textbf{normal or Gaussian distribution} with mean $\mu$ and variance $\sigma^2$ if it has density function $f$, given by

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$
We write \( \mathsf{X} \sim \mathcal{N}(\mu, \sigma^2) \).

If \( \mu = 0 \) and \( \sigma = 1 \) then
\[
f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},
\]
and the distribution is known as a **standard normal distribution**.

The normal distribution is the most important distribution in the study of Statistics, whereas the exponential distribution is the main distribution in Applied Probability.

- Many textbooks have tables for the cdf of the standard normal distribution. Engineers often tabulate
\[
Q(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2} u^2} du,
\]
for various values of \( x \geq 0 \). Hence, if \( \mathsf{X} \sim \mathcal{N}(0, 1) \), then
\[
Q(x) = \mathbb{P}(\mathsf{X} > x).
\]

### 9 Functions of a RV

Let \( \mathsf{X} \) be a random variable and let \( g: \mathbb{R} \to \mathbb{R} \) be a function.

Then \( \mathsf{Y} := g(\mathsf{X}) \) is again a random variable.

**Examples.**

1. Let \( \mathsf{X} \) be the voltage output to a half-wave rectifier, then \( \mathsf{Y} = |\mathsf{X}| \) is the output.

2. A voice transmission system can transmit up to \( m \) voice signals. Let \( \mathsf{X} \) be the number of active sources, then the number of transmitted sources is \( \min(\mathsf{X}, m) \), and the number of discarded voices \( \mathsf{Y} \) is therefore \( \max(\mathsf{X} - m, 0) \).

Suppose the sources work “independent” of each other and are active with probability \( p \). Then a reasonable model is to assume that \( \mathsf{X} \sim \text{Bin}(n, p) \).
What is the pmf of $Y$?

$Y$ can take values in $\{0, 1, \ldots, n-m\}$. For $k = 1, \ldots, n-m$ we have

$$P(Y = k) = P(X = m + k) = p_n(m + k),$$

where by the Binomial formula

$$p_n(m + k) = \binom{n}{m + k} p^{m+k}(1-p)^{n-m-k}.$$

For $k = 0$, we have

$$P(Y = 0) = P(X \leq m) = \sum_{i=0}^{m} p_n(i).$$

3. A radio-active source is put at a distance of 1 (unit) in front of a screen. Let $Y$ be the position on the screen of an emitted particle.

We can express $Y$ as a function of the angle $X \in (-\pi/2, \pi/2)$.

Namely,

$$Y = \tan X.$$
We “know” that $X$ has a Uniform distribution on $(-\pi/2, \pi/2)$.

We first express the cdf of $Y$, say $F_Y$, in terms of the cdf of $X$, say $F_X$.

We have for each fixed position $y$

\[
F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\tan X \leq y) = \mathbb{P}(X \leq \arctan y) = \frac{1}{\pi} \arctan y + \frac{1}{2}.
\]

We get the pdf of $Y$, say $f_Y$, by differentiating $F_Y$. Thus,

\[
f_Y(y) = \frac{d}{dy} \frac{1}{\pi} \arctan y = \frac{1}{\pi(1 + y^2)}, \ y \in \mathbb{R}.
\]

This is the pdf of the so-called Cauchy distribution.

**Affine transformation**

It often happens that

\[
Y = aX + b.
\]

This is called an affine transformation. If $a > 0$, then

\[
F_Y(y) = F_X \left( \frac{y - b}{a} \right).
\]

Hence, if $X$ has a density $f_X$, then

\[
f_Y(y) = \frac{1}{a} f_X \left( \frac{y - b}{a} \right).
\]

If $a < 0$, then, (check this yourself!)

\[
f_Y(y) = -\frac{1}{a} f_X \left( \frac{y - b}{a} \right).
\]

Hence, for general $a$, when $f_X$ exists

\[
f_Y(y) = \frac{1}{|a|} f_X \left( \frac{y - b}{a} \right). \quad (2)
\]
Example. Let $X \sim N(0, 1)$. Thus the pdf $f_X$ is given by

\[ f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad x \in \mathbb{R}. \]

Now consider the random variable

\[ Y = aX + b, \]

with $a > 0$. Using (2) above, we have

\[ f_Y(y) = f_X \left( \frac{y - b}{a} \right) \frac{1}{a} = \frac{1}{a \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{y - b}{a} \right)^2}, \quad y \in \mathbb{R}. \]

This is exactly the pdf of the $N(b, a^2)$-distribution. We thus have shown:

A r.v. $Y$ has an $N(\mu, \sigma^2)$-distribution if and only if

\[ Y = \mu + \sigma X \]

for some $X \sim N(0, 1)$.

This is the way to think of any Gaussian r.v.: it is a simple (affine) transformation of a standard normal Gaussian r.v.

Exercise. Let $X \sim N(0, 1)$. Prove that $Y = X^2$ has a $\chi^2_1$ distribution.

Exercise. Let $Y = \cos X$, where $X \sim U(0, 2\pi)$. $Y$ can be viewed as the sample of a sinusoidal waveform at a “random instant of time”.

Find the pdf of $Y$. 

© STAT2202 75

© STAT2202 76
10 Expectation

Definition. Let $X$ be a discrete random variable. The expected value (or mean value of $X$), denoted by $\mathbb{E}X$, is defined by

$$\mathbb{E}X = \sum_x x \mathbb{P}(X = x).$$

This number, sometimes written as $\mu_X$, is an indication for the “mean” of the distribution.

Example. Find $\mathbb{E}X$ if $X$ is the outcome of a toss of a fair die.

Since $\mathbb{P}(X = 1) = \ldots \mathbb{P}(X = 6) = 1/6$

$$\mathbb{E}X = 1 \left(\frac{1}{6}\right) + 2 \left(\frac{1}{6}\right) + \ldots + 6 \left(\frac{1}{6}\right) = \frac{7}{2}.$$

Note: $\mathbb{E}X$ is not necessarily a possible outcome of the random experiment as in the previous example.

Definition. If $X$ is a discrete random variable, then for any real-valued function $g$

$$\mathbb{E}g(X) = \sum_x g(x) \mathbb{P}(X = x).$$

Example. Find $\mathbb{E}X^2$ if $X$ is the outcome of the toss of a fair die. We have

$$\mathbb{E}X^2 = 1^2 \frac{1}{6} + 2^2 \frac{1}{6} + 3^2 \frac{1}{6} + \ldots + 6^2 \frac{1}{6} = \frac{91}{6}.$$

Note: $\frac{91}{6} = \mathbb{E}X^2 \neq (\mathbb{E}X)^2 = \left(\frac{7}{2}\right)^2 = \frac{49}{4}$.

By simply replacing the probability mass function with the probability density function and the summation with an integration we find the expectation of a (function of a) continuous random variable. In particular if the continuous variable has density function $f$, then

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x) f(x) \, dx.$$
Definition. The variance of a random variable $X$, denoted by $\text{Var}(X)$ is defined by

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2.$$ 

This number, sometimes written as $\sigma^2_X$, measures the spread or dispersion of the distribution of $X$.

It may be regarded as a measure of the consistency of outcome, a smaller value of $\text{Var}(X)$ implies that $X$ is more often near $\mathbb{E}X$ than for a larger value of $\text{Var}(X)$.

The square root of the variance is called the standard deviation.

It is sometimes also useful to know $\mathbb{E}X^r$, which is called the $r$th moment of $X$.

Remark. The expectation, or moment of a random variable need not always exist or can be $\pm \infty$.

We finally list some properties for expectations and variances involving discrete or continuous random variables:

<table>
<thead>
<tr>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{E}(aX + b) = a\mathbb{E}X + b$</td>
</tr>
<tr>
<td>$\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$</td>
</tr>
<tr>
<td>$\text{Var}(aX + b) = a^2 \text{Var}(X)$</td>
</tr>
</tbody>
</table>

Exercise. Prove this, using the definitions of expectation and variance.
Discrete Distributions

We give the expectation and variance for some known discrete distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean $\mu$</th>
<th>Variance $\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin$(n, p)$</td>
<td>$np$</td>
<td>$np(1 - p)$</td>
</tr>
<tr>
<td>G$(p)$</td>
<td>$\frac{1}{p}$</td>
<td>$\frac{1 - p}{p^2}$</td>
</tr>
<tr>
<td>Po$(\lambda)$</td>
<td>$\lambda$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td>Hyp$(n, pN, N)$</td>
<td>$np$</td>
<td>$np(1 - p) \frac{N - n}{N - 1}$</td>
</tr>
</tbody>
</table>

Exercise. What are the expectation and variance for a Bernoulli random variable with success probability $p$?

Continuous Distributions

Here are the expectations and variances for some known continuous distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean $\mu$</th>
<th>Variance $\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>U$(a, b)$</td>
<td>$\frac{a + b}{2}$</td>
<td>$\frac{(b - a)^2}{12}$</td>
</tr>
<tr>
<td>Exp$(\lambda)$</td>
<td>$\frac{1}{\lambda}$</td>
<td>$\frac{1}{\lambda^2}$</td>
</tr>
<tr>
<td>Gam$(\alpha, \lambda)$</td>
<td>$\frac{\alpha}{\lambda}$</td>
<td>$\frac{\alpha}{\lambda^2}$</td>
</tr>
<tr>
<td>N$(\mu, \sigma^2)$</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>

Exercise. Prove this.
Let $X$ and $Y = g(X)$, be continuous r.v.'s. Note that we can calculate $\mathbb{E}Y$ in two ways:

$$\mathbb{E}Y = \int_{-\infty}^{\infty} y f_Y(y) \, dy ,$$

and

$$\mathbb{E}Y = \mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx .$$

It depends on the situation which one is the easiest to use.

**Example.** Let $Y = a \cos(\omega t + X)$, be the value of a sinusoid signal at time $t$ with random phase $X \sim U(0, 2\pi]$. The expected value of the signal at time $t$ is

$$\mathbb{E}Y = \mathbb{E}a \cos(\omega t + X) = \frac{1}{2\pi} \int_{0}^{2\pi} a \cos(\omega t + x) \, dx$$

$$= \frac{a}{2\pi} \sin(\omega t + x) \bigg|_{0}^{2\pi} = 0 ,$$

as is to be expected. More important is the average power of the signal, i.e. $\mathbb{E}Y^2$. We have,

$$\mathbb{E}Y^2 = a^2 \mathbb{E} \cos^2(\omega t + X)$$

$$= \frac{a^2}{2} \mathbb{E}[1 + \cos(2\omega t + 2X)]$$

$$= \frac{a^2}{2} + \frac{a^2}{4\pi} \int_{0}^{2\pi} \cos(2\omega t + 2x) \, dx$$

$$= \frac{a^2}{2} .$$
11 Bounds and Transforms

Bounds

The mean and the variance do not give us (in general) enough information to determine the distribution of a random variable. However, they may provide us with bounds. We discuss two such bounds.

Suppose $X$ can only take non-negative values and has pdf $f$. For any $x > 0$, we can write

$$
\mathbb{E}X = \int_0^x tf(t) \, dt + \int_x^\infty tf(t) \, dt \\
\geq \int_x^\infty tf(t) \, dt \geq \int_x^\infty xf(t) \, dt \\
= x \mathbb{P}(X \geq x).
$$

We thus have proved the Markov inequality: Suppose $X \geq 0$, then for all $x > 0$,

$$
\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}X}{x}.
$$

If we also know the variance of a random variables, we can give a “tighter” bound.

Chebyshev Inequality: For any random variable $X$ with mean $\mu$ and variance $\sigma^2$, we have

$$
\mathbb{P}(|X - \mu| \geq x) \leq \frac{\sigma^2}{x^2}.
$$

Proof. Let $D^2 = (X - \mu)^2$, then by the Markov inequality and the definition of the variance

$$
\mathbb{P}(D^2 \geq x^2) \leq \frac{\sigma^2}{x^2},
$$

where the event \{ $D^2 \geq x^2$ \} is equivalent to the event \{ $|X - \mu| \geq x$ \}. 

Example. The lifetime of a certain machine component is on average 10 days, with a standard deviation of less than 20 days. How likely is the machine to survive 100 days?

Let $X$ be the lifetime. Since we know nothing about the pdf of $X$, we cannot calculate $\Pr(X \geq 100)$. However, irrespective of this pdf, the Markov inequality gives us the bound:

$$\Pr(X \geq 100) \leq \frac{10}{100} = 0.1.$$  

And, using the Chebyshev inequality, we have

$$\Pr(X \geq 100) \leq \frac{20^2}{90^2} \approx 0.05.$$  

Transforms

Many calculations and manipulations involving probability distributions are facilitated by the use of transforms. We discuss a number of these transforms.

Probability Generating Function

Let $N$ be a non-negative and integer-valued random variable.

The probability generating function (PGF) of $N$ is the function $G : [0, 1] \to [0, 1]$ defined by

$$G(z) := \mathbb{E} z^N = \sum_{k=0}^{\infty} z^k \Pr(N = k).$$
**Example.** Let $N \sim \text{Po}(\mu)$, then the PGF of $N$ is given by

$$G(z) = \sum_{k=0}^{\infty} z^k \frac{e^{-\mu} \mu^k}{k!}$$

$$= e^{-\mu} \sum_{k=0}^{\infty} \frac{(z\mu)^k}{k!}$$

$$= e^{-\mu} e^{z\mu} = e^{-\mu(1-z)} .$$

Knowing only the PGF of $N$, we can easily obtain the pmf:

$$\mathbb{P}(N = k) = \frac{1}{k!} \frac{d^k}{dz^k} G(z) \bigg|_{z=0} .$$

Thus we have the **uniqueness** property: two pmf’s are the same if and only if their pgf’s are the same.

Another useful property of the PGF is that we can obtain the moments of $N$ by differentiating $G$ and evaluating it at $z = 1$.

Differentiating $G(z)$ w.r.t. $Z$ gives

$$G'(z) = \frac{d}{dz} \mathbb{E} z^N = \mathbb{E} z^N \mathbb{E} z^{N-1} .$$

$$G''(z) = \frac{d^2}{dz^2} \mathbb{E} N z^{N-1} = \mathbb{E} N (N-1) z^{N-2} .$$

$$G'''(z) = \mathbb{E} N (N-1)(N-2) z^{N-3} .$$

Etcetera.

In particular,

$$\mathbb{E} N = G'(1) ,$$

and

$$\text{Var}(N) = G'''(1) + G'(1) - (G'(1))^2 .$$

**Exercise.** Evaluate the mean and variance of a $\text{Po}(\mu)$ random variable via its PGF.
Laplace Transform

Let $X$ be a non-negative random variable with distribution function $F$. The Laplace transform of $X$ is the function $L : \mathbb{R}_+ \to [0, 1]$ defined by

$$L(s) := \mathbb{E} e^{-sX}.$$ 

When $X$ has a continuous distribution with density function $f$, the Laplace transform becomes

$$L(s) = \int_0^\infty e^{-sx} f(x) \, dx,$$

In other words, the Laplace transform of $X$ is the “ordinary” Laplace transform of the pdf $f$.

For a discrete random variable $X$, we have

$$L(s) = \sum_x e^{-sx} \mathbb{P}(X = x).$$

As for the PGF, Laplace transforms have the uniqueness property: Two Laplace transforms are the same if and only if their corresponding distribution functions are the same.

Similar to the PGF, the moments of $X$ follow from the derivatives of $L$:

If $\mathbb{E}X^n$ exists, then $L$ is $n$ times differentiable, and

$$\mathbb{E}X^n = (-1)^n L^{(n)}(0).$$

**Example.** The Laplace transform of $X \sim \text{Gam}(\alpha, \lambda)$ is given by

$$\mathbb{E} e^{-sX} = \int_0^\infty \frac{e^{-\lambda x} \lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-sx} \, dx$$

$$= \left( \frac{\lambda}{\lambda + s} \right)^\alpha \int_0^\infty \frac{e^{-(\lambda+s)x} (\lambda + s)^\alpha x^{\alpha-1}}{\Gamma(\alpha)} \, dx$$

$$= \left( \frac{\lambda}{\lambda + s} \right)^\alpha.$$
Characteristic Function

For a random variable $X$ that can also take negative values, the Laplace transform is ill-defined.

However, we can instead define the Fourier transform or characteristic function of $X$, by

$$
\Phi(\omega) := \mathbb{E} e^{j\omega X}, \ \omega \in \mathbb{R},
$$

where $j = \sqrt{-1}$.

When $X$ has pdf $f$, we have

$$
\Phi(\omega) = \int_{-\infty}^{\infty} e^{j\omega x} f(x) \, dx .
$$

In other words, $\Phi$ is the Fourier transform of the function $f$. In this case, $f$ can be found from $\Phi$ by the inversion formula:

$$
f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi(\omega) e^{-j\omega x} \, d\omega .
$$

Example.

The Fourier transform has the same kind of properties as the Laplace transform. The main difference is that we now have to work in the complex plane.

- The moment theorem states that

$$
\mathbb{E} X^n = \frac{1}{j^n} \Phi^{(n)}(0) .
$$

- The uniqueness theorem states that two distributions are the same if the characteristic functions are the same.

Example. The characteristic function of an $\text{Exp}(\lambda)$ random variable is given by

$$
\Phi(\omega) = \int_{0}^{\infty} \lambda e^{-\lambda x} e^{j\omega x} \, dx
$$

$$
= \frac{\lambda}{\lambda - j\omega} .
$$

Finally, we mention the following important result.
The characteristic function of a \( N(0, 1) \) random variable is given by \( \Phi(\omega) = e^{-\frac{1}{2} \omega^2} \).

**Exercise.** Derive from this the CF of a \( N(\mu, \sigma^2) \) random variable.

**Remark.** The transforms discussed here are particularly useful when dealing with sums of independent random variables (to be discussed later).

**Remark.** Simulation provides a way to quickly perform random experiments and to gain “insight” into the random mechanism.

The starting point of any simulation is to generate “random numbers”.

One way to generate (pseudo) random numbers is to use the power residue method, which involves the recursion:

\[ z_i = \alpha z_{i-1} \mod M, \]

where \( \alpha \) is an integer and \( M \) is a large integer.
where \( \alpha \) is a carefully chosen integer between 1 and \( M \), and \( M \) is a prime number. The resulting numbers are in the range 0 to \( M - 1 \).

Particularly “good” choices are: \( \alpha = 7^5 \) and \( M = 2^{31} - 1 \).

If we divide the \( z_1, z_2, \ldots \) by \( M \), we produce numbers between 0 and 1, which may be viewed as (independent) samples from the Uniform distribution on \([0,1]\).

Such a procedure is implemented in any decent calculator or computer as a random number generator.

Here is an example:

\[
\begin{align*}
0.0100569 & \quad 0.287304 & \quad 0.544480 & \quad 0.377250 \\
0.3338210 & \quad 0.709423 & \quad 0.479557 & \quad 0.228095 \\
0.7559040 & \quad 0.130341 & \quad 0.361738 & \quad 0.564670 \\
0.0617347 & \quad 0.408820 & \quad 0.736501 & \quad 0.104207 \\
0.5496780 & \quad 0.398980 & \quad 0.946124 & \quad 0.338100 \\
\end{align*}
\]

**Transformation method**

We have seen how we can generate (pseudo) random numbers from a \( U[0,1] \) distribution.

How can we generate random numbers from another distribution? One approach is given by the following result.

Suppose \( F \) is a cdf with inverse \( F^{-1} \), such that \( F^{-1}(F(x)) = x \). Let \( U \sim U[0,1] \) and define

\[
X := F^{-1}(U).
\]

Then,

\[
\mathbb{P}(X \leq x) = \mathbb{P}(U \leq F(x)) = F(x).
\]

Hence, \( X \) has cdf \( F \).
Thus, we can generate random numbers from the cdf $F$ as follows:

1. Generate $u$ from the uniform random generator.

2. Output $F^{-1}(u)$.

This is called the **transformation method** of generating random numbers from a certain distribution.

**Example.** For the exponential distribution we have

$$F(x) = 1 - e^{-\lambda x}, \ x > 0,$$

so that for $y \in (0, 1)$,

$$F^{-1}(y) = -\frac{1}{\lambda} \log(1 - y).$$

Hence, output $x := -\frac{1}{\lambda} \log(1 - u)$.

We could also output $-\frac{1}{\lambda} \log u$ (why?).

Below a sample of size 30 from an $\text{Exp}(1)$ distribution is plotted.
Let $T \geq 0$ be the lifetime of a component, with $f$ and $F$ being its pdf and cdf, respectively.

1. Generate $u$ from the uniform random generator.
2. Output $x$ for which $F(x) - u < u \leq F(x)$.

Other important quantities are $R(t) = P(T > t)$ and $r(t)$, the failure rate, given by

$$r(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)},$$

for $t \geq 0$, that is,

$$r(t) = \frac{d}{dt} \left[ -\log R(t) \right].$$

Clearly $R(t) = -f(t)$, and so

$$R(t) = \exp \left[ -\int_0^t r(u) \,du \right].$$

Example. To generate a sample from a Bernoulli random variable with success probability $p = 0.3$, we can draw $u$ from the uniform random generator and output 1 if $u \leq 0.3$ and 0 else.
Example. Suppose the component has a power law failure rate: \( r(t) = \alpha \beta t^{\beta - 1} \), where \( \alpha, \beta > 0 \).

\[
R(t) = \exp \left( - \int_0^t r(u) \, du \right) = \exp \left( -\alpha t^\beta \right).
\]

Also, for \( t \geq 0 \),

\[
F(t) = 1 - e^{-\alpha t^\beta} \quad \text{and} \quad f(t) = \alpha \beta t^{\beta - 1} e^{-\alpha t^\beta}.
\]

\( T \) has a Weibull distribution (if \( \beta = 1 \) we get the exponential distribution). It can be shown that the MTTF is \( \mathbb{E}T = \alpha^{-1/\beta} \Gamma(1 + 1/\beta) \).

Example. A system consists of \( n \) independent components in series. Since the system fails if and only if at least one component fails, the reliability function for the system is given by

\[
R(t) = R_1(t)R_2(t) \cdots R_n(t),
\]

where \( R_i \) is the reliability function for component \( i \). If the \( n \) independent components were in parallel, we would have

\[
1 - R(t) = (1 - R_1(t))(1 - R_2(t)) \cdots (1 - R_n(t)),
\]

because now the system fails if and only if all components fail.

14 Pairs of Random Variables

Many random experiments involve not just one but \textbf{multiple} random variables.

Examples.

1. We randomly select a person from a large population and measure his/her weight \( X \) and height \( Y \).

2. We shoot at a two-dimensional target. Let \( X \) and \( Y \) be the coordinates of the point of impact.

3. We randomly select 20 people from a large population and ask their preference for a political party. Number the people from 1 to 20, and let \( X_1, \ldots, X_{20} \) be the measurements.
How can we specify a model for the experiments above?

We cannot just specify the pdf or pmf of the individual random variables.

We also need to specify the “interaction” between the random variables. E.g., in Example 1, if the height $Y$ is large, we expect that $X$ is large as well.

We need to specify the joint distribution of all the random variables $X_1, \ldots, X_n$ involved in the experiment.

Alternatively, we need to specify the distribution of the random vector $X := (X_1, \ldots, X_n)$.

We first show how this works for pairs of random variables.

**Joint pmf**

We will write $\mathbb{P}(X = x, Y = y)$ for the probability of the event $\{X = x\} \cap \{Y = y\}$.

**Definition.** Let $(X, Y)$ be a discrete random vector. The function $(x, y) \mapsto \mathbb{P}(X = x, Y = y)$ is called the joint probability mass function of $X$ and $Y$.

**Example.** In a box are three dice. Die 1 is a normal die; die 2 has no 6 face, but instead two 5 faces; die 3 has no 5 face, but instead two 6 faces.

The experiment consists of selecting a die at random, followed by a toss with that die.

Let $X$ be the die number that is selected, and let $Y$ be the face value of that die. The joint pmf of $X$ and $Y$ is specified below.
The pmf’s of $X$ and $Y$, the so-called marginal pmf’s, can be found by summing up over respectively the $y$’s and the $x$’s, e.g.,

$$
P(X = x) = \sum_y P(X = x, Y = y).$$

Let $S_{X,Y}$ be the set of possible outcomes of $(X,Y)$. We have for all $B \subset S_{X,Y}$,

$$
P((X, Y) \in B) = \sum_{(x,y) \in B} P(X = x, Y = y).$$

An important way of creating joint pmf’s is by starting with the marginal pmf’s of $X$ and $Y$ and then to define the events $\{X = x\}$ and $\{Y = y\}$ to be independent, for all $x$ and $y$.

We then have (definition of independent events)

$$
P(X = x, Y = y) = P(X = x) P(Y = y).$$

**Example.** Repeat the experiment above with three normal dice. Since the events $\{X = x\}$ and $\{Y = y\}$ should be independent, each entry in the pmf table is $\frac{1}{3} \times \frac{1}{6}$.

Clearly in the first experiment not all events $\{X = x\}$ and $\{Y = y\}$ are independent (why not?).
Joint pdf

The analogue of the pmf for continuous \( X \) and \( Y \) is the joint pdf.

**Definition.** We say that the random variables \( X \) and \( Y \) have a joint probability density function \( f \) if for all events \( \{(X, Y) \in A\} \), where \( A \) is a subset of \( \mathbb{R}^2 \) (the plane), we have

\[
P((X, Y) \in A) = \int \int_A f(x, y) \, dx \, dy.
\]

We often write \( f_{X,Y} \) for \( f \).

Note that the calculation of probabilities has been reduced to *integration* over a set \( A \).

**Example.** We select randomly a point in the unit square. Let \( X \) be the x-coordinate and \( Y \) the y-coordinate. The joint pdf of \( X \) and \( Y \) is simply

\[
f(x, y) = \begin{cases} 
1, & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\
0, & \text{otherwise}.
\end{cases}
\]

What is the probability that the point is at least a distance 1 away from the origin? Let \( A = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \geq 1\} \), and let \( S \) be the unit square. Then,

\[
P(X^2 + Y^2 \geq 1) = \int \int_{A \cap S} 1 \, dx \, dy = 1 - \pi/4.
\]
**Example.** We say that $Z_1$ and $Z_2$ have a bivariate Gaussian (or normal) distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ and $\rho$ if the joint density function is given by

$$f(z_1,z_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left( \frac{(z_1-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(z_1-\mu_1)(z_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(z_2-\mu_2)^2}{\sigma_2^2} \right) \right\}.$$  

**Properties of pdf.** (similar to the 1-dimensional case:)

1. $f(x,y) \geq 0$, for all $x$ and $y$.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) \, dx \, dy = 1$.

- Any function $f$ satisfying these conditions, and for which the integral is well-defined, can be a joint pdf.
- $f(x,y)$ can be interpreted as the “infinitesimal” probability that $X = x$ and $Y = y$:
  $$\mathbb{P}(x \leq X \leq x+h, \, y \leq Y \leq y+h) = \int_x^{x+h} \int_y^{y+h} f(u,v) \, du \, dv \approx h^2 \, f(x,y).$$

- We can obtain the marginal pdf of $X$, say $f_X$, by integrating $f$ over all $y$:
  $$f_X(x) = \int_{-\infty}^{\infty} f(x,y) \, dy.$$  
  (Similar for $Y$).
Joint cdf

Finally, we can define the (less interesting) joint cdf.

Let $X$ and $Y$ be two random variables observed from the same random experiment. We define the joint cumulative distribution function as

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

If $X$ and $Y$ are continuous random variables, the joint pdf is then

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

15 Independence and conditional distributions

Let $X$ and $Y$ be two random variables.

**Definition:** We say $X$ and $Y$ are independent random variables if any event defined by $X$ is independent of every event defined by $Y$.

That is if $X$ and $Y$ are independent we can always say that

$$\mathbb{P}(\{a < X \leq b\} \cap \{c < Y \leq d\}) = \mathbb{P}(a < X \leq b) \mathbb{P}(c < Y \leq d)$$

for any possible choice of $a$, $b$, $c$ and $d$.

It follows that $X$ and $Y$ are independent if and only if $F_{X,Y}(x, y) = F_X(x)F_Y(y)$.
Hence, for continuous random variables:

\[ f_{X,Y}(x,y) = f_X(x) f_Y(y) \]

For discrete random variables \( X \) and \( Y \) are independent if and only if for all \( x \) and \( y \)

\[ P(X = x, Y = y) = P(X = x) P(Y = y) \]

**Example.** We draw at random a point \((X,Y)\) from the 16 points on the square \(E\) below.

Clearly \( X \) and \( Y \) are independent.

**Example.** (Lifetime distributions.)

A component has an exponential lifetime distribution with mean 1 year. If it fails there is a spare component to replace it immediately. Find the probability that one spare is enough to allow the machine to function continuously for at least a year.

Let \( X_1 \) be the lifetime of the component and \( X_2 \) the lifetime of the spare. We need to find

\[ P(X_1 + X_2 \geq 1). \]

We assume the component and spare have independent lifetimes, in which case their joint pdf is given by

\[ f(x_1, x_2) = e^{-x_1 - x_2} \]

for \( x_1 \) and \( x_2 \) both positive.

Consider the complementary event:

\[ P(X_1 + X_2 \leq 1) = \int \int_{x_1 + x_2 \leq 1} e^{-x_1 - x_2} \, dx_1 dx_2 \]
Writing the double integral as a repeated integral:

\[
P(X_1 + X_2 \leq 1) = \int_0^1 \left( \int_0^{1-x_2} e^{-x_1} \, dx_1 \right) e^{-x_2} \, dx_2
\]

\[= \int_0^1 \left( 1 - e^{-(1-x_2)} \right) e^{-x_2} \, dx_2
\]

\[= (1 - 2e^{-1})
\]

Hence

\[
P(X_1 + X_2 \geq 1) = 2/e \approx 0.73575888\ldots
\]

Exercise. Let \(X\) and \(Y\) be independent standard normal random variables. Determine the joint pdf, \(f\) say, of \(X\) and \(Y\), and give a 3-dimensional sketch of its graph.

Exercise. Let \(X \sim \text{Po}(\lambda)\) and \(Y \sim \text{Po}(\mu)\) be independent. Show that \(X + Y \sim \text{Po}(\lambda + \mu)\).
Conditional distribution

Example. We draw at random a point \((X, Y)\) from the 10 points on the triangle \(D\) below.

![Diagram of a triangle with 10 points labeled: 0, 1, 2, 3, 4, 0, 1, 2, 3, 4.]

The joint and marginal pmf's are easy to determine:

\[
P(X = i, Y = j) = \frac{1}{10}, \quad (i, j) \in D,
\]

and

\[
P(X = i) = \frac{5 - i}{10}, \quad i \in \{1, 2, 3, 4\},
\]

\[
P(Y = j) = \frac{j}{10}, \quad j \in \{1, 2, 3, 4\}.
\]

Clearly \(X\) and \(Y\) are not independent. In fact, if we know that \(X = 2\), then \(Y\) can only take the values \(j = 2, 3\) or 4. The corresponding probabilities are

\[
P(Y = j \mid X = 2) = \frac{P(Y = j, X = 2)}{P(X = 2)} = \frac{1/10}{3/10} = \frac{1}{3}.
\]

We thus have determined the conditional pmf of \(Y\) given \(X = 2\).

**Definition.** If \(X\) and \(Y\) are discrete and \(P(X = x) > 0\), then the probabilities

\[
P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)},
\]

for all \(y\), give the conditional pmf of \(Y\) given \(X = x\).
We can extend this to general $Y$:

**Definition.** If $X$ is discrete with $\mathbb{P}(X = x) > 0$, then

$$\mathbb{P}(Y \leq y \mid X = x) = \frac{\mathbb{P}(X = x, Y \leq y)}{\mathbb{P}(X = x)}$$

for all $y$, gives the **conditional cdf** of $Y$ given $X = x$.

We write $F_Y(y \mid x)$.

The corresponding density (if it exists) is the **conditional pdf** of $Y$ given $X = x$; Notation $f_Y(y \mid x)$.

When $X$ is *continuous*, we can not longer directly apply the definition of conditional probability (why not?)

Instead, we define the **conditional cdf** of $Y$ given $X = x$ as the limit

$$F_Y(y \mid x) := \lim_{h \to 0} F_Y(y \mid x < X \leq x + h).$$

**Exercise.** Show that this limit is equal to

$$\int_{-\infty}^{y} \frac{f_{X,Y}(x, v)}{f_X(x)} dv.$$

The corresponding density is called the **conditional pdf** of $Y$ given $X = x$:

$$f_Y(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$
Conditional Expectation

For discrete $X$ and $Y$, $y \mapsto \mathbb{P}(Y = y \mid X = x)$ is a genuine pmf, for each fixed $x$. Hence, we can assign to it an expectation:

$$\mathbb{E}[Y \mid x] := \sum_y y \mathbb{P}(Y = y \mid X = x).$$

Similarly, in the continuous case we can define

$$\mathbb{E}[Y \mid x] := \int_{-\infty}^{\infty} y f_Y(y \mid x) \, dy.$$

This number $\mathbb{E}[Y \mid x]$ is called the conditional expectation of $Y$ given $X = x$.

16 Multiple Random Variables

Suppose $X_1, X_2, \ldots, X_n$ are random variables pertaining to some random experiment. The joint cdf $F$ is defined by

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

which completely specifies the probability distribution of the vector $(X_1, X_2, \ldots, X_n)$.

However, if the $X_i$’s are discrete, it suffices to only know the joint pmf $p$, defined by

$$p(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Similarly, when $X_i$’s are continuous the probability distribution is completely specified by the joint pdf $f$ (if it exists):

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}.$$

Integration of $f$ over a subset $A$ of $\mathbb{R}^n$ gives the probability that the vector $(X_1, \ldots, X_n)$ lies in $A$. 
Discrete random variables $X_1, \ldots, X_n$ are said to be independent if

$$
P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1) P(X_2 = x_2) \cdots P(X_n = x_n),$$

for all $x_1, x_2, \ldots, x_n$.

Similarly, for continuous random variables with a joint density function $f$, independence is equivalent to

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

An infinite sequence $X_1, X_2, \ldots$ of random variables is called independent if for any finite choice of parameters $i_1, i_2, \ldots, i_n$ (none of them the same) the random variables $X_{i_1}, \ldots, X_{i_n}$ are independent.

**Important remarks**

- More often than not, the independence of random variables is a model assumption, rather than a consequence.

- Instead of describing a random experiment via an explicit description of $S$ and $P$, we will usually model the experiment through one or more (independent) random variables.

**Example.** The experiment where we flip a coin indefinitely is modelled by random variables $X_1, X_2, \ldots$ which are independent and Bernoulli distributed with some success parameter $p$ (known or unknown), i.e.,

$$P(X_i = 1) = p = 1 - P(X_i = 0), \quad i = 1, 2, \ldots$$

We can interpret $\{X_i = 1\}$ as the event that the $i$th result is “Heads”.
Suppose $X_1, \ldots, X_n$ are the measurements on a random experiment. Often we are interested in functions of the measurement. Examples are:

1. $X_1, \ldots, X_n$ are repeated measurements of a certain quantity. Then $(X_1 + \cdots + X_n)/n$ is what we are really interested in.

2. $X_1, \ldots, X_n$ are the lifetimes of the components in a series system. Then lifetime of the system is $\min(X_1, \ldots, X_n)$.

Let the random variable $Z$ be defined as a function of several random variables:

$$Z = g(X_1, \ldots, X_n).$$

How can we find the pmf, pdf and/or cdf of $Z$?

There are several methods.

Suppose $X_1, \ldots, X_n$ have joint density $f$. Then, the cdf of $Z$ is given by

$$F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(X \in A) \quad = \quad \int \cdots \int_A f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n,$$

where $A := \{(x_1, \ldots, x_n) : g(x_1, \ldots, x_n) \leq z\}$.

By differentiating $F_Z$ w.r.t. $z$ we can find the density $f_Z$ of $Z$.

**Example.** Consider the sum $Z = X + Y$ of two continuous random variables $X$ and $Y$ with joint pdf $f$. We have,

$$\mathbb{P}(Z \leq z) = \mathbb{P}((X, Y) \in A) = \int \int_A f(x, y) \, dx \, dy,$$

where $A = \{(x, y) : x + y \leq z\}$. Hence,

$$\int \int_A f(x, y) \, dx \, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y) \, dy \, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z} f(x, u - x) \, du \, dx$$

$$= \int_{-\infty}^{z} \int_{-\infty}^{\infty} f(x, u - x) \, dx \, du,$$
whence, by differentiation

\[ f_Z(z) = \int_{-\infty}^{\infty} f(x, z - x) \, dx, \quad z \in \mathbb{R}. \]

If, moreover, \( X \) and \( Y \) are independent, then

\[ f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \, dx, \quad z \in \mathbb{R}. \]

Hence, the density of \( Z \) is the \textit{convolution} of the densities of \( X \) and \( Y \).

**Exercise.** Let \( X \) and \( Y \) be independent and uniformly distributed on the interval \([0,1]\). Determine the density of \( Z = X + Y \).

**Example.** Let \( X_1, \ldots, X_n \) be the lifetimes of \( n \) components in a parallel system. Assume that the lifetimes are independent and exponentially distributed with parameter \( c \). The lifetime of the system is

\[ Z = \max(X_1, \ldots, X_n). \]

What is the pdf of \( Z \)?

Note first that the event \( \{ \max(X_1, \ldots, X_n) \leq z \} \) is equivalent to the intersection of all events \( \{ X_1 \leq z \}, \ldots, \{ X_n \leq z \} \).

Thus, we have

\[
\begin{align*}
P(Z \leq z) &= P(X_1 \leq z, \ldots, X_n \leq z) \\
&= P(X_1 \leq z) \cdots P(X_n \leq z) \\
&= (1 - e^{-cz})^n,
\end{align*}
\]

whence

\[ f_Z(z) = n(1 - e^{-cz})^{n-1}ce^{-cz}, \quad (z \geq 0). \]
Linear Transformation

Let \( x = (x_1, \ldots, x_n)^T \) be a (column) vector in \( \mathbb{R}^n \) and \( A \) an \( (n \times n) \)-matrix. The mapping \( x \mapsto z \), with

\[
  z = Ax
\]

is called a linear transformation. Suppose \( A \) is invertible. Hence,

\[
  x = A^{-1}z .
\]

\[\text{•} \quad \text{Any } n\text{-dimensional rectangle with “volume” } V \text{ is transformed into a } n\text{-dimensional parallelepiped with volume } V |A|, \text{ where } |A| := |\det(A)|.\]

Now consider a random (column) vector \( X = (X_1, \ldots, X_n)^T \). Let

\[
  Z = AX .
\]

If \( X \) has joint density \( f_X \), what is the joint density \( f_Z \) of \( Z \)?

Consider a fixed \( x \). Let \( z = Ax \). Hence, \( x = A^{-1}z \). Define the \( n \)-dimensional cube \( C = [z_1, z_1 + h] \times \cdots \times [z_n, z_n + h] \).

Let \( D \) be the image of \( C \) under \( A^{-1} \), i.e., the parallelepiped of all points \( x \) such that \( Ax \in C \).

Then,

\[
  \mathbb{P}(Z \in C) \approx h^n f_Z(z) .
\]

But also,

\[
  \mathbb{P}(Z \in C) = \mathbb{P}(X \in D) \approx h^n |A^{-1}| f_X(x) .
\]

We conclude that

\[
  f_Z(z) = \frac{f_X(A^{-1}z)}{|A|}, \quad z \in \mathbb{R}^n . \quad (3)
\]
17 Multiple Random Variables
Continued

General transformations

Let \( x = (x_1, \ldots, x_n)^T \) be a (column) vector.

We can apply the same technique as for the linear transformation to general transformations \( x \mapsto g(x) \).

Written out,

\[
\begin{pmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_n 
\end{pmatrix}
\mapsto
\begin{pmatrix}
  g_1(x) \\
  g_2(x) \\
  \vdots \\
  g_n(x) 
\end{pmatrix}.
\]

For a fixed \( x \), let \( z = g(x) \). Suppose \( g \) is invertible, hence, \( x = g^{-1}(z) \).

- Any infinitesimal \( n \)-dimensional rectangle at \( x \) with volume \( V \) is transformed into a \( n \)-dimensional parallelepiped at \( z \) with volume \( V \, |J_x(g)| \), where \( J_x(g) \) is the Jacobian at \( x \) of the transformation \( g \):

\[
J_x(g) = \det \begin{bmatrix}
\frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n}
\end{bmatrix}
\]

Now consider a random (column) vector \( Z = g(X) \).

Let \( C \) be a small cube around \( z \) with volume \( h^n \). Let \( D \) be image of \( C \) under \( g^{-1} \). Then, as in the linear case,

\[
\mathbb{P}(Z \in C) \approx h^n \, f_Z(z) \\
\approx h^n \, |J_z(g^{-1})| \, f_X(x) .
\]
Hence, we have the transformation rule

\[ f_Z(z) = f_X(g^{-1}(z)) \cdot |J_z(g^{-1})|, \quad z \in \mathbb{R}^n. \]

(Note: \( |J_z(g^{-1})| = 1/|J_x(g)| \))

**Important Remark.** In most coordinate transformations, it is \( g^{-1} \) rather than \( g \) which is given, i.e., we are given an expression for \( x \) as a function of \( z \).

**Example.** Let \( X \) and \( Y \) be two independent standard normal random variables. \((X, Y)\) is a random point in the plane. Let \((R, \Theta)\) be the corresponding polar coordinates. Determine the joint pdf \( f_{R,\Theta} \) of \( R \) and \( \Theta \).

Specify a point \((x, y)\) in polar coordinates:

\[ x = r \cos \theta \]

and

\[ y = r \sin \theta. \]

Note that we have specified \( x \) and \( y \) in terms of \( r \) and \( \theta \) rather than the other way around (that is we have specified \( g^{-1} \) rather than \( g \)). The corresponding Jacobian is

\[
\begin{vmatrix}
\frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\
\frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta}
\end{vmatrix} = \begin{vmatrix}
\cos \theta & -r \sin \theta \\
\sin \theta & r \cos \theta
\end{vmatrix} = r.
\]

The joint pdf of \( X \) and \( Y \) is given by (check!)

\[ f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2}. \]

Hence, using the transformation rule,

\[ f_{R,\Theta}(r, \theta) = \frac{1}{2\pi} e^{-r^2/2} r, \]

for \( r \geq 0 \) and \( \theta \in [0, 2\pi) \).

**Exercise.** Derive the marginal pdf’s of \( R \) and \( \Theta \), and show that \( R \) and \( \Theta \) are independent. Is that to be expected?
Expected value of functions of random variables

Similar to the 1-dimensional case, the expected value of \( Z = g(X_1, \ldots, X_n) \) can be evaluated in the discrete case as

\[
\mathbb{E}[Z] = \sum_x g(x, \ldots, x_n) \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n),
\]

and in the continuous case as

\[
\mathbb{E}[Z] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n,
\]

where \( f \) is the joint pdf of \( X_1, \ldots, X_n \).

**Example.** In the continuous case, find the expectation of \( X + Y \). (Do the discrete case yourself).

Let \( f \) be the joint pdf of \( X \) and \( Y \), then

\[
\mathbb{E}(X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f(x, y) \, dx \, dy
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) \, dx \, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) \, dx \, dy
\]

\[
= \int_{-\infty}^{\infty} x f_X(x) \, dx + \int_{-\infty}^{\infty} y f_Y(y) \, dy
\]

\[
= \mathbb{E}[X] + \mathbb{E}[Y].
\]

Note that \( X \) and \( Y \) do not have to be independent.

This is easily generalized to the following result:
Suppose $X_1, X_2, \ldots, X_n$ are random variables measured on the same random experiment, with means $\mu_1, \mu_2, \ldots, \mu_n$. Let

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

where $a, b_1, b_2, \ldots, b_n$ are constants. We have

$$\mathbb{E}Y = a + b_1\mathbb{E}X_1 + \cdots + b_n\mathbb{E}X_n$$

$$= a + b_1\mu_1 + \cdots + b_n\mu_n$$

That is, substitute the mean for each $X$.

Another important result is:

If $X_1, \ldots, X_n$ are independent, then

$$\mathbb{E}X_1X_2\cdots X_n = \mathbb{E}X_1\mathbb{E}X_2\cdots \mathbb{E}X_n.$$  

**Exercise.** Prove this for the continuous case with $n = 2$.

---

**Definition.** The **covariance** of two random variables $X$ and $Y$ is defined as the number

$$\text{cov}(X, Y) := \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$

It is a measure for the amount of linear dependency between the variables.

Closely related to this is the **correlation** of $X$ and $Y$, defined as

$$\rho(X, Y) := \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}X} \sqrt{\text{Var}Y}}.$$  

- The covariance (and correlation) of two independent random variables is 0.
Here are some more properties:

<table>
<thead>
<tr>
<th>Rules for Variance and Covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Var($X$) = $\mathbb{E}X^2 - (\mathbb{E}X)^2$.</td>
</tr>
<tr>
<td>2 Var($aX + b$) = $a^2$Var($X$).</td>
</tr>
<tr>
<td>3 cov($X, Y$) = $\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$.</td>
</tr>
<tr>
<td>4 cov($X, Y$) = cov($Y, X$).</td>
</tr>
<tr>
<td>5 cov($aX + bY, Z$) = $a$ cov($X, Z$) + $b$ cov($Y, Z$).</td>
</tr>
<tr>
<td>6 cov($X, X$) = Var($X$).</td>
</tr>
<tr>
<td>7 Var($X + Y$) = Var($X$) + Var($Y$) + 2 cov($X, Y$).</td>
</tr>
<tr>
<td>8 $X$ and $Y$ indep. $\implies$ cov($X, Y$) = 0.</td>
</tr>
</tbody>
</table>

As a consequence, we have

If $X_1, X_2, \ldots, X_n$ are independent

\[
\text{Var}(Y) = b_1^2 \text{Var}(X_1) + \cdots + b_n^2 \text{Var}(X_n)
\]

\[
= b_1^2 \sigma_1^2 + \cdots + b_n^2 \sigma_n^2
\]

18 Jointly Gaussian Random Variables

First, recall the 1-dimensional case:

Let $X \sim \text{N}(0, 1)$. Then, $X$ has density $f_X$ given by

\[
f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.
\]

Consider the transformation $Z = \mu + \sigma X$.

Then, $Z$ has density

\[
f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}.
\]

In other words, $Z$ has a Gaussian distribution with expectation $\mu$ and variance $\sigma^2$.

Let’s generalize this to $n$ dimensions.
The \( n \)-dimensional density of the random vector \( \mathbf{X} = (X_1, \ldots, X_n)^T \) (column vector), with \( X_1, \ldots, X_n \) independent and standard normal, is

\[
f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}e^{-\frac{1}{2}x^T x}.
\]

Consider the transformation \( \mathbf{Z} = \mathbf{\mu} + B \mathbf{X} \). The density of \( \mathbf{Z} \) is

\[
f_{\mathbf{Z}}(z) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(z-\mathbf{\mu})^T \Sigma^{-1}(z-\mathbf{\mu})},
\]

where \( \Sigma = BB^T \).

**Exercise.** Check this, using the transformation rule.

\( \mathbf{Z} \) is said to have a **multi-variate Gaussian** (or normal) distribution with **expectation vector** \( \mathbf{\mu} \) and **covariance matrix** \( \Sigma \).

**Example.** Consider the 2-dimensional case with \( \mathbf{\mu} = (\mu_1, \mu_2)^T \), and

\[
B = \begin{pmatrix}
\sigma_1 & 0 \\
\sigma_2 \rho & \sigma_2 \sqrt{1 - \rho^2}
\end{pmatrix}.
\]

The covariance matrix is now

\[
\Sigma = \begin{pmatrix}
\sigma_1^2 & \rho \sigma_1 \sigma_2 \\
\rho \sigma_1 \sigma_2 & \sigma_2^2
\end{pmatrix}.
\]

Therefore, the density is

\[
f_{\mathbf{Z}}(z) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left( \frac{(z_1 - \mu_1)^2}{\sigma_1^2} - 2\rho \frac{(z_1 - \mu_1)(z_2 - \mu_2)}{\sigma_1 \sigma_2} + \frac{(z_2 - \mu_2)^2}{\sigma_2^2} \right) \right\}.
\]

This is the pdf of the bi-variate Gaussian distribution which we encountered before.

Here are some pictures of the density, for \( \mu_1 = \mu_2 = 0 \) and \( \sigma_1 = \sigma_2 = 1 \), and for various \( \rho \).
Let, as before,

$$Z = \mu + B\mathbf{X}.$$ 

For any random vector $\mathbf{V}$ define

$$\mathbb{E}\mathbf{V} = (\mathbb{E}V_1, \ldots, \mathbb{E}V_n)^T$$

Then,

$$\mathbb{E}Z = \mu + B\mathbb{E}\mathbf{X} = \mu.$$ 

Hence the name *expectation vector*.

For any matrix $M$ of random variables, define $\mathbb{E}M$ as the matrix of expectations.

The covariance of $Z_i$ and $Z_j$ is given by $\mathbb{E}(Z_i - \mu_i)(Z_j - \mu_j)$, which is the $(i, j)$th element of the matrix

$$\mathbb{E}(B\mathbf{X})(B\mathbf{X})^T = B\mathbb{E}\mathbf{X}\mathbf{X}^TB^T = BIB^T = \Sigma.$$ 

Hence, the name *covariance matrix*.
Affine combinations

A most important property of the normal distribution is the following:

If $X_i \sim N(\mu_i, \sigma_i^2)$, independently, for $i = 1, 2, \ldots, n$, then

$$Y = a + \sum_{i=1}^{n} b_i X_i \sim N\left(a + \sum_{i=1}^{n} b_i \mu_i, \sum_{i=1}^{n} b_i^2 \sigma_i^2\right)$$

The easiest way to prove this is by using characteristic functions.

(1) The characteristic function of $X_i$ is given by

$$\Phi_{X_i}(\omega) = e^{j \mu_i \omega - \frac{1}{2} \sigma_i^2 \omega^2}$$

(2) Since $X_1, \ldots, X_n$ are independent, also $e^{j b_1 X_1 \omega}$, $\ldots$, $e^{j b_n X_n \omega}$ are independent, and hence

$$\Phi_Y(\omega) = E e^{j (a + \sum_{i=1}^{n} b_i X_i) \omega}$$

$$= e^{j a \omega} \prod_{i=1}^{n} e^{j b_i X_i \omega} = e^{j a \omega} \prod_{i=1}^{n} \Phi_{X_i}(b_i \omega)$$

$$= e^{j a \omega} \prod_{i=1}^{n} \exp\left\{j \mu_i (b_i \omega) - \frac{1}{2} \sigma_i^2 (b_i \omega)^2\right\}$$

$$= \exp\left\{ j \left( a + \sum_{i=1}^{n} b_i \mu_i \right) \omega - \frac{1}{2} \left( \sum_{i=1}^{n} b_i^2 \sigma_i^2 \right) \omega^2 \right\},$$

which is the characteristic function of a Gaussian distribution with the parameters stated above.

Example. A machine produces ball bearings with a $N(1,0.01)$ diameter (cm). The balls are placed on a sieve with a $N(1.1,0.04)$ diameter. The diameter of the balls and the sieve are assumed to be independent of each other.

What is the probability that a ball will fall through?

Let $X \sim N(1,0.01)$ and $Y \sim N(1.1,0.04)$. We need to calculate $P(Y > X) = P(Y - X > 0)$. 
But, $Z := Y - X \sim N(0.1, 0.05)$. Hence
\[
\mathbb{P}(Z > 0) = \mathbb{P} \left( \frac{Z - 0.1}{\sqrt{0.05}} > \frac{-0.1}{\sqrt{0.05}} \right) = 1 - Q(0.447) \approx 0.67.
\]

**Exercise.** Let $X = (X_1, X_2)^T$ have a bi-variate normal distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ and $\rho$. Prove the following.

1. $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$.

2. $\rho(X_1, X_2) = \rho$.

3. The conditional distribution of $X_2$ given $X_1 = x_1$ is
   \[
   N \left( \mu_2 + \frac{\sigma_2 \rho}{\sigma_1} (x_1 - \mu_1), \frac{\sigma_2^2 (1 - \rho^2)}{\sigma_1^2} \right).
   \]

\section*{19 Sums of Independent Random Variables}

In this section we discuss two celebrated theorems in probability: the Law of Large Numbers and the Central Limit Theorem.

Both theorem deal with \textit{sums of independent random variables}. They arise for example in the following situations:

1. We flip a (biased) coin infinitely many times. Let $X_i = 1$ if the $i$th flip is “heads” and $X_i = 0$ otherwise.
   
   In general we do not know $p = \mathbb{P}(X_i = 1)$. However, using the outcomes $x_1, \ldots, x_n$, we would estimate $p$ by $(x_1 + \cdots + x_n)/n$.

2. A certain machine needs to work continuously. The machine has one component that is very unreliable. This component is replaced immediately upon failure. Suppose there are $n$ such (spare) components. If we denote the component lifetimes by
$X_1, \ldots, X_n$, then the lifetime of the machine is given by $X_1 + \cdots + X_n$.

3. We weigh 20 randomly selected people. The average weight of the group is $(X_1 + \cdots + X_{20})/20$.

Let $X_1, X_n, \ldots$ be independent and identically distributed random variables. For each $n$ let

$$S_n = X_1 + \cdots + X_n.$$ 

Let $\mathbb{E}X_i = \mu$ and $\text{Var}(X_i) = \sigma^2$ (assuming that these are finite).

What can we say about the random variable $S_n$? Some easy results are:

$$\mathbb{E}S_n = n \mathbb{E}X_1 = n\mu.$$ 

and, by the independence,

$$\text{Var}(S_n) = n \text{Var}(X_1) = n\sigma^2.$$ 

If we know the pdf or pmf of $X_i$, then we can (in principle) determine the pdf or pmf of $S_n$.

The easiest way is to use transform techniques (Laplace transform, Characteristic function, etc.)

An important property of these transforms is that the transform of the sum of independent random variables is equal to the product of the individual transforms. Can you see why? The general procedure will be clear from the following example.

Example. Suppose each $X_i \sim \text{Exp}(\lambda)$. The Laplace transform of $X_i$, say $L$, is given by

$$L(s) = \mathbb{E}e^{-sX_i} = \frac{\lambda}{\lambda + s}.$$ 

The Laplace transform of $S_n$, is given by

$$\mathbb{E}e^{-sS_n} = \mathbb{E}e^{-s(X_1 + \cdots + X_n)}$$ 

$$= \mathbb{E}e^{-sX_1} \cdots \mathbb{E}e^{-sX_n} = (L(s))^n$$ 

$$= \left( \frac{\lambda}{\lambda + s} \right)^n.$$ 

Using the uniqueness of Laplace transforms, this shows that $S_n$ has a $\text{Gam}(n, \lambda)$ distribution (Erlang distribution).
Law of Large Numbers

Consider the coin flip example. We expect that $S_n/n$ is close to the unknown $p$ for large $n$. We know this happens “empirically”:

![Graph showing empirical distribution of $S_n/n$ over 100 coin flips]

In general, we expect $S_n/n$ to be close to $\mu$. Does this happen in our mathematical model?

By Chebyshev’s inequality we have for all $\epsilon > 0$,

$$\mathbb{P} \left( \left| \frac{S_n}{n} - \mu \right| > \epsilon \right) \leq \frac{\text{Var}(S_n/n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0,$$

as $n \to \infty$.

In other words the probability that $S_n/n$ is more than $\epsilon$ away from $\mu$ can be made arbitrarily small by choosing $n$ large enough.

This is the Weak Law of Large Numbers.

There is also a Strong Law of Large Numbers:

$$\mathbb{P} \left( \lim_{n \to \infty} \frac{S_n}{n} = \mu \right) = 1,$$

as $n \to \infty$. 

© STAT2202 153 © STAT2202 154
Central Limit Theorem

The Central Limit Theorem states, roughly, this:

The sum of a large number of iid random variables has approximately a Gaussian distribution.

More precisely, it states that for all $x$

$$\lim_{n \to \infty} \mathbb{P} \left( \frac{S_n - n\mu}{\sigma \sqrt{n}} \leq x \right) = \Phi(x),$$

where $\Phi$ is the cdf of the standard normal distribution.

Let us see the CLT in action.

The next picture shows the pdf’s of $S_1, \ldots, S_4$ for the case where the $X_i$ have a $U[0, 1]$ distribution.

The same, but now for an $\text{Exp}(1)$ distribution.
Note that the CLT is not restricted to continuous distributions. The next picture shows the cdf of $S_{30}$ in the case where the $X_i$ have a Bernoulli distribution with success probability $1/2$.

![CDF of $S_{30}$](image1)

Note that $S_{30} \sim \text{Bin}(30, 1/2)$.

Using the CLT we thus find:

Let $X \sim \text{Bin}(n, p)$. For large $n$, we have

$$
P(X \leq k) \approx P(Y \leq k),$$

where $Y \sim N(np, np(1-p))$.

For example,

$$
P(X = k) \approx P(k - \frac{1}{2} \leq Y \leq k + \frac{1}{2}).$$

As a rule of thumb, the approximation is accurate if both $np$ and $n(1-p)$ are larger than 5.

We can improve on this somewhat by using a *continuity correction*, as illustrated by the following graph for the pmf of the $\text{Bin}(10, 1/2)$ distribution.

![PMF of Bin(10, 1/2)](image2)
Example.

In Lecture 5 we calculated, using the computer, the probability $P(X \leq 99)$, where $X \sim \text{Bin}(200, 0.51)$. Let $Y \sim \text{N}(200 \times 0.51, 200 \times 0.51 \times 0.49)$, and let $Z$ be standard normal. Using the CLT we have

$$P(X \leq 99) \approx P(Y \leq 99)$$
$$= P\left(\frac{Y - 102}{\sqrt{49.98}} \leq \frac{99 - 102}{\sqrt{49.98}}\right)$$
$$= P(Z \leq -0.4243) = 1 - P(Z \leq 0.4243)$$
$$= 0.3357 .$$

Using the continuity correction we find

$$P(X \leq 99) \approx P(Y \leq 99 + 1/2) = 0.3618 .$$

(Exact answer: 0.361704 ... )

Exercise. The number of calls $X$ arriving at a call centre during an hour has a Po(100) distribution.

Show, using probability generating functions, that $X$ has the same distribution as $X_1 + \cdots + X_{100}$, where $X_1, \ldots, X_{100}$ are independent Po(1) distributed random variables.

Use this fact to approximate (with the CLT) the probability that there are more than 130 arrivals during an hour.

(Answer: 0.00135)