

EE2 Mathematics

Probability and Statistics

Dr Bruno Clerckx

Department of Electrical and Electronic Engineering, Imperial College London

October 2017

b.clerckx@imperial.ac.uk

Room 816, EEE Building

Course materials on Blackboard page.

1/321

2/321

Teaching details:

- ▶ Autumn 2017 term
 - ▶ 15 Lectures
 - ▶ 7 Classes
- ▶ Summer 2018 term
 - ▶ 2 Lectures (for Revision)

Course content:

- ▶ Probability & Statistics

Exam Paper: 2 prob/stats questions, closed-book

Texts

The general content of the course is covered in many books.

Look for “Statistics For Engineers” type titles:

- ▶ Montgomery, D.C. and Runger, G.C. (2011) *Applied Statistics and Probability for Engineers*, 5th edition, Wiley.
- ▶ Devore, J.L (2004) *Probability and Statistics for Engineering and the Sciences*, 7th edition, Thomson/Brooks/Cole.
- ▶ Vining, G.G. (2006) *Statistical Methods for Engineers*, 2nd edition, Brooks/Cole.

More mathematical:

- ▶ Rice, J. (1993) *Mathematical Statistics and Data Analysis*, Wadsworth.

For EEE graduates (reference for the 4th year course)

- ▶ Papoulis, A and Pillai, S. (2002) *Probability, Random Variables and Stochastic Processes*, McGraw-Hill.

3/321

4/321

Introduction I

Probability is the mathematical science of *UNCERTAINTY*, which provides precise rules for analysing and understanding our ignorance about uncertain situations.

Diverse practical applications, including:

- ▶ modeling and control of financial instruments
- ▶ object tracking
- ▶ queuing theory
- ▶ statistics
- ▶ telecommunications

5/321

Statistical Analysis

A statistical analysis often consists of one or more of the following:

- ▶ Data Summary - “what is the average salary of Imperial graduates?”
- ▶ Prediction - “Can we determine that a component will fail?”
- ▶ Decision making - “Which site should we select to drill for oil?”
- ▶ Answering specific questions - “studying fuel supplement performance... is a new supplement really better than an existing one?”

Often, a statistical analysis can be viewed as attempting to separate a *signal* from *noise*. This gives a simple model for reasoning about data

$$\text{DATA} = \text{SIGNAL} + \text{NOISE}$$

7/321

Introduction II

Statistics is the science (and *art*) of reasoning about *DATA* in the presence of *UNCERTAINTY*. This uncertainty can arise for example when a signal is corrupted by *NOISE* or when we cannot observe key components of a process.

Example

Sampling uncertainty



Example

Credit Scoring



It is the dual components of data and uncertainty that characterise statistics.

6/321

Contents

Events, Probability and Sets

Random Variables and Probability Distributions

Systems and Component Reliability

Jointly Distributed Random Variables

Law of Large Numbers and Central Limit Theorem

Statistics

8/321

Events, Probability and Sets

Events, Probability and Sets

Set Theory

Sample Spaces and Events

Probability Axioms

Conditional Probability

Probability Tables

Total Probability

Bayes Theorem

Random Variables and Probability Distributions

Systems and Component Reliability

Jointly Distributed Random Variables

Law of Large Numbers and Central Limit Theorem

Events, Probability and Sets

9/321

Set Theory

A *set* is a collection of unordered, distinct objects. The objects in a set are called *elements*, denoted ω . If object ω belongs to set A , we say “ ω belongs to A ” and write $\omega \in A$. Conversely, if ω is not in A , we write $\omega \notin A$.

The contents of a set are enclosed in curly braces. For example, $A = \{\omega\}$, is a set with *cardinality* 1. Cardinality is a measure of the number of elements of a set. A set with a single element is referred to as a *singleton*.

Note that elements of a set can be any mixture of different objects, such that

$$B = \{1, 2, 3, A, \omega\}$$

is a set that contains the integers 1,2 and 3, the set A , and the object ω .

Events and Probability

In many real world situations, the outcome of an action is uncertain.

However, we can often list all outcomes that could happen, and then make statements concerning which outcomes were more or less likely.

Probability is the tool we use to make such statements, and set theory is the mathematical framework in which probability is formalised.

We begin with a review of relevant set theory.

Events, Probability and Sets

10/321

Sometimes, when we do not wish to explicitly name the elements of a set, we use the generic element ω , with a numeric subscript to imply a label

$$B = \{\omega_1, \omega_2, \dots, \omega_n\}$$

If it is tiresome, or impossible, to list all the elements of a set, we can use either a clear textual description or a mathematical description.

Example

Playing cards with face values: =

{playing cards with face values}

Number of accidents in a year: $\{0, 1, 2, \dots\}$

Height: $\{x : x > 0\}$ ■

Important sets

The *empty set*, denoted $\emptyset = \{\}$, is the unique set that contains no elements.

For application purposes, we restrict the type of elements a set can contain to a collection of plausible and relevant values. This collection is itself a set, called the *universal set*, and denoted Ω . In the previous example, all three sets could be regarded as universal sets.

Example

Throw a die



Set Operations

Set theory provides operations for manipulating collections of sets.

UNION

The *union* of two sets, A and B , is defined to be the set containing all elements in A alone, all elements in B alone, and all elements shared by both A and B . This is written $A \cup B$, and formally defined as

$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$$

Example

Consider throwing a standard die. Define A as throwing an even number, and B as throwing a number greater than 3. Then

$$A \cup B = \{\square, \blacksquare, \blacksquare, \blacksquare\} \cup \{\blacksquare, \blacksquare, \blacksquare, \blacksquare\} = \{\square, \blacksquare, \blacksquare, \blacksquare\}$$



Set relations

As in arithmetic, where we can compare numbers, set theory has means of comparing sets. We say that “ A is a subset of B ”, and write

$$A \subseteq B$$

if *all* elements of A are also members of B .

If it transpires that $B \subseteq A$ and $A \subseteq B$, then A and B have identical elements, and are said to be equal, $A = B$.

It is always the case that $\emptyset \subseteq A$, for all $A \neq \emptyset$.

Example

$$\begin{aligned} A = \{1, 2\}, B = \{2, 1\} & \quad A = B \\ A = \{1, 2\}, B = \{2, 1, 3\} & \quad A \subseteq B \end{aligned}$$



For events A and B , the union has the following properties

$$\begin{aligned} A \cup \emptyset &= A & \text{identity law} \\ A \cup A &= A & \text{idempotent law} \\ A \cup \Omega &= \Omega & \text{domination law} \\ A \cup B &= B \cup A & \text{commutative law} \end{aligned}$$

More generally, for a list of sets A_1, A_2, \dots, A_n , the union is the set that contains all elements that belong to at least one of the n sets, written

$$\bigcup_{i=1}^n A_i = \{\omega \in \Omega : \text{for some } i, \omega \in A_i\}$$

INTERSECTION

The *intersection* of two sets, A and B , is defined to be the set containing all elements that belong to both A and B . This is written $A \cap B$, and formally defined as

$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$$

Example

Consider throwing a standard die. Define A as throwing an even number, and B as throwing a number greater than 3. Then

$$A \cap B = \{\square, \blacksquare, \boxtimes\} \cap \{\blacksquare, \boxtimes, \boxdot\} = \{\boxtimes, \boxdot\}$$

■

We also need to consider how to combine these operations. For three sets, A , B and C , the *distributive* law is

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \end{aligned}$$

Example

Consider $A = \{1, 2, 3\}$, $B = \{2, 3, 4\}$ and $C = \{1, 5\}$. Then

$$\begin{aligned} A \cap (B \cup C) &= \{1, 2, 3\} \cap (\{2, 3, 4\} \cup \{1, 5\}) \\ &= \{1, 2, 3\} \cap \{1, 2, 3, 4, 5\} = \{1, 2, 3\} \\ (A \cap B) \cup (A \cap C) &= (\{1, 2, 3\} \cap \{2, 3, 4\}) \cup (\{1, 2, 3\} \cap \{1, 5\}) \\ &= \{2, 3\} \cup \{1\} = \{1, 2, 3\} \end{aligned}$$

■

For events A and B , the intersection has the following properties

$$\begin{aligned} A \cap \emptyset &= \emptyset && \text{domination law} \\ A \cap A &= A && \text{idempotent law} \\ A \cap \Omega &= A && \text{identity law} \\ A \cap B &= B \cap A && \text{commutative law} \end{aligned}$$

More generally, for a list of sets A_1, A_2, \dots, A_n , the intersection is the set that contains all elements that belong to all of the n sets.

$$\bigcap_{i=1}^n A_i = \{\omega \in \Omega : \text{for all } i, \omega \in A_i\}$$

Now, from inspection of the definitions of union and intersection, we have that

$$A \cap B \subseteq A \subseteq A \cup B$$

For sets A_1, A_2, \dots, A_n , an interesting case occurs if pairs of sets A_i and A_j share no elements, for all i, j and $i \neq j$. More formally, if

$$A_i \cap A_j = \emptyset$$

for all pairs i, j , then the sets are said to be *disjoint*. Disjoint sets will prove important in the mathematical development of probability.

A further interesting case occurs if a collection of sets are disjoint, but the union of the sets is the universal set. That is, for A_1, A_2, \dots, A_n disjoint, if

$$\bigcup_{i=1}^n A_i = \Omega$$

then the sets A_1, A_2, \dots, A_n form a *partition* of Ω . This case will also prove important in the development of probability.

Complements and differences

The final set operators we require are complements and differences.

The *complement* of a set A is defined to be the set that contains all elements of Ω that do not belong to A . The complement is denoted \bar{A} and we say “not A ”.

Properties

$$\begin{aligned}\overline{(\bar{A})} &= A & \bar{\emptyset} &= \Omega \\ A \cup \bar{A} &= \Omega & A \cap \bar{A} &= \emptyset\end{aligned}$$

Example

Let A be the set of odd numbered elements when throwing a die. Then

$$\bar{A} = \{\square, \blacksquare, \boxtimes\}$$



Useful identities can be found starting with operations on the universal set.

$$\begin{aligned}A &= A \cap \Omega & A \cup B &= (A \cup B) \cap \Omega \\ &= A \cap (B \cup \bar{B}) & &= (A \cup B) \cap (B \cup \bar{B}) \\ &= (A \cap B) \cup (A \cap \bar{B}) & &= B \cup (A \cap \bar{B})\end{aligned}$$

Note that in both cases, we are left with a *disjoint union*, since the operands of the union include the complementary events B and \bar{B} .

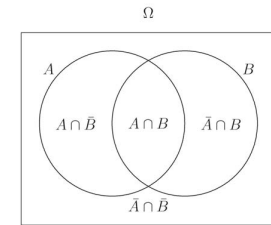
Closely related to the complement is the *difference*, A/B , which selects the subset of elements of A which do not belong to B . Formally,

$$A/B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$$

This gives another view of disjoint sets. A and B are disjoint if and only if $A/B = A$.

Another way of thinking about differences is to note the following

$$A/B = A \cap \bar{B} \quad \bar{A} = \Omega/A$$



Finally, we need tools for mixing complements, unions and intersections. The rules for this are given by *De Morgan's laws* which state

$$\begin{aligned}\overline{(A \cup B)} &= \bar{A} \cap \bar{B} \\ \overline{(A \cap B)} &= \bar{A} \cup \bar{B}\end{aligned}$$

Example

Consider $\Omega = \{1, 2, \dots, 10\}$, $A = \{1, 2, 3, 4, 5, 6\}$ and $B = \{5, 6, 7, 8, 9\}$, then

- ▶ $(A \cup B) = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$
- ▶ $\overline{(A \cup B)} = \bar{A} \cap \bar{B} = \{10\}$
- ▶ $(A \cap B) = \{5, 6\}$
- ▶ $\overline{(A \cap B)} = \bar{A} \cup \bar{B} = \{1, 2, 3, 4, 7, 8, 9, 10\}$



Of course, our objective in using set theory is describe aspects of the world in a manipulable notation.

Example

An aircraft has 3 engines A , B and C , each of which either works or fails. Ω has 8 outcomes.

Denote by A the event that engine A works, etc. (and note the slight abuse of notation).

Then

- ▶ All 3 engines work: $A \cap B \cap C$
- ▶ All 3 engines fail: $\overline{A} \cap \overline{B} \cap \overline{C} = \overline{(A \cup B \cup C)}$
- ▶ Exactly one engine works:

$$(A \cap \overline{B} \cap \overline{C}) \cup (\overline{A} \cap B \cap \overline{C}) \cup (\overline{A} \cap \overline{B} \cap C)$$

- ▶ At least two engines work:

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

Sample Spaces and Events

We now consider a *random experiment*, where the consequences of an action, the *outcome*, is unknown, but all possible outcomes can be described by a non-empty set S , called the *sample space*. In thinking about random experiments, outcomes are elements of S , which takes the role of the universal set.

Example

Toss of a coin: $S = \{HEAD, TAIL\}$

Roll of a single die: $S = \{\square, \blacksquare, \blacklozenge, \blacktriangle, \blacktriangledown, \blacktriangleright\}$

Example

An experiment involves selecting a molded plastic connector and measuring its thickness. We might choose to define the sample space S as

$$S = \{x : x > 0\}$$

since a negative value can never occur.

Exercise

Let A, B and C be three arbitrary events. Using only the operations of union, intersection and complement, write down expressions for the events that, of A, B, C ;

- (a) Only A occurs.
- (b) Both A and B , but not C occurs.
- (c) All three events occur.
- (d) At least one event occurs.
- (e) At least two events occur.
- (f) One and only one event occurs.
- (g) Exactly two events occur.
- (h) No events occur.
- (i) Not more than two events occur.

Perhaps the connectors are *known* to be constructed within fixed tolerances, say

$$S = \{x : 20 < x < 30\}$$

Alternatively, the allowed range may be categorised, such that we could consider

$$S = \{low, medium, high\}$$

It will usually be convenient to distinguish between *discrete* samples spaces, that are finite, or countably infinite, and *continuous* sample spaces that involve intervals of real numbers.

Events

Now that we have a means for describing all outcomes of a random experiment, we can define *events* as *subsets* of the sample space. That is, an event is a subset of the possible outcomes.

Example

Coin tossing: $E = \{\text{HEAD}\}, E = \{\text{TAIL}\}$

Die rolling: $E = \{\square\}, E = \{\text{Even number}\} = \{\square, \blacksquare, \boxtimes\}$

Tossing two coins:

$E = \{\text{Head on first toss}\} = \{(\text{HEAD}, \text{HEAD}), (\text{HEAD}, \text{TAIL})\}$

Molded plastic connectors:

$E = \{\text{connector medium or high quality}\} = \{\text{medium, high}\}$ ■

Probability Axioms

To characterise the uncertainty of an event, we define a set function P , called a probability function (or more formally, a probability measure) that takes a set as argument, and returns a value. For this set function to be a probability, for any event $E \subseteq S$

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. if $E \cap F = \emptyset$ then
 $P(E \cup F) = P(E) + P(F)$

For disjoint subsets $E_1, E_2, \dots \in S$, axiom 3 generalises to

$$P\left(\bigcup_i E_i\right) = \sum_i P(E_i).$$

Since any elementary outcome ω is an event, it has probability $P(\omega)$.

Special Events

In the context of a random experiment,

- ▶ the empty set, \emptyset is described as the *null event*.
- ▶ An event that is a singleton subset of S is called an *elementary event* of S .
- ▶ The sample space S consists of the union of all elementary events, and is referred to as the *universal event*. This means that the universal event will *always* occur (that is, at least one elementary outcome must occur).
- ▶ Similarly, the null event *never* occurs.

Once the experiment has been conducted, the outcome will be $\omega^* \in S$.

- ▶ For any event $E \subseteq S$ we say that E has occurred if and only if $\omega^* \in E$.
- ▶ The purpose of probability is to quantify the uncertainty of events E between the null event and the universal event.

Various useful results arise from the axioms. These often concern spotting disjoint unions and using axiom 3. For example, consider the event E and its complement

$$E \cup \overline{E} = S$$

Since this is a disjoint union, we have

$$\begin{aligned} P(E) + P(\overline{E}) &= P(S) = 1 \\ P(E) &= 1 - P(\overline{E}) \end{aligned}$$

This is a very useful result relating the probability of complementary events. Sometimes it will be difficult to compute $P(E)$ but straightforward to compute $P(\overline{E})$.

Another interesting result concerns the union of arbitrary events E and F . Consider

$$\begin{aligned} E \cup F &= E \cup (\bar{E} \cap F) \\ P(E \cup F) &= P(E) + P(\bar{E} \cap F) \end{aligned} \quad (\text{A})$$

Also

$$\begin{aligned} F &= (E \cap F) \cup (\bar{E} \cap F) \\ P(F) &= P(E \cap F) + P(\bar{E} \cap F) \end{aligned} \quad (\text{B})$$

Rearrange (B) and substitute into (A), to obtain

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Known as Boole's relationship. Note that if $E \cap F = \emptyset$, this reduces to axiom 2. Easily interpreted using a Venn diagram.

Since $P(E \cap F) \geq 0$,

$$P(E \cup F) \leq P(E) + P(F).$$

If $E \subseteq F$,

$$F = (F \cap E) \cup (F \cap \bar{E}) = E \cup (F \cap \bar{E})$$

with $E \cap (F \cap \bar{E}) = \emptyset$. Hence, $P(F) = P(E) + P(F \cap \bar{E})$ and

$$P(E) \leq P(F).$$

For any two events E and F ,

$$\max(P(E), P(F)) \leq P(E \cup F) \leq P(E) + P(F)$$

$$P(E) + P(F) - 1 \leq P(E \cap F) \leq \min(P(E), P(F))$$

Proof:

$$\begin{aligned} P(E \cap F) &= P(E) + P(F) - P(E \cup F) \geq P(E) + P(F) - 1 \\ P(E \cap F) &= P(E) + P(F) - P(E \cup F) \\ &\leq P(E) + P(F) - \max(P(E), P(F)) \\ &= \min(P(E), P(F)) \end{aligned}$$

■

Example

Consider events A , B and C , with

$$\begin{aligned} P(A) &= P(B) = P(C) = 0.90 \\ P(A \cap B) &= P(B \cap C) = P(A \cap C) = 0.85 \\ P(A \cap B \cap C) &= 0.83 \end{aligned}$$

Using the fact that $A = (A \cap \bar{B}) \cup (A \cap B)$, it follows that

1. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 2 \times 0.9 - 0.85 = 0.95$
2. $P(\bar{A} \cap B) = P(B) - P(A \cap B) = 0.05$

More difficult

$$\begin{aligned}
 P(A \cup B \cup C) &= P(A \cup B) + P(C) - P(C \cap (A \cup B)) \\
 &= P(A \cup B) + P(C) - P((C \cap A) \cup (C \cap B)) \\
 &= P(A \cup B) + P(C) \\
 &\quad - P(C \cap A) - P(C \cap B) + P(C \cap A \cap B) \\
 &= 0.95 + 0.90 - 0.85 - 0.85 + 0.83 \\
 &= 0.98
 \end{aligned}$$



Poincare's Formula:

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

Boole's inequality:

$$P(\cup_{j=1}^n A_j) \leq \sum_{j=1}^n P(A_j).$$

Computing Probabilities

In the classical context, we consider a sample space S consisting of n *equally likely* elementary events, and the probability of event $E \subseteq S$ is

$$P(E) = \frac{\text{\#elementary events in } E}{n}$$

Example

Roll of a die.

$$S = \{ \begin{smallmatrix} \blacksquare \\ \bullet \end{smallmatrix}, \begin{smallmatrix} \blacksquare \\ \bullet \bullet \end{smallmatrix}, \begin{smallmatrix} \blacksquare \\ \bullet \bullet \bullet \end{smallmatrix}, \begin{smallmatrix} \blacksquare \\ \bullet \bullet \bullet \bullet \end{smallmatrix}, \begin{smallmatrix} \blacksquare \\ \bullet \bullet \bullet \bullet \bullet \end{smallmatrix}, \begin{smallmatrix} \blacksquare \\ \bullet \bullet \bullet \bullet \bullet \bullet \end{smallmatrix} \}$$

Now

$$P(\begin{smallmatrix} \blacksquare \\ \bullet \end{smallmatrix}) = P(\begin{smallmatrix} \blacksquare \\ \bullet \bullet \end{smallmatrix}) = \dots = P(\begin{smallmatrix} \blacksquare \\ \bullet \bullet \bullet \bullet \bullet \end{smallmatrix}) = \frac{1}{6} \quad \text{by symmetry}$$

and

$$P(\text{Even number}) = P(\begin{smallmatrix} \blacksquare \\ \bullet \bullet \end{smallmatrix} \cup \begin{smallmatrix} \blacksquare \\ \bullet \bullet \bullet \bullet \end{smallmatrix} \cup \begin{smallmatrix} \blacksquare \\ \bullet \bullet \bullet \bullet \bullet \bullet \end{smallmatrix}) = P(\begin{smallmatrix} \blacksquare \\ \bullet \bullet \end{smallmatrix}) + P(\begin{smallmatrix} \blacksquare \\ \bullet \bullet \bullet \bullet \end{smallmatrix}) + P(\begin{smallmatrix} \blacksquare \\ \bullet \bullet \bullet \bullet \bullet \bullet \end{smallmatrix}) = \frac{1}{2}$$

Example

Testing the lifetime of 200 light bulbs:

Lifetime (h)	Number of bulbs	Proportion
< 1000	45	0.225
[1000, 1500]	80	0.400
> 1500	75	0.375

Compute the probability that a bulb lasts less than 1500 hours. Denote lasting less than 1000 hours as Sh (for short), between 1000 and 1500 as M , and more then 1500 as L . Then, since the categories are disjoint

$$\begin{aligned}
 P(\text{bulb lasts} < 1500) &= P(Sh \cup M) \\
 &= P(Sh) + P(M) = 0.225 + 0.4 = \underline{0.625}
 \end{aligned}$$

Note that events form a partition

$$S = Sh \cup M \cup L$$

To compute the probability of the union of arbitrary events, disjoint or otherwise, then we require the *addition rule of probability*, derived earlier

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

Heuristically, the final term takes out the overlap between the events that we would otherwise count twice.

This result can be generalised to $n > 2$ events.

Example

A card is selected at random from a standard deck. Let A refer to the event of drawing a heart, and B refer to the event of drawing a face card, $\{J, Q, K\}$ in any suit

Then

$$P(A) = \frac{13}{52} = \frac{1}{4} \quad P(B) = \frac{12}{52} = \frac{3}{13}$$

A and B are not disjoint, since some face cards are hearts.

$$P(A \cap B) = \frac{3}{52}$$

Now, the addition rule provides the means to compute the probability of A or B

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{4} + \frac{3}{13} - \frac{3}{52} = \frac{11}{26}$$



Conditional Probability

We now turn to the key concept of *conditional probability*. This is concerned with describing the probability of an event, given that another event has occurred. This is of central importance, since events are often influenced by other factors.

Example

Gender gap in politics: $P(\text{labour}) \neq P(\text{labour}|\text{female voter})$

Multiple component failure: $P(\text{engine stops}|\text{plug failure})$ ■

We use the notation

$$P(A|B)$$

to mean “the probability that event A occurs, given that event B has occurred”, or more concisely, “the probability of A given B ”.

For events A and B , with $P(B) > 0$, the conditional probability $P(A|B)$ is

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

Conditional probability considers the reduced subset of the sample space given by the conditioning event.

Multiplication Law of probability

If $B \subseteq A$,

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

If $A \cap B = \emptyset$,

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

By rearranging the conditional probability, we have the *multiplication law* of probabilities

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

From which it follows that

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

This illustrates the relationship between the conditional probabilities, and the role of the unconditional probabilities, $P(A)$ and $P(B)$.

Note that the only situation that the conditional probabilities are equal is when $P(A) = P(B)$.

Example

Throw dice. Let $A = \{\text{score an even number}\}$ and $B = \{\text{score} \geq 3\}$.

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{2}{3}, \quad P(A \cap B) = \frac{1}{3}$$

then

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

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



Independence

Conditional probability allows us to define the concept of *independence*, which indicates that the occurrence of one event does not change the probability of the occurrence of another event.

Events A and B are said to be *independent* if

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

For independent events, using the multiplication rule gives

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

This feature, independent events having multiplicative probabilities, will prove useful in statistical analysis.

For n events, we have *mutual independence* if

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

Example

Throw dice. Let $A = \{\square, \blacksquare, \blacksquare\}$ and $B = \{\blacksquare, \blacksquare, \blacksquare\}$.

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

then

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

and

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

Thus A and B are independent.



Example

Two independently operating missiles, M_1 and M_2 are aimed at a target, with $P(M_i \text{ hits}) = 0.7$, for $i = 1, 2$. Then

$$P(\text{both hit}) = P(M_1 \text{ hits} \cap M_2 \text{ hits}) = (0.7)^2 = 0.49.$$

$$P(\text{neither hits}) = (0.3)^2 = 0.09.$$

$$P(\text{at least one hits}) = 1 - 0.09 = 0.91.$$

Note that the last case could be calculated as

$$P(M_1 \text{ hits} \cup M_2 \text{ hits}) = 0.7 + 0.7 - 0.49 = 0.91$$



Standard results for probability extend to the conditional probability, such that conditional probabilities behave like ordinary probabilities. For example, for events A and B

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

Recalling that

$$B = (A \cap B) \cup (\bar{A} \cap B) \implies P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

then

$$P(\bar{A} \cap B) = P(B) - P(A \cap B)$$

divide both sides by $P(B)$

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

Probability Tables

We can now consider the probabilities of pairs of events, and their complements. Such information, can be conveniently represented as a *probability table*, as follows

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

Note that this table represents disjoint unions in a simple way. For example $P(A) + P(\bar{A}) = 1$, or $P(A) = P(A \cap B) + P(A \cap \bar{B})$. Moreover, conditional probability information is also readily computed. Recall the definition of the conditional probability, $P(A|B)$. Then we restrict attention to the B row, and the specific cell $A \cap B$.

Example

An order of 100 girders arrives at a building site. The girders are checked for two defects, A or B . Two are found to have both type A and type B defects ($A \cap B$), 6 have just type A defects ($A \cap \bar{B}$) and 4 have just type B defects ($\bar{A} \cap B$). (Note that this works with counts, as well as proportions and probabilities)

	A	\bar{A}	6
B	2	4	6
\bar{B}	6	?	100 - 6
	8	100 - 8	100

The problem structure is now clearly delineated and conditional probability calculations are straightforward.

For example, we have

$$P(A|B) = \frac{1}{3} \quad P(B|A) = \frac{2}{8} = \frac{1}{4}$$

Or, the probability of *exactly* one fault, is

$$P((A \cap \bar{B}) \cup (\bar{A} \cap B)) = P(A \cap \bar{B}) + P(\bar{A} \cap B) = \frac{6}{100} + \frac{4}{100} = \frac{1}{10}$$

■

Such representations can be extended to more complicated layouts, provided the disjoint union structure is preserved.

Total Probability

Recall that a collection of disjoint sets A_1, A_2, \dots, A_k , forms a partition of S , if $A_i \cap A_j = \emptyset$ and

$$S = \bigcup_{i=1}^k A_i$$

Now, for any event B , we can write

$$B = (B \cap A_1) \cup \dots (B \cap A_k) \implies P(B) = P(B \cap A_1) + \dots (B \cap A_k)$$

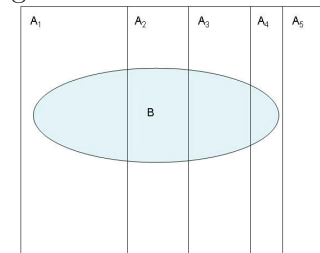
Now, we can write each event in terms of the multiplication law, to yield

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_k)P(A_k) \\ &= \sum_{i=1}^k P(B|A_i)P(A_i) \end{aligned}$$

For events A_1, A_2, \dots, A_k such that $A_i \cap A_j = \emptyset$ for all $i, j, i \neq j$, and $\bigcup_{i=1}^k A_i = S$, and event B , the theorem of total probability states

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

e.g.



This provides a method of re-assembling an unconditional probability from specified conditional probabilities.

Example

A factory uses 3 machines, X , Y and Z to produce a specific component. Suppose

1. Machine X produces 50% of the components, of which 3% are defective
2. Machine Y produces 30% of the components, of which 4% are defective
3. Machine Z produces 20% of the components, of which 5% are defective

Compute the probability that a randomly selected item is a defective.

Let D denote the event that an item is defective,

$$D = (D \cap X) \cup (D \cap Y) \cup (D \cap Z).$$

By the law of total probability

$$\begin{aligned} P(D) &= P(D|X)P(X) + P(D|Y)P(Y) + P(D|Z)P(Z) \\ &= 0.03(0.5) + 0.04(0.3) + 0.05(0.2) = \underline{0.037} \end{aligned}$$

■

Bayes Theorem

We now turn to the rule for switching conditional probabilities, *Bayes theorem*. For events A_1, \dots, A_k forming a partition of S , and any other event B , the multiplication rule for conditional probability states

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

Therefore

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

Note, that the term in the denominator can be expressed in terms of conditional probabilities via the theorem of total probability.

For events A_1, A_2, \dots, A_k that form a partition of S , and event B , Bayes rule is

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{P(A_1)P(B|A_1) + \dots + P(A_k)P(B|A_k)}$$

Another way to think of Bayes theorem is as follows. If we regard the events A_i as possible causes of the event B , Bayes theorem enables us to determine the probability that a particular A occurred, given that B occurred.

Example

Continuing the previous example (defective components). Suppose a defective component is found among the output of the factory. What is the probability that it came from each of the machines X , Y and Z .

We seek $P(X|D)$, $P(Y|D)$ and $P(Z|D)$. Earlier we found $P(D) = 0.037$.

$$P(X|D) = \frac{P(D|X)P(X)}{P(D)} = \frac{0.03(0.5)}{0.037} = 0.4054$$

$$P(Y|D) = \frac{P(D|Y)P(Y)}{P(D)} = \frac{0.04(0.3)}{0.037} = 0.3243$$

$$P(Z|D) = \frac{P(D|Z)P(Z)}{P(D)} = \frac{0.05(0.2)}{0.037} = 0.2703$$

Note that (of course) $P(X|D) + P(Y|D) + P(Z|D) = 1$. ■

Example

Diagnostics. Congestive heart disease (CHD) is a disorder of the heart associated with thickening of the arterial walls. The standard, non-invasive diagnostic test involves a set time on a treadmill.

Failing the treadmill test is taken as suggestive of congestive heart disease, and further tests are required. However, the test is known to be inaccurate, and misses some true sufferers, and falsely diagnoses some non-sufferers.

Now, let C denote having congestive heart disease, and T denote failing the treadmill test. Suppose that there is 5% incidence of CHD in the population ($P(C) = 0.05$), the test has 99% accuracy for CHD sufferers ($P(T|C) = 0.99$) and the test has a 20% false positive rate ($P(T|\bar{C}) = 0.2$).

We can now compute the probability of CHD if a positive test is seen.

$$\begin{aligned} P(C|T) &= \frac{P(T|C)P(C)}{P(T)} = \frac{P(T|C)P(C)}{P(T|C)P(C) + P(T|\bar{C})P(\bar{C})} \\ &= \frac{0.99(0.05)}{0.99(0.05) + 0.2(0.95)} \approx 0.207 \end{aligned}$$

So, the probability of CHD is small when a treadmill test is failed. This happens because of the small incidence in the population ($P(C)$) and the relatively high false positive probability ($P(T|\bar{C})$). ■

Of course, this information can be represented perfectly well in a probability table

	T	\bar{T}	
C	0.0495	0.0005	0.050
\bar{C}	0.1900	0.7600	0.9500
	0.2395	0.7605	1

$$P(C|T) = \frac{P(T \cap C)}{P(T)} = \frac{0.0495}{0.2395} = 0.207$$

Random Variables and Probability Distributions

Events, Probability and Sets

Random Variables and Probability Distributions

Discrete Random Variables

Continuous Random Variables

Systems and Component Reliability

Jointly Distributed Random Variables

Law of Large Numbers and Central Limit Theorem

Statistics

Example

- *Distinct* pairs in a poker hand. A poker hand consists of a draw of 5 cards from the deck. Any hand is an outcome ω from the sample space S of all hands. The number of pairs depends on the outcome ω . Denote the number of pairs by the random variable X , then

$$X(\omega) \in \{0, 1, 2\}$$

since a hand can have 0, 1, or 2 pairs.

- Darts. Throw one dart at a dartboard. The sample space S consists of all possible points of impact. This is uncountable since it includes all possible points on the board. However, the score $X(\omega)$ is one of the finite set of integers from 1 and 60.



Random Variables and Probability Distributions

Random variables are a fundamental concept in probability and statistics. A random variable is a special kind of function.

A *random variable*, X , on a sample space S is a rule that assigns a numerical value to each outcome of S , or, in other words, a function from S into the set of real numbers.

NOTATION: We use uppercase letters for random variables,

and lower case letters to denote particular values a random variable can take. Thus to say that random variable X takes the specific value x , we write $(X = x)$.

The *range* of the random variable is the collection of values the random variable takes.

As noted above, the expression

$$(X = x)$$

refers to the set of all elements in S assigned the value x by the random variable X . Since a random variable refers to elements in S , we can also refer to

$$P(X = x)$$

or sometimes, more concisely, $P(x)$, the probability that the random variable X takes the value x .

When the range of a random variable can be counted, we have a *discrete random variable*, otherwise we have a *continuous random variable* (cf. data types).

Discrete Random Variables

The *probability distribution* for a discrete random variable X is the collection of probabilities assigned by the random variable to its range, and can be represented by a formula, a table, or a graph, each of which provides the probabilities corresponding to each x .

As x varies across the range of the random variable X , if $P(X = x)$ can be described as a function of x (and perhaps depending on other values), then this function is called the *probability mass function* (PMF) of the random variable X and is denoted by $f_X(x)$.

Example

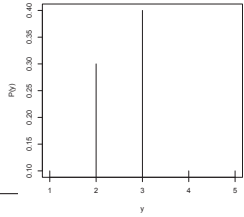
Consider the discrete random variable Y , with range $\{1, 2, \dots, 5\}$ and corresponding probability distribution

y	1	2	3	4	5
$f_Y(y)$	0.1	0.3	0.4	0.1	0.1

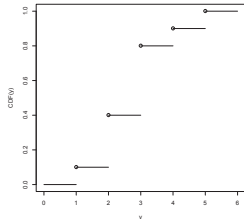
Note that this satisfies the conditions for a discrete probability distribution.

Also note that implicit in such definitions is that probability 0 is assigned to all other values.

PMF



CDF



For a discrete random variable X , with range x_1, x_2, \dots , letting

$$p_i = f_X(x_i) = P(X = x_i) \quad i = 1, 2, \dots$$

the following must be true

1. $p_i \geq 0$
2. $\sum_{i=1}^{\infty} p_i = 1$

The *cumulative distribution function* (CDF), $F_X(x)$, of a discrete random variable X is equal to $P(X \leq x)$, that is, the probability of the random variable taking a value *less than or equal to* x . We can also write

$$F_X(x_j) = P(X \leq x_j) = \sum_{i=1}^j p_i = p_1 + p_2 + \dots + p_j$$

Note that $F_X(x)$ is a non-decreasing function, which must satisfy $F_X(x) = 0$ for $x < x_1$. Also, if x_k is the largest value X can take, then $F_X(x) = 1$ for $x \geq x_k$.

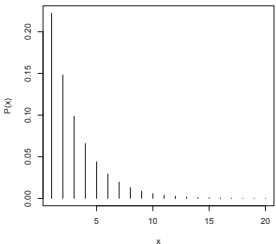
Sometimes, we write $f_X(x; \theta)$ to indicate that the PMF depends on other *parameters* θ .

Example

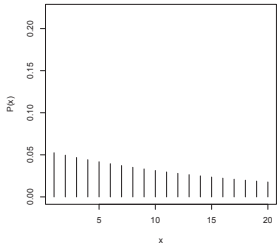
Consider the discrete random variable X with range $\{1, 2, \dots\}$ and PMF

$$f_X(x; \theta) = (1 - \theta)^{x-1} \theta$$

For $\theta = 1/3$ the PMF is



While for $\theta = 1/18$ the PMF is



Note that the vertical scales are the same in both plots.

Theoretical Mean and Variance

Just like summarising a data sample, the probability distribution of a random variable can be usefully described by a small collection of numbers, the most important of which are the (theoretical) mean and (theoretical) variance.

The *expected value*, or *expectation* or *theoretical mean* of a discrete random variable with range $\{x_1, x_2, \dots\}$

$$E(X) = \sum_x x f_X(x)$$

where the summation is over the range of the random variable.

E is the expectation operator. Interpret it as computing the sum of whatever function of X is in the brackets weighted by the corresponding probability. Here $E(X)$ is simply a weighted average of the values in the range.

We often write $\mu = E(X)$ and call μ the population mean.

Properties of expectation:

1. $E(aX + b) = aE(X) + b$ for any $a, b \in \mathbb{R}$
2. $E(X + Y) = E(X) + E(Y)$ for any r.v.s X, Y
3. $E[g(X)] = \sum_x g(x) f_X(x)$ for any function $g(x)$

However, in general, expectation is not multiplicative:

$$E(XY) \neq E(X)E(Y)$$

So, for example, $E(X^2) = E(XX) \neq E(X)E(X) = E(X)^2$.

Example

Consider again the discrete random variable Y , with range $1, 2, \dots, 5$ and corresponding probability distribution

y	1	2	3	4	5
$f_Y(y)$	0.1	0.3	0.4	0.1	0.1

The expectation is

$$E(Y) = \sum_y y f_Y(y) = 1(0.1) + 2(0.3) + 3(0.4) + 4(0.1) + 5(0.1) = 2.8.$$



Note that the population mean need not be part of the range of a discrete random variable (consider the number of children in the average family).

Continuing with the analogy to summarising sample data, we can consider the *theoretical variance*, which measures spread about the mean.

The *theoretical variance* of a discrete random variable is

$$\text{Var}[X] = E[(X - \mu)^2] = \sum_x (x - \mu)^2 f_X(x)$$

where the summation is over the range of the random variable, and μ is the population mean, $\mu = E(X)$.

We often write $\sigma^2 = \text{Var}[X]$, and refer to σ^2 as the *population variance*. This is a weighted average of the squared deviations from the population mean. The population standard deviation σ , is the positive square root of the population variance.

By expanding the definition of the theoretical variance, we can obtain another representation, more suitable for calculations.
For a discrete random variable X

$$\begin{aligned}
 \text{Var}[X] &= E[(X - \mu)^2] = \sum_x (x - \mu)^2 f_X(x) \\
 &= \sum_x (x^2 - 2\mu x + \mu^2) f_X(x) \\
 &= \sum_x x^2 f_X(x) - 2\mu \sum_x x f_X(x) + \mu^2 \sum_x f_X(x) \\
 &= E(X^2) - 2\mu E(X) + \mu^2 \\
 &= E(X^2) - 2\mu^2 + \mu^2 \\
 &= E(X^2) - E(X)^2
 \end{aligned}$$

We always have $\text{Var}[X] \geq 0$. Hence $E(X^2) \geq E(X)^2$.

Example

Consider the discrete random variable Y , with range $1, 2, \dots, 5$ and corresponding probability distribution

y	1	2	3	4	5
$f_Y(y)$	0.1	0.3	0.4	0.1	0.1

Earlier, we saw that $\mu = E(Y) = 2.8$. For the population variance we have (from the definition)

$$\begin{aligned}
 \text{Var}(Y) &= \sum_y (y - \mu)^2 f_Y(y) \\
 &= (1 - 2.8)^2(0.1) + (2 - 2.8)^2(0.3) + (3 - 2.8)^2(0.4) \\
 &\quad + (4 - 2.8)^2(0.1) + (5 - 2.8)^2(0.1) \\
 &= 1.16
 \end{aligned}$$

The population standard deviation is $\sqrt{1.16} \approx 1.08$.

Properties of variance:

1. $\text{Var}(aX + b) = a^2 \text{Var}(X)$ for any $a, b \in \mathbb{R}$

Proof:

$$\begin{aligned}
 \text{Var}(aX + b) &= E[(aX + b - E(aX + b))^2] \\
 &= E[(aX + b - aE(X) - b)^2] \\
 &= a^2 E[(X - E(X))^2] = a^2 \text{Var}(X)
 \end{aligned}$$

■

2. $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ only if the r.v.s X, Y are *uncorrelated* (more on this later)

Important discrete distributions

There are a number of discrete distributions that are widely applicable in practice.

Discrete Uniform Distribution

Appropriate for situations in which an experiment has k equally likely outcomes, denoted by the integers $1, 2, \dots, k$. Archetypal examples include coin tossing and throwing dice. The PMF of the discrete uniform distribution is

$$f_X(x) = \frac{1}{k} \quad \text{for } x = 1, 2, \dots, k$$

The CDF has a particularly simple form

$$F_X(x) = \sum_{i=1}^x \frac{1}{k} = \frac{x}{k} \quad \text{for } x = 1, 2, \dots, k$$

The mean of the discrete uniform distribution is

$$\begin{aligned} E(X) &= \sum_x x f_X(x) = \sum_{x=1}^k \frac{x}{k} \\ &= \frac{1}{k} \sum_{x=1}^k x = \frac{1}{k} \frac{k(k+1)}{2} \\ &= \frac{k+1}{2} \end{aligned}$$

And the variance follows from the computational formula $\text{Var}[X] = E(X^2) - E(X)^2$.

$$\begin{aligned} E(X^2) &= \sum_x x^2 f_X(x) = \sum_{x=1}^k \frac{x^2}{k} \\ &= \frac{1}{k} \sum_{x=1}^k x^2 \end{aligned}$$

Example

Consider throwing a fair die. The random variable X counts the number of spots on the face that falls up. Then we have

$$E(X) = \frac{k+1}{2} = 7/2 = 3.5 \quad \text{Var}(X) = \frac{k^2-1}{12} = \frac{36-1}{12} \approx 2.29$$



Now, we have the result that

$$\sum_{x=1}^k x^2 = \frac{k(k+1)(2k+1)}{6}$$

giving $E(X^2) = (k+1)(2k+1)/6$. Substituting this into the formula for the variance gives

$$\text{Var}(X) = \frac{(k+1)(2k+1)}{6} - \left(\frac{k+1}{2}\right)^2$$

Further manipulation yields

$$\text{Var}(X) = \frac{k^2-1}{12}$$

Binomial Distribution

Many experiments result in two mutually exclusive outcomes. For example, components leaving a production line are either defective or not. If a series of n such components are constructed, such that the manufacture of each can be treated as identical and independent, we will often be interested in the *number* of defectives in the production run.

This scenario is called a *binomial experiment* and the key ingredients are

- ▶ dichotomous outcomes, generically *success* and *failure*
- ▶ n identical trials
- ▶ independence between trials
- ▶ constant probability of success, p (failure probability $q = 1 - p$)
- ▶ X counts the number of successes

Denote a success in a binomial experiment as S and a failure as F . The sample space consists of n -tuples involving S and F . A representative outcome may be

$$\overbrace{SSFFFS \dots FS}^n$$

where the letter in the i th position refers to the i th trial. A typical outcome consisting of x successes is described by the event $X = x$. For example, the outcome

$$\overbrace{SSS \dots S}^x \overbrace{FFF \dots F}^{n-x}$$

is the intersection of n independent trials: x success, and $n - x$ failures. Hence, its probability is

$$\overbrace{ppp \dots p}^x \overbrace{(1-p)(1-p)(1-p) \dots (1-p)}^{n-x} = p^x(1-p)^{n-x}$$

It follows that the PMF of X is given by

$$\begin{aligned} f_X(x; p, n) &= P(X = x) \\ &= P[(S_1 \cap \dots \cap S_x \cap F_{x+1} \cap \dots \cap F_n) \\ &\quad \cup (S_1 \cap \dots \cap S_{x-1} \cap F_x \cap S_{x+1} \cap F_{x+2} \cap \dots \cap F_n) \\ &\quad \cup \dots \cup (F_1 \cap \dots \cap F_{n-x} \cap S_{n-x+1} \cap \dots \cap S_n)] \\ &= \binom{n}{x} p^x (1-p)^{n-x} \end{aligned}$$

for $x = 0, 1, 2, \dots, n$, and zero otherwise. We write

$$X \sim \text{Bin}(n, p)$$

to denote that the random variable X is binomial with parameters n (the number of trials) and p (the probability of success).

Every other outcome described by the event $X = x$ is simply a rearrangement of the x successes and $n - x$ failures and is assigned the same probability. A standard combinatoric result

states that the number of distinct arrangements of the x successes and $n - x$ failures is

$${}^nC_x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

where $n! = n(n-1)(n-2) \dots (2)1$, and $0! = 1$.

Example

A fair coin is tossed 6 times; call heads success. This is a binomial experiment with $n = 6$ and $p = q = 1/2$.

a). The probability that we get exactly 2 heads is

$$\begin{aligned} f_X(2; n = 6, p = 1/2) &= \binom{6}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^4 \\ &= \frac{6!}{2!(6-2)!} \left(\frac{1}{4}\right) \left(\frac{1}{16}\right) \approx 0.23 \end{aligned}$$

b). Let E denote the event that we get at least 4 heads (i.e. 4, 5, or 6). Then

$$\begin{aligned} P(E) &= \sum_{x \in \{4, 5, 6\}} f_X(x; n = 6, p = 1/2) \\ &= \binom{6}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2 + \binom{6}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right) + \binom{6}{6} \left(\frac{1}{2}\right)^6 \left(\frac{1}{2}\right)^0 \end{aligned}$$

$$P(E) = \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{11}{32} \approx 0.34$$

c). The probability of getting no heads (ie. all failures) is

$$(1-p)^6 = \left(\frac{1}{2}\right)^6 = \frac{1}{64}$$

so the probability of *at least* one head is

$$1 - (1-p)^6 = 1 - \frac{1}{64} = \frac{63}{64} \approx 0.98$$

■

Now we consider the expected value of the binomial distribution. First, recall the binomial theorem

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

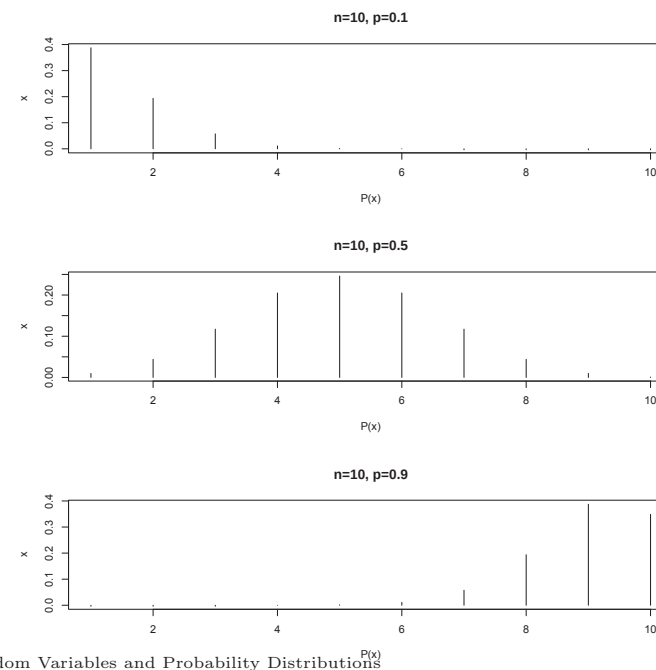
Now, we seek

$$\begin{aligned} E(X) &= \sum_x x f_X(x) \\ &= \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^n x \binom{n}{x} p^x (1-p)^{n-x} \quad \text{change index for } x=0 \end{aligned}$$

Recall the combinatoric identity

$$x \binom{n}{x} = n \binom{n-1}{x-1}$$

Example



Then

$$E(X) = np \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x}$$

Now, let $y = x - 1$ (and so $x = y + 1$)

$$= np \underbrace{\sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y}}_{\text{Bin}(n-1,p)} = np [(p + 1 - p)^{n-1}]$$

which follows by comparison with the binomial theorem. Thus

$$E(X) = np [(p + 1 - p)^{n-1}] = np.$$

Computing the variance of the binomial distribution is more involved. We will derive it later, when we utilise other properties of expectation, and the character of binomial experiments.

For the binomial random variable $X \sim \text{Bin}(n, p)$, we have

$$\sigma^2 = \text{var}[X] = np(1 - p)$$

Example

6 tosses of a fair coin; success refers to heads.

$$E(X) = 3 \quad \text{Var}[X] = 3/2$$



The Geometric Distribution

A different question we might ask when observing a sequence of identical and independent dichotomous trials is “how many trials are required before the first success?” If p is the probability of a success, then to achieve a first success at trial x , we must have observed $x - 1$ failures. The probability of this event is

$$\begin{aligned} f_X(x; p) &= P(F_1 \cap \dots \cap F_{x-1} \cap S_x) \\ &= (1 - p)^{x-1} p \quad \text{for } x = 1, 2, \dots \end{aligned}$$

The random variable X that counts the number of trials until the first success has a *geometric distribution*, $X \sim \text{Geo}(p)$.

Note that

$$\sum_x f_X(x; p) = \sum_{x=1}^{\infty} (1 - p)^{x-1} p = \frac{p}{1 - (1 - p)} = 1$$

Example

Noisy communications. Suppose that the probability that a bit transmitted along a digital communication channel is received in error is 0.05, and that the transmission of bits is independent. What is the probability of a single error in the next 8 bits? What is the expected number of errors, and the theoretical variance?

Let X denote the number of bits in error in the next 8 bits transmitted.

$$P(X = 1; n = 8, p = 0.05) = \binom{8}{1} 0.05^1 0.95^7 \approx 0.279$$

$$E(X) = np = 8 \times 0.05 = 0.4$$

$$\text{Var}[X] = np(1 - p) = 8 \times 0.05 \times 0.95 = 0.38$$



The expected value of a geometric distribution is $E(X) = 1/p$, and the variance is $\text{Var}(X) = (1 - p)/p^2$. These are involved to obtain, so we will derive the expected value for a specific example in a tutorial.

Example

Consider throwing a fair coin.

(a). What is the probability that the first head occurs on the third throw?

$$f_X(3; p = 1/2) = \left(1 - \frac{1}{2}\right)^2 \frac{1}{2} = \left(\frac{1}{4}\right) \frac{1}{2} = \frac{1}{8}$$

(b) What is the expected number of throws before observing a head?

$$E(X) = \frac{1}{p} = \frac{1}{1/2} = 2$$



The Poisson Distribution

Many physical problems are concerned with events occurring independently of one another in time and space (or any other medium).

Example

Counts measured by a Geiger counter in an 8-minute interval.
Number of aircraft accidents in a set time.
Distribution of non-contagious disease in a geographical region. ■

A discrete random variable X is said to have a *Poisson* distribution with parameter $\lambda > 0$ if it has PMF

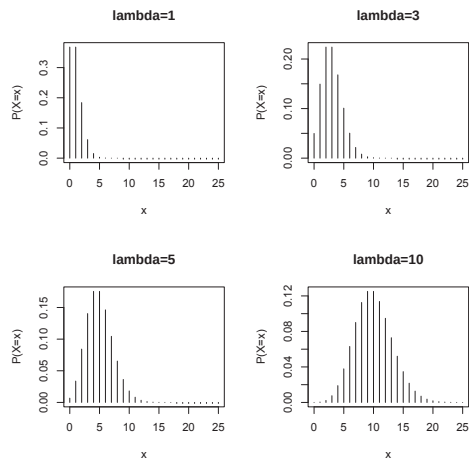
$$f_X(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

and zero otherwise. We write $X \sim \text{Poisson}(\lambda)$.

This distribution often provides a good model for the probability distribution of the number of events that occur infrequently in space and time, and the parameter λ is associated with the mean number of events per unit time.

Example

PMFs for Poisson distributions for a selection of values of λ .



Clearly the PMF of the Poisson distribution assigns non-negative probabilities. If we recall the series expansion of e^λ

$$\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^\lambda$$

then

$$\begin{aligned} \sum_x f_X(x; \lambda) &= \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\ &= e^{-\lambda} e^\lambda = 1 \end{aligned}$$

To compute the mean of a Poisson distribution, again recall that $e^\lambda = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$

$$\begin{aligned} E(X) &= \sum_x x f_X(x; \lambda) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} \\ &= e^{-\lambda} \left[0 + \lambda + \frac{2\lambda^2}{2!} + \frac{3\lambda^3}{3!} + \dots \right] \\ &= \lambda e^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2!} + \dots \right] \\ &= \lambda e^{-\lambda} e^\lambda = \lambda \end{aligned}$$

So the expected value of the Poisson distribution is λ .

To compute the variance of the Poisson distribution, notice that

$$\begin{aligned} E[X(X-1)] - E(X)[E(X)-1] &= E(X^2) - E(X) - [E(X)^2 - E(X)] \\ &= \text{Var}(X) \end{aligned}$$

We then have

$$\begin{aligned} E[X(X-1)] &= \sum_x x(x-1)f_X(x) = \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \left[0 + 0 + \lambda^2 + \lambda^3 + \frac{\lambda^4}{2!} + \dots \right] \\ &= \lambda^2 e^{-\lambda} \left[1 + \lambda + \frac{\lambda^2}{2!} + \dots \right] \\ &= \lambda^2 e^{-\lambda} e^{\lambda} = \lambda^2 \end{aligned}$$

So the variance is

$$\text{Var}(X) = E[X(X-1)] - E(X)[E(X)-1] = \lambda^2 - \lambda(\lambda-1) = \lambda$$

(B). What is the probability of at least two accidents in a 1 year period?

Let Y denote the number of accidents in a year,
 $Y \sim \text{Poisson}(2)$. So,

$$f_Y(y) = \begin{cases} \frac{e^{-2}2^y}{y!} & y = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} P(Y > 1; \lambda = 2) &= 1 - P(y \leq 1; \lambda = 2) \\ &= 1 - \{f_Y(0; \lambda = 2) + f_Y(1; \lambda = 2)\} \\ &= 1 - \left\{ \frac{e^{-2}}{0!} + \frac{e^{-2}2}{1!} \right\} \\ &= 1 - 3e^{-2} \approx 0.594 \end{aligned}$$



Example

Suppose accidents occur following a Poisson distribution at a rate of 2 per year.

(A). What is the probability of more than one accident in a six-month period?

Let X denote the number of accidents in six months,
 $X \sim \text{Poisson}(2 \cdot \frac{1}{2})$.

So,

$$f_X(x; \lambda = 1) = \begin{cases} \frac{e^{-1}1^x}{x!} & x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} P(X > 1; \lambda = 1) &= 1 - P(X \leq 1; \lambda = 1) \\ &= 1 - \{f_X(0; \lambda = 1) + f_X(1; \lambda = 1)\} \\ &= 1 - \left\{ \frac{e^{-1}}{0!} + \frac{e^{-1}}{1!} \right\} \\ &= 1 - 2e^{-1} \approx 0.264 \end{aligned}$$

The Poisson distribution can be thought of as an approximation for a binomial random variable with parameters n and p when n is large and p is small enough to make np small.

Suppose $X \sim \text{Bin}(n, p)$ and let $\lambda = np$. Then

$$\begin{aligned} f_X(x) &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n(n-1)(n-2)\dots(n-x+1)}{n^x} \frac{\lambda^x (1-\lambda/n)^n}{x! (1-\lambda/n)^x} \end{aligned}$$

For large n and moderate λ

$$\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda} \quad \frac{n(n-1)(n-2)\dots(n-x+1)}{n^x} \approx 1$$

and

$$\left(1 - \frac{\lambda}{n}\right)^x \approx 1$$

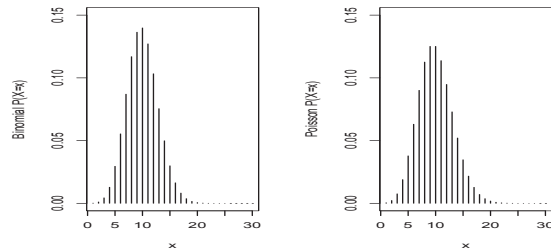
Hence, for large n and moderate p

$$f_X(x) \approx \frac{e^{-\lambda} \lambda^x}{x!}$$

Thus, with appropriate n and p , the number of successes is approximately Poisson, with parameter $\lambda = np$.

Example

Binomial with $n = 50$ and $p = 0.05$, and Poisson approximation.



In this case, X is said to be a continuous random variable. The function f_X is called the *probability density function* (PDF) of X , and (in analogy to the discrete case), it must satisfy

- ▶ $f_X(x) \geq 0$
- ▶ $\int_{-\infty}^{\infty} f_X(x) dx = 1$

That is, X is non-negative and the total area under the graph is 1. Note that the PDF can take values > 1 .

For any interval A we have

$$P(X \in A) = \int_A f_X(x) dx$$

Continuous Probability Distributions

Suppose that X is a random variable on a sample space S that is a continuum, such as an interval of the real line.

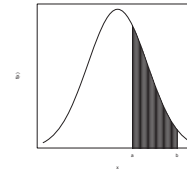
Example

X = time to failure of an engine component

X = aircraft wing span



The set $\{a \leq X \leq b\}$ is an event in S and so the probability of $\{a \leq X \leq b\}$ is defined. Now, assume that there is a function $f_X : \mathbb{R} \rightarrow \mathbb{R}$ such that $P(a \leq X \leq b)$ is defined to be the area under the curve $f_X(x)$ between $x = a$ and $x = b$.



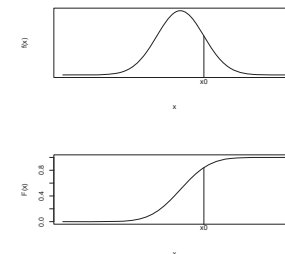
The CDF, F_X , of a continuous random variable X is defined exactly as in the discrete case:

$$F_X(a) = P(X \leq a)$$

If X has PDF f_X then

$$F_X(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f_X(x) dx$$

Note that F_X is monotonically non-decreasing, $F_X(a) \leq F_X(b)$ whenever $a < b$, and we must also have $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.



For any continuous random variable X , we have $P(X = x) = 0$ for any real value x . This is necessary to ensure that F_X is continuous. Practically speaking, the fact that continuous random variables have zero probability for discrete points is of little concern. Consider measuring daily rainfall. What is the probability that we observe a rainfall measurement of *exactly* 2.193 cm? It is unlikely we would ever observe this exact value.

The relationship between the distribution function and the PDF is

$$f_X(x) = \frac{dF_X(x)}{dx}$$

whenever the derivative exists.

We will often be interested in evaluating probabilities of the form $P(a < X \leq b)$:

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F_X(b) - F_X(a) = \int_a^b f_X(x) dx \end{aligned}$$

If b is sufficiently close to a ,

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx \approx f_X(a)(b - a)$$

Note that $f_X(x) \geq 0$ since $F_X(x)$ is non-decreasing.

Example

Given

$$f_X(x) = \begin{cases} cx^2 & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Find the value of c for which f_X is a valid density, obtain the CDF, and compute $P(1 < X \leq 2)$.

We require

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

So

$$\int_0^2 cx^2 dx = \left[\frac{cx^3}{3} \right]_0^2 = \left(\frac{8}{3} \right) c$$

Thus, $\frac{8c}{3} = 1$, so $c = 3/8$.

For $0 \leq x \leq 2$, the CDF is

$$F_X(u) = \int_0^u \frac{3x^2}{8} dx = \frac{3}{8} \int x^2 dx = \frac{3}{8} \left[\frac{x^3}{3} \right]_0^u = \frac{u^3}{8}$$

The complete form of the CDF is thus

$$F_X(u) = \begin{cases} 0 & u < 0 \\ \frac{u^3}{8} & 0 \leq u \leq 2 \\ 1 & u > 2 \end{cases}$$

Now

$$P(1 < X \leq 2) = F_X(2) - F_X(1) = \frac{2^3}{8} - \frac{1^3}{8} = \frac{7}{8}$$

Of course, this could be obtained directly from the PDF

$$\int_1^2 \frac{3}{8} x^2 dx = \frac{3}{8} \left[\frac{x^3}{3} \right]_1^2 = \frac{1}{8} (8 - 1) = \frac{7}{8}$$

Theoretical mean and variance

We can define the expected value and variance of a continuous random variable X in direct analogy to the discrete case.

The *expected value* (or *expectation* or *theoretical mean*) of a continuous random variable X is

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

This can be viewed as a weighted average. Again, we will sometimes write $\mu = E(X)$, and refer to this as the population mean.

Similarly, we have the theoretical variance

The *theoretical variance* of a continuous random variable X is

$$\text{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$$

Again, we will sometimes write $\sigma^2 = \text{Var}(X)$, and refer to this as the population variance. Note that the variance must be non-negative.

Here also,

$$\begin{aligned} \text{Var}(X) &= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f_X(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f_X(x) dx - 2\mu \int_{-\infty}^{\infty} x f_X(x) dx + \mu^2 \int_{-\infty}^{\infty} f_X(x) dx \\ &= E(X^2) - E(X)^2 \geq 0 \end{aligned}$$

Example

Continuing the previous example, the random variable X has PDF

$$f(x) = \begin{cases} \frac{3x^2}{8} & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Its expectation is

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^2 x \frac{3x^2}{8} dx \\ &= \frac{3}{8} \int_0^2 x^3 dx = \frac{3}{8} \left[\frac{x^4}{4} \right]_0^2 \\ &= \frac{3(4)}{8} = \frac{3}{2} \end{aligned}$$

Example

Continuing the previous example, the random variable X has PDF

$$f_X(x) = \begin{cases} \frac{3x^2}{8} & 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Then $\text{Var}(X) = E(X^2) - E(X)^2$, and

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx \\ &= \int_0^2 x^2 \frac{3x^2}{8} dx \\ &= \frac{3}{8} \int_0^2 x^4 dx = \frac{3}{8} \left[\frac{x^5}{5} \right]_0^2 \\ &= \frac{3}{8} \frac{32}{5} = \frac{12}{5} \implies \text{Var}(X) = 12/5 - (3/2)^2 = 3/20 \end{aligned}$$

Continuous Uniform Distribution

Suppose we have a continuous random variable X that is equally likely to take a value in any fixed size interval within the range $[a, b]$. X has PDF

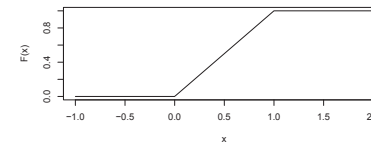
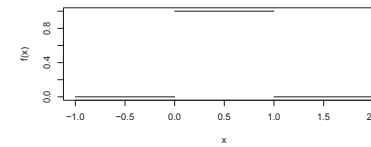
$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

and we write $X \sim \text{Unif}(a, b)$. The CDF of X is

$$F_X(u) = \begin{cases} 0 & u < a \\ \int_a^u \frac{1}{b-a} dx = \frac{u-a}{b-a} & a \leq u \leq b \\ 1 & u > b \end{cases}$$

Example

$X \sim \text{Unif}(0, 1)$



Note in this case for $0 \leq u \leq 1$, $F_X(u) = u$.

■

The mean of the continuous uniform distribution is

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_a^b \frac{x}{b-a} dx \\ &= \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b \\ &= \frac{1}{b-a} \left(\frac{b^2}{2} - \frac{a^2}{2} \right) = \frac{1}{b-a} \frac{(b+a)(b-a)}{2} \\ &= \frac{a+b}{2} \end{aligned}$$

The variance of the continuous uniform distribution is $\text{Var}(X) = E(X^2) - E(X)^2$, and

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_a^b \frac{x^2}{b-a} dx \\ &= \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b \\ &= \frac{1}{b-a} \left(\frac{b^3}{3} - \frac{a^3}{3} \right) \\ &= \frac{(b-a)(b^2 + ab + a^2)}{3(b-a)} = \frac{b^2 + ab + a^2}{3} \end{aligned}$$

And hence

$$\text{Var}(X) = \frac{b^2 + ab + a^2}{3} - \left[\frac{a+b}{2} \right]^2 = \frac{(b-a)^2}{12}$$

Example

Suppose $X_1 \sim \text{Unif}(0, 1)$ and $X_2 \sim \text{Unif}(0, 1)$, and X_1 and X_2 are independent.

Compute the probability that $P(X_1 \leq 1/2 \cap X_2 \leq 1/2)$.

Using the CDF obtained earlier, we have

$$P(X_i \leq 1/2) = F_X(1/2) = 1/2$$

for $i = 1, 2$. Now, since X_1 and X_2 are independent, we have

$$P(X_1 \leq 1/2 \cap X_2 \leq 1/2) = F_X(1/2)F_X(1/2) = 1/4$$

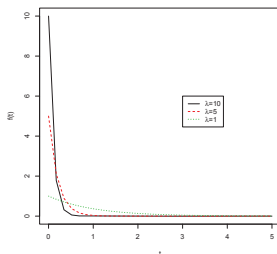
■

To verify that this is a valid probability density function, note that the density is non-negative, and

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x; \lambda) dx &= \int_0^{\infty} \lambda e^{-\lambda x} dx \\ &= \left[-e^{-\lambda x} \right]_0^{\infty} = (0 - (-1)) = 1 \end{aligned}$$

Example

Exponential PDFs for a selection of values of λ .



Exponential distribution

The exponential distribution is often useful for modelling the length of the lifetime of electronic components. A random variable X has an exponential distribution with parameter $\lambda > 0$ if its PDF is

$$f_X(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The CDF for the exponential distribution is

$$F_X(u; \lambda) = \begin{cases} 0 & u \leq 0 \\ \int_0^u \lambda e^{-\lambda x} dx = [-e^{-\lambda x}]_0^u = 1 - e^{-\lambda u} & u > 0 \end{cases}$$

Some components, particularly electronics, do not wear. For such a component, the amount of time it has operated does not impact the probability that it will continue to function. Specifically, we are interested in the probability that a component operates for at least b more time units, given that it has already operated for a time units.

Now suppose the random lifetime of such a component, T , follows an exponential distribution. Then

$$\begin{aligned} P(T > a + b | T > a) &= \frac{P(T > a + b \cap T > a)}{P(T > a)} \\ &= \frac{P(T > a + b)}{P(T > a)} \end{aligned}$$

since $(T > a + b) \subseteq (T > a)$. We previously determined the distribution function of the exponential distribution

$$F_T(u; \lambda) = 1 - e^{-\lambda u}, \quad \text{for } u > 0$$

$$\begin{aligned}
 P(T > a + b | T > a) &= \frac{1 - F_T(a + b; \lambda)}{1 - F_T(a; \lambda)} \\
 &= \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} \\
 &= 1 - F_T(b) = P(T > b)
 \end{aligned}$$

The exponential is the only continuous distribution with this property, which is known as the *memoryless* property.

Example

Suppose that user queries at a call centre have random lengths (in minutes) with exponential density

$$f_T(t; \lambda = 1/5) = \frac{1}{5} e^{-\frac{t}{5}}$$

for $t > 0$ and $f_T(t; \lambda = 1/5) = 0$ otherwise.

What is the probability that a call lasts less than two minutes?

We require $P(T < 2)$.

$$P(T < 2) = F_T(2) = 1 - e^{-2\lambda} = 1 - e^{-\frac{2}{5}} \approx 0.3297.$$

■

To obtain the expected value of an exponential distribution we require

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx$$

Using integration by parts, with

$$u = x \quad \frac{du}{dx} = 1 \quad \frac{dv}{dx} = \lambda e^{-\lambda x} \quad v = -e^{-\lambda x}$$

Then

$$E(X) = -x e^{-\lambda x} \Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx$$

The first term goes to zero (since the exponential dominates), so

$$\begin{aligned}
 E(X) &= 0 + \left[\frac{e^{-\lambda x}}{-\lambda} \right]_0^{\infty} \\
 &= [(0) - (-1/\lambda)] \\
 &= 1/\lambda
 \end{aligned}$$

For the variance, we use

$$\text{Var}(X) = E(X^2) - E(X)^2 = E(X^2) - (1/\lambda)^2$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx$$

Again, we use integration by parts, with

$$u = x^2 \quad \frac{du}{dx} = 2x \quad \frac{dv}{dx} = \lambda e^{-\lambda x} \quad v = -e^{-\lambda x}$$

so we have

$$\begin{aligned} E(X^2) &= -x^2 e^{-\lambda x} \Big|_0^{\infty} + 2 \int_0^{\infty} x e^{-\lambda x} dx \\ &= 0 + 2 \int_0^{\infty} x e^{-\lambda x} dx \end{aligned}$$

We could proceed by using integration by parts again, but note that we have already determined that

$$\int_0^{\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

thus

$$\int_0^{\infty} x e^{-\lambda x} dx = \frac{1}{\lambda^2}$$

and hence

$$E(X^2) = \frac{2}{\lambda^2}$$

Substituting this into the formula for the variance, we have

$$\text{Var}(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Change of Variable: One function of one RV

Let X be a continuous random variable and $g(x)$ a strictly monotonic function (such that we have a one-to-one correspondence). The random variable $Y = g(X)$ has PDF

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| \quad \text{where } x = g^{-1}(y).$$

Outline proof: If $g(x)$ is an increasing function, then

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq g(x)) = P(X \leq x) = F_X(x).$$

Differentiating both sides w.r.t. y gives:

$$f_Y(y) = \frac{dF_X(x)}{dy} = \frac{dF_X(x)}{dx} \frac{dx}{dy} = f_X(x) \frac{dx}{dy},$$

and $\frac{dx}{dy} = \frac{d}{dy} g^{-1}(x) > 0$, so the result holds.

If $g(x)$ is an decreasing function, we end up with

$$f_Y(y) = -f_X(x) \frac{dx}{dy},$$

and $\frac{dx}{dy} < 0$, so the result holds.

Example

Let $X \sim \text{Exp}(\lambda)$ and $g(x) = cx$, where $c > 0$. The random variable $Y = cX$ has PDF

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = f_X\left(\frac{y}{c}\right) \left| \frac{d}{dy} \frac{y}{c} \right| = \lambda e^{-\lambda \frac{y}{c}} \frac{1}{c} = \frac{\lambda}{c} e^{-\frac{\lambda}{c} y},$$

from which we can deduce that $Y \sim \text{Exp}(\lambda/c)$.

Let now $X \sim \text{Exp}(\lambda)$ and $g(x) = cx$, where $c < 0$. By application of the result, the random variable $Y = cX$ has PDF

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = f_X\left(\frac{y}{c}\right) \left| \frac{d}{dy} \frac{y}{c} \right| = \lambda e^{-\lambda \frac{y}{c}} \left| \frac{1}{c} \right| = \left| \frac{\lambda}{c} \right| e^{-\frac{\lambda}{c} y}.$$

Note that Y is always negative.

Proof:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(cX \leq y) = P(-dX \leq y) = P(dX \geq -y) \\ &= 1 - P(dX \leq -y) = 1 - P(X \leq -y/d) \end{aligned}$$

$$\begin{aligned} f_Y(y) &= \frac{d F_Y(y)}{dy} = -f_X\left(\frac{-y}{d}\right) \left(\frac{-1}{d}\right) = \frac{\lambda}{d} e^{\lambda \frac{y}{d}} \\ &= -\frac{\lambda}{c} e^{-\lambda \frac{y}{c}} = \left| \frac{\lambda}{c} \right| e^{-\frac{\lambda}{c} y} \end{aligned}$$

■

Is it a valid PDF? Let us show that

$$I = \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$$

Proof:

$$\begin{aligned} I &= 2 \int_0^{+\infty} e^{-\frac{x^2}{2}} dx = 2J \\ J^2 &= \int_0^{+\infty} e^{-\frac{x^2}{2}} dx \int_0^{+\infty} e^{-\frac{y^2}{2}} dy \\ &= \int_0^{+\infty} \int_0^{+\infty} e^{-\frac{x^2+y^2}{2}} dx dy \\ &= \int_{\theta=0}^{\frac{\pi}{2}} \int_{r=0}^{\infty} e^{-\frac{r^2}{2}} r dr d\theta = \frac{\pi}{2} \end{aligned}$$

■

The normal Distribution

A random variable X follows a normal distribution with parameters μ and σ^2 if it has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We denote this by $X \sim N(\mu, \sigma^2)$. The simplest case, known as the standard normal, is when $\mu = 0$ and $\sigma^2 = 1$. We commonly use Z to denote standard normal random variables, and $\phi(z)$ and $\Phi(z)$ for their PDF and CDF, respectively:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad \Phi(z) = \int_{-\infty}^z \phi(u) du$$

A standard normal random variable has mean 0 and variance 1.

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} z \phi(z) dz = \int_{-\infty}^{\infty} \frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty} = 0 \end{aligned}$$

$$\begin{aligned} E(Z^2) &= \int_{-\infty}^{\infty} z^2 \phi(z) dz = \int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= - \int_{-\infty}^{\infty} \frac{z}{\sqrt{2\pi}} \left(e^{-\frac{z^2}{2}} \right)' dz \\ &= - \left[\frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1, \end{aligned}$$

so $\text{Var}(Z) = 1$.

If $Z \sim N(0, 1)$ and $X = \sigma Z + \mu$, then $X \sim N(\mu, \sigma^2)$.

Apply the change-of-variable formula:

$$f_X(x) = f_Z(z) \left| \frac{dz}{dx} \right| = \phi \left(\frac{x - \mu}{\sigma} \right) \left| \frac{d}{dx} \frac{x - \mu}{\sigma} \right|$$

$$= \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

which we identify as the PDF of $N(\mu, \sigma^2)$.

Since $E(X) = E(\sigma Z + \mu) = \mu$ and $\text{Var}(X) = \text{Var}(\sigma Z + \mu) = \sigma^2$, we can see that the parameters μ and σ^2 are the mean and variance of the normal distribution.

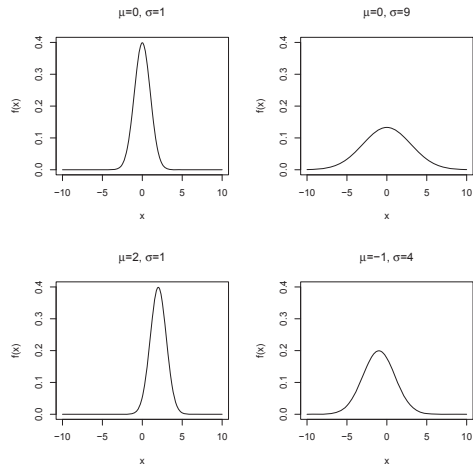
It is often the case that measurements of many random phenomena appear to have been generated from mechanisms that are closely approximated by a normal distribution.

Example

- Height of men (or women).
- Diffusion velocity of molecules in gas.
- PDF of the ground state in a quantum harmonic oscillator.

The normal CDF cannot be obtained in closed form. In order to make probability calculations we can use tables obtained by numerical integration.

Selection of normal distributions. Note that the distribution is symmetric about the mean μ .



Two tables of probabilities are provided, one explicitly prepared for the exam.

THE STANDARD NORMAL DISTRIBUTION FUNCTION

STANDARD NORMAL DENSITY

Entries in table are probabilities p such that $\Phi(z) = p$

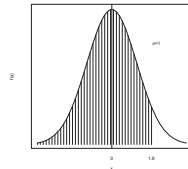
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	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
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

TABLE OF THE STANDARD NORMAL CDF

The table is a tabulation such that $\Phi(z) = p$ for $z \geq 0$. To compute $\Phi(z)$ for $z > 0$, use the table directly.

Example

$$P(Z \leq 1.6) = \Phi(1.6) = 0.9452$$

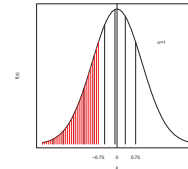


For $z < 0$, we exploit the symmetry of the standard normal about its mean, $\mu = 0$.

$$\Phi(z) + \Phi(-z) = 1 \implies \Phi(-z) = 1 - \Phi(z)$$

Example

$$\begin{aligned} P(Z \leq -0.75) &= 1 - \Phi(0.75) \\ &= 1 - 0.7734 = 0.2266 \end{aligned}$$



The following table is from the formula sheet provided in the exam. Note that it only runs up to the first decimal place - any exam question will be adequately answered by rounding to the nearest position.

y	$\phi(y)$	$\Phi(y)$	y	$\phi(y)$	$\Phi(y)$	y	$\phi(y)$	$\Phi(y)$	y	$\Phi(y)$
0	.399	.5	.9	.266	.816	1.8	.079	.964	2.8	.997
.1	.397	.540	1.0	.242	.841	1.9	.066	.971	3.0	.999
.2	.391	.579	1.1	.218	.864	2.0	.054	.977	0.841	.8
.3	.381	.618	1.2	.194	.885	2.1	.044	.982	1.282	.9
.4	.368	.655	1.3	.171	.903	2.2	.035	.986	1.645	.95
.5	.352	.691	1.4	.150	.919	2.3	.028	.989	1.96	.975
.6	.333	.726	1.5	.130	.933	2.4	.022	.992	2.326	.99
.7	.312	.758	1.6	.111	.945	2.5	.018	.994	2.576	.995
.8	.290	.788	1.7	.094	.955	2.6	.014	.995	3.09	.999

Example

$$\begin{aligned} P(-2 < Z < 1) &= \Phi(1) - \Phi(-2) \\ &= \Phi(1) - (1 - \Phi(2)) \\ &= 0.841 - (1 - 0.977) = 0.818 \end{aligned}$$



How is the standard normal distribution spread around 0?

$$\begin{aligned} P(|Z| < 1) &= P(-1 < Z < 1) = \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) \\ &= 2\Phi(1) - 1 = 0.6826 \end{aligned}$$

$$P(|Z| < 2) = 0.9544$$

$$P(|Z| < 3) = 0.9974.$$

As a rule of thumb, Z is concentrated between -3 and 3 (three-sigma rule).

Working backwards, if $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1),$$

so any normal distribution can be transformed to a standard normal by subtracting the mean μ , and dividing by the standard deviation σ . This transformation is called *standardising*.

Similarly, as a rule of thumb, X is concentrated between $\mu - 3\sigma$ and $\mu + 3\sigma$ (three-sigma rule).

Example

Suppose $X \sim N(3, 5^2)$. Now $Z = (X - 3)/5 \sim N(0, 1)$

a). Compute $P(X < 5)$.

$$P(X < 5) = P\left(\frac{X - 3}{5} < \frac{5 - 3}{5}\right) = P(Z < 0.4) = \Phi(0.4) = 0.655$$

b). Compute $P(0.5 < X < 5.5)$.

$$\begin{aligned} P(0.5 < X < 5.5) &= P\left(\frac{0.5 - 3}{5} < \frac{X - 3}{5} < \frac{5.5 - 3}{5}\right) \\ &= \Phi(0.5) - \Phi(-0.5) = \Phi(0.5) - (1 - \Phi(0.5)) \\ &= 2\Phi(0.5) - 1 = 2(0.6915) - 1 = 0.3830 \end{aligned}$$



Exercise (Exam Question May 2014)

By making use of the standard normal table, compute the following integral

$$\int_{-\infty}^{2.35} \sqrt{\frac{2}{\pi}} e^{-2(u-2)^2} du.$$

Provide your reasoning.

The Chi-Square Distribution

Let $Z \sim N(0, 1)$ and take $U = Z^2$.

$$\begin{aligned} F_U(u) &= P(U \leq u) = P(Z^2 \leq u) \\ &= \begin{cases} 0, & \text{if } u < 0, \\ P(-\sqrt{u} \leq Z \leq \sqrt{u}) = \int_{-\sqrt{u}}^{\sqrt{u}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz, & \text{if } u \geq 0. \end{cases} \end{aligned}$$

Differentiating both sides w.r.t. u gives

$$f_U(u) = \begin{cases} \frac{1}{2\sqrt{u}\sqrt{2\pi}} e^{-\frac{u}{2}} + \frac{1}{2\sqrt{u}\sqrt{2\pi}} e^{-\frac{u}{2}} = \frac{1}{\sqrt{u}\sqrt{2\pi}} e^{-\frac{u}{2}}, & \text{if } u > 0, \\ 0, & \text{if } u \leq 0. \end{cases}$$

U is the square of a standard normal random variable and is called Chi-Square with 1 degree of freedom.

The Log-Normal Distribution

Let $X \sim N(\mu, \sigma^2)$ and take $Y = e^X$.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= \begin{cases} 0, & \text{if } y \leq 0, \\ P(e^X \leq y) = P(X \leq \ln y), & \text{if } y > 0 \end{cases} \end{aligned}$$

with

$$P(X \leq \ln y) = \int_{-\infty}^{\ln y} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Differentiating both sides w.r.t. y gives

$$f_Y(y) = \begin{cases} \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln y - \mu}{\sigma}\right)^2}, & \text{if } y > 0, \\ 0, & \text{if } y \leq 0. \end{cases}$$

$Y = e^X$ is a log-normal random variable, i.e. $X = \ln Y$ is a normal random variable.

Chebyshev's Inequality

If X is a random variable with $\mu = E(X)$ and $\sigma^2 = \text{Var}(X)$, then

$$\forall a > 0, \quad P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

or equivalently

$$\forall a > 0, \quad P(|X - \mu| < a) \geq 1 - \frac{\sigma^2}{a^2}.$$

Valid for any distribution of X .

Slightly more general form:

$$\forall b, \forall a > 0, \quad P(|X - b| \geq a) \leq \frac{1}{a^2} E[(X - b)^2].$$

■

Proof: Let us define the random variable Y , function of the random variable X , as

$$Y = \begin{cases} a^2, & \text{if } |X - b| \geq a, \\ 0, & \text{if } |X - b| < a. \end{cases}$$

We then write

$$E(Y) = a^2 P(|X - b| \geq a) + 0 P(|X - b| < a)$$

Moreover, we can always write $(X - b)^2 \geq Y$.

Hence

$$E[(X - b)^2] \geq E(Y) = a^2 P(|X - b| \geq a)$$

If $E(X) = \mu$ and $\text{Var}(X) = 0$, then $P(X = \mu) = 1$.

Proof: Take $a = \frac{1}{n}$, $n \in N^*$, such that

$$0 \leq P(|X - \mu| \geq \frac{1}{n}) \leq 0, \quad \forall n \in N^*.$$

Hence,

$$P(|X - \mu| > 0) = \lim_{n \rightarrow \infty} P(|X - \mu| \geq \frac{1}{n}) = 0$$

$$P(|X - \mu| \leq 0) = P(|X - \mu| = 0) = P(X = \mu) = 1.$$

■

When a random variable has a variance 0, its PDF is concentrated on a single point, i.e. its mean.

Example

Show that $E(X^2) = 0$ leads to $E(X) = 0$ and $P(X = 0) = 1$.

To show this, recall that $\text{Var}(X) = E(X^2) - E(X)^2 \geq 0$. If $E(X^2) = 0$, $E(X) = 0$ and $\text{Var}(X) = 0$. From previous result, $P(X = \mu) = P(X = 0) = 1$.

Particular case of Chebyshev's inequality:

Assuming $\sigma^2 \neq 0$, set $a = k\sigma$ ($k > 0$) and write

$$\forall k > 0, \quad P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Hence,

$$k = 1, \quad P(|X - \mu| < \sigma) = P(\mu - \sigma < X < \mu + \sigma) \geq 0,$$

$$k = 2, \quad P(|X - \mu| < 2\sigma) = P(\mu - 2\sigma < X < \mu + 2\sigma) \geq \frac{3}{4},$$

$$k = 3, \quad P(|X - \mu| < 3\sigma) = P(\mu - 3\sigma < X < \mu + 3\sigma) \geq \frac{8}{9}.$$

Valid for any probability distribution. When the distribution is known, more accurate results can be obtained, e.g. see Normal distribution.

Systems and Component Reliability

Events, Probability and Sets

Random Variables and Probability Distributions

Systems and Component Reliability

Time-to-failure distributions

Hazard rate functions

Commonly used life distributions

Mean time to failure

Jointly Distributed Random Variables

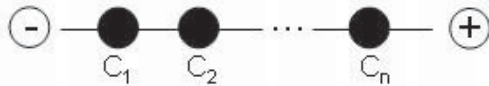
Law of Large Numbers and Central Limit Theorem

Statistics

Systems and Component Reliability

153/321

A *series systems* consisting of n identical components, C_i , can be represented as



In this case, if any component fails, the system fails. Suppose now, that the components operate independently, and let C_i be the event that component C_i fails, for $i = 1, 2, \dots, n$.

Let $P(C_i) = \theta$, then $P(\overline{C}_i) = 1 - \theta$.

Now,

$$\begin{aligned} P(\text{system functions}) &= P(\overline{C}_1 \cap \overline{C}_2 \cap \dots \cap \overline{C}_n) \\ &= P(\overline{C}_1)P(\overline{C}_2) \dots P(\overline{C}_n) \\ &= (1 - \theta)^n \end{aligned}$$

Systems and Component Reliability

155/321

Systems and Component Reliability

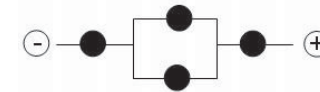
Systems

Consider a system of components put together such that the entire system works only if certain combinations of the components work.

Example

Computer network; car engine; software system; human body.

It is often convenient to represent such a system as a *circuit*.



The system functions if there is a functioning path from $-$ to $+$. It is useful to consider series, parallel and mixed systems separately.

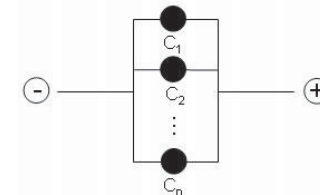
Systems and Component Reliability

154/321

Example

Consider a series system with $n = 3$ and $\theta = 0.1$. The probability that the system functions is $(1 - 0.1)^3 = 0.9^3 = 0.729$. ■

The simplest parallel system consisting of n identical components can be represented as



The system functions if there is a working path from $-$ to $+$. Again, suppose the n components operate independently, and the components all have the same probability of failure $P(C_i) = \theta$. The system fails only if all n components fail.

Systems and Component Reliability

156/321

$$\begin{aligned}
 P(\text{system function}) &= 1 - P(\text{system fails}) \\
 &= 1 - P(C_1 \cap C_2 \cap \dots \cap C_n) \\
 &= 1 - P(C_1)P(C_2) \dots P(C_n) \\
 &= 1 - \theta^n
 \end{aligned}$$

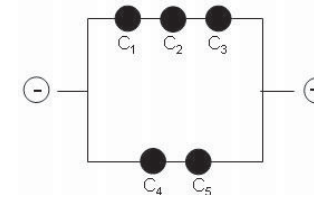
Example

consider a parallel system with $n = 3$ and $\theta = 0.1$. The probability that the system functions is $1 - 0.1^3 = 0.999$. ■

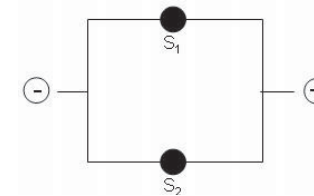
To deal with more complicated cases, it is useful to decompose the system into series and parallel paths.

Example

Consider the following mixed system, where 5 identical components operate independently.



To reason about the system functioning, it is sufficient to consider this representation



Then

$$\begin{aligned}
 P(\text{system function}) &= 1 - P(\text{system fails}) \\
 &= 1 - (P(S_1)P(S_2))
 \end{aligned}$$

and the results derived earlier can be used to compute $P(S_i)$, $i = 1, 2$.

Time-to-failure distributions

Let T denote the *random time to failure* of a unit under study (e.g. gears, semi-conductors). The random variable T is non-negative and has distribution function

$$F_T(t) = P(T \leq t)$$

which is called the *failure time distribution*, with

$$F_T(t) = P(T \leq t) = \int_0^t f_T(u) du$$

where $f_T(t)$ is the *failure time density*.

Since a component either fails or it does not, we also have

$$R_T(t) = P(T > t) = 1 - F_T(t)$$

where $R_T(t)$ is called the *reliability function*. $R_T(t)$ is the probability that a unit does *not* fail in the interval $(0, t]$, or equivalently that the unit is still functioning at time t .

As usual

$$f_T(t) = \frac{dF_T(t)}{dt} = -\frac{dR_T(t)}{dt}$$

Note that

$$f_T(t)\delta t \approx P(t < T \leq t + \delta t),$$

the probability that a unit fails in the short interval $(t + \delta t]$.

We can consider the probability of failure as the unit gets older, that is, the probability that the unit will fail in the short interval $(t + \delta t]$, given that it has survived to time t .

Let A be the event “unit fails in $(t + \delta t]$ ” ($\{t < T \leq t + \delta t\}$).

Let B be the event “unit not failed by time t ” ($\{T > t\}$)

Then,

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)}{P(B)} \quad \text{since } A \subseteq B \\ &\approx \frac{f_T(t)\delta t}{1 - F_T(t)} \quad \text{recall } F_T(t) = P(T \leq t) \end{aligned}$$

Hence

$$P(A|B) \approx \frac{f_T(t)\delta t}{R_T(t)}$$

Regard this conditional probability, $P(A|B)$, as the probability of imminent failure at time t .

The quantity

$$z_T(t) = \frac{f_T(t)}{R_T(t)} \propto P(A|B)$$

is called the *hazard rate* (or *failure rate*) of the unit.

The hazard function is an indicator of the proneness to failure of a unit after a time t has elapsed.

The *cumulative hazard function* is

$$H_T(t) = \int_0^t z_T(u) du$$

and this is related to the reliability function as

$$R_T(t) = e^{-H_T(t)}$$

This follows from standard integration results, since

$$H_T(t) = \int_0^t z_T(u) du = \int_0^t \frac{f_T(t)}{R_T(t)} dt = [-\ln R_T]_0^t = -\ln R_T(t)$$

This material is simply standard probability results, applied in a specific context. The context has introduced new structures, as follows

F Failure time distribution

f Failure time density

R Reliability function

z Hazard rate function

H Cumulative hazard function

Note that F, f, R, z, H give mathematically equivalent descriptions of t in the sense that given any one of these functions the others may be deduced.

Example

Consider units with hazard rate given by

$$z_T(t) = \lambda.$$

Then:

$$H_T(t) = \int_0^t \lambda du = \lambda t \quad \text{cumulative hazard}$$

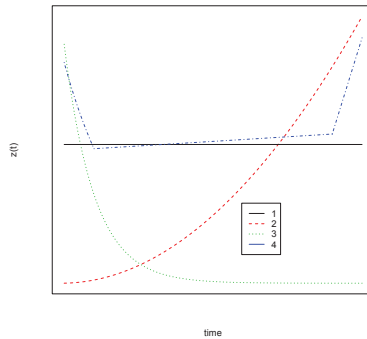
$$R_T(t) = e^{-H_T(t)} = e^{-\lambda t} \quad \text{reliability function}$$

$$f_T(t) = \frac{-dR_T(t)}{dt} = \lambda e^{-\lambda t} \quad \text{failure density}$$

Note that this is the density function of the exponential distribution. Recall the memoryless property of this distribution.

Hazard rate functions

Knowledge about an item's hazard rate often helps us to select the appropriate failure time distribution for the item.



1). Constant Hazard.

Here $z_T(t) = \lambda$. Proneness to failure at any time is constant, and therefore not related to time. This is suitable for components that do *not* age, the primary examples of which are semi-conductor components.

2). Increasing hazard.

If $z_T(t)$ is an increasing function of t , then T is said to have an increasing failure rate. This is appropriate for items that age or wear.

3). Decreasing hazard.

If $z_T(t)$ is a decreasing function of time, then T has a decreasing failure rate. This could happen when a manufacturing process produces low-quality units – many will fail early.

4). Bathtub hazard.

Named for the shape of the hazard function. High “infant mortality”, followed by period of stabilisation (sometimes called the chance failure period), followed by a wear-out period. Possibly good life distribution for humans?

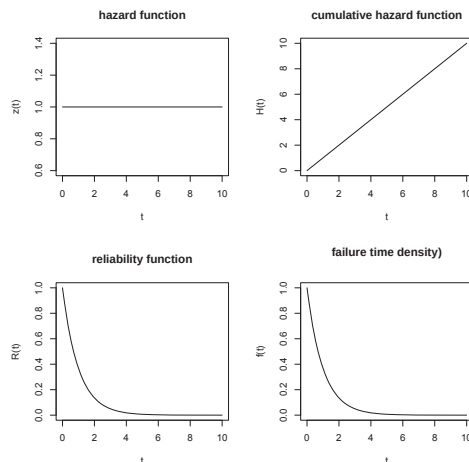
Commonly used life distributions

Exponential distribution

Explored in detail already: PDF leads to constant hazard.

Example

$T \sim \text{Exponential}(1)$



Two-parameter Weibull distribution

A positive random variable T has a Weibull distribution with parameters λ, β (both > 0) if its CDF is

$$F_T(t; \lambda, \beta) = \begin{cases} 1 - e^{-(\lambda t)^\beta} & t > 0 \\ 0 & \text{otherwise} \end{cases}$$

By differentiating, we obtain the PDF:

$$f_T(t; \lambda, \beta) = \begin{cases} \lambda \beta (\lambda t)^{\beta-1} e^{-(\lambda t)^\beta} & t > 0 \\ 0 & \text{otherwise} \end{cases}$$

β is called the *shape* parameter, and λ is called the *scale* parameter.

It can be shown that if $X \sim \text{Expo}(\lambda^\beta)$, $Y = X^{1/\beta} \sim \text{Weibull}(\lambda, \beta)$.

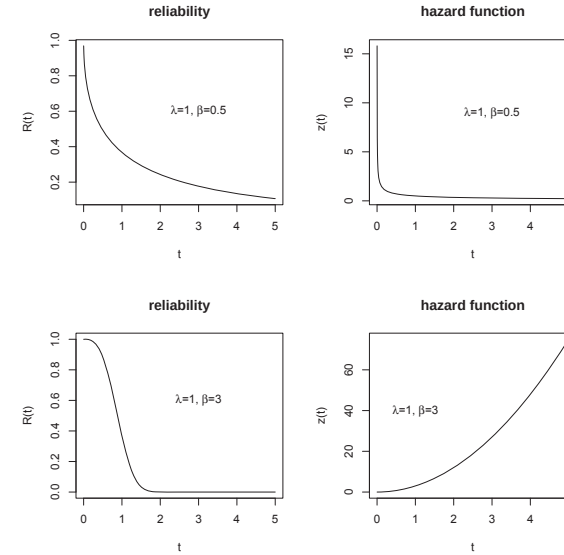
The hazard function is

$$\begin{aligned} z_T(t; \lambda, \beta) &= \frac{f_T(t)}{R_T(t)} = \frac{f_T(t)}{1 - F_T(t)} \\ &= \frac{\lambda\beta(\lambda t)^{\beta-1}e^{-(\lambda t)^\beta}}{1 - (1 - e^{-(\lambda t)^\beta})} = \lambda\beta(\lambda t)^{\beta-1} \end{aligned}$$

The cumulative hazard is thus

$$\begin{aligned} H_T(t; \lambda, \beta) &= \int_0^t z_T(u) du = \int_0^t \lambda\beta(\lambda u)^{\beta-1} du \\ &= \lambda^\beta \beta \left[\frac{u^\beta}{\beta} \right]_0^t = \lambda^\beta t^\beta = (\lambda t)^\beta \end{aligned}$$

This is a very flexible distribution that can be used to describe both decreasing and increasing rate of failure. Some examples of Weibull distributions:



Example

A certain component is known to have a Weibull failure density with $\beta = 2$ and $\lambda = 10^{-3}$. What is the probability that a such a component survives longer than 500 hours?

$$R_T(t; \lambda, \beta) = 1 - F_T(t; \lambda, \beta) = e^{-(\lambda t)^\beta}$$

So

$$\begin{aligned} R_T(500; \lambda = 10^{-3}, \beta = 2) &= e^{-(10^{-3}(500))^2} \\ &= e^{-\left(\frac{1}{2}\right)^2} = e^{-\frac{1}{4}} \\ &\approx 0.7788 \end{aligned}$$

■

Mean time to failure

The mean time to failure (MTTF) of a unit is defined as

$$MTTF = E[T] = \int_0^\infty t f_T(t) dt$$

Since $f_T(t) = -R'_T(t)$

$$MTTF = - \int_0^\infty t R'_T(t) dt$$

Integration by parts gives

$$MTTF = - [t R_T(t)]_0^\infty + \int_0^\infty R_T(t) dt$$

It can be shown that if $MTTF < \infty$, then the first term is zero. Thus, the MTTF is obtained directly from the integral of $R_T(t)$.

Jointly Distributed Random Variables

Events, Probability and Sets

Random Variables and Probability Distributions

Systems and Component Reliability

Jointly Distributed Random Variables

Discrete Random Variables

Continuous Random Variables

Independent Random Variables

Conditional Distributions

Expectation, Variance, Covariance, Correlation

Joint Normal (Gaussian) Distribution

Moments

Sums of random variables

Change of Variables

Example

MTTF: Exponential distribution: $f_T(t) = \lambda e^{-\lambda t}$

$$\begin{aligned} MTTF &= \int_0^{\infty} R_T(t) dt = \int_0^{\infty} e^{-\lambda t} dt \\ &= \left[\frac{e^{-\lambda t}}{-\lambda} \right]_0^{\infty} = \left(0 - \frac{1}{-\lambda} \right) = \frac{1}{\lambda} \end{aligned}$$

Jointly Distributed Random Variables

So far, we have considered only single random variables. It is often useful to make probability statements concerning more than one random variable. We will mostly focus on pairs of random variables, that is, bivariate distributions.

For random variables X and Y , the *joint cumulative distribution function* is defined as

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) \quad -\infty < a, b < \infty.$$

Note that we read a comma as *and*, i.e. \cap .

In practice it can sometimes be difficult to manipulate the joint CDF.

It will be convenient to distinguish between discrete and continuous random variables.

Discrete Random Variables

The PMF of a single discrete random variable X assigns probability to every value x in the range of the random variable. For a pair of random variables X and Y , the sample space S consists of all pairs (x, y) that can be derived from the ranges of X and Y .

Example

Y is tossing a coin. X is throwing a dice. The sample space consists of all pairs (x, y) :

$$S = \left\{ \begin{array}{llllll} (\square, H), & (\square, H), & (\square, H), & (\square, H), & (\square, H), & (\square, H), \\ (\square, T), & (\square, T), & (\square, T), & (\square, T), & (\square, T), & (\square, T) \end{array} \right\}$$

■

The *joint probability mass function* determines how probability is assigned to all pairs of values (x, y) .

For random variables X and Y with sample space S , the joint probability mass function is defined for each pair (x, y) in the sample space as

$$f_{X,Y}(x,y) = P(X = x, Y = y) .$$

Note that this must satisfy the characteristics of discrete probability: the PMF must be non-negative, and the sum of probabilities must be 1. For a set A consisting of pairs (x, y) , the probability

$$P[(X, Y) \in A]$$

is obtained by summing the joint PMF over pairs in A :

$$P[(X, Y) \in A] = \sum_{(x,y) \in A} \sum f_{X,Y}(x,y) .$$

Jointly Distributed Random Variables

177/321

Jointly Distributed Random Variables

178/321

Marginal distribution

The PMF of a single variable is obtained by summing the joint PMF across the entire range of the other variable. In this context, this result yields a *marginal* PMF. If the joint PMF is represented as a rectangular array, the marginal PMFs are simply the row and column totals.

The *marginal* PMFs of X and Y , denoted by $f_X(x)$ and $f_Y(y)$ respectively, are obtained from the joint PMF by

$$f_X(x) = \sum_y f_{X,Y}(x,y) ,$$

$$f_Y(y) = \sum_x f_{X,Y}(x,y) .$$

Where summation is across the appropriate range.

Jointly Distributed Random Variables

179/321

Jointly Distributed Random Variables

180/321

Example

Suppose that X and Y have joint PMF represented by the joint probability table

		X		
		1	2	3
Y	5	0.2	0.1	0.0
	6	0.2	0.1	0.1
	7	0.1	0.1	0.1

A) $P(X = 1, Y = 6) = f_{X,Y}(1, 6) = 0.2$

B) Let $A = \{(1, 5), (2, 6), (3, 7)\}$. Then

$$\begin{aligned}
 P[(X, Y) \in A] &= f_{X,Y}(1, 5) + f_{X,Y}(2, 6) + f_{X,Y}(3, 7) \\
 &= 0.2 + 0.1 + 0.1 = 0.4 .
 \end{aligned}$$

Example

Suppose that X and Y have joint PMF represented by the joint probability table

		X			
		1	2	3	
Y	5	0.2	0.1	0.0	0.3
	6	0.2	0.1	0.1	0.4
	7	0.1	0.1	0.1	0.3
		0.5	0.3	0.2	1.0

Of course, both marginals PMFs must satisfy the conditions for probabilities. Also, we should be careful with the ranges of the marginal distributions.

x	1	2	3	y	5	6	7
$f_X(x)$	0.5	0.3	0.2	$f_Y(y)$	0.3	0.4	0.3

Continuous Random Variables

If X and Y are continuous, we define the *joint probability density function* in a similar way.

For continuous random variables X and Y , the function $f_{X,Y}(x, y)$ is the *joint PDF* of X and Y if, for any two-dimensional set A ,

$$P[(X, Y) \in A] = \iint_A f_{X,Y}(x, y) dx dy.$$

In particular, if $A = \{(x, y) : a \leq x \leq b, c \leq y \leq d\}$, i.e. a rectangle, then

$$P[(X, Y) \in A] = \int_a^b \int_c^d f_{X,Y}(x, y) dy dx.$$

This must satisfy the usual conditions: the joint PDF must be non-negative and the total area must be 1:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy dx = 1.$$

Example

Consider continuous random variables X and Y with joint PDF

$$f_{X,Y}(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

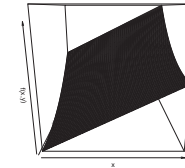
First, let us verify that this is a valid joint PDF.

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy dx &= \frac{6}{5} \int_0^1 \int_0^1 (x + y^2) dy dx \\ &= \frac{6}{5} \int_0^1 \left[xy + \frac{y^3}{3} \right]_0^1 dx \\ &= \frac{6}{5} \int_0^1 \left(x + \frac{1}{3} \right) dx \\ &= \frac{6}{5} \left[\frac{x^2}{2} + \frac{x}{3} \right]_0^1 = \frac{6}{5} \left(\frac{1}{2} + \frac{1}{3} \right) = 1 \end{aligned}$$

If X and Y are continuous with a joint cumulative distribution function $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$, the *joint probability density function* is obtained as

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$$

The joint PDF can be represented as



We evaluate probabilities directly from the joint PDF, e.g.,

$$\begin{aligned} P\left(0 \leq X \leq \frac{1}{4}, 0 \leq Y \leq \frac{1}{4}\right) &= \int_0^{\frac{1}{4}} \int_0^{\frac{1}{4}} \frac{6}{5}(x + y^2) dy dx \\ &= \frac{6}{5} \int_0^{\frac{1}{4}} \left[xy + \frac{y^3}{3} \right]_0^{\frac{1}{4}} dx \\ &= \frac{6}{5} \int_0^{\frac{1}{4}} \left(\frac{x}{4} + \frac{1}{192} \right) dx \\ &= \frac{6}{5} \left[\frac{x^2}{8} + \frac{x}{192} \right]_0^{\frac{1}{4}} \approx 0.0109 \end{aligned}$$

Marginal distributions

In direct analogy to the discrete case, we can reason about the marginal distributions when X and Y are continuous.

The *marginal probability density functions* of X and Y , denoted by $f_X(x)$ and $f_Y(y)$ respectively, are obtained from the joint PDF by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \quad \text{for } -\infty < x < \infty,$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx \quad \text{for } -\infty < y < \infty.$$

Note the use of the subscript to identify the marginal.

Knowing $f_{X,Y}(x,y)$, we can find $f_X(x)$ and $f_Y(y)$. The opposite is in general not true.

Example

Consider continuous random variables X and Y with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy = \frac{6}{5} \int_0^1 (x+y^2) dy$$

$$= \frac{6}{5} \left[xy + \frac{y^3}{3} \right]_0^1 = \frac{6}{5}x + \frac{2}{5}$$

for $0 \leq x \leq 1$, and 0 otherwise.

Also

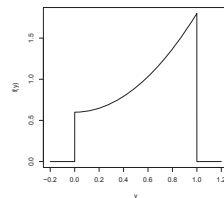
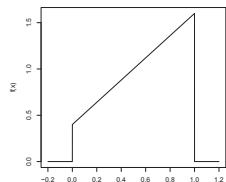
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx = \frac{6}{5} \int_0^1 (x+y^2) dx$$

$$= \frac{6}{5} \left[\frac{x^2}{2} + y^2x \right]_0^1 = \frac{6}{5}y^2 + \frac{3}{5}$$

for $0 \leq y \leq 1$, and zero otherwise.

Of course, we can compute probabilities from marginal distributions in the usual manner. For example

$$P\left(\frac{1}{4} \leq Y \leq \frac{3}{4}\right) = \int_{\frac{1}{4}}^{\frac{3}{4}} f_Y(y) dy = \frac{37}{80} \approx 0.4265.$$



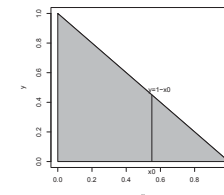
If the region of interest is not rectangular, we have to work a bit harder to evaluate probabilities or marginals.

Example

Consider the continuous random variables X and Y with joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} 24xy & 0 \leq x \leq 1, 0 \leq y \leq 1, x+y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Note the extra constraint $y \leq 1 - x$.



First, let us establish that this is a valid PDF. We need to take care with the limits of integration, as the region is triangular.

$$\begin{aligned}\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy &= \int_0^1 \left\{ \int_0^{1-x} 24xy dy \right\} dx \\ &= \int_0^1 24x \left[\frac{y^2}{2} \right]_{y=0}^{y=1-x} dx \\ &= \int_0^1 12x(1-x)^2 dx = 1.\end{aligned}$$

Now suppose $A = \{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq 1, x+y \leq 0.5\}$,

$$\begin{aligned}P[(X,Y) \in A] &= \iint_A f_{X,Y}(x,y) dx dy = \int_0^{0.5} \int_0^{0.5-x} 24xy dy dx \\ &= \int_0^{0.5} 24x \left[\frac{y^2}{2} \right]_{y=0}^{y=0.5-x} dx = \int_0^{0.5} 12x(1/2-x)^2 dx \\ &= 0.0625\end{aligned}$$

Independent Random Variables

So far, we have mentioned the concept of independent random variables only in passing, e.g. when defining the binomial distribution as a sum of independent Bernoullis. We are now in a position to give a formal definition, which stems from the concept of independent events. The idea is the same: if X and Y are independent, then knowing about one does not tell us anything about the other.

Random variables X and Y are *independent* if for all x and y

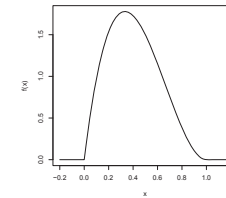
$$P(X \leq x \cap Y \leq y) = P(X \leq x)P(Y \leq y)$$

If this condition is not satisfied, then X and Y are *dependent*.

Finally, we can obtain the marginal distribution of X as

$$\begin{aligned}f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \\ &= \int_0^{1-x} 24xy dy \\ &= 12x(1-x)^2\end{aligned}$$

for $0 \leq x \leq 1$ and 0 otherwise.



■

Independence can be also be written equivalently in terms of PMF and PDF

Random variables X and Y are *independent* if for all x and y

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

(continuous random variable)

$$P(X = x \cap Y = y) = P(X = x)P(Y = y)$$

(discrete random variable)

If this condition is not satisfied, then X and Y are *dependent*.

Thus, random variables are independent if the joint PMF or PDF can be expressed as a product of the marginal PMFs or PDFs.

Example

Consider the earlier example (*discrete* random variables), with

		X			$f_Y(y)$
		1	2	3	
Y	5	0.2	0.1	0.0	0.3
	6	0.2	0.1	0.1	0.4
	7	0.1	0.1	0.1	0.3
$f_X(x)$		0.5	0.3	0.2	1.0

In this example, we have

$$f_X(1)f_Y(7) = 0.5 \times 0.3 = 0.15 \neq 0.1 = f_{X,Y}(1,7)$$

thus, X and Y are dependent. ■

Example

Consider *continuous* random variables X and Y with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The marginal PDFs are

$$f_X(x) = \frac{6}{5}x + \frac{2}{5} \quad \text{and} \quad f_Y(y) = \frac{6}{5}y^2 + \frac{3}{5}.$$

The product of the marginals is

$$f_X(x)f_Y(y) = \left(\frac{6}{5}x + \frac{2}{5}\right) \left(\frac{6}{5}y^2 + \frac{3}{5}\right),$$

which is different to the joint density, and therefore X and Y are dependent. ■

More than 2 variables

The concepts introduced here extend readily to more than two variables. The joint PMF, PDF, and CDF are defined in a straightforward manner.

- ▶ Joint CDF: $F_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$

- ▶ Joint PDF:

$$f_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = \frac{\partial^n F_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_n}$$

For more than two variables, we have (mutual) independence if every subset of variables (pair, triplet, etc) are independent, that is, the joint PMF (PDF) can be expressed as a product of appropriate marginals.

- ▶ Independent

$$F_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2) \dots F_{X_n}(x_n)$$

$$f_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_n}(x_n)$$

- ▶ i.i.d. (independent and identically distributed): RVs are independent and have the same distribution

Example

Consider X_1, X_2, \dots, X_n independent, with $X_i \sim \text{Poisson}(\lambda)$. Then, from independence,

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

From the definition of the Poisson distribution, this gives

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{x_1! x_2! \dots x_n!}$$

In the context of statistical inference, this is the likelihood function (more on this later). ■

Conditional Distributions

We can extend the concept of conditional probability to jointly distributed random variables. In this context, we want to consider how one random variable behaves when we condition on a specific realisation of the other random variable.

For discrete/continuous random variables X and Y with joint PMF/PDF $f_{X,Y}(x,y)$ and marginals $f_X(x)$ and $f_Y(y)$, with $f_X(x) > 0$, the *conditional* PMF/PDF of Y given that $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

$f_{Y|X}(y|x)$ is a valid probability mass/density function:

$$\int_{-\infty}^{+\infty} f_{Y|X}(y|x) dy = \frac{1}{f_X(x)} \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy = \frac{f_X(x)}{f_X(x)} = 1.$$

Moreover,

$$P(Y \leq y | X = x) = \int_{-\infty}^y f_{Y|X}(y|x) dy.$$

If X, Y are independent, conditional PMF/PDF are equal to marginal PMF/PDF

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y), \quad \forall x, y$$

Example

Consider continuous random variables X and Y with joint PDF

$$f_{X,Y}(x,y) = \begin{cases} \frac{6}{5}(x+y^2) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The conditional PDF of X given $Y = 0.3$ is

$$f_{X|Y}(x|y) = \frac{f(x, 0.3)}{f_Y(0.3)} = \frac{\frac{6}{5}(x + 0.3^2)}{\frac{6}{5}(0.3^2) + \frac{3}{5}} = \frac{100}{59}(x + 0.3)$$

for $0 \leq x \leq 1$. ■

Expectation

The result for the expected value of a function of a random variable extends to joint distributions. Again, the result states that we do not require the distribution of the transformed variable, we simply weight the PDF (or PMF) appropriately.

For random variables X and Y , the expected value of $g(X, Y)$ is

$$E[g(X, Y)] = \begin{cases} \sum_x \sum_y g(x, y) f_{X,Y}(x, y) & X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy & X, Y \text{ continuous} \end{cases}$$

For $g(X, Y) = X$,

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \right] dx \\ &= \int_{-\infty}^{\infty} x f_X(x) dx \end{aligned}$$

Example

Consider the earlier example, with

		X			
		1	2	3	$f_Y(y)$
Y	5	0.2	0.1	0.0	0.3
	6	0.2	0.1	0.1	0.4
	7	0.1	0.1	0.1	0.3
$f_X(x)$		0.5	0.3	0.2	1.0

Compute $E[g(X, Y)]$, where $g(X, Y) = X + Y$. No need to think about the distribution of $g(X, Y)$, simply use the result

$$\begin{aligned} E[g(X, Y)] &= \sum_x \sum_y g(x, y) f_{X,Y}(x, y) \\ &= 6(0.2) + 7(0.2) + 8(0.1) + \dots + 9(0.1) + 10(0.1) = 7.7. \end{aligned}$$



Example

Recall previous example with the conditional PDF of X given $Y = 0.3$ being

$$f_{X|Y}(x|y) = \frac{100}{59} (x + 0.3)$$

for $0 \leq x \leq 1$. We can also consider the *conditional expectation* of X given that $Y = 0.3$,

$$E[X|Y = 0.3] = \int_{-\infty}^{\infty} x f_{X|Y}(x|0.3) dx = \frac{100}{77} x^3 + \frac{9}{118} x^2 \Big|_0^1 = \frac{227}{354}.$$



Conditional Expectation

For random variables X and Y , the *conditional expectation* of X w.r.t. Y , denoted as $E[X|Y]$, is the random variable $g(Y)$, function of Y , whose value at point y of Y is given by

$$g(y) = E[X|Y = y] = \begin{cases} \sum_i x_i P(X = x_i | Y = y), & \text{discrete,} \\ \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx, & \text{continuous.} \end{cases}$$

Note that $E(X)$ is a number while $E[X|Y]$ is a random variable!

If X, Y are independent, then $E[X|Y] = E(X)$.

Proof:

$$\begin{aligned} \forall y, \quad E[X|Y = y] &= \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx \\ &= \int_{-\infty}^{+\infty} x f_X(x) dx \quad (\text{by independence}) \\ &= E(X) \end{aligned}$$



For random variables X and Y ,

$$E_Y[E[X|Y]] = E(X)$$

Proof:

$$\begin{aligned} E_Y[E[X|Y]] &= \int_{-\infty}^{+\infty} E[X|Y = y] f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f_{X|Y}(x|y) dx f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f_{X,Y}(x, y) dx dy \\ &= E(X) \end{aligned}$$



Other properties:

$$\begin{aligned} E[ag(X) + h(X)|Y] &= a E[g(X)|Y] + E[h(X)|Y], \\ E[g(X)h(Y)|Y = y] &= h(y) E[g(X)|Y = y]. \end{aligned}$$

Conditional Variance

For random variables X and Y , the *conditional variance* of X w.r.t. Y , denoted as $\text{Var}[X|Y]$, is the random variable, function of Y , whose value at point $Y = y$ is given by

$$\text{Var}[X|Y = y] = \begin{cases} \sum_i [x_i - E[X|Y = y]]^2 P(X = x_i|Y = y), & \text{disc.} \\ \int_{-\infty}^{+\infty} [x - E[X|Y = y]]^2 f_{X|Y}(x|y) dx, & \text{cont.} \end{cases}$$

For random variables X and Y ,

$$\text{Var}(X) = E[\text{Var}[X|Y]] + \text{Var}[E[X|Y]]$$

Proof: Write

$$\begin{aligned} \text{Var}(X) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - E(X))^2 f_{X,Y}(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - E[X|Y = y] + E[X|Y = y] - E(X))^2 f_{X,Y}(x, y) dx dy \end{aligned}$$

Covariance

It is often useful to characterise the nature of the dependence between dependent random variables X and Y . The *covariance*, another property defined in terms of expectation, measures the strength of such dependencies.

The *covariance* between random variables X and Y is

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)],$$

where μ_x and μ_y are the expected value of X and Y respectively. Thus

$$\text{Cov}(X, Y) = \begin{cases} \sum_x \sum_y (x - \mu_x)(y - \mu_y) f_{X,Y}(x, y) & \text{disc.} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f_{X,Y}(x, y) dx dy & \text{cont.} \end{cases}$$

where μ_x and μ_y are the expected values of X and Y respectively.

and expand each term ...

Noting that

$$\int_{-\infty}^{+\infty} (x - E[X|Y = y]) f_{X|Y}(x|y) dx = 0,$$

the double product is equal to zero and $\text{Var}(X)$ is left with the following two terms

$$\begin{aligned} E[\text{Var}[X|Y]] &= \int_{-\infty}^{+\infty} \text{Var}[X|Y = y] f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - E[X|Y = y])^2 f_{X|Y}(x|y) dx f_Y(y) dy \\ \text{Var}[E[X|Y]] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (E[X|Y = y] - E_Y[E[X|Y]])^2 f_{X,Y}(x, y) dx dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (E[X|Y = y] - E(X))^2 f_{X,Y}(x, y) dx dy \end{aligned}$$

■

Some useful properties:

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

$$\text{Cov}(X, X) = \text{Var}(X)$$

$$\text{Cov}(X, a) = 0, \text{ for any constant } a$$

$$\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$$

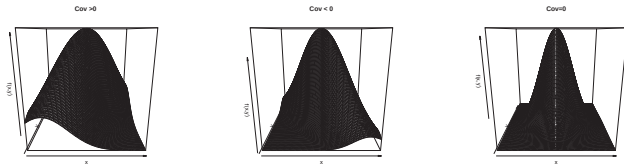
Expanding the covariance formula gives

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

If $\text{Cov}(X, Y) > 0$ then large values of X tend to be associated with large values of Y . The higher the covariance, the stronger the relationship. Conversely, if $\text{Cov}(X, Y) < 0$, then large values of X tend to be associated with small values of Y , and vice versa.

A covariance near zero indicates that there is no simple linear relationship between the variables.

Example



Example

Consider discrete random variables X and Y with joint PMF

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{2} & x = 3, y = 4 \\ \frac{1}{3} & x = 3, y = 6 \\ \frac{1}{6} & x = 5, y = 6 \\ 0 & \text{otherwise} \end{cases}$$

Now $E(X) = \frac{1}{2}3 + \frac{1}{3}3 + \frac{1}{6}5 + 0 = 10/3$ and $E(Y) = 5$. The covariance is then

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - \mu_x)(Y - \mu_y)] \\ &= \frac{(3 - 10/3)(4 - 5)}{2} + \frac{(3 - 10/3)(6 - 5)}{3} \\ &\quad + \frac{(5 - 10/3)(6 - 5)}{6} + 0 \\ &= 1/6 - 1/9 + 5/18 + 0 = 1/3 \end{aligned}$$

Example

Recall the example with X and Y continuous, with joint PDF

$$f_{X,Y}(x, y) = 24xy$$

for $0 \leq x \leq 1$, $0 \leq y \leq 1$, $x + y \leq 1$, and 0 otherwise.

$$\begin{aligned} \mu_x &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dy dx = \int_0^1 \int_0^{1-x} x 24xy dy dx \\ &= \int_0^1 24x^2 \left[\frac{y^2}{2} \right]_0^{1-x} dx = \int_0^1 (12x^2 - 24x^3 + 12x^4) dx = 2/5. \end{aligned}$$

Similarly, $\mu_y = 2/5$.

Then

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x, y) dy dx = \int_0^1 \int_0^{1-x} xy 24xy dy dx \\ &= 8 \int_0^1 x^2 (1-x)^3 dx = 2/15 \end{aligned}$$

The covariance is

$$\text{Cov}(X, Y) = E(XY) - \mu_x \mu_y = \frac{2}{15} - \left(\frac{2}{5} \right)^2 = -\frac{2}{75}$$



If X and Y are independent, then

$$E(XY) = E(X) E(Y)$$

Proof: the proof is straightforward; we show the discrete case:

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy f_{X,Y}(x, y) = \sum_x \sum_y xy f_X(x) f_Y(y) \\ &= \left(\sum_x x f_X(x) \right) \left(\sum_y y f_Y(y) \right) = E(X) E(Y) \end{aligned}$$

■

Thus, for X and Y independent

$$\text{Cov}(X, Y) = E(XY) - E(X) E(Y) = 0.$$

However, note that the converse does not apply!

An example of uncorrelated but dependent RVs

Example

Consider the RV Θ uniformly distributed in $[0, 2\pi]$

$$f_{\Theta}(\theta) = \frac{1}{2\pi} \quad \text{for } 0 \leq \theta \leq 2\pi.$$

Define

$$X = \cos(\Theta), \quad Y = \sin(\Theta).$$

Clearly, X and Y are not independent. But they are uncorrelated:

$$E[X] = \frac{1}{2\pi} \int_0^{2\pi} \cos(\theta) d\theta = 0$$

$$E[Y] = \frac{1}{2\pi} \int_0^{2\pi} \sin(\theta) d\theta = 0$$

$$E[XY] = \int_0^{2\pi} \sin(\theta) \cos(\theta) f_{\Theta}(\theta) d\theta = \frac{1}{4\pi} \int_0^{2\pi} \sin(2\theta) d\theta = 0.$$

If X and Y are independent, then

$$E[g(X)h(Y)] = E[g(X)] E[h(Y)]$$

Proof:

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) f_{X,Y}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) f_X(x) f_Y(y) dx dy \\ &= E[g(X)] E[h(Y)] \end{aligned}$$

■

Example

Consider discrete random variables X and Y , with joint PMF given by

$$f_{X,Y}(x, y) = \begin{cases} 1/4 & x = 3, y = 5 \\ 1/4 & x = 4, y = 9 \\ 1/4 & x = 6, y = 9 \\ 1/4 & x = 7, y = 5 \\ 0 & \text{otherwise} \end{cases}$$

Now, $E(X) = 5$ and $E(Y) = 7$. Also,

$$E(XY) = \frac{3(5)}{4} + \frac{4(9)}{4} + \frac{7(5)}{4} + \frac{6(9)}{4} = 35.$$

Thus, $\text{Cov}(X, Y) = E(XY) - E(X) E(Y) = 35 - 35 = 0$.

Now $P(X = 4) > 0$ and $P(Y = 5) > 0$ (compute the marginals to verify), however, $P(X = 4, Y = 5) = 0 \neq P(X = 4)P(Y = 5)$.

■

Correlation

A deficiency of covariance is that it depends on the units of measurement. For example, the covariance of kX and kY ($k \neq 0$) is equal to $k^2 \text{Cov}(X, Y)$. For this reason, we prefer to work with the *correlation* of X, Y .

The correlation coefficient of variables X and Y is

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

provided that both variances are finite.

The correlation coefficient characterises the strength of the *linear* relationship between the variables X and Y !

Property of the correlation coefficient: $-1 \leq \rho \leq 1$

Proof: Take $U = Y - E(Y)$ and $V = X - E(X)$ and apply Cauchy-Schwartz's inequality

$$\begin{aligned} (E(UV))^2 &\leq E(U^2) E(V^2) \\ (E[(Y - E(Y))(X - E(X))])^2 &\leq E[(Y - E(Y))^2] E[(X - E(X))^2] \\ (\text{Cov}(Y, X))^2 &\leq \text{Var}(X) \text{Var}(Y) \\ \rho^2 &\leq 1. \end{aligned}$$

Equality $\rho^2 = 1$ corresponds to the double root

$$\alpha_0 = \frac{E(UV)}{E(V^2)} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \rho \frac{\sigma_Y}{\sigma_X},$$

which gives, for $\rho = \pm 1$

$$P\left(Y - E(Y) = \rho \frac{\sigma_Y}{\sigma_X} (X - E(X))\right) = 1$$

or alternatively

$$P\left(\frac{Y - E(Y)}{\sigma_Y} = \pm \frac{X - E(X)}{\sigma_X}\right) = 1.$$

Lemma - Cauchy-Schwartz's inequality: For random variables U and V ,

$$(E(UV))^2 \leq E(U^2) E(V^2).$$

Equality occurs only if it exists α_0 such that $P(U = \alpha_0 V) = 1$.

Proof: For any real α , we can always write

$$0 \leq E[(U - \alpha V)^2] = E(U^2) - 2\alpha E(UV) + \alpha^2 E(V^2).$$

This is reminiscent of the quadratic equation $a\alpha^2 + b\alpha + c \geq 0$, which is possible only if the discriminant $b^2 - 4ac \leq 0$. Hence,

$$(E(UV))^2 - E(U^2) E(V^2) \leq 0.$$

If equality occurs, there exists a double root $\alpha_0 = \frac{E(UV)}{E(V^2)}$ for which

$$E[(U - \alpha_0 V)^2] = 0.$$

From Chebyshev's inequality, this implies

$$E(U - \alpha_0 V) = 0, \quad \text{Var}(U - \alpha_0 V) = 0, \quad P(U - \alpha_0 V = 0) = 1.$$

The proof highlights that the correlation coefficient characterises the strength of the *linear* relationship between the variables X and Y !

Equality, $\rho = \pm 1$, occurs only when $Y = aX + b$ for $a \neq 0$, a perfect linear relationship. In this case, $\rho = 1$ when $a > 0$, and $\rho = -1$ when $a < 0$.

$\rho > 0$ (resp. $\rho < 0$) indicates that X and Y evolve in the same (resp. opposite) direction.

Note that, if X and Y are independent, then $\rho = 0$. However, $\rho = 0$ does *not* imply independence. When $\rho = 0$ we say that the random variables are *uncorrelated*.

If X and Y are uncorrelated, there is no linear relationship between them. However, a non-linear relationship may still exist even though $\rho = 0$!

Example

Assume $Z \sim N(0, 1)$ and take $X = Z$ and $Y = Z^2$:

$$\text{Cov}(X, Y) = E(Z^3) - E(Z) E(Z^2) = 0$$

because $E(Z^3) = E(Z) = 0$. Hence $\rho = 0$ even though $Y = X^2$.

■

Example

Earlier, we considered discrete random variables X and Y , with joint PMF given by

$$f_{X,Y}(x, y) = \begin{cases} 1/4 & x = 3, y = 5 \\ 1/4 & x = 4, y = 9 \\ 1/4 & x = 6, y = 9 \\ 1/4 & x = 7, y = 5 \\ 0 & \text{otherwise} \end{cases}$$

We found that the covariance was 0. Also, $\text{Var}(X) = 2.5$ and $\text{Var}(Y) = 4$. Thus the correlation of X and Y is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = 0.$$

■

Example

Earlier, we considered discrete random variables X and Y , with joint PMF given by

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{2} & x = 3, y = 4 \\ \frac{1}{3} & x = 3, y = 6 \\ \frac{1}{6} & x = 5, y = 6 \\ 0 & \text{otherwise} \end{cases}$$

We found that the covariance was $1/3$. Also,

$$\text{Var}(X) = E(X^2) - E(X)^2 = \frac{1}{2}9 + \frac{1}{3}9 + \frac{1}{6}25 - \left(\frac{10}{3}\right)^2 = \frac{5}{9}$$

and $\text{Var}(Y) = 1$. Thus the correlation of X and Y is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{1/3}{\sqrt{5/9}} \approx 0.447.$$

Covariance and Correlation Matrices

Stacking up X and Y in a vector, it is very common to work with a *covariance matrix*

$$\begin{aligned} \mathbf{R} &= E \left[\begin{bmatrix} X - E(X) \\ Y - E(Y) \end{bmatrix} \begin{bmatrix} X - E(X) & Y - E(Y) \end{bmatrix} \right] \\ &= \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{bmatrix} \end{aligned}$$

and a *correlation matrix*

$$\mathbf{C} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Joint Normal (Gaussian) Distribution

X and Y are said to be jointly normal (Gaussian) distributed $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ if

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}$$

for $-\infty < x < +\infty$, $-\infty < y < +\infty$, $|\rho| < 1$.

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy = \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} N(\mu_X, \sigma_X^2)$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dx = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}} N(\mu_Y, \sigma_Y^2)$$

The marginals alone do not tell us everything about the joint PDF, except when X, Y are independent. Note the equivalence between uncorrelated and independent for Normal distribution!

Jointly Distributed Random Variables

225/321

4. Compute the variance of X and the variance of Y , i.e. $\text{Var}(X)$ and $\text{Var}(Y)$, the covariance between X and Y , i.e. $\text{Cov}(X, Y)$, and the correlation coefficient between X and Y , i.e. $\text{Corr}(X, Y)$.
5. Are X and Y uncorrelated? Independent? Provide your reasoning.
6. Compute the conditional probability mass function of X given that $Y = 0, 1, 2$.
7. Compute the conditional expectation of X given that $Y = 0, 1, 2$.
8. Relying on your result in 7), compute the expectation of X , i.e. $E(X)$.

Exercise (Exam Question May 2014)

Consider two discrete random variables X and Y that have joint probability mass function represented by the joint probability table

		Y		
		0	1	2
X	0	0.05	0.05	0.15
	1	0.05	0.05	0.25
	2	0.15	0.20	0.05

1. Compute the probability that X is smaller or equal to Y , i.e. $P(X \leq Y)$ and the probability that X is strictly smaller than Y , i.e. $P(X < Y)$.
2. Compute the marginal probability mass function of X and Y .
3. Compute the expectation of X , i.e. $E(X)$, and the expectation of Y , i.e. $E(Y)$.

Jointly Distributed Random Variables

226/321

Moments

For $r = 1, 2, \dots$, the quantities

$$m_r = E[X^r]$$

are referred to as the *moments* of the random variable. The first moment, m_1 , is just the mean of X . The second moment, together with the first, yields the variance, via

$$\text{Var}(X) = E[X^2] - E[X]^2 = m_2 - m_1^2.$$

The third moment provides information about the skewness of the distribution, and so on.

For many important distributions the full sequence of moments can be obtained from a special function, called the *moment generating function* (MGF).

Moment Generating Functions

The MGF of a random variable X is defined as

$$m_X(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} f_X(x) & X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx & X \text{ continuous} \end{cases}$$

whenever this expectation exists. Note that the MGF is a function of parameter t .

The MGF has several merits: it provides the full sequence of moments, uniquely identifies the distribution function of random variables, and provides useful results for sums of random variables.

Example

Consider the random variable $X \sim \text{Poisson}(\lambda)$, with PMF

$$f_X(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots$$

The MGF is

$$\begin{aligned} m_X(t) &= E(e^{tX}) = \sum_x e^{tx} f_X(x) \\ &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda} e^{\lambda e^t} = \exp(\lambda(e^t - 1)) \end{aligned}$$



Example

Consider the random variable $X \sim N(\mu, \sigma^2)$, with PDF

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The MGF is

$$\begin{aligned} m_X(t) &= E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{x^2-2(\mu+t\sigma^2)x+\mu^2}{\sigma^2}} dx \\ &= e^{t\mu+t^2\sigma^2/2} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{x^2-2(\mu+t\sigma^2)x+\mu^2+2t\mu\sigma^2+t^2\sigma^4}{\sigma^2}} dx \\ &= e^{t\mu+t^2\sigma^2/2} \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-(\mu+t\sigma^2)}{\sigma}\right)^2}}_{N(\mu+t\sigma^2, \sigma^2)} dx = e^{t\mu+t^2\sigma^2/2} \end{aligned}$$



The link between the moments of X and the MGF stems from its power series expansion:

$$\begin{aligned} m_X(t) &= E(e^{tX}) = E\left(\sum_{r=0}^{\infty} \frac{(tX)^r}{r!}\right) = \sum_{r=0}^{\infty} \frac{E(X^r)}{r!} t^r \\ &= 1 + \frac{E(X)}{1!} t + \frac{E(X^2)}{2!} t^2 + \frac{E(X^3)}{3!} t^3 + \dots \\ &= 1 + \frac{m_1}{1!} t + \frac{m_2}{2!} t^2 + \frac{m_3}{3!} t^3 + \dots \end{aligned}$$

In this power series, the coefficient of t^r is $m_r/r!$. If we can expand the MGF of X in such a way, we can obtain the moments of X simply by equating coefficients. If expanding the MGF is difficult, we can still obtain the moments by differentiation.

Provided that $m_X(t) < \infty$ for $t \in (-\epsilon, \epsilon)$, with $\epsilon > 0$, we can evaluate the derivatives of $m_X(t)$ at $t = 0$. First, notice that

$$\begin{aligned}\frac{d}{dt}m_X(t) &= m_1 + m_2t + \frac{m_3}{2}t^2 + \dots \\ \frac{d^2}{dt^2}m_X(t) &= m_2 + m_3t + \frac{m_4}{2}t^2 + \dots \\ &\vdots \\ \frac{d^r}{dt^r}m_X(t) &= m_r + m_{r+1}t + \frac{m_{r+2}}{2}t^2 + \dots\end{aligned}$$

Evaluating each of these derivatives at $t = 0$ leaves us with just the constant term, so we can compute the r th moment of X as

$$m_r = \left. \frac{d^r}{dt^r}m_X(t) \right|_{t=0} = m_X^{(r)}(0).$$

Example

We found that the Poisson distribution has MGF $m_X(t) = \exp(\lambda(e^t - 1))$. To compute the first and second moments, we find

$$\begin{aligned}m_X'(t) &= \lambda e^t \exp(\lambda(e^t - 1)) \\ m_X''(t) &= \lambda^2 e^{2t} \exp(\lambda(e^t - 1)) + \lambda e^t \exp(\lambda(e^t - 1))\end{aligned}$$

So that

$$\begin{aligned}E(X) &= m_X'(0) = \lambda \\ E(X^2) &= m_X''(0) = \lambda^2 + \lambda \Rightarrow \text{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda\end{aligned}$$

■

Example

We found that the Normal distribution has MGF $m_X(t) = e^{t\mu + t^2\sigma^2/2}$. To compute the first and second moments, we find

$$\begin{aligned}m_X'(t) &= e^{t\mu + t^2\sigma^2/2}(\mu + 2t\sigma^2/2) \\ m_X''(t) &= e^{t\mu + t^2\sigma^2/2}(\mu + 2t\sigma^2/2)^2 + e^{t\mu + t^2\sigma^2/2}\sigma^2\end{aligned}$$

So that

$$\begin{aligned}E(X) &= m_X'(0) = \mu \\ E(X^2) &= m_X''(0) = \mu^2 + \sigma^2 \Rightarrow \text{Var}(X) = \mu^2 + \sigma^2 - \mu^2 = \sigma^2\end{aligned}$$

■

The MGF uniquely identifies the distribution of a random variable. Thus, if we are told that a random variable X has MGF $m_X(t) = \exp(\lambda(e^t - 1))$, then it *must* be a Poisson random variable with mean λ .

Now consider the sum of two independent random variables X and Y . From the properties of expectation and independence, we have

$$\begin{aligned}m_{X+Y}(t) &= E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) \\ &= E(e^{tX})E(e^{tY}) = m_X(t)m_Y(t).\end{aligned}$$

For example, applying this formula when X and Y are both normally distributed yields the MGF of another normal distribution. This result, that the sum of independent normal random variables is itself a normal random variable, will prove very important in statistical analysis.

Example

Consider independent random variables X and Y , with $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, and let $Z = X + Y$. The MGF of Z is

$$\begin{aligned} m_Z(t) &= m_{X+Y}(t) = m_X(t)m_Y(t) \\ &= \exp(t\mu_1 + t^2\sigma_1^2/2)\exp(t\mu_2 + t^2\sigma_2^2/2) \\ &= \exp(t(\mu_1 + \mu_2) + t^2(\sigma_1^2 + \sigma_2^2)/2). \end{aligned}$$

This is the MGF of the Normal distribution with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. Thus $Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. ■

Example

Consider independent random variables X and Y , with $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, and let $Z = X + Y$. The MGF of Z is

$$\begin{aligned} m_Z(t) &= m_{X+Y}(t) = m_X(t)m_Y(t) = \exp(\lambda(e^t - 1))\exp(\mu(e^t - 1)) \\ &= \exp(e^t(\lambda + \mu) - (\lambda + \mu)). \end{aligned}$$

This is the MGF of the Poisson distribution with mean $\lambda + \mu$. Thus $Z \sim \text{Poisson}(\lambda + \mu)$. ■

Example

Consider independent random variables X and Y , with $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$, and let $Z = aX + bY$. The MGF of Z is

$$\begin{aligned} m_Z(t) &= m_{aX+bY}(t) = E[e^{atX}] E[e^{btY}] \\ &= \exp(at\mu_1 + a^2t^2\sigma_1^2/2)\exp(bt\mu_2 + b^2t^2\sigma_2^2/2) \\ &= \exp(t(a\mu_1 + b\mu_2) + t^2(a^2\sigma_1^2 + b^2\sigma_2^2)/2). \end{aligned}$$

This is the MGF of the Normal distribution with mean $a\mu_1 + b\mu_2$ and variance $a^2\sigma_1^2 + b^2\sigma_2^2$. Thus $Z \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$. ■

Relationship with Fourier Transform

If $t = jw$, the MGF of a random variable X is often called characteristic function of X

$$\phi_X(w) = m_X(jw) = E(e^{jwX}) = \int_{-\infty}^{\infty} f_X(x)e^{jwx} dx.$$

The *characteristic function* is nothing else than a Fourier transform with $-w$ instead of w .

Hence, the PDF could be obtained from the characteristic function

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(w)e^{-jwx} dw.$$

For two independent RVs X, Y , $m_{X+Y}(t) = m_X(t)m_Y(t)$. Hence,

$$\phi_{X+Y}(w) = \phi_X(w)\phi_Y(w).$$

Since a multiplication in the transformed domain (Fourier) corresponds to a convolution in the direct domain, we have

$$f_{X+Y}(z) = f_X(z) \otimes f_Y(z) = \int_u f_X(z - u)f_Y(u)du.$$

Sums of random variables

In statistical applications, we are often concerned with sums of random variables. Consider random variables X_1, X_2, \dots, X_n . We have already seen, using the properties of sums/integrals, that

$$E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n) = \sum_{i=1}^n E(X_i).$$

For *mutually independent* random variables X_1, X_2, \dots, X_n , we have

$$\begin{aligned} \text{Var}(X_1 + X_2 + \dots + X_n) &= \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n) \\ &= \sum_{i=1}^n \text{Var}(X_i). \end{aligned}$$

We can easily prove these results for two random variables.

The result for the variance of a weighted sum of two independent random variables, $Z = a_1X + a_2Y$, can be obtained directly from the definition:

$$\begin{aligned} \text{Var}(a_1X + a_2Y) &= E[(a_1X + a_2Y - (a_1\mu_X + a_2\mu_Y))^2] \\ &= E[(a_1(X - \mu_X) + a_2(Y - \mu_Y))^2] \\ &= E[a_1^2(X - \mu_X)^2 + a_2^2(Y - \mu_Y)^2 \\ &\quad + 2a_1a_2(X - \mu_X)(Y - \mu_Y)] \\ &= E[a_1^2(X - \mu_X)^2] + E[a_2^2(Y - \mu_Y)^2] \\ &\quad + E[2a_1a_2(X - \mu_X)(Y - \mu_Y)] \\ &= a_1^2 \text{Var}(X) + a_2^2 \text{Var}(Y) + 2a_1a_2 \text{Cov}(X, Y) \end{aligned}$$

Since X and Y are independent, $\text{Cov}(X, Y) = 0$, and thus

$$\text{Var}(a_1X + a_2Y) = a_1^2 \text{Var}(X) + a_2^2 \text{Var}(Y)$$

Consider the weighted sum $Z = a_1X + a_2Y$, where X and Y are continuous random variables, and a_1, a_2 are constants.

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (a_1x + a_2y) f_{X,Y}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a_1x f_{X,Y}(x, y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a_2y f_{X,Y}(x, y) dx dy \\ &= a_1 \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \right) dx \\ &\quad + a_2 \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{X,Y}(x, y) dx \right) dy \end{aligned}$$

Note that each of the inner integrals returns a marginal distribution, thus,

$$E(Z) = a_1 \int_{-\infty}^{\infty} x f_X(x) dx + a_2 \int_{-\infty}^{\infty} y f_Y(y) dy = a_1 E(X) + a_2 E(Y)$$

Set $a_1 = a_2 = 1$ for the sum.

Summary: For two independent random variables X and Y ,

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

$$E(XY) = E(X) E(Y)$$

$$\text{Cov}(X, Y) = 0$$

$$\rho = 0$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$$

$$m_{X+Y}(t) = m_X(t) m_Y(t)$$

$$E[X|Y] = E(X)$$

Example

Consider independent random variables, X_1 and X_2 , with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. Then

$$\begin{aligned} E(X_1 + X_2) &= E(X_1) + E(X_2) = \mu_1 + \mu_2, \\ \text{Var}(X_1 + X_2) &= \text{Var}(X_1) + \text{Var}(X_2) = \sigma_1^2 + \sigma_2^2. \end{aligned}$$

From the result stated earlier, MGF arguments also tell us that the sum of independent normal random variables is a normal random variable, thus,

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

■

The general result for independent random variables X_1, X_2, \dots, X_n , with

$$X_i \sim N(\mu_i, \sigma_i^2)$$

and non-zero constants a_1, a_2, \dots, a_n , is that the random variable Y defined by

$$Y = \sum_{i=1}^n a_i X_i$$

is also normally distributed:

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

This result will provide the framework for much of our inference.

Example

Differences of independent random variables. Consider

$$Y = a_1 X_1 + a_2 X_2$$

where $a_1 = 1$ and $a_2 = -1$. Using earlier results, we have

$$\begin{aligned} E(Y) &= a_1 E(X_1) + a_2 E(X_2) = \mu_1 - \mu_2, \\ \text{Var}(Y) &= a_1^2 \text{Var}(X_1) + a_2^2 \text{Var}(X_2) = \sigma_1^2 + \sigma_2^2. \end{aligned}$$

Moreover, if $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, we know that $-X_2 \sim N(-\mu_2, \sigma_2^2)$ (from the properties of the normal distribution: $E[e^{-tX}] = e^{-t\mu + t^2\sigma^2/2}$), so we have

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

■

Example

Consider a sequence of Bernoulli trials

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases}$$

with $P(X_i = 1) = p$, for $i = 1, 2, \dots, n$. The binomial random variable X is then

$$X = X_1 + X_2 + \dots + X_n$$

The mean and variance of each X_i are

$$E[X_i] = 1p + 0(1 - p) = p$$

and

$$\text{Var}(X_i) = (1-p)^2p + (0-p)^2(1-p) = p(1-p)$$

Using the result for the expectation of a sum of random variables we have

$$E[X] = E[X_1] + E[X_2] + \dots + E[X_n] = np$$

Similarly, the variance of the sum of independently distributed random variables is the sum of the variances:

$$\begin{aligned} \text{Var}(X) &= \text{Var}(X_1 + X_2 + \dots + X_n) \\ &= \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n) \\ &= p(1-p) + p(1-p) + \dots + p(1-p) \\ &= np(1-p) \end{aligned}$$

Thus, the variance of the binomial distribution is $np(1-p)$.

Hence, we saw two ways of finding the $\text{Var}(X)$ of a Bin: 1) using the property $E(X^2) - E(X)^2$, 2) using a sum of Bernoulli RV. ■

3. We are interested in computing the error probability of this communication system. The error probability is the probability of wrongly decoding the transmitted signal and can be approximated as the moment generating function of P evaluated at the point $t = -d$ for $d > 0$. Making use of such approximation, find the error probability. Provide your reasoning.
4. From the results in 3), find the expected value of the received power after selection. Provide your reasoning.

Exercise (Exam Question May 2014)

Consider a communication system where the transmitter is equipped with one antenna and the receiver with two antennas. The power of the signal received at antenna i is denoted as P_i , $i = 1, 2$, and is modeled as an exponentially distributed random variable with parameter $\lambda > 0$. The receiver only uses one antenna at a time and selects the antenna with the largest power. Hence, the power of the signal after selection is given by $P = \max_{i=1,2} P_i$. We assume the receive antennas are deployed such that P_1 and P_2 are independent.

1. Find the probability that the power of the signal after selection, P , falls below a certain level S . Provide your reasoning.
2. Find the probability density function of P . Provide your reasoning.

Exercise (Exam Question May 2014)

- Assume $X \sim N(\mu, \sigma^2)$. Find the moment generating function $m_X(t)$ of X . Provide your reasoning.
- Consider the following statement: If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ and X_1, X_2 are independent random variables, we have $2X_1 - X_2 \sim N(2\mu_1 - \mu_2, 2\sigma_1^2 - \sigma_2^2)$. Is the statement correct? If yes, provide a proof. If not, correct the statement and provide a proof.

Change of Variables: One function of two RVs

Given two random variables X and Y characterized by the joint PDF $f_{X,Y}(x,y)$ and a function $g(x,y)$, we form a new random variable Z as $Z = g(X,Y)$. What is the PDF $f_Z(z)$?

Example

A receiver output signal Z usually consists of the desired signal X buried in noise Y : $Z = X + Y$. ■

Example

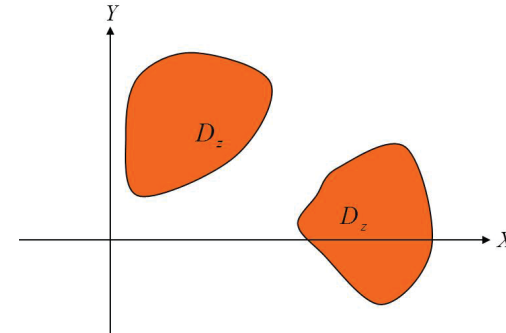
Other possible functions: $X - Y$, XY , X/Y , $\max(X,Y)$, $\min(X,Y)$, $\sqrt{(X^2 + Y^2)}$,... ■

We start with

$$F_Z(z) = P(g(X,Y) \leq z) = P((X,Y) \in D_z),$$

$$= \int \int_{x,y \in D_z} f_{XY}(x,y) dx dy$$

where D_z in the XY plane represents the region such that $g(x,y) \leq z$ is satisfied.



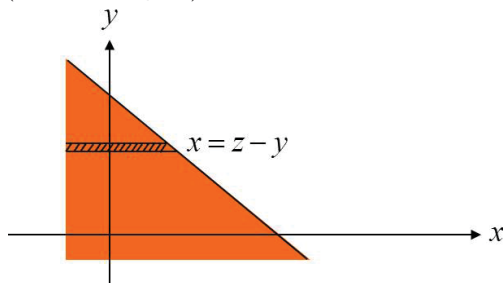
Example

$Z = X + Y$. Find $f_Z(z)$.

$$F_Z(z) = P(X + Y \leq z) = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{z-y} f_{XY}(x,y) dx dy.$$

The region $D_z : x + y \leq z$ is shaded.

Integrating over the horizontal strip along the x -axis first (inner integral) followed by sliding that strip along the y -axis from $-\infty$ to $+\infty$ (outer integral) we cover the entire shaded area.



We can find $f_Z(z)$ by differentiating $F_Z(z)$.

Recall the Leibnitz's differentiation rule: Suppose

$$H(z) = \int_{a(z)}^{b(z)} h(x,z) dx,$$

then

$$\frac{dH(z)}{dz} = \frac{db(z)}{dz} h(b(z), z) - \frac{da(z)}{dz} h(a(z), z) + \int_{a(z)}^{b(z)} \frac{\partial h(x,z)}{\partial z} dx.$$

Using this, we get

$$f_Z(z) = \int_{-\infty}^{+\infty} \left(\frac{\partial}{\partial z} \int_{-\infty}^{z-y} f_{XY}(x,y) dx \right) dy$$

$$= \int_{-\infty}^{+\infty} \left(f_{XY}(z-y, y) - 0 + \int_{-\infty}^{z-y} \frac{\partial f_{XY}(x,y)}{\partial z} dx \right) dy$$

$$= \int_{-\infty}^{+\infty} f_{XY}(z-y, y) dy.$$

Integration can also be carried out first along the y-axis followed by the x-axis

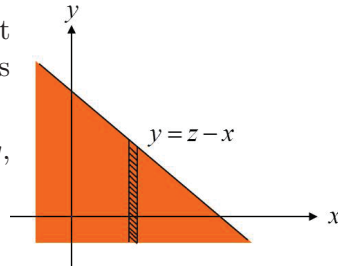
$$F_Z(z) = \int_{x=-\infty}^{+\infty} \int_{y=-\infty}^{z-x} f_{XY}(x, y) dy dx,$$

and differentiation gives

$$\begin{aligned} f_Z(z) &= \frac{dF_Z(z)}{dz} = \int_{x=-\infty}^{+\infty} \left(\frac{\partial}{\partial z} \int_{y=-\infty}^{z-x} f_{XY}(x, y) dy \right) dx \\ &= \int_{x=-\infty}^{+\infty} f_{XY}(x, z-x) dx. \end{aligned}$$

If X and Y independent, $f_{XY}(x, y) = f_X(x)f_Y(y)$ and we get

$$f_Z(z) = \int_{y=-\infty}^{+\infty} f_X(z-y)f_Y(y)dy = \int_{x=-\infty}^{+\infty} f_X(x)f_Y(z-x)dx.$$

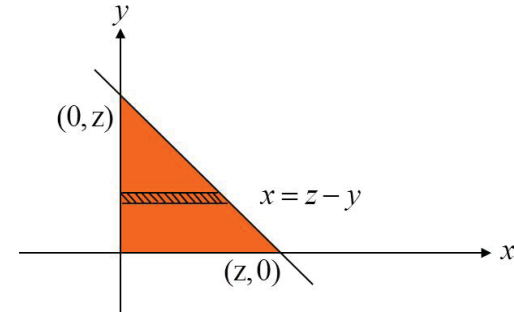


The above integral is the standard convolution of the functions $f_X(z)$ and $f_Y(z)$ expressed two different ways.

Hence, **If two RVs are independent, then the density of their sum equals the convolution of their density functions:** $f_Z = f_X \otimes f_Y$.

Surprising? No, recall that this was already observed from the characteristic functions!

As a special case, suppose that $f_X(x) = 0$ for $x < 0$ and $f_Y(y) = 0$ for $y < 0$, then the new limits for D_z are



In that case,

$$F_Z(z) = \int_{y=0}^z \int_{x=0}^{z-y} f_{XY}(x, y) dx dy$$

or

$$f_Z(z) = \int_{y=0}^z \left(\frac{\partial}{\partial z} \int_{x=0}^{z-y} f_{XY}(x, y) dx \right) dy = \begin{cases} \int_0^z f_{XY}(z-y, y) dy, & z > 0 \\ 0, & z \leq 0. \end{cases}$$

By considering vertical strips, we get

$$F_Z(z) = \int_{x=0}^z \int_{y=0}^{z-x} f_{XY}(x, y) dy dx$$

or

$$f_Z(z) = \int_{x=0}^z f_{XY}(x, z-x) dx = \begin{cases} \int_{x=0}^z f_X(x)f_Y(z-x) dx, & z > 0, \\ 0, & z \leq 0, \end{cases}$$

if X and Y are independent RVs. ■

Example

Consider two independent exponential RVs X, Y with common parameter λ . Determine $f_Z(z)$ for $Z = X + Y$?

$$f_X(x) = \lambda e^{-\lambda x} U(x),$$

$$f_Y(y) = \lambda e^{-\lambda y} U(y),$$

$$f_Z(z) = \int_0^z \lambda^2 e^{-\lambda x} e^{-\lambda(z-x)} dx = \lambda^2 e^{-\lambda z} \int_0^z dx = z \lambda^2 e^{-\lambda z} U(z).$$

■

Change of Variables: Two functions of two RVs

Consider two random variables X and Y characterized by the joint PDF $f_{X,Y}(x, y)$. We want to perform a change of variables and consider random variables U and V obtained as

$$U = R(X, Y), \quad V = S(X, Y).$$

What is the joint PDF $f_{U,V}(u, v)$ of U and V ? Same procedure as that for one function can be applied. However, if the following two conditions are satisfied:

1. we can express X and Y as a function of U and V as $X = L(U, V)$ and $Y = T(U, V)$ (one-to-one correspondence),
2. the determinant of the Jacobian J exists (functions are continuous and differentiable) and is non-zero

$$J = \begin{bmatrix} \frac{\partial X}{\partial U} & \frac{\partial X}{\partial V} \\ \frac{\partial Y}{\partial U} & \frac{\partial Y}{\partial V} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial U} & \frac{\partial L}{\partial V} \\ \frac{\partial T}{\partial U} & \frac{\partial T}{\partial V} \end{bmatrix},$$

then

$$f_{U,V}(u, v) = |\det(J)| f_{X,Y}(x, y).$$

For X and Y independent, the joint PDF is written as

$$f_{X,Y}(x, y) = \begin{cases} e^{-x}e^{-y} = e^{-(x+y)} & x > 0, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

We are looking for the marginal PDF of $U = \frac{X}{X+Y}$.

Option 1: Let us add the RV $V = X + Y$ such that

$$X = UV, \quad Y = V - UV = V(1 - U)$$

$$\det(J) = \det \left(\begin{bmatrix} \frac{\partial X}{\partial U} & \frac{\partial X}{\partial V} \\ \frac{\partial Y}{\partial U} & \frac{\partial Y}{\partial V} \end{bmatrix} \right) = \det \left(\begin{bmatrix} V & U \\ -V & 1 - U \end{bmatrix} \right) = V.$$

We get

$$f_{U,V}(u, v) = \begin{cases} ve^{-v} & v > 0, 0 < u < 1 \\ 0 & \text{otherwise} \end{cases}$$

and the marginal PDF of U

$$f_U(u) = \begin{cases} \int_0^\infty ve^{-v} dv = 1 & 0 < u < 1 \\ 0 & \text{otherwise.} \end{cases}$$

U is uniformly distributed over $[0,1]$.

Common application: given X and Y , find the PDF of $U = R(X, Y)$. In such case, we add an auxiliary RV V that is a simple function of X and Y (i.e. allows an easy computation of J). We then compute $f_{U,V}(u, v)$ and derive the marginal PDF $f_U(u)$ of U .

Example

The lifetime of a machine is distributed as EXPO(1). A manufacture has two machines that work independently. Note their respective lifetime X and Y . Find the PDF of the ratio of the lifetime X of the first machine to the total lifetime $X + Y$.

We can also compute the marginal PDF of V as

$$f_V(v) = \begin{cases} ve^{-v} & v > 0 \\ 0 & \text{otherwise.} \end{cases}$$

We note that U and V are independent.

Option 2: We could have chosen the RV $V = X$ such that

$$f_{U,V}(u, v) = \begin{cases} \frac{v}{u^2} e^{-\frac{v}{u}} & v > 0, 0 < u < 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_U(u) = \begin{cases} 1 & 0 < u < 1 \\ 0 & \text{otherwise.} \end{cases}$$

$$f_V(v) = \begin{cases} e^{-v} & v > 0 \\ 0 & \text{otherwise.} \end{cases}$$

We note that U and V are here not independent. ■

The Rayleigh distribution

Assume X and Y independent and both distributed $\sim N(0, \sigma^2)$, such that

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

We are interested in $U = \sqrt{X^2 + Y^2}$.

Let us add the RV $V = \tan^{-1}\left(\frac{Y}{X}\right)$ such that

$$X = U \cos V, \quad Y = U \sin V$$

$$\det(J) = \det\left(\begin{bmatrix} \frac{\partial X}{\partial U} & \frac{\partial X}{\partial V} \\ \frac{\partial Y}{\partial U} & \frac{\partial Y}{\partial V} \end{bmatrix}\right) = \det\left(\begin{bmatrix} \cos V & -U \sin V \\ \sin V & U \cos V \end{bmatrix}\right) = U.$$

We get

$$f_{U,V}(u,v) = \frac{u}{2\pi\sigma^2} e^{-\frac{u^2}{2\sigma^2}}, \quad 0 < u < \infty, -\pi \leq v \leq \pi.$$

Thus,

$$f_U(u) = \int_{-\pi}^{\pi} f_{U,V}(u,v) dv = \frac{u}{\sigma^2} e^{-\frac{u^2}{2\sigma^2}}, \quad 0 < u < \infty.$$

This is the PDF of a Rayleigh RV with parameter σ^2 , and

$$f_V(v) = \int_0^{\infty} f_{U,V}(u,v) du = \frac{1}{2\pi}, \quad -\pi \leq v \leq \pi.$$

This is the uniform distribution over $[-\pi, \pi]$.

We note that U and V are independent.

If X and Y are zero mean independent Normal random variables with common variance, then $\sqrt{X^2 + Y^2}$ has a Rayleigh distribution and $\tan^{-1}\left(\frac{Y}{X}\right)$ has a uniform distribution. Moreover these two derived RVs are independent.

Alternatively, for X and Y zero mean independent Normal random variables, $X + jY$ represents a complex Normal RV. It follows that the magnitude and phase of a complex Normal RV are independent with Rayleigh and uniform distributions respectively.

Law of Large Numbers and Central Limit Theorem

Events, Probability and Sets

Random Variables and Probability Distributions

Systems and Component Reliability

Jointly Distributed Random Variables

Law of Large Numbers and Central Limit Theorem

Weak Law of Large Numbers

Central Limit Theorem

Statistics

Law of Large Numbers and Central Limit Theorem

We study two fundamental theorems, namely Law of Large Numbers and Central Limit Theorem, concerned with sequences of random variables satisfying the following properties:

- ▶ RVs X_1, \dots, X_n are independent
- ▶ RVs have the same expectation: $E(X_i) = \mu, \forall i$
- ▶ RVs have the same variance: $\text{Var}(X_i) = \sigma^2, \forall i$

Results focus on

$$S_n = X_1 + \dots + X_n, \quad E S_n = n\mu, \quad \text{Var } S_n = n\sigma^2$$
$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}, \quad E \bar{X}_n = \mu, \quad \text{Var } \bar{X}_n = \frac{\sigma^2}{n}$$

Proof: Use Chebyshev's inequality

$$\forall \epsilon > 0 \quad P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \rightarrow 0$$

if $n \rightarrow \infty$. ■

When we make n independent measurements X_1, \dots, X_n of the same quantity μ with the same accuracy $\frac{1}{\sigma^2}$, then the arithmetic mean (or sample mean) of those measurements converges in probability towards μ .

Weak Law of Large Numbers

Definition: A sequence $\{U_n\}$ of RVs U_1, \dots, U_n converges in probability towards θ if

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} P(|U_n - \theta| > \epsilon) = 0$$

or equivalently

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} P(|U_n - \theta| \leq \epsilon) = 1.$$

We denote this convergence as $U_n \xrightarrow{p} \theta$.

Weak Law of Large Numbers (LLN): If $\{X_n\}$ is a sequence of independent RVs with same mean μ and variance σ^2 , then \bar{X}_n converges in probability towards μ , $\bar{X}_n \xrightarrow{p} \mu$.

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| \leq \epsilon) = 1.$$

Bernoulli Theorem: If $Y_1, \dots, Y_n \sim \text{Bernoulli}(p)$, independent, then

$$X = Y_1 + \dots + Y_n \sim \text{Binomial}(n, p)$$

and

$$\frac{X}{n} = \bar{Y}_n = \frac{Y_1 + \dots + Y_n}{n} \xrightarrow{p} p$$

Proof: Use Chebyshev's inequality

$$\forall \epsilon > 0 \quad P\left(\left|\frac{X}{n} - p\right| > \epsilon\right) \leq \frac{np(1-p)}{n^2\epsilon^2} \rightarrow 0$$

if $n \rightarrow +\infty$. ■

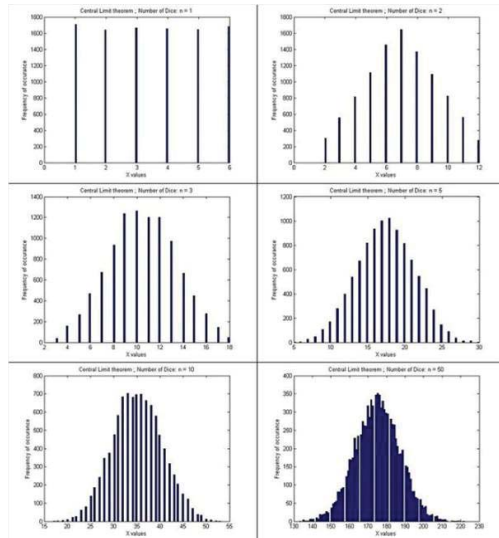
When n is large enough, the relative frequency $\frac{X}{n}$ of an event becomes as close as we want to the probability p of that event.

Central Limit Theorem

For i.i.d. RVs,
 $S_n = X_1 + \dots + X_n$
tends to Gaussian
(Normal) as n goes to
infinity.

Useful in commu-
nications

That's why noise
is usually Gaussian.



Suppose we have n random variables, X_i for $i = 1, 2, \dots, n$, mutually independent and identical (that is each random variable has the same distribution), each having mean μ and variance σ^2 . We can see that

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) = n\mu$$

and

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) = n\sigma^2.$$

It can be shown that, as $n \rightarrow \infty$ the distribution of this sum converges to the normal distribution, that is,

$$X_1 + X_2 + \dots + X_n \rightarrow N(n\mu, n\sigma^2).$$

This is a special case of the *central limit theorem* (CLT).

The central limit theorem is an important result in probability, and is the basis for a big part of classical statistical inference. We can also use it to approximate distributions.

Example

Consider $X_1, \dots, X_n \sim \text{Bernoulli}(p)$, independent, and set

$$Y_n = \sum_{i=1}^n X_i \sim \text{Binomial}(n, p)$$

The distribution of this sum converges to $N(np, np(1-p))$ as $n \rightarrow \infty$. Thus, for large n , the normal distribution provides a reasonable approximation to the binomial distribution.



Example

Consider a consignment of 100 components, and suppose that $P(\text{defective}) = 0.005$. Let X be the number of non-defective components. Then $X \sim \text{Bin}(100, 0.995)$. Computing a probability like $P(X > 75)$ is messy:

$$P(X > 75) = P(X = 76) + P(X = 77) + \dots + P(X = 100),$$

as it requires evaluation of a large number of individual binomial probabilities. However X is approximately normal:

$$N(100 \times 0.995, 100 \times 0.995 \times 0.005) = N(99.5, 0.4975)$$

so

$$P(X > 75) \approx 1 - \Phi\left(\frac{75 - 99.5}{\sqrt{0.4975}}\right)$$



Central Limit Theorem (CLT): If $\{X_n\}$ is a sequence of independent and identically distributed (i.i.d.) with mean μ and variance σ^2 , then

$$P(S_n \leq a) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{a - n\mu}{\sigma\sqrt{n}}\right) \rightarrow P\left(Z \leq \frac{a - n\mu}{\sigma\sqrt{n}}\right)$$

when $n \rightarrow +\infty$ (with $Z \sim N(0, 1)$).

When we want to deal with a RV given by the sum of i.i.d. RVs, we can find for any a an approximation of the probability $P(S_n \leq a)$ by evaluating a probability of a standard normal RV.

Moivre-Laplace Theorem: (Approximation of a Binomial RV by a standard normal) If $X \sim \text{Binomial}(n, p)$, then $X = Y_1 + \dots + Y_n$ is a sum of independent Bernoulli with parameter p , and

$$P(X \leq a) = P\left(\frac{X - np}{\sqrt{npq}} \leq \frac{a - np}{\sqrt{npq}}\right) \rightarrow P\left(Z \leq \frac{a - np}{\sqrt{npq}}\right)$$

if $n \rightarrow +\infty$, with $q = 1 - p$.

Example

Determine the number n of independent measurements of μ , made with the same precision in order to have

$$P\left(|\bar{X} - \mu| \leq \frac{|\mu|}{100}\right) \geq 0.95.$$

By LLN,

$$P\left(|\bar{X} - \mu| \leq \frac{|\mu|}{100}\right) \geq 1 - \frac{\sigma^2}{n\left(\frac{|\mu|}{100}\right)^2} \geq 0.95,$$

which leads to $\frac{\sigma^2}{n}\left(\frac{100}{\mu}\right)^2 \leq 0.05$ and $n \geq 200000\left(\frac{\sigma}{\mu}\right)^2$.

By CLT,

$$\begin{aligned} P\left(|\bar{X} - \mu| \leq \frac{|\mu|}{100}\right) &= P\left(\left|\frac{S_n - n\mu}{\sqrt{n}\sigma}\right| \leq \frac{|\mu|}{100} \frac{\sqrt{n}}{\sigma}\right) \\ &\approx P\left(|Z| \leq \frac{|\mu|}{100} \frac{\sqrt{n}}{\sigma}\right) \geq 0.95. \end{aligned}$$

Since $P(|Z| \leq 1.96) \approx 0.95$, we find $\frac{|\mu|}{100} \frac{\sqrt{n}}{\sigma} \approx 1.96$, i.e. about $n \geq 40000\left(\frac{\sigma}{\mu}\right)^2$. ■

Events, Probability and Sets

Random Variables and Probability Distributions

Systems and Component Reliability

Jointly Distributed Random Variables

Law of Large Numbers and Central Limit Theorem

Statistics

- Statistics and Sampling Distributions
 - Point Estimation
 - Properties of Estimators
 - Method of Moments
 - Maximum Likelihood Estimation

Example

Suppose we independently draw $n = 10$ observations from a standard normal distribution (that is a population with a $N(0, 1)$ distribution), and compute the sample mean and sample standard deviation. Moreover, suppose we repeat this procedure 5 times. In an experiment this gave

Sample	\bar{x}	s
1	0.0757	0.8914
2	0.5549	1.4599
3	-0.066	0.7606
4	0.7433	0.6039
5	-0.5988	1.0473

Note that all these summaries are different, in every case.



Earlier, we discussed various functions as methods of summary for a sample denoted x_1, x_2, \dots, x_n . For example, the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i .$$

We now wish to consider samples as being drawn from a larger population, and using the sample to make statements about unknown parameters (such as the mean) of the population. Before we obtain the data, there is uncertainty about precisely what values will be observed, and we can regard each observation as a random variable, X_1, X_2, \dots, X_n . Before the observations are obtained, there is uncertainty about the value of \bar{x} (or any other summary).

Just as the observations vary from sample to sample, so does any value computed from a sample. Capturing this uncertainty is a key component of statistical analysis. With this concept of uncertainty in mind, we now define a statistic.

A *statistic* is any quantity calculated from sample data. Prior to obtaining data, there is uncertainty as to which value of the statistic will occur. Therefore a statistic is a *random variable*.

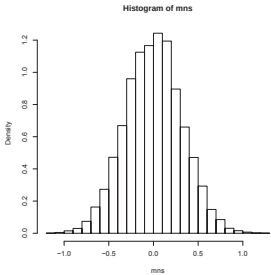
Regarding the sample mean as a statistic means that we can write

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i .$$

When we observe the sample data, we obtain \bar{x} . The same concept applies to other such quantities. Since a statistic is a random variable it has a probability distribution. The probability distribution of a statistic is often called the *sampling distribution*, to stress that it reflects how the statistic varies in value across the possible samples that could be selected.

Example

Consider the means of 10,000 samples of size $n = 10$ drawn from a standard normal distribution. The sampling distribution of the mean in this case can be approximated by the histogram of the collection of means. Note that we know the exact distribution in this case.



The expected value, and theoretical variance of the sampling distribution will often prove important. ■

The sampling distribution of a statistic depends on the parent population, the sample size, and the sampling mechanism. In practical situations, we will often be concerned with samples obtained as follows:

A *random sample* of size n from a distribution with probability mass or density function $f_X(x)$ is a set of independently and identically distributed random variables X_1, X_2, \dots, X_n , each with mass/density function $f_X(x)$.

Sometimes this is abbreviated as ' X_1, X_2, \dots, X_n are iid from $f_X(x)$ '. In practice, we might appeal to randomisation to induce independence (more on this later). In some cases, it is possible to derive the sampling distribution of a statistic analytically.

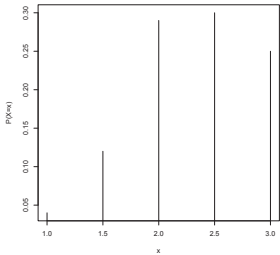
Example

Consider the random variable X , with PMF given by

x	1	2	3
$f_X(x)$	0.2	0.3	0.5

Suppose we draw two numbers X_1 and X_2 independently, according to $f_X(x)$, and are interested in the mean $\bar{X} = (X_1 + X_2)/2$. Independence means we can compute $f_{X_1, X_2}(x_1, x_2) = P(X_1 = x_1 \cap X_2 = x_2)$. The small range of X allows us to consider every pair (X_1, X_2) , and the corresponding mean. The expected value of \bar{X} is $\mu = 2.3$, and its variance is $\sigma^2/n = 0.305$.

x_1	x_2	$p(x_1, x_2)$	\bar{x}
1	1	0.04	1.0
2	1	0.06	1.5
3	1	0.10	2.0
1	2	0.06	1.5
2	2	0.09	2.0
3	2	0.15	2.5
1	3	0.10	2.0
2	3	0.15	2.5
3	3	0.25	3.0



Notice that, on average, the sample mean gives the population value $E(X)$, and the variability of the sample mean is less than the variability of the original distribution. ■

The Distribution of the Sample Mean

Many important statistical procedures are concerned with making statements about the population mean, and are based on the properties of the distribution of the sample mean. From earlier results, if X_1, X_2, \dots, X_n is a random sample from a population with mean μ and variance σ^2 , and \bar{X} is the sample mean, then

$$\begin{aligned} E[\bar{X}] &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \left(\sum_{i=1}^n E(X_i)\right) \\ &= \frac{n\mu}{n} = \mu. \end{aligned}$$

For the variance, we have

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n \text{Var}(X_i)\right) \quad (\text{Independence}) \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

The standard deviation, $\frac{\sigma}{\sqrt{n}}$, is called the *standard error* of the sample mean. Notice that this quantity decreases as the sample size increases.

Statistics

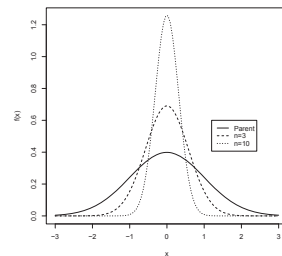
289/321

When the parent population is normal, such that X_1, X_2, \dots, X_n are iid from $N(\mu, \sigma^2)$, then

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

In this case, we know everything about the sampling distribution of \bar{X} , and can compute probabilities of the type $P(a \leq \bar{X} \leq b)$ by suitable standardising.

The figure shows a parent population, and the corresponding sampling distributions for \bar{X} , for samples of size 3 and 10. Notice that, as the sample size increases, the sampling distribution becomes increasingly concentrated about its mean.



Statistics

291/321

Statistics

290/321

Example

The total daily output of a titanium oxide (white pigment) facility is well modelled by a normal distribution, with mean 500 tons, and standard deviation 50 tons. On 25 randomly selected days per year the output is precisely measured, and the sample mean of the 25 measurements recorded as an annual QC measure. What is the probability that this sample mean is between 490 and 510?

Denote the daily output as X , with $X \sim N(500, 50^2)$. We have a random sample of size $n = 25$, so

$$\bar{X} \sim N(500, 50^2/25) \quad \text{and} \quad Z = \frac{\bar{X} - 500}{10} \sim N(0, 1).$$

We can now compute

$$\begin{aligned} P(490 \leq \bar{X} \leq 510) &= P(-1 \leq Z \leq 1) \\ &= \Phi(1) - \Phi(-1) = 0.682. \end{aligned}$$

Statistics

292/321

The Central Limit Theorem

We have seen that the CLT can be used to approximate the sum of independent random variables from any distribution. Using the properties of the normal distribution, we can extend this result to the sample mean: given a random sample of size n from a distribution with mean μ and variance σ^2 , the distribution of \bar{X} is approximately $N(\mu, \sigma^2/n)$ when n is large. The quality of the approximation improves with increasing n .

Statistical inference is usually based on the sampling distribution of a statistic, but there are relatively few choices of parent distribution and statistic for which the sampling distribution can be computed exactly.

However, if our random variables are iid, we will always be able to appeal to the CLT to obtain an approximation to the sampling distribution of the sample mean. A widely accepted heuristic is that for $n \geq 30$ the approximation is sufficiently good to yield reliable results.

Statistics

293/321

In all the examples we have considered so far, the parameters and distribution were known, and we could easily evaluate probabilities such as $P(X > 3)$. Statistical problems involve the reverse situation: we obtain realisations (that is, sample observations) for some unknown probability model, and we wish to estimate the values of the model parameters. Our development will consider situations involving a known distribution, with unknown parameters. Very typically, we wish to estimate the mean of the underlying distribution.

Having observed the sample x_1, x_2, \dots, x_n , we may be interested in:

- ▶ Point estimation, that is, estimating the value of unknown parameters.
- ▶ Interval estimation, that is, estimating an interval that contains the unknown parameters with high probability.

Statistical Analysis

Probability theory provides us with the tools and models to reason about random experiments. When we observe x , we think of it as only one possible value of X , a random variable. The objective of statistical inference is to *learn* about aspects of the probability distribution which produced x , to allow us to draw *inference* about the nature of the process under study.

In this context, inference is concerned with making generalisations from the specific (the sample) to the general (the population). We will often use the notation

$$f(x|\theta)$$

to represent the probability model for x , with θ representing the parameters of the model (for example n and p for a binomial random variable, λ for an exponential). The purpose of this notation is to make explicit that we need to know the value of θ in order to evaluate the probability of a particular observation

x .

Statistics

294/321

Point Estimation

The objective of point estimation is to use a sample of data to provide a *guess* of an unknown population parameter. We want to ensure that our guesses are as good as possible, in a number of senses.

Example

Suppose we are interested in the true average lifetime of a specific type of spark plug. Call this population mean value μ . We then obtain a random sample x_1, x_2, \dots, x_9 of $n = 9$ sparkplugs. The sample mean \bar{x} can then be used to make statements about μ . Other statistics can also be used to make related statements; for example the sample variance s^2 provides information about the population variance σ^2 . ■

Sample realisations and statistics are usually denoted with Roman symbols, like \bar{x}, s , whereas unknown population parameters are denoted with Greek symbols, like μ, θ .

A *point estimate* of an unknown population parameter θ is a single number which represents our “best guess” of θ . A point estimate arises from selecting a suitable statistic and computing its value from a sample. The selected statistic is called a *point estimator* for θ , and is commonly denoted by $\hat{\theta}$ (read ‘theta-hat’).

As stated previously, a statistic is a function that refers to a random sample. An *estimator* is a statistic used to guess the value of some unknown parameter.

This description raises the important point that a number of possible estimators may be suitable for the same parameter (cf. different measures of central tendency).

Example

A chemical plant produces a compound that contains a particular constituent B . A random sample of size $n = 11$ batches of this compound was obtained and the percentages of B in each were measured:

5.3, 4.9, 6.2, 5.7, 4.8, 5.4, 6.1, 6.3, 5.6, 5.5, 7.2

Assume that the normal distribution with mean μ provides a plausible model for the percentage of B . Since this distribution is symmetric, we could estimate μ using the sample mean, $\bar{x} = \frac{1}{n} \sum_{i=1}^{11} x_i = 5.72$, or the sample median $\tilde{x} = (x_5 + x_6)/2 = 5.6$, or the midpoint between the extremes $(x_{11} + x_1)/2 = 6$, or many others.

All these estimators are plausible. We cannot assess their quality directly, since the value of the parameter is unknown, so we are forced to select among estimators using other criteria.

Properties of Estimators

The statistical solution to this problem is to recall that all statistics, as functions of random samples, have a sampling distribution. The sampling distribution provides a description of the random behaviour of the statistic, and thereby facilitates choice of estimator.

Example

Shooting analogy – are the sights consistently off?
Are the shots very erratically spread around the target?

Two criteria that sometimes lead to accurate estimators are unbiasedness and minimum variance.

Estimator bias

A point estimator $\hat{\theta}$ is an *unbiased* estimator of θ if $E(\hat{\theta}) = \theta$, for all θ . The difference $E(\hat{\theta}) - \theta$ is known as the *bias* of $\hat{\theta}$ and thus unbiased estimators have bias 0.

Note that we use the generic notation $\hat{\theta}$ to refer to an estimator of θ . The definition states that the sampling distribution of an unbiased estimator has expected value equal to θ , the population value. Thus, if we could repeat the experiment an infinite number of times, on average an unbiased estimator would yield the correct answer.

Provided we are confident about the distribution of the parent population, we can assess the bias of $\hat{\theta}$, without knowing the value of θ .

Example

Components of a certain type are either operable or defective. Suppose we are interested in the random variable X that counts the number of defectives in a standard batch of 50 components. Now, a batch is randomly selected, the components tested, and $X = 3$ defectives found. This is a binomial experiment, and we wish to estimate the unknown p . A plausible estimator is

$$\hat{p} = X/n,$$

yielding the estimate $3/50 = 0.06$. Now,

$$E(\hat{p}) = E(X/n) = E(X)/n = np/n = p$$

Thus \hat{p} is an unbiased estimator of p . ■

We have seen earlier that $E(\bar{X}) = \mu$, so the sample mean is always an unbiased estimator of the population mean μ , whatever the distribution.

Statistics

301/321

Statistics

302/321

$$\begin{aligned} &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2(\bar{X} - \mu)n(\bar{X} - \mu) \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right]. \end{aligned}$$

Now, taking expectations,

$$\begin{aligned} E(S^2) &= \frac{1}{n-1} \left(\sum_{i=1}^n E[(X_i - \mu)^2] - n E[(\bar{X} - \mu)^2] \right) \\ &= \frac{1}{n-1} (n\sigma^2 - n\text{Var}[\bar{X}]) && \text{since } E(\bar{X}) = \mu \\ &= \frac{1}{n-1} \left(n\sigma^2 - n\frac{\sigma^2}{n} \right) && \text{since } \text{Var}(\bar{X}) = \sigma^2/n. \end{aligned}$$

Statistics

303/321

Statistics

304/321

When choosing among estimators, all else being equal, we generally prefer those that are unbiased, since they are, on average, correct. Recall the formula for the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Treating this as an estimator of σ^2 is often preferred to the version with denominator n , for the following reason. First, we formulate this as a statistic, and obtain the computational formula

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu + \mu - \bar{X})^2 \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 + \sum_{i=1}^n (\bar{X} - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) \right] \end{aligned}$$

Thus,

$$E(S^2) = \frac{1}{n-1} [\sigma^2(n-1)] = \sigma^2.$$

and we see that the sample variance is unbiased. The alternative formula, with n in the denominator, is biased, though there is little difference between the two for large n .

When we look at interval estimation for normal distributions we will find other reasons to consider the sample variance as an estimator for σ^2 .

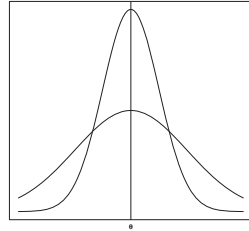
Note that unbiasedness can be difficult to establish for many types of estimator.

Minimum variance estimators

For choosing among estimators, the next characteristic of the sampling distributions to consider is the variance.

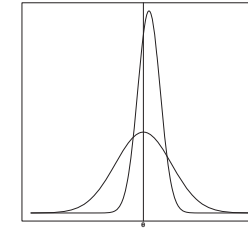
When selecting among unbiased estimators of θ , we often favour the estimator with minimum variance; the *minimum variance unbiased estimator* (MVUE) of θ , since the minimum variance estimator has the most concentrated distribution.

A key aspect of mathematical statistics is concerned with developing tools for identifying MVUEs.



Other estimation issues

While the MVUE is desirable, it is often possible to obtain an estimator with small bias, but appreciably smaller variance (see figure).



For this reason, we prefer (wherever possible) to choose the estimator which minimises the *mean squared error* (MSE), defined as

$$\text{MSE}_\theta(\hat{\theta}) = \text{E}[(\hat{\theta} - \theta)^2].$$

MSE, variance and bias

Notice that

$$\begin{aligned} \text{MSE}_\theta(\hat{\theta}) &= \text{E} \left[\left\{ (\hat{\theta} - \text{E}(\hat{\theta})) + (\text{E}(\hat{\theta}) - \theta) \right\}^2 \right] \\ &= \text{E} \left[(\hat{\theta} - \text{E}(\hat{\theta}))^2 + 2(\hat{\theta} - \text{E}(\hat{\theta}))(\text{E}(\hat{\theta}) - \theta) + (\text{E}(\hat{\theta}) - \theta)^2 \right] \\ &= \text{E} \left[(\hat{\theta} - \text{E}(\hat{\theta}))^2 \right] + 2(\text{E}(\hat{\theta}) - \theta) \text{E}(\hat{\theta} - \text{E}(\hat{\theta})) + (\text{E}(\hat{\theta}) - \theta)^2 \end{aligned}$$

The first of these terms is the variance of $\hat{\theta}$, the second is zero, because $\text{E}(\hat{\theta} - \text{E}(\hat{\theta})) = 0$, and the third is the square of the bias of $\hat{\theta}$. We write

$$\text{MSE}_\theta(\hat{\theta}) = \text{Var}(\hat{\theta}) + \left(\text{Bias}_\theta(\hat{\theta}) \right)^2.$$

Standard errors

Earlier, we mentioned the standard error of the sample mean; the standard deviation of the sampling distribution of the mean. This concept extends to any estimator:

The *standard error* of estimator $\hat{\theta}$ is the standard deviation $\sqrt{\text{Var}(\hat{\theta})}$. Sometimes, the standard error will depend on unknown parameters, that must be estimated. In this case, we have the *estimated standard error* of the estimator.

We will use the generic notation $\sigma_{\hat{\theta}}$ to refer to the standard error, and either $\hat{\sigma}_{\hat{\theta}}$ or $s_{\hat{\theta}}$ for the estimated standard error. This seems complicated, but keep in mind that we are just talking about the standard deviation of the probability distribution of a statistic.

Example

Recall the example involving the compound with constituent B . We had the following data

5.3, 4.9, 6.2, 5.7, 4.8, 5.4, 6.1, 6.3, 5.6, 5.5, 7.2

We assumed that the parent population was normal with unknown mean μ and unknown variance σ^2 .

We know that $\bar{X} \sim N(\mu, \sigma^2/n)$. However, σ^2 is unknown, so we use the sample variance as an estimator of σ^2 :

$$s^2 = \frac{1}{n-1} \sum_{x=1}^n (x_i - \bar{x})^2 = 0.476.$$

Thus, the estimated standard error is $s_{\bar{X}} = 0.218$.



We have talked about desirable characteristics of estimators, without specifying how to construct estimators. A number of methods are available for automatically constructing estimators, including

- ▶ Method of moments
- ▶ Least squares
- ▶ Maximum likelihood (ML)

In general, ML estimation is preferred in statistics because of desirable theoretical properties, although they can sometimes require significant computation in realistic contexts.

Method of Moments

The method of moments attempts to equate sample characteristics with corresponding population theoretical values. The method of moment estimators are the solutions to such equations.

Consider a random sample X_1, X_2, \dots, X_n from PDF (or PMF) $f(x)$. Recall that the k th moment of X is $m_k = E(X^k)$. The sample equivalent of the theoretical moments are the *sample moments*,

$$\frac{1}{n} \sum_{i=1}^n X_i^k, \quad \text{for } k = 1, 2, \dots$$

The first theoretical moment is $E(X) = \mu$ and the corresponding sample moment is $\frac{1}{n} \sum_{i=1}^n X_i$. The second theoretical moment is $E[X^2]$ and the corresponding sample moment is $\frac{1}{n} \sum_{i=1}^n X_i^2$.

Example

Consider a random sample $X_1, \dots, X_n \sim \text{Exp}(\lambda)$. This could refer to waiting times at a service center, for example. The first population moment is the mean,

$$m_1 = E(X) = \frac{1}{\lambda}$$

and the first sample moment is \bar{X} . To obtain the method of moments estimator, we equate these:

$$\bar{X} = \frac{1}{\hat{\lambda}}$$

and thus the estimator is $\hat{\lambda} = 1/\bar{X}$.



The problem gets more complicated when there are more parameters to estimate since we need to obtain sufficient equations to identify the required number of parameters.

Example

Consider a random sample of size n from $N(\mu, \sigma^2)$. We have theoretical moments

$$m_1 = E(X) = \mu \quad \text{and} \quad m_2 = E(X^2) = \sigma^2 + \mu^2,$$

and corresponding sample moments $\frac{1}{n} \sum_{i=1}^n X_i$ and $\frac{1}{n} \sum_{i=1}^n X_i^2$. The method-of-moments estimator for the population mean is thus

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X},$$

unsurprisingly.

For the variance, we have:

$$\begin{aligned} \hat{\sigma}^2 + \hat{\mu}^2 &= \frac{1}{n} \sum_{i=1}^n X_i^2 \\ \Rightarrow \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 \quad \text{subst. } \hat{\mu} \\ &= \frac{1}{n} \left[\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right] \end{aligned}$$

This is the usual weighted sum of squared deviations about the mean formulation. Note that the method-of-moments estimator for the variance is biased, i.e. $E(\hat{\sigma}^2) = \sigma^2 + \mu^2 - E(\bar{X}^2) \neq \sigma^2$. ■

Maximum Likelihood Estimation

Generally, maximum-likelihood estimation is the recommended approach, since the resulting estimators have desirable properties. The approach is built around the joint probability or density of the collection of random variables in a sample.

Consider the sample X_1, X_2, \dots, X_n with joint density or PMF

$$L(\theta) = f(X_1 = x_1, \dots, X_n = x_n | \theta)$$

where θ refers to unknown parameters. When the joint probability is regarded as a function of θ , it is called the *likelihood function*.

The value of θ that maximises the likelihood function is known as the *maximum-likelihood estimators* (MLE) of θ . The estimates evaluated for observed values yield *maximum-likelihood estimates*.

The maximum-likelihood estimators give the values of θ that agree mostly closely with the observed data. By independence, the joint density/mass function of a random sample decomposes into a product of the marginal density/mass functions.

Example

Consider a random sample of size n from an exponential distribution with parameter λ . The maximum likelihood estimator for λ is obtained from the likelihood function:

$$\begin{aligned} L(\lambda) &= f(x_1, x_2, \dots, x_n | \lambda) = \prod_{i=1}^n f(x_i | \lambda) \\ &= \left(\lambda e^{-\lambda x_1} \right) \left(\lambda e^{-\lambda x_2} \right) \dots \left(\lambda e^{-\lambda x_n} \right) \\ &= \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \end{aligned}$$

We seek to maximise the likelihood as a function of λ .

A useful step here is to take logs to obtain the so-called *log-likelihood* function. Since logs are monotonic, the maximum of $L(\theta)$ will coincide with the maximum of $\ell(\theta) = \log L(\theta)$.

$$\ell(\lambda) = \log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n x_i$$

Since we seek a maximum, we differentiate

$$\frac{d \log L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$

and equate with zero

$$\frac{n}{\hat{\lambda}} - \sum_{i=1}^n x_i = 0 \implies \hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i}$$

Of course, we should verify that this is a maximum:

$$\frac{d^2 \log L(\theta)}{d\theta^2} = -\frac{n}{\lambda^2} < 0$$

Note that this is in agreement with the method-of-moments estimator. ■

Example

Consider the random sample X_1, X_2, \dots, X_n from a normal distribution, $N(\mu, \sigma^2)$. Recall

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

The likelihood function is then

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2} (\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i-\mu)^2}$$

The log-likelihood is

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

We seek the values of μ and σ^2 that jointly maximise the log-likelihood. Consider μ first:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu).$$

Equating with zero yields

$$\frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (x_i - \hat{\mu}) = 0 \implies \sum_{i=1}^n x_i - n\hat{\mu} = 0,$$

and the MLE is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Checking the second derivative verifies that this is a maximum. For σ^2 we have

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2,$$

which we equate with zero to obtain

$$-\frac{n}{2} \frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^n (X_i - \hat{\mu})^2 = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Again, we examine the second derivative to ensure we have a maximum. Note the estimator for the variance is again, biased. ■

The general ML procedure is then

1. write, and simplify, the likelihood
2. take logs
3. differentiate and equate the resulting equations with zero and solve
4. examine the second derivative