Mahidol University

MATHEMATICAL MODELING And RISK ANALYSIS

Man Van Minh Nguyen

This page is left blank intentionally.

MATHEMATICAL MODELING And

RISK ANALYSIS

© Man VM. Nguyen

Department of Mathematics

Faculty of Science

This page is left blank intentionally.

Contents

List of Figures

This page is left blank intentionally.

List of Tables

Chapter 5

PETRI NETWORKS - *From Process Mining View*

[[42]]

INTRODUCTION

KEY TERMS reminded

- `Data Science` aims to answer the following four questions.

  1. Reporting: What happened?
  2. Diagnosis: Why did it happen?
  3. Prediction: What will happen?
  4. Recommendation: What is the best that can happen?

- `Process Science` refers to the broader discipline that combines knowledge from *information technology* (IT) and knowledge from *management sciences* (ManSci) to improve and run operational processes.

  WHY? Business processes have become more complex, heavily rely on information systems, and may span multiple organizations. Therefore, process modeling has become of the utmost importance.

OUTLINE

1. INTRODUCTION to Process Mining
2. PETRI NETWORKS- Background

3. BEHAVIOR of PETRI NETS
4. NETWORK STRUCTURES and TYPICAL PROBLEMS in Petri Nets

5. SUMMARIZED OUTCOMES and REVIEWED PROBLEMS
6. ASSIGNMENT on MODELING by PETRI NETS
7. INSTRUCTIONS for working out the assignment [on health service management]

Key References

- R1: Part I of **PROCESS MINING**, 2nd edition, 2016, Springer, by `W.M.P. Aalst, van der`[1]
- R2: **Process mining techniques and applications**- A systematic mapping study, in Expert Systems With Applications, **Vol 133** (2019) 260-295, Elsevier, by `Cleiton dos Santos Garcia et. al.`
- R3: Chapter 4-9 of the text **DATA MINING**- Concepts, Models, Methods, and Algorithms,

  3rd Edition, 2020, IEEE press and Wiley, by `Mehmed Kantardzic`.
- R4: **Modeling Business Processes**: A Petri Net-Oriented Approach, 2011 Massachusetts Institute of Technology, by Wil van der Aalst and Christian Stahl

[1] W.M.P. Aalst is a full professor in Process Analytics and a full professor in Process Science, both at TUe (the Eindhoven University of Technology). He is also a full professor at RWTH Aachen University.

5.1 INTRODUCTION to Process Mining

5.1.1 *WHAT is Process Mining (PM)?*

Process mining emerged as a new research field that focuses on the analysis of processes using **event data** since 1990s. Historically it was *work-flow mining*. Most relevant disciplines are

1. *Data science*: an interdisciplinary field aiming to turn data into real value. It is the core of many partially overlapping (sub)disciplines:
   - algorithms, business models, data mining
   - data transformation, storage and retrieval of information
   - (computing) infrastructures: databases, distributed systems
   - privacy- security law, predictive analytics (learning, predictions)
   - statistics (modeling+ inference), can be viewed as the origin of data science (**DS**)...

2. *Process science*- **PS**, Figure 5.1, a broad term aims to combine

   knowledge from information technology and knowledge from management sciences

   to run and improve operational processes.

   In process science, basic sub-processes are **event**, not number.

   Research objective is changed [from sample numerical/digital data to event data], so the coupling methods are changed, upgraded to a more sophisticated scale, and as a result, applications are broader.

3. *Process Mining*- the missing link between DS and PS.

   With the key study objective of **event & event data**, Process Mining (**PM**) can be seen as a mean (a research methodology with mixture of modern disciplines) to bridge the gap between data science and process science.

   The concept of EVENT (in process science), popularly called **Internet of Events**- IoE, [2] is a newly structural extension of *classical sample data*, i.e. *numerical-digital observations* in Data Science.

[2] coined from 2014, early beginning of the AI-based disruptive technology era

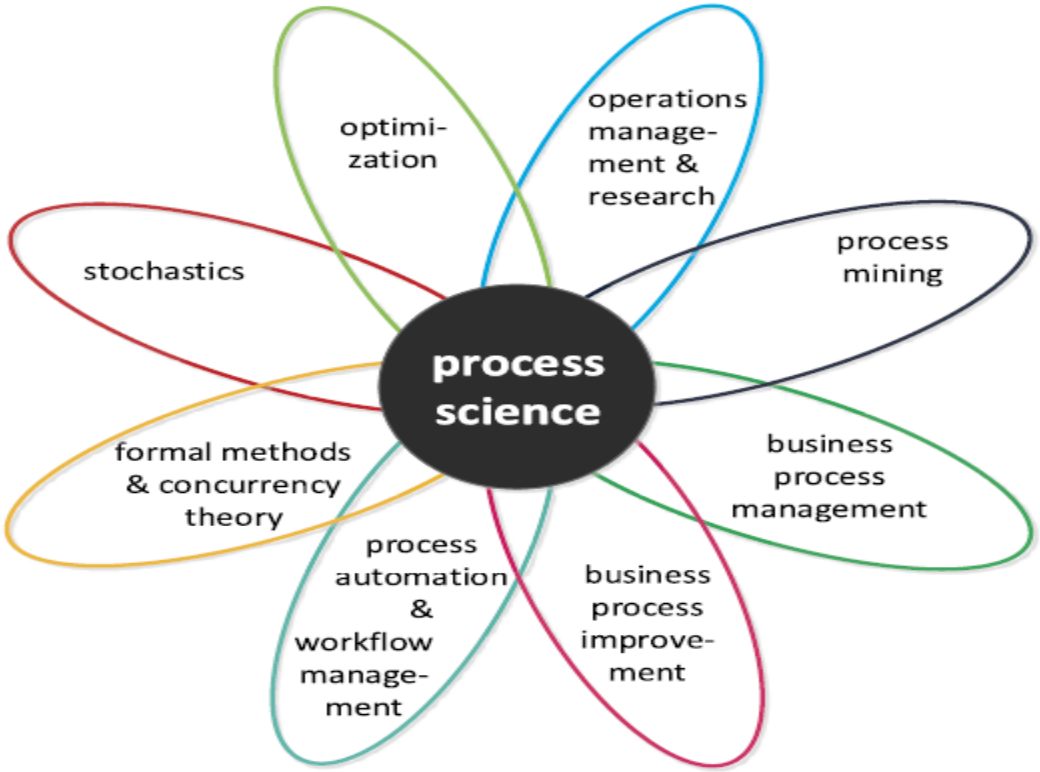
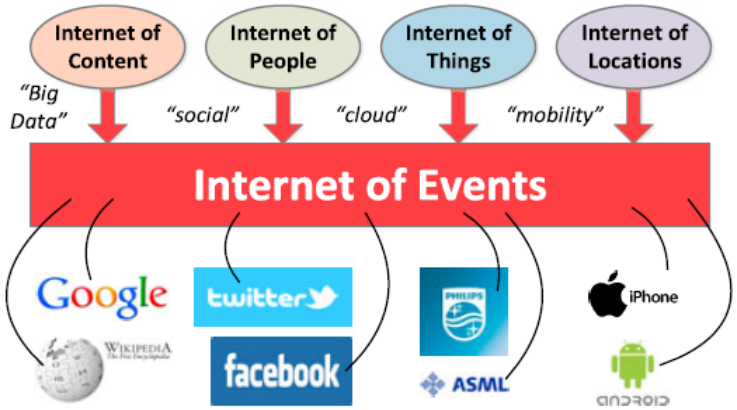


Figure 5.1: *Process science and its constituents*

In general, the **IoE** include 4 dimensions (constituents), possibly overlapping, as follows.

Content, i.e., all information created by humans to increase knowledge on partic-



Event data nowadays are generated from a variety of sources connected to the Internet [courtesy Wil van der Aalst (2016)]

ular subjects. The IoC includes traditional web pages, articles, encyclopedia-Wikipedia, YouTube, e-books, newsfeeds...

People, i.e., all data related to personal, human-being subject with their social interaction (e-mail, Facebook, Twitter, forums, LinkedIn, etc.

The Internet of Things (IoT), i.e., all physical objects connected to the network, or broadly, relevant hardware which allow an event or sub-process happens.

Locations refers to all data that have a geographical or geo-spatial dimension. [3]

[3] In Data Science and Data Mining (now a mature discipline) we have mostly focused on few dimensions, say Content and Locations, of the newly

Discussion on drawbacks of approaches:

Data science and its key component, data mining, however are data-centric, **not** process-centric.

Data mining, Statistics, Machine learning and the likes technically do not consider end-to-end process models. Process science approaches are process-centric, but often focus on modeling rather than learning from event data, see next figures.

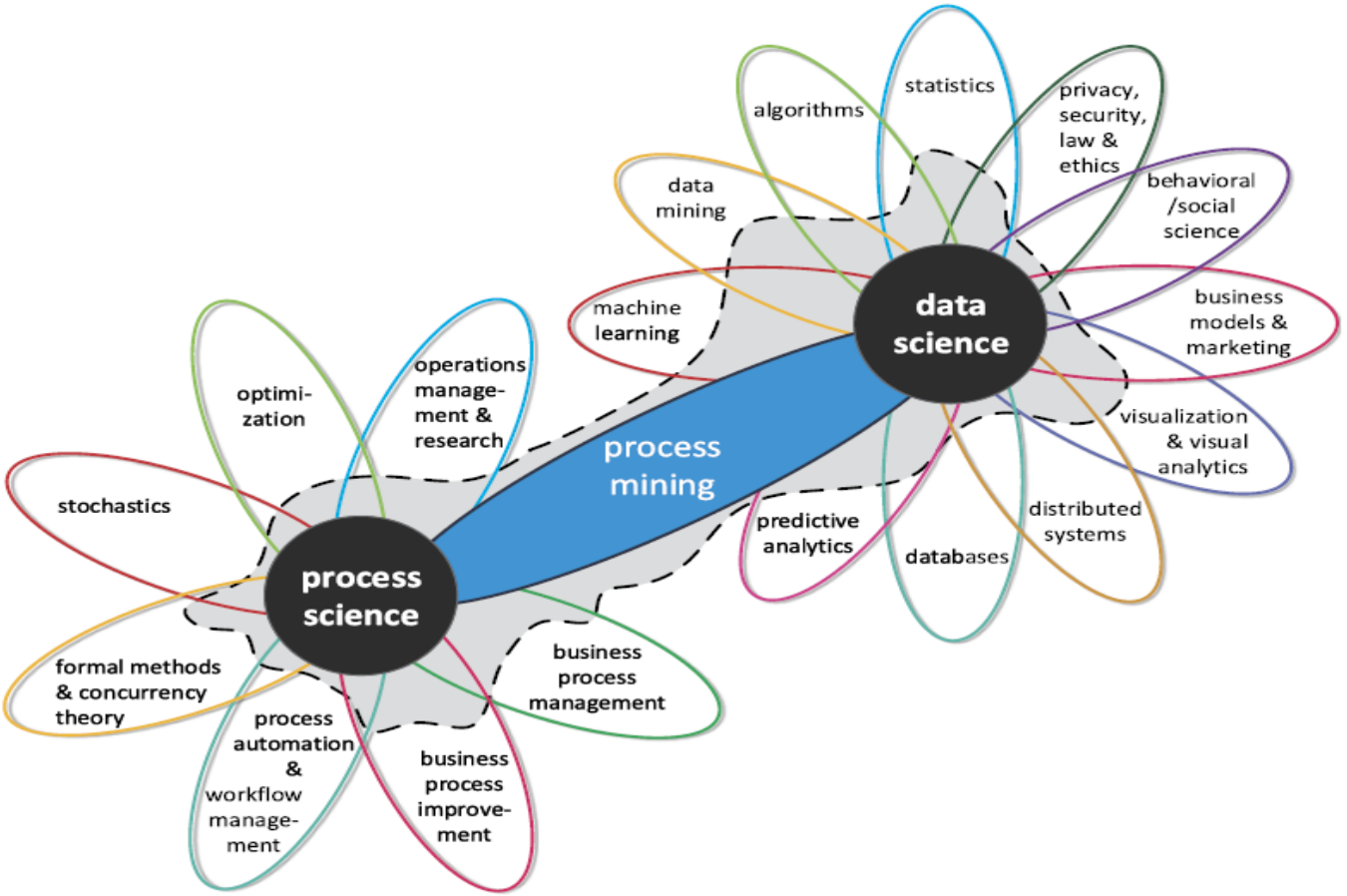
5.1.2 *WHY? DS and PS to Process Mining (PM)*

Structurally, PM is viewed as the core of a flower, with the branches

- business process management and improvement (BPM & I), business intelligence (BI), as Petri nets, Lean Six Sigma, TQM
- formal methods and concurrency theory, machine learning
- optimization, operations (management and research),
- process automation and work-flow management,

defined fours.

- stochastic process and analysis, as Markov models,
- statistical inference and optimization, time series model and analysis, and so on.



Process mining as the bridge between data science and process science

Data science revisited, aims to answer the following four questions.

1. Reporting: What happened?
2. Diagnosis: Why did it happen?
3. Prediction: What will happen?
4. Recommendation: What is the best that can happen?

- However, without incorporating the methods of PS with those of DS, the answers of the above key fours could be **not fully** adequate. To adequately answer such questions there is **not just** a need for raw data and computing power.
- Expertise in data+process mining: probability/statistics/stochastics, inference and causality analysis, and visualization to **efficiently decode diverse complex datasets** from biology, chemistry, computing, environment, epidemiology, medicine, plant science, urban traffic, virology... are vital.

A brief spectrum of techniques for old and new applications

PM only recently emerged as a subdiscipline of both data science and process science,

but the corresponding techniques can be applied to any type of operational processes (organizations/systems), as

- analyzing treatment processes in hospitals,
- improving customer service processes in a multinational corporation,
- understanding the behavior of customers choosing an insurance firm,
- improving the **efficacy of a new vaccine** to cope with fatal pandemics, by firstly well knowing lab's experimental designs, better modeling data obtained from labs/field trips/factories, in order to perform convincing statistical analyses, and finally to make meaningful guidelines or (nearly) optimal decisions.

A fact-based conclusion:

The coworkers must know things at process level (at least at 4 dimensions of EVENT DATA $\sim$ [Content, People, Technology/tool, Location]) to do the right thing!

5.1.3 *HOW to bring PM to life? Main challenges*

Possibly to be figured out by the upcoming interdisciplinary teams, or center built up to cope with urgent problems, to maintain Univ. X position ...**What are key challenges?**

I) Define the PM research spectrum (breath of topics),

II) Create a consistent and explicit process model given an *event log*,

III) Diagnose issues observing dynamic behavior with the use of tools?

1. Furthermore, processes and information need to be aligned perfectly in order to meet requirements related to compliance[4], efficiency, and customer service. How to define and build clear causal bonds so that decisions could be made?
2. The identification of issues and diagnoses needs explore
   the causal and casual (occasional) relations between activities,
   and this functionality is **not** present in a traditional *Workflow Management System* (WFMS) or Business Process Management System (BPMS).

[4] the action or fact of complying with a wish or command

5.1.4 *Process Mining - A brief development history and Top applications*

First of first, we need to provide an overview of development history and then describe the **process mining spectrum** (research topics). (See a brief story in ref R2 summarizing these previous studies between 2003 and 2018.) Notably

- 2015 - *Compliance monitoring in business processes*: Functionalities, application, and tool-support
- 2016 - *The State of the Art of Business Process Management Research*: related to research methods, quality discussion, maturity, citations index, and progress in the business process management
- 2016 - *Process mining in healthcare*, Biomedical Informatics.
- 2018 - *A systematic mapping study of process mining*: maps the relationship between data mining tasks in the process mining context...

Top seven areas with $> 80\%$ publications related to PM applications [ref. R2]:

α **Healthcare:** covering clinical path, patient treatment, or processes of a hospital

β **ICT:** related to software development, IT operation services

γ **Manufacturing:** in industrial activities, realized by a factory that usually receives material and delivers partial or finished products

δ **Education:** e-learning, scientific applications, and centers with innovation process management.

ε **Finance and** η **Logistics:** see more in R1 and R2

λ **Robotics / Smart:** advanced technologies related to smart buildings, industry 4.0...

5.1.5 *Process Mining (PM) key research topics*

1. Process Discovery
2. Process Conformance
3. Process Enhancement... (kindly see ref. R1 and R2)

From the Quality Engineering view, including process control and improvement, PM **does not** replace the traditional process improvements methods, such as Business Process Improvement (BPI),

Continuous Process Improvement (CPI), Corporate Performance Management (CPM), Total Quality Management (TQM), Six Sigma, and others.

However, process miners are able (a) to check compliance, diagnose deviations, point out bottlenecks, and then (b) to perform, integrate, accelerate process improvements, as well as recommend actions and, last but not least redesign systems.

Process Discovery [see R1]

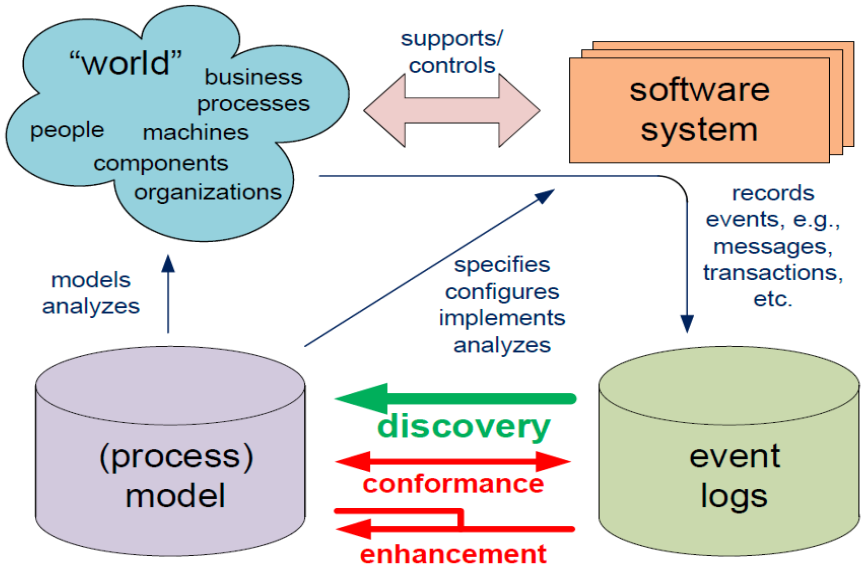
The process discovery should balance four competing quality criteria:

simplicity, fitness (able to replay event log)

precision (avoid underfitting) and generalization (not overfitting the log).

5.1.6 *Conclusion*

1. Process mining is a new research discipline as well as cutting edge technology enabling evidence-based process analysis. The three basic types of process mining currently are Process Discovery, Process Conformance and Process Enhancement.



Process Discovery

2. Nevertheless, there are still many open scientific challenges and most end-user organizations are not yet aware of the potential of process mining.
3. To cope with new real live challenges, IEEE, with Data Mining Technical Committee of the Computational Intelligence Society (CIS) created a **Task Force on Process Mining** since 2010s, see more at `https://www.tf-pm.org/`.
4. Two major concerns in process mining - **Concurrency**[5] **and Causality** will be stud-

[5] the fact of two or more events or circumstances happening or existing at the same time/ in Computing: the ability to execute more than one

ied.

For concurrency (i.e., parallelism), one of the most essential problems in PM, we are going to investigate in details the methodology and key techniques of **Petri Net** in next sections. The matter of causality will be the subject in Chapter **??**.

program or task simultaneously..

PETRI NETWORKS

[[42]]

5.2 PETRI NETWORKS- Background

5.2.1 *The Art of Modeling- motivated from Operation Research*

Operation research (OpRe), is a branch of management science heavily relying on **modeling**. Here a variety of mathematical models ranging from

- (I. deterministic modeling): integer linear programming, dynamic programming, to
- (II. stochastic modeling): Markov chains, queueing models, to stochastic dynamic programming, and
- (III. mixed type one): as simulation [discrete event simulation, MCMC simulation].

ELUCIDATION

- Models are used to reason about *processes* (redesign) and to make decisions *inside processes* (planning and control). The models used in operations management are typically tailored towards a particular analysis technique and only used for answering a specific question.

In contrast, process models in Business Process Management typically serve *multiple purposes*.

- However, creating such models is therefore a difficult and delicate task, since concurrency exists. In such complex process models, **concurrency** must be handled properly. The most popular theory for studying concurrency is **Petri Net**, named after the German Carl Adam Petri. [6]

Carl Adam Petri (1926- 2010)

A German scientist and engineer,
one of the renown pioneers in *Computing*,
and a visionary who founded an extraordinarily
fruitful domain of study in the field of
distributed discrete event systems.

[*courtesy* Karsten Wolf, [39]]

[6] Carl Adam Petri (1926- 2010), the first computer scientist to identify *concurrency* as a fundamental aspect of computing (sketched largely in his seminal PhD thesis, title **Communication with Automata**, submitted to the Science Faculty of Darmstadt Technical University in 1962, where in fact he outlined a whole new foundations for computer science). Petri's father was a serious scholar. He had a PhD in mathematics and had met Minkowski and Hilbert.

IMPACTS on Science & Engineering

Petri nets nowadays have brought engineers a breakthrough in their treatment of *discretely controlled systems*. Petri nets are a key to solve the *design problem*, as this is the first technique to allow for a unique description, as well as powerful analysis of **discrete control systems**.

■ CONCEPT 1. *WHAT is Petri Net (PetN)?*

A **Petri net** is a graphical tool [a bipartite graph consisting of *places* and *transitions*] for the description and analysis of *concurrent processes* which arise in systems with many components (distributed systems).

The graphics, together with the rules for their coarsening and refinement, were invented in August 1939 by Carl Adam Petri.

5.2.2 *Formal definitions*

Definition 5.1 (Transition system or State transition system)**.**

A transition system is a triplet $TS = (S, A, T)$ where S is the set of *states*, $A \subseteq \mathcal{A}$ is the set of *activities* (often referred to as *actions*), and $T \subseteq S \times A \times S$ is the set of *transitions*.

The following subsets are defined implicitly,

- $S^{start} \subseteq S$ is the set of *initial states* (sometimes referred to as 'start' states), and $S^{end} \subseteq S$ is the set of *final states* (sometimes referred to as 'accept' states).

For most practical applications the state space S is finite. In this case the transition system is also referred to as a Finite-State Machine or a finite automaton (FA), see Section **??**.

- WHY transition systems? The goal of (using transition systems in) a **process model** is to decide *which activities* need to be executed and in *what order*.

Activities can be executed sequentially, activities can be optional or concurrent, and the repeated execution of the same activity may be possible.

BEHAVIOR of a transition system

Given a transition system one can reason about its **behavior**.

The transition starts in one of the initial states.

Any path in the graph starting in such a state corresponds to a possible *execution sequence*.

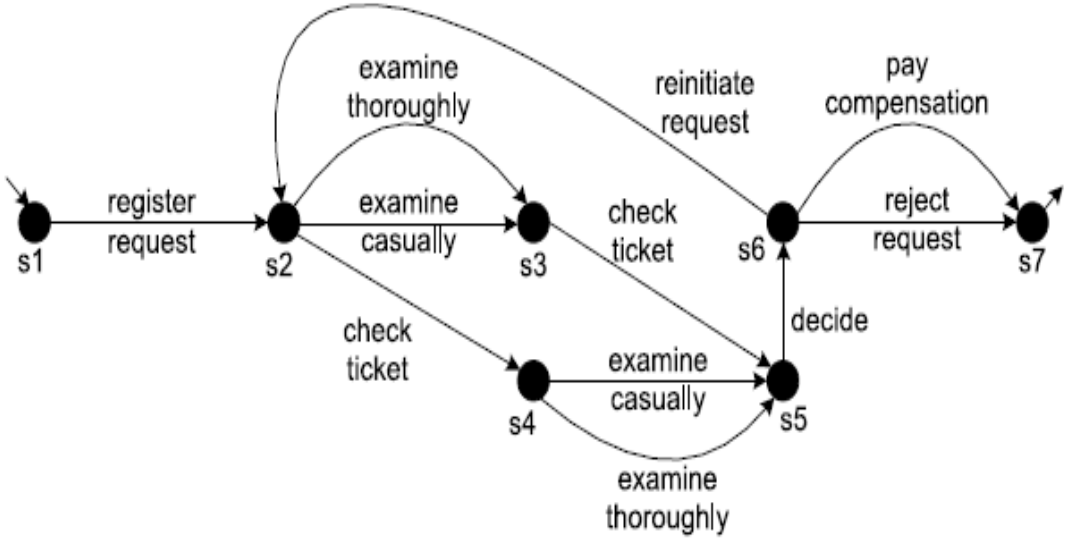
* A path *terminates successfully* if it ends in one of the final states.

* A path *deadlocks* if it reaches a non-final state without any outgoing transitions.

(Note that the absence of deadlocks does not guarantee successful termination).

♦ **EXAMPLE 5.1.**

Observe a transition system in Figure 5.2 we see the state space $S = \{s_1, s_2, \ldots, s_7\}$.



A transition system having one initial state and one final state

Figure 5.2: *A small size transition system*

Here, $S^{start} = \{s_1\}$, $S^{end} = \{s_7\}$.

Could the reader fill in fully the set of *activities*

$A = \{$ register request, examine thoroughly, examine casually, $\ldots\}$, and completely determine the set $T = \{(s_1, \text{register request}, s_2), (s_2, \text{examine casually}, s_3), \ldots\}$ of all *transitions*? ■

NOTES on using transition system:

1. **Any process model with executable semantics** can be mapped onto a transition system. Therefore, many notions defined for transition systems can easily be translated to higher-level languages such as Petri nets...
2. Transition systems, however, are simple but have problems expressing concurrency succinctly, as 'state explosion'. But a **Petri Net** can be used much more compactly and efficiently. Indeed, suppose that there are n parallel activities, i.e., all n activities need to be executed but any order is allowed. [7]

On the set of all multisets $\mathcal{M}$ over a domain D

Given a finite domain $D = \{x_1, x_2, \cdots, x_k\}$,

a map $X : D \longrightarrow \mathbb{N}$ defines a multi-set on D as follows:

for each $x \in D$, $X(x) = m$ denotes the number of times x is included in the multi-set,

[7] There are $n!$ possible execution sequences. The transition system requires 2^n states and $n \times 2^{n-1}$ transitions. When $n = 10$, the number of reachable states is $2^n = 1024$, and the number of transitions is $n \times 2^{n-1} = 5120$. A **Petri Net** needs only 10 transitions and 10 places to model the 10 parallel activities.

i.e.,

$$M = \{\underbrace{x_1, \ldots, x_1}_{m_1 \ times}, \ \cdots, \underbrace{x_k, \ldots, x_k}_{m_k \ times}\}. \tag{5.1}$$

Evidently, $support(M) \subseteq D$ and we could use multiplicative format for

$$M = \{x_1^{m_1}, x_2^{m_2}, \ldots, x_k^{m_k}\} \longrightarrow M = [m_1, m_2, \ldots, m_k],$$

meaning M is identified with the list $[m_1, m_2, \ldots, m_k]$.

Here frequencies $m_i \geq 0$, $m_i = 0$ means that x_i does not appear in M, and $support(M)$ consists of **different elements** in the multi-set M.

♦ EXAMPLE 5.2.

A multi-set (also referred to as *bag*) is like a set in which each element may occur multiple times, and the order is **not** matter.

Given domain $D = \{a, b, c, d, e\}$, $k = |D| = 5$, we observe a multi-set with $n = 9$ elements: one a, two b's, three c's, two d's, and one e: $M = [a, b, b, c, c, d, d, e, c] = \{a, b^2, c^3, d^2, e\} \equiv [1, 2, 3, 2, 1\}$.

Definition 5.2 (**Petri Net** is a bipartite directed graph N of *places* and *transitions*)**.**

A Petri net is a triplet $N = (P,\ T,\ F)$ where P is a finite set of *places*,

T is a finite set of *transitions* such that $P \cap T = \emptyset$, and

$F \subseteq (P \times T)\ \cup (T \times P)$ is a set of directed arcs, called the *flow relation*.

1. A *token* is a special *transition node*, being graphically rendered as a black dot, •. The symbolic tokens generally denote elements of the real world. Places can contain tokens, and transitions **cannot**.
2. A transition is *enabled* if each of its input places contains a **token**.

   [Example: node `enter` in figure 5.3 has input places `wait` and `free`; node `leave` is not enabled]

3. *Firing*: An **enabled transition** can *fire*, thereby consuming (*energy* of) one token [e.g., node `enter` in figure 5.3] from each *input place* and producing at least one token for each *output place* next.
4. A *marking* is a distribution of tokens across places.

   A *marking* of net N is a function $m : P \to \mathbb{N}$, assigning to

   each place $p \in P$ the number $m(p)$ of tokens at this place. [e.g. $m(\texttt{wait}) = 3$.]

   Denote $M = m(P)$, the range of map m, viewed as a multiset.
5. A *marked* Petri net is a pair (N, M), where $N = (P, T, F)$ is a Petri net and where M is a *multi-set* (or *bag*, defined generally in Equation 5.1) over P denoting the *marking* of the net.

   ♦ We write the set of all multisets over P as $\mathcal{M}(P)$ or $\mathcal{M}$ for short.

   ♦ The set of all marked Petri nets is denoted $\mathcal{N}$.

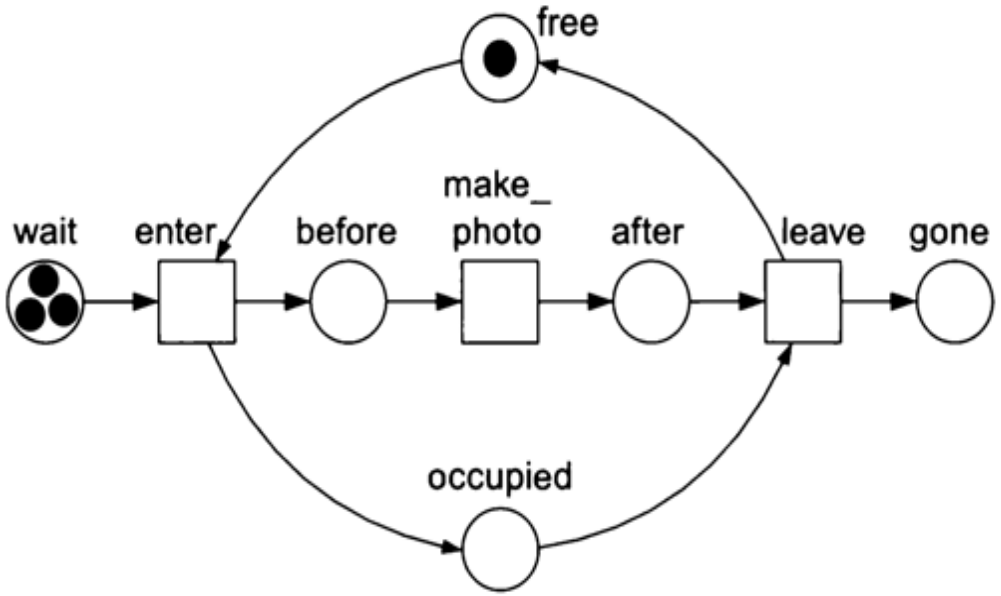


Figure 5.3: *A Petri net for the process of an X-ray machine*

♦ EXAMPLE 5.3.

The Petri net in figure 5.3 has three transitions (drawn as □):

$$T = \{enter,\ make - photo,\ leave\} \equiv \{t_1, t_2, t_3\}.$$

- Transition `enter` is **enabled** if there is at least one token in place `wait` and at least

one token in place `free`. In the marking of this net, these conditions are fulfilled. Transition `make-photo` is enabled if place `before` holds at least one token. This condition is **not** fulfilled.

QUESTIONS:

- List the places P, and give the marking of the Petri Net in figure 5.3.

$$P = \{\texttt{wait}, \texttt{before}, \texttt{after}, \ldots, \texttt{occupied}\} \equiv \{p_1, p_2, \ldots, p_6\}.$$

$$M = m(P) = Range(m) = [m(\texttt{wait}), m(\texttt{before}), \ldots, m(\texttt{occupied})$$

$$\Longrightarrow \; M = m(P) = Range(m) = [3, 0, 0, 0, 1, 0]$$

- Determine fully the flow relation $F \subseteq (P \times T) \cup (T \times P)$. ■

ELUCIDATION

1. PLACES: In a Petri net, graphically, a place $p \in P$ is represented by a circle or ellipse.

   A place p always models a *passive* component:

   p can store, accumulate or show things. A place has discrete states.

2. TRANSITIONS: The second kind of elements of a Petri net are *transitions*.

   Graphically, a transition $t \in T$ is represented by a square or rectangle.

   A transition t always models an *active* component:

   t can produce things/tokens, consume, transport or change them.

   After each firing of a transition [consuming energy of token] the tokens are reallocated on places, henceforth building up the dynamic of **Petri net**.

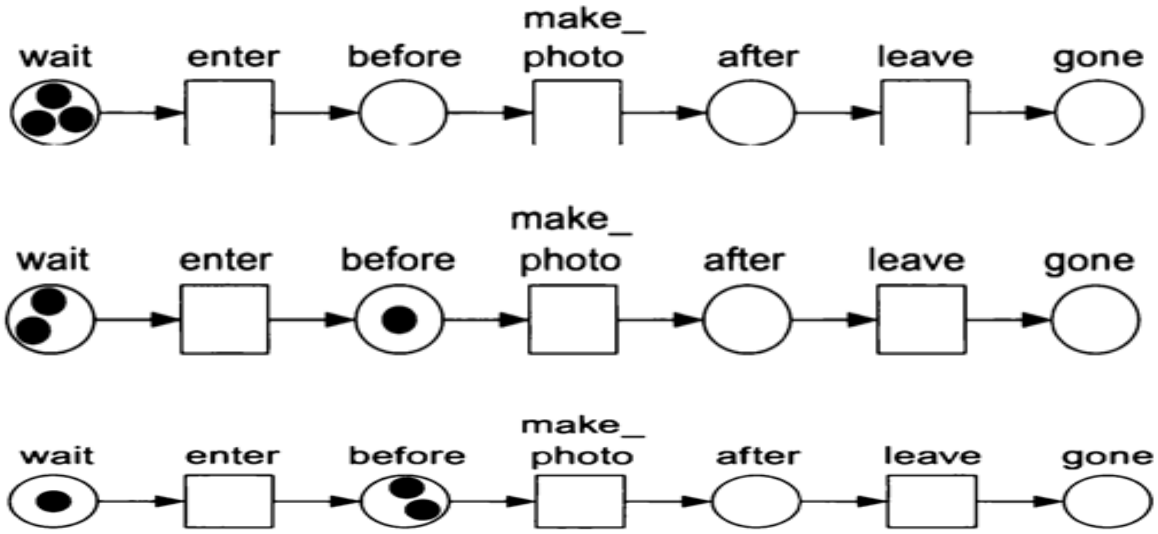
   A redistribution or reallocation of tokens across places is a **marking** [Figure 5.4].

3. ARCS: Places and transitions are connected to each other by directed arcs, graphically, represented by an arrow. An arc **never** models a system component, but an abstract, sometimes only notional relation between components such as *logical connections*, or *access rights*.

4. OPERATIONS: The sum of two multi-sets ($A \uplus B$), the difference ($A \setminus B$), the presence of an element in a multi-set ($x \in M$), and the notion of subset ($X \leq Y$) are defined in the classic way of set theory. [8]

[8] What is multi-set used for? A marking corresponds to a multi-set of **tokens**. However, multi-sets are **not only** used to represent markings; later

♦ **EXAMPLE 5.4.** *Figure 5.4 shows a* **Petri net** *with three different markings.*



The first three markings in a process of the X-ray machine

a) [top, **transition *enter*** not fired**];** b) [middle**, transition *enter*** fired**];** and

c) [down, **transition *enter*** has fired again**]**

Figure 5.4: *Three different markings on a Petri net, modeling of a process of an X-ray machine*

Find the places P, and give the transitions T of the net.

Write down completely three different markings in format of lists or tables.

we will use multi-sets to model *event logs* where the same trace may appear multiple times.

Definition 5.3 (Input is place, output is transition.)**.**

Let $N = (P, T, F)$ be a Petri net. Elements of $P \cup T$ are called **nodes**.

- A node x is an *input node* of another node y if and only if there is a directed arc from x to y (i.e., $(x, y) \in F$).

  Node x is an *output node* of y if and only if $\exists\ (y, x) \in F$.

- For any $x \in P \cup T$, write $\bullet x = \{y |\ (y, x) \in F\}$, and call **the preset** of x, and $x \bullet = \{y |\ (x, y) \in F\}$- **the postset** of x.

- For a set X, define X^* to be the set of sequences containing elements of X, i.e., $(x_1, x_2, \cdots, x_n) \in X^*$ for any $n \in \mathbb{N}$ and entries $x_1, x_2, \cdots, x_n \in X$.

Q: Can you give $\bullet\, c1 = ?$, $c5 \bullet = ?$ in Figure 5.5.

♦ **EXAMPLE 5.5** (On Enabled transition and Marking changes)**.**

The marked Petri net in Figure 5.5 has the (initial) marking M_0 with only one token, node **start**:

$$M_0 = m(P) = [1, 0, 0, 0, 0, 0, 0] \equiv [\texttt{start}].$$

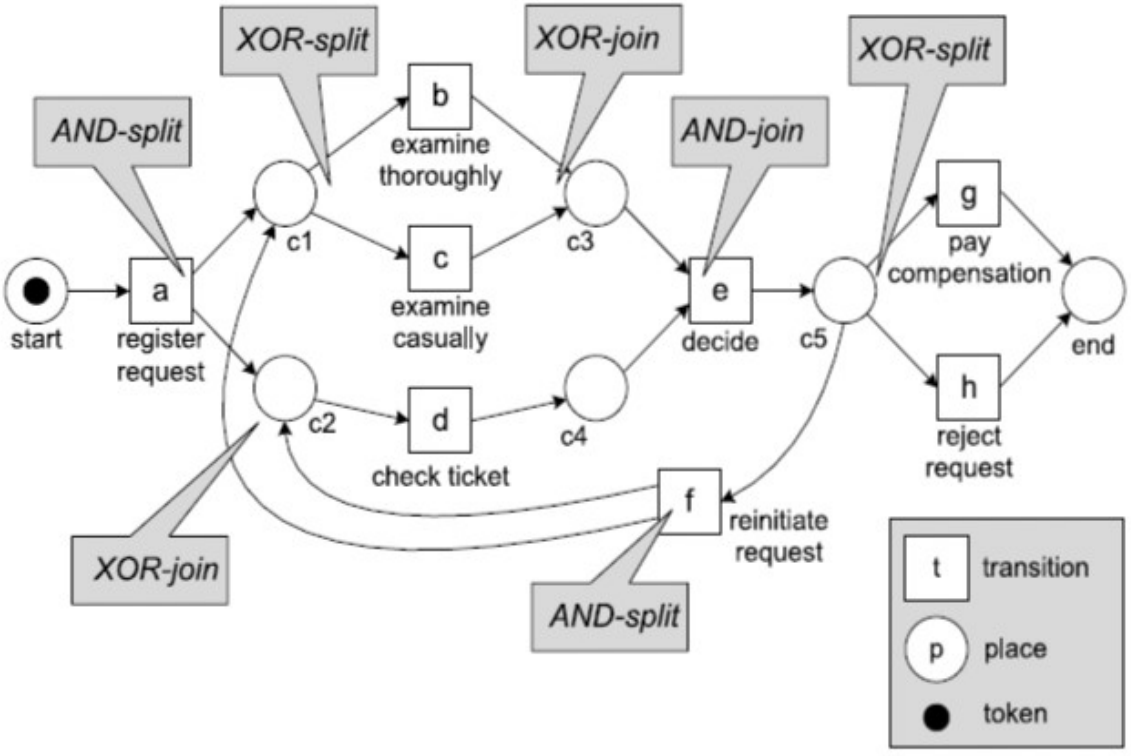


Figure 5.5: *A marked Petri net with one initial token*

[*courtesy* Wil van der Aalst, [**?**]]

- Hence, transition **a** is enabled at marking [start], now a becomes new token (with full energy)! Firing **a** results in the marking [c1, c2]: one token is consumed and two tokens are produced.

  At marking [c1, c2], transition **a** is no longer enabled (spent all energy now).

  However, transitions **b**, **c**, and **d** have become enabled.

- From marking [c1, c2], firing **b** results in marking [c2, c3] = [0, 0, 1, 1, 0,0,0].

  Here, **d** is still enabled, but **b** and **c** not anymore.

  Because of the loop construct involving f there are infinitely many firing sequences starting in [*start*] and ending in [*end*].

- Now with multiple token at the beginning, assume that the initial marking is: [$start^5$].

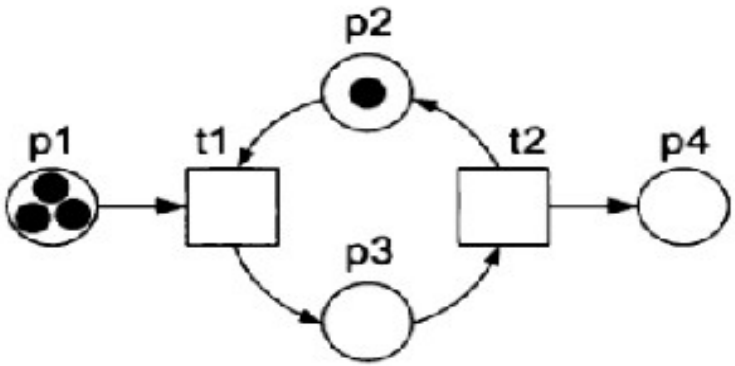
  Firing **a** now results in the marking [$start^4$, c1, c2]. At this marking **a** is still enabled.

  Firing **a** again results in marking [$start^3$, $c1^2$, $c2^2$] = [3,2,2, 0,0,0,0].

  Transition **a** can fire 5 times in a row resulting in [$c1^5$, $c2^5$]. Note, after the 1st occurrence of **a**, also **b**, **c**, and **d** are enabled and fire, but only (**b**, **d**) concurrently...

PRACTICE 5.1.

Consider the Petri net in figure below.



A Petri net showing transitions t1 and t2.

Figure 5.6: *A simple Petri net, with only two transitions*

1. Define the net formally as a triple (P, T, F).
2. List presets and postsets for each transition.
3. Determine the marking of this net.
4. Are the transitions t1 and t2 enabled in this net.

5.2.3 *Important usages of* **Petri net** *via explaining Figure 5.5*

Important usages: Information systems (IS) and business process modeling (BPM).

Business Process (BP)

- An *organization* is a system consisting of humans, machines, materials, buildings, data, knowledge, rules and other means, with a set of goals to be met. Most organizations have, as one of their main goals, the *creation* or *delivery* of (physical) *products* or (abstract) *services*.
- The creation of services and products is performed in *business processes* (BP).
  A BP is a set of *tasks* with *causal dependencies* between tasks.

Task ordering principles The five basic task ordering principles (Figure 5.5) are

1. *Sequence* pattern: putting tasks in a linear order;
2. *XOr-split* pattern: selecting only one branch to execute; node `c1`,
3. *And-split* patterns: all branches will be executed; node `f`

4. *XOr-join* patterns: only one of the incoming branches should be ready in order to continue, node `c3`; and

5. *And-join* pattern: all incoming branches should be ready in order to continue, node `e`

Execution of tasks For the execution of tasks *resources* are required.

- Resources can either be *durable* or *consumable*.

  The 1st kind - *Durable*- is available again after execution of one or more tasks.

  Examples of this kind are humans, machines, computers, tools, information & knowledge.

  Two kinds of durable resources are of particular importance: the humans as a resource, called *human resources*, and information and knowledge, which we will call *data resources*.

  The 2nd, Consumable resources disappear during the task execution. Examples are energy, money, materials, components and data.

- The results or output of a task can be considered as resources for subsequent tasks or as final products or services.
- Since human activities are sometimes replaced by computer systems,

  we use the term *agents* as a generic term for human and system resources.

Definition 5.4 (Information Systems and Modeling Business Processes)**.**

To study **Petri net** we first formalize the above concepts.

1. A **business process** consists of a *set of activities* that is performed in an organizational and technical environment. These activities are coordinated to jointly realize a business goal.

   Each business process is enacted by a single organization, but it may interact with business processes performed by other organizations.
2. An **information system** is a software system to capture, transmit, store, retrieve, manipulate, or display information, thereby supporting people, organizations, or other software systems.

The awareness of the importance of business processes has triggered the introduction of the concept of *process-aware information systems*.

The most notable implementations of the concept of process-aware information systems are workflow management systems. A workflow management system is configured with a **process model**, its graphical visualization is **workflow net**.

SUMMARY 1.

A **Petri net** is a triplet structure (P, T, F). The structure of a Petri net is determined if we know the places P, the transitions T, and the flow relation F of the ways in which they are connected with each other (i.e. arcs connecting places and transitions.).

1. A **Petri net** contains zero or more places. Each place has a unique name, the place label. We can describe the places of a Petri net by the set P of *place labels*.

   We can describe the transitions in a Petri net in the same way. Each transition has a unique name, the transition label.

We describe the transitions of a Petri net by the set T of *transition labels*.

2. **Transitions** are the *active* nodes of a Petri net, because they can change the marking through firing. We therefore name transitions with *verbs* to express action.

   E.g., see node **b**, **c**, and **d** in Figure 5.7.

   Places are the *passive* nodes of a Petri net because they **cannot** change the marking.

   We name places using nouns, adjectives, or adverbs.

3. In addition to the places and transitions, we must describe the *arcs*. Like a transition in a transition system, we can represent an arc as an ordered pair (x, y). The set of arcs is a *binary relation*.

   As a **Petri net** has two kinds of arcs, we obtain two binary relations.

   (i) The binary relation $R_I \subseteq P \times T$ contains all arcs connecting transitions and their input places.

   (ii) The binary relation $R_O \subseteq T \times P$ contains all arcs connecting transitions and their

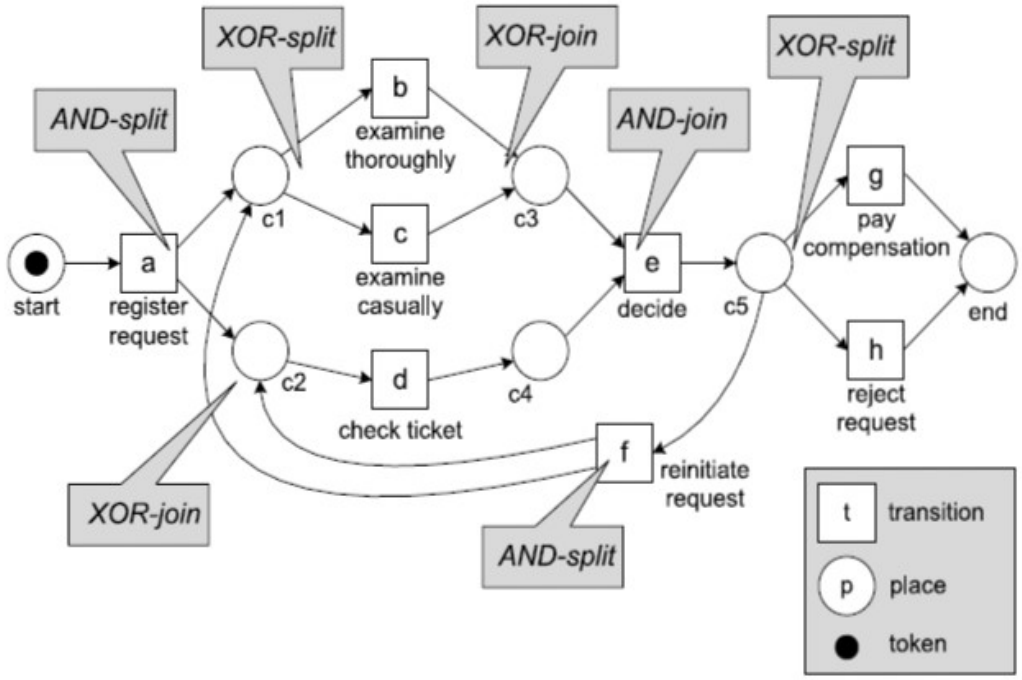


Figure 5.7: *A marked Petri net with one initial token*

output places.

The union $R_I \cup R_O$ represents all arcs of a Petri net. This union is again a relation

$$F = R_I \cup R_O \subseteq (P \times T) \cup (T \times P) \tag{5.2}$$

called the **flow relation**, where $(p, t) \in F$ defines the arc from p to t.

4. **Labeling of Arcs and Transitions**:

   Arcs and transitions can be labeled with *expressions* (for instance, $-$, a subtraction, and variables x and y). These expressions have two central properties:

   (i) If all *variables* in an expression are replaced by *elements*, it becomes possible to evaluate the expression in order to obtain yet another element.

   (ii) The variables in these expressions are *parameters* describing different instances ("modes") of a transition.

   Such a transition can **only occur** if its labeling evaluates to the logical value '**true**'.

5. We can summarize the possible roles of tokens, places, and transitions with the following modeling guideline:

   We represent events as transitions, and we represent states as places and tokens.

   We represent the evolving states of a system by the distribution of tokens over the places. Each token in a place is part of the state. A token can model

a physical object,

information,

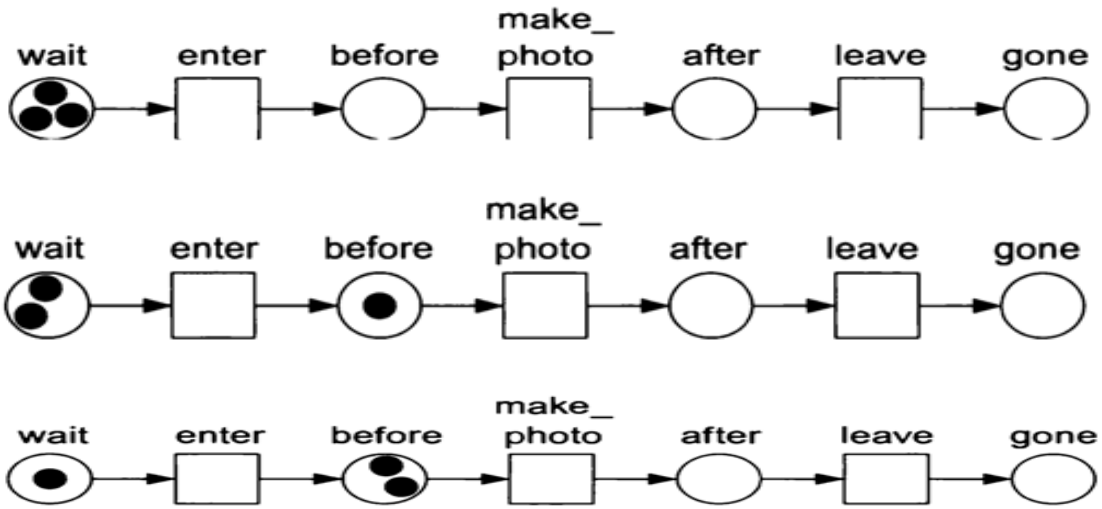
or a combination of the two,

but it can also model the state of an object or a condition. See more in Section 5.6.

In the remainder we only consider Petri nets (special class of workflow nets), to model BPs. For many purposes it is sufficient to consider classical Petri nets, i.e. with '**black**' tokens. **The modeling of *colored Petri nets* is postponed till next texts**.

Practical Problem 1.

Given a process of a X-ray machine in which we assume the first marking in Figure 5.8.a shows that there are three patients in the queue waiting for an X-ray. Figure 5.8.b depicts the next marking, which occurs after the firing of transition enter.



The first three markings in a process of the X-ray machine

a) [top, **transition *enter*** not fired**];** b) [middle**, transition *enter*** fired**];** and

c) [down, **transition *enter*** has fired again**]**

Figure 5.8: *A Petri net model of a business process of an X-ray machine*

1. Determine the two relations R_I and R_O, and the flow relation $F = R_I \cup R_O$. [HINT: find sets P, T.]
2. A patient may enter the X-ray room only after the previous patient has left the room. We must make sure that places **before** and **after** together do not contain more than one token.

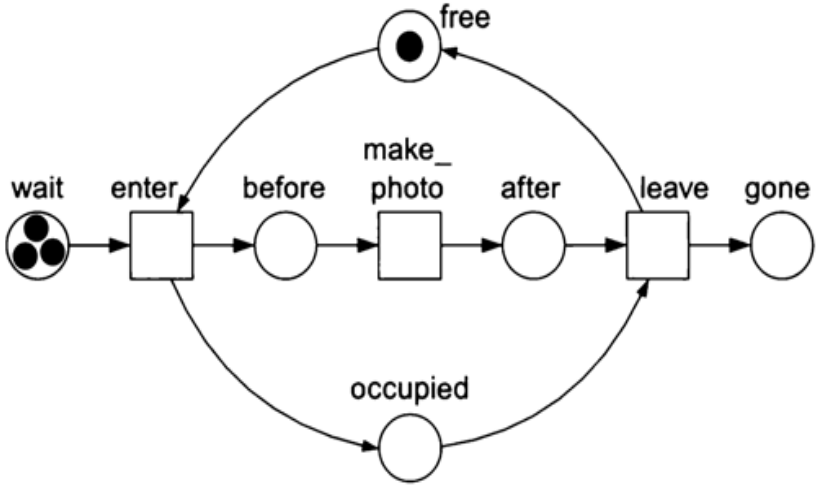


Figure 5.9: *An improved Petri net for the business process of an X-ray machine*

There are two possible states: the room can be **free** or **occupied**. We model this by adding theses two places to the model, to get the improved the Petri net, in Figure

5.9. Now for this Petri net, can place **before** contain more than one token? Why? Rebuild the set P of place labels.

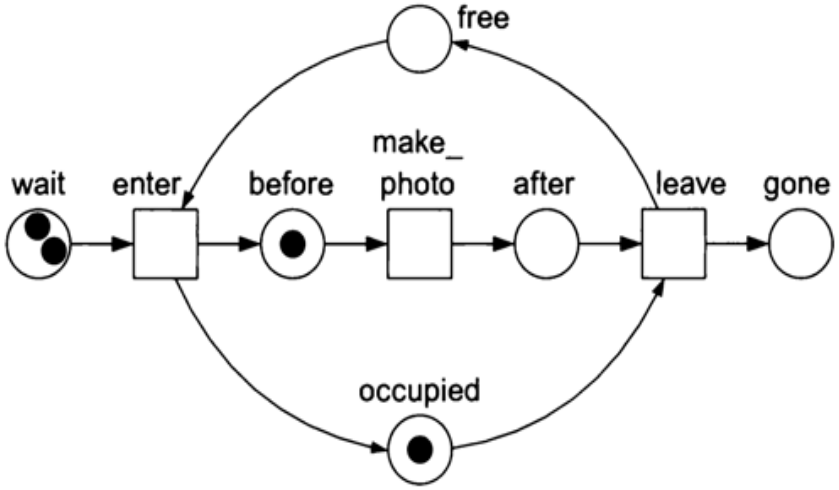


Figure 5.10: *The marking of the improved Petri net for the working process of an X-ray room after transition* ***enter*** *has fired.*

3. As long as there is no token in place **free** [Figure 5.10], can transition **enter** fire again? Explain why or why not. Remake the two relations R_I and R_O.

The Behavior of Petri Nets

[[42]]

The behavior of a **Petri Net** is defined by the net structure, the distribution of tokens over the places P, and the firing of transitions T.

5.3 BEHAVIOR of PETRI NETS

5.3.1 *From Firing, Reachability to Labeled Petri net*

A token is graphically rendered as a black dot in the graph of a Petri net.

Definition 5.5 (*Firing rule*- as an ordering binary relation on $\mathcal{N}$)**.**

Let $(N, M) \in \mathcal{N}$ be a marked Petri net with $N = (P, T, F)$ and $M \in \mathcal{M}$.

- A transition is enabled if there is **at least one token in each of its input places**.
- Transition $t \in T$ is *enabled* at marking M, denoted $(N, M)\,[t\rangle$,

  if and only if $\bullet t \leqslant M$.
- The firing rule $\alpha \quad [t\rangle \quad \beta \subseteq \mathcal{N} \times T \times \mathcal{N}$ is the smallest relation satisfying

$$(N, M)[t\rangle \implies \underbrace{(N, M)}_{\alpha} \quad [t\rangle \quad \underbrace{\left[N, (M \setminus \bullet t) \uplus t \bullet\right]}_{\beta} \tag{5.3}$$

  for any $(N, M) \in \mathcal{N}$ and any $t \in T$.

♣ **OBSERVATION 1.**

1. Places can contain tokens, transitions cannot. But transitions can change the marking through firing. That is places are passive, and transitions are active.

   To fire a transition, it must be enabled.

2. **Enabledness**: A marking $M = [m_1, m_2, \ldots, m_k] \equiv m : P \to \mathbb{N}$, by (5.1), has the form

$$M = \{\underbrace{x_1, \ldots, x_1}_{m_1 \ times}, \ \cdots \ , \underbrace{x_k, \ldots, x_k}_{m_k \ times}\},$$

   so we can equivalently say $t \in T$ is *enabled* at M, written $\bullet t \leqslant M$

   if and only if for all places $p \in \bullet t$, then $m(p) > 0$.

3. An enabled transition t can fire, thereby changing the marking M to a marking $M_1 = (M \setminus \bullet t) \uplus t \bullet$.

   If an enabled transition fires, then it consumes one token from each of its input places and produces one token in each of its output places.

4. $(N, M)[t\rangle$ means that transition t is enabled at marking M.

   E.g., $(N, [start])\ [a\rangle$ in Figure 5.11.

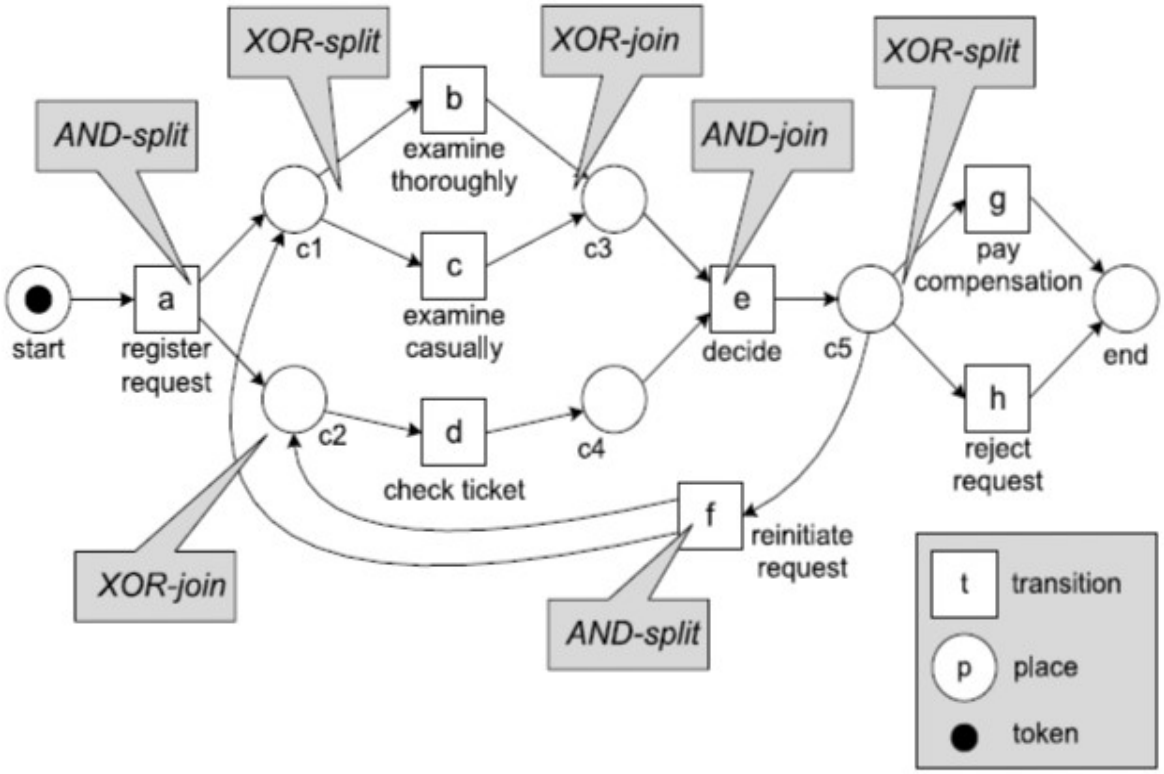


Figure 5.11: *A marked Petri net with one initial token*

5. $(N, M)\ [t\rangle\ (N, M_1)$ denotes that firing this enabled transition t results in marking M_1.

   For example, $(N, [start])\ [a\rangle\ (N, [c1, c2])$, how about $(N, [c3, c4])\ [e\rangle\ (N, [c5])$?

Definition 5.6 (*Firing sequence*).

Let $(N, M_0) \in \mathcal{N}$ be a marked Petri net with $N = (P, T, F)$.

1. A sequence $\sigma \in T^*$ is called a *firing sequence* of (N, M_0) **if and only if**,

   for some natural number $n \in \mathbb{N}$, there exist markings $M_1, M_2, \ldots, M_n$

   and transitions $T_1, T_2, \ldots, T_n$ such that

   - $\sigma = (t_1, t_2, \ldots, t_n) \in T^*$
   - and for all i with $0 \leq i < n$, then $(N, M_i)[t_{i+1}\rangle$ and $(N, M_i)[t_{i+1}\rangle\,(N, M_{i+1})$.

2. A marking M is *reachable* from the initial marking M_0 **if and only if** there exists a sequence of enabled transitions whose firing leads from M_0 to M.

   The set of reachable markings of (N, M_0) is denoted $[N,\ M_0\rangle$.

   [E.g., the marked Petri net shown in Fig. 5.11 has seven reachable markings.]

3. (Petri net system) A Petri net system (P, T, F, M_0) consists of a Petri net (P, T, F) and a distinguished marking M_0, the *initial marking*.

♦ **EXAMPLE 5.6.**

In Fig. 5.11, write marking $M_0 = [start] = [1, 0, 0, 0, 0, 0, 0]$,

we get the marked Petri net (N, M_0) and see that

- The empty sequence $\sigma_0 = \langle\ \rangle$ - being enabled in (N, M_0)- is a firing sequence of (N, M_0).
- The sequence $\sigma_1 = \langle a,\ b\rangle$ is also enabled in (N, M_0), and firing σ_1 results in marking $[c2, c3]$. We can write $(N, [start])\ [a\ b\rangle\ (N, [c2, c3])$ or $(N, M_0)\ [\sigma_1\rangle\ (N, [c2, c3])$.
- The sequence $\sigma_2 = \langle a,\ b,\ d,\ e\rangle$ is another possible firing sequence,

  and should we get $(N, M_0)\ [\sigma_2\rangle\ (N, [c5])$?

- Is $\sigma = \langle a,\ c,\ d,\ e,\ f,\ b,\ d,\ e,\ g\rangle$ a firing? What is the reachable marking M in the output $(N, M_0)\ [\sigma\rangle\ (N, M)$?

- Check that the set $[N,\ M_0\rangle$ (of reachable markings of (N, M_0)) has seven reachable markings. ■

Definition 5.7 (*Labeled Petri net*)**.**

Often transitions are identified by a single letter, but also have a longer label describing the corresponding activity. A **labeled Petri net** with $N = (P, T, F, A, l)$ where (P, T, F) is a Petri net as defined in Definition 5.2.

- $A \subseteq \mathcal{A}$ is a set of *activity labels*, and the map $l \in \{L : T \longrightarrow A\}$ is a *labeling function*. [One can think of the transition label as the *observable action*. Sometimes one wants to express that particular transitions are **not** observable, or invisible.]
- Use the label τ for a special activity label, called 'invisible'. A transition $t \in T$ with $l(t) = \tau$ is said to be unobservable, silent or invisible.

♣ **OBSERVATION 2.**

1. The X-ray example in Practical Problem 1 illustrates that it is possible to go through several markings in a Petri net by a series of firings.

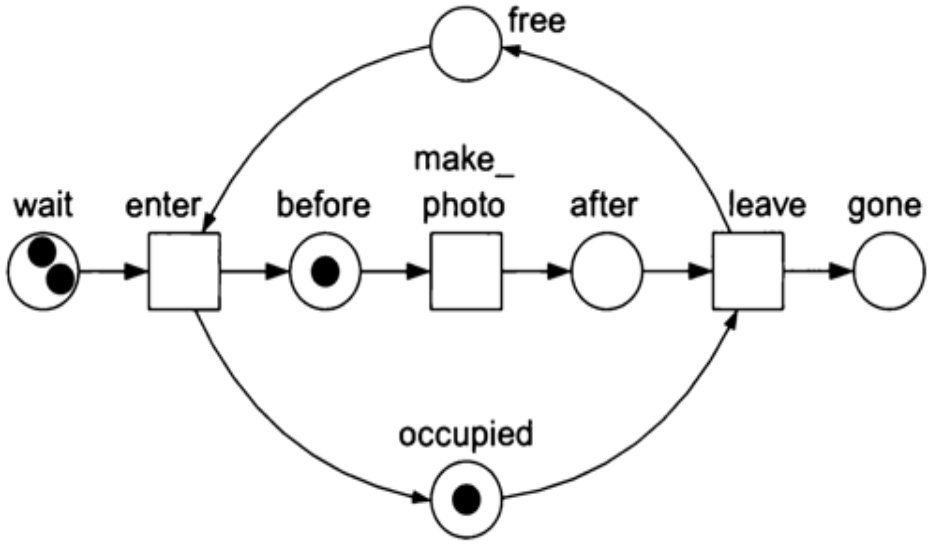


Figure 5.12: *The marking of the improved Petri net for the working process of an X-ray room after transition* ***enter*** *has fired.*

The transitions keep firing until the net reaches a marking that **does not** enable any transition. Like the terminal state in transition systems [Definition 5.1], this marking is a *terminal marking*.

2. We may convert any Petri net into a labeled Petri: just take $A = T$ and $l(t) = t$ for any $t \in T$. The reverse is **not always** possible, (many transitions have the same label).

♣ **QUESTION 5.1.** *On markings in nets, when we have modeled a system as a* **Petri net** *system* (N, M_0) *(see Definition 5.6.3) then some matter occur, including*

1. How many markings are reachable?
2. Which markings are reachable?
3. Are there any reachable terminal markings?

As we know the *initial marking* M_0 for the given system (N, M_0), we answer such questions by calculating the set of markings reachable from M_0. We represent this set as a graph- the **reachability graph** of the net. Its nodes correspond to the reachable markings and its edges to the transitions moving the net from one marking to another.

The key structure is **reachability graph** via transition systems.[9]

♦ **EXAMPLE 5.7** (Reachability graph)**.**

Consider the **Petri net** system in Figure 5.13 modeling the four seasons.

[9]Transition system is the primal model of process modeling, they are the most elementary formalism with which we can describe systems.

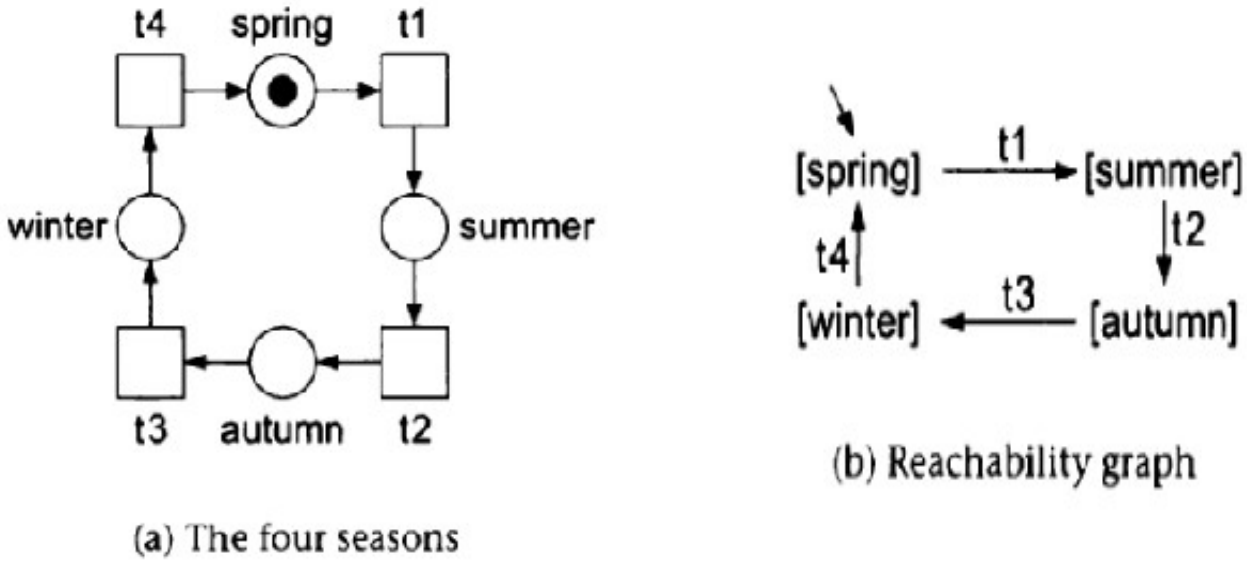


Figure 5.13: *A Petri net system and its reachability graph*

Recall that, each of the reachable markings is represented as a *multiset* (where the same element may appear multiple times). Multiset `[spring]` thus represents the marking in figure 5.13(a).

- The incoming edge **without source** pointing to this node denotes that this marking is the initial marking. We labeled each edge of the reachability graph on the

right with the transition that fired in the corresponding marking.

- Figure 5.13(b) depicts the accompanying reachability graph that represents the set of markings that are reachable from the initial marking shown in figure 5.13(a). We can conclude that the net in figure 5.13(a) has four reachable markings.
- If a marking M is reachable from the initial marking M_0, then the reachability graph has a path from the start node to the node representing marking M. This path represents a sequence of transitions that **have to be fired** to reach marking M from M_0.

We refer to this transition sequence as a *run* (as an execution in finite automaton). A run is *finite* if the path and hence the transition sequence is finite.

Otherwise, the run is *infinite*.

The path from marking [`spring`] to marking [`winter`] is a finite run (t1, t2, t3) of the net in figure 5.13(a). Does it have infinite run? ■

5.3.2 *Representing Petri Nets as Special Transition Systems*

Our discussion shows that we can verify certain properties of a Petri net system by inspecting its reachability graph. For a simple Petri net system, it is easy to construct the accompanying reachability graph, but for more complex nets, reachability graphs can become huge, and it is possible to **forget markings**.

GENERIC AIM:

We describe the behavior of a Petri net system $(N, M_0) \equiv (P, T, F, \ M_0)$ as a state transition system (S, TR, S_0) by showing how to determine the state space S, the transition relation TR, and the initial state S_0 for the system (P, T, F, M_0).

Why are state transition systems suitable for representing Petri Nets?

The transition system represents the state space of the modeled system, thus representing all possible markings M of the net.

Definition 5.8 (Reachability graph).

Let (N, M_0) with $N = (P, T, F, A, l)$ be a marked labeled Petri net.

(N, M_0) defines a transition system $TS = (S, A_1, TR)$ with

$S = [N, M_0\rangle$, $S^{start} = \{M_0\}$, $A_1 = A$, and

$TR = \{(M, M_1) \in S \times S \mid \exists t \in T \quad (N, M)[\, t\rangle \quad (N, M_1)\}$, or with label $l(t)$:

$$TR = \{(M, l(t), M_1) \in S \times A \times S \mid \exists t \in T \quad (N, M)[\, t\rangle \quad (N, M_1)\}. \tag{5.4}$$

TS is often referred to as the **reachability graph** of (N, M_0).

ELUCIDATION

- Elements of A in net N are *labels*, but when transformed to A_1 they are called *actions* in the output transition system TS.
- The initial state S_0 is defined by the initial marking : $S_0 = S^{start} = \{M_0\}$.

  The *markings* M (multisets) in N becomes the *states* in TS.
- The description of the transition relation TR for the Petri net system is more delicate.

  Let us consider two arbitrary states in state space S- that is, two markings M and M_1.

- Transition (M, M_1) is an element of the transition relation TR if there is a transition $t \in T$ enabled at marking M, and the firing of t in marking M yields marking M_1.

  We formalize this by defining transition relation TR as the set of

  all pairs $(M, M_1) \in S \times S$ satisfying

$$\exists t \in T : \quad (N, M)[\, t\rangle\, (N, M_1).$$

  Otherwise, transition (M, M_1) is **not** possible, and (M, M_1) is **not** an element of TR.

- The set $\mathcal{M}$ (all markings) contains markings that are reachable from the given initial marking M_0, but also markings that are not. As a result, for a given marked Petri net (N, M_0), its reachability graph TS is a subgraph of the full transition system.

PRACTICE 5.2.

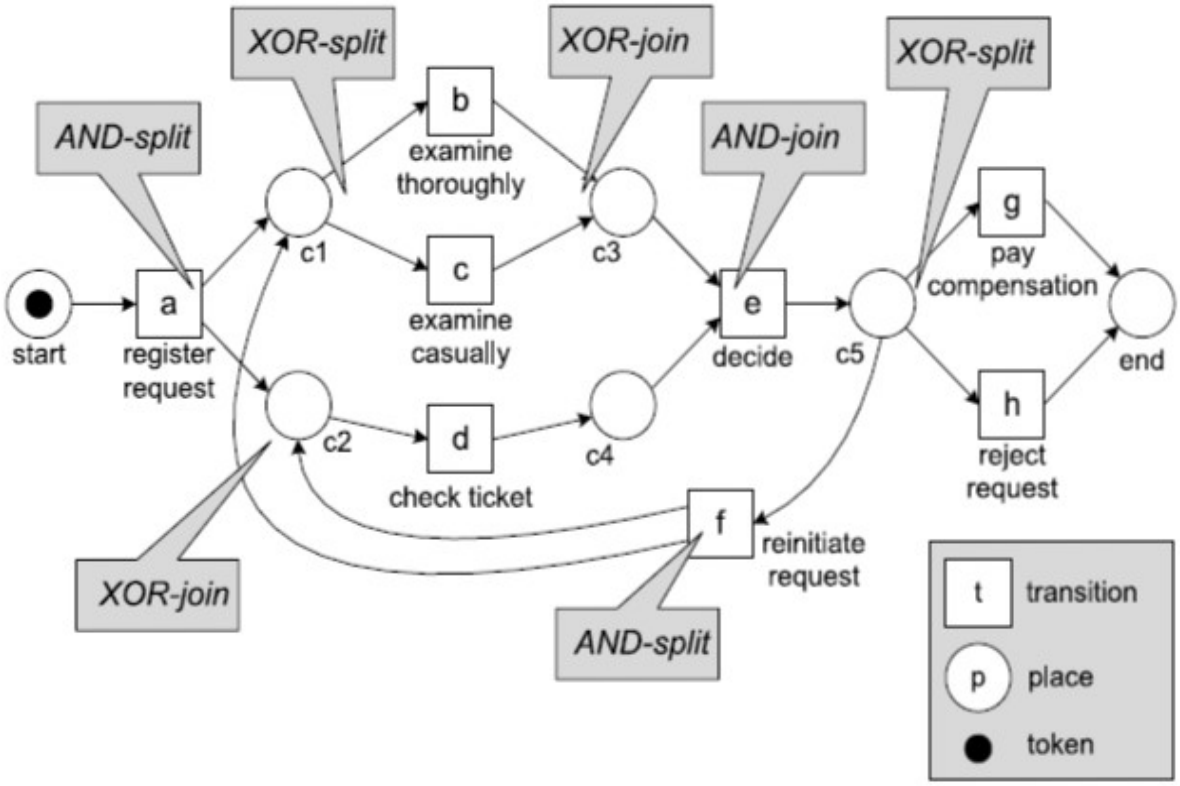


Figure 5.14: *A labeled marked Petri net*

Build up the transition system TS generated from the labeled marked Petri net shown in Fig. 5.14.

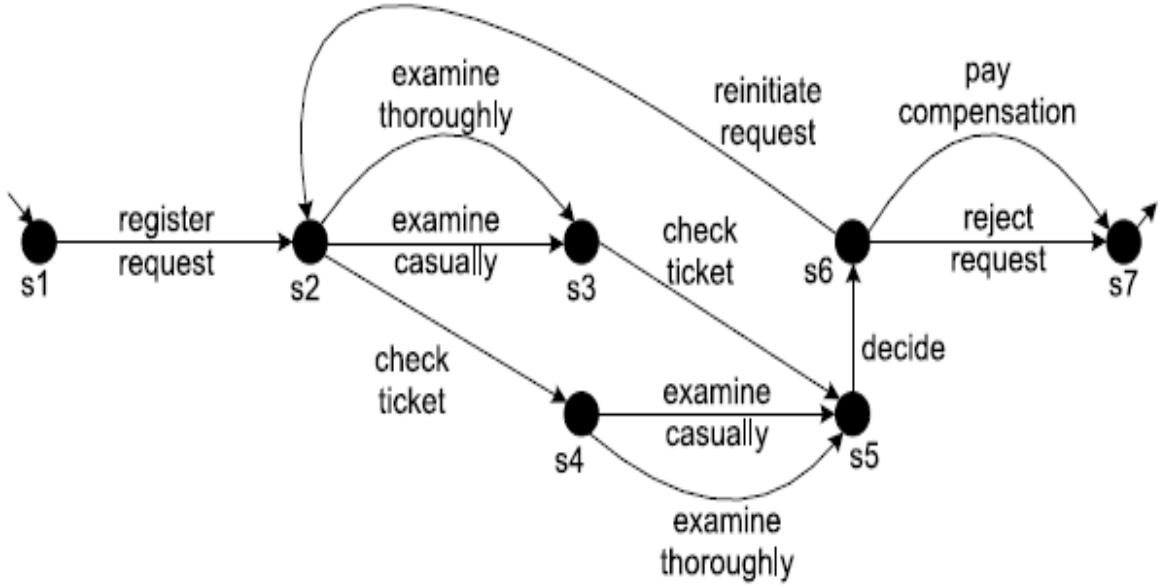
HINT: DIY first, follow ideas in EXAMPLE 5.7.

States of TS correspond to reachable markings, i.e., multi-sets of tokens.

Note that $S^{start} = \{[start]\}$ is a singleton containing the initial marking of the Petri net. The Petri net does not explicitly define a set of final markings S^{end}.

However, in this case it is obvious to take $S^{end} = \{[end]\}$.

The outcome should be as follows



A transition system having one initial state and one final state

Figure 5.15: *The **reachability graph** TS of the above labeled marked Petri net*

but you must give specific values of $s2$, $s3$, $s4$, $s5$, $s6$ in terms of places ci. ■

5.4 ON NETWORK STRUCTURES and TYPICAL PROBLEMS in Petri net

Places and transitions in a Petri net are connected by *arcs*.

Connecting of nodes determines the behavior of the network.

Formally, the way in which transitions are connected determines the order in which they can fire. First we recall key terms, rules and relevant ideas, in a **Petri net** $N = (P, T, F)$.

1. When a transition t fires, the resulting number of tokens in any place p is equal to the initial number of tokens **minus** the number of consumed tokens plus the number of produced tokens.
2. The total number of tokens in the net changes if the number of input places of transition t is **not** the same as the number of output places of transition t. Accordingly, the *firing of a transition* may increase or decrease the overall number of tokens.

3. When several transitions are *enabled* at the same moment, it is **not** determined which

of them will fire. This situation is a nondeterministic choice.

Even though we **do not** know in this case which transition will fire, we know that one of them will be fired.

5.4.1 *Causality, Concurrency and Synchronization*

In general, transitions represent *events*. Let us assume that there are three events: x, y and z. We first discuss typical network structures to state that:

1. Event y happens after event x.
2. Event x and event y take place concurrently (at the same time or in any order).
3. Event z happens after both event x and y.

The first case of causality is defined by Item 1. and shown in figure 5.16(left).

■ **CONCEPT 2.**

1. **Causality** [10] is formally understood as a relationship between two events in a system that must take place in a certain order. In a **Petri net** N, we may represent this relationship by two transitions connected through an intermediate place.

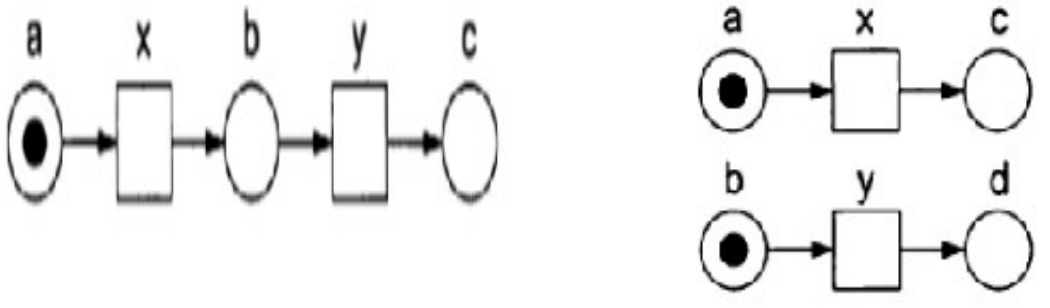


Figure 5.16: *Causality and Concurrency in a* **Petri net**

2. **Concurrency**[11] is an important feature of information systems.

   In a concurrent system, several events **can occur simultaneously**. For example,

[10] 1. the **relationship** between cause and effect; 2. the **principle** that everything has a cause;

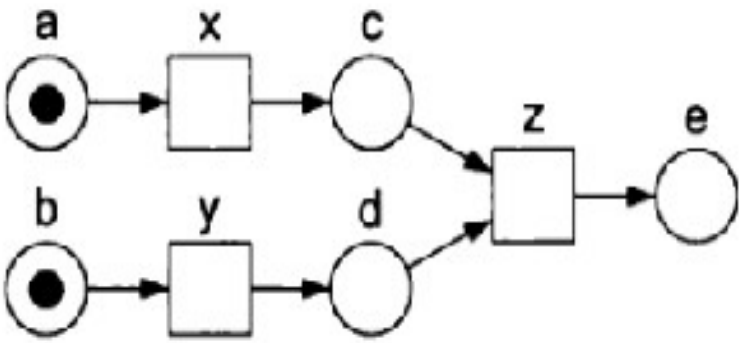
[11] i.e., parallelism, the fact of two or more events or circumstances happening or existing at the same time

several users may access an information system like a database at the same time.

- In Figure 5.16(right) transitions x and y can fire independently of each other; that is, there is **no causal relationship** between the firing of x and of y. With network structures, as those shown in Figure 5.16(right), we can model concurrency. In concurrent models, there is often a need for **synchronization**.

  Question: Can we find the reachability graph of the net N in Figure 5.16(right)?
  Hint: N has 4 places, initial marking is $M_0 = [1, 1, 0, 0]$, two transitions $x, y \in T$.

- We can model synchronization in a **Petri net** as a transition with at least two input places. In figure 5.17, transition z has two input places and **can only fire** after transitions x and y have fired.



Synchronization:
Transition z occurs after the concurrent transitions x and y.

Figure 5.17: *Synchronization in a* **Petri net**

- In industry or any process, assume that transitions x and y represent two concurrent production steps. Transition z can then represent an assembly step that can take place only after the results of the two previous production steps are available, see fig. 5.17.

5.4.2 *Effect of Concurrency*

Figure 5.16(right) shows a simple concurrency occurring in a Petri net.

♣ **QUESTION 5.2.** *Could we quantify the concurrency for a given process or* **Petri net***?*

The answer is yes, and the tool is transition system. Remind that, a transition system is formally a triplet $TS = (S, A, T)$ where S is the set of *states*, $A \subseteq \mathcal{A}$ is the set of *activities* (often referred to as *actions*), and $T \subseteq S \times A \times S$ is the set of *transitions*.

Concurrency via transition systems

Fact 5.1.

If the model of a process contains a lot of concurrency or multiple tokens reside in the same place, then the transition system TS is much bigger than the Petri net $N = (P, T, F)$.

Generally, a marked Petri net (N, M_0) may have infinitely many reachable states.

Problem 5.3 illustrates this fact.

5.5 SUMMARIZED OUTCOMES and REVIEWED PROBLEMS

After studying this chapter you should be able to:

1. Explain the terms

   "place," "transition," "flow relation," "token," "input place," "output place,"

   "preset," "postset," "enabled," "firing," "consumption," "production,"

   "nondeterminism," "initial marking," "terminal marking," and "reachable marking."

2. Model simple systems and business processes as Petri nets.
3. Draw the accompanying Petri net system when place set P, transition set T, flow relation F, and the initial marking M_0 are given.
4. Indicate for a given Petri net system

   which transitions are enabled, which transitions can fire, and

   which markings are reachable from a given marking.

PROBLEM 5.1. *Explain the following terms for Petri nets:*

1. "enabled transition”

2. "firing of a transition”

3. "reachable marking,"

4. "terminal marking," and

5. "nondeterministic choice"?

PROBLEM 5.2. *Consider the Petri net system in figure below.*

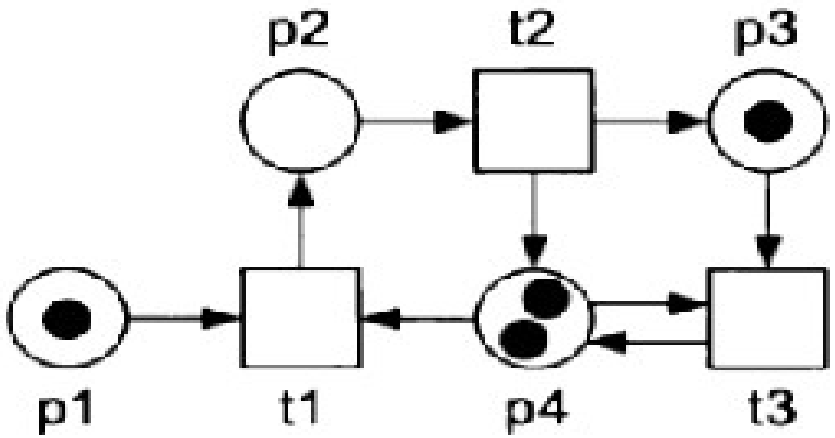


Figure 5.18: *A Petri net with small numbers of places and transitions*

1. Formalize this net as a quadruplet (P, T, F, M_0).
2. Give the preset and the postset of each transition.
3. Which transitions are enabled at M_0?

4. Give all reachable markings. What are the reachable terminal markings?

5. Is there a reachable marking in which we have a nondeterministic choice?

6. Does the number of reachable markings increase or decrease if we remove

   (1) place $p1$ and its adjacent arcs and

   (2) place $p3$ and its adjacent arcs?

PROBLEM 5.3 (From small marked Petri net to bigger transition system)**.**

Consider a marked Petri net $N = (P, T, F)$ with $|P| = 4 = n$, with the 4 start places each initially get 3 tokens, and the exit node is specially designated as place `OUT`. Besides, assume $|T| = 4$, and the initial marking is M_0, as seen in Figure 5.19.

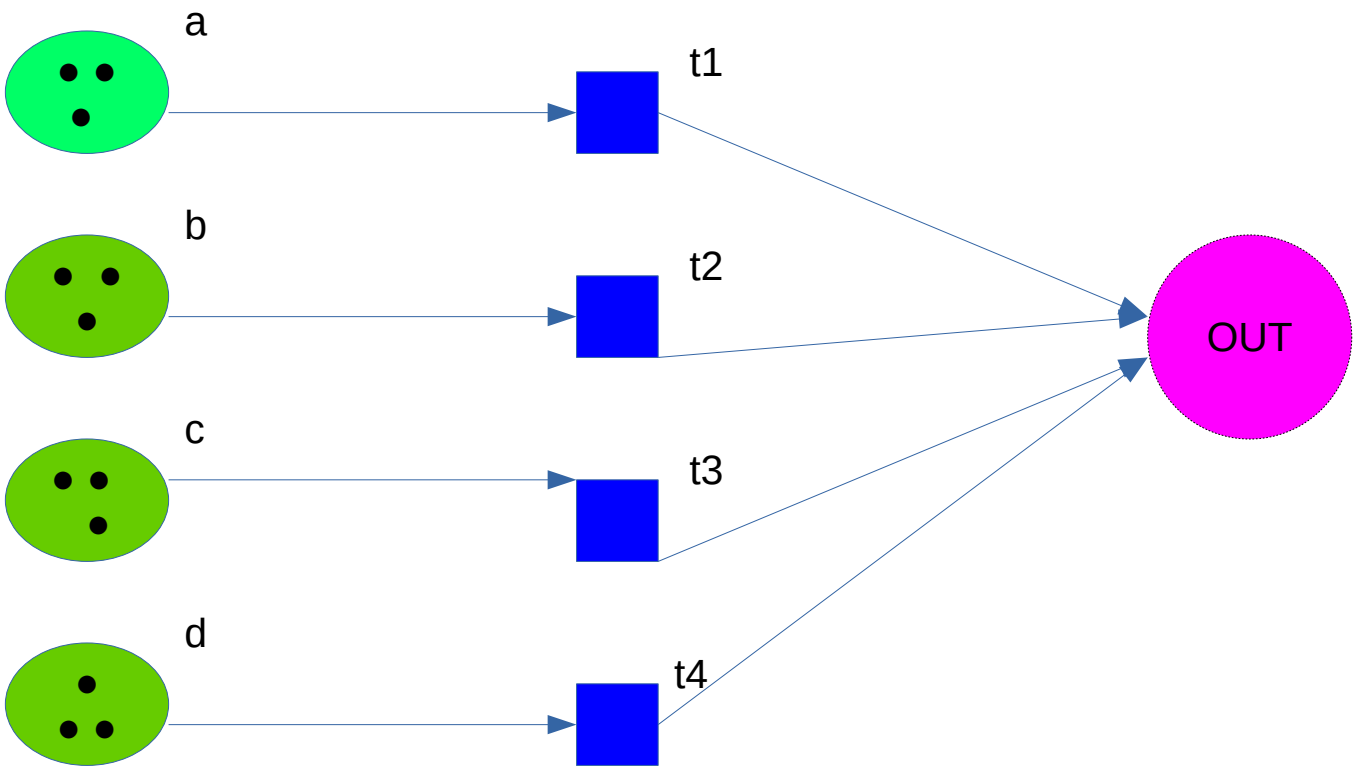


Figure 5.19: *A* **Petri net** *allows concurrency*

Denote TS for the whole transition system made by the Petri net $N = (P, T, F)$.

1. Write down M_0, P, T of N.

2. If **not allow concurrency** in this process (marked Petri net) then how many states of the transition system TS can be created? ? How many transitions are there?

   NOTE: The states of TS are in fact the *reachable markings* of the Petri net N.

   HINT:

   Extend the ideas of Hamming distance in the hypercube $H_n = (\{0,1\}^n, E)$ to "quaternary cube" $K_n = (\{0,1,2,3\}^n, E)$, and modify the concept of edge in K_n to capture the transitions in TS.

3. (* Optional) If we **allow concurrency** in this net, how many states and how many transitions of the transition system TS could be built? ■

Part C:
Advanced Methods and Models

This part introduces various advanced algebraic models and other methods to solve optimization problems.

Chapter ??: Causality with Graphical Models with Bayesian Networks

Chapter 6: Statistical Inference Methods with Bayesian Inference

This page is left blank intentionally.

Chapter 6

Statistical Inference Methods

Bayesian Inference

[Source [42]]

6.1 Introduction to Bayesian Inference Framework

Before embarking into a long journey, let's first of first look at a simple problem.

♦ **EXAMPLE 6.1** (Salaries)**.**

What do you think is the average starting annual salary of a Math-Stat graduate? Is it $ 20,000 per year? Unlikely, that's too low. Perhaps, $ 200,000 per year?

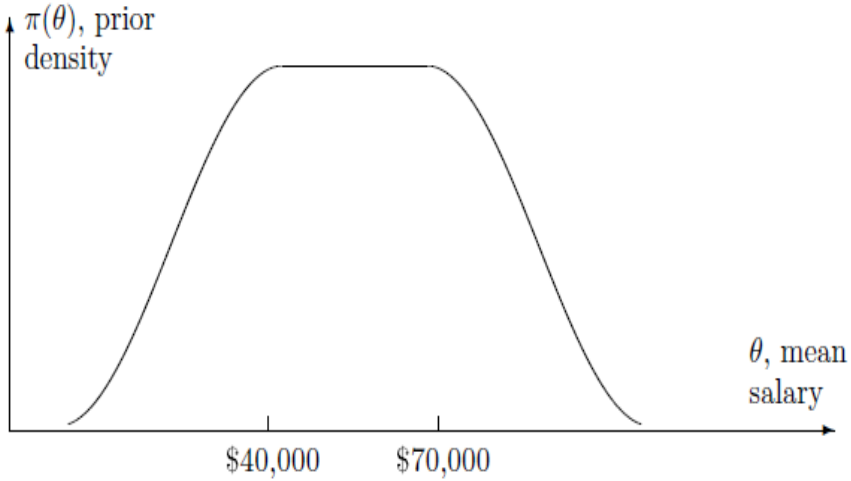
No, that's too high for a fresh graduate. **Then what is the best guess**?

ELUCIDATION

- Between $ 40,000 and $ 70,000 sounds like a reasonable range. We can **express it as some distribution** with the most likely range between $ 40,000 and $ 70,000, see Figure 6.1.
- We can certainly collect data on 100 recent graduates,

  compute their average salary, and use it as an estimate.

  But before that, we already have our beliefs on what the mean salary may be.



Our prior distribution for the average starting salary.

Figure 6.1: *Illustration a specific prior-distribution*

- **Collected data may force us to change our initial idea** about the unknown parameter θ. Probabilities of different values of θ may change.

  * Then we'll have a **posterior distribution** of θ. ■

After study of this chapter, you should be able to

1. describe the Bayesian machinery connecting a prior with its posterior distribution
2. formulate the Bayes' Rule, and find marginal distribution (predictive p.d.f.) of a data
3. understand empirical Bayes inference via using three popular conjugate families of (Binomial, Beta), (Poisson, Gamma) and (Normal, Normal)
4. conduct a Bayesian estimation using the Bayes estimator (posterior mean), tow cases of risk of an estimator
5. perform Bayesian Testing.

Interesting results and many new statistical methods can be obtained when we take a rather different look at statistical problems. The difference is in our treatment of **uncertainty**. So far, random samples were the only source of uncertainty in all the discussed statistical problems. The only distributions, expectations, and variances

considered so far were distributions, expectations, and variances of data and various statistics computed from data.

Now we view parameters of a model as random variables as well. In Bayesian inference, the parameter is treated as a random variable, and Bayes Theorem plays a crucial role.

6.1.1 *The frequentist approach and the Bayesian approach*

Population parameters were considered fixed. Statistical procedures were based on the distribution of data given these parameters (belong to a parameter space $\mathcal{M}$),

$$f(\boldsymbol{X};\theta) = f(X_1, X_2, \cdots, X_n;\theta);\ \theta \in \mathcal{M}$$

The frequentist approach

According to this approach, all probabilities refer to random samples of data and possible long-run frequencies, and so do such concepts as *unbiasedness*, *consistency*, *confidence level*, and *significance level*; reminded as follows.

1. An estimator $\widehat{\theta}$ is unbiased if in a long run of random samples, it averages to the parameter θ, that is $\mathbf{E}[\widehat{\theta}] = \theta$.
2. A test has significance level α if in a long run of random samples, $(100\alpha)\%$ of times the true hypothesis H_0 is rejected.
3. An interval has confidence level $(1-\alpha)$ if in a long run of random samples, $(1-\alpha)100\%$ of obtained confidence intervals contain the parameter, as shown in Figure 6.2; etc.

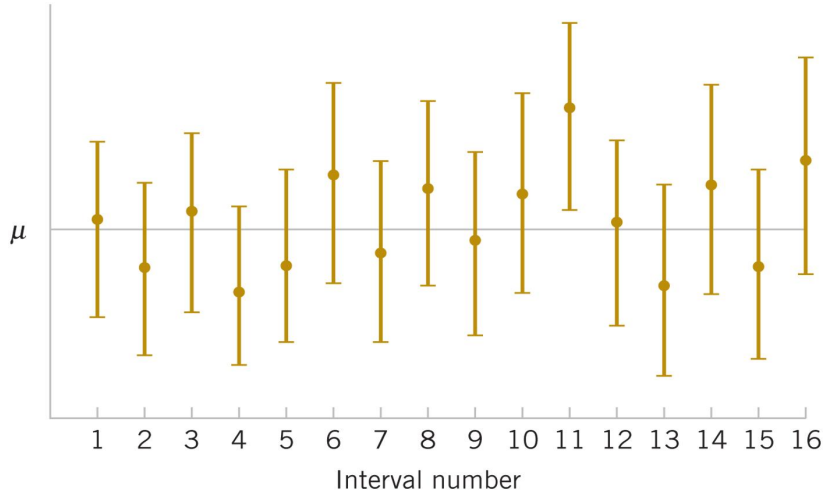


Figure 6.2: *Many CI of μ show that $\overline{X}$ is a random variable*

The Bayesian approach

However, there is another approach: the **Bayesian approach**.

One benefit of the Bayesian approach- via EXAMPLE 6.1 - is that we no longer have to explain our results in terms of a "long run." Often we collect just one sample for our analysis and don't experience any long runs of samples.

Instead, the *Bayesian approach* basically says

'Observed data add information about the parameter'.

With the Bayesian approach, uncertainty affects on the unknown parameter θ as well. Some values of θ are more likely than others. Then, as long as we talk about the likelihood, we can define a whole distribution of values of θ.

a/ Let us call it a **prior distribution**. It reflects our ideas, beliefs, and past experiences about the parameter before we collect and use the data.

b/ The updated knowledge about θ given the data $\mathbf{x} = x_1, x_2, \cdots, x_n$ can be expressed as the **posterior distribution**, via the Bayesian formula

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}.$$

Hence, we can state the result or make decision in terms of the **posterior distribution** of parameter θ.

6.1.2 *Prior and posterior in Bayesian inference*

We have two information sources for the Bayesian inference:

i) collected and observed data [gives the **conditional likelihood**];

ii) and prior distribution of the parameter [gives the **Prior**].

Combining the **data** (conditional likelihood) with the **prior** gives us the **posterior**.

A/ Prior distribution of the parameter

1. Prior to the experiment, our knowledge about the parameter θ is expressed in terms of the **prior distribution** (prior pdf) $h(\theta)$.
2. The observed sample of data $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ [generally be a random sample] has distribution

$$f(\mathbf{x}\,|\boldsymbol{\theta}) = f(\mathbf{x}\,|\theta_1, \theta_2, \cdots, \theta_k) = f(x_1, x_2, \cdots, x_n|\theta_1, \theta_2, \cdots, \theta_k)$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k) \in \mathcal{M}$, a vector of k parameters.

When $k = 1$ write

$$f(\mathbf{x}\,|\theta) = f(x_1, x_2, \cdots, x_n|\theta);$$

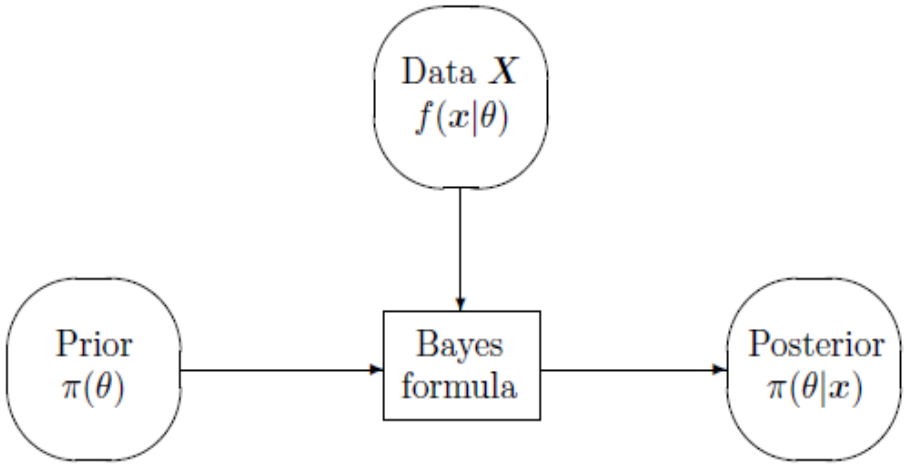
and this distribution is conditional on θ. That is, different values of the parameter θ generate different distributions of data. And thus, conditional probabilities about $\mathbf{X}$ generally depend on the condition, θ. See functions $f(x|\theta)$ in Example 6.2.

3. Obviously then $f(\mathbf{x}\,|\theta)\ h(\theta)$ is the joint pdf $f(\mathbf{x}; \theta)$.

CONVENTION: In this chapter, depending on specific contexts, we use notation

$h(\theta)$ and $\pi(\theta)$ equivalently for **prior** distributions, and

$h(\theta|\ \mathbf{x})$ and $\pi(\theta|\ \mathbf{x})$ equivalently for **posterior** distributions.



Two sources of information about the parameter θ.

Figure 6.3: *The Bayesian machinery*

Principles:

1. 'Observed data add information about the parameter'.
2. Computationally, given the data $\mathbf{x} = x_1, x_2, \cdots, x_n$, the updated knowledge about

θ can be expressed as the **posterior distribution**, via the Bayesian machinery, (see Figure 6.3).

B/ Bayes's formula and the posterior distribution

The Bayesian machinery means

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}. \tag{6.1}$$

Recall the elementary Bayes's formula, and think data $A = \theta,\ \mathbf{x} = B$:

$$\mathbb{P}[A\ B] = \mathbb{P}[A|B]\ \mathbb{P}[B] = \mathbb{P}[B|A]\ \mathbb{P}[A] \Rightarrow \mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\ \mathbb{P}[A]}{\mathbb{P}[B]}.$$

- Specifically, the posterior distribution updates knowledge about parameter θ given a data $\mathbf{x}$ by the following rule

Bayes' Rule

$$\pi(\theta|\,\mathbf{x}) = \pi(\theta|\mathbf{X}=\mathbf{x}) = \frac{f(\mathbf{x};\theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}\,|\theta)\;h(\theta)}{m(\mathbf{x})} = \frac{h(\theta)\;f(\mathbf{x}\,|\theta)}{m(\mathbf{x})}. \tag{6.2}$$

- The denominator $m(\mathbf{x})$ is called the **marginal distribution** or the **predictive p.d.f** of the data $\mathbf{x}$. It represents the *unconditional distribution* of data $\mathbf{x}$.
- For $m(\mathbf{x})$, being unconditional means that it is constant (just the Evidence) for different values of the parameter θ. How do we compute $m(\mathbf{x})$?

Just think similarly as using the Total Probability (General union rule) for a data B,

$$\mathbb{P}[B] = \mathbb{P}\left[\bigcup_i^n B\;E_i\right] = \sum_i^n \mathbb{P}[B\cap E_i] = \sum_i^n \mathbb{P}[B|E_i]\;\mathbb{P}[E_i].$$

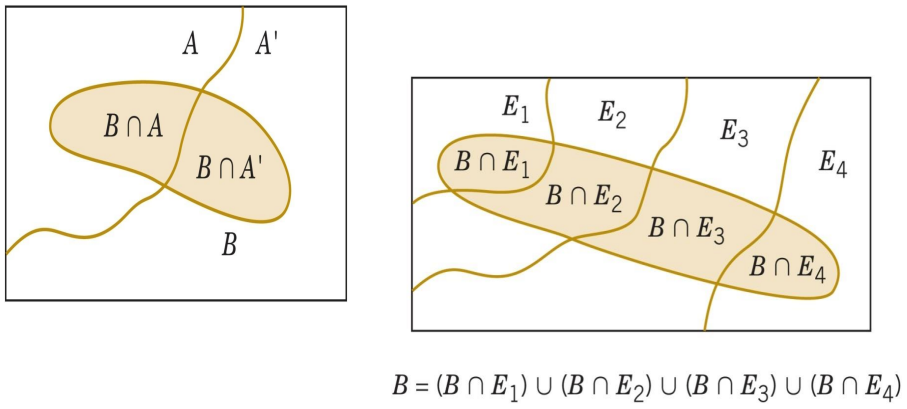


Figure 6.4: *General union rule*

C/ The marginal distribution (predictive p.d.f.) of a data

How do we find $m(\mathbf{x})$ from a data $\mathbf{x} = x_1, x_2, \cdots, x_n$?

The marginal distribution $m(\mathbf{x})$- also **predictive p.d.f.** of X-

is the average p.d.f. of X, with respect to the prior p.d.f $h(\theta)$

expressed from $\mathbf{x}$ by the **total probability** with respect to the model parameter θ.

$$m(\mathbf{x}) = \sum_{\theta} f(\mathbf{x}\,|\theta)\; h(\theta)$$

for discrete prior distributions $h()$

$$m(\mathbf{x}) = \int_{\theta} f(\mathbf{x}\,|\theta)\; h(\theta)\; d\theta \tag{6.3}$$

for continuous prior distributions $h()$

♦ **EXAMPLE 6.2** (Quality inspection).

A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 10%.

We have to decide whether **to accept or to reject** the shipment based on the model parameter θ, the proportion of defective parts.

Before we see the real data, let's assign a 50-50 chance to both suggested values of θ,

i.e., the prior distribution of θ is given as

$$h(\theta = \theta_1 = 0.05) = h(\theta = \theta_2 = 0.10) = 0.5.$$

A random sample of 20 parts has 3 defective ones.

Question: Calculate the **posterior distribution** of θ.

GUIDANCE for solving. We apply the Bayes formula (6.1).

Denote X for the number of defective parts in the shipment. Given θ, the distribution of X is binomial $\mathbf{Bin}(n = 20, \theta)$. Using table or soft R we calculate the densities

- $n = 20;\ x = 3$;
- $a = dbinom(x, n, 0.05) = 0.0596$; # the pmf $f(x|\theta_1) = f(3|0.05)$
- $b = dbinom(x, n, 0.10) = 0.1901$; # the pmf $f(x|\theta_2) = f(3|0.10)$

The marginal distribution of data X (for $x = 3$), by Equation (6.3) is

$$m(3) = \sum_{\theta} f(x|\theta)\ h(\theta) = f(x|\theta_1)\ h(\theta_1) + f(x|\theta_2)\ h(\theta_2)$$

$$= f(3|\ 0.05)\ h(0.05) + f(3|\ 0.10)\ h(0.10) = (0.0596)(0.5) + (0.1901)(0.5) = 0.12485. \tag{6.4}$$

Posterior probabilities of parameter θ for

$\theta = \theta_1 = 0.05$ and $\theta = \theta_2 = 0.10$

are now computed, by Equation (6.2) as

$$\pi(\theta_1|X = x = 3) = \frac{f(x|0.05)\ h(0.05)}{m(3)} = \frac{(0.0596)(0.5)}{0.1248} = 0.2387;$$

$$\pi(\theta_2|X = x = 3) = \frac{f(x|0.10)\ h(0.10)}{m(3)} = \frac{(0.1901)(0.5)}{0.1248} = 0.7613.$$

CONCLUSION.

1. At first, we had no preference between the two suggested values of θ.
2. Then we observed a rather high proportion of defective parts, $3/20 = 15\%$.
3. Taking this into account, $\theta = 0.10$ is now about three times as likely than $\theta = 0.05$. ■

Table 6.1: *Summary*

NOTATION - Formula	**Meaning**
$\underline{X}\|\theta \sim f(\mathbf{x}\,\|\,\theta),\ \theta \in \mathcal{M}$	condition distribution of $\underline{X}$ given $\theta \in \mathcal{M}$
$\mathbf{x} \in R_{\underline{X}}$	range of the data
$\theta \sim h(\theta), \theta \in \mathcal{M}$	another probability distribution
	where $h(\theta)$ is prior distribution
$f(\mathbf{x}\,\|\theta)\ h(\theta) = f(\mathbf{x},\ \theta)$	joint distribution of $\mathbf{x}, \theta$
$m(\mathbf{x}) = \sum_{\theta\in\mathcal{M}}\ f(\mathbf{x}\,\|\theta)\ h(\theta)$	the marginal distribution of $\underline{X}$
	for discrete prior distributions h
$m(\mathbf{x}) = \int_{\theta\in\mathcal{M}}\ f(\mathbf{x}\,\|\theta)\ h(\theta)\ \mathrm{d}\theta$	the marginal distribution of $\underline{X}$
	for continuous prior distributions h
$\pi(\theta\|\,\mathbf{x}) = \pi(\theta\|\mathbf{X} = \mathbf{x})$ $= \dfrac{f(\mathbf{x}\,\|\theta)\ h(\theta)}{m(\mathbf{x})}$	posterior distribution of $\theta\|\mathbf{X} = \mathbf{x}$

6.2 Review of important distributions

We wrote a data vector by $\mathbf{X}$ or $\underline{X}$ both indicate $X_1, X_2, \cdots, X_n$;

a sample vector either by $\mathbf{x}$ or $\underline{x}$, both stand for $x_1, x_2, \cdots, x_n$.

Denote $R_{\underline{X}} := \mathrm{Range}(X)$ = the range of the data.

Parameter θ is taken from the space $\mathcal{M}$. Hence, allowing the parameter θ takes value in the parameter space $\mathcal{M}$, posterior distribution is an updated knowledge about θ given the data $\mathbf{X} = X_1, X_2, \cdots, X_n$. Kindly see summary tables below for key probability distributions.

Particularly, three important distributions for studying the **reliability and failure rates** of systems are the gamma and the Weibull distributions. These distributions are discussed here as further examples of continuous distributions.

6.2.1 *Beta distribution*

The probability density function of $\mathbf{Beta}(\nu_1, \nu_2)$ is

$$f(x; \nu_1, \nu_2) = \begin{cases} \dfrac{1}{B(\nu_1, \nu_2)} x^{\nu_1 - 1}(1-x)^{\nu_2 - 1}, & \text{when } 1 > x > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{6.5}$$

Name of distribution	**Notation**	**Parameters**	**Expectation** $\mathbf{E}[X] = \mu$	**Variance** $\mathbf{V}[X] = \sigma^2$
Bernoulli	$X \sim \mathbf{B}(p)$	p- probability of success	p	$p(1-p)$.
Binomial	$X \sim \mathbf{Bin}(n,p)$	n, p	np	$np(1-p)$
Poisson	$X \sim \text{Pois}(\lambda)$	λ	λ	λ
Gauss	$X \sim \mathbf{N}(\mu, \sigma^2)$	μ, σ^2	μ	σ^2
Exponential	$X \sim \mathbf{Exp}(\beta)$	β	β	β^2
χ^2 (with $\mathbf{df}_{=}k$)	$X \sim \chi^2_k$	k	k	$2k$
Student	$T \sim t_{n,p}$	n, p	$\mu_T = 0$	$n/((n-2)$
Beta	$X \sim \mathbf{Beta}(\nu_1, \nu_2)$	ν_1, ν_2	$\frac{\nu_1}{\nu_1 + \nu_2}$	$\frac{v\nu_2}{(\nu_1+\nu_2)^2(\nu_1+\nu_2+1)}$
Gamma	$X \sim \boldsymbol{G}(\alpha, \lambda)$	α, λ	$\mu_G = \alpha\ \lambda$	$\sigma^2_G = \alpha\ \lambda^2$
Weibull	$X \sim \boldsymbol{W}(\alpha, \beta)$	α, β		

with shape parameters $\nu_1, \nu_2 > 0$; the integration

$$B(\nu_1, \nu_2) = \int_0^1 x^{\nu_1 - 1}(1-x)^{\nu_2 - 1} dx. \tag{6.6}$$

When $\nu_1 = \nu_2 = 1$, Beta becomes the uniform $U(0, 1)$.

The probability cumulative function of $\mathbf{Beta}(\nu_1, \nu_2)$ is

$$I_x(\nu_1, \nu_2) = \frac{1}{B(\nu_1, \nu_2)} \int_0^x u^{\nu_1 - 1}(1-u)^{\nu_2 - 1} du, \tag{6.7}$$

with $0 \leq x \leq 1$. Note that $I_x(\nu_1, \nu_2) = 1 - I_{1-x}(\nu_2, \nu_1)$.

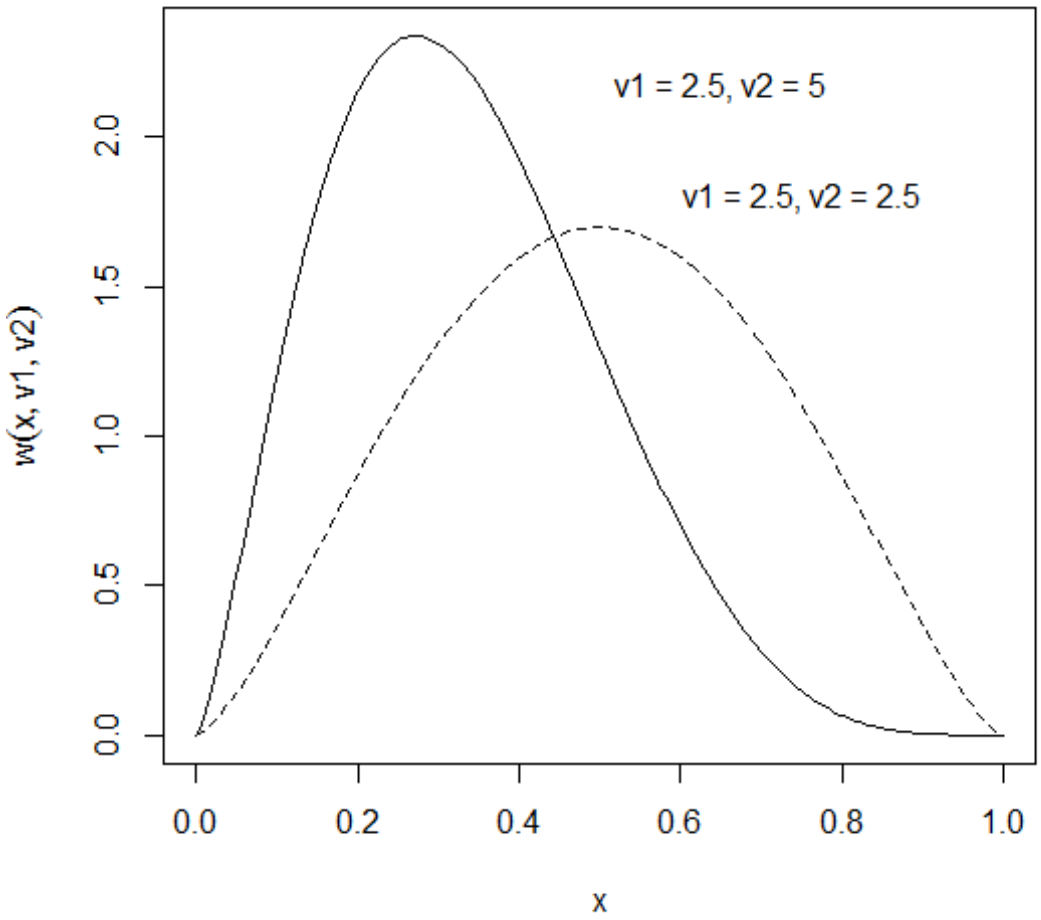


Figure 6.5: *The pdf* $f(x;\nu_1,\nu_2)$ *of* $\mathbf{Beta}(\nu_1,\nu_2)$ *when* $\nu_1 = 2.5, \nu_2 = 2.5; \nu_1 = 2.5, \nu_2 = 5.00$.

Figure 6.5 shows the graph f of $\mathbf{Beta}(2.5, 5.0)$ and $\mathbf{Beta}(2.5, 2.5)$. If $\nu_1 = \nu_2$ then the

pdf f of Beta is symmetric via the vertical line $\mu = \frac{1}{2}$. We don't have the explicit mgf of $\mathbf{Beta}(\nu_1, \nu_2)$, but we know the m-th moment

$$\begin{aligned} \mu_m &= \frac{1}{B(\nu_1,\nu_2)} \int_0^1 u^{m+\nu_1-1}(1-u)^{\nu_2-1} du = \frac{B(\nu_1+m,\nu_2)}{B(\nu_1,\nu_2)} \\ &= \frac{\nu_1(\nu_1+1)\cdots(\nu_1+m-1)}{(\nu_1+\nu_2)(\nu_1+\nu_2+1)\cdots(\nu_1+\nu_2+m-1)}. \end{aligned} \tag{6.8}$$

The beta distribution has an important role in the theory of statistics. As will be seen later, many methods of statistical inference are based on the order statistics. The distribution of the order statistics is related to the beta distribution. Moreover, since the beta distribution can have a variety of shapes, it has been applied in many cases in which the variable has a distribution on a finite domain. By introducing a location and a scale parameter, one can fit a shifted-scaled beta distribution to various frequency distributions.

6.2.2 *Gamma distribution*

Two important distributions for studying the reliability and failure rates of systems are the gamma and the Weibull distributions. We will need these continuous distributions in our study of reliability methods.

♣ Practical motivation 1.

Suppose we use in a manufacturing process a machine which mass-produces a particular part. In a random manner, it produces defective parts at a rate of λ per hour.

The number of defective parts produced by this machine in a time period $[0, t]$ is a random variable $X(t)$ having a Poisson distribution with mean $\lambda\, t$.

By Lemma **??**, the probability density function of of $X(t)$ is

$$\mathbb{P}[X(t) = j] = e^{-\lambda t}\, \frac{(\lambda\, t)^j}{j!}, \quad j = 0, 1, 2, ...$$

Now we wish to study the distribution of the time until the k-th defective part is produced. Call this **continuous** random variable Y_k.

We use the fact that the k-th defect will occur before time t (i.e., $Y_k \leq t$) if and only if at least k defects occur up to time t (i.e. $X(t) \geq k$). Therefore,

$$Y_k \leq t \Leftrightarrow X(t) \geq k,$$

thus the c.d.f. for Y_k is

$$\begin{aligned} G(t;k,\lambda) = \mathbb{P}[Y_k \leq t] = \mathbb{P}[X(t) \geq k] &= 1 - \sum_{j=0}^{k-1} \mathbb{P}(X(t) = j) \\ &= 1 - \sum_{j=0}^{k-1} \frac{(\lambda t)^j \, e^{-(\lambda t)}}{j!} \end{aligned} \tag{6.9}$$

The corresponding p.d.f. for Y_k is

$$g(t;k,\lambda) = \begin{cases} \dfrac{\lambda^k}{(k-1)!} \, t^{k-1} \, e^{-(\lambda t)}, & \text{when } \ t \geq 0, \\ 0, & \text{when } \ t < 0. \end{cases} \tag{6.10}$$

This p.d.f. is a member of a general family of distributions gamma $\boldsymbol{G}(\alpha, \beta)$ which depend on two parameters α and β. The probability density function of $\boldsymbol{G}(\alpha, \beta)$, gen-

eralized from Equation 6.23, is

$$g(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-x/\beta}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{6.11}$$

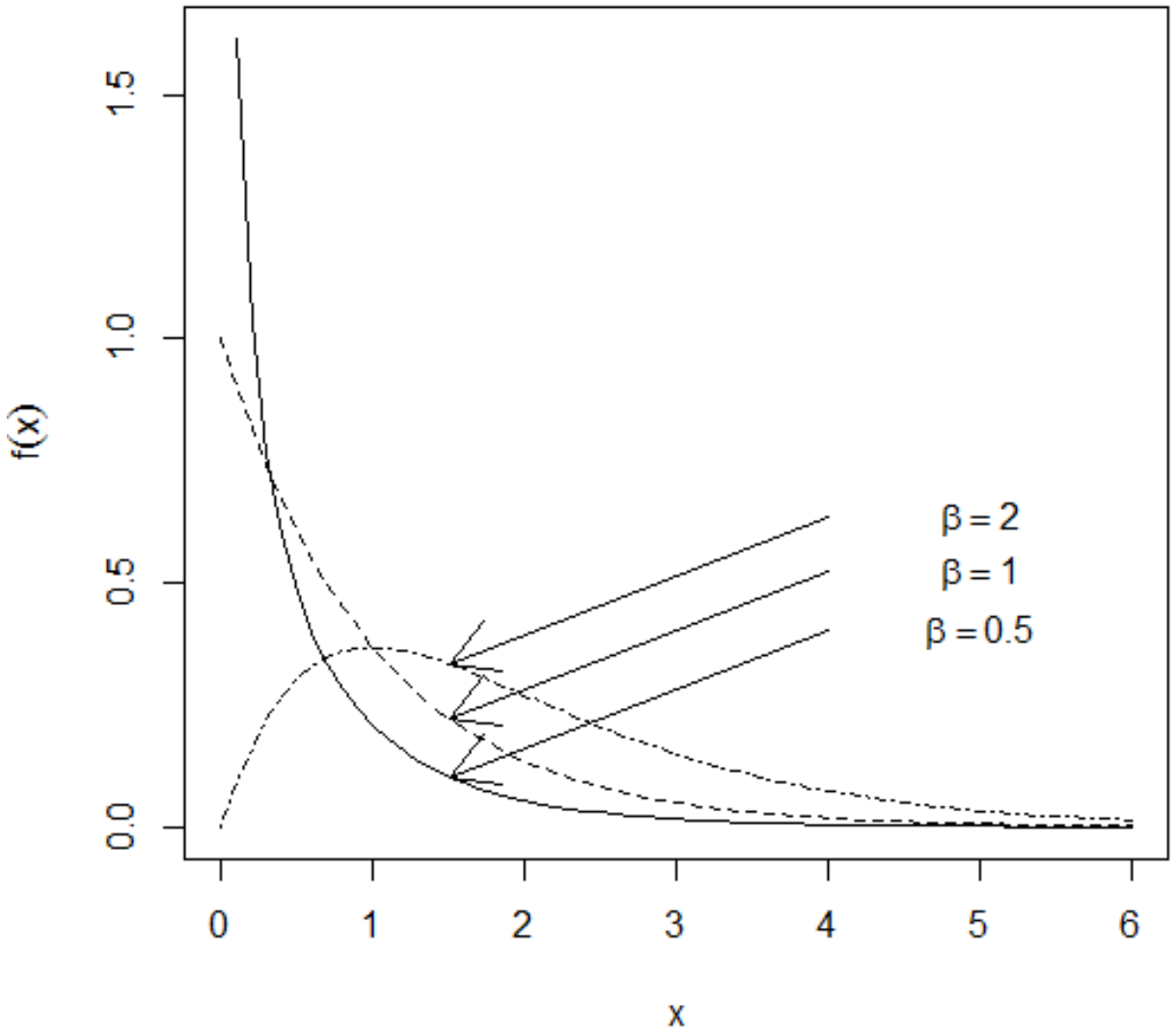


Figure 6.6: *The pdf with* $\beta = 1$ *and* $\alpha = 0.5,\ 1,\ 2$.

In soft R, function `pgamma` computes c.d.f of a gamma distribution having the shape α, and scale β; $0 < \alpha, \beta < \infty$. If we use $\alpha = shape = 1 = \beta = scale$ then the cdf $F_{\boldsymbol{G}}(1) = 0.6321206$.

```
> pgamma(q=1, shape=1, scale=1)

    [1] 0.6321206
```

The expected value and variance of the gamma distribution $\boldsymbol{G}\,\alpha, \beta)$ are, respectively,

$$\begin{aligned} \mu &= \alpha\ \beta \\ \sigma^2 &= \alpha\ \beta^2. \end{aligned} \tag{6.12}$$

Property 6.1.

1. $\Gamma(\alpha)$ is called the **gamma function** of α and is defined as the integral

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt; x > 0 \tag{6.13}$$

The gamma function satisfies the relationship $\Gamma(1) = 1$ and

$$\Gamma(x+1) = x\Gamma(x),\ \forall x > 1. \tag{6.14}$$

Hence, for every positive integer $x = n \in \mathbb{N}$ then $\Gamma(n+1) = n!$. Besides, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

We note also that the exponential distribution $\mathbf{Exp}(\beta)$ is a special case of the gamma distribution with $\alpha = 1$, write $\mathbf{Exp}(\beta) = \boldsymbol{G}(1, \beta)$.

2. If $X_i \sim \boldsymbol{G}(\alpha_i, \beta)$ are independent, then

$$\sum_{i=1}^{n} X_i \sim \boldsymbol{G}(\sum_{i=1}^{n} \alpha_i, \beta).$$

Therefore, if particularly, $X_i \sim \mathbf{Exp}(\beta) = \boldsymbol{G}(1, \beta)$ are iid exponential, the sum $\sum_{i=1}^{n} X_i \sim \boldsymbol{G}(n, \beta)$. Hence, the sum

$$T = t(\mathbf{X}) = \sum_{i=1}^{n} X_i = t$$

has pdf

$$f_T(t; n, \beta) = \begin{cases} \dfrac{1}{\beta^n \Gamma(n)}\, t^{n-1}\, e^{-t/\beta}, & \text{if } t \geq 0. \\ \text{or } \dfrac{\theta^n}{\Gamma(n)}\, t^{n-1}\, e^{-\theta\, t}, & \text{if put } \theta = 1/\beta \end{cases} \tag{6.15}$$

6.3 Empirical Bayes Inference

The Empirical Bayes Inference approach allows us to choose a suitable specific prior from a large family of prior distributions. Here, the data $\mathbf{X} = X_1, X_2, \cdots, X_n$ is used to choose this prior.

♣ QUESTION 6.1.

In an empirical Bayes, what kind of prior family of distributions do we use for θ?

Answer: We often choose a prior family of distributions such that both the prior and the posterior have the same structure, which is called a "conjugate family".

(A) In the Binomial model, Beta is a conjugate family.

(B) In the Poisson model, Gamma is a conjugate family.

(C) In the Normal model, Normal is the conjugate family.

Definition 6.1 (KEY: Conjugate distribution and posterior mean)**.**

In Bayesian inference, the followings are essential concepts.

1. A family of prior distributions π is **conjugate to the model** $f(x|\theta)$ if the posterior belongs to the same family of π.
2. The **posterior mean** (or posterior expectation) of a parameter θ is the conditional mean of the posterior $h(\theta|\,\mathbf{x})$, given by

   $\mathbf{E}_h[\theta \mid \boldsymbol{x}] = \mathbf{E}[h(\theta|\,\mathbf{x}) \mid \boldsymbol{x}]$ as a function of data $\mathbf{x}$, explicitly

$$\mathbf{E}_h[\theta \mid \boldsymbol{x}] = \begin{cases} \sum_\theta \theta\ h(\theta|\,\mathbf{x}) = \dfrac{\sum_\theta \theta\ f(\boldsymbol{x}|\,\boldsymbol{\theta})\ h(\boldsymbol{\theta})}{m(\mathbf{x})}, & \theta \text{ is discrete} \\ \\ \int_\theta \theta\ h(\theta|\,\mathbf{x})\ d\theta = \dfrac{\int_\theta \theta\ f(\boldsymbol{x}|\,\boldsymbol{\theta})\ h(\boldsymbol{\theta})}{m(\mathbf{x})}, & \theta \text{ is continuous} \end{cases} \tag{6.16}$$

CONVENTION for conjugate distribution families

From now on, we fix random observations $\mathbf{X} = X_1, X_2, \cdots, X_n$, and write $\boldsymbol{x} = x_1, x_2, \cdots, x_n$ for its observed random values. For single parameter θ we equivalently use pairs of symbols

- $m(\mathbf{x})$ and $f_h(\mathbf{x})$ as the joint predictive p.d.f. (**the Evidence**) of data $\mathbf{X} = X_1, X_2, \cdots, X_n$, with $h(\boldsymbol{\theta})$ is the prior,

$$m(\mathbf{x}) = f_h(\mathbf{x}) = \int_{\boldsymbol{\theta}\in\mathcal{M}} f(\mathbf{x}\,|\boldsymbol{\theta})\; h(\boldsymbol{\theta})\; \mathrm{d}\boldsymbol{\theta}$$

$$= \int_{\theta_1} \cdots \int_{\theta_k} \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})\; h(\boldsymbol{\theta}) d\theta_1 \cdots d\theta_k,$$

- $\pi(\theta|\,\mathbf{x}) = \pi(\theta|\mathbf{X} = \mathbf{x})$ and $h(\theta|\,\mathbf{x}) = h(\theta|\mathbf{X} = \mathbf{x})$ as the posterior p.d.f. (**the Posterior**) of θ.

In brief, use $\pi(\theta|\,\mathbf{x})$ with $m(\mathbf{x})$; and use $h(\theta|\,\mathbf{x})$ with $f_h(\mathbf{x})$.

Computing the Posterior:

With random data $\boldsymbol{x} = x_1, x_2, \cdots, x_n$ from a distribution $f(x; \boldsymbol{\theta})$, the posterior p.d.f.of $\boldsymbol{\theta}$, corresponding to the prior p.d.f. $h(\boldsymbol{\theta})$ can be written by

either

$$\pi(\boldsymbol{\theta}|\,\mathbf{x}) = \frac{f(\boldsymbol{x}; \boldsymbol{\theta})}{m(\mathbf{x})} = \frac{f(\boldsymbol{x}|\;\boldsymbol{\theta})\; h(\boldsymbol{\theta})}{m(\mathbf{x})} = \frac{\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})\; h(\boldsymbol{\theta})}{m(\mathbf{x})} \tag{6.17}$$

or

$$h(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}; \boldsymbol{\theta})}{f_h(\mathbf{x})} = \frac{f(\boldsymbol{x}|\;\boldsymbol{\theta})\; h(\boldsymbol{\theta})}{f_h(\mathbf{x})} = \frac{\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})\; h(\boldsymbol{\theta})}{f_h(\mathbf{x})}. \tag{6.18}$$

FRAMEWORK OF INVESTIGATION

We follow the following pattern when investigating three popular conjugate distribution families.

- Conclusion about pairing; and reminder of distributions (if necessary)
- DATA and LIKELIHOOD

  Discussion on likelihood (if necessary)
- PRIOR function $h(\boldsymbol{\theta})$
- POSTERIOR function $\pi(\theta|\,\mathbf{x}) = \pi(\theta|\mathbf{X} = \mathbf{x}) = h(\theta|\,\mathbf{x})$,

$$h(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\,\boldsymbol{\theta})\; h(\boldsymbol{\theta})}{f_h(\mathbf{x})}$$

- POSTERIOR MEAN, the conditional mean of the posterior (optional)

$$\mathbf{E}_h[\theta \mid \boldsymbol{x}] = \mathbf{E}[h(\theta|\,\mathbf{x}) \mid \boldsymbol{x}].$$

6.3.1 *A/ Binomial Distributions* $X \sim \mathbf{Bin}(k;\theta)$, $0 < \theta < 1$

Beta family is conjugate to the Binomial model.

- DATA and LIKELIHOOD $f(\mathbf{x}\,|\theta)$:

The p.d.f.of $X \sim \mathbf{Bin}(k;\theta)$, (assume k is known) is

$$f(x|\theta) = \binom{k}{x}\theta^x(1-\theta)^{k-x}, \quad x = 0, \cdots, k.$$

A sample $\mathbf{x} = x_1, x_2, \cdots, x_n$ from binomial $\mathbf{Bin}(k;\theta)$ has the probability mass function

$$f(\mathbf{x}\,|\theta) = \prod_{i=1}^{n}\binom{k}{x_i}\theta^{x_i}(1-\theta)^{k-x_i} \sim \theta^{\sum x_i}(1-\theta)^{nk-\sum x_i}.$$

- PRIOR $h(\theta)$:

  Suppose that θ has a prior beta distribution $\mathbf{Beta}(\alpha,\beta)$, with p.d.f.

$$b(\theta;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}, \text{ when } 0<\theta<1, \tag{6.19}$$

  where $\alpha,\beta > 0$, and $B(a,b)$ is the complete beta function

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

  So the **prior density** of $\mathbf{Beta}(\alpha,\beta)$ has the same form, as a function of θ

$$h(\theta) \sim \theta^{\alpha-1}(1-\theta)^{\beta-1} \text{ for } 0<\theta<1.$$

- POSTERIOR $h(\theta|x)$:

  Then the **posterior density** of θ, given $X = x$, is

$$h(\theta|x) = \frac{1}{B(\alpha+x,\beta+k-x)}\,\theta^{\alpha+x-1}\,(1-\theta)^{\beta+k-x-1}, \tag{6.20}$$

or given $\mathbf{X} = \mathbf{x} = x_1, x_2, \cdots, x_n$ is

$$h(\theta|\,\mathbf{x}) \sim f(\mathbf{x}\,|\theta)\; h(\theta) \sim \theta^{\alpha+\sum\; x_i-1}(1-\theta)^{\beta+nk-\sum x_i-1};\ \text{when } 0<\theta<1.$$

Therefore, we recognize the posterior density also as a $\mathbf{Beta}(\alpha_x, \beta_x)$ with new parameters $\alpha_x = \alpha + \sum_{i=1}^{n} x_i$ and $\beta_x = \beta + nk - \sum_{i=1}^{n} x_i$.

- POSTERIOR MEAN:

Fact: Any $\mathbf{Beta}(\alpha, \beta)$ has the mean: $\mu = \alpha/(\alpha+\beta)$.

The posterior density

$$h(\theta|\,\mathbf{x}) \sim \theta^{\alpha+\sum\; x_i-1}(1-\theta)^{\beta+nk-\sum x_i-1}$$

is a $\mathbf{Beta}(\alpha + \sum\; x_i, \beta + nk - \sum x_i)$. Hence, the posterior mean of θ, given $\mathbf{X} = \mathbf{x}$, is

$$\begin{aligned}\mathbf{E}[\theta|x] &= \frac{1}{B(\alpha_x,\beta_x)}\int_0^1 \theta^{\alpha_x}\,(1-\theta)^{\beta_x-1}d\theta \\ &= \frac{\alpha_x}{\alpha_x+\beta_x} = \frac{\alpha+\sum_{i=1}^{n} x_i}{\alpha+\beta+nk}.\end{aligned} \tag{6.21}$$

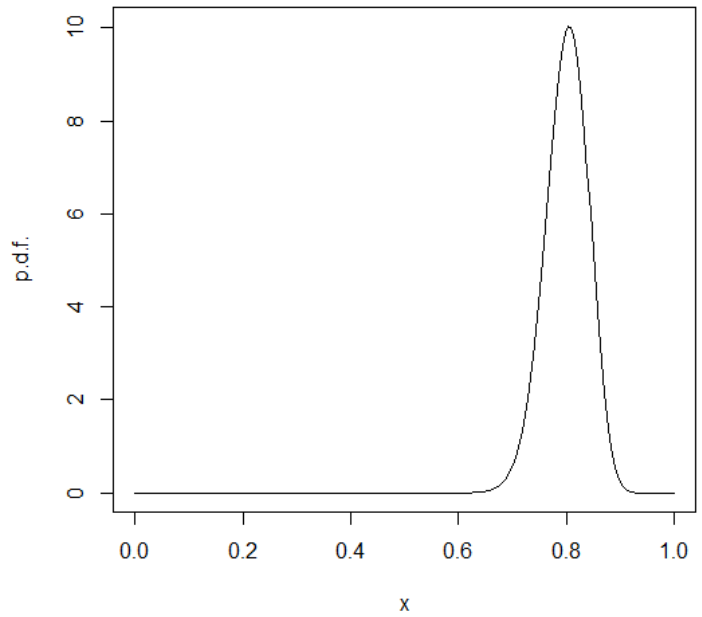
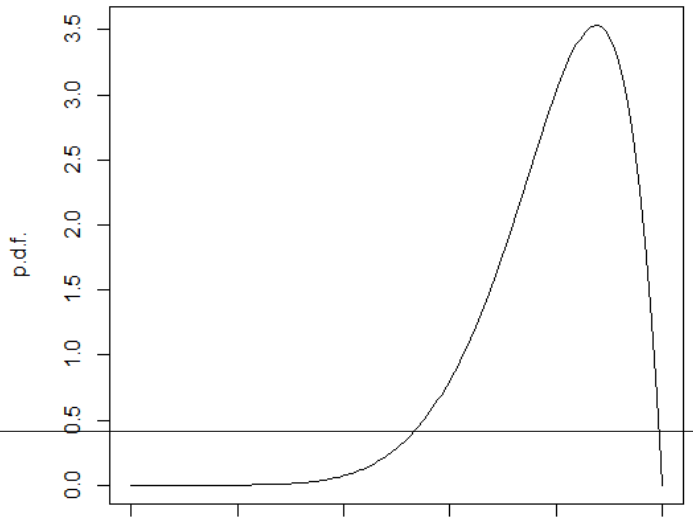


Figure 6.7: *The prior p.d.f. of* $\mathbf{Beta}(80, 20)$



♦ **EXAMPLE 6.3.**

Suppose that X is a discrete random variable having a binomial distribution $\mathbf{Bin}(n, \theta)$; n is known, but θ is unknown. The parameter space is $\mathcal{M} = \{\theta : 0 < \theta < 1\}$.

Suppose we believe that θ is close to 0.8, with small dispersion around this value. In Figure 6.7 we illustrate the prior p.d.f. of a Beta distribution $\mathbf{Beta}(80, 20)$, whose functional form is

$$h(\theta; 80, 20) = \frac{99!}{79!19!}\, \theta^7 9\ (1-\theta)^{19},\ 0 < \theta < 1.$$

If we wish, however, to give more weight to small values of θ, we can choose the $\mathbf{Beta}(8, 2)$ as a prior density, i.e.,

$$h(\theta; 8, 2) = 72\ \theta^7\ (1-\theta),\ 0 < \theta < 1.$$

Using the prior $h(\theta; 8, 2)$ and Equation 6.20 we can prove that the posterior $h(\theta|x)$ of θ is again the p.d.f. of a Beta distribution $\mathbf{Beta}(8 + x, n - x + 2)$. [**Why**?]

Hint: use the marginal distribution $m(\mathbf{x})$

$$m(\mathbf{x}) = \int_{\theta \in \mathcal{M}} f(x|\theta)\ h(\theta)\ d\theta.$$

in

$$h(\theta|x) = \frac{f(x|\ \theta)\ h(\theta; 8, 2)}{m(\mathbf{x})}.$$

In Figure 6.9 we present some of these posterior p.d.f. for the case of $n = 10$ and $X = 6, 7, 8$. ■

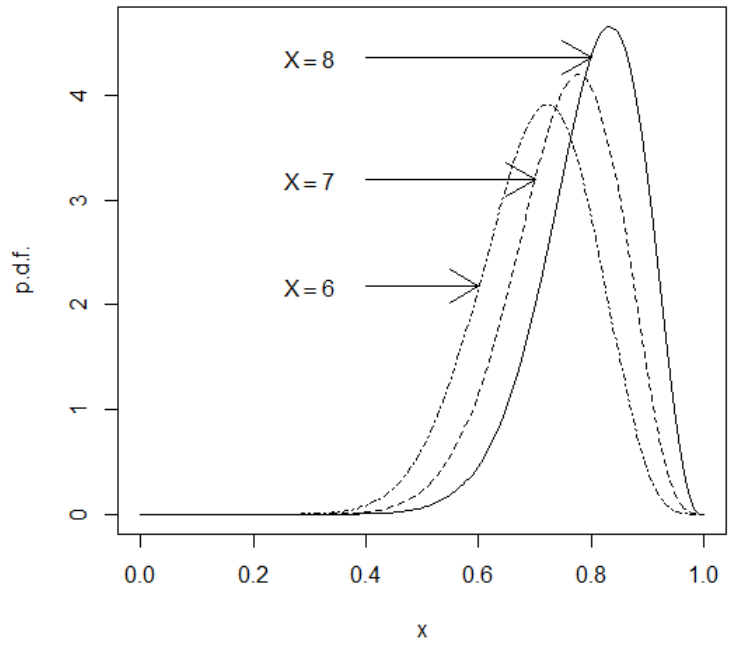
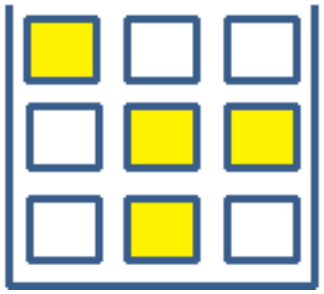


Figure 6.9: *The posterior density $h(\theta|x)$ of θ, with $n = 10$, $X = 6, 7, 8$.*

♦ **EXAMPLE 6.4** (MARKETING RESEARCH- from SMA3, 2017, Vietnam).

Let

- Large Box = Population;



- N = total number of tickets in the box = Population size;
- θ = proportion of households using our products, then

  $\theta \in \mathcal{M} = [0, 1]$.

Take a random sample of k tickets from the box

$\Rightarrow X =$ number of colored tickets in the sample.

Think $k = 500, X = 100$.

$X|\theta \sim \mathbf{Bin}(k, \theta) \Longrightarrow \ f(x|\ \theta) = \binom{k}{x}\theta^x(1-\theta)^{k-x}$ is the conditional pdf.

Classic approach: $\hat{\theta} = \dfrac{X}{k}$.

Bayesian approach: the prior is $\pi(\theta)$- probability distribution on $\mathcal{M} = [0, 1]$- assumed to be distribution Beta(a, b), see Section 6.2.1.

* No knowledge: $\text{Beta}(a,b) = \text{Beta}(1,1)$ = Uniform

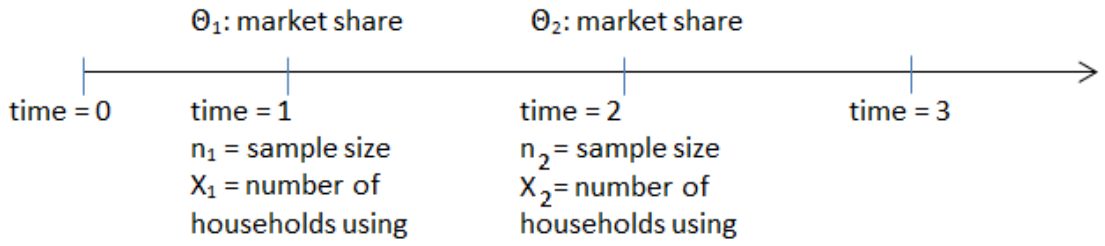
* With knowledge: $X|\theta \sim \text{Bin}(k,\theta)$

$$\theta \sim \pi(\theta|\ a,b) \equiv \text{Beta}(a,b) \equiv \frac{1}{\text{B}(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

$\Longrightarrow$ Posterior pdf = $h(\theta|X,a,b) \equiv \text{Beta}(a+X, b+k-X)$ or

$$h(\theta|\ X,a,b) = \pi(\theta|\ X,a,b) \equiv \text{Beta}(X+a, k-X+b). \tag{6.22}$$

Note that a and b are unknown; we need to estimate them by $\hat{a}$ and $\hat{b}$.



Now the sample size k get values $n_1, n_2, \ldots$ at time points 1, 2,...

time = 1: $X_1|\theta_1 \sim \textbf{Bin}(n_1,\theta_1)$, and

the prior $\theta_1 \sim \pi(\theta_1) = \text{Beta}(a, b)$, so

the posterior of θ_1 is $h(\theta_1|X_1)$:

$$h(\theta_1|X_1) \sim \text{Beta}(X_1 + a, n_1 - X_1 + b)$$

Posterior mean $\mathbf{E}[$ the posterior $h(\theta_1|X_1)] = \dfrac{X_1 + a}{n_1 + a + b}$; see Eq. 6.21.

If $a = b = 1$ (uniform), posterior mean $= \dfrac{X_1 + 1}{n_1 + 2}$

(Classical estimator $= \dfrac{X_1}{n_1}$)

time = 2: $X_2|\theta_2 \sim \mathbf{Bin}(n_2, \theta_2)$, and $a = b = 1$ then

the prior $\theta_2 \sim \pi(\theta_2) = \text{Beta}(X_1 + 1, n_1 - X_1 + 1)$, so

the posterior of $\theta_2|X_2$ is $h(\theta_2|X_2)$:

$$h(\theta_2|X_2) \sim \text{Beta}(X_2 + X_1 + 1, n_2 - X_2 + n_1 - X_1 + 1)?$$

Posterior mean = ?

time = 3: Could we work out the case time $= 3$? ■

6.3.2 *B/ Poisson Distributions*

Gamma family is conjugate to the Poisson model.

Reminder: The probability density function of $\boldsymbol{G}(\alpha, \lambda)$, Gamma distribution, is

$$g(x; \alpha, \lambda) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-\lambda\, x}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases} \tag{6.23}$$

Note: $\Gamma(1+k) = k\Gamma(k)$.

DATA: Let $\mathbf{X} = X_1, X_2, \cdots, X_n$ be a sample from Poisson(θ) distribution with a Gamma $\boldsymbol{G}(\alpha, \lambda)$ prior distribution of θ. Then

$$f(\mathbf{x}\,|\theta) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{e^{-\theta}\ \theta^{x_i}}{x_i!} \sim e^{-n\theta}\ \theta^{\sum_i x_i}. \tag{6.24}$$

DISCUSSION:

1. We dropped ($x_i!$) and wrote that the result is "proportional" ($\sim$) to $e^{-n\theta}\ \theta^{\sum_i x_i}$ (don't contain θ often simplifies the computation).

2. The form of the posterior distribution

$$h(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\ \boldsymbol{\theta})\ h(\boldsymbol{\theta})}{f_h(\mathbf{x})}$$

can be obtained without the constant term, and if needed, we can eventually evaluate the normalizing constant in the end, making

$$h(\theta|x) = \pi(\theta|x)$$

a fine distribution with the total probability 1. In particular, the marginal distribution $m(\mathbf{x}) = f_h(\mathbf{x})$ can be dropped because it is θ-free.

But keep in mind that in this case we obtain the posterior distribution "up to a constant coefficient."

- **PRIOR**: The Gamma *prior distribution* of θ has density

$$h(\theta) = \theta^{\alpha-1}\, e^{-\lambda\theta}. \tag{6.25}$$

As a function of θ, this prior density has the same form as the model $f(x|\theta)$ – a power of θ multiplied by an exponential function.

This is the general idea behind conjugate families.

- **POSTERIOR**: Then, the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$ is

$$\pi(\theta|\,\mathbf{x}) = \pi(\theta|\mathbf{X} = \mathbf{x}) \sim f(\boldsymbol{x}|\ \boldsymbol{\theta})\ h(\theta),$$

hence

$$\pi(\theta|\,\mathbf{x}) \sim \left(e^{-n\theta}\,\theta^{\sum_i x_i}\right)\,\left(\theta^{\alpha-1}\,e^{-\lambda\theta}\right). \tag{6.26}$$

Therefore, , the posterior

$$\pi(\theta|\,\mathbf{x}) \sim \theta^{\alpha+\sum_i x_i-1}\;e^{-(\lambda+n)\theta}$$

Comparing with the general form of a Gamma density (6.23) we see that

$\pi(\theta|\,\mathbf{x})$ is the Gamma distribution with new parameters,

$$\alpha_x = \alpha + \sum_i x_i; \quad \lambda_x = \lambda + n. \tag{6.27}$$

SUMMARY 2. *We can conclude that:*

1. Gamma prior distributions is conjugate to Poisson models.
2. Having observed a Poisson sample $\mathbf{X} = \mathbf{x} = x_1, x_2, \cdots, x_n$ we update the Gamma $\boldsymbol{G}(\alpha, \lambda)$ prior distribution of θ to the posterior $\boldsymbol{G}(\alpha + \sum_i x_i, \lambda + n)$.

6.3.3 *C/ Normal Distributions*

Normal family is conjugate to the Normal model.

Consider now a sample $\mathbf{x} = x_1, x_2, \cdots, x_n$ from Normal distribution $\mathbf{N}(\theta, \sigma^2)$ with an unknown mean θ and a known variance σ^2. The pdf

$$f(\mathbf{x}\,|\theta) = (2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} \exp\left\{ -\frac{(x_i-\theta)^2}{2\sigma^2} \right\}$$

$$\sim \exp\left\{ -\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2\sigma^2} \right\}$$

hence

$$f(\mathbf{x}\,|\theta) \sim \exp\left\{ \theta \frac{\sum x_i}{\sigma^2} - \theta^2 \, \frac{n}{2\sigma^2} \right\} = \exp\left\{ \left(\theta\,\overline{x} - \frac{\theta^2}{2}\right) \, \frac{n}{\sigma^2} \right\}$$

- PRIOR: If the prior distribution of θ is also Normal $\mathbf{N}(\mu, b^2)$, with prior mean μ and prior variance b^2 then

$$\pi(\theta) \sim \exp\left\{ -\frac{(\theta-\mu)^2}{2b^2} \right\} \sim \exp\left\{ \left(\theta\mu - \frac{\theta^2}{2}\right) \, \frac{1}{b^2} \right\},$$

  and again, it has a similar form as $f(\mathbf{x}\,|\theta)$.

- POSTERIOR: The posterior density of θ equals

$$\pi(\theta|\ \mathbf{x}) \sim f(\mathbf{x}\,|\theta)\ \pi(\theta) \sim \exp\left\{ -\frac{(\theta-\mu_x)^2}{2b_x^2} \right\}$$

where the **posterior mean** and **posterior variance** respectively are

$$\mu_x = \mathbf{E}[\theta|\,\mathbf{x}] = \frac{n\,\overline{X}\,/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2} \tag{6.28}$$

$$\text{and}\;\; b_x^2 = \mathbf{V}[\theta|\,\mathbf{x}] = \frac{1}{n/\sigma^2 + 1/b^2}$$

This posterior distribution then is certainly $\mathbf{N}\,(\mu_x, b_x^2)$.

When $n = 1$ (univariate case), put $A = b^2$ then

$$\mu_x = \frac{x/\sigma^2 + \mu/A}{1/\sigma^2 + 1/A} \;\text{and}\; b_x^2 = \frac{1}{1/\sigma^2 + 1/b^2} = \frac{A\,\sigma^2}{A + \sigma^2} \tag{6.29}$$

SUMMARY 3 (Empirical Bayes Inference).

We can conclude that:

1. **Normal family** of prior distributions is conjugate to the Normal model with unknown mean;
2. **Posterior parameters** are given by Equation 6.28.
3. How will the **posterior mean** behave when it is computed from a large sample?

   As the sample size n increases, we get more information from the data, and as a result, the frequentist estimator will dominate. According to Equation (6.28), the posterior mean converges to the sample

mean $\overline{X}$ as $n \to \infty$.

4. **Posterior mean** $\mathbf{E}_h[\theta \mid \boldsymbol{x}]$ will converge to $\overline{X}$ when $\tau \to \infty$.

   Large τ means a lot of *uncertainty* in the prior distribution of θ; thus, naturally, we should rather use observed data as a *more reliable source of information* in this case.

ASSIGNMENT 1 (Rare events with Poisson process)**.**

The number of network blackouts each week at MU has $\text{Pois}(\theta)$ distribution. The weekly rate of blackouts θ is not known exactly, but according to the past experience with similar networks, it averages $\mu = \mathbf{E}[\theta] = 4$ blackouts with a standard deviation of 2. After two weeks, we observed sample $\mathbf{x} = (x_1, x_2)$.

There exists a Gamma distribution with the given mean $\mu = \alpha/\lambda$, and standard deviation $\sigma = \sqrt{\alpha}/\lambda$.

1. Determine the Gamma distribution $\boldsymbol{G}(\alpha, \lambda)$ as the **prior distribution** of θ. **Hint**: you can choose a specific root of (α, λ).
2. Suppose there were $x_1 = 2$ blackouts in the 1st week. Given that, find the **posterior distribution** of θ. **Hint**: use formula 6.27 to get Gamma posterior $\boldsymbol{G}(\alpha_x = \alpha + \sum_i x_i, \lambda_x = \lambda + n)$.
3. Suppose there were $x_2 = 0$ blackouts in the 2nd week. Update the posterior distribution of θ. Compute the average weekly rate of blackouts per week from this posterior distribution. What is

your conclusion? ■

6.4 Bayesian Decision Procedures

6.4.1 *Introductory Bayesian Decision Theory*

Decision theory is concerned with the problem of making decisions, and **statistical decision theory** is concerned with optimal decision making under uncertainty or when statistical knowledge is available only on some of the uncertainties involved in the decision problem. Statistical decision theory tries to combine other relevant information with the sample information to arrive at the optimal decision. Therefore, a Bayesian setting seems to be more appropriate for decision theory. Abraham Wald (1902-1950) laid the foundation for statistical decision theory.

This section covers three topics, motivated by few corresponding advantages.

- *Bayesian Estimation*: An important source of information that decision theory utilizes is the **prior information**. Prior information could be based on past experiences of similar situations or on expert opinion.

- *Loss functions and Risk of an estimator*: Another piece of relevant information is the possible consequences of the decisions. Often these consequences can be quantified. That is, the **loss or gain** of each decision can be expressed as a number (called the loss or utility).
- Bayesian Testing: via the Bayes decision function and Bayes risk.

The framework- Things start from statistical distributions

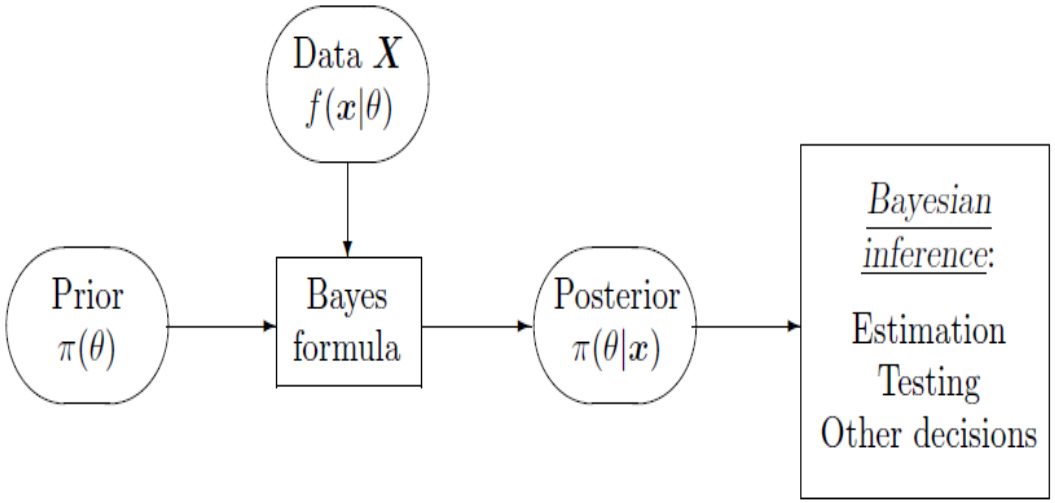
It is often the case that optimal decision depend on unknown parameters of statistical distributions. The **Bayesian decision framework** provides us the tools

- to integrate information that one may have on the unknown parameters with the information obtained from the observed sample, in such a way
- that the expected loss due to erroneous decisions will be minimized.

6.4.2 *Bayesian Estimation*

Prior and posterior- The role of posterior distribution

In a Bayesian framework we express our prior belief (based on prior information) on which θ values are plausible, by a p.d.f. on $\mathcal{M}$, which is called the **prior** probability density function $h(\boldsymbol{\theta})$, see from Section 6.1.2.



Grand scheme for Bayesian inference

Figure 6.10: *Posterior distribution is the basis for Bayesian inference.*

We have already completed the most important step in Bayesian inference. We obtained the posterior distribution. All the knowledge about the unknown parameter is now **included in the posterior**, and that is what we'll use for further statistical analysis.

Definition 6.2 (The **Bayes estimator**- a posterior mean)**.**

In an estimation problem, the decision function is an estimator

$\widehat{\boldsymbol{\theta}}_B = \widehat{\boldsymbol{\theta}}(x_1, \cdots, x_n)$ of $\boldsymbol{\theta}$, which yields a point in the parameter space $\boldsymbol{\Theta}$.

To estimate $\boldsymbol{\theta}$ we simply compute the *posterior mean* of θ given data $\mathbf{X}$, now named the **Bayes estimator**, that $\mathbf{E}_h[\theta \mid \mathbf{X} = \boldsymbol{x}] = \mathbf{E}[h(\theta|\,\mathbf{x}) \mid \boldsymbol{x}]$, explicitly given as

$$\widehat{\boldsymbol{\theta}}_B = \mathbf{E}_h[\theta \mid \mathbf{X} = \boldsymbol{x}] = \begin{cases} \sum_\theta \theta\, h(\theta|\,\mathbf{x}) & \theta \text{ is discrete} \\ \\ \int_\theta \theta\, h(\theta|\,\mathbf{x})\, d\theta & \theta \text{ is continuous} \end{cases} \tag{6.30}$$

The estimator $\widehat{\theta}_B$ is what we "expect" θ to be, after observed $\mathbf{x}$.

6.4.3 *Loss functions and Risk of an estimator*

A **loss** or **utility** to a decision maker is the effect of the interaction of two factors:

(1) the decision or action selected by the decision maker; and

(2) the event or state of the world that actually occurs.

Classical statistics does not explicitly use a loss function or a utility (payoff) function.

We have considered several point estimators such as the *maximum likelihood estimator*, the method of *moments estimator*, and the *posterior mean*. In fact, there are many other ways to generate estimators. **How do we choose among them?** The answer is found in *Statistical Decision Theory* which is a formal theory for comparing statistical procedures.

- Consider a parameter θ which lives in a parameter space Θ. Let $\widehat{\theta}$ be an estimator of θ. In the language of decision theory, an estimator is sometimes called a *decision rule* and

  the possible values of the decision rule are called *actions*.
- Loss function $L(\widehat{\theta}, \theta)$ measures the discrepancy between θ and $\widehat{\theta}$. Some examples include

1. $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$- the squared error loss,

2. $L(\widehat{\theta},\theta) = |\widehat{\theta} - \theta|^p$- L_p loss,

3. $L(\widehat{\theta},\theta) = |\widehat{\theta} - \theta|$- the absolute error loss,

4. $L(\widehat{\theta},\theta) = \mathbf{E}_\theta\left[\log \dfrac{f(X;\theta)}{f(X;\widehat{\theta})}\right] = \int \log\left(\dfrac{f(x;\theta)}{f(x;\widehat{\theta})}\right) f(x;\theta)\, dx$- the Kullback-Leibler loss.

Definition 6.3. *(Risk of an estimator is the mean of a loss function)*

1. The **risk of an estimator** $\widehat{\theta}$ is

$$\begin{aligned} R(\theta) = R(\theta,\widehat{\theta}) &= \mathbf{E}_\theta\Big[L(\widehat{\theta},\theta)\Big] \\ &= \int L(\widehat{\theta},\theta)\, f(x)\, dx. \end{aligned} \tag{6.31}$$

   where $\mathbf{E}_\theta$ represents the expectation w.r.t. the true distribution f.

2. The **posterior risk** as the conditional mean of L on data:

$$\rho(\widehat{\theta}) = \mathbf{E}[L(\widehat{\theta},\theta)|\mathbf{X} = \mathbf{x}].$$

Two important cases of the loss function

CASE A. Parameter $\boldsymbol{\theta}$ is real ($\boldsymbol{\theta} = \theta \in \mathbb{R}$) with (squared error) loss function $L(\widehat{\theta},\theta) = (\widehat{\theta} - \theta)^2$.

i) The **risk** is just the mse (*mean squared error*):

$$R(\theta, \widehat{\theta}) = \mathbf{E}[(\widehat{\theta} - \theta)^2] = \mathbf{E}[(\widehat{\theta} - \theta)^2] = \mathbf{MSE} = \mathbf{V}_\theta[\widehat{\theta}] + \mathbf{Bias}_\theta(\widehat{\theta})^2.$$

ii) The **posterior risk** as the conditional mean of L:

$$\rho(\widehat{\theta}) = \mathbf{E}[L(\widehat{\theta}, \theta)|\mathbf{X} = \mathbf{x}] = \mathbf{E}[(\widehat{\theta} - \theta)^2|\mathbf{X} = \mathbf{x}].$$

We choose an estimator which minimizes this squared-error posterior risk $\rho(\widehat{\theta})$.

iii) In this case, the **Bayes estimator** $\widehat{\theta}_B$ of θ, given by Equation 6.30, gives the lowest squared-error posterior risk $\rho(\widehat{\theta})$; and it becomes the posterior variance:

$$\begin{aligned} \rho(\widehat{\theta}_B) &= \mathbf{E}\Big[(\widehat{\theta}_B - \theta)^2 | \, \mathbf{x}\Big] = \mathbf{E}\Big[\big(\mathbf{E}_h[\theta \mid \mathbf{x}] - \theta\big)^2 | \, \mathbf{x}\Big] \\ &= \mathbf{E}\Big[\big(\theta - \mathbf{E}_h[\theta \mid \mathbf{x}]\big)^2 | \, \mathbf{x}\Big] = \mathbf{V}_h[\theta \mid \boldsymbol{x}] \end{aligned} \tag{6.32}$$

which measures variability of θ around $\widehat{\theta}_B$, according to the posterior $h(\theta|\,\mathbf{x})$ of θ.

CASE B. Parameter $\boldsymbol{\theta}$ is real ($\boldsymbol{\theta} = \theta \in \mathbb{R}$) with loss function

$$L(\widehat{\theta}, \theta) = c_1(\widehat{\theta} - \theta)^+ + c_2(\theta - \widehat{\theta})^+, \quad \text{with } c_1, c_2 > 0, \text{ and } \ (a)^+ = \max(a, 0).$$

PROPERTIES OF THE ESTIMATOR $\widehat{\theta}$ [All cases]

1. The risk of the estimator $\widehat{\theta}$ is $R(\theta) = \mathbf{E}_\theta\Big[L(\widehat{\theta}, \theta)\Big]$ where $\mathbf{E}_\theta$ represents the expectation with respect to the true distribution f.
2. In general, parameter θ may be a quantity being defined or derived from data X. In such scenarios, writing the risk $R(\theta)$ is less likely possible, we have to find $R(\theta)$ by ad-hoc analysis, see Example 6.6.
3. When an estimator $\widehat{\theta}$- a decision rule, is computed/updated from an observed data $\mathbf{x}$, such that the risk $R(\theta)$, or some induced quantities (from $R(\theta)$) being minimized to be $R(\widehat{\theta})$, we can call possible values of the decision rule $\widehat{\theta}$ the posterior actions, see Example 6.6.

♦ **EXAMPLE 6.5** (The posterior risk- Normal case).

The Bayes estimator of the mean θ of a normal $\mathbf{N}(\theta, \sigma^2)$ distribution, where θ has a normal $\mathbf{N}(\mu, b^2)$ prior, is just the *posterior mean*. So, by Eq. 6.28:

$$\widehat{\theta}_B = \mu_x = \frac{n\,\overline{X}\,/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2}, \tag{6.33}$$

and its posterior risk - in **CASE A**- is

$$\rho(\widehat{\theta}_B) = \mathbf{V}_h[\theta \mid \boldsymbol{x}] = b_x^2 = \frac{1}{n/\sigma^2 + 1/b^2} \tag{6.34}$$

As expected, this risk decreases to 0 as the size $n \to \infty$. ■

In order to illustrate CASE B in a decision problem consider the following.

♦ **EXAMPLE 6.6** (**Inventory Management** with optimal stopping point).

- The following is the simplest inventory problem that is handled daily by organizations of all sizes world wide. One such organization is *Bread-Talk Express* that supplies bread to a large community in BKK.
- Every night, the shift manager has to decide how many loafs of bread, s, to

bake for the next day's consumption.

- Let X (a random variable) be the number of units demanded during the day. If a manufactured unit is left at the end of the day we lose \$c1 on that unit. On the other hand, if a unit is demanded and is not available, due to shortage, the loss is \$c2.

Questions:

1. Assume that the daily loss includes both over-supply loss and shortage loss, what is the **loss function** $L(s, X)$ at the end of the day?
2. How many units, s, should be manufactured so that the total expected loss due to overproduction or to shortages will be minimized? Investigate solutions when the distribution F_X is known, and then unknown.

BRAINSTORMING

- Which are random variables?

- Pattern of the risk $L(s, X)$? Need a diagram for getting revenue?

- Pattern of the *expected loss* via the risk?

- When to stop produce more breads?

- How to compute optimal value s_0 of s?

GUIDANCE for solving.

1. The *loss* at the end of the day, including both over-supply loss and shortage loss is

$$L(s, X) = c_1(s - X)^+ + c_2(X - s)^+, \tag{6.35}$$

where $a^+ = \max(a, 0)$. The loss $L(s, X)$ is a random variable.

2. If the p.d.f. of X is $f(x)$, $x = 0, 1, \ldots,$ the *expected loss* $R(s) = \mathbf{E}_\theta[L(s, X)]$, is a function of the quantity s,

$$\begin{aligned} R(s) &= \sum_x L(s,x)\ f(x) = c_1 \sum_{x=0}^{s} (s-x) f(x) + c_2 \sum_{x=s+1}^{+\infty} (x-s) f(x) \\ &= c_2 \mathbf{E}[X] - (c_1 + c_2) \sum_{x=0}^{s} x f(x) + s(c_1 + c_2) F(s) - c_2 s, \end{aligned} \tag{6.36}$$

where $F(s)$ is the c.d.f. of X at $X = s$ and $\mathbf{E}[X]$ is the expected demand.

3. The optimal value s_0 of s, is the smallest integer s for which

$R(s+1) - R(s) \geq 0$. For $s = 0, 1, \cdots,$ we have

$$R(s+1) - R(s) = (c_1 + c_2) F(s) - c_2,$$

we find that (the so-called **optimal stopping point**)

$$s^0 = \textbf{ smallest non-negative integer } s, \textbf{ such that}$$

$$F(s) \geq \frac{c_2}{c_1 + c_2}. \tag{6.37}$$

In other words s^0 is the $\frac{c_2}{c_1+c_2}$-th quantile of $F(x)$.

4. We have seen that the optimal decision is a function of $F(x)$.

   If this distribution is unknown, or only partially known, one cannot determine the optimal value s^0. However, after observing a large number N of X values, by considering the empirical distribution, $F_N(x)$, of the demand and determine the level $S^0(F_N)$- one can find the smallest s value such that the

$$F_N(s) \geq \frac{c_2}{c_1+c_2}.$$

♦ **EXAMPLE 6.7. Inventory Management recalled** *for* ***CASE B***

As shown in EXAMPLE 6.6, from Equation 6.35, the Bayes estimator is

$$\widehat{\theta}_B(\boldsymbol{x}) = \textbf{the } \frac{c_2}{c_1+c_2} \textbf{ th quantile of}$$

$$\textbf{the posterior distribution of } \theta, \textbf{ given } \boldsymbol{x}_n.$$

ASSIGNMENT 2 (Defective items in Manufacturing).

- A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 7%.
- We have to decide whether **to accept or to reject** the shipment based on the model parameter θ, the proportion of defective parts.
- Before we see the real data, let's assign a 50-50 chance to both suggested values of θ, i.e., the prior distribution of θ is given as

$$h(\theta = \theta_1 = 0.05) = h(\theta = \theta_2 = 0.07) = 0.5.$$

A random sample **x** of $n = 20$ parts has 2 defective ones.

a/ Compute the marginal distribution $m(x)$ of data X for $x = 2$.

b/ Find posterior probabilities $\pi(\theta|X = x = 2)$ of parameter θ for $\theta = \theta_1 = 0.05$ and $\theta = \theta_2 = 0.07$. What possibly is your conclusion?

6.4.4 *Bayesian Testing*

New look at testing hypotheses

We discuss testing hypotheses as a **Bayesian decision problem**.

- Suppose that we consider a null hypothesis H_0 concerning a parameter θ of the p.d.f. of X.
- Suppose also that the parameter space $\boldsymbol{\Theta}$ is partitioned to two sets $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$, where
  $\boldsymbol{\Theta}_0$ is the set of values $\boldsymbol{\theta}$ corresponding to H_0 and
  $\boldsymbol{\Theta}_1$ is the complementary set of of $\boldsymbol{\Theta}_0$ in $\boldsymbol{\Theta}$.

Why? In non-Bayesian statistics, θ was not random, thus H_0 and H_A were either true (with probability 1) or false (with probability 1).

- **Principle**:

  For Bayesian tests, in order for H_0 to have a meaningful, non-zero probability, it often represents a set of parameter values instead of just one θ_0, and we are testing

$$H_0 : \theta \in \Theta_0 \quad \text{(Null hypothesis)} \tag{6.38}$$

$$H_A : \theta \in \Theta_1 \quad \text{(Alternative hypothesis)}. \tag{6.39}$$

- **Idea**:

  Comparing posterior probabilities of H_0 and H_A,

$$\mathbb{P}[\Theta_0|\mathbf{X} = \mathbf{x}] \quad \text{and} \quad \mathbb{P}[\Theta_1|\mathbf{X} = \mathbf{x}],$$

  we decide whether $\mathbb{P}[\Theta_1|\mathbf{X} = \mathbf{x}]$ is large enough to present significant evidence and to reject the null hypothesis.

The statistician has to make a decision whether H_0 is true or H_1 is true.

1. If $h(\boldsymbol{\theta})$ is a prior p.d.f. of $\boldsymbol{\theta}$ then the **prior probability** that H_0 is true is

$$\pi = \int_{\Theta_0} h(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{6.40}$$

   and the prior probability that H_1 is true is $1 - \pi = \int_{\Theta_1} h(\boldsymbol{\theta}) d\boldsymbol{\theta}$.

2. Let $d(\pi)$ be a **decision function**, assuming the values 0 and 1, i.e.,

$$d(\pi) = \begin{cases} 0, & \text{decision to accept } H_0 \ (H_0 \text{ true}) \\ 1, & \text{decision to reject } H_0 \ (H_1 \text{ true}). \end{cases}$$

3. Let w be an **indicator** of the true situation, i.e.,

$$w = \begin{cases} 0, & \text{if } H_0 \text{ true} \\ 1, & \text{if } H_1 \text{ true}. \end{cases}$$

Loss and risk

Often one can anticipate the consequences of Type I and Type II errors in hypothesis testing and assign a loss $L(\theta, \alpha)$ associated with each possible error. Here θ is the

parameter, and α is our action, the decision on whether we accept or reject the null hypothesis.

Definition 6.4 (Loss function and prior risk)**.**

We hence impose a **loss function** for erroneous decision

$$L(d(\pi); w) = \begin{cases} 0, & \text{if } d(\pi) = w \\ r_0, & \text{if } d(\pi) = 0,\ w = 1 \\ r_1, & \text{if } d(\pi) = 1,\ w = 0, \end{cases} \tag{6.41}$$

where r_0, r_1 are finite positive constants.

The **prior risk** associated with the decision function $d(\pi)$ is

$$\begin{aligned} R(d(\pi), \pi) &= d(\pi)\ r_1\ \pi + (1 - d(\pi))\ r_0\ (1 - \pi) \\ &= r_0(1 - \pi) + \underline{d(\pi)}\ [\pi\ (r_0 + r_1) - r_0]. \end{aligned} \tag{6.42}$$

We wish to choose a decision function which minimizes the prior risk $R(d(\pi), \pi)$.

■ **CONCEPT 3.** *Such a decision function is called the Bayes decision function, and the prior risk associated with the Bayes decision function is called the* ***Bayes risk****.*

According to the above formula of $R(d(\pi), \pi)$ we should choose $d(\pi)$ to be 1 if, and only if $\pi(r_0 + r_1) - r_0 < 0$.

Accordingly, the Bayes decision function is

$$d^0(\pi) = \begin{cases} 0, & \text{if } \pi \geq \dfrac{r_0}{r_0 + r_1} \\ 1, & \text{if } \pi < \dfrac{r_0}{r_0 + r_1}. \end{cases} \tag{6.43}$$

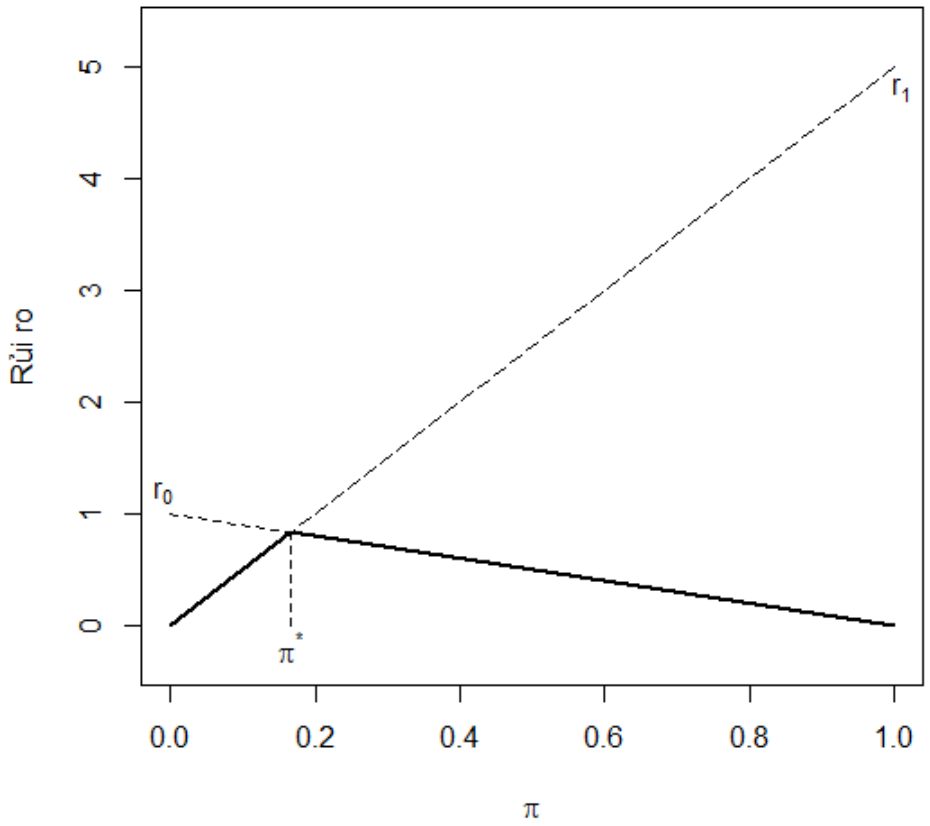


Figure 6.11: *The Bayes risk function*

Let

$$\pi^* = \frac{r_0}{r_0 + r_1}$$

and define the indicator function

$$I(\pi;\pi^*) = \begin{cases} 1, & \text{if } \pi \geq \pi^* \\ 0, & \text{if } \pi < \pi^*. \end{cases}$$

Then, the Bayes risk is

$$R^0(\pi) = r_0(1-\pi)I(\pi;\pi^*) + \pi r_1(1 - I(\pi;\pi^*)). \tag{6.44}$$

- In Fig. 6.11 we present the graph of the Bayes risk function $R^0(\pi)$ with $r_0 = 1$ and $r_1 = 5$. We see that the function $R^0(\pi)$ attains its maximum at $\pi = \pi^*$.

  The maximal Bayes risk is

$$R^0(\pi^*) = \frac{r_0 r_1}{r_0 + r_1} = \frac{5}{6}.$$

- If the value of π is close to π^* the Bayes risk is close to $R^0(\pi^*)$. The analysis above can be performed even before observations commenced.

- If π is close to 0 or to 1 the Bayes risk $R^0(\pi)$ is small we may reach decision concerning the hypotheses without even making observations.
- Observations cost money, and it might not be justifiable to spend this money.

  On the other hand, if the cost of observations is negligible compared to the loss due to erroneous decision, it might be prudent to take as many observations as required to reduce the Bayes risk.

DISCUSSION. *More on the Bayes risk* $R^0(\pi)$

1. After observing a random sample, $x_1, x_2, \cdots, x_n$, we convert the prior p.d.f. of $\boldsymbol{\theta}$ to posterior, and determine the posterior probability of H_0, namely

$$\pi_n = \int_{\Theta} h(\boldsymbol{\theta} \mid x_1, ..., x_n)\, d\boldsymbol{\theta}.$$

2. The analysis then proceeds as before, replacing π with the posterior probability π_n. Accordingly, the Bayes decision function and

   the Bayes posterior risk respectively are

$$d_0(x_1, ..., x_n) = \begin{cases} 0, & \text{if } \pi_n \geq \pi^* \\ 1, & \text{if } \pi_n < \pi^*, \end{cases}$$

$$R^0(\pi_n) = r_0(1 - \pi_n)I(\pi_n; \pi^*) + \pi_n r_1(1 - I(\pi_n; \pi^*)).$$

3. Under certain regularity conditions,

   if H_0 is true then $\lim_{n\to\infty} \pi_n = 1$; if H_0 is false then $\lim_{n\to\infty} \pi_n = 0.$.

We illustrate this with a simple example.

♦ **EXAMPLE 6.8.**

Suppose that X has a normal distribution, with known $\sigma^2 = 1$ and unknown mean μ. We wish to test $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$.

Suppose that the prior distribution of μ is also normal, $\mathbf{N}(\mu^*, \tau^2)$.

Find the posterior probability of H_0, namely π_n.

The posterior distribution of μ given $X_1, \cdots, X_n$, is normal with posterior mean

$$\mathbf{E}[\mu \mid X_1, ..., X_n] = \mu^* \frac{1}{1+n\tau^2} + \overline{X}_n \frac{n\tau^2}{1+n\tau^2},$$

and posterior variance

$$\mathbf{V}[\mu \mid X_1, ..., X_n] = \frac{n\tau^2}{1+n\tau^2}.$$

Accordingly, the posterior probability of H_0 is

$$\pi_n = \Phi\left(\frac{\mu_0 - \mu^* \dfrac{1}{1+n\tau^2} - \overline{X}_n \dfrac{n\tau^2}{1+n\tau^2}}{\sqrt{\dfrac{\tau^2}{1+n\tau^2}}} \right).$$

According to the Law of Large Numbers,

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu \right) = 1,$$

here μ is the true mean. Hence,

$$\lim_{n\to\infty} \pi_n = \begin{cases} 1, & \text{if } \mu < \mu_0 \\ \frac{1}{2}, & \text{if } \mu = \mu_0 \\ 0, & \text{if } \mu > \mu_0. \end{cases}$$

DISCUSSION.

- Notice that the prior probability that $\mu = \mu_0$ is 0. Thus, if $\mu < \mu_0$ or $\mu > \mu_0$ then $\lim_{n\to\infty} R^0(\pi_n) = 0,$ with probability one. That is, if n is sufficiently large, the Bayes risk is, with probability close to one, smaller than some threshold r^*.

  This suggests to continue, stepwise or sequentially, collecting observations, until the Bayes risk $R^0(\pi_n)$ is, for the first time, smaller than r^*.

- At stopping, $\pi_n \geq 1 - \frac{r^*}{r_0}$ or $\pi_n \leq \frac{r^*}{r_1}$. We obviously choose $r^* < \frac{r_0 r_1}{r_0 + r_1}$. ■

SUMMARY 4.

Methodology: Bayesian inference combines the information contained in the data and in the prior distribution of the unknown parameter. It is based on the **posterior distribution**, which is the conditional distribution of the unknown parameter, given the data.

The most commonly used Bayesian parameter estimator is the **posterior mean**.

NOTATIONS: In Bayesian (as in frequentist) arguments, the density of data X will be denoted by $X \sim f(x|\theta)$ for single parameter θ. It is convenient to use boldface to denote vectors, for example $\mathbf{x} = x_1, x_2, \cdots, x_n$ is data of size n, multi-parameters are denoted by $\boldsymbol{\theta}$.

- **Prior distributions** will typically be denoted by Θ with their density functions being either $h(\theta)$ or $h(\theta|\gamma)$ where γ is another parameter (sometimes called a *hyperparameter*).

- **Likelihood** is the joint conditional pdf of $\mathbf{X}$, given $\Theta = \theta$,

$$L(\theta) = f(\boldsymbol{x}|\theta) = f(x_1|\theta)\ f(x_2|\theta) \cdots\ f(x_n|\theta). \tag{6.45}$$

The joint pdf of data $\mathbf{X}$ and prior Θ is

$$g(\boldsymbol{x}, \theta) = L(\theta)\ h(\theta). \tag{6.46}$$

- **The posterior p.d.f.** of θ given data $\mathbf{x}$, corresponding to the prior p.d.f. $h(\theta)$ (or also $\pi(\theta)$) can be written by $\pi(\boldsymbol{\theta}|\,\mathbf{x})$ or $\pi(\boldsymbol{\theta}|\,\mathbf{x}, \gamma)$:

$$\begin{aligned} \pi(\theta|\,\mathbf{x}) &= \frac{g(\boldsymbol{x}, \theta)}{m(\mathbf{x})} = \frac{L(\theta)\ h(\theta)}{m(\mathbf{x})} \\ &= \frac{\prod_{i=1}^{n} f(x_i|\theta)\ h(\theta)}{m(\mathbf{x})} \end{aligned} \tag{6.47}$$

where $m(\mathbf{x})$ is the marginal distribution or the *joint predictive p.d.f.* (the Evi-

dence) of data,

$$m(\mathbf{x}) = \int_{\theta \in \mathcal{M}} f(\mathbf{x}\,|\theta)\; h(\theta)\; \mathrm{d}\theta = \int_{\theta \in \mathcal{M}} f(\mathbf{x}\,|\theta)\; \pi(\theta)\; \mathrm{d}\theta. \tag{6.48}$$

6.5 Chapter's Final Review and Problems

PROBLEM 6.1.

Suppose that $X_1, X_2, \cdots, X_n$ are **i.i.d.** $X \sim \text{Poiss}(\lambda)$, and denote an estimator

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Find the **Bias**, the standard error se and the **MSE** of this estimator.

PROBLEM 6.2. *(Network blackouts, solution of ASSIGNMENT 1)*

The number of network blackouts each week at MU has Pois(θ) distribution. The weekly rate θ of blackouts θ is not known exactly, but according to the past experience with similar networks, it averages $\mu = \mathbf{E}[\theta] = 4$ blackouts with a $\sigma = 2$.

After two weeks of data, we observed sample $\mathbf{x} = (2, 0)$.

GUIDANCE for solving.

There exists a Gamma distribution with the given mean $\mu = \alpha/\lambda$, and the standard deviation $\sigma = \sqrt{\alpha}/\lambda = 2$. As a result,

$$\mu = 4 = \frac{\alpha}{\lambda}; \quad \sigma = 2 = \sqrt{\alpha}/\lambda.$$

- A specific root of this system is $(\alpha = 4, \lambda = 1)$.
- Hence, we can assume the Gamma$(\alpha = 4, \lambda = 1)$ prior distribution for θ. After two weeks of data, the weekly rate of network blackouts, according to Problem **??** has Gamma posterior distribution $\boldsymbol{G}(\alpha_x = \alpha + \sum_i x_i, \lambda_x = \lambda + n)$.
- Since $x_1 = 2$ blackouts this week, the posterior distribution of θ is Gamma with parameters $\alpha_x = \alpha + \sum_i x_i = 6, \lambda_x = \lambda + n = 2)$.
- Since $x_2 = 0$ blackouts the 2nd week, the posterior distribution of θ is updated to a Gamma with parameters

  $\alpha_x = \alpha + \sum_i x_i = 4 + 2 + 0, \lambda_x = \lambda + n = 1 + 2)$, or parameters $\alpha_x = 6$; $\lambda_x = 3$.

SUMMARY: The Bayes estimator of the weekly rate θ is

$$\widehat{\theta}_B = \mathbf{E}[\theta|\,\mathbf{x}] = \frac{\alpha_x}{\lambda_x} =? \text{ blackouts per week}$$

with a posterior risk $\rho(\widehat{\theta}_B) = \mathbf{V}_h[\theta \mid \boldsymbol{x}] = \dfrac{\alpha_x}{\lambda_x^2} =?$ What is your conclusion?

This posterior distribution has the average weekly rate $\widehat{\theta}_B = 6/3 = 2$ blackouts per week. Two weeks with very few blackouts reduced our estimate of the average rate from 4 to 2. ■

PROBLEM 6.3.

Let $X_1, X_2, \cdots, X_n \sim_{iid} X$ be a random sample, where both $\theta = \mathbf{E}[X]$ and $\sigma^2 = \mathbf{V}[X]$ are unknown. Denote the sample variance by

$$S^2 = \frac{1}{n-1}\sum_i [X_i - \overline{X}]^2.$$

B1 Prove mathematically that the sample standard deviation S is a *consistent* estimator of σ, if $\mathbf{E}[X^4] < \infty$.

B2 Now suppose that $X \sim \mathbf{N}(\theta, \sigma^2)$ is a Gaussian. Consider the sequence $\{\overline{X}_{\mathrm{n}}\}_n$ defined by $\overline{X}_{\mathrm{n}} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Prove explicitly that, as $n \longrightarrow \infty$,

$$\mathbb{P}\Big[\,|\,\overline{X}_{\mathrm{n}} - \theta|\,< \varepsilon\Big] \longrightarrow 1$$

to conclude that $\overline{X}_{\mathrm{n}}$ is a *consistent* sequence of estimators of θ.

GUIDANCE for solving.

B1 The sample std. deviation S is a *consistent* estimator of σ? We prove that

$$S^2 = S_n^2 = \frac{1}{n-1}\sum_i [X_i - \overline{X}]^2 \xrightarrow{P} \sigma^2.$$

$\mathbf{E}[X^4] < \infty$ implies that $\mathbf{V}[S^2] < \infty$. First, the sample variance

$$S^2 = \frac{1}{n-1}\sum_i [X_i - \overline{X}]^2 = (n-1)^{-1}\Big(\sum_{i=1}^{n} X_i^2 - n\ \overline{X}^2\Big), \tag{6.49}$$

By the Weak LLN on the 2nd sample moment $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ and Equation 6.49 we have

the following

$$S_n^2 = \frac{1}{n-1}\sum_i [X_i - \overline{X}_{\rm n}]^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_{\rm n}^{\,2}\right) \xrightarrow{P} 1 \cdot \left[\mathbf{E}[X^2] - \mu^2\right] = \sigma^2. \tag{6.50}$$

From the discussion above, we have immediately that $S_n \xrightarrow{P} \sigma$, so $S = S_n$ is a *consistent* estimator of σ.

PROBLEM 6.4.

ASEAN nations- in particular the *Mekong River Basin* with countries of Lao, Thailand, Cambodia and Vietnam- are experiencing extreme natural threats or disasters (to be defined as either *flooding* or *drought*) for years. Those extreme phenomena are called **rare events**, and a Poisson distribution $\text{Poiss}(\theta)$ is used to model them.

Scientists of the *Mekong River Committee* (MRC) observed X, the monthly number of natural disasters continuously in the last 5 years in the entire region. They code threat levels in five codes of 0, 1, 2, 3, 4 where the larger value means the more serious level the disaster is. The following table shows observed data representing a summary of a random sample of size $n = 60$.

A1/ Find the maximum likelihood estimate (MLE) $\widehat{\theta}$ of θ.

A2/ The most fatal events are expressed by codes 3 and 4.

Compute the probability $\mathbb{P}(X \geq 3)$ (measuring how much likelihood the most fatal

events happened in the past 5 years), from which the MRC could workout risk-control solutions.

GUIDANCE for solving.

A1/ (4 points)

Carry out the ML Estimation process to get $\lambda_{ML} = \widehat{\theta} = \overline{X}$:

- Since X_i are IID, we can write the likelihood function (of observing data vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$) with respect to θ by the formula

$$L(\theta) = f(\boldsymbol{X}; \theta) = \prod_i p(X_i; \theta). \ (1 \ point)$$

- Take derivative of $L_* = \log L(\theta)$ with respect to variable θ and solve equation

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0 \Leftrightarrow \widehat{\theta} = \overline{X} \, .$$

Check that $\dfrac{\partial^2 L_*(\widehat{\theta})}{\partial \theta^2} < 0$. (2 points)

The estimate $\theta_{ML} = \widehat{\theta} = \overline{X} = (\sum x_i)/n = (1.14 + ... + 4.11)/60 = 2.03?$ (1 point)

x	0	1	2	3	4
Frequency	9	14	14	12	11

A2/ (4 points)

The probability $\mathbb{P}[X \geq 3]$ of most fatal events that is associated with the ML estimate λ_{ML} is

$$\mathbb{P}[X \geq 3] = \sum_{3 \leq x} \mathbb{P}[X = x] = 1 - \mathbb{P}(X < 3] = 1 - F(2),$$

where

$$\mathbb{P}[X = x] = \frac{\widehat{\lambda}^x \, e^{-\widehat{\lambda}}}{x!} \qquad x = 0, 1, 2, ... \tag{6.51}$$

is just the pmf of the Poisson distribution $\text{Pois}(\widehat{\lambda})$ (3 points).

So $\mathbb{P}(X \geq 3) = 1 - F(2) = 1 - [\mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2)] = 0.3314$

PROBLEM 6.5.

- A home appliance firm orders a shipment of spare items from a manufacturer.
- The manufacturer claims that 5% of the items in the shipment is defective, but the firm's inspector feels that in fact it is 8%.

  We have to decide whether **to accept or to reject** the shipment based on the model parameter θ, the proportion of defective items.
- Before we see the real data, let's assign a 40-60 chance to both suggested values of θ, i.e., the prior distribution of θ is given as

$$h(\theta = \theta_1 = 0.05) = 0.4; \text{ and } h(\theta = \theta_2 = 0.08) = 0.6.$$

  A random sample **x** of $n = 20$ parts has 3 defective ones.

C1/ Compute the marginal distribution $m(x)$ of data X for $x = 3$.

C2/ Find posterior probabilities $\pi(\theta|X = x = 3)$ of parameter θ for $\theta = \theta_1$ and $\theta = \theta_2$.

What possibly is your conclusion?

GUIDANCE for solving.

C1/ (2 points) The prior densities of θ are

$$h(\theta_1 = 0.05) = 0.4; \text{ and } \ h(\theta_2 = 0.08) = 0.6.$$

The marginal distribution of data $X = x$ is

$$m(x) = \sum_{\theta} f(x|\theta)\ h(\theta) = f(x|\theta_1)\ h(\theta_1) + f(x|\theta_2)\ h(\theta_2).$$

So the marginal distribution of data $X = x = 3$, with $\theta_1 = 0.05,\ \theta_2 = 0.08$ is

$$\begin{aligned} m(3) &= \sum_{\theta} f(x|\theta)\ h(\theta) \\ &= f(3|\ 0.05)\ h(0.05) + f(3|\ 0.08)\ h(0.08) = 0.108 \end{aligned}$$

C2/ The posterior probability of parameter θ is

$$\pi(\theta|\mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x};\theta)}{m(\mathbf{x})} = \frac{f(\mathbf{x}\,|\theta)\ h(\theta)}{m(\mathbf{x})},$$

as a result, posterior probabilities of parameter θ for $\theta = \theta_1 = 0.05$ and $\theta = \theta_2 = 0.08$ are

$$\pi(\theta_1|X = x = 3) = \frac{f(x|0.05)\ h(0.05)}{m(3)} = 0.219$$

$$\pi(\theta_2|X = x = 3) = \frac{f(x|0.08)\ h(0.08)}{m(3)} = 0.780$$

(3 points)

Conclusion:

At first, we had no preference between the two suggested values of θ.

Then we observed a rather high proportion of defective parts, $3/20 = 15\%$.

Taking this into account, the likelihood of choosing $\theta = 0.08$ is now around 780/219= 3.5 times that of $\theta = 0.05$. (3 points)

PROBLEM 6.6.

Bread-Boutique firm supplies bread to a large community in Bangkok. Every night, the shift manager has to decide how many loafs of bread, $s \in \mathbb{N}$, to bake for the next day's consumption.

Let Y be the number of units demanded during the day. If a manufactured unit is left at the end of the day we lose c_1 baht on that unit. On the other hand, if a unit is demanded and is not available, due to shortage, the loss is c_2 baht.

D1/ Suppose that the daily loss includes both over-supply loss and shortage loss, what is the **loss function** $L(s, Y)$ at the end of the day? (4 points)

D2/ Let $f(y)$ and $F(y)$ ($y = 0, 1, 2, \ldots$) be the p.d.f. and c.d.f of the **bread demand** Y, and let $R(s) = \mathbf{E}[L(s, Y)]$ be the **expected loss** - as a function of the quantity s.

The optimal stopping value $s_0 \in \mathbb{N}$ is the smallest value of s such that

$$R(s+1) - R(s) \geq 0.$$

Find s_0, when the unit over-supply loss $c_1 = 4$ baht, the unit shortage loss $c_2 = 21$ baht, and the demand $Y \sim \mathbf{N}(\mu, \sigma^2)$ where $\mu = 2000,\ \sigma = 100$. (6 points)

GUIDANCE for solving.

D1/ The **loss function** $L(s, Y)$ at the end of the day? (4 points)

The daily loss includes both over-supply loss $L_0 = c_1(s - Y)^+$ and shortage loss $L_1 = c_2(Y - s)^+$, hence (2 points)

$$L(s, Y) = L_0 + L_1 = c_1(s - Y)^+ c_2(Y - s)^+, \quad \text{where} \quad a^+ = \max(a, 0).$$

The loss $L(s, Y)$ is a random variable, because Y is a random variable, while s is just a variable. (2 points)

D2/ The optimal stopping value s_0? (6 points)

– Compute the **expected loss** $R(s) = \mathbf{E}[L(s, Y)]$ as

$$R(s) = c_1 \sum_{y=0}^{s} (s - y) f(y) + c_2 \sum_{y=s+1}^{+\infty} (y - s) f(y) \tag{6.52}$$

(1 point) then obtain that

$$R(s) = c_2 \mathbf{E}[Y] - (c_1 + c_2) \sum_{y=0}^{s} y f(y) + s(c_1 + c_2) F(s) - c_2 s,$$

where $F(s)$ is the c.d.f. of Y at $Y = s$. (1 point)

– We know that the optimal value s_0 of s is the smallest integer s for which

$R(s+1) - R(s) \geq 0$. (1 point) Simplify

$$R(s+1) - R(s) = \ldots = (c_1 + c_2) F(s) - c_2$$

– Conclude that the optimal stopping point

$$s^0 = \textbf{ smallest non-negative integer } s, \textbf{ such that}$$

$$F(s) \geq \frac{c_2}{c_1 + c_2}. \tag{6.53}$$

Hence s^0 is the $\frac{c_2}{c_1 + c_2}$-th quantile of $F(y)$. (1 point)

– If the unit over-supply loss $c_1 = 4$ baht, the unit shortage loss $c_2 = 21$ baht, then $\frac{c_2}{c_1 + c_2} = 0.84$.

The bread demand $Y \sim \mathbf{N}(\mu, \sigma^2)$ where $\mu = 2000,\ \sigma = 100$, therefore s^0 is the 84%-th quantile of $F(y)$.

– Finally, the domain bounded by the Gaussian curve $f(y)$ within the two lines $\mu - \sigma, \mu + \sigma$ gives an area of 0.68. Hence the 84%-th quantile

$$s^0 = \Phi^{-1}(0.84) = \mu + \sigma = 2000 + 100 = 2100$$

loafs of bread. (2 points)

This page is left blank intentionally.

Bibliography

[1] ALEXANDER HOLMES, Introductory Business Statistics, OpenStax, Rice University, 2017

[2] Anna Mikusheva, course materials for 14.384 Time Series Analysis, Fall 2007. MIT OpenCourseWare (http://ocw.mit.edu), Massachusetts Institute of Technology, nology. Downloaded on [01 June 2021]

Annette J. Dobson and Adrian G. Barnett,

An Introduction to Generalized Linear Models, Third Edition, CRC (2008)

[3] Annette J. Dobson and Adrian G. Barnett,

An Introduction to Generalized Linear Models, Third Edition, CRC (2008)

[4] Antal Kozak, Robert A. Kozak, Christina L. Staudhammer, Susan B. Watts *Introductory Probability and Statistics Applications for Forestry and Natural Sciences*, CAB International (2008)

[5] *Canvas paintings* by Australian artists of ethnic minorities, Australian National Museum

[6] **Practical Optimization: a Gentle Introduction**, John W. Chinneck, 2000

[7] **Introduction to Linear Optimization**, Dimitris Bertsimas and John N. Tsitsiklis, Athena Scientific, 1997

[8] U. N. Bhat. *A controlled transportation queueing process*. Management Science, 16(7): 446-452, 1970.

[9] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*, Second Ed. Springer, 2006.

[10] David S. Moore, George P. McCabe and Bruce A. Craig.
Introduction to the Practice of Statistics, 6th edition, (2009) W. Freeman Company, New York

[11] Man VM Nguyen. *Data Analytics- Foundation: Inference, Regression and Stochastic Processes*. Saarbrücken Germany: LAP LAMBERT Academic Publishing (2020) ISBN 978-620-2-79791-7

[12] Douglas C. Montgomery, George C. Runger,
Applied Statistics and Probability for Engineers, Sixth Edition, (2014) John Wiley & Sons

[13] Jay L. Devore and Kenneth N. Berk,
Modern Mathematical Statistics with Applications, 2nd Edition, Springer (2012)

[14] Glonek G.F.V. and Solomon P.J. *Factorial and time course designs for cDNA microarray experiments, Biostatistics* **5**, 89-111, 2004.

[15] Hedayat, A. S., Sloane, N. J. A. and Stufken, J. *Orthogonal Arrays, Springer-Verlag*, 1999.

[16] Robert V. Hogg, Joseph W. McKean, Allen T. Craig *Introduction to Mathematical Statistics,* Seventh Edition *Pearson*, 2013.

[17] David Ruppert and David S. Matteson. *Statistics and Data Analysis for Financial Engineering with R examples*, Second Edition. Springer (2015)

[18] Härdle, Wolfgang, and Léopold Simar. Applied multivariate statistical analysis. 2nd. Springer, 2007.

[19] Inada, T., Shimamura, Y., Todoroki, A., Kobayashi, H., and Nakamura, H., Damage Identification Method for Smart Composite Cantilever Beams with Piezoelectric Materials, Structural Health Monitoring 2000, Stanford University, Palo Alto, California, 1999,pp. 986-994.

[20] Jolliffe, I. T. Principal component analysis. 2nd. Springer, 2002.

[21] Lapin, L.L. , Probability and Statistics for Modern Engineering, PWS-Kent Publishing, 2nd Edition, Boston, Massachusetts,1990.

[22] Ljung, L. System identification: theory for the user, Prentice Hall, Englewood Cliffs, NJ, 1987

[23] Mahmut Parlar, *Interactive operations research with Maple: methods and models*, (2000) Springer

[24] Paul Mac Berthouex, Linfield C. Brown, *Statistics for Environmental Engineers*, 2nd Edition, LEWIS PUBLISHERS, CRC Press, 2002

[25] Madhav, S. P., *Quality Engineering using robust design, Prentice Hall*, 1989.

[26] Michael Baron, *Probability and Statistics for Computer Scientists*, 2nd Edition (2014), CRC Press, Taylor & Francis Group

[27] R. H. Myers, Douglas C. Montgomery and Christine M. Anderson-Cook
Response Surface Methodology : Process and Product Optimization Using Designed Experiments, Wiley, 2009.

[28] Man Nguyen, Tran Vinh Tan and Phan Phuc Doan,
Statistical Clustering and Time Series Analysis for Bridge Monitoring Data, Recent Progress in Data Engineering and Internet Technology, Lecture Notes in Electrical Engineering 156, (2013) pp. 61 - 72, Springer-Verlag

[29] Nguyen Van Minh Man,
Computer-Algebraic Methods for the Construction of Designs of Experiments, Ph.D. thesis, 2005

[30] Nguyen, Man V. M.
Some New Constructions of strength 3 Orthogonal Arrays,
the Memphis 2005 Design Conference Special Issue of the **Journal of Statistical Planning and Inference**, Vol 138, Issue 1 (Jan 2008) pp. 220-233.

[31] Nathabandu T. Kottegoda, Renzo Rosso. *Applied Statistics for Civil and Environmental Engineers*, 2nd edition (2008), Blackwell Publishing Ltd and The McGraw-Hill Inc

[32] Man Nguyen (2005) *Computer-algebraic Methods for the Construction of Design of Experiments*, Ph.D thesis, Eindhoven Univ. Press, 2005

[33] Paul Mac Berthouex. L. C. Brown. *Statistics for Environmental Engineers*; 2nd edition (2002), CRC Press

[34] A. Ravi Ravindran (editor). *Operations Research and Management Science Handbook*, CRC, 2008

[35] Alvin C. Rencher and William F. Christensen, *Methods of Multivariate Analysis*, Wiley, 2012

[36] Ron S. Kenett, Shelemyahu Zacks.
Modern Industrial Statistics with Applications in R, MINITAB, 2nd edition, (2014), Wiley

[37] Robert H. Shumway and David S. Stoffer.
Time Series Analysis and Its Applications: With R Examples, **Springer Texts in Statistics**, 3rd Edition, (2011)

[38] W.M.P. Aalst, van der PROCESS MINING, 2nd edition, 2016, Springer

[39] Kurt Jensen, Wil M.P. van der Aalst, Gianfranco Balbo, Maciej Koutny, Karsten Wolf (Eds.).
Transactions on Petri Nets and Other Models of Concurrency VII LNCS 7480, Springer, 2013 .

[40] Sheldon M. Ross. *Introduction to probability models*, 10th edition, (2010), Elsevier Inc.

[41] Sloane N.J.A., `http://neilsloane.com/hadamard/index.html/`

[42] Google Earth, Digital Globe, 2014- 2019

[43] Larry Wasserman, *All of Statistics- A Concise Course in Statistical Inference*, **Springer**, (2003)

[44] C.F. Jeff Wu, Michael Hamada *Experiments: Planning, Analysis and Parameter Design Optimization,* Wiley, 2000.

[45] Wendy L. Martinez and Angel R. Martinez, *Computational Statistics Handbook with MATLAB*, CHAPMAN & HALL/CRC, 2002