Dan Sasaki
Institute of Social Science
dsasaki@iss.u-tokyo.ac.jp

Welcome to

28 February 2011

Bayesian vs non-Bayesian statistical thinking

A quick sneak preview of contents:

0. Introduction.

Please try warming up your statistical mind by exercise examples in 0.0.

The remainder, 0.1 through 0.3, should be read only if you are curious.

1. Probability and statistics.

Introduction of somewhat technical mathematical concepts in 1.1 through 1.3 is mostly for completeness and also for your possible intellectual curiosity.

The practice problems in 1.4 can be done as long as you have good sense of statistical humour, that is, even without reading through 1.1 through 1.3.

2. Statistics and econometrics.

This chapter may be useful especially if your statistics/econometrics has been rusty.

Even if you are a regular user of statistical packages and/or data, it is always recommendable to ascertain that you comprehend the theoretical workmanship behind the scene.

3. Information and decisions.

The dichotomy between the Bayesian and the non-Bayesian (frequentist) extends beyond narrowly defined statistics and econometrics, to carry over to decision theory, game theory, and microeconomic theory in general. By and large, microeconomic game theory is predominantly Bayesian whilst mainstream statistics in mathematical and engineering fields tends to be non-Bayesian.

This chapter sheds light on how statistical decisions are to be made in various practical settings, which are not necessarily confined to what we would ordinarily regard as statistical or econometric research.

4. Rationality.

Econometrics often assumes that not only the researcher but also the subjects must be economically rational. What do we mean by this?

We shall discuss this chapter only if time allows.

0. Introduction

0.0. Bayesing-up exercise

0.0.1. Two identical sealed envelopes are put on the table. You are told by the instructor that each envelope contains a cheque, one of them carrying 20 euros more than the other. However, neither you nor the instructor can foretell which one contains more than the other (both cheques are supposed to carry positive amounts). You are asked to take whichever one of the two envelopes you like.

Determine whether each of the following statements (in *slanted fonts*) is true, or false.

0.0.1.i. After you have picked up one of the two envelopes but *before* you open it, the instructor offers you that you may, if you wish to, exchange the envelopes so as to take the other one instead. *Your exchanging the envelopes will neither increase nor decrease the expected amount of money you receive.*

0.0.1.ii. Now you have opened one of the envelopes to find therein a cheque which is written so calligraphically that it is almost illegible to you, but the instructor has pronounced it for you which sounded "cent (100) euros." Conditional upon this information, *the other envelope contains either 80 euros with probability one half, or 120 euros with probability one half. The expected amount in the other envelope is therefore* $0.5 \times 80 + 0.5 \times 120 = 100$ *euros.*

0.0.1.iii. *In general, if one of the envelopes contains an amount* x *, the conditional expected amount in the other envelope is also* $0.5(x-20) + 0.5(x+20) = x$ *.*

0.0.1.iv. Now it has turned out, however, that the instructor actually said "cinq (5) euros" instead. Conditional upon this revised information, *the other envelope cannot carry* -15 *euros and thus must carry 25 euros.*

0.0.1.v. *All in all, if one of the envelopes contains an amount* $x > 20$ *, the conditional expected amount in the other envelope is also* $0.5(x-20) + 0.5(x+20) = x$ *, whilst if the former carries* $y < 20$ *euros the other must carry* $y + 20$*.*

0.0.1.vi. By 0.0.1.v, your exchanging the envelopes can never decrease your expected gain and can possibly increase it. Therefore, as in 0.0.1.i, *you should always exchange the envelopes even before opening one.*

0.0.2. You have taken an IQ test, of which the score is known to be distributed approximately symmetrically around the true IQ of the person, and your score is 140. Which one of the following ⟨a⟩ through ⟨e⟩ describes the uncensored truth?

⟨a⟩ Your true IQ is more likely to be below, than above, 140.

⟨b⟩ Your true IQ is approximately equally likely to be either below or above 140.

⟨c⟩ Your true IQ is more likely to be above, than below, 140.

⟨d⟩ Your true IQ is more likely to be 140 than either above or below.

⟨e⟩ Your true IQ is approximately symmetrically distributed around 140.

0.0.3. Our allegedly well educated instincts seem to tell us invariably that stupid people outnumber wise people in our troubled world. If so, why is it always the majority opinion, not the minority opinion, that democracy is supposed to heed? Choose the correct explanation from among the following ⟨a⟩ through ⟨e⟩.

⟨a⟩ Democracy, by definition, is for popularity not for correctness. Were we interested in making correct decisions, we would respect only those opinions of the wisest few and ignore the rest of the public.

⟨b⟩ Should the opinions split sharply betwixt the wise on one side and the stupid on the other, then we should undoubtedly listen to the former in order to achieve a wise decision. In reality, however, wise people disagree amongst themselves, so do stupid people, thus on average those of one opinion differ little in terms of wisdom or stupidity from those of another opinion.

⟨c⟩ Wise people can often influence the public opinion, so that they can in effect exercise more votes than the rest of the crowd.

⟨d⟩ When a sufficiently large majority agree, their opinion is more likely to be correct than the opposing opinion even when the supporters of the latter are substantially wiser than those of the former, and when each individual form her/his opinion independently.

⟨e⟩ All of the above.

0.0.4. Two coins have been tossed.

0.0.4.i. Someone whispers that one is a head. Assuming that this information is reliable, what is the probability that the other is also a head ?

0.0.4.ii. You have taken a glimpse at one of the coins, which is a head. What is the probability that the other is also a head ?

0.0.4.iii. Someone reports that at least one is a head. Assuming that this report is truthful, what is the probability that both are heads ?

0.0.5. Two identical sealed envelopes are put on the table. You are told by the instructor that each envelope contains a cheque, one of them exactly twice the amount of the other. However, neither you nor the instructor can foretell which one contains more than the other (both cheques are supposed to carry positive amounts). You are asked to take whichever one of the two envelopes you like.

Determine whether each of the following statements (in *slanted fonts*) is true, or false.

0.0.5.i. After you have picked up one of the two envelopes but *before* you open it, the instructor offers you that you may, if you wish to, exchange the envelopes to take the other one instead. *Your exchanging the envelopes will neither increase nor decrease the expected amount of money you receive.*

0.0.5.ii. Now you have opened one of the envelopes to find out that the cheque inside it is, say, ten euros. Conditional upon this information, *the other envelope contains either five euros with probability one half, or twenty euros with probability one half. The expected amount in the other envelope is therefore* $5 \times 0.5 + 20 \times 0.5 = 12.5$ *euros.*

0.0.5.iii. *In general, if one of the envelopes contains an amount* x *, the conditional expected amount in the other envelope is always* $1.25x$ *.*

0.0.5.iv. Now that you have seen the amount x in one of the envelopes, the instructor tells you that you may exchange the envelopes and have the other one instead. *Insofar as* $x > 0$ *, your exchanging the envelopes will increase the conditional expected amount of money you receive.*

— *Please try all the exercise examples on your own, including the above, prior to the intensive sessions scheduled on 28 February 2011.* —

0.1. Objectives and the scope

Econometrics is a practical, quantitatively analytical science to take a close look at how the economy as a whole is performing, how specific industries and markets are functioning, how traders (firms, consumers, etc.) interact either strategically or otherwise, and how exogenous environments including policy variables affect the economy and the participants therein. In brief, it provides a framework for practical analysis through which basic economic theory can be applied to useful economic problems. This, obviously, encompasses both the descriptive side, how to make sense of real economic phenomena in the light of economic theory, as well as the more affirmative side, how to build and solve a model which predicts the expected outcomes of a hypothetical economic decision. The lectures are to overview *both* the basic nuts and bolts of econometric theory, especially in the contexts of Bayesian and non-Bayesian statistical thoughts, *and* some of those commonly debated econometric issues inasmuch as time allows.

0.2. Aims and expected learning outcomes

The lectures cover some of the most elementary and the most typical issues in econometrics, through which a practically useful "package" of knowledge, thinking, and analytical tools, will be built. The knowledge includes not only what has been popularly studied by contemporary econometricians, but more importantly those key terms and concepts in econometrics which can almost instantly form part of in-depth scientific understanding of our contemporary economy and various economic issues and decisions therein. The thinking, obviously the most important of all, is about how to link basic economic theory with more concrete economic issues. This should include both the "deductive" ability to apply general abstract theory to specific practical problems, and the "inductive" ability to distill a set of seemingly complicated practical economic problems into a simple general principle, often focusing on one aspect (such as one specific variable) of practical problems at a time. Finally, the analytical tools include some of the basic mathematical skills particularly useful in analysing interactions between economic bodies such as traders and/or policy makers. Among these tools are probability theory and statistical decision theory.

Although most of those specific issues explicitly discussed in these lectures are confined within the scopes of economics and econometrics, part of the ultimate goal of this lecture series is to cultivate the aforementioned package into a highly practical toolbox which can also be applied to many other problems in and outside economics. Such a package

is highly "portable" across disciplines as well as across different cultures and working environments, thereby forms part of personal transferable skills.

0.3. Learning and instruction methods

All lectures are made available in print. However, these are by no means a comprehensive "textbook". Reference to more formal textbooks and other sources is strongly encouraged whenever necessary, although it is generally *un*necessary to read (or even worse, try to memorise) a thick textbook from cover to cover. It is also unnecessary, and more often harmful than useful, to memorise every material listed in lectures. Essentially, the lectures should be treated as a "glossary" for you to overview which topics are covered, as well as for me to remember what to mention in explanatory sessions.

Collective sessions are treated as "tutorials" meant to give an audience-friendly commentary on lecture materials. It is often difficult, especially for first-time learners, to digest a written material alone. Part of collective sessions shall be spent on various discussions related, but not strictly confined, to the lecture materials.

Comments and questions on these lectures, and on other materials related to the subject, are more than welcome and positively encouraged.

1. Probability and statistics

1.1. Elementary probability theory

Partition

A partition of a set Ω is a collection of subsets of Ω that is disjoint and exhaustive. Namely, $\Pi = \{P_i\}_i$ is a partition of Ω if :

- $P_i \subseteq \Omega$ for all i ,
- $P_i \cap P_j = \phi$ for all $i \neq j$, and
- $\bigcup_i P_i = \Omega$.

A partition can be defined over any set. More than one partition may be definable over the same set. For any set Ω , for example, $\Pi = \{\Omega\}$ defines a **trivial partition** of Ω .

When there are two partitions $\Pi^1 = \{P_i^1\}_i$ and $\Pi^2 = \{P_j^2\}_j$ defined over the same set Ω , the former is a **coarsening** of the latter and the latter is a **refinement** of the former if, for any j , there exists an i such that $P_j \subseteq P_i$.

Example : When the **universal set** is $\Omega \equiv \{$, $\}$, there can be defined two partitions over this set. One is the *trivial* partition $\Pi^T \equiv \Big\{\{$, $\}\Big\}$, and the other is the finest partition (sometimes referred to as the **discrete partition**) $\Pi^F \equiv \Big\{\{\}, \{\}\Big\}$.

σ-algebra (σ-field)

A σ-algebra or a σ-field defined over a set Ω is a collection of subsets of Ω , denoted by $\Sigma = \{S_i\}_i$ hereinafter, that satisfies the following requirements.

- $S_i \subseteq \Omega$ for all i .
- Closed under complements, i.e., $(\Omega \setminus S_i) \in \Sigma$ for all i .
- Closed under *countable* (finite and countably infinite) unions, i.e., $\bigcup_{i \in I} S_i \in \Sigma$ for any *countable* set I.
- $\phi \in \Sigma$ and $\Omega \in \Sigma$.

Like a partition, a σ-algebra can also be defined over any set. More than one σ-algebra may be definable over the same set. For any set Ω, $\mathrm{T} = \{\Omega, \phi\}$ defines a **trivial σ-algebra** over Ω.

When there are two σ-algebras Σ^1 and Σ^2 defined over the same set Ω, the former is a **coarsening** of the latter and the latter is a **refinement** of the former if $\Sigma^1 \subseteq \Sigma^2$. Therefore, a coarsening of a σ-algebra is also called a **sub σ-algebra**. Note in particular that a sub σ-algebra over Ω does not refer to a σ-algebra defined over a subset of Ω.

When the set Ω is countable, practically the most important σ-algebra is the **discrete σ-algebra**, which is the collection of all subsets of Ω, denoted by 2^Ω. When the set Ω is uncountable and a metric space (Ω, d) is defined, the **Borel σ-algebra** is commonly used, which is the smallest σ-algebra (or more formally, the *intersection* of all σ-algebras) containing all open subsets of Ω.

Example : When the **universal set** is $\Omega \equiv \{$ ☻, ☹ $\}$, there can be defined two σ-algebras over this set. One is the *trivial* σ-algebra $\mathrm{T} \equiv \big\{\{\,\}, \{$ ☻, ☹ $\}\big\}$, and the other is the *discrete* σ-algebra $2^\Omega \equiv \big\{\{\,\}, \{$ ☻ $\}, \{$ ☹ $\}, \{$ ☻, ☹ $\}\big\}$.

Measure

A measure is a function μ defined over a σ-algebra $\Sigma = \{S_i\}_i$ such that :

- $\mu[S_i] \geq 0$ for all i,
- $\mu[\phi] = 0$, and
- if $S_i \cap S_j = \phi$ then $\mu[S_i] + \mu[S_j] = \mu[S_i \cup S_j]$.

When a σ-algebra Σ is defined over a set Ω, a measure μ defined over Σ is a **probability measure** if, in addition to the above three requirements, it satisfies the additional condition

- $\mu[\Omega] = 1$.

Depending upon the context, especially when the set Ω is a subset of $\mathbb{R}^k$ and the σ-algebra Σ is the Borel σ-algebra over Ω, a measure often refers to the **Lebesgue measure** unless otherwise specified. Intuitively, a Lebesgue measure is the (k-dimensional) *volume* of each subset included in Σ. An arc has a positive length but its area is zero. Therefore, its one-dimensional Lebesgue measure is positive, whilst its multi-dimensional Lebesgue measures

are zeros. A surface has a positive area but a zero volume. Thus, its two-dimensional Lebesgue measure is positive but its three- and higher-dimensional Lebesgue measures are zeros. In general, a k-dimensional **hypersurface** has a positive k-dimensional Lebesgue measure, while its $(k+1)$- and higher-dimensional Lebesgue measures are all zeros.

When Ω is the real line $\mathbb{R}$, the Lebesgue measure of any countable subset is zero. There also are uncountable subsets of which the Lebesgue measure is zero.

Probability space

A probability space is defined by a triple (Ω, Σ, μ), where Ω is a set, Σ is a σ-algebra over Ω, and μ is a probability measure over Σ. Each element of Ω is called a **state**, therefore a probability space is also called a **state space**. On the other hand, each element of Σ, which is a subset of Ω by definition, is called an **event**.

A **measure space** can be defined likewise except, of course, that the measure μ need not be a probability measure.

Measurability

A function $f : \Omega \to Q$ is measurable with respect to a partition $\Pi = \{P_i\}_i$ of Ω if f is *constant* over P_i for every i.

A function $f : \Omega \to Q$ is measurable with respect to a σ-algebra $\Sigma = \{S_i\}_i$ over Ω if $f^{-1}[q] \in \Sigma$ for any $q \in Q$.

In economic theory, **information** can be defined by either a partition or a σ-algebra, respectively called an **information partition** and an **information field**. A decision maker's action must be a measurable function of the state with respect to his or her information field or information partition.

When there are two sets of information, if the partition or σ-field associated with one is a refinement of that with the other, then these two sets of information are **ranked** (or **nested**). Otherwise, if neither of the two partitions or σ-fields is a refinement of the other, then the two sets of information are **differential** (or **nonnested**).

Note that, if Ω is countable (either finite or countably infinite), an information partition and an information field are equivalent. Namely, any information expressed with a partition can be rewritten with a σ-algebra, and vice versa. If Ω is uncountable, on

the other hand, there exists such information that can be expressed only by means of a σ-algebra, not by means of a partition. For example, suppose that Ω is a unit segment $[0, 1]$ and that Σ is the smallest σ-algebra (i.e., the intersection of all σ-algebras) containing all singletons in Ω. Then, Σ consists only of all *countable* subsets of Ω and their complements. Therefore, a function over Ω measurable with respect to Σ is a function which is constant except at countably many points. In order for *all* of such functions to be measurable with respect to a partition of Ω, the partition needs to be the finest partition consisting of all singletons over the unit segment, $\Pi = \{P_i \,|\, P_i = \{i\}, i \in [0, 1]\}$. However, any arbitrary function over $[0, 1]$ is indeed measurable with respect to this partition. Hence there exists no partition that can express information represented by Σ.

In fact, if Ω is uncountable, there is no partition usable to express information represented by a Borel σ-algebra. For practical purposes, therefore, an information field is more commonly used than an information partition when Ω is uncountable.

Genericity

When a measure-metric space is defined by (Ω, Σ, μ, d), a set $Q \in \Sigma$ is a **generic set** if Q is an **open subset** (a subset which *excludes* its **boundaries** and hence consists only of **interior points**) of Ω, and $\mu[\Omega \setminus Q] = 0$. The complement $\Omega \setminus Q$, which is a closed subset of Ω, is then called a **non-generic set**. Note that an arbitrary subset of Ω that is not a generic set is not referred to as a "non-generic" set.

Likewise, when (Ω, Σ, μ, d) is a probability-metric space, where μ is a probability measure, an event $Q \in \Sigma$ is a **generic event** if Q is an open subset of Ω, and $\mu[Q] = 1$. The complement, which has the probability $\mu[\Omega \setminus Q] = 0$, is a **non-generic event**. A generic event is a stricter requirement than an **almost sure** event, the latter referring simply to an event with probability one.

1.2. Stochastic variable

In a probability space (Ω, Σ, μ) where Ω is the set of states, Σ is a σ-algebra defined over Ω, and μ is a probability measure defined over Σ, any function $X[\omega]$ of the state $\omega \in \Omega$ is called a **stochastic variable**.

Probability function

A **probability function** is defined as

$$\text{Prob}\,\{X[\omega] \in X^*\} \;=\; \mu\,[\{\omega \,|\, X[\omega] \in X^*\}]$$

where X^* is a set such that $\{\omega \,|\, X[\omega] \in X^*\} \in \Sigma\,.$

Cumulative distribution function

When $X[\omega] \in I\!R^n$ is a stochastic variable, its **cumulative distribution** $F[x]$ is defined as

$$F[x] \;=\; \text{Prob}\,\{X[\omega] \leq x\} \;=\; \mu\,[\{\omega \,|\, X[\omega] \leq x\}]$$

where the vector inequality $X[\omega] \leq x$ indicates element-by-element inequality. Obviously, a cumulative distribution function (often abbreviated as **c.d.f.**) is a special form of a probability function.

Probability density function

The **probability density** of a stochastic variable $X[\omega] \in I\!R^n$ is defined as

$$f[x] \;=\; \lim_{\Delta x \downarrow 0} \frac{\text{Prob}\,\{x \leq X[\omega] \ll x + (\Delta x\,, \cdots, \Delta x)\}}{(\Delta x)^n}$$

where the weak vector inequality $\leq$ is used for element-by-element weak inequality, and the strict vector inequality $\ll$ for element-by-element strict inequality. A probability density function (abbreviated as **p.d.f.**) exists only if the c.d.f. is continuous.

In particular, when $X[\omega] \in I\!R\,,$

$$\frac{dF[x]}{dx} \;=\; f[x]\,, \qquad\qquad \int_{z=-\infty}^{x} f[z]dz \;=\; F[x]\,.$$

Stochastic dominance

A stochastic variable $X_1[\omega] \in I\!R$ **first-order stochastic dominates** another stochastic variable $X_2[\omega] \in I\!R$ if their c.d.f.'s $F_1[x]$ and $F_2[x]$ satisfy

$$F_1[x] \leq F_2[x] \quad \text{ for all } \quad x \in I\!R\,.$$

Likewise, $X_1[\omega]$ **second-order stochastic dominates** $X_2[\omega]$ if

$$\int_{z=-\infty}^{x} F_1[z]\,dz \leq \int_{z=-\infty}^{x} F_2[z]\,dz \quad \text{ for all } \quad x \in I\!R\,.$$

Higher-order stochastic dominance can be defined similarly.

Note that lower-order stochastic dominance automatically implies higher-order stochastic dominance, but the converse does not hold. For instance, first-order stochastic dominance is a stricter requirement than second-order stochastic dominance.

Theorem i: If and only if $X_1[\omega]$ first-order stochastic dominates $X_2[\omega]$,

$$E\left[t\left[X_1[\omega]\right]\right] \;\geq\; E\left[t\left[X_2[\omega]\right]\right] \qquad \text{for } \textit{any increasing} \text{ function } t[x].$$

Proof: Using *integration by parts*,

$$\begin{aligned} E\left[t[X_1[\omega]]\right] - E\left[t[X_2[\omega]]\right] &= \int_{x=-\infty}^{\infty} t[x]\, dF_1[x] - \int_{x=-\infty}^{\infty} t[x]\, dF_2[x] \;= \\ &= \int_{x=-\infty}^{\infty} t[x]\, d(F_1[x] - F_2[x]) \;= \\ &= \Big[t[x]\,(F_1[x] - F_2[x])\Big]_{x=-\infty}^{x=\infty} - \int_{x=-\infty}^{\infty} (F_1[x] - F_2[x])\, dt[x] \end{aligned}$$

where $F_1[-\infty] - F_2[-\infty] = F_1[\infty] - F_2[\infty] = 0$ for any well-defined distributions. Hence

$$E\left[t[X_1[\omega]]\right] - E\left[t[X_2[\omega]]\right] \;=\; -\int_{x=-\infty}^{\infty} (F_1[x] - F_2[x])\, dt[x].$$

For this to be positive for *any* increasing utility $t[x]$, it is necessary and sufficient that $F_1[x] - F_2[x] \leq 0$ for all x.

Theorem ii: If and only if $X_1[\omega]$ second-order stochastic dominates $X_2[\omega]$,

$$E\left[u\left[X_1[\omega]\right]\right] \;\geq\; E\left[u\left[X_2[\omega]\right]\right] \qquad \text{for } \textit{any increasing concave} \text{ function } u[x].$$

Proof is analogous to Theorem i (see the exercise example).

In economic terms, if the distribution of a risky income X_1 first-order stochastic dominates the distribution of another risky income X_2, then a decision maker with *any increasing utility function* (a utility function is assumed to be increasing by default) must prefer X_1 to X_2. If X_1 second-order stochastic dominates X_2, then any *risk averse* decision maker must prefer X_1 to X_2.

Example i : Any risky income is second-order stochastic dominated, but not first-order stochastic dominated, by its expected amount as a riskless income.

Example ii : Country Humpkin Co. Ltd., whose share price will be
£0 with probability 20%, £50 with probability 60%, £100 with probability 20%,
first-order stochastic dominates Country Dumpkin Co. Ltd. whose share price will be
£0 with probability 40%, £50 with probability 40%, £90 with probability 10%,
£100 with probability 10%.

Warning : Second-order stochastic dominance is *not* about second-order *moments* (variances).
For instance, the share of Country Bumpkin Co. Ltd.:
£1 with probability 50% , £99 with probability 50% ,
has a higher variance than, but is not second-order stochastic dominated by, the share of Country Humpkin Co. Ltd. in above Example ii.

Hedging

Hedging, or more explicitly **risk hedging**, refers to an investment decision which is designed so as to reduce the risk in terms of second-order stochastic dominance, without changing the expected return from investment.

Example : Shares of Country Bumpkin Co. Ltd. and Country Pumpkin Co. Ltd. have independent and identical probability distributions of prices
£1 with probability 50% , £99 with probability 50% .

If an investor buys two shares of Country Bumpkin Co. Ltd., the total value will be
£2 with probability 50% , £198 with probability 50% .
Obviously, the same distribution arises when the investor buys two shares of Country Pumpkin Co. Ltd.

On the other hand, if the investor buys one share of Country Bumpkin Co. Ltd. and one share of Country Pumpkin Co. Ltd., then the total value will be
£2 with probability 25% , £100 with probability 50% , £198 with probability 25% .

The latter distribution second-order stochastic dominates the former.

Exercise example

Complete the proof of Theorem ii.
(Hint : You may assume differentiability of $u[x]$, with $u'[x] \geq 0$ and $u''[x] \leq 0$ for all x .)

1.3. Conditional probability, prior, and posterior

Given a probability space (Ω, Σ, μ) , the probability of an event $\Omega_1 \in \Sigma$ **conditional upon** another event $\Omega_2 \in \Sigma$ is defined, whenever $\mu[\Omega_2] > 0$, as

$$\widetilde{\mu}[\Omega_1 | \Omega_2] \;=\; \frac{\mu[\Omega_1 \cap \Omega_2]}{\mu[\Omega_2]} .$$

The function $\widetilde{\mu}[\cdot|\Omega_2]$ is the **conditional probability measure** given Ω_2. The conditional probability is also called the **posterior probability** or simply the **posterior**. On the other hand, the original probability measure μ is called the **unconditional probability**, the **prior probability**, or simply the **prior**.

Based upon a stochastic variable $X[\omega]$ where $\omega \in \Omega$ and sets X^*, X^{**} such that $\{\omega \,|\, X[\omega] \in X^*\} \in \Sigma$, $\{\omega \,|\, X[\omega] \in X^{**}\} \in \Sigma$, a **conditional probability function** is defined as

$$\text{Prob}\,\{X[\omega] \in X^* \,|\, X[\omega] \in X^{**}\} \;=\; \frac{\text{Prob}\,\{X[\omega] \in X^* \cap X^{**}\}}{\text{Prob}\,\{X[\omega] \in X^{**}\}}$$

whenever $\text{Prob}\{X[\omega] \in X^{**}\} > 0$. Similarly, based upon two stochastic variables $X[\omega]$, $Y[\omega]$ and two sets X^*, Y^* such that $\{\omega \,|\, X[\omega] \in X^*\} \in \Sigma$, $\{\omega \,|\, Y[\omega] \in Y^*\} \in \Sigma$, the conditional probability can be written as

$$\text{Prob}\,\{X[\omega] \in X^* | Y[\omega] \in Y^*\} \;=\; \frac{\text{Prob}\,\{(X[\omega], Y[\omega]) \in X^* \times Y^*\}}{\text{Prob}\,\{Y[\omega] \in Y^*\}}$$

These two stochastic variables are **independent** if and only if

$$\text{Prob}\,\{X[\omega] \in X^* | Y[\omega] \in Y^*\} \;=\; \text{Prob}\,\{X[\omega] \in X^*\}$$

for all Y^* such that $\text{Prob}\{Y[\omega] \in Y^*\} > 0$.

Note: Conditional probabilities can be defined even when the unconditional probability cannot be defined. For instance, a stochastic variable $X[\omega]$ distributed uniformly over the entire real line $\mathbb{R}$ cannot be defined by means of an unconditional probability function, but can be defined by the conditional c.d.f.

$$\widetilde{F}\,[x \,|\, l \le X[\omega] < r] = \left\{ \begin{array}{ll} 0 & x < l \\ \dfrac{r-l}{x-l} & l \le x < r \\ 1 & x \ge r \end{array} \right\} \qquad -\infty < l < r < \infty\,.$$

Bayes' theorem

When $X[\omega]$ and $Y[\omega]$ are stochastic variables,

$$\text{Prob}\,\{X[\omega] \in X^* \,|\, Y[\omega] \in Y^*\} \;=\; \frac{\text{Prob}\,\{X[\omega] \in X^*\}\;\text{Prob}\,\{Y[\omega] \in Y^* \,|\, X[\omega] \in X^*\}}{\text{Prob}\,\{Y[\omega] \in Y^*\}}$$

for any sets X^*, Y^* such that $\text{Prob}\{X[\omega] \in X^*\} > 0$, $\text{Prob}\{Y[\omega] \in Y^*\} > 0$.

Likelihood

So far, the conditional probability $\widetilde{\mu}[\Omega_1|\Omega_2]$ of an event Ω_1 given another event Ω_2 has been considered as a function of Ω_1, and thereby the conditional probability measure $\widetilde{\mu}[\cdot|\Omega_2]$ has been defined. The same function $\widetilde{\mu}[\Omega_1|\Omega_2]$ is called the **likelihood** of Ω_2 upon the occurrence of Ω_1, when reinterpreted as a function of Ω_2. Note that the likelihood function $\widetilde{\mu}[\Omega_1|\cdot]$ is generally not a probability measure.

1.4. Practice problem

1.4.1. Is each of the following statements true, or false?

1.4.1.i through 1.4.1.iii pertain to a stochastic variable X distributed according to a uniform distribution between k and $k+1$, where k is an unknown real number. Based upon one and only one observation $X = 4.03$,

1.4.1.i. the interval $[3.055, 4.005]$ gives a 95% **confidence set** in estimating the unknown parameter k.

1.4.1.ii. there are many different 95% confidence sets for k.

1.4.1.iii. the only **unbiased test** is to reject any **null hypothesis** that k is either lower than 3.055 or higher than 4.005, whilst accepting any null that k is between 3.055 and 4.005.

In 1.4.1.iv through 1.4.1.vi, a coin is repeatedly tossed until it lands with the head on top. *A priori* the coin may or many not be fair, in that the probability of heads is an unknown parameter h or, in other words, the probability of tails is $1 - h$.

1.4.1.iv. The **point estimator** that $h = 1$ if the head materialises upon the first toss and that $h = 0$ otherwise, is an **unbiased estimator** for h.

1.4.1.v. Any point estimator for h other than the one is described in 1.4.1.iv, is biased.

1.4.1.vi. The estimator described in 1.4.1.iv is a **maximum likelihood estimator**.

1.4.2. Choose one from among ⟨a⟩ through ⟨e⟩ to complete each of the following statements.

1.4.2.i. Two dice C and D are tested for fairness. Die C has been rolled 60 times, out of which the deuce ("2") materialised 11 times. Die D has been rolled 30000 times, out of which the deuce occurred 5500 times.

⟨a⟩ Die D has been tossed many more times, thus likely to be far fairer, than die C.

⟨b⟩ The two dice have the same empirical frequency of the deuce, hence they are equally likely to be (un)fair.

⟨c⟩ Die D is far less likely to be fair than die C.

⟨d⟩ Both dice should be regarded fair.

⟨e⟩ Both dice should be regarded unfair.

1.4.2.ii. To research socioeconomic gender disparity among our University of Tokyo degree holders, we interview 100 recent graduates. Unfortunately for our purpose, even our relatively new graduates are still disproportionately male dominated, scarcely 20% of whom are women. To maximise statistical accuracy, we should randomly select :

⟨a⟩ 20 women and 80 men, proportional to their respective population sizes.

⟨b⟩ 50 from among female graduates and 50 from among male graduates.

⟨c⟩ 80 women and 20 men, in order to compensate for their unequal population sizes.

⟨d⟩ slightly more women than men (e.g., 52 women and 48 men), reflecting slightly higher mortality rates among males than among females.

⟨e⟩ any 100 graduates irrespective of gender, to maximise randomness.

1.4.3. Consider a possibly unfair coin which, when tossed, lands with its head on top with probability h. The prior distribution of h is uniform between 0 and 1.

1.4.3.i. Let $f[h]$ denote the prior probability *density* of h. Express $f[h]$ as a function of h.
(Hint : $f[h]$ needs to be identified only for the ranges $h < 0$, $0 < h < 1$ and $h > 1$. It need not be specified when $h = 0$ and when $h = 1$.)

1.4.3.ii. Let $F[h]$ denote the prior *cumulative distribution* of h. Express $F[h]$ as a function of h.
(Hint : Unlike previously, $F[h]$ can be identified for the whole range $-\infty < h < \infty$.)

Now, the coin has been tossed once, and has landed with its head on top.

1.4.3.iii. Let $\ell[h|H]$ denote the *likelihood*, i.e., the prior probability that the coin lands with its head on top *given that the coin is of the type* h. Express $\ell[h|H]$ as a function of h (where $0 < h \leq 1$).

1.4.3.iv. Let $f[h|H]$ denote the *posterior density*, that is, $f[h|H] = \dfrac{\ell[h|H]f[h]}{\displaystyle\int_{h=0}^{1} \ell[h|H]f[h]\,dh}$. Using 1.4.3.i and 1.4.3.iii, express $f[h|H]$ as a function of h (where $0 < h < 1$).

1.4.3.v. Compute the conditional (= posterior) expectation $E[h|H] = \displaystyle\int_{h=0}^{1} hf[h|H]\,dh$.

2. Statistics and econometrics

2.0. Data

Generally, **data** refer to a set of **observations**. A **census** is a set of all available observations, whilst a **sample** is a subset of potentially available observations, called the **population**. The number of observation is called the **size** of the data (or of the population).

When the population is large, sampling not only saves time and costs to observe the whole population, but also suffers little loss of accuracy in that it depends more on the size of the sample than the size of the population how accurately the sample represents the population. For this reason, a sample is usually taken by selecting randomly a predetermined number of observations out of the population in question.

Statistical theory regards data as **stochastic** (or **random**) **variables**. The underlying mechanism through which the data is created is called the **data generating process**. Essentially, any empirical statistical analysis is a task to reconstruct the data generating process by observing the data. This task is sometimes referred to as **statistical inference**, or **estimation**.

2.1. Statistics

A **statistic** refers to a function of the data. For example, the **moments**, the **quantiles**, and the **mode** of a sample are statistics. On the other hand, those of a probability distribution are **functionals** of the distribution function.

Moments

Generally, the k-th order moment of a sample $\{X_i\}_{i=1}^n$ is defined as $\frac{1}{n}\sum_{i=1}^{n}(X_i)^k$. More specifically, this is called the moment **around zero**. In particular, the first order moment is known as the **average** or also as the **mean**. Analogously, the moment of a distribution $F[x]$ is defined as $\int_{x=-\infty}^{\infty} x^k dF[x]$.

For $k \neq 1$, the moment is ofttimes defined **around the mean** as $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$ for the sample, where $\overline{X}$ is the sample mean, and also as $\int_{x=-\infty}^{\infty}(x-\mu)^k dF[x]$ for the

distribution, where μ is the mean of the distribution. The most frequently used moment around the mean is the **variance**, which is the second order moment. The square root of the variance is referred to as the **standard deviation**. The sign of the third order moment indicates the asymmetry of the distribution, whilst the fourth order moment indicates the comparative thickness of tails.

Mean : Sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, true (structural) mean $\mu = \int_{x=-\infty}^{\infty} x\, dF[x]$.

Variance : Sample variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ (maximum likelihood) or $\overline{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ (unbiased), true variance $\sigma^2 = \int_{x=-\infty}^{\infty}(x-\mu)^2 dF[x]$.

Standard deviation : Sample standard deviation $s = \sqrt{s^2}$ or $\overline{s} = \sqrt{\overline{s}^2}$, true standard deviation $\sigma^2 = \sqrt{\sigma^2}$.

Skewness : Sample skewness $= \frac{1}{ns^3}\sum_{i=1}^{n}(X_i - \overline{X})^3$, true skewness $\frac{1}{\sigma^3}\int_{x=-\infty}^{\infty}(x-\mu)^3 dF[x]$.

Kurtosis : Sample kurtosis $\frac{1}{ns^4}\sum_{i=1}^{n}(X_i - \overline{X})^4$, true kurtosis $\frac{1}{\sigma^4}\int_{x=-\infty}^{\infty}(x-\mu)^4 dF[x]$.

Exercise examples

2.1.M.1. A group of a dozen law professors had their BMI calculated. A half dozen of them scored 20, whilst the remaining half dozen scored 26. The mean, the variance, the skewness, and the kurtosis of their scores are...

2.1.M.2. A group of a dozen economics professors had their BMI calculated. Two of them scored 18, whilst the other ten scored 24. The mean, the variance, the skewness, and the kurtosis of their scores are...

2.1.M.3. The standard uniform distribution, that is the uniform distribution between 0 and 1, has a mean, a variance, a skewness, and a kurtosis respectively of...

2.1.M.4. The standard normal distribution, that is the normal distribution with a zero mean and a unit variance, has a skewness and a kurtosis respectively of...

2.1.M.5. The standard exponential distribution, that is the exponential distribution with a unit mean and a unit variance, has a skewness and a kurtosis respectively of...

Table of the standard normal cumulative $\int_{y=-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{y^2}{2}\right] dy$

z	•.0	•.1	•.2	•.3	•.4	•.5	•.6	•.7	•.8	•.9
0.•	.5000	.5398	.5793	.6179	.6554	.6915	.7257	.7580	.7881	.8159
1.•	.8413	.8643	.8849	.9032	.9192	.9332	.9452	.9554	.9641	.9713
2.•	.9772	.9821	.9861	.9893	.9918	.9938	.9953	.9965	.9974	.9981
3.•	.9986	.9990	.9993	.9995	.9997	.9998	.9998	.9999	.9999	1.000

Quantiles

For a sample or for a distribution, a **quantile** is conceptualised as the inverse function of the **cumulative distribution**. Namely, the q-quantile marks where the fraction q of the sample, or of the distribution, lies below.

Median : $F^{-1}[1/2]$.

Quartiles : the lower (bottom) quartile $F^{-1}[1/4]$, the upper (top) quartile $F^{-1}[3/4]$.

Quintiles : $F^{-1}[1/5], \cdots, F^{-1}[4/5]$.

Deciles : $F^{-1}[1/10], \cdots, F^{-1}[9/10]$.

Percentiles : $F^{-1}[1/100], \cdots, F^{-1}[99/100]$.

Exercise examples

2.1.Q.1. The upper and the lower quartiles of the samples in 2.1.M.1 and in 2.1.M.2 respectively, are...

2.1.Q.2. The x-quantile of the standard uniform distribution is...

2.1.Q.3. The 0.5-, 1-, 2.5-, 5-, 50- (median), 95-, 97.5-, 99-, and 99.5-percentiles of the standard normal distribution, are...

2.1.Q.4. The median and the 63.21-percentile of the standard exponential distribution are...

Mode

The mode is where the most frequent observations congregate, defined formally as $\underset{x}{\operatorname{argmax}} f[x]$ where $f[x]$ is the probability mass function if the variable in question is discrete, or the probability density function if the variable is continuous.

The sample mode can be defined similarly if the variable is discrete, whilst it cannot be precisely defined when the variable is continuous, as no two observations can exactly coincide with positive probabilities. Realistically, therefore, sample observations are categorised into intervals so as to identify the mode in approximation.

Exercise examples

2.1.D.1. The mode of each of those data and distributions in 2.1.M.1 through 2.1.M.5 is...

2.1.D.2. Among the aforesaid five data and distributions, which ones are unimodal and which ones are bimodal?

2.2. Estimation

Obviously, as the size of the data grows, the data generating process is revealed with asymptotic (i.e., as the data size tends to infinity) certainty. However, by any finite data, the generating process can be reconstructed only probabilistically, i.e., with stochastic errors which are built in the data. The reconstruct is called the **estimate** and is a function of the data. This function is called the **estimator**.

For most practical purposes, we are interested only in some **parameters** of the data generating process, rather than in identifying its every detail. Thus we **parametrise** the data generating process by defining those parameters, which are **functionals** of the data generating process, that are useful for specific purposes. Then a **statistic**, that is a function of the data, is sought so as to serve as an estimator for each of the parameters in question. Such a procedure is called **parametric estimation**.

For any parametric estimator, if a statistic determines the estimate uniquely, such a statistic is said to be a **sufficient statistic**. The estimator itself is an obvious sufficient statistic; so is the entirety of the data. Among these sufficient statistics, one with the fewest dimensions is called a **minimum sufficient statistic**.

Point estimation and set estimation : An estimator for the data generating process or for its specific parameter, can take either a single value, or a set of values. The former is called a **point estimator**, and its value is the **point estimate**. The latter gives the **confidence set**, in which the conditional probability that the true data generating process or its parameter lies, given the observed data, is equal to the pre-determined **confidence level**. The higher the confidence level, the larger

the confidence set. A confidence set is also known as a **confidence interval** when estimating a unidimensional parameter.

Point estimation and statistical inference

Unbiased estimation

A point estimator for a parameter is said to be **unbiased** if, for any feasible value of the true parameter, the expectation of the point estimate is equal to the true parameter value.

Namely, given any feasible value of the true parameter θ, if the expectation of the estimator $\widehat{\theta}$ for the parameter coincides with its true value, i.e., $E[\widehat{\theta}|\theta] = \theta$, such an estimator is unbiased.

Exercise examples

2.2.UB.1. A coin has been tossed 20 times, out of which are 11 occurrences of the head and 9 occurrences of the tail. The unbiased estimate for the true probability of the head is...

2.2.UB.2. Four newspapers have independently conducted public opinion polls, according to which 32%, 26%, 29% and 33% of the respondents expressed their support to the conservative party. Assuming that these four polls are *a priori* equally reliable, the unbiased estimate for the true fraction of voters who would vote for the Tories is...

2.2.UB.3. The long-distance bus service from city X to city Y operates daily, and takes five hours under normal driving conditions, which is how it ran on 477 out of the recent 500 days. It has taken six hours on 13 days, seven hours on 4 days, eight hours thrice, eleven hours once, and on the remaining two days it failed to reach the destination, city Y, due to an accident and a mechanical engine failure en route. The unbiased expected journey time of this bus service tomorrow is...

2.2.UB.4. A sample of eight students have been randomly selected and their body weights are recorded as 136, 168, 162, 155, 180, 191, 144, and 132 pounds respectively. The estimated mean and variance of the body weights of all students in this school are...

Pitfall in unbiased estimation : Unbiased estimation is also known as **rational expectation** in economics, forming part of foundation in economic theory. It is not,

however, almighty. When used in statistical inference, depending upon the context it may produce a totally nonsensical estimator.

Example : A possibly unfair coin, which lands with its head on top with probability h, is tossed repeatedly until it indeed lands with its head on top. Then, the number of tosses N is recorded. Construct an *unbiased estimator* of h as a function of N.

Solution: The definition of an unbiased estimator is

$$E[\widehat{h}] = h \qquad \text{for any } h \in (0, 1] .$$

Note first that this relation should hold when $h = 1$, in which case $N = 1$ with probability 1, hence $\widehat{h}$ should be 1 if $N = 1$. Given this, however, $\widehat{h}$ must be set to 0 for any $N \geq 2$ to be unbiased, as $\text{Prob}\{N = 1\} = h$ for any h. Hence the only unbiased estimator is

$$\widehat{h} = \begin{cases} 1 & \text{if} \quad N = 1 , \\ 0 & \text{if} \quad N \geq 2 . \end{cases}$$

Note: The **maximum likelihood estimator** $\widehat{h} = \dfrac{1}{N}$ seems more sensible in this case, although it is biased.

The pitfall here reflects the "inflexible" definition of an unbiased estimator, that is, it must be *a priori* unbiased *given any true value of the parameter in question.*

Likelihood

The probability that a statistic, such as an estimator $\widehat{\theta}$, takes a certain value *conditional upon the true value of the parameter* θ, can be viewed either as a function of the statistic, or as a function of the parameter value. The former is the *conditional probability* $\Pr[\widehat{\theta}|\theta]$, whilst the latter is the *likelihood* $L[\theta|\widehat{\theta}]$.

Maximum likelihood estimation : An estimator is a **maximum likelihood estimator** if the likelihood that the data generating process, or its parameter, is equal to the estimate is higher than that to any other estimate.

Namely, given any feasible value of the estimator $\widehat{\theta}$, if the likelihood $L[\theta|\widehat{\theta}]$ is maximised at $\theta = \widehat{\theta}$, then such an estimator is *maximum likelihood.*

Exercise examples

2.2.ML.1. A coin has been tossed 20 times, out of which are 11 occurrences of the head and 9 occurrences of the tail. The maximum likelihood estimate for the true probability of the head is...

2.2.ML.2. In a typical Japanese provincial jurisdiction, a driver's licence must be renewed once every five years, and the renewal must be done between a month before and a month after the licence holder's birthday. Prof. Dr. Jelly Bean renewed his licence on the 6th of January in 1993, 1998 and 2008, and on the 6th of March in 2003. The maximum likelihood estimate for his birthday is...

2.2.ML.3. The long-distance bus service from city X to city Y operates daily, and takes five hours under normal driving conditions, which is how it ran on 477 out of the recent 500 days. It has taken six hours on 13 days, seven hours on 4 days, eight hours thrice, eleven hours once, and on the remaining two days it failed to reach the destination, city Y, due to an accident and a mechanical engine failure en route. The maximum likelihood estimate for the journey time on this bus service tomorrow is...

2.2.ML.4. A sample of eight students have been randomly selected and their body weights are recorded as 136, 168, 162, 155, 180, 191, 144, and 132 pounds respectively. The estimated mean and variance of the body weights of all students in this school are...

Consistent estimation

An estimator is said to be **consistent** if it asymptotically **converges** (**in probability**) to the true data generating process, or its parameter, as the data size grows. *Maximum likelihood estimation is consistent.*

Set estimation and confidence levels

Confidence set : A set of parameter values where the likelihood, given the observed value of the relevant statistic, is higher than a certain threshold (*confidence level*).

Exercise examples

2.2.S.1. Statistically, 12.69% of the University of Tokyo graduates eventually end their lives by suicide. Assuming that the standard error of this estimation is 0.44%, the 90%-, the 95%-, and the 99%-confidence intervals for the true suicide rate among the University of Tokyo graduates are respectively...

2.2.S.2. A sample of eight students have been randomly selected and their body weights are recorded as 136, 168, 162, 155, 180, 191, 144, and 132 pounds respectively. The 90%- and the 95%-confidence intervals for the true average body weights of all students in this school are...

2.2.S.3. A coin has been tossed 20 times, out of which are 11 occurrences of the head and 9 occurrences of the tail. The 95%-confidence set for the true probability of the head is...

2.2.S.4. The alcohol content of a certain brand of wine is known to follow a distribution that is approximately normal with mean 13.4% and standard deviation 0.3%. When the local authority inspects the wine, they open randomly selected bottles to measure their alcohol contents, and if more than 1% of inspected bottles fall below the alcohol content indicated on the label, the winery is to be fined for false labelling. To avoid the fine, the alcohol content on the label should not exceed...

Confidence level : The threshold likelihood above which the parameter value is included in the *confidence set*.

Exercise examples

2.2.L.1. Statistically, 12.69% of the University of Tokyo graduates eventually end their lives by suicide. Assuming that the standard error of this estimation is 0.44%, the confidence level that the true suicide rate lies between $12.69\% - 0.44\% = 12.25\%$ and $12.69\% + 0.44\% = 13.13\%$ is...

2.2.L.2. Trains call at a certain station with a constant interval of one hour. From the viewpoint of a passenger randomly arriving at this station, the confidence level that the last train left more than 20 minutes ago is...

2.2.L.3. Trains call at a certain station randomly with an average frequency of one per hour. From the viewpoint of a passenger randomly arriving at this station, the confidence level that the last train left more than 20 minutes ago is...

2.3. Hypothesis testing

A **hypothesis** is a tentative statement about the data generating process. Typically, it involves a statement concerning a certain parameter. Hypothesis testing is a task to assess how likely or unlikely the hypothesis is to be true, given the data. The hypothesis being tested is called the **null hypothesis**.

Conventionally, hypothesis testing is either to **reject** the null hypothesis, or to **accept** it, given a pre-determined **significance level**. That is, the hypothesis is accepted if and only if it lies within the confidence set with the confidence level equal to the complement of the significance level (i.e., 1− the significance level). The higher the significance level, the easier it is to reject the null hypothesis.

Faulty rejection, that is to reject the null hypothesis when it is true, is called the **type I error**, whilst **faulty acceptance**, to accept the hypothesis when it is false, is called the **type II error**. The significance level is interpreted as the maximum admissible probability of the type I error. Obviously, there is an inevitable trade-off between these two types of errors: the lower the significance level, the less probable the type I error whilst the more probable the type II error.

Simple hypotheses and composite hypotheses : The hypothesis to be tested can be either simple, that is to specify one value for the data generating process or its parameter, or composite, specifying the set in which the true process or parameter to lie. A typical example of simple hypothesis testing is a **two-sided test**, where the null hypothesis is that the parameter is equal to a certain value, which is rejected if either the estimate based upon the observed data is either too far above or too far below the hypothesised value. A common example of composite hypothesis testing is a **one-sided test**, where the null hypothesis claims that the parameter is greater than (or less than) a certain threshold.

Unbiased testing : If a test is constructed such that the null is rejected if and only if its likelihood, given the observed data, is below a certain level, such a test is said to be **unbiased**.

Nested and non-nested testing : The most standard hypothesis testing is either to accept or reject the null. This is called **nested testing**. **Non-nested testing** is to compare multiple hypotheses to determine which one is more likely to be true than the other(s).

Nested hypothesis testing and significance levels

Null hypothesis : The hypothesis that the parameter takes a certain benchmark value, or that two data sets are drawn from structurally the same distribution.

Alternative hypothesis : The hypothesis that the parameter differs from a certain benchmark value, or that two data sets are drawn from structurally distinct distributions.

Type-I error : To reject the null when it is true. The probability of the type-I error is referred to as the *size* of the test.

Type-II error : To fail to reject the null when it is false. The probability of the type-II error varies depending upon the true value of the parameter, but it does not exceed the complement of the test size.

Significance level : Also known as the *size* of the test.

p-value : The minimum test size that the null is rejected.

Exercise examples

2.3.1. Japanese chess players A and B played 100 matches in the past, out of which B won 63 games. The null hypothesis that A and B are equally competent, is...

2.3.2. A random number generator produces i.i.d. (independently and identically distributed) numbers according to the uniform distribution between 0 and an unknown parameter θ. Four numbers produced by this generator have been 0.800, 0.223, 0.614, and 0.009. The null hypothesis that $\theta = 2$ is...

2.3.3. Among those eight students sampled in 2.2.S.2, the middle four, weighing 162, 155, 180, and 191 pounds respectively, are seniors whilst the other four, weighing 136, 168, 144, and 132, are freshmen. The null hypothesis that studying in this school does not make students fatter, is...

2.3.4. A referendum for the constitutional amendment has collected between 46 million and 48 million votes, among which 10000 have been counted, with 4649 pros and 5351 cons. To announce the predicted rejection of the said amendment, the p-value is...

2.3.5. A student has scored 120 in an IQ test which is known to have a standard error of 8 marks. If the null is that the student has an average IQ of 100, the p-value is...

Unbiased test

Unbiasedness of a statistical test is somewhat distinct from that of an estimator. If the **set estimator** $\widetilde{\Theta}[x;\omega]$ (which gives a **confidence interval** or **confidence set**) for a parameter θ satisfies

$$\inf_{\theta\in\widetilde{\Theta}[x;\omega]} \text{Prob}\,\{X[\omega] = x \,|\, \theta\} \;\geq\; \sup_{\theta\notin\widetilde{\Theta}[x;\omega]} \text{Prob}\,\{X[\omega] = x \,|\, \theta\}$$

where $X[\omega]$ is the data used in this test and x is its realisation, then such a test is called an **unbiased test**. When the test is parametric and the distribution of data is symmetric

around the parameter to be estimated (e.g., estimating the mean of a normal distribution), the acceptance set becomes symmetric around the singleton null hypothesis.

Example : When n i.i.d. observations drawn from a normal distribution with mean μ and variance σ^2 have the arithmetic average m, the 95% confidence interval (set estimator) for μ given σ^2 is taken as approximately

$$m - \frac{1.956}{\sqrt{n}}\,\sigma \leq \mu \leq m + \frac{1.956}{\sqrt{n}}\,\sigma$$

when n is sufficiently large. The test which accepts a null hypothesis $\mu = \mu_0$ when and only when μ_0 falls within this confidence interval, is an unbiased test. On the other hand, a "test" which randomly accepts any null hypothesis with probability 95% and rejects it with probability 5%, has the same **size** (5%) but is not an unbiased test.

2.4. Dependence

Many data consist of observations in more than one variable. For many practical purposes, it is important to determine whether, and how, these variables depend upon one another. In abstract general terms, variables are **dependent** if the data generating process is not **separable** between them.

Correlation

When two variables are **linearly dependent**, they are said to be **correlated**. If one variable tends to take high values when the other is high (resp., low), they are said to be **positively** (resp., **negatively**) correlated. If one is a linear function of the other, they are said to be **perfectly correlated**, with the **correlation coefficient** being equal to either 1 (perfect positive correlation) or -1 (perfect negative correlation). Note that when dependency between the two variables is nonlinear, there may be little correlation even when they are strongly dependent.

Covariance : $\sigma_{XY} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ (sample), or

$$\mathrm{Cov}[X,Y] = \int_{y=-\infty}^{\infty}\int_{x=-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)\,dF_X[x,y]\,dF_Y[x,y] \text{ (structural).}$$

Note : $\mathrm{Cov}[X,Y] = \mathrm{Cov}[Y,X]$ identically holds.

(Co)variance matrix : $\mathrm{Var}[X,Y] = \begin{bmatrix} \mathrm{Var}[X] & \mathrm{Cov}[X,Y] \\ \mathrm{Cov}[Y,X] & \mathrm{Var}[Y] \end{bmatrix}$.

Determination coefficient : $R^2[X,Y] = \dfrac{(\mathrm{Cov}[X,Y])^2}{\mathrm{Var}[X]\ \mathrm{Var}[Y]}$.

Correlation : $\mathrm{Corr}[X,Y] = \sqrt{R^2[X,Y]} = \dfrac{\mathrm{Cov}[X,Y]}{\mathrm{Std}[X]\ \mathrm{Std}[Y]}$.

Exercise examples

2.4.C.1. Consider all points on the unit circle $x^2 + y^2 = 1$. Then

the means $\mu_X = \dfrac{1}{2\pi}\displaystyle\int_{\theta=0}^{2\pi} \cos\theta\, d\theta = ...$, $\quad \mu_Y = \dfrac{1}{2\pi}\displaystyle\int_{\theta=0}^{2\pi} \sin\theta\, d\theta = ...$

the variances $\mathrm{Var}[X] = \dfrac{1}{2\pi}\displaystyle\int_{\theta=0}^{2\pi} (\cos\theta - \mu_X)^2 d\theta = ...$, $\mathrm{Var}[Y] = \dfrac{1}{2\pi}\displaystyle\int_{\theta=0}^{2\pi} (\sin\theta - \mu_Y)^2 d\theta = ...$

and the covariance $\mathrm{Cov}[X,Y] = \dfrac{1}{2\pi}\displaystyle\int_{\theta=0}^{2\pi} (\cos\theta - \mu_X)(\sin\theta - \mu_Y)\, d\theta = ...$

Hence the determination coefficient and the correlation coefficient between the x- and the y-coordinates of these points obtain

$R^2[X,Y] = \dfrac{(\mathrm{Cov}[X,Y])^2}{\mathrm{Var}[X]\ \mathrm{Var}[Y]} = ...$, $\quad \mathrm{Corr}[X,Y] = \dfrac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\ \mathrm{Var}[Y]}} = ...$

2.4.C.2. Consider uniformly distributed points over the triangle $0 \le y \le x \le 1$, where

the means $\mu_X = 2\displaystyle\int_{x=0}^{1}\int_{y=0}^{x} x\, dy\, dx = ...$, $\quad \mu_Y = 2\displaystyle\int_{x=0}^{1}\int_{y=0}^{x} y\, dy\, dx = ...$

the variances $\mathrm{Var}[X] = 2\displaystyle\int_{x=0}^{1}\int_{y=0}^{x} (x - \mu_X)^2 dy\, dx = ...$,

$\mathrm{Var}[Y] = 2\displaystyle\int_{x=0}^{1}\int_{y=0}^{x} (y - \mu_Y)^2 dy\, dx = ...$

and the covariance $\mathrm{Cov}[X,Y] = 2\displaystyle\int_{x=0}^{1}\int_{y=0}^{x} (x - \mu_X)(y - \mu_Y)\, dy\, dx = ...$

The correlation coefficient between the x- and the y-coordinates of these points thereby obtains $\mathrm{Corr}[X,Y] = \dfrac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\ \mathrm{Var}[Y]}} = ...$

2.4.C.3. The population densities and the GDP per capita of the G7 countries have been reported as follows:

countries	US	Canada	UK	France	Germany	Italy	Japan
km^{-2}	31	3	246	110	232	193	339
US\$ 10^3	45	43	46	40	40	36	34

the means $\overline{X} = \dfrac{31 + 3 + \cdots + 339}{7} = ...$, $\quad \overline{Y} = \dfrac{45 + 43 + \cdots + 34}{7} = ...$

the variances $\mathrm{Var}[X] = \dfrac{(31 - \overline{X})^2 + (3 - \overline{X})^2 + \cdots + (339 - \overline{X})^2}{7} = ...$,

$$\text{Var}[Y] = \frac{(45-\overline{Y})^2 + (43-\overline{Y})^2 + \cdots + (34-\overline{Y})^2}{7} = \ldots$$

and the covariance $\text{Cov}[X,Y] = \dfrac{(31-\overline{X})(45-\overline{Y}) + \cdots + (339-\overline{X})(34-\overline{Y})}{7} = \ldots$

The correlation coefficient between the population densities and the income per capita is $\text{Corr}[X,Y] = \dfrac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]\ \text{Var}[Y]}} = \ldots$

(Note: These moments are computed without weighting the countries according to their areas and population sizes. For instance, the mean population density $\overline{X}$ differs from the average population density of all these seven countries combined.)

2.4.C.4. The population densities and the GDP per capita of the eight regions in Japan have been reported as follows:

regions	Hokkaido	Tohoku	Kanto	Chubu	Kinki	Chugoku	Shikoku	Kyushu
km^{-2}	67	151	1298	359	692	240	215	355
JPY 10^3	3395	3518	4567	4295	3959	3940	3397	3275

The correlation coefficient between the population densities and the income per capita is...

Regression

When, ofttimes through interpretative details external to the quantitative data, it is suspected that one of the variables is unilaterally dependent upon the other(s), a model that re-expresses the **dependent variable** as a function of the other variable(s) is hypothesised. Such a function is called the **regression equation**, and the variables serving as arguments of the function are called **independent variables**, **explanatory variables**, or **regressors**. Regression equations can be either linear or nonlinear.

Independent variables : Also referred to as *explanatory variables* or *regressors*, on the right-hand side of the *regression equation*.

Dependent variable : on the left-hand side of the regression equation.

Residual : That component of the variation in the dependent variable which remained unexplained by the regressors. Also known as the *error term*.

Least square : Those estimated regression coefficients minimising the *sum of squared residuals*.

Linear OLS (*ordinary least square*) : The regression equation is $Y = X\beta + \varepsilon$, where the estimated coefficients are $\widehat{\beta} = (X'X)^{-1}X'Y$.

Exercise examples

2.4.R.1. When regressing y on x in 2.4.C.1, the estimated coefficient obtains $\left(\int_{\theta=0}^{2\pi} \cos^2\theta\, d\theta\right)^{-1} \int_{\theta=0}^{2\pi} \cos\theta \sin\theta\, d\theta = ...$

2.4.R.2. When regressing y on x and the constant in 2.4.C.1, the estimated coefficient and intercept simultaneously obtain

$$\begin{pmatrix} \int_{\theta=0}^{2\pi} \cos^2\theta\, d\theta & \int_{\theta=0}^{2\pi} \cos\theta\, d\theta \\ \int_{\theta=0}^{2\pi} \cos\theta\, d\theta & \int_{\theta=0}^{2\pi} d\theta \end{pmatrix}^{-1} \begin{pmatrix} \int_{\theta=0}^{2\pi} \cos\theta \sin\theta\, d\theta \\ \int_{\theta=0}^{2\pi} \sin\theta\, d\theta \end{pmatrix} = ...$$

2.4.R.3. When regressing x on y in 2.4.C.1, the estimated coefficient obtains $\left(\int_{\theta=0}^{2\pi} \sin^2\theta\, d\theta\right)^{-1} \int_{\theta=0}^{2\pi} \sin\theta \cos\theta\, d\theta = ...$

2.4.R.4. When regressing y on x in 2.4.C.2, the estimated coefficient obtains $\left(\int_{x=0}^{1}\int_{y=0}^{x} x^2 dy\, dx\right)^{-1} \int_{x=0}^{1}\int_{y=0}^{x} xy\, dy\, dx = ...$

2.4.R.5. When regressing y on x and the constant in 2.4.C.2, the estimated coefficient and intercept simultaneously obtain

$$\begin{pmatrix} \int_{x=0}^{1}\int_{y=0}^{x} x^2 dy\, dx & \int_{x=0}^{1}\int_{y=0}^{x} x\, dy\, dx \\ \int_{x=0}^{1}\int_{y=0}^{x} x\, dy\, dx & \int_{x=0}^{1}\int_{y=0}^{x} dy\, dx \end{pmatrix}^{-1} \begin{pmatrix} \int_{x=0}^{1}\int_{y=0}^{x} xy\, dy\, dx \\ \int_{x=0}^{1}\int_{y=0}^{x} y\, dy\, dx \end{pmatrix} = ...$$

2.4.R.6. When regressing x on y in 2.4.C.2, the estimated coefficient obtains $\left(\int_{x=0}^{1}\int_{y=0}^{x} y^2 dy\, dx\right)^{-1} \int_{x=0}^{1}\int_{y=0}^{x} yx\, dy\, dx = ...$

2.4.R.7. When regressing the GDP per capita on the population density in 2.4.C.3, the estimated coefficient obtains $\dfrac{31 \times 45 + 3 \times 43 + \cdots + 339 \times 34}{31^2 + 3^2 + \cdots + 339^2} = ...$

2.4.R.8. When regressing the GDP per capita on the population density and the constant in 2.4.C.3, the estimated coefficient and intercept obtain

$$\begin{pmatrix} 31^2 + \cdots + 339^2 & 31 + \cdots + 339 \\ 31 + \cdots + 339 & 1 + \cdots + 1 \end{pmatrix}^{-1} \begin{pmatrix} 31 \times 45 + \cdots + 339 \times 34 \\ 45 + \cdots + 34 \end{pmatrix} = ...$$

Elasticity : In economic statistics, it is a vastly common practice to take logarithms of each variable and then to regress one on others. When the regression equation itself is linear, it is said to be **loglinear**, and the **regression coefficients** represent **elasticities**, i.e., the proportional increment in the dependent variable associated with those in the independent variables.

2.5. Prior and posterior

The **prior belief**, or simply the **prior**, is a hypothetical belief about the data generating process before observing any data. It can be either subjective to the statistician, or formed through reasoning external to the data. The **posterior** (**belief**) is the belief **updated** through the observed data.

Bayesian statistics is the school which emphasises the role of the prior, as opposed to **frequentist** statistics.

2.6. Common difficulties in statistical analysis

- **Selection bias :** The sample may not be collected truly randomly from the population. In particular, when values within a certain range are not observed, the data is said to be **truncated**. When the occurrences of those values within a certain range are observed but not their raw values, the data is **censored**.
- **Model selection :** Selection of regressors and the form of regression equation. In particular, if regressors are too interdependent, their coefficients become indeterminate, the situation called (**multi**) **collinearity**.
- **Endogeneity :** Dependency may be mutual not unilateral, in which case the regression analysis may not be the optimal statistical model. When this is suspected, the regressor needs to be replaced by a truly independent variable, called an **instrumental variable**.
- **Error distribution :** The **error terms**, the discrepancy between the observed values of the independent variable and their theoretical values derived from the regression equation, may follow a probability distribution that differs from what the model assumes. For example, they may be **serially dependent**, or **heteroscedastic**, in which cases the regression analysis may require technical adjustments.

2.7. Practice problems

2.7.1. Is each of the following statements true, or false ?

2.7.1.i. The correlation coefficient in 2.4.C.1 is (strictly) positive.

2.7.1.ii. The correlation coefficient in 2.4.C.2 is (strictly) positive.

2.7.1.iii. The correlation coefficient in 2.4.C.4 is (strictly) positive.

2.7.1.iv. When regressing the GDP per capita on the population density in 2.4.C.4, the coefficient has the opposite sign from when regressing the population density on the GDP per capita instead.

2.7.2. Choose one from among ⟨a⟩ through ⟨e⟩ to complete each of the following statements.

2.7.2.i. The regression coefficients differ between
⟨a⟩ 2.4.R.1 and 2.4.R.2.
⟨b⟩ 2.4.R.1 and 2.4.R.3.
⟨c⟩ 2.4.R.4 and 2.4.R.5.
⟨d⟩ 2.4.R.4 and 2.4.R.6.
⟨e⟩ 2.4.R.7 and 2.4.R.8.

2.7.2.ii. The signs of the correlation coefficient in 2.4.C.1 and of the regression coefficient in 2.4.R.2 are
⟨a⟩ both zero, as the two variables X and Y are mutually independent.
⟨b⟩ both positive, as the two variables X and Y are mutually dependent.
⟨c⟩ both zero, even though the two variables X and Y are mutually dependent.
⟨d⟩ both positive, even though two variables X and Y are mutually independent.
⟨e⟩ different, as the two variables X and Y are uncorrelated but mutually dependent.

2.7.2.iii. The signs of the correlation coefficient in 2.4.C.3 and of the regression coefficient in 2.4.R.7 are
⟨a⟩ the same, as always the case when based upon the same data set.
⟨b⟩ the same, coincidentally in this example.
⟨c⟩ opposite, as always the case when based upon the same data set.
⟨d⟩ coincidentally opposite, as income and demography are mutually independent.
⟨e⟩ coincidentally opposite, as 2.4.R.7 lacks the constant as a regressor.

Consider a **binary** variable which takes either 1 with probability ρ, or 0 with probability $1-\rho$.

2.7.2.iv. The *median* is ⟨a⟩ ρ ⟨b⟩ $1/2$ ⟨c⟩ $1-\rho$
⟨d⟩ 0 if $\rho < 1/2$, 1 if $\rho > 1/2$ ⟨e⟩ anywhere between 0 and 1

2.7.2.v. The *variance* is ⟨a⟩ ρ ⟨b⟩ $1/2$ ⟨c⟩ $1/4$ ⟨d⟩ $(1-\rho)\rho$ ⟨e⟩ $\sqrt{(1-\rho)\rho}$

2.7.3. Which of the following seven is the approximate fraction of those students who score above 55 in a test wherein the average score and the standard deviation are respectively known to be 50 and 10?

15.87% 30.85% 45.00% 50.00% 55.00% 69.15% 84.13%

2.7.4. Statistically, the fraction of smokers amongst the University of Tokyo students is estimated to be 9.95%, with the standard error 0.05%. Which of the following ten approximates the 95% confidence interval?

$[0.025\%, 0.975\%]$ $[0.05\%, 0.95\%]$ $[0.05\%, 9.95\%]$ $[4.95\%, 14.95\%]$ $[4.95\%, 13.95\%]$
$[5.00\%, 9.95\%]$ $[5.00\%, 95.00\%]$ $[9.85\%, 10.05\%]$ $[9.90\%, 10.00\%]$ $[9.95\%, 95.00\%]$

2.7.5. In a typical distribution of income, the mean is (Choose one from the following seven):

⟨a⟩ above the median. ⟨d⟩ the same as the median. ⟨c⟩ below the median.
⟨b⟩ the same as the mode. ⟨e⟩ below the mode. ⟨f⟩ undefined.
⟨g⟩ between the median and the mean, whichever is higher.

2.7.6. 200 students took a test. 100 of them scored 78 marks whilst the other 100 scored 62 marks. What is the **variance** of this grade distribution?

□ 8 □ 16 □ 49 □ 49.64 □ 64 □ 70 □ 140 □ 256 □ 1600 □ 9800 □ 9928
□ 20 if scores are out of 100 marks. □ 400 if scores are out of 100 marks.
□ 464 if scores are out of 100 marks. □ 964 if scores are out of 100 marks.

2.7.7. A variable X is distributed symmetrically (such as the normal distribution). The distribution of X^2 is (Choose one):

□ also symmetrical, with a larger mean than that of X.
□ also symmetrical, with a larger variance than that of X.
□ also symmetrical, with a larger mean and a larger variance than that of X.
□ generally asymmetrical, its upper tail being larger than its lower tail.
□ generally asymmetrical, its lower tail being larger than its upper tail.
□ generally asymmetrical, though the direction of asymmetry depends upon the distribution.

3. Information and decisions

3.1. Bayesian decision theory

Tree

In a tree, **nodes** are interconnected by branches. There is only one **initial node**, from which the tree originates. Every node has a *unique* path from the initial node. If the path from the initial node N_0 to a node N_a passes another node N_b, then N_a is called a **successor** of N_b, and N_b is a **predecessor** of N_a. The initial node is a predecessor of any other node in the tree. A terminal node, on the other hand, is a node where the tree ends. Obviously, a terminal node has no successors.

Decision tree : A tree in which every terminal node carries an outcome (typically with well-defined payoffs) whereas every non-terminal node is assigned either to the decision maker or to nature. The former indicates active decision making, whilst the latter indicates randomisation by nature.

Information set : In a decision tree, if the decision maker does not know whether she/he is at one node or another, these two nodes belong to the same *information set*. A **perfect information** tree is a tree in which all information sets are **singletons**. If a tree contains a multiple-node information set, it is an **imperfect information** tree.

Action : Each branch in a decision tree represents an *action*. In other words, an action is a contingent move *given that the information set* (from which the branch stems) *has been reached*.

Strategy : A complete contingent plan. Note that a strategy is one of the two alternative ways to materialise a **decision rule**, a function which maps from the tree to the complete contingent plan.

Strategy space : A decision maker's strategy space is the set of all **mixed strategies** available to her/him. Note that a strategy space is a *convex set*. More precisely, if there are k pure strategies available to the decision maker, her/his strategy space is $(k-1)$-dimensional **simplex**. Note also that, as long as there are only finitely many pure strategies available to the decision maker, her/his strategy space is also a *compact set*.

History : Each node is associated with its unique path from the initial node of the tree. This path is referred to as a history.

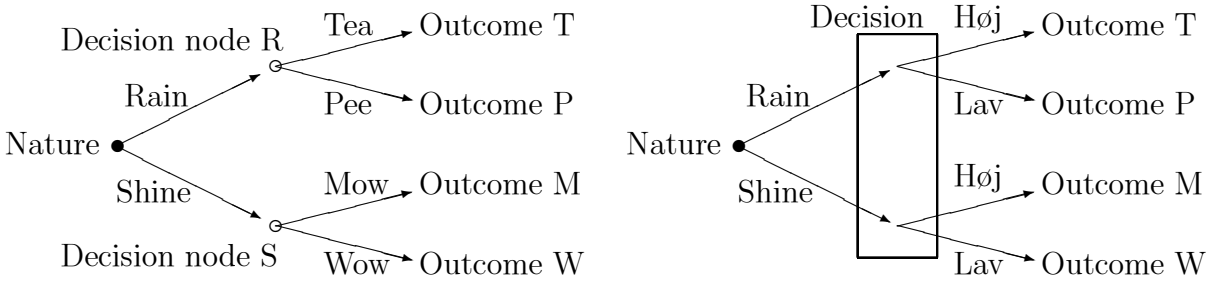
Behaviour : A behaviour, or a **behavioural strategy**, is an alternative way to materialise a decision rule. A decision maker's contingent plan is represented by a function of all information sets assigned to the decision maker. If this function always maps from an information set to a pure action, then such a behaviour is a **pure behaviour**.

Note that the (mixed) **behaviour space** is smaller than the (mixed) strategy space in the sense that more than one strategies may correspond to one behaviour. For, a mixed behaviour cannot incorporate any stochastic dependence between the randomisation of actions at one information set and that at another information set.

Subtree : In a tree, for any node N which is not a terminal node, the set $S[N]$ consisting of the node itself and all of its successors can be uniquely defined. Also its complement $\bar{S}[N]$, the set consisting of all other nodes in the tree, can be uniquely defined. If the tree does not involve any information set that includes nodes from both of these two sets, then $S[N]$ alone can be reinterpreted as a tree. This is called the *subtree* commencing from N.

Note that, when N is the initial node of a tree, the whole tree is a subtree commencing from N. Any subtree that is strictly smaller than the whole tree is called a **proper subtree**.

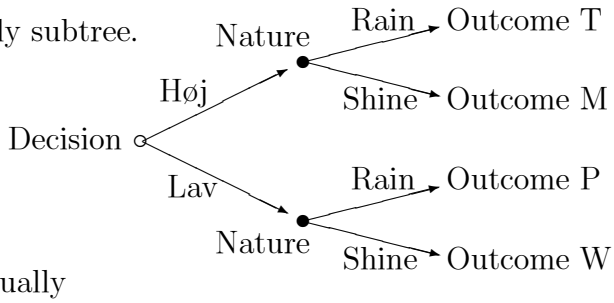
Example : In the left decision tree, the behaviour space is $[0,1]^2$ (a unit square) to which any behaviour (t, m) belong, where the probability of "Tea" is t and that of "Pee" is $1-t$ if it Rains, the probability of "Mow" is m and that of "Wow" is $1-m$ if it Shines. The strategy space, on the other hand, is a three-dimensional **simplex** (regular tetrahedron) $\left\{(v,x,y,z) \in I\!R^4_+ \,|\, v+x+y+z=1\right\}$, where v, x, y and z respectively denotes the probabilities of "Tea, Mow", "Tea, Wow", "Pee, Mow" and "Pee, Wow".



In this decision tree, the **state of nature**, either "Rain" or "Shine", reveals after the **chance move** (the random move by Nature), so that the decision maker's **information partition** is {{Rain}, {Shine}} (or his/her **information field** is {{ }, {Rain}, {Shine}, {Rain, Shine}}). That is, he/she knows whether it Rains or

Shines by the time he/she makes a move. The part of the tree commencing from Decision node R forms a *proper subtree*, so does the part commencing from Decision node S.

Were the decision maker not to know the state of nature before making a move, then the tree should be drawn as the right tree, where the two decision nodes belong to the same *information set*. In the right decision tree, there are only two *pure actions* ("Høj" and "Lav") available, whereby the strategy space and the behaviour space coincide. In this tree, there is no proper subtree, and hence the entirety of the tree is its only subtree.



Depending upon the structure of the problem, it is occasionally possible to redraw the tree so as to avoid multi-node information sets and thereby to simplify visually all the decisions and actions involved in the tree. For instance, the right tree in the above example can be redrawn in this way without altering the structure of the problem.

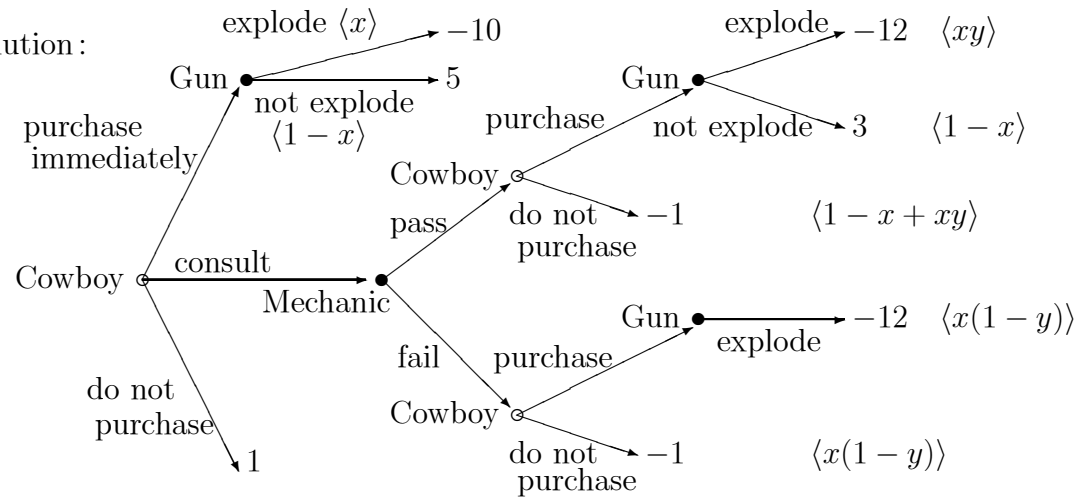
Example : A risk-seeking Western cowboy is wondering whether to buy an antique handgun or not. However, the gun can explode with prior probability $x > 0$. If the cowboy does not buy the gun, his von Neumann Morgenstern utility is 1. If he buys the gun, if it explodes then his von Neumann Morgenstern utility is -10, otherwise if it does not explode his von Neumann Morgenstern utility is 5. Before deciding whether to buy the gun or not, the cowboy can consult a mechanic for a professional advice on the danger of explosion, although the macho cowboy's von Neumann Morgenstern utility decreases by 2 if he consults the mechanic, reflecting his reluctance to admit his inability to inspect the gun by himself. If the gun is not explosive it will certainly pass the mechanical test, whilst even if the gun is explosive it can still pass the mechanical test with conditional probability $y > 0$.

Question i : What is the probability of the gun being explosive *conditional upon* passing the mechanical test ?

Solution : The gun can pass the test in two mutually exclusive events. One is that it is explosive and yet passes the test, which has the probability xy. The other event is that it is not explosive, of which the probability is $1 - x$. Thus, by Bayes' rule, the probability in question is $\dfrac{xy}{1 - x + xy}$.

Question ii : Draw a decision tree describing the decision problem of the cowboy.

Solution :



In the tree, inside angled brackets ⟨ ⟩ are *unconditional* (prior) *probabilities*. Note that the random draw deciding whether the gun explodes occurs at the end of the tree, in order to avoid multi-node information sets. Note also that, in decision-theoretic terms, "Mechanic" and "Gun" are *nature moves*. Namely, the mechanic's making a mistake or not isn't his/her active decision.

Question iii : Identify the range of probabilities x and y where it is optimal for the cowboy to consult the mechanic and then to decide whether to buy the gun or not, and graph the range.

Solution : If the gun fails the test, by assumption it is explosive with certainty. Hence, only the following three strategies can be optimal.

- If the cowboy never buys the gun, his (expected) utility is

$$U_N \;\equiv\; 1.$$

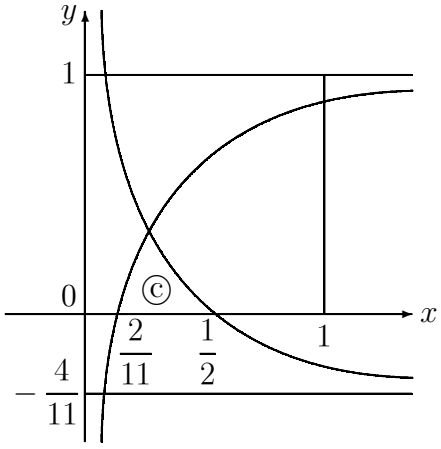
- If the cowboy buys the gun without consulting the mechanic, his expected utility is

$$U_B \;\equiv\; -10x + 5(1-x) \;=\; 5 - 15x\,.$$

- If the cowboy consults the mechanic and then buys the gun if and only if it passes the test, then the cowboy's expected utility is

$$U_C \;\equiv\; x(1-y) - 10xy + 5(1-x) - 2 \;=\; -4x - 11xy + 3.$$

The consultation strategy is optimal if and only if $U_C \geq \max\{U_N, U_B\}$, namely $y \leq \min\left\{-\frac{4}{11} + \frac{2}{11x}\,,\, 1 - \frac{2}{11x}\right\}$, which is the area below both the hyperbola $y = -\frac{4}{11} + \frac{2}{11x}$ and also the hyperbola $y = 1 - \frac{2}{11x}$ (the area ⓒ in the diagram).



Question iv : How would the answer to Question iii change if the cowboy were risk-neutral? What if he were risk-averse?

Solution: The cowboy's risk attitude, whether risk-seeking, risk-neutral, or risk-averse, is *already incorporated in his utility levels*, as his utility is defined as von Neumann Morgenstern utility. Therefore, changing the assumption about his risk attitude would make no difference in the answer to Question iii. (This is a *trick question*!!)

Belief

In a decision tree, whenever a decision maker is at a multi-node information set, he/she must have a *subjective probability distribution over the nodes in the information set*. This distribution embodies his/her **belief** concerning at which node the present status is.

Belief system : A decision maker must have a belief in every information set assigned to him/her. The complete collection of a decision maker's beliefs exhausting all of his/her information sets in the tree is his/her **belief system**.

Assessment : In every information set, the decision maker assigned to the information set must [i] hold a belief, and [ii] choose an action. The pair consisting of the belief and the action is sometimes called an assessment.

Consistency of a belief system

If a decision maker's belief system conforms to Bayes' rule, then it is said to be **Bayesian consistent**, or **Bayesian rational** (or simply "**Bayesian**", "**consistent**", or "**rational**", all the same meaning).

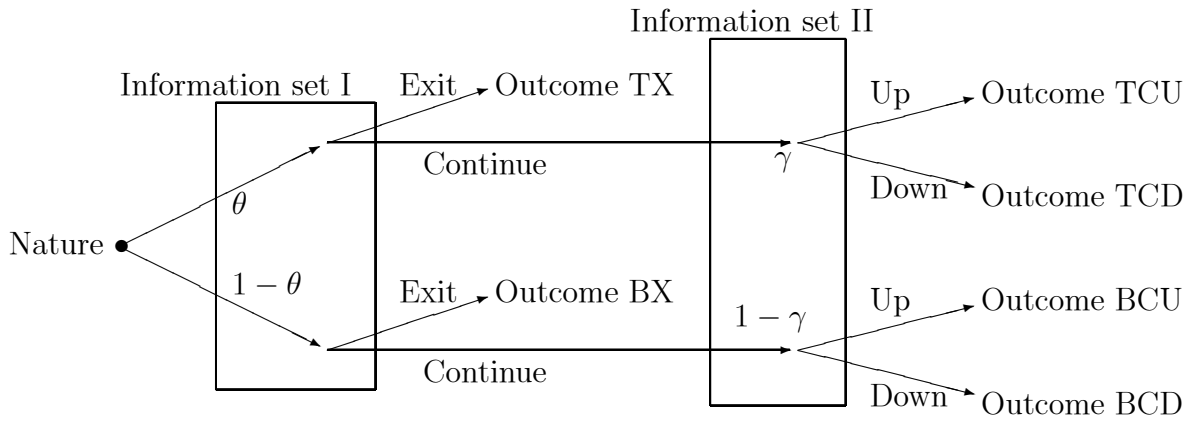
Sequential rationality : Bayes' rule is inapplicable to an event of which the prior probability is nil. Any arbitrary probability distribution *conditional upon a zero-probability event* is consistent with Bayes' rule. Thereby Bayesian consistency provides no regulation on beliefs **off the equilibrium path**, i.e., in those information sets not reached with strictly positive probabilities.

Sequential rationality, or **sequential consistency**, is a stricter requirement than Bayesian consistency. It regulates beliefs off the equilibrium path as follows.

A belief system β is **sequentially rational**, or **sequentially consistent**, if there is a *sequence* of decision rules and beliefs $\{\alpha_\nu\,,\beta_\nu\}_{\nu=1}^{\infty}$ converging to $\alpha\,,\ \beta$ (in distribution), in which each decision-belief pair $\alpha_\nu\,,\ \beta_\nu$ enables *every information*

set to be reached with a strictly positive probability and also the belief system β_ν is Bayesian consistent.

Example : In the decision tree below, if the decision maker chooses "Continue" with any strictly positive probability in Information set I, then for his/her belief system to be *Bayesian consistent* he/she must believe $\gamma = \theta$.



If the decision maker chooses "Exit" with certainty, then the prior probability that Information set II is reached is zero. Hence, any belief γ is *Bayesian consistent*. However, to be *sequentially consistent*, the decision maker still has to adhere to $\gamma = \theta$.

3.2. Induction

There are generally two alternative methods of decision making, known as the **normal form** and the **extensive form**.

Normal form decision : A *complete contingent plan* is made *in advance*. Once a part of the plan has been executed, the decision maker is to adhere to the remainder of the plan invariably. In other words, the normal form is a framework in which any dynamic decision is reduced into a one-time-for-all, instantaneous decision.

In decision theory and game theory, such a complete contingent plan is referred to as a **strategy**.

Extensive form decision : A series of temporary decisions are made one after another as time passes. At each given time, the decision maker reassesses the present situation, called the **history**, and makes a decision for the *immediate next step*.

In decision theory and game theory, each of the temporary decisions made along an extensive form is referred to as an **action**. The whole series of actions, constituting

the entirety of the extensive decision, is called a **behaviour**, which is regarded as a function (a **decision rule** to be more specific) that maps from the set of all feasible histories to the set of admissible actions.

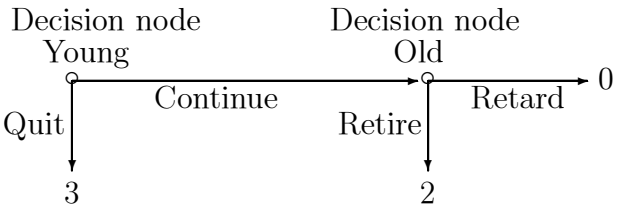
In general, whenever the normal form is applicable, it tends to have the advantage of being simpler to compute than the extensive form. On the other hand, the extensive form is more versatile, useful even to those problems to which the normal form is difficult to apply.

Backward induction

Backward induction is a solution concept making use of the one-way flow of time. If a decision maker has a plan such that, at any interim point in its execution, the remainder of the plan is optimal (i.e., payoff maximal) for the decision maker, then the entire plan is also optimal. Note that the converse does not automatically follow.

For a decision rule to be **backward induction proof**, or **inductively optimal**, the decision must be optimal *in every subtree*. Note that this is a stricter requirement than the optimality (payoff maximality) of the decision. That is, there are decision rules which are optimal yet not inductively optimal.

Example : In the decision tree below, obviously, any decision rule that assigns probability 1 to "Quit" is optimal. Namely, any *strategy* which assigns the probability mass of 1 between "Quit, Retire" and "Quit, Retard" whilst zero probabilities to "Continue, Retire" and "Continue, Retard", is optimal. Also, any *behaviour* which assigns probabilities 1 to "Quit" and 0 to "Continue", and any arbitrary probabilities between "Retard" and "Retire", is optimal.



However, assigning any strictly positive probability to "Retard" is suboptimal *within the subtree* commencing from the Decision node Old. This confines the inductively optimal *strategy* to "Quit, Retire" with probability one (pure strategy). Likewise, the only backward induction proof *behaviour* is to "Quit" with probability 1, then to "Retire" with probability 1.

As in this example, outcomes in a decision tree are often represented by their respective von Neumann Morgenstern utility from the decision maker's point of view, unless specified otherwise.

Mathematical induction

Backward induction is in fact closely related to the general technique of **mathematical induction** used in various mathematical proofs. Namely, in order for the lemma $f[n] = g[n]$ to be true for *every positive integer* n,

- the lemma has to hold with $n = 1$, that is, $f[1] = g[1]$;
- if the lemma holds with $n = k$ $(k = 1, 2, 3, \cdots)$, i.e., $f[k] = g[k]$, then the lemma should also hold with $n = k + 1$, i.e., $f[k + 1] = g[k + 1]$.

Example : Prove the formula $\displaystyle\sum_{\ell=1}^{n} \ell^4 = \frac{1}{30} n(n+1)(2n+1)(3n^2+3n-1)$.

Inductive proof: When $n = 1$, the formula trivially holds.

Assuming that the formula holds when $n = k$, that is,

$$\sum_{\ell=1}^{k} \ell^4 = \frac{1}{30} k(k+1)(2k+1)(3k^2+3k-1),$$

then, when $n = k + 1$,

$$\begin{aligned} \sum_{\ell=1}^{k+1} \ell^4 &= \sum_{\ell=1}^{k} \ell^4 + (k+1)^4 = \frac{1}{30} k(k+1)(2k+1)(3k^2+3k-1) + (k+1)^4 = \\ &= \frac{1}{30}(k+1)((k+1)+1)(2(k+1)+1)(3(k+1)^2+3(k+1)-1), \end{aligned}$$

hence the formula also holds with $n = k + 1$. Q.E.D.

Note: An inductive proof is an *extensive form* proof. The normal form counterpart of the above proof is as follows.

Summing the binomial expansion

$$(x+1)^5 = x^5 + 5x^4 + 10x^3 + 10x^2 + 5x + 1$$

over $x = 1, \cdots, n$, we obtain

$$\sum_{x=1}^{n} (x+1)^5 = \sum_{x=1}^{n} x^5 + 5\sum_{x=1}^{n} x^4 + 10\sum_{x=1}^{n} x^3 + 10\sum_{x=1}^{n} x^2 + 5\sum_{x=1}^{n} x + \sum_{x=1}^{n} 1.$$

Subtracting $\displaystyle\sum_{x=2}^{n} x^5$ from either side,

$$(n+1)^5 = 1 + 5\sum_{x=1}^{n} x^4 + 10\sum_{x=1}^{n} x^3 + 10\sum_{x=1}^{n} x^2 + 5\sum_{x=1}^{n} x + \sum_{x=1}^{n} 1. \qquad \text{☻}$$

Hence, substituting the *known* formulae

$$\sum_{x=1}^{n} 1 = n\,, \quad \sum_{x=1}^{n} x = \frac{n(n+1)}{2}\,, \quad \sum_{x=1}^{n} x^2 = \frac{n(n+1)(2n+1)}{6}\,, \quad \sum_{x=1}^{n} x^3 = \left(\frac{n(n+1)}{2}\right)^2$$

into $\overset{\bullet\bullet}{\smile}$,

$$\begin{aligned}\sum_{x=1}^{n} x^4 &= \frac{1}{5}\left((n+1)^5 - 1 - 10\sum_{x=1}^{n} x^3 - 10\sum_{x=1}^{n} x^2 - 5\sum_{x=1}^{n} x - \sum_{x=1}^{n} 1\right) = \\ &= \frac{n(n+1)(2n+1)}{6} \cdot \frac{3n^2+3n-1}{5}\,.\end{aligned}$$

3.3. Signal extraction

Signal extraction is a typical form of statistical inference. When more than one **noisy signal** has been received, the decision maker must weigh these signals to form a reasonably updated belief. It can interpretatively be either Bayesian or non-Bayesian.

Gaussian (normal) signal extraction : When n independent normal signals $q_1\,, \cdots, q_n$ estimate the same (real-valued) quantity, the posterior distribution of the true quantity is also normal with

$$\text{mean} \quad \frac{\sum_{i=1}^{n} \frac{q_1}{{\sigma_i}^2}}{\sum_{i=1}^{n} \frac{1}{{\sigma_i}^2}}\,, \qquad \text{variance} \quad \frac{1}{\sum_{i=1}^{n} \frac{1}{{\sigma_i}^2}}\,,$$

where ${\sigma_i}^2$ is the unconditional **standard error** of the signal q_i. Its inverse $\dfrac{1}{{\sigma_i}^2}$ is called the **precision** of signal q_i. The mean of the posterior, which is the unbiased **point estimator** of the true quantity, is in fact the *precision-weighted average* of signals, and the precision of the posterior is the sum of the precisions of signals.

Example : Suppose it is statistically established that the distribution of *true* IQs is normal with mean 100, standard deviation 20. Consider an IQ test of which the score is normally distributed with mean equal to the true IQ of the person and standard error 10. If a student obtains an IQ score of 140 in this exam, what is the posterior distribution of his/her true IQ? If another student obtains a score of 60, what is his/her expected IQ?

Solution : The posterior of the true IQ of the student who has scored 140 is a normal distribution with

$$\text{mean } \frac{\frac{100}{20^2}+\frac{140}{10^2}}{\frac{1}{20^2}+\frac{1}{10^2}} = 132\,, \qquad \text{standard deviation } \frac{1}{\sqrt{\frac{1}{20^2}+\frac{1}{10^2}}} = 4\sqrt{5} \approx 8.94427191\,.$$

The student who has managed to score 60 is, unfortunately, incapable of comprehending even a relatively straightforward concept such as signal extraction. ☹

Binary signal extraction : When an unobservable outcome can be either O_1 or O_2, and when there are $m+n$ independent signals $b_1\,,\cdots,b_{m+n}$ which indicate

$$b_1 = \cdots = b_m = \{O_1\} \qquad \text{and} \qquad b_{m+1} = \cdots = b_{m+n} = \{O_2\}\,.$$

The probability that the outcome is O_1 is given by

$$\text{Prob}[O_1] \;=\; \frac{\prod\limits_{i=1}^{m} p_i \prod\limits_{i=m+1}^{m+n} (1-p_i)}{\prod\limits_{i=1}^{m} p_i \prod\limits_{i=m+1}^{m+n} (1-p_i) + \prod\limits_{i=1}^{m}(1-p_i) \prod\limits_{i=m+1}^{m+n} p_i}$$

where p_i is the unconditional probability that the signal b_i is correct.

When the opinion of person i $(i = 1,\cdots,m+n)$ is considered as a noisy signal about the (socially) correct choice, and if everyone has the same probability of being correct, with $p_1 = \cdots = p_{m+n} > \frac{1}{2}$, then

$$\text{Prob}[O_1] \geq \frac{1}{2} \quad \text{ if and only if } \quad m \geq n\,.$$

This gives a justification for the **simple majority rule** in **social choice**.

Example : You and your best friend have just finished taking the same examination and are now discussing a particularly difficult problem given in the exam. From the track records, the unconditional probability that your answer is correct is 75% whilst the unconditional probability that your friend's answer is correct is 70%. The discussion has just revealed that, as far as this particular problem is concerned, your friend has written essentially the same answer as yours. Given this information, what is the probability that your answer to this problem has been correct ? Assume that there are only correct and incorrect answers but no partially correct ones.

Solution : $\dfrac{0.75\times 0.70}{0.75\times 0.70 + (1-0.75)(1-0.70)} = \dfrac{7}{8} = 87.5\%\,.$

3.4. Practice problems

3.4.1. After the example in 1.4.3, the same coin has been tossed another 19 times, so as to attain 20 tosses in total, out of which 11 heads and 9 tails have been observed.

Determine whether each of the following statements is true, or false.

3.4.1.i. The *likelihood* has now become $h^{11}(1-h)^9$.

3.4.1.ii. The *posterior density* has now become $h^{11}(1-h)^9$.

3.4.1.iii. The maximum likelihood estimate for h is $11/20$.

3.4.1.iv. The unbiased estimate for h, computed as the conditional (= posterior) expectation $E[h|H^{11}T^9] = \int_{h=0}^{1} h f[h|H^{11}T^9]\, dh$ (where $f[h|H^{11}T^9]$ denotes the posterior density), is $11/20$.

(Hint : The difference here from examples 2.2.ML.1 and 2.2.UB.1, respectively, in chapter 2, section 2.2, is that we now explicitly define the prior. In other words, 1.4.3 provides an easy example illustrating the Bayesian view of statistics and its difference from the **frequentist** view of statistics exemplified in 2.2. The main theme in our practice 3.4.1 is to probe when, and how, these distinct views materialise.)

3.4.2. It is known that, when a variable X has a normal distribution $\mathrm{N}[\mu_X, \sigma_X^2]$ with mean μ_X and variance σ_X^2, and variable Y has another normal distribution $\mathrm{N}[\mu_Y, \sigma_Y^2]$ with mean μ_Y and variance σ_Y^2 which is independent of X, the sum of these two variables $X+Y$ has a normal distribution $\mathrm{N}[\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2]$ with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$. Given the observation $X + Y = 2.718$, the (conditional) expectation of X is

⟨a⟩ μ_X ⟨b⟩ $2.718 - \mu_Y$ ⟨c⟩ $\dfrac{\mu_X}{\sigma_X^2} + \dfrac{2.718 - \mu_Y}{\sigma_Y^2}$

⟨d⟩ $\dfrac{\mu_X\sigma_X^2 + (2.718 - \mu_Y)\sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}$ ⟨e⟩ $\dfrac{\mu_X\sigma_Y^2 + (2.718 - \mu_Y)\sigma_X^2}{\sigma_X^2 + \sigma_Y^2}$

3.4.3. A risk-neutral early Victorian gold digger has the opportunity to buy a mining site near Melbourne. The total sum of the purchase price of the site, transport expenses, and actual digging costs, is 1200 pounds sterling. However, the mining site actually contains gold only with prior probability g. The gold digger assesses the gross revenue from gold to be 6000 pounds sterling (hence the nett profit 4800 pounds sterling) if the site contains gold, whilst no revenue can be raised if the site is devoid of gold. Before deciding whether to purchase the site, the gold digger has the option to consult a geologist

at a fee of f (> 0) pounds sterling. However, the geologist makes a misjudgment (either to misjudge a site full of gold as containing no gold, or *vice versa*) with probability m.

3.4.3.i. What is the conditional probability of the site containing gold when the geologist says so?

3.4.3.ii. Draw a decision tree describing the decision problem of the gold digger.

3.4.3.iii. According to the tree drawn in 3.4.3.ii, how many *pure strategies* are available for the gold digger?

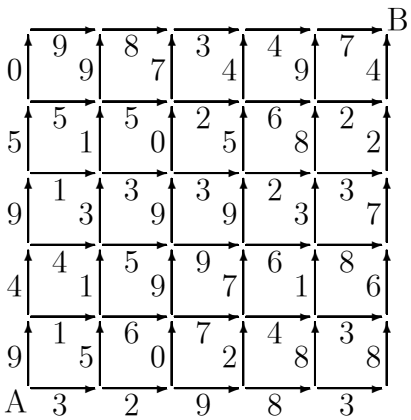
3.4.3.iv. Can it ever be optimal for the gold digger to consult the geologist and then to buy always (or to buy never) regardless of his advice? Is this useful in precluding any of the pure strategies listed in 3.4.3.iii and 3.4.3.iv? If so, which pure strategies are precluded?

3.4.3.v. For each of those pure strategies *not* precluded in 3.4.3.iv, compute the (*ex ante*) expected nett profit for the gold digger.

3.4.3.vi. For what ranges of the probabilities g, m and the fee f is it optimal for the gold digger to consult the geologist and then to decide whether to purchase the site or not?

(Hint: Your answer to 3.4.3.vi changes its algebraic form depending upon whether g is greater or less than 0.2, and also upon whether m is greater or less than 0.5.)

3.4.4. Choose one from among ⟨a⟩ through ⟨e⟩ to complete each of the following statements.



3.4.4.i. In the diagramme, how many distinct routes are available from A to B?

⟨a⟩ ${}_{10}\mathrm{C}_5 = \dfrac{10 \times 9 \times 8 \times 7 \times 6}{5 \times 4 \times 3 \times 2 \times 1} = 252.$

⟨b⟩ ${}_{10}\mathrm{P}_5 = 10 \times 9 \times 8 \times 7 \times 6 = 30240.$

⟨c⟩ ${}_{10}\mathrm{H}_5 = \dfrac{10 \times 11 \times 12 \times 13 \times 14}{1 \times 2 \times 3 \times 4 \times 5} = 2002.$

⟨d⟩ $10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 3628800.$

⟨e⟩ $5! \times 5! = (5 \times 4 \times 3 \times 2 \times 1)^2 = 14400.$

3.4.4.ii. In the diagramme, the number on each arrow indicates the transport cost. How much does the most economical route from A to B cost?

⟨a⟩ 27 ⟨b⟩ 28 ⟨c⟩ 29 ⟨d⟩ 30 ⟨e⟩ 31

The unconditional probabilities that you are correct is 75%, that your friend is correct is 30%. You and your friend disagree upon:

3.4.4.iii. a true-or-false question. Assuming that "True" and "False" have the same *prior* (50% each) to be correct, the *posterior* that you are correct is

3.4.4.iv. a multiple-choice question. Assuming that all five alternatives have the same *prior* (20% each) to be correct, the *posterior* that you are correct is

⟨a⟩ 7/8. ⟨b⟩ 28/39. ⟨c⟩ 21/31. ⟨d⟩ 21/40. ⟨e⟩ 7/15.

4. Rationality

No man-made theory can ever be flawless. Decision theory is no exception. It has been widely recognised that there are a number of shortfalls in decision theory, some inherent and theoretical, some others when applied to various economic problems. Accordingly, there have been quite a few attempts to modify, to deviate from, and/or to test against the traditional framework of decision theory.

4.1. Decision-theoretic modelling

The ground rule: Use decision theory if and only if the following checklist is satisfied. Otherwise, stay open to the possibility that decision theory (at least in its ordinary definition) may not be the most suitable analytical tool.

Checklist :

- *The decision maker must have a well defined objective function to maximise.*

  If the decision maker behaves according to some specific *behavioural assumption*, then his/her payoff function or strategy space should be constructed in a way that justifies his/her specific behavioural code as a consequence, or as a means, of payoff maximisation. Should there be an economic agent whose behaviour cannot be explained by payoff maximisation (i.e., if it is impossible to construct a payoff function and/or a strategy space that is consistent with his/her behavioural code), then such an agent should not be regarded as a "decision maker."

- *The decision maker's payoff function must be exogenous.*

  The decision maker cannot *directly* influence his/her **payoff structure**. His/her payoff structure is ordinarily defined by his/her **payoff function**, a function which maps either from his/her decision rule or from resulting outcomes to payoffs. Namely, the *decision tree* (including the payoffs at its terminal nodes) should be taken as absolutely given. If there are factors that affect the decision maker's payoff function, it should be explicitly built into the function either *as an argument* or *as a parameter*, so as to make sure that the *payoff function need not be altered after part of the decision has been executed.*

- *The structure of the decision problem* (tree) *must be known to the decision maker.*

For example, if payoffs are **state contingent**, the **chance move** (nature's random move to determine the state) should be modelled as part of the problem, and the *probability distribution of the chance move must be* (at least subjectively) *known to the decision maker.*

If *information is incomplete* in that the decision maker does not know his/her exact payoff function, then the *prior distribution* of relevant states (e.g., his/her **type**) must be known.

- *The decision maker cannot choose his/her beliefs.*

  Beliefs can be formed endogenously. However, decision maker cannot *directly choose* his/her beliefs as if he/she chooses his/her actions and strategies. After all, beliefs must be *consistent with actions and strategies*; any strategy that is chosen based upon inconsistent beliefs cannot be deemed as an optimal strategy.

  Also importantly, the decision maker *must have beliefs*. Not having a belief, or refusing to have any belief whatsoever, is not a decision-theoretically admissible option.

4.2. Full rationality, or bounded rationality ?

Full rationality may appear natural and plausible at the very first glance, yet it is in fact a very strong assumption. This assumption technically simplifies theory, but its practical plausibility is much in question.

Full rationality : A model in which decision making procedures are assumed to be perfectly costless and swift, where computational costs, memory constraints, and possibilities of errors are *unquestioned*.

Bounded rationality : A framework which *explicitly models* procedural limitations in decision making. Bounded rationality encompasses a large variety of models, which include (but are not confined to) the following.

Imperfect recall : As part of the decision is executed, the decision maker may *lose* some of the information acquired in an earlier stage. This includes (but again, is not confined to) roughly three subcases.

- Forgetting the decision maker's *own action* taken earlier.

- Forgetting which information sets have been passed through.
- Forgetting whether *the same* information set has been reached once. In this case, the information set contains two nodes that are mutually a predecessor and a successor. This class of imperfect recall is referred to as **absent-mindedness**.

Information processing : Taking computational costs and delays explicitly into the model. One of the typical sub-fields of applied decision theory where this type of "bounded rationality" is heavily studies is *organisational design*.

Artificial intelligence : Taking into account the fact that no human brain, as well as no other kinds of brains, can have unbounded memory. A typical model in this category is the **finite automaton**, which is a decision-making algorithm that recognises only a finite number of states (and/or histories).

Many situations and models that are broadly categorised as "bounded rationality" can indeed be transcribed into a **full rationality** model. Examples include, but are not confined to, the following.

- *Limited computational ability :*

  Perhaps the most typical way to incorporate limitations in the decision maker's computational capacity is to introduce *costs* associated with computation, and subtract them from the decision maker's payoff. Another alternative in transcribing computational incapacity into a full rationality model is to introduce *errors* caused by miscomputation.

- *Careless decision maker :*

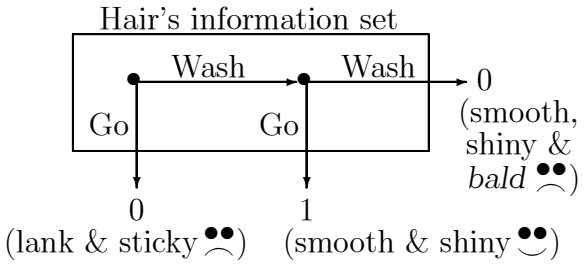
  Probability of mistakes can be modelled by adding *nature's stochastic moves* to determine whether the decision maker makes an error or not. An alternative technique is to modify final payoffs from the decision maker's raw payoffs to his/her *expected* payoffs taking into account those payoff losses caused by mistakes.

- *Ignorance about the payoff structure :*

  This can be modelled as the decision maker's *type*: he/she can either be informed of his/her own payoff function, or not. If he/she is uninformed, then he/she is incapable of making decisions on his/her own actions and strategies. This is modelled as the decision maker's "exiting" the system, receiving a certain exogenously given payoff. If and only if he/she is informed, he/she stays in the system and undertakes a strategy based upon a conscious, active decision.

In other words, the so-called "full rationality" framework is indeed capable of modelling those situations which are seemingly believed as bounded rationality.

However, some other kinds of bounded rationality situations are more intrinsically against the full rationality framework. Perhaps one of the most typical situations of this sort is *imperfect recall*. The difficulty in transcribing an imperfect recall model into a perfect recall model is apparently related to the fact that the former shows a number of distinctive properties which are theoretically proven to be impossible in the latter. These properties include the curious fact that a decision maker with imperfect recall can be *strictly* better off adopting a strictly mixed strategy than adopting any pure strategy. This can be visualised by the following simple deterministic (without Nature's chance moves) example, where the decision maker is "absent minded." The pure strategy Wash entails a zero payoff, and the other pure strategy Go also entails a zero payoff, whilst mixing between Wash with probability w and Go with probability $1 - w$ can raise an expected payoff of $w(1 - w)$. The payoff maximal strategy is to mix between Wash and Go with probability one half each, entailing an expected payoff of one quarter.



4.3. Is decision theory empirically testable ?

It would be careless to prejudge as if *all* decision-theoretic models were empirically not testable. Furthermore, it is one of the most common — and the most harmful — misunderstandings in economics to believe as if every theory should be empirically testable. Examples of useful but not empirically testable theories abound. There are also quite a number of theoretical models and frameworks, including the expected utility theory (*à la* von Neumann and Morgenstern) and the permanent income hypothesis, that have been repeatedly rejected by various econometric tests and yet remain as broadly accepted standards in economic studies.

It is noteworthy, however, that decision theory is not a kind of discipline inherently geared toward empirical testing. This stems largely from the fact that decision theory is about *strategies*, neither *outcomes* nor *actions*. Strategies, by definition, are *complete contingent plans* that encompass not only the **equilibrium paths** (those actions actually taken with strictly positive probabilities) but also those actions that are *off the equilibrium paths*, i.e., those hypothetical actions that are never meant to take place. When it comes to empirical testing, this inevitably entails the problem of observability : strategies are not directly observable.

In a purely static decision problem, there are no off-equilibrium-path actions, and therefore the aforementioned observability problem is relatively less serious. In a dynamic problem, on the other hand, it is often a laborious — although rare to be absolutely impossible — task to infer the decision maker's strategies based upon his/her *observed* actions and outcomes. In a subject where empirical testing is part of the main interest, the practicability of such inference can indeed be a determinant factor in deciding whether to adopt a decision-theoretic model or not.

Once again, it obviously depends upon the specific context whether empirical testability should be regarded as a relevant question. Insofar as the subject remains prescriptive or normative, the question of empirical testability is often not a major concern.

4.4. Practice problem

4.4.1. Determine whether each of the following preferences is "rational" according to our standard framework of decision-theoretic rationality.

4.4.1.i. A nun, who is most devoted into the doctrine of frugality, says the less money she has, the happier she feels.

4.4.1.ii. The strange-looking man next door says he prefers coffee to tea on even-number mornings, but *vice versa* on odd-number mornings.

4.4.1.iii. A beauty contest referee reports that Miss Pimple beats Miss Wrinkle, that Miss Wrinkle beats Miss Freckle, but that Miss Freckle beats Miss Pimple.

4.4.1.iv. The menu in a small local restaurant lists only three items, all priced at £5 per portion. The first item is beef steak, the second is vegetable salad, and the third is "the meal of the day" which can be either beef steak or vegetable salad with probability 0.5 each. Every day, quite a number of customers choose "the meal of the day."

4.4.1.v. A child abuser, when arrested, confessed to the local police that his newborn baby's agonised look was a major source of his satisfaction.

4.4.1.vi. An antique dealer in wealthy Lower Manhattan offers to buy a 50 cent coin featuring late John F. Kennedy for 65 cents.

4.4.2. Contemplate diagnosis D which informs you that you presently face a 20% risk of death by a sudden heart attack within one week, but that an immediate surgical

operation can entirely eradicate the said risk. Diagnosis E, on the other hand, warns you that the emergency care unit wherein you are currently hospitalised can offer little help with your anticipated heart attack which will kill you within one week with probability 70%, except that a high-tech open heart surgery will reduce the probability of your death down to 50%. According to our standard decision theory, your willingness to pay for the surgical treatment recommended by diagnosis D must be

⟨a⟩ higher than that by diagnosis E.

⟨b⟩ the same as that by diagnosis E.

⟨c⟩ lower than that by diagnosis E.

⟨d⟩ higher than that by diagnosis E if you are *risk averse*, *vice versa* if you are *risk seeking*.

⟨e⟩ higher than that by diagnosis E if you are *risk seeking*, *vice versa* if you are *risk averse*.

4.4.3.

Teacher : "Listen, kids ! At least one of you has a stain on your face. Can everyone of you tell if you have it or not ?"

Kids : "No, we can't !"

Teacher : "Listen, kids ! At least one of you has a stain on your face. Can everyone of you tell if you have it or not ?"

Kids : "No, we can't !"

Teacher : "Listen, kids ! At least one of you has a stain on your face. Can everyone of you tell if you have it or not ?"

Kids : "No, we can't !"

Teacher : "Listen, kids ! At least one of you has a stain on your face. Can everyone of you tell if you have it or not ?"

Kids : "No, we can't !"

Teacher : "Listen, kids ! At least one of you has a stain on your face. Can everyone of you tell if you have it or not ?"

Kids : "No, we can't !"

Teacher : "Listen, kids ! At least one of you has a stain on your face. Can everyone of you tell if you have it or not ?"

Kids : "Yes, we can !!!"

Question : How many kids had stained faces ? Explain how you derive your conclusion.

4.4.4. The following excerpt, supposedly from a newspaper article published on 23 September 1929, contains an obvious error. Locate it.

Since the end of the first world war, the world economy has grown unprecedentedly, both in the sheer volume of production and in the depth of international inter-dependence. Does this trend enlarge the role which our national macroeconomic policy is expected to play ? The answer to this question is hardly trivial. It is true that we are now required to be responsible to the entire world, much unlike our Victorian fathers. On the other hand, our very sovereignty begins to be questioned in that we can no longer singlehandedly decide what we do in our homeland as we did yesteryear.