

2 General inference problem

- 2.1 Measurement precision
- 2.2 Statistical models
- 2.3 Inference problem
- 2.4 Goals in statistical inference
- 2.5 Statistical decision theoretic approach to inference

2.1 Measurement precision

We studied different probability models because they can be used to describe a population of interest. Finding such a good model is our final goal. On the way towards this goal, we use the data to identify the parameters that are used to describe the models.

The statistical inference problem arises ^{恰好地}precisely because we do not know the exact value of the parameter in the model description and we use the data to work out a ^{代表}proxy for the parameter.

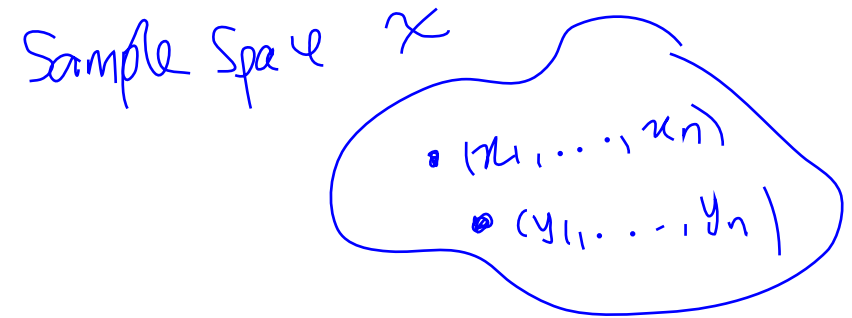
The statistician is ^{面对}confronted with the problem of drawing a conclusion about the population by using the limited information available from the dataset.



The purpose in Statistical Inference is to draw conclusions from data.

The conclusions might be about predicting further outcomes, evaluating risks of events, testing hypotheses, among others.

In all cases, inference about the **population** is to be drawn from limited information contained in the sample.



The most common situation in Statistics:

- an experiment has been performed with independent and identically distributed (i.i.d.) observations;
- the possible results are real numbers that form a vector of observations $\mathbf{x} = (x_1, x_2, \dots, x_n)$;
- the appropriate sample space \mathcal{X} is \mathbb{R}^n ;
- there is typically a “hidden” mechanism that generates the data and we are looking for ways to identify it.

Models will describe this “hidden” mechanism in some simplistic but hopefully useful way.

可靠的

For the model to be more trustworthy, continuous variables, such as time, interval measurements, etc. should be treated as such, where feasible. However, in practice, only discrete events can actually be observed.

测量工具

Thus we record with some *unit of measurement*, Δ , determined by the precision of the measuring instrument. This unit of measurement is always finite in any real situation.

$$P(X \leq \pi) = P(X < \pi)$$

"continuous"

$$X \sim N(0, 1)$$

$$P(X = 7) = 0$$

If ^{以实验为依据的}empirical observations were truly continuous, then, with probability one, no two observed responses would ever be ^{相同的}identical. This fact will sometimes be used in our theoretical derivations.

On the other hand, the *real life empirical observations are discrete*. This fact will be utilized by us to keep some of the proofs simpler. In many cases we will be dealing with the discrete case only, thus avoiding more involved measure-theoretic arguments.

2.2 Statistical models

Having collected the observations we would like to calculate the joint density (in the continuous case):

$$L_X(\mathbf{x}) = f_X(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \dots f_{X_n}(x_n) \quad (1)$$

联合密度

$= \prod_{i=1}^n f_{X_i}(x_i)$

In the discrete case this will be just the product of the probabilities for each of the measurements to be in a suitable interval of length Δ .

If the observations are **independent and identically distributed (i.i.d.)** then all densities in (1) are the same:

$$f_{X_1}(x) = f_{X_2}(x) = \dots = f_{X_n}(x) = f(x).$$

$$L_X(\mathbf{x}) = \prod_{i=1}^n f_X(x_i)$$

This is the most typical situation we will be discussing in our course.



The need of **Statistical Inference** arises since typically, our knowledge about $f_X(x_1, x_2, \dots, x_n)$ is **incomplete**.

Given an inference problem and having collected some data, we construct one or more set of possible **models** which may help us to understand the data generating mechanism.



Basically, statistical models are working assumptions about how the dataset was obtained.



Example 2.9

If our data were counts of accidents within $n = 10$ consecutive weeks on a busy crossroad, it may be reasonable to assume that a Poisson distribution with an unknown parameter λ has given rise to the data. That is, we may assume that we have 10 independent realisations of a $\text{Poisson}(\lambda)$ random variable.

Example 2.10

If, on the other hand, we measured the lengths X_i of 20 female babies at age 4 weeks, it would be reasonable to assume normal distribution for these data. Symbolically we denote this as follows:

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, 2, \dots, 20.$$

The models we use, as seen in the examples above, are usually about the **shape of the density** or of the cumulative distribution function (cdf) of the population from which we have sampled.

These models should represent, as much as possible, the available prior theoretical knowledge about the data generating mechanism.

It should be noted that in most cases, we do not exactly know which population distribution to assume for our model.

Suggesting the set of models to be validated, is a difficult matter and there is always a risk involved in this choice.

设想

The reason is that if the contemplated set of models is “too large”, many of them will be similar and it will be difficult to single out the model that is best supported by the data.

On the other hand, if the contemplated set of models is “too small” there exists the risk that none of them gives an adequate description of the data.

Choosing the most appropriate model usually involves a close collaboration between the statistician and the people who formulated the inference problem.

In general, we can view the statistical model as the triplet $(\mathcal{X}, \mathcal{P}, \Theta)$ where:

- \mathcal{X} is the sample space (the set of all possible realizations $\mathbf{X} = (X_1, X_2, \dots, X_n)$)
- \mathcal{P} is a family of model functions $P_\theta(\mathbf{X})$ that depend on the unknown parameter θ ;
- Θ is the set of possible θ -values (the parameter space indexing the models).

2.3 Inference problem

The statistical inference problem can now be formulated:

Once the random vector \mathbf{X} has been observed, what can be said about which members of \mathcal{P} best describe how it was generated?

The reason we are speaking about a problem here is that we do **not** know the exact shape of the distribution that generated the data.

The reason that there exists a possibility of making inference rests in the fact that typically a given observation is much more probable under some distributions than under others (i.e. the observations give **information** about the distribution).

This information should be combined with the *a priori* information about the distribution to do the inference. Always there is some a priori information. It could be more or less specific.

Parametric Inference:

When it is specific to such an extent that the shape of the distribution is known up to some finite number of parameters i.e. the parameter θ is finite-dimensional, we have to conduct parametric inference.

Most of the classical statistical inference techniques are based on fairly specific assumptions regarding the population distribution and most typically the description of the population is in a parametric form.

In introductory textbooks, the student just practices applying standard parametric techniques. However, to be successful in practical statistical analysis, one has to be able to deal with situations where standard parametric assumptions are not justified.

Non-parametric Inference:

A whole set of methods and techniques are available that may be classified as nonparametric procedures. We will be dealing with them in the later parts of the course.

These procedures allow us to make inferences without or with a very limited amount of assumptions regarding the functional form of the underlying population distribution.

If Θ could only be specified as a *infinite dimensional function space*, we speak about *non-parametric* inference.

Non-parametric procedures are applicable in more general situations (which is good). However, if they are applied to a situation where a particular parametric distributional shape indeed holds, the nonparametric procedures may not be as efficient as compared to a procedure specifically tailored for the corresponding parametric case which would be bad if the specific parametric model indeed holds.

Robustness approach:

The situation, in practice, might be even more ^{模糊的} blurred. We may know that the population is close to parametrically describable but yet “deviates a bit” from the parametric family.

^{偏离}

Going over in such cases directly to a purely nonparametric approach would not properly address the situation since the idea about a relatively small deviation from the baseline parametric family will be lost. Hence we can use the robustness approach where we still keep the idea about the “ideal” parametric model but allow for *small* deviations from it.

^{中间的}

The aim is in such “intermediate” situations to be close to efficient if the parametric model holds but at the same time to be “less sensitive” to small deviations from the ideal model. These important issues will be discussed later in the course.

<https://www.cnblogs.com/hxsyl/p/4876465.html>

Bayesian Inference:

Another way to classify the Statistical Inference procedures is by the way we treat the *unknown* parameter θ .

classical / frequentist

If we treat it as unknown but deterministic then we are in a **Non-Bayesian** setting. If we consider the set of θ -values as quantities that before collecting the data, have different probabilities of occurring according to some (*a priori*) distribution, then we are speaking about *Bayesian inference*.

$\theta \sim \text{random variable}$

$\theta \sim X$

<https://www.cnblogs.com/hizhaolei/p/8196096.html>

Bayesian approach allows us to introduce and utilise any additional (prior) information (when such information is available). This information is entered through the **prior distribution** over the set Θ of

3. 贝叶斯决策的一般过程

(1) 估计先验概率: ①根据实际情况做经验估计; ②根据样本分布的频率估计概率。

(2) 计算类条件概率密度: ①参数估计: 类条件概率分布类型已知, 参数未知, 通过训练样本来估计 (最大似然法、Bayes估计); ②非参数估计: 不判断类条件概率分布类型, 直接根据训练样本来估计 (Parzen窗、k-近邻法)。

(3) 计算后验概率。

(4-1) 若进行最小错误率决策, 根据后验概率即可作出决策。

parameter values and reflects our prior belief about how likely any of the parameter values is before obtaining the information from the data.

2.4 Goals in statistical inference

Following are the most common activities in statistical inference:

- Estimation
- Confidence set construction
- Hypothesis testing

2.4.1 Estimation



We want to calculate a number (or a k -dimensional vector, or a single function) as an approximation to the numerical characteristic in question.

But let us point out immediately that there is little value in calculating an approximation to an unknown quantity without having an idea of how “good” the approximation is and how it compares with other approximations. Hence, immediately questions about **confidence interval** (or more generally, **confidence set**) construction arise.

To quote the famous statistician A.N. Whitehead, in Statistics we always have to “**seek simplicity and distrust it**”.

2.4.2 Confidence set construction

After the observations are collected, further information about the set Θ is added and it becomes ^{貌似合理的}plausible that the true distribution belongs to a smaller family than it was originally ^{假设}postulated, i.e., it becomes clear that the unknown θ -value belongs to a subset of Θ .

The problem of confidence set construction arises: i.e., determining a (possibly small) plausible set of θ -values and ^{澄清}clarifying the sense in which the set is plausible.

2.4.3 Hypothesis testing

$$H_0: \mu > 150$$

$$\mu \in \mathbb{R}$$

$$(X_1, \dots, X_n)$$

An experimenter or a statistician sometimes has a theory which when suitably translated into mathematical language becomes a statement that the true unknown distribution belongs to a smaller family than the originally postulated one.

One would like to formulate this theory in the form of a hypothesis. The data can be used then to infer whether or not his theory ^{服从}complies ^{混乱}with the observations or is in such a serious disarray that it would indicate that the hypothesis is false.

Deeper insight in all of the above goals of inference and deeper understanding of the nature of problems involved in them is given by **Statistical Decision Theory**.

Here we define in general terms what a *statistical decision rule* is and it turns out that all of the procedures discussed above can be viewed as a suitably defined decision rule.

Moreover, defining optimal decision rules as solutions to suitably formulated **constrained mathematical optimization problems** will help us to find “best” decision rules in many practical and relevant situations.

2.5 Statistical decision theoretic approach to inference

Statistical Decision Theory studies all inference problems (estimation, confidence set construction, hypothesis testing) from a unified point of view.

All parts of the decision making process are formally defined, a desired optimality ^{标准}criterion is formulated and a decision is considered optimal if it optimizes the criterion.

2.5.1 Introduction

Statistical Decision Theory may be considered as the theory of a two-person game with one player being the statistician and the other one being the nature. To specify the game, we define:

- Θ - set of states (of nature);
- \mathcal{A} - set of actions (available to the statistician);
- $L(\theta, a)$ - real-valued function (loss) on Θ \times \mathcal{A}

There are some important differences between mathematical theory of games (that only involves the above triplet) and Statistical Decision Theory. The most important differences are:

- In a two-person game both players are trying to maximize their winnings (or to minimize their **losses**), whereas in decision theory **nature** chooses a state without this view in mind. **Nature** can not be considered an "intelligent opponent" who would behave "rationally".
- There is no complete information available (to the **statistician**) about **nature's** choice.

- In Statistical Decision Theory **nature** always has the first move in choosing the "true state" θ .
- The **statistician** has the chance (and this is *most important*) to gather *partial information* on **nature's** choice by sampling or performing an experiment. This gives the **statistician** data $\mathbf{X} = (X_1, X_2, \dots, X_n)$ that has a distribution $L(\mathbf{X}|\theta)$ *depending on θ* . This is used by the **statistician** to work out his decision.

$$L(\mathbf{x}, \theta) = \prod_{i=1}^n f_{\mathbf{X}}(x_i)$$

Definition 2.1

A (deterministic) *decision function* is a function $d : \mathcal{X} \rightarrow \mathcal{A}$ from the sample space to the set of actions.

There is a non-negative **loss** (a random variable) $L(\theta, d(\mathbf{X}))$ incurred by this action. 引起

We define the RISK to be:

$$\mathbb{E}_{\theta}[L(\theta, d(\mathbf{X}))] = R(\theta, d).$$

For a fixed decision, this is a function (risk function) depending on θ . $R(\theta, d)$ is the average loss of the **statistician** when the **nature** has a true state θ and the **statistician** uses decision d .

2.5.2 Examples

Example 2.11 (Hypothesis testing)

Assume that a data vector $\mathbf{X} \sim f(\mathbf{X}, \theta)$. Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ where $\theta \in \mathbb{R}^1$ is a parameter of interest.

Let $\mathcal{A} = \{a_1, a_2\}$ and $\Theta = \mathbb{R}^1$ where a_1 denotes the action "accept H_0 " and a_2 is the action "Reject H_0 ". Let

$D = \{\text{Set of all functions from } \mathcal{X} \text{ into } \mathcal{A}\}.$

Define

$$L(\theta, a_1) = \begin{cases} 1 & \text{if } \theta > \theta_0, \\ 0 & \text{if } \theta \leq \theta_0 \end{cases}$$

$$L(\theta, a_2) = \begin{cases} 0 & \text{if } \theta > \theta_0, \\ 1 & \text{if } \theta \leq \theta_0 \end{cases}$$

"0-1 loss"

wrong
correct

correct
wrong

Then we have

RISK:

$$\begin{aligned} R(\theta, d) &= \mathbb{E}[L(\theta, d(\mathbf{X}))] \\ &= L(\theta, a_1)P_\theta(d(\mathbf{X}) = a_1) + L(\theta, a_2)P_\theta(d(\mathbf{X}) = a_2) \\ &= \begin{cases} P_\theta(d(\mathbf{X}) = a_1) & \text{if } \theta > \theta_0, \\ P_\theta(d(\mathbf{X}) = a_2) & \text{if } \theta \leq \theta_0. \end{cases} \end{aligned}$$

Hence

- if $\theta \leq \theta_0$: $R(\theta, d) = P_\theta(\text{reject } H_0) = \text{Error of type I},$
- if $\theta > \theta_0$: $R(\theta, d) = P_\theta(\text{accept } H_0) = \text{Error of type II}.$

Example 2.12 (Estimation)

Let now $\mathcal{A} = \Theta$ with the interpretation that each action corresponds to selecting a point $\theta \in \Theta$. Every $d(\mathbf{X})$ maps \mathcal{X} into Θ and if we chose

$$L(\theta, d(\mathbf{X})) = (\theta - d(\mathbf{X}))^2 \quad (\text{quadratic loss})$$

then the decision rule d (which we can call an estimator) has a risk function

$$R(\theta, d) = \mathbb{E}_{\theta}[(d(\mathbf{X}) - \theta)^2] = \text{MSE}_{\theta}(d(\mathbf{X})).$$

$$\mathbb{E}_{\theta}[L(\theta, d(\mathbf{X}))]$$

2.5.3 Randomized decision rule

We will see later when studying optimality in hypothesis testing context that the set of deterministic decision rules D is not ^{凸面的}convex and it is difficult to develop a decent mathematical optimization theory over it.

This set is also very small and examples show that very often a simple randomization of given deterministic rules gives better rules in the sense of risk minimization. This explains the reason for the introduction of the randomized decision rules.

$$\delta = \begin{cases} d_1 & w_1 & w_1 \\ d_2 & w_2 & w_2 \\ \vdots & \vdots & \vdots \\ d_m & w_m & w_m \end{cases}$$

$$\mathcal{D} = \{d_1, d_2, \dots, d_m\}$$

Definition 2.2

A rule δ which chooses d_i with probability w_i is a randomized decision rule where $\sum w_i = 1$.

For the randomized decision rule δ we have:

$$L(\theta, \delta(X)) = \sum w_i L(\theta, d_i(X)) \quad \text{and} \quad R(\theta, \delta) = \sum w_i R(\theta, d_i).$$

The set of all randomized decision rules generated by the set D in the above way will be denoted by \mathcal{D} .

D \mathcal{D}

2.5.4 Optimal decision rules

Given a game (Θ, \mathcal{A}, L) and a random vector X whose distribution depends on $\theta \in \Theta$ what (randomized) decision rule δ should the statistician choose to perform “optimally”?

This is a question that is easy to pose but usually difficult to answer.

$$\theta \in \Theta$$

The reason is that usually uniformly best decision rules (that minimize the risk uniformly for all θ -values) do not exist! It leads us to the following two ways out:

First way out:

Constraining the set of decision rules and try to find uniformly best in this smaller set. This corresponds to looking for optimality under restrictions. We can eliminate some of the decision rules since they do not satisfy the restrictions by hoping, in the smaller set of remaining rules to be able to find a uniformly best.

Sensible constraints that we introduce in the estimation context are usually unbiasedness or invariance.

unbiased: $E(\hat{\theta}) = \theta$

Definition 2.3

A decision rule d is *unbiased* if

$$\mathbb{E}_{\theta'}[L(\theta', d(\mathbf{X}))] \geq \mathbb{E}_{\theta}[L(\theta, d(\mathbf{X}))] \text{ for all } \theta, \theta' \in \Theta$$

holds.

$$L(\theta, d) = (\theta - d(x))^2$$

Exercise 2.9 (at lecture)

Show that in the context of estimation of a parameter θ with quadratic loss function, the above definition is tantamount to the requirement

$$E_{\theta}[d(\mathbf{X})] = \theta \quad \text{for all } \theta \in \Theta,$$

that is, the new definition is equivalent to the unbiasedness from classical statistical estimation theory.

$$\text{e.g. } L(\theta, d) = |\theta - d(x)|$$

It should be noted that the new definition of unbiasedness is more general and can be applied to broader class of loss functions.

The same definition also makes sense in hypothesis testing where we can also introduce unbiased tests in the same way (see later the separate lecture about optimality in hypothesis testing) and then look for optimality amongst all unbiased α level tests.

Second way out:

Reformulating the optimality criterion in a new way. Since the “uniformly best” no matter what θ -value is too strong a requirement, we can introduce:

- Bayes risk
- Minimax risk

of a decision rule and try to find the rules that minimize these risks.

This leads to Bayesian and to minimax decision rules.

2.5.5 Bayesian and minimax decision rules

Bayesian rule:

Think of the parameter θ as random variable with a given (known) prior density $\tau(\theta)$ on Θ .

Define the **Bayesian risk of the decision rule δ** with respect to the prior $\tau(\theta)$:

$$r(\tau, \delta) = \mathbb{E}[R(T, \delta)] = \int_{\Theta} R(\theta, \delta) \tau(\theta) d\theta$$

"Integrated out θ "

where T is a random variable over Θ with a density τ .

Then the **Bayesian rule** δ_τ with respect to the prior $\tau(\theta)$ is defined as:

$$\underline{r(\tau, \delta_\tau)} = \overset{\text{min}}{\inf_{\delta \in \mathcal{D}}} \underline{r(\tau, \delta)} \quad \delta \in \mathcal{D}$$

Sometimes a Bayesian rule may not exist and so we ask for an ϵ -Bayes rule.

For $\epsilon > 0$, this is any rule $\delta_{\epsilon\tau}$ that satisfies

$$r(\tau, \delta_{\epsilon\tau}) \leq \inf_{\delta \in \mathcal{D}} r(\tau, \delta) + \epsilon.$$

Minimax rule:

Instead of considering uniformly best rules we consider rules that minimize the supremum^{上确界} of the values of the risk over the set Θ . This means safeguarding against the worst possible performance.

The value

$$\sup_{\theta \in \Theta} R(\theta, \delta)$$

$$\delta \in \mathcal{D} \quad \mathcal{D} \subseteq \mathcal{D}$$

$$\delta \in \mathcal{D}$$

is called **minimax risk** of the decision rule δ . Then the rule δ^* is called **minimax** in the set \mathcal{D} if

$$\sup_{\theta \in \Theta} R(\theta, \delta^*) = \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta) = \text{minimax value of the game.}$$

Again, like in the Bayesian case, even if the minimax value is finite there may not be a minimax decision rule. Hence we introduce the notion of ϵ -minimax rule δ_ϵ such that

$$\sup_{\theta \in \Theta} R(\theta, \delta_\epsilon) \leq \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta) + \epsilon.$$

Note that sometimes choosing a minimax rule may turn out to be a too pessimistic strategy but experience shows that in most cases minimax rules are good.

o.g. worst case scenario occurs with extremely probability.

2.5.6 Least favorable prior distribution

Now we define the **least favorable distribution** (i.e. least favorable prior $\tau^*(\theta)$ over the set Θ) as:

$$\inf_{\delta \in \mathcal{D}} r(\tau^*, \delta) = \sup_{\tau} \inf_{\delta \in \mathcal{D}} r(\tau, \delta)$$

Prior that leads
to the largest
Bayesian risk.

It indeed deserves its name. If the **statistician** were told which prior distribution nature was using, he would like least to be told that τ^* was the **nature's** prior (since given that he always performs in an optimal way by choosing the corresponding Bayesian rule, he still has the highest possible value of the Bayesian risk as compared to the other priors).

2.5.7 Geometric interpretation for finite Θ

几何理解

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$$

Definition 2.4

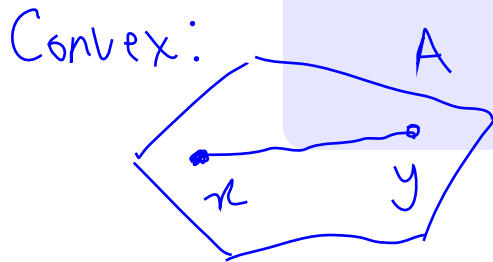
A set $A \subset \mathbb{R}^k$ is ^{凸面的}convex if for all vectors $\vec{x} = (x_1, x_2, \dots, x_k)'$ and $\vec{y} = (y_1, y_2, \dots, y_k)'$ in A and all $\alpha \in [0, 1]$ then

凸集是对于集合内的每一对点，连接该对点的直线段上的每个点也在该集合内

Not convex



$$\alpha \vec{x} + (1 - \alpha) \vec{y} \in A.$$



Now let's assume that Θ has k elements only. Then define the **risk set** of a set D of decision rules as the set of all **risk points** $\{R(\theta, d), \theta \in \Theta\}$, $d \in D$. For a fixed d , each such risk point belongs to \mathbb{R}^k and by “moving” d within D , we get a set of such k -dimensional vectors.

Theorem 2.6

The risk set of a set \mathcal{D} of randomized decision rules generated by a given set D of non-randomized decision rules is convex.

Proof.

If \vec{y} and \vec{y}^* are the risk points of δ and $\delta^* \in D$, correspondingly, then any point in the form

$$\vec{z} = \alpha \vec{y} + (1 - \alpha) \vec{y}^*$$

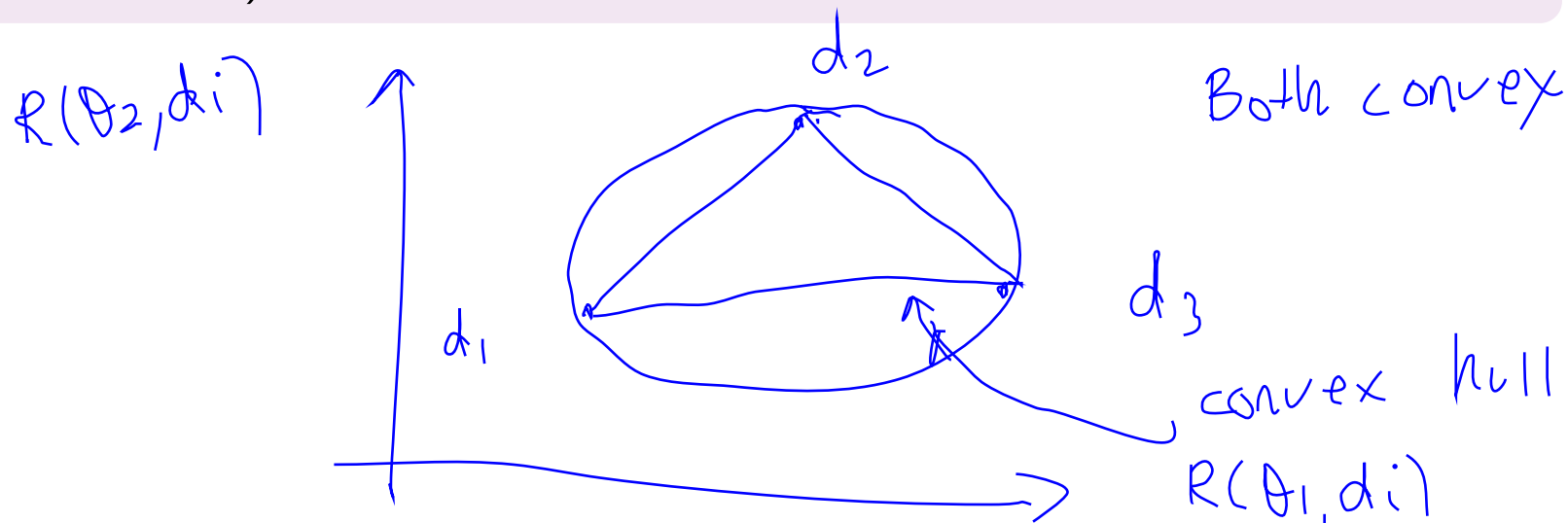
$$\alpha \in (0, 1]$$

corresponds to (is the risk point of) the randomized decision rule $\delta_\alpha \in \mathcal{D}$ that chooses δ with probability α and the rule δ^* with probability $(1 - \alpha)$. Hence any such \vec{z} belongs to the risk set of \mathcal{D} . \square

Remark 2.1

The risk set of the set of all randomized rules \mathcal{D} generated by the set D is *the smallest convex set containing the risk points of all of the non-randomized rules in D (i.e. the convex hull of the set of risk points of D).*

点集Q的凸包 (convex hull) 是指一个最小凸多边形, 满足Q中的点或者在多边形边上或者在其内。



How to illustrate Bayes rules:

Since $\Theta = (\theta_1, \theta_2, \dots, \theta_k)$ then the prior $\tau = (p_1, p_2, \dots, p_k)$ in the case we are dealing with ($p_i \geq 0, \sum_{i=1}^k p_i = 1$). The Bayes risk of any rule δ with respect to the prior τ is $r(\tau, \delta) = \sum_{i=1}^k p_i R(\theta_i, \delta)$. $= b$

All points \vec{y} in the risk set, corresponding to certain rules δ^* for which

$$\sum_{i=1}^k p_i y_i = r(\tau, \delta^*) = \text{the same value} = b,$$

give rise to the same value b of the Bayesian risk and hence are equivalent from a Bayesian point of view. The value of their risk can be easily illustrated and (at least in case of $k = 2$), one can easily illustrate the point in the convex risk set that corresponds to (is the risk point of the) Bayesian rule with respect to the prior τ .

How to illustrate Minimax rules:

In a similar way, the minimax rule can be illustrated in the case of finite Θ . Three cases can be illustrated:

- i) The minimax decision rule corresponding to the point at the lower intersection ^{交点} of S with the line $R_1 = R_2$;
- ii) When S lies entirely to the left of the line $R_1 = R_2$ so that $R_1 < R_2$ for every point in S , and therefore the minimax rule is simply that which minimises R_2 ;
- iii) When S lies entirely to the right of the line $R_1 = R_2$ so that $R_1 > R_2$ for every point in S , and therefore the minimax rule is simply that which minimises R_1 .

2.5.8 Example

Let the set $\Theta = \{\theta_1, \theta_2\}$. Let X have possible values 0, 1 and 2; the set $\mathcal{A} = \{a_1, a_2\}$ and let

$$L(\theta_1, a_1) = L(\theta_2, a_2) = 0, \quad L(\theta_1, a_2) = 1, \quad L(\theta_2, a_1) = 3.$$

The distributions of X are tabulated as follows:

x	0	1	2
$P(x \theta_1)$	0.81	0.18	0.01

x	0	1	2
$P(x \theta_2)$	0.25	0.5	0.25

Interpretation: an attempt by the **statistician** to guess the state of **nature**. If their guess is correct, they do not lose anything; if they are wrong, they loses \$1 or \$3 depending on the type of error they made. In their guess they are supported by one observation X that has a different distribution under θ_1 and under θ_2 .

Exercise 2.10 (at lecture)

Consider all possible non-randomized decision rules based on one observation:

$$2^3 = 8$$

$$2 \times 2 \times 2$$

x	$d_1(x)$	$d_2(x)$	$d_3(x)$	$d_4(x)$	$d_5(x)$	$d_6(x)$	$d_7(x)$	$d_8(x)$
0	a_1	a_1	a_1	a_1	a_2	a_2	a_2	a_2
1	a_1	a_1	a_2	a_2	a_1	a_1	a_2	a_2
2	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2

a) Sketch the risk set of all randomized rules generated by

$$\mathcal{D} = \{d_1, d_2, \dots, d_8\}$$

b) Find the minimax rule δ^* in \mathcal{D} and compute its risk.

c) For what prior is δ^* a Bayes rule with respect to that prior (i.e. what is the least favorable distribution)?

d) Find the Bayes rule for the prior $\{1/3, 2/3\}$ over $\{\theta_1, \theta_2\}$. Compute the value of its Bayes risk.

2.5.9 Fundamental lemma

$$R(\theta, \delta_{\tau^*}) = c_0 \quad \forall \theta \in \Theta$$

Lemma

If $\tau^*(\theta)$ is a prior on Θ and the Bayes rule δ_{τ^*} has a constant risk with respect to θ (i.e. if $R(\theta, \delta_{\tau^*}) = c_0$ for all $\theta \in \Theta$) then:

- i) δ_{τ^*} is minimax;
- ii) $\tau^*(\theta)$ is the least favorable distribution.

Proof.

minimises the Bayesian risk

i) We compute the minimax risk of δ_{τ^*} and compare it to the minimax risk of any other rule δ :

(call c_0 for all $\theta \in \Theta$)

$$c_0 = \sup_{\theta \in \Theta} R(\theta, \delta_{\tau^*})$$

$$= \int_{\Theta} R(\theta, \delta_{\tau^*}) \tau^*(\theta) d\theta \quad \text{since constant for all } \theta \quad (\text{multiply by one})$$

$$\leq \int_{\Theta} R(\theta, \delta) \tau^*(\theta) d\theta \quad \text{since } \delta_{\tau^*} \text{ Bayes with respect to } \tau^*$$

$$\leq \sup_{\theta \in \Theta} R(\theta, \delta)$$

$$\sup_{\theta \in \Theta} R(\theta, \delta_{\tau^*}) \leq \sup_{\theta \in \Theta} R(\theta, \delta)$$

which means that δ_{τ^*} is minimax.

for any $\delta \in \mathcal{D}$

ii) Now take any other prior $\tau(\theta)$:

Bayesian decision rule
prior τ :

def

$$\inf_{\delta} r(\tau, \delta) = \inf_{\delta} \int_{\Theta} R(\theta, \delta) \tau(\theta) d\theta$$

small as possible

$$\begin{aligned} &\leq \int_{\Theta} R(\theta, \delta_{\tau^*}) \tau(\theta) d\theta = R(\theta, \delta_{\tau^*}) \underbrace{\int_{\Theta} \tau(\theta) d\theta}_1 \\ &= \int_{\Theta} R(\theta, \delta_{\tau^*}) \tau^*(\theta) d\theta \\ &= r(\tau^*, \delta_{\tau^*}), \end{aligned}$$

$\int_{\Theta} \tau^*(\theta) d\theta$

hence $\tau^*(\theta)$ is least favorable. (Leads to the largest Bayesian risk)

Note here we have used the fact that $R(\theta, \delta_{\tau^*})$ is constant and

$$\int_{\Theta} \tau^*(\theta) d\theta = \int_{\Theta} \tau(\theta) d\theta = 1.$$

□

Remark 2.2

The lemma provides a hint on how to find minimax estimators. The minimax estimators are (special) Bayes estimators with respect to the least favorable prior. $\tau^*(\theta)$

First, we can obtain the general form of the Bayes estimator with respect to any given prior. Then we choose a prior for which the corresponding Bayes rule has its (usual) risk independent of θ , i.e. constant with respect to θ .

2.5.10 Finding Bayes rules analytically

This is important in its own right but also as a device to be utilized in the search for minimax rules.

Given the prior and the observations $\mathbf{X} = (X_1, X_2, \dots, X_n)$ we can find the Bayes rule point-wise (i.e. for any given $\mathbf{X} = \mathbf{x}$) by solving a certain minimization problem.



Important notation:

- $f(\mathbf{X}|\theta)$ is the conditional density of $\mathbf{X} = (X_1, X_2, \dots, X_n)$ given the parameter θ ;
- $\tau(\theta)$ is the prior density on θ ; 先验密度
- $g(\mathbf{X})$ is the marginal density of \mathbf{X} , i.e. $g(\mathbf{X}) = \int_{\Theta} f(\mathbf{X}|\theta)\tau(\theta)d\theta$; 边际密度
- $f(\mathbf{X}, \theta)$ is the joint density of \underline{X} and $\underline{\theta}$

$$\int_{\Theta} f(\mathbf{X}|\theta)d\theta$$

$$f(\mathbf{X}, \theta) = f(\mathbf{X}|\theta)\tau(\theta) = h(\theta|\mathbf{X})g(\mathbf{X}).$$

- $h(\theta|\mathbf{X})$ is the posterior density of θ given $\mathbf{X} = (X_1, X_2, \dots, X_n)$; 后验密度

$$h(\theta|\mathbf{X}) = \frac{f(\mathbf{X}, \theta)}{g(\mathbf{X})} = \frac{f(\mathbf{X}|\theta)\tau(\theta)}{\int_{\Theta} f(\mathbf{X}|\theta)\tau(\theta)d\theta}$$

"Bayes Theorem"

Now we formulate a **General Theorem** regarding the calculation of Bayesian decision rules.

Theorem 2.7

For $X \in \mathcal{X}$, $a \in \mathcal{A}$ and for a given prior $\tau(\theta)$ we define:

$$Q(\mathbf{X}, a) = \int_{\Theta} \underline{L(\theta, a)} \underline{h(\theta|\mathbf{X})} d\theta,$$

where $L(\cdot, \cdot)$ is a particular **loss** function.

Suppose that for each $\mathbf{X} \in \mathcal{X}$, there exists a rule $a_{\mathbf{X}} \in \mathcal{A}$ such that

$$Q(X, a_{\mathbf{X}}) = \inf_{a \in \mathcal{A}} Q(\mathbf{X}, a)$$

If $\delta_{\tau}(\mathbf{X}) = a_{\mathbf{X}}$ belongs to \mathcal{D} then $\delta_{\tau}(\mathbf{X}) = a_{\mathbf{X}}$ is the (point wise defined) Bayes decision rule with respect to the prior $\tau(\theta)$.

Proof.

$$E[R(\tau, \delta)]$$

$$R(\tau, \delta) = E_{\theta}[L(\theta, \delta)]$$

For any decision rule δ we have its Bayesian risk:

$$\begin{aligned}
 r(\tau, \delta) &= \int_{\Theta} \underline{R(\theta, \delta)} \tau(\theta) d\theta \\
 &= \int_{\Theta} \left[\int_{\mathcal{X}} \underline{L(\theta, \delta(\mathbf{X}))} \underline{f(\mathbf{X}|\theta)} d\mathbf{X} \right] \tau(\theta) d\theta \\
 &= \int_{\mathcal{X}} \left[\int_{\Theta} \underline{L(\theta, \delta(\mathbf{X}))} \underline{h(\theta|\mathbf{X})} d\theta \right] \underline{g(\mathbf{X})} d\mathbf{X} \\
 &= \int_{\mathcal{X}} Q(\mathbf{X}, \delta(\mathbf{X})) g(\mathbf{X}) d\mathbf{X}
 \end{aligned}$$

$$\begin{aligned}
 f(\mathbf{X}, \theta) &= f(\mathbf{X}|\theta) \tau(\theta) \\
 &= h(\theta|\mathbf{X}) g(\mathbf{X})
 \end{aligned}$$

where we use the short-hand notation $d\mathbf{X} := dX_1 dX_2 \dots dX_n$.

$$\min_{\delta \in \mathcal{D}} r(\tau, \delta) \Leftrightarrow \min_{\delta \in \mathcal{D}} Q(\mathbf{X}, \delta(\mathbf{X}))$$

But for every fixed \mathbf{X} -value, $Q(\mathbf{X}, \delta(\mathbf{X}))$ is smallest when $\delta(\mathbf{X}) = a_{\mathbf{X}}$. Making that way our “best choice” for each \mathbf{X} -value, we will, of course, minimize the value of $r(\tau, \delta)$. Hence, we should be looking for an action $a_{\mathbf{X}}$ that gives an infimum to

$$\inf_{a \in \mathcal{A}} \int_{\Theta} L(\theta, a) h(\theta | \mathbf{X}) d\theta. \quad \therefore \inf_{a \in \mathcal{A}} Q(\mathbf{X}, a) \quad \square$$

We now will apply the general theorem to estimation and to hypothesis testing.

$$\Lambda = \Theta$$

Theorem 2.8 (case of estimation)

Consider a point estimation problem for a real-valued parameter θ . The prior over θ is denoted by $\tau(\theta)$. Then:

- a) For a squared error loss $L(\theta, a) = (\theta - a)^2$:

$$\delta_\tau(\mathbf{X}) = \mathbb{E}(\theta|\mathbf{X}) = \int_{\Theta} \theta h(\theta|\mathbf{X}) d\theta.$$

"posterior mean"

The Bayesian estimator with respect to quadratic loss is the conditional expected value of the parameter given the observed data.

- b) For an absolute error loss $L(\theta, a) = |\theta - a|$:

"posterior median"

$$\int_{-\infty}^{\delta_\tau} h(\theta|\mathbf{X}) d\theta = 0.5$$

$$\delta_\tau(\mathbf{X}) = \text{median of } h(\theta|\mathbf{X}).$$

The Bayesian estimator with respect to absolute value loss is the median of the conditional distribution of the parameter given the observed data.

Example 2.13

Show that for a given random variable Y with a finite second moment, the function $q_1(a) = \mathbb{E}[(Y - a)^2]$ is minimised for $a^* = \mathbb{E}(Y)$.

Solution:

Setting the derivative with respect to a to zero we get

$$\frac{\partial}{\partial a} q_1(a) = \frac{\partial}{\partial a} \mathbb{E}[(Y - a)^2] = \frac{\partial}{\partial a} [\cancel{\mathbb{E}(Y^2)} - 2\mathbb{E}(Y)a + a^2] = -2\mathbb{E}(Y) + 2a = 0$$

from which we deduce that the stationary point is $a^* = \underline{\mathbb{E}(Y)}$ and this stationary point gives rise to a minimum since

$$\frac{\partial^2 q_1(a)}{\partial a^2} = \frac{\partial^2}{\partial a^2} \mathbb{E}[(Y - a)^2] = 2 > 0.$$

Example 2.14

Show that for a given random variable Y with $E|Y| < \infty$, the function $q_2(b) = \mathbb{E}[|Y - b|]$ is minimised for $b^* = \text{median}(Y)$.

$$|Y - b| = \begin{cases} y - b & y \geq b \\ b - y & y < b \end{cases}$$

$$E[|Y - b|] = \int_b^{\infty} (y - b) f(y) dy + \int_{-\infty}^b (b - y) f(y) dy$$

Solution (continuous case for simplicity):

Let the density of Y be denoted by $f(y)$ and the cdf by $F(y)$. Having in mind the definition of absolute value we have:

$$\begin{aligned}
 \frac{\partial}{\partial b} \mathbb{E}(|Y - b|) &= \frac{\partial}{\partial b} \left[\int_{-\infty}^b (b - y)f(y)dy + \int_b^{\infty} (y - b)f(y)dy \right] \\
 &= \frac{\partial}{\partial b} \left[\underbrace{bF(b)} - \underbrace{\int_{-\infty}^b yf(y)dy} + \underbrace{\int_b^{\infty} yf(y)dy} - \underbrace{b(1 - F(b))} \right] \\
 &= \underbrace{1}_{\text{from } \frac{\partial}{\partial b} bF(b)} F(b) + \underbrace{bf(b)}_{\text{from } -\frac{\partial}{\partial b} \int_{-\infty}^b yf(y)dy} - \underbrace{bf(b)}_{\text{from } \frac{\partial}{\partial b} \int_b^{\infty} yf(y)dy} - \underbrace{(1 - F(b))}_{\text{from } -\frac{\partial}{\partial b} b(1 - F(b))} + \underbrace{bf(b)}_{\text{from } -\frac{\partial}{\partial b} b(1 - F(b))} \\
 &= 2F(b) - 1 \\
 &= 0
 \end{aligned}$$

$F(b^*) = \frac{1}{2}$
 $\frac{\partial^2 \ell_2(b)}{\partial b^2} = 2f(b) > 0$

from which we deduce that the stationary point b^* satisfies $F(b^*) = 0.5$, i.e., b^* is the median. The stationary point b^* gives rise to a minimum but you should check that the second derivative is positive.

Remark 2.3

A case in which $\delta_\tau \in \mathcal{D}$ could not be satisfied is in point estimation with $\Theta \equiv \mathcal{A}$ - a finite set. Then $\mathbb{E}(\theta|\mathbf{X})$ might not belong to \mathcal{A} , hence $\mathbb{E}(\theta|\mathbf{X})$ would not be a function $\mathcal{X} \rightarrow \mathcal{A}$ and δ_τ would not be a legitimate estimator. But if $\Theta \equiv \mathcal{A}$ is *convex*, it can be shown that *always* $\mathbb{E}(\theta|\mathbf{X}) \in \mathcal{A}$.

Bayesian hypothesis testing with a generalized 0 – 1 loss

A prior $\tau(\theta)$ is given on Θ . Assume that the parameter space Θ is subdivided into two complementary subsets $\Theta = \Theta_0 \cup \Theta_1$ and we are testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1.$$

Two actions a_0 (accept H_0) and a_1 (reject H_0) are possible and the losses when using these actions are given by:

$$L(\theta, a_0) = \begin{cases} 0 & \text{if } \theta \in \Theta_0 \\ c_2 & \text{if } \theta \in \Theta_1 \end{cases} \quad \text{and} \quad L(\theta, a_1) = \begin{cases} c_1 & \text{if } \theta \in \Theta_0 \\ 0 & \text{if } \theta \in \Theta_1. \end{cases}$$

These losses make sense in hypothesis testing because the correct guess of H_0 or of H_1 should not involve any loss (so the loss is set to zero) whereas an incorrect guess should involve some positive loss.)

The loss when a first type error occurs is denoted as c_1 and the loss when a second type error occurs is denoted as c_2 . Since the consequences of the two types of error may not be equally heavy, $c_1 \neq c_2$ in general.)

Theorem 2.9 (case of hypothesis testing)

Assume that the parameter space Θ is subdivided into two complementary subsets $\Theta = \Theta_0 \cup \Theta_1$ and we are testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1.$$

with a generalised 0-1 loss function. Then the test

$$\varphi^* = \begin{cases} \text{Reject } H_0 & \text{if } P(\theta \in \Theta_0 | \mathbf{X}) < \frac{c_2}{c_1 + c_2} \\ \text{Accept } H_0 & \text{if } P(\theta \in \Theta_0 | \mathbf{X}) > \frac{c_2}{c_1 + c_2} \end{cases}$$

is a Bayesian rule (Bayesian test) for the above testing problem with respect to the prior $\tau(\theta)$.

Proof.

According to the General Theorem about Bayesian inference, we have to compare the two quantities $Q(\mathbf{X}, a_0)$ and $Q(\mathbf{X}, a_1)$ and take as our action the one that gives the smaller value. In this way we are minimising the Bayesian risk for the given prior, hence we are deriving the optimal Bayesian decision in the context of hypothesis testing.

$$\begin{aligned}
Q(\mathbf{X}, a_0) &= \int_{\Theta} L(\theta, a_0) h(\theta|\mathbf{X}) d\theta & Q(\mathbf{X}, a_1) &= \int_{\Theta} L(\theta, a_1) h(\theta|\mathbf{X}) d\theta \\
&= \int_{\Theta_1} c_2 h(\theta|\mathbf{X}) d\theta & &= \int_{\Theta_0} c_1 h(\theta|\mathbf{X}) d\theta \\
&= c_2 P(\theta \in \Theta_1|\mathbf{X}) & &= c_1 P(\theta \in \Theta_0|\mathbf{X}) \\
&= c_2 (1 - P(\theta \in \Theta_0|\mathbf{X})) & &
\end{aligned}$$

Hence we would reject H_0 when $Q(\mathbf{X}, a_1) < Q(\mathbf{X}, a_0)$, i.e. for

$$\{\mathbf{X} : c_1 P(\theta \in \Theta_0|\mathbf{X}) < c_2 (1 - P(\theta \in \Theta_0|\mathbf{X}))\}$$

which is equivalent to

$$\left\{ \mathbf{X} : P(\theta \in \Theta_0|\mathbf{X}) < \frac{c_2}{c_1 + c_2} \right\} //$$

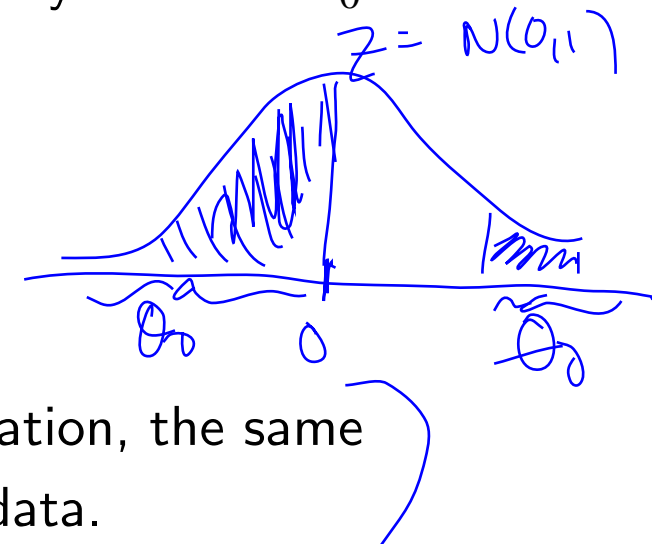
$$P(Z < 0) = \int_{-\infty}^0 f(z) dz$$

This means that to perform Bayesian hypothesis testing of

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

we need to calculate the posterior conditional probability that $\theta \in \Theta_0$ given the data:

$$P(\theta \in \Theta_0) = \int_{\Theta_0} h(\theta|\mathbf{X}) d\theta$$



For this calculation we need, as in the case of estimation, the same posterior density $h(\theta|\mathbf{X})$ of the parameter given the data.

Then we compare this posterior conditional probability with the threshold $c_2 / (c_1 + c_2) \in (0, 1)$ and reject H_0 when this posterior probability is not large enough, i.e., is below the threshold.

This makes perfect sense. We also note that the threshold $c_2 / (c_1 + c_2)$ to compare with, when the two types of errors are equally weighted (i.e., when $c_1 = c_2$ is chosen, is just equal to $1/2$). —

The case $c_1 = c_2$ (so that the ratio $c_2 / c_1 = 1$) is referred to as the usual zero-one loss in Bayesian hypothesis testing. The general case of different c_1 and c_2 is referred to as the generalised zero-one loss. \square

2.5.11 Examples

Example 2.15 (at lecture)

Given θ , the distribution of each X_i , $i = 1, 2, \dots, n$ are Bernoulli with parameter θ , i.e.

$$f(\mathbf{X}|\theta) = \theta^{\sum X_i} (1 - \theta)^{n - \sum X_i} = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1 - x_i}$$

Assume a beta prior $\tau(\theta)$ for the (random variable) θ over $(0, 1)$:

$$\tau(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1} I_{(0,1)}(\theta)$$

$0 < \theta < 1$
 $L(\theta, a) = (\theta - a)^2$

Show that the Bayesian estimator $\hat{\theta}_B$ with respect to quadratic loss is:

$$\hat{\theta}_B = \frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}$$

Hence, calculate the minimax estimator for the probability of success in the Bernoulli experiment.

We recall first the definition and some properties of the Beta function that is used in the definition of the Beta density. In particular, there is the following relation between $B(\alpha, \beta)$ and the Gamma function $\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Normalizing constant

Since the gamma function satisfies $\Gamma(a) = (a-1)\Gamma(a-1)$, after substitution we get the following recurrent relation for the Beta function:

$$B(\alpha, \beta) = \frac{\alpha-1}{\alpha+\beta-1} B(\alpha-1, \beta) \Rightarrow \frac{B(\alpha, \beta)}{B(\alpha-1, \beta)} = \frac{\alpha-1}{\alpha+\beta-1}$$

Solution (at lecture):

For the beta prior and after a bit of work, we can find that:

$$h(\theta|\mathbf{X}) = \frac{f(\mathbf{X}|\theta)\tau(\theta)}{\int_0^1 f(\mathbf{X}|\theta)\tau(\theta)d\theta} = \frac{\theta^{\sum X_i + \alpha - 1} (1 - \theta)^{n - \sum X_i + \beta - 1}}{B(\sum X_i + \alpha, n - \sum X_i + \beta)}$$

which is again a beta density.

$$\theta | \mathbf{X} \sim \text{Beta} \left(\sum_{i=1}^n X_i + \alpha, n - \sum_{i=1}^n X_i + \beta \right)$$

Hence, the Bayesian estimator

$$\begin{aligned}
 \hat{\theta}_\tau &= \int_0^1 \theta h(\theta|\mathbf{X}) d\theta \\
 &= \frac{\Gamma(\sum X_i + \alpha + 1) \Gamma(n + \alpha + \beta)}{\Gamma(n + 1 + \alpha + \beta) \Gamma(\sum X_i + \alpha)} \\
 &= \frac{B(\sum X_i + \alpha + 1, n - \sum X_i + \beta)}{B(\sum X_i + \alpha, n - \sum X_i + \beta)} \\
 &= \text{by the above property of the Beta function} \\
 &= \frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n} \\
 &= \frac{\bar{X} + \alpha/n}{1 + \frac{\alpha + \beta}{n}}
 \end{aligned}$$

divide by n
 $\frac{\bar{X} + 0}{1 + 0} = \bar{X} \quad \text{as } n \rightarrow \infty$

The above derivation holds for any $\text{Beta}(\alpha, \beta)$ prior.

also MLE

 θ

Compare the estimator obtained with the UMVUE (to be discussed in Lecture 3) \bar{X} and appreciate the effect of the prior on the form of the estimator for small and for large sample size n . The UMVUE also coincides with the MLE here.

 $\sim (\alpha, \beta)$ $\theta \sim \text{Beta}(\alpha, \beta)$

In particular, when the sample size n is small, the effect of the prior (via the values of the parameters α and β) may be significant and the Bayesian estimator $\hat{\theta}_\tau$ may be very different from the UMVUE ^{/MLE} thus expressing the influence of the prior information on our decision.

On the other hand, when the sample size n is very large, we see that $\hat{\theta}_\tau \approx \bar{X}$ holds no matter what the prior. That is, when the sample size increases, the prior's effect on the estimator starts disappearing!

$$\text{Var}(aY) = a^2 \text{Var}(Y)$$

$$Y = \sum X_i \sim \text{Binomial}(n, \theta) \quad \begin{aligned} \mathbb{E}(Y) &= n\theta \\ \text{Var}(Y) &= n\theta(1-\theta) \end{aligned}$$

Let us calculate the (usual) risk with respect to quadratic loss of any such Bayes estimator:

$$R(\theta, \hat{\theta}_\tau) = \mathbb{E}_\theta [L(\theta, \hat{\theta}_\tau)]$$

$$\text{MSE} = \text{variance} + \text{bias}^2$$

$$\hat{\theta}_\tau = \frac{\sum X_i + \alpha}{\alpha + \beta + n}$$

$$= \text{Var}_\theta(\hat{\theta}_\tau) + (\theta - \mathbb{E}_\theta(\hat{\theta}_\tau))^2$$

$$= \frac{n\theta(1-\theta)}{(n+\alpha+\beta)^2} + \left(\frac{n\theta + \alpha}{\alpha + \beta + n} - \theta \right)^2$$

= ...

= ... more tedious algebra that we can skip ...

try this.

= ...

$$= \frac{n\theta - n\theta^2 + (\alpha + \beta)^2\theta^2 + \alpha^2 - 2\alpha(\alpha + \beta)\theta}{(n + \alpha + \beta)^2}$$

numerator:

$$(-n + (\alpha + \beta)^2)\theta^2 + (n - 2\alpha(\alpha + \beta))\theta$$

t^*

$$\frac{n^2}{2\alpha^2} = n$$

$$\alpha = \frac{\sqrt{n}}{2}$$

For this risk not to depend on θ , it must hold that:

$$\begin{cases} (\alpha + \beta)^2 = n \\ 2\alpha(\alpha + \beta) = n \end{cases}$$

solve it

rewrite 2nd eq

$$\alpha + \beta = \frac{n}{2\alpha} \quad \text{sub into ①}$$

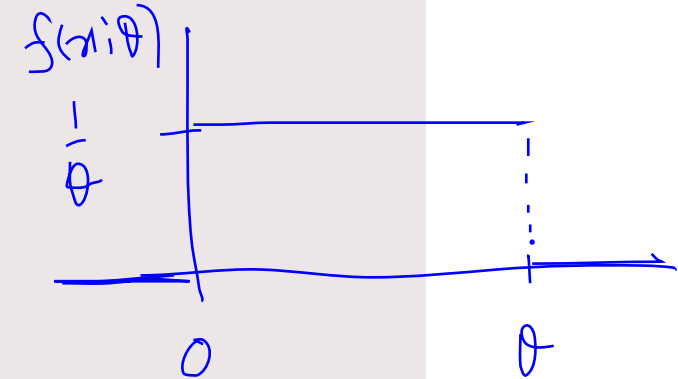
The solution to this system is $\alpha = \beta = \sqrt{n}/2$. Hence the minimax estimator of θ is

$$\hat{\theta}_{\text{minimax}} = \frac{\sum X_i + \sqrt{n}/2}{n + \sqrt{n}}.$$

Exercise 2.11 (at lecture)

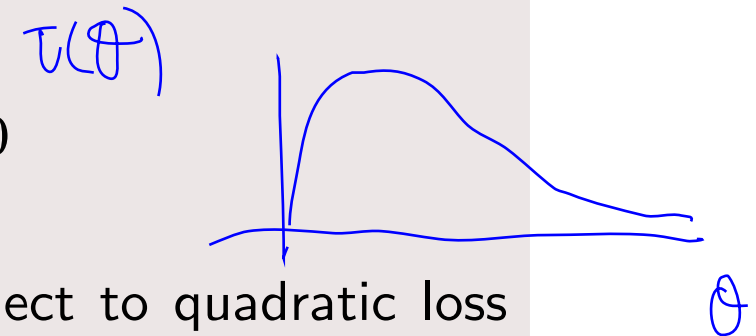
Suppose a single observation x is available from the uniform distribution with a density

$$f(x|\theta) = \frac{1}{\theta} I_{(x, \infty)}(\theta), \quad \theta > 0$$



The prior on θ has density:

$$\theta \sim \text{Gamma}(2, 1) \quad \tau(\theta) = \theta \exp(-\theta), \quad \theta > 0$$



- i) Find the Bayes estimator of θ with respect to quadratic loss $L(\theta, a) = (\theta - a)^2$.
- ii) Find the Bayes estimator of θ with respect to absolute value loss $L(\theta, a) = |\theta - a|$.

$$X_i \sim \text{Exp}(\theta)$$

Example 2.16

Suppose X_1, X_2, \dots, X_n have conditional joint density:

$$f_{X_i}(x_i) = \theta e^{-\theta x_i}$$

$$f_{X|\Theta}(x_1, x_2, \dots, x_n|\theta) = \theta^n e^{-\theta \sum_{i=1}^n x_i},$$

where $x_i > 0$ for $i = 1, \dots, n$ and $\theta > 0$. The prior density is given by $\tau(\theta) = \underline{k}e^{-k\theta}$ for $\theta > 0$, where $k > 0$ is a known constant i.e. the observations are exponentially distributed given θ , and the prior on θ is also exponential but with a different parameter.

- i) Calculate the posterior density of Θ given $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ $h(\theta|x)$
- ii) Find the Bayesian estimator of θ with respect to squared error loss.

$$h(\theta|x) = \frac{f(x|\theta)\tau(\theta)}{g(x)} \propto f(x|\theta)\tau(\theta)$$

Solution:

i) It turns out we do not need to calculate the normalising constant and can actually shortcut the solution. Looking at the joint density

$$f(\mathbf{x}|\theta)\tau(\theta) = k\theta^n e^{-\theta(\sum_{i=1}^n x_i + k)} \propto e^{-\theta(\sum x_i + k)}$$

we see that up to a normalising constant this is a

$$\theta|x \sim \text{Gamma}\left(n+1, \frac{1}{\sum_{i=1}^n x_i + k}\right)$$

density which must be the posterior $h(\theta|x)$.

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

then

$$f(x) \propto x^{\alpha-1} e^{-x/\beta}$$

$$n = \alpha - 1 \text{ or } \alpha = n + 1$$

$$\frac{1}{\beta} = \frac{\sum x_i + k}{1}$$

$$\text{or } \beta = \frac{1}{\sum x_i + k}$$

$$h(\theta|\mathbf{X}) \propto f(\mathbf{X}|\theta)L(\theta)$$

$$\theta|\mathbf{X} \sim \text{family}$$

ii) For a Bayes estimator with respect to quadratic loss, we have $\hat{\theta} = \mathbb{E}(\theta|\mathbf{X})$, and for a $\text{Gamma}(\alpha, \beta)$ density it is known that the expected value is equal to $\alpha\beta$, hence we get immediately

$$\hat{\theta} = \mathbb{E}(\theta|\mathbf{X}) = \frac{n+1}{\sum_{i=1}^n x_i + k} = \alpha\beta$$

We could also calculate this directly:

$$\hat{\theta} = \int_0^\infty \theta h(\theta|\mathbf{x}) d\theta = \frac{(\sum_{i=1}^n x_i + k)^{n+1}}{\Gamma(n+1)} \int_0^\infty \theta^{n+1} e^{-\theta(\sum_{i=1}^n x_i + k)} d\theta$$

and after changing variables: $\theta(\sum_{i=1}^n x_i + k) = y$ with $d\theta = \frac{dy}{(\sum_{i=1}^n x_i + k)}$
we can continue the evaluation:

$$\hat{\theta} = \frac{\int_0^\infty e^{-y} y^{n+1} dy}{\Gamma(n+1)(\sum_{i=1}^n x_i + k)} = \frac{\Gamma(n+2)}{\Gamma(n+1)(\sum_{i=1}^n x_i + k)} = \frac{n+1}{\sum_{i=1}^n x_i + k}$$

We arrive at the same answer but of course the shortcut solution is much simpler. Note however that the shortcut solution does not always work. In general, it is not always possible to guess the posterior from the joint density $f(\mathbf{x}|\theta)\tau(\theta)$.

$w(\theta|x)$ $\tau(\theta)$

If the posterior distribution belongs to the same family as the prior, the prior and posterior are then called conjugate distributions. The prior itself is called a conjugate prior and, in such cases, the shortcut approach works.

This is the reason that Bayesian modellers are often looking for conjugate priors trying to simplify the calculations. If such priors are difficult to find or are not reasonable suggestions for a prior in a particular situation, then one needs to resort to the full-scale Bayesian estimation instead of the shortcut approach.

Very often, there is no need to calculate explicitly the marginal $g(\mathbf{X})$ in the formula for the Bayes estimator. This is an important observation since the calculation of the integral that defines the marginal $g(\mathbf{X})$ may be difficult so it would be good if it could be avoided.

Using the symbol \propto to denote proportionality up to a constant between two functions, we can write:

$$h(\theta|\mathbf{X}) = \frac{f(\mathbf{X}|\theta)\tau(\theta)}{g(\mathbf{X})} \propto f(\mathbf{X}|\theta)\tau(\theta)$$

Hence, the shape is that posterior $h(\theta|\mathbf{X})$ (which conditionally on \mathbf{X} is a function of θ only), is determined with or without knowing $g(\mathbf{X})$ since the latter only serves to norm the conditional density to integrate to one.

We could guess the shape of the density $h(\theta|\mathbf{X})$ by just analysing the product $f(\mathbf{X}|\theta)\tau(\theta)$ as a function of θ .

In our previous example, we have

$$f(\mathbf{X}|\theta)\tau(\theta) = \theta^{\sum_{i=1}^n X_i + \alpha - 1} (1 - \theta)^{n - \sum_{i=1}^n X_i + \beta - 1}$$

This already identifies $h(\theta|\mathbf{X})$ as being

$$\theta|\mathbf{X} \sim \text{Beta}\left(\sum_{i=1}^n X_i + \alpha, n - \sum_{i=1}^n X_i + \beta\right)$$

distributed. But for any Beta distributed random variable with parameters a, b it is known that the expected values is equal to $\frac{a}{a+b}$. Hence we get immediately the Bayes estimator

$$\hat{\theta}_B = \mathbb{E}(\theta|\mathbf{X}) = \frac{\sum_{i=1}^n X_i + \alpha}{\alpha + \beta + n}$$

without the need to analyse and calculate the marginal $g(\mathbf{X})$.

Example 2.17

Let X_1, X_2, \dots, X_n be a random sample from the normal density with mean μ and variance 1. Consider estimating μ with a squared-error loss. Assume that the prior $\tau(\mu)$ is a normal density with mean μ_0 and variance 1.

Show that the Bayes estimator of μ is given by

$$\frac{\mu_0 + \sum_{i=1}^n X_i}{n + 1}.$$

Solution:

Let $\mathbf{X} = (X_1, \dots, X_n)$ be the random variables. Setting $\mu_0 = x_0$ for convenience of the notation, we can write:

$$h(\mu|\mathbf{X} = \mathbf{x}) \propto \exp\left(-\frac{1}{2} \sum_{i=0}^n (x_i - \mu)^2\right) \propto \exp\left(-\frac{n+1}{2} \left[\mu^2 - 2\mu \frac{\sum_{i=0}^n x_i}{n+1}\right]\right)$$

By completing the square with the expression that does not depend on μ , we get the following

$$h(\mu|\mathbf{X} = \mathbf{x}) \propto \exp\left(-\frac{n+1}{2} \left[\mu - \frac{\sum_{i=0}^n x_i}{n+1}\right]^2\right)$$

which implies that $h(\mu|\mathbf{X} = \mathbf{x})$, (being a density), must be the density of a

$$N\left(\frac{\sum_{i=0}^n x_i}{n+1}, \frac{1}{n+1}\right).$$

$$\pi_0 = \mu_0$$

Hence, the Bayes estimator with respect to the quadratic loss (being the posterior mean) would be

$$\hat{\mu}_{\text{bayes}} = \frac{1}{n+1} \sum_{i=0}^n x_i = \frac{1}{n+1} \left(\mu_0 + \sum_{i=1}^n x_i \right) = \frac{1}{n+1} \mu_0 + \frac{n}{n+1} \bar{X},$$

that is, the Bayes estimator is a convex combination of the mean of the prior and of \bar{X} . In this combination, the weight of the prior information diminishes quickly when the sample size increases. The same estimator is obtained with respect to absolute value loss.

Example 2.18

As part of a quality inspection program, five components are selected at random from a batch of components to be tested. From past experience, the parameter θ (the probability of failure), has a beta distribution with density

$$\theta \sim \text{Beta}(2, 5)$$

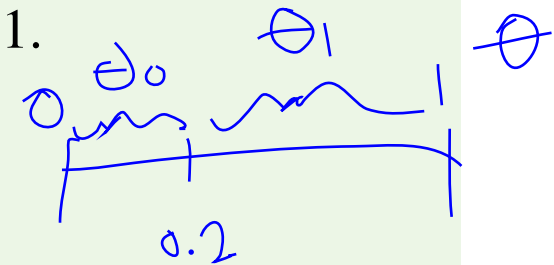
$$\tau(\theta) = 30\theta(1 - \theta)^4, \quad 0 \leq \theta \leq 1.$$

We wish to test the hypothesis

$$H_0 : \theta \leq 0.2 \quad \text{against} \quad H_1 : \theta > 0.2$$

using Bayesian hypothesis testing with a zero-one loss. Find the decision if:

- i) In a batch of five there were no failures found.
- ii) In a batch of five there was one failure found.



Solution:

i) $X \sim \text{Bin}(5, \theta)$. We have:

$$P(X = 0 | \theta) = (1 - \theta)^5$$

$$P(X = x | \theta) = \binom{5}{x} \theta^x (1 - \theta)^{5-x}$$

$$x = 0, 1, \dots, 5$$

which means that the posterior of θ given the sample is

$$\propto P(X=0|\theta) \pi(\theta)$$

$h(\theta | X = 0) \propto (1 - \theta)^5 \theta (1 - \theta)^4 = \theta (1 - \theta)^9$. Hence:

$$\theta | X=0 \sim \text{Beta}(2, 10)$$

$$h(\theta | X = 0) = 110 \theta (1 - \theta)^9$$

Note: $\frac{\Gamma(12)}{\Gamma(10)\Gamma(2)} = \frac{11!}{9!1!} = 110$

$$\frac{1}{B(2, 10)} \quad \Gamma(n) = (n-1)!$$

Then we get for the posterior probability given the sample:

$$P(\theta \leq 0.2 | X = 0) = \int_0^{0.2} 110 \theta (1 - \theta)^9 d\theta = 0.6779$$

$$> 1/2$$

accept H_0

and we accept H_0 since the above posterior probability is $> \frac{1}{2}$.

ii) Now:

$$P(X = 1|\theta) = 5(1 - \theta)^4\theta$$

$$\frac{1}{5} = 0.2$$

which implies that the posterior of θ given the sample is

$$h(\theta|X = 1) \propto \underbrace{(1 - \theta)^4\theta}_{(30)} \underbrace{(1 - \theta)^4\theta}_{(30)} = (1 - \theta)^8\theta^2.$$

$$\text{Beta}(3, 9)$$

Hence:

$$h(\theta|X = 1) = \frac{\Gamma(12)}{\Gamma(9)\Gamma(3)} (1 - \theta)^8\theta^2 = 495\theta^2(1 - \theta)^8$$

$$B(3, 9) = \frac{1}{8!2!} = 495$$

Then we get for the posterior probability given the sample:

$$P(\theta \leq 0.2 | X = 1) = \int_0^{0.2} 495\theta^2(1 - \theta)^8 d\theta = 0.3826 < \frac{1}{2}$$

reject H_0

$H_1: \theta > 0.2$

and we reject H_0 since the above posterior probability is $< \frac{1}{2}$.

Exercise 2.12 (at lecture)

In a sequence of consecutive years $1, 2, \dots, T$, an annual number of high-risk events is recorded by a bank. The random number N_t of high-risk events in a given year is modelled via $\text{Poisson}(\lambda)$ distribution. This gives a sequence of independent counts n_1, n_2, \dots, n_T . The prior on λ is $\text{Gamma}(a, b)$ with known $a > 0, b > 0$:

$$\tau(\lambda) = \frac{\lambda^{a-1} e^{-\lambda/b}}{\Gamma(a) b^a}, \lambda > 0.$$

- i) Determine the Bayesian estimator of the intensity λ with respect to quadratic loss.
- ii) Assume that the parameters of the prior are $a = 2, b = 2$. The bank claims that the yearly intensity λ is no more than $\bar{2}$. Within the last six years counts were 0, 2, 3, 3, 2, 2. Test the bank's claim via Bayesian testing with a zero-one loss.

Sample size

data
x's

$P(\lambda < 2 | N)$ compare 1/2

$T = 6$

$\sum n_i = 12$

$H_0: \lambda \leq 2$

$H_a: \lambda > 2$

2.5.12 What to do if a Bayes rule can not be found analytically (additional material)

Integration plays a significant role in analytic determination of the Bayesian estimators and tests. Integration may be difficult to get in closed form and numerical methods need to be applied in such situations. Simple Monte Carlo methods to calculate the integrals can always be applied to:

$$\int_{\Theta} \theta f(X|\theta) \tau(\theta) d\theta \quad \text{and} \quad \int_{\Theta} f(X|\theta) \tau(\theta) d\theta.$$

However, besides the simple Monte Carlo methods, there are more complicated Monte Carlo procedures which are specific and very useful in Bayesian inference. To motivate these procedures we first consider a simplified general example given in the following Lemma.

Lemma

Suppose we generate random variables by the following algorithm:

- i) Generate $Y \sim f_Y(y)$;
- ii) Generate $X \sim f_{X|Y}(x|Y)$.

Then $X \sim f_X(x)$.

Proof.

For the cumulative distribution function $F_X(x)$ we have:

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \mathbb{E}[F_{X|Y}(x|y)] \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^x f_{X|Y}(t|y) dt \right] f_Y(y) dy \\ &= \int_{-\infty}^x \left[\int_{-\infty}^{\infty} f_{X|Y}(t|y) f_Y(y) dy \right] dt \\ &= \int_{-\infty}^x \left[\int_{-\infty}^{\infty} f_{X,Y}(t,y) dy \right] dt \\ &= \int_{-\infty}^x f_X(t) dt. \end{aligned}$$

Hence, the random variable X generated by the algorithm has a density $f_X(x)$. □

The above Lemma tells us that if we wanted to calculate an expected value $\mathbb{E}[W(X)]$ for any function $W(X)$ with $\mathbb{E}[W^2(X)] < \infty$ then we can generate independently the sequence $(Y_1, X_1), (Y_2, X_2), \dots, (Y_m, X_m)$ for a specified large value m and then by the *Law of Large Numbers* we will have

$$\bar{W} \approx \mathbb{E}[W(X)].$$

The above simple observation can be generalized in the following algorithm of the **Gibbs sampler**. Let m be a positive integer and X_0 an initial value. Then for $i = 1, 2, \dots, m$:

- i) Generate $Y_i | X_{i-1} \sim f_{Y|X}(y|x)$
- ii) Generate $X_i | Y_i \sim f_{X|Y}(x|y)$.

It can be shown that $Y_i \xrightarrow{d} f_Y(y)$ and $X_i \xrightarrow{d} f_X(x)$ as $i \rightarrow \infty$. Therefore, intuitively, a convergence of the Gibbs sampler could be argued about in a manner similar to the Lemma.

The rigorous justification is slightly more involved. The reason is that the pairs

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_k, Y_k), (X_{k+1}, Y_{k+1})$$

generated by the Gibbs sampler are **not** generated independently. However, we need only the pair (X_k, Y_k) (and none of the previous $(k-1)$ pairs) to generate (X_{k+1}, Y_{k+1}) . Hence having a **Markov chain** and for it, under quite general conditions, the distribution stabilizes (reaches an equilibrium).

The application of the Gibbs sampler in Bayesian inference can help in overcoming one of the major obstacles of this inference, namely, the fact that the prior may not be precisely known. We can allow more freedom to ourselves by modeling the prior itself using another random variable. We get the so-called **hierarchical Bayes** model if we assume:

$$X|\theta \sim f(x|\theta), \quad \Theta|\gamma \sim q(\theta|\gamma), \quad \Gamma \sim \psi(\gamma)$$

with $q(.|.)$ and $\psi(.)$ known density functions. Here γ is called the *hyperparameter*. We keep in mind that $f(X|\theta)$ does **not** depend on γ . Keeping $g(.)$ as a generic notation for a density, we get using the Bayes formula:

$$g(\theta, \gamma|x) = \frac{f(x|\theta)q(\theta|\gamma)\psi(\gamma)}{g(x)}.$$

This conditional joint density is proportional to the product of known densities $f(x|\theta)q(\theta|\gamma)\psi(\gamma)$ hence $g(\theta|x, \gamma)$ and $g(\gamma|x, \theta)$ can (in principle) be determined. When there is no easy analytic way of doing this then there is the *Metropolis-Hastings* algorithm to help us simulate from the conditionals. The Metropolis-Hastings algorithm is discussed in a Bayesian statistics course and we will avoid discussing it here.

We can then start a Gibbs sampler with an initial value γ_0 as follows:

- i) $\Theta_i|X, \gamma_{i-1} \sim g(\theta|X, \gamma_{i-1})$
- ii) $\Gamma_i|X, \theta_i \sim g(\gamma|X, \theta_i)$.

In other words, we simulate from the *full conditionals* - the conditional distributions of each parameter given the other parameters and the data.

Taking sufficiently large repetitions the algorithm will converge under suitable conditions as follows:

$$\Theta_i \xrightarrow{d} h(\theta|X), \quad \Gamma_i \xrightarrow{d} g(\gamma|X) \quad \text{as} \quad i \rightarrow \infty.$$

Hence the simple arithmetic average of the Θ_i values (after possibly discarding some initial iterates before stabilization has occurred) will converge towards the Bayes estimator with respect to quadratic loss for the given hierarchical Bayes model.

In practice, we would generate the stream of values $(\theta_1, \gamma_1), (\theta_2, \gamma_2), \dots$. Then choosing large values of m and $B > m$, our Bayes estimate of θ will be the average

$$\frac{1}{B-m} \sum_{i=m+1}^B \theta_i.$$

Remark 2.4

The Gibbs sampler works fine when indeed the conditional distributions are completely known. The conditional distributions are often only known up to a (normalizing) proportionality constant.

Interestingly, the Gibbs sampler can still be used also in these cases but drawing from the conditional distribution is more involved. The best algorithm for this case: the **Metropolis-Hastings** algorithm. For details (separate course MATH5960 in Bayesian inference).

Here we present the essence of the algorithm. Suppose that a density $f(x)$ is only known up to a normalizing constant, i.e. $f(x) = c\tilde{f}(x)$ where \tilde{f} is known. Choose an arbitrary, completely known, so-called proposal density $u(x'|x)$. Let the t th generated data point is $x = X^t$. For this given x define the set of points

$$A_x = \{x' : \tilde{f}(x')u(x|x') < \tilde{f}(x)u(x'|x)\}.$$

- i) Generate a value x' randomly from $u(\mathbf{X}'|x)$.
- ii) If x' is not in A_x then we put $X^{t+1} = x'$ as the new simulated point. However, if x' is in A_x then perform a further randomization and accept x' with probability $\tilde{f}(x')u(x|x') / \tilde{f}(x)u(x'|x)$. If it is accepted, again put $X^{t+1} = x'$. Otherwise, put $X^{t+1} = x$.

The theory of the algorithm requires only mild conditions on $u(x'|x)$ to work but practice shows that in terms of computing time needed to run it. For generating a “well mixing” sequence of simulated values, the choice of $u(x'|x)$ needs to be done carefully.