2 Lecture 2: THE GENERAL INFERENCE PROBLEM

2.1 Measurement precision

The purpose in Statistical Inference is to draw relevant conclusions from given data. The conclusions could be about predicting further outcomes, evaluating risks of events, testing hypotheses, etc. In all cases, inference about the **population** is to be drawn but only a limited information contained in the sample is available. The most common situation in Statistics is the one in which an experiment has been repeatedly performed, the replicates being independent of one another. The possible results are real numbers that form a vector of observations $\mathbf{x}=(x_1, x_2, \ldots, x_n)$. The appropriate sample space is R^n. There is typically a "hidden" mechanism that generates the data and one is looking for suitable ways to identify it. **Models** will describe this mechanism in some simplistic but hopefully useful way. 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. Thus, the observations will be recorded 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.

If empirical observations were truly continuous, then, with probability one, no two observed responses would ever be identical. This fact will sometimes be used in our theoretical derivations. On the other hand, as pointed out above, the *real life empirical observations are indeed 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 got the vector of observations we can calculate the joint density (in the continuous case)

$$L_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1).f_{X_2}(x_2)\ldots f_{X_n}(x_n) \tag{1}$$

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 were **independent identically distributed (i.i.d.)** then all densities in (1) would be the same: $f_{X_1}(x) = f_{X_2}(x) = \cdots = f_{X_n}(x) = f(x)$. This is the most typical situation we will be discussing in our course. The need of **Statistical Inference** arises since typically, our knowledge about $f_{\mathbf{X}}(x_1, x_2, \ldots, x_n)$ 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. These models are usually about the shape of the density or of the cumulative distribution function. They should represent, as much as possible, the available prior theoretical knowledge about the

data generating mechanism. The suggestion of the set of models to be validated, is the first step, perhaps the most difficult: to think of a set of suitable model functions which might reasonably describe the data generating mechanism. The step usually involves a close *collaboration* between the statistician and the people who formulated the problem. We can view the statistical model as the triplet $(\mathcal{X}, \mathcal{P}, \Theta)$ where :

$-\mathcal{X}$ is the sample space (i.e.. the set of all possible realizations $\mathbf{X}$=$(X_1, X_2, \ldots, X_n)$

$-\mathcal{P}$ is a family of model functions $P_\theta(\mathbf{X})$ that depend on the unknown parameter θ;

$-\Theta$ is the set of possible θ-values, i.e.. the parameter space indexing the models.

2.3 Inference problem

The statistical inference problem can be formulated as follows:

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 make the inference. Note that we *always* have some a priori information. It could be more or less specific. When it is specific to such an extent that the shape of the distribution is known up to some *finite* number of parameters, we have to conduct *parametric* inference. In this case Θ is at most a (subset of) finite-dimensional Euclidean space. If Θ could only be specified as a certain *infinite dimensional function space*, we speak about *non-parametric* inference. Needless to say that nonparametric inference procedures are applicable in more general situations (which is good). However if they are applied for a situation where a parametric distributional shape gives adequate enough description of the data, nonparametric procedures may not be as efficient as s specifically tailored parametric procedure (which would be bad if the specific parametric model indeed holds).

The situation in practice is often more blurred: we may know that the populations is "close" to parametrically describable and yet "deviates a bit" from the parametric family. Going over in such cases to purely nonparametric approach would not properly address the situation since the idea about *relatively small* deviation from the baseline parametric family will be lost. The proper approach in such cases would be the *robustness* approach where we still keep the idea about the "ideal" parametric model but allow for *small* deviations from it. 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 in the course.

Another way to classify the Statistical Inference procedures is determined by the way we treat the *unknown* parameter θ. If we treat it as unknown but deterministic (fixed)

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.* The Bayesian approach allows us to introduce and effectively utilise any additional (prior) information about the model when such information is available. This information is entered in the model through the **prior distribution** over the set Θ of parameter values and reflects our prior belief about how likely any of the parameter values is before obtaining the information from the data. This topic will also be discussed shortly in our course.

2.4 Goals in Statistical Inference

Following are the most common goals in inference:

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 have been made, further information to our a priori knowledge about the set Θ has been added, so it becomes plausible that the true distribution belongs to a smaller family than it was originally postulated, i.e. it becomes clear that the unknown θ-value belongs to a *subset* of Θ. The problem of confidence set construction arises. It means, determining a (possibly small) plausible set of θ-values and clarifying the sense in which the set is plausible.

2.4.3 Hypotheses testing

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 complies with the observations or is in such a serious disarray that 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 any 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 practically relevant situations.

2.5 Statistical Decision Theoretic Approach to Inference

2.5.1 Introduction

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 criterion is formulated and a decision is considered optimal if it optimizes the criterion.

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 $\Theta \times \mathcal{A}$.

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 as an "intelligent opponent" who would behave "rationally". Also, 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 him the data $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ that has a distribution $L(\mathbf{X}|\theta)$ *depending* on θ. This is used by the statistician to work out his decision.

Definition 1. A (deterministic) *decision function* is a function $d : \mathcal{X} \to \mathcal{A}$ from the sample space to the set of actions.

There is a non-negative loss $L(\theta, d(\mathbf{X}))$ incurred by this action. Of course this is a random variable. Hence one defines the *RISK* $E_\theta L(\theta, d(\mathbf{X})) = R(\theta, d)$. For a fixed decision , this is a function (risk function) depending on θ. $R(\theta, d)$ is interpreted as the average loss of the statistician when the nature has a true state θ and the statistician uses decision *d.*

2.5.2 Examples.

Assume that a data vector $\mathbf{X} \sim f(\mathbf{X}, \theta)$.

a) Hypothesis testing. Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ where $\theta \in R^1$ is a parameter of interest. Let $\mathcal{A} = \{a_1, a_2\}, \Theta = R^1$. Here a_1 denotes the action "accept H_0 " whereas a_2 is the action "Reject H_0." Let

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

Define

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

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

Then

$$R(\theta, d) = EL(\theta, d(\mathbf{X})) = L(\theta, a_1)P_\theta(d(\mathbf{X}) = a_1) + L(\theta, a_2)P_\theta(d(\mathbf{X}) = a_2) =$$

$$\begin{cases} P_\theta(d(\mathbf{X}) = a_1) & \text{if } \theta > \theta_0, \\ P_\theta(d(\mathbf{X}) = a_2) & \text{if } \theta \leq \theta_0. \end{cases}$$

Hence

i) if $\theta \leq \theta_0 : R(\theta, d) = P_\theta($ reject $H_0) =$ Error of I type ,

ii) if $\theta > \theta_0 : R(\theta, d) = P_\theta($ accept $H_0) =$ Error of II type .

We will see later when studying optimality in hypothesis testing context that the set of deterministic decision rules D is not convex and it is difficult to develop a decent mathematical optimization theory over it. It has to be extended by including the so-called randomized decision rules if we want to formulate and solve such problems.

b) 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 $L(\theta, d(\mathbf{X})) = (\theta - d(\mathbf{X}))^2$ (*quadratic loss*) then the decision rule d (which we can call *estimator*) has a risk function

$$R(\theta, d) = E_\theta(d(\mathbf{X}) - \theta)^2 = MSE_\theta(d(\mathbf{X})).$$

2.5.3 Randomized decision rule

The set D of deterministic decision rules is not 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.

Definition 2. A rule δ which chooses d_i with probability $w_i, \sum w_i = 1$, is a randomized decision rule.

For the randomized decision rule δ we have:

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

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

2.5.4 Optimal decision rules

Given a game $(\Theta, \mathcal{A}, L)$ 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. The reason is that usually *uniformly* best decision rules (that minimize the risk uniformly for all θ-values) do not exist! This observations is easy to understand and explanation will be given during the lecture. It leads us to the following two ways out:

- *First way out.*

  a) 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 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.

  Definition 3. A decision rule d is *unbiased* if

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

  holds.

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

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

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

  It is obvious 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

a) Bayesian rule. Let us think of the θ-parameter now as of being a random variable with a given (known) prior density τ on Θ. Define the

Bayesian risk of the decision rule δ with respect to the prior τ:

$r(\tau, \delta) = E[R(T, \delta)] = \int_\Theta R(\theta, \delta)\tau(\theta)d\theta$ (here T is a random variable over Θ having a distribution with a density τ.)

Then the *Bayesian rule δ_τ with respect to the prior τ* is defined as:

$$r(\tau, \delta_\tau) = \inf_{\delta \in \mathcal{D}} r(\tau, \delta)$$

Sometimes a Bayesian rule may not exist and then we could ask for an ϵ-Bayes rule. For $\epsilon > 0$, this is any rule $\delta_{\epsilon\tau}$ that satisfies $r(\tau, \delta_{\epsilon\tau}) \leq \inf_{\delta \in \mathcal{D}} r(\tau, \delta) + \epsilon$.

b) Minimax rule. Instead of considering uniformly best rules (that usually do not exist), one can alternatively consider rules that minimize the *supremum* of the values of the risk over the set Θ. This would mean safeguarding against the worst possible performance.

The value $\sup_{\theta \in \Theta} R(\theta, \delta)$ is called *minimax risk* of the decision rule δ. Then the rule δ^* is called *minimax* in the set $\mathcal{D}$ if

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

Note that again, like in the Bayesian case, even if the minimax value is finite there may not be a minimax decision rule. Hence again 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 too pessimistic a strategy but experience shows that in most cases minimax rules are good rules.

2.5.6 Least favorable prior distribution

After the above definitions have been given, one can easily define the least favorable distribution (i.e. least favorable prior τ^* over the set Θ) as:

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

It indeed deserves its name. From the above definition we see that 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 of the decision rules in the case of finite Θ

Definition 4. A set $A \subset R^k$ is *convex* if for all vectors $\vec{x} = (x_1, x_2, \ldots, x_k)'$ and $\vec{y} = (y_1, y_2, \ldots, y_k)'$ in A and all $\alpha \in [0,1] : \alpha\vec{x} + (1-\alpha)\vec{y} \in$ A.

Assume that Θ has k elements all together. Now let us define the *risk set* of a set D of decision rules. This is simply 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 R^k and by "moving"d within D, we get a set of such k-dimensional vectors.

Theorem 2.1. *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: It is easy to see that if $\vec{y}$ and $\vec{y}'$ are the risk points of the decision rules δ and $\delta' \in D$, correspondingly, then any point in the form $\vec{z} = \alpha\vec{y} + (1-\alpha)\vec{y}'$ corresponds to (is the risk point of) the randomized decision rule $\delta_\alpha \in \mathcal{D}$ that chooses the rule δ with probability α and the rule δ' with probability $(1-\alpha)$. Hence any such $\vec{z}$ belongs to the risk set of $\mathcal{D}$.

Remark: In fact, 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).

How to illustrate Bayes rules: Since $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ then the prior $\tau = (p_1, p_2, .., p_k)$ 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 w.r. to the prior τ is $r(\tau,\delta) = \sum_{i=1}^k p_i R(\theta_i,\delta)$. All points $\vec{y}$ in the risk set, corresponding to certain rules δ^* for which $\sum_{i=1}^k p_i y_i = r(\tau,\delta^*) =$ 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. (See illustration handed out).

In a similar way, the minimax rule can be illustrated in the case of finite Θ at least (See illustration).

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 $L(\theta_1, a_1) = L(\theta_2, a_2) = 0, L(\theta_1, a_2) = 1, L(\theta_2, a_1) = 3$. The distributions of X are tabulated as follows:

$$\left| \begin{array}{cccc} x & 0 & 1 & 2 \\ P(x|\theta_1) & .81 & .18 & .01 \end{array} \right| \left| \begin{array}{cccc} x & 0 & 1 & 2 \\ P(x|\theta_2) & .25 & .5 & .25 \end{array} \right|$$

The decision problem formulated above can be interpreted as an attempt by the statistician to guess the right state of nature. If his guess is correct, he does not lose anything but if he is wrong, he either loses \$1 or \$3 depending on the type of error he has made. In his guess he is supported by one observation X that has a different distribution under θ_1 and under θ_2.

Now, consider all possible non-randomized decision rules based on one observation:

$$\left[\begin{array}{cccccccc} x & d_1(x) & d_2(x) & d_3(x) & d_4(x) & d_5(x) & d_6(x) & d_7(x) & d_8(x) \\ 0 & a_1 & a_1 & a_1 & a_1 & a_2 & a_2 & a_2 & a_2 \\ 1 & a_1 & a_1 & a_2 & a_2 & a_1 & a_1 & a_2 & a_2 \\ 2 & a_1 & a_2 & a_1 & a_2 & a_1 & a_2 & a_1 & a_2 \end{array} \right]$$

The following questions have to be answered:

a) sketch the risk set of all randomized rules generated by $d_1, d_2, .., d_8$;

b) find the minimax rule δ^*(in $\mathcal{D}$) and compute its risk;

c) for what prior is δ^* a Bayes rule w.r. to that prior (i.e., what is the least favorable distribution) ;

d) What is the Bayes rule for the prior $\{1/3, 2/3\}$ over $\{\theta_1, \theta_2\}$? What is the value of its Bayes risk?

Solution : to be discussed at lecture.

2.5.9 Fundamental Lemma

Lemma 2.2. *If τ^* is a prior on Θ and the Bayes rule δ_{τ^*} has a constant risk w.r. to θ(i.e. if $R(\theta, \delta_{\tau^*}) = c_0$ for all $\theta \in \Theta$) then:*

a) δ_{τ^*} is minimax;

b) τ^* is the least favorable distribution.

Proof: a) Let us compute the minimax risk of δ_{τ^*} and compare it to the minimax risk of any other rule δ:

$c_0 = \sup_{\theta\in\Theta} R(\theta, \delta_{\tau^*}) =$ (since constant for all θ) $= \int_\Theta R(\theta, \delta_{\tau^*})\tau^*(\theta)d\theta \leq$ (since δ_{τ^*} Bayes w.r. to τ^*) $\leq \int_\Theta R(\theta, \delta)\tau^*(\theta)d\theta \leq \sup_{\theta\in\Theta} R(\theta, \delta)$,

which means that δ_{τ^*} is minimax.

b) Now take any other prior τ. Then we have:

$\inf_\delta r(\tau,\delta) = \inf_\delta \int_\Theta R(\theta,\delta)\tau(\theta)d\theta \leq \int_\Theta R(\theta,\delta_{\tau^*})\tau(\theta)d\theta =$ (since $R(\theta,\delta_{\tau^*})$ constant and

$$\int_\Theta \tau^*(\theta)d\theta = \int_\Theta \tau(\theta)d\theta = 1) = \int_\Theta R(\theta,\delta_{\tau^*})\tau^*(\theta)d\theta = r(\tau^*,\delta_{\tau^*}),$$

hence τ^* is least favorable.

Remark. The above lemma gives a hint how to find minimax estimators. The minimax estimators turn out to be (special) Bayes estimators with respect to the least favorable prior. First we can obtain the general form of the Bayes estimator with respect to ANY given prior. Then in order to get the minimax estimator we should choose such 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 seen in 2.5.9, as a device to be utilized in the search for minimax rules. It turns out that given the prior and the observations $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ we can find the Bayes rule point-wise (i.e. for any given $\mathbf{X}$=$\mathbf{x}$) by solving a certain minimization problem. In many practically relevant cases the solution can even be given in a closed form.

To see how, let us first introduce the following notation:

$-f(\mathbf{X}|\theta)$ is the conditional density of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ given θ;

$-\tau(\theta)$ is the prior density on θ;

$-g(\mathbf{X})$ is the marginal density of $\mathbf{X}$, i.e. $g(\mathbf{X}) = \int_\Theta f(\mathbf{X}|\theta)\tau(\theta)d\theta$;

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

$-f(\mathbf{X},\theta)$ is the joint density of X and θ. Note: $f(\mathbf{X},\theta) = f(X|\theta)\tau(\theta) = h(\theta|\mathbf{X})g(\mathbf{X})$

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

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

Theorem 2.3. *Define for $X \in \mathcal{X}, a \in \mathcal{A}$ and for a given prior τ, define:*

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

(remember that $L(.,.)$ was the loss function)).

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

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

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

Proof: For any decision rule δ we have:

$$r(\tau, \delta) = \int_\Theta R(\theta, \delta)\tau(\theta)d\theta = \int_\Theta [\int_{\mathcal{X}} L(\theta, \delta(\mathbf{X}))f(\mathbf{X}|\theta)d\mathbf{X}]\tau(\theta)d\theta =$$

$$\int_{\mathcal{X}} [\int_\Theta L(\theta, \delta(\mathbf{X}))h(\theta|\mathbf{X})d\theta]g(\mathbf{X})d\mathbf{X} =$$

$$\int_{\mathcal{X}} Q(\mathbf{X}, \delta(\mathbf{X}))g(\mathbf{X})d\mathbf{X}$$

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

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

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

Now let us apply the above general theorem to the cases of estimation and hypothesis testing.

Theorem 2.4. *(case of estimation). Consider a point estimation problem for a real-valued parameter θ. The prior over θ is denoted by τ. Then:*
a) for a squared error loss $L(\theta, a) = (\theta - a)^2 : \delta_\tau(\boldsymbol{X}) = E(\theta|\boldsymbol{X}) = \int_\Theta \theta h(\theta|\boldsymbol{X})d\theta$;
b) for an absolute error loss $L(\theta, a) = |\theta - a| : \delta_\tau(\boldsymbol{X}) =$ median of $h(\theta|\boldsymbol{X})$.

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

Theorem 2.5. *(case of Bayesian hypothesis testing with a generalized $0-1$ loss). Let $\Theta = \Theta_0 \cup \Theta_1, \Theta_0 \cap \Theta_1 = \oslash$ and we are testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$. A prior τ is given on Θ. Two non-randomized actions a_0(accept H_0) and a_1(reject H_0) are possible and the losses are given by: $L(\theta, a_0) = \begin{cases} 0 \text{ if } \theta \in \Theta_0 \\ c_2 \text{ if } \theta \in \Theta_1 \end{cases}$, $L(\theta, a_1) = \begin{cases} c_1 \text{ if } \theta \in \Theta_0 \\ 0 \text{ if } \theta \in \Theta_1 \end{cases}$. Then the test $\varphi^* = \begin{cases} \text{Reject } H_0 \text{ if } P(\theta \in \Theta_0 | \boldsymbol{X}) < c_2/(c_1 + c_2) \\ \text{Accept } H_0 \text{ if } P(\theta \in \Theta_0 | \boldsymbol{X}) > c_2/(c_1 + c_2) \end{cases}$ is a Bayesian rule (Bayesian test) for the above testing problem w.r. to the prior τ.*

Proof: According to the General Theorem, we have to compare $Q(\mathbf{X}, a_0)$ and $Q(\mathbf{X}, a_1)$ and take as our action the one that gives the smaller value. Now:

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

$$Q(\mathbf{X}, a_1) = \int_{\Theta} L(\theta, a_1) h(\theta|\mathbf{X}) d\theta = \int_{\Theta_0} c_1 h(\theta |\ \mathbf{X}) d\theta = c_1 P(\theta \in \Theta_0 |\ \mathbf{X})$$

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

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

2.5.11 Example about calculating minimax estimator for the probability of success in the Bernoulli experiment.

Let given θ, the distribution of each $X_i, i = 1, 2, \ldots, n$ be Bernoulli with parameter θ, i.e. $f(\mathbf{X}|\theta) = \theta^{\Sigma X_i}(1 - \theta)^{n - \Sigma X_i}$ and assume a beta-prior τ for the (random variable) θ over $(0, 1) : \tau(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}.I_{(0,1)}(\theta)$.

Hereby $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1 - x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function (with the obvious property $B(\alpha, \beta) = \frac{\alpha-1}{\alpha+\beta-1}.B(\alpha - 1, \beta)$).

For the above prior we can find easily

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

(which is again a beta-density). Hence the Bayesian estimator is

$$\hat{\theta}_\tau = \int_0^1 \theta h(\theta|\mathbf{X}) d\theta = \frac{B(\Sigma X_i + \alpha + 1, n - \Sigma X_i + \beta)}{B(\Sigma X_i + \alpha, n - \Sigma X_i + \beta)} = \text{by the property of the beta-function} =$$

$$= \frac{\Sigma X_i + \alpha}{\alpha + \beta + n}.$$

The above derivation holds for any beta prior Beta(α, β).

Very improtant note: We do NOT in fact need to calculate the normalization $g(\mathbf{X})$ explicitly. Once we realized that $f(\mathbf{X}|\theta)\tau(\theta) \propto \theta^{\Sigma X_i + \alpha - 1}(1 - \theta)^{n - \Sigma X_i + \beta - 1}$ (hence up to a norming constant we are dealing with a Beta density with parameters $\Sigma X_i + \alpha$ and $n - \Sigma X_i + \beta$) then the conditional $h(\theta|\mathbf{X})$, being a density, MUST be the Beta density with these parameters. Knowing how the expected value of the beta distribution depends on the paramteres of the distribution we can immediately get the Bayesian estimator as $\frac{\Sigma X_i + \alpha}{\alpha + \beta + n}$. Such type of arguments are routinely used in Bayesian inference and save a lot of unnecessary computations of norming constants! See the tutorial problems for further illustrations of this approach.

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

$$R(\theta,\hat{\theta}_\tau)=E(\hat{\theta}_\tau-\theta)^2=\mathrm{Var}_\theta(\hat{\theta}_\tau)+(\theta-E_\theta\hat{\theta}_\tau)^2=\frac{n\theta(1-\theta)}{(n+\alpha+\beta)^2}+(\frac{n\theta+\alpha}{\alpha+\beta+n}-\theta)^2=\ldots=$$

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

For this risk not to depend on θ, it has to hold: $\begin{cases}(\alpha+\beta)^2=n\\ 2\alpha(\alpha+\beta)=n\end{cases}$

The solution to this system is $\alpha=\beta=\sqrt{n}/2$. Hence (see 2.5.9) the minimax estimator of θ is

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

2.5.12 What to do if a Bayes rule can not be found analytically.

As seen, integration techniques play a significant role in analytic determination of the Bayesian estimators and tests. Integration may be difficult to get in closed form in most of the cases and some numerical methods need to be applied in such situations. Simple Monte Carlo methods to calculate the integrals $\int_\Theta \theta f(X|\theta)\tau(\theta)d\theta$ and $\int_\Theta f(X|\theta)\tau(\theta)d\theta$ can always be applied. 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:

$$F_X(x)=P(X\le x)=E[F_{X|Y}(x|y)]=\int_{-\infty}^{\infty}[\int_{-\infty}^{x}f_{X|Y}(t|y)dt]f_Y(y)dy=$$

$$\int_{-\infty}^{x}[\int_{-\infty}^{\infty}f_{X|Y}(t|y)f_Y(y)dy]dt=\int_{-\infty}^{x}[\int_{-\infty}^{\infty}f_{X,Y}(t,y)dy]dt=\int_{-\infty}^{x}f_X(t)dt.$$

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 $E[W(X)]$ for any function $W(X)$ with $E[W^2(X)]<\infty$ then we can generate independently the

sequence $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_m, X_m)$ for a specified large value m and then by the *Law of Large Numbers* we will have

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

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 $i = 1, 2, \ldots, m$:

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

In more advanced texts, it can be shown that $Y_i \to^d f_Y(y)$ and $X_i \to^d f_X(x)$ as $i \to \infty$. Therefore, intuitively, a convergence of the Gibbs sampler could be argued about in a manner similar to the Lemma.

The rigorous justification of the latter convergence is in fact a bit more involved. Indeed the Gibbs sampler algorithm is similar but not quite the same as the one in the Lemma. Let us examine the pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_k, Y_k), (X_{k+1}, Y_{k+1})$ generated by the Gibbs sampler. On one hand, they are not generated independently, on the other hand, 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}). With other words, given the present, the future of the sequence is independent of the past. Such a sequence is called 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 in fact 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:

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

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) $\Gamma_i|X, \theta_i \sim g(\gamma|X, \theta_i)$.

(With 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: $\Theta_i \to^d h(\theta|X), \Gamma_i \to^d g(\gamma|X)$ as $i \to \infty$. Hence the simple arithmetic

average of the Θ_i values (after possibly discarding some initial iterates before stabilization has occured) 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 $(\theta_1, \gamma_1), (\theta_2, \gamma_2), \ldots$. Then choosing large values of m and $B > m$, our Bayes estimate of θ will be the average

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