8 Decision theory

Given an outcome x of an experiment modeled by $X \sim P_\theta$, with $\theta \in \Omega$ unknown, we can make a *decision* $d = \delta(x)$ based on x. The possible values that d can take are D. The consequence of making a decision is measured by a loss function $L(\theta, d)$. Viewing $\delta(X)$ as a random vector, taking it's values in the space D, and $L(\theta, \delta(X))$ as a random variable we can use the risk

$$R(\theta, \delta) \quad = \quad E_\theta L(\theta, \delta(X)),$$

as a measure of the loss made, when basing the decision on the decision rule δ. One can then try to find the decision δ that minimizes this overall loss, i.e.

$$\hat{\delta} \quad = \quad \mathrm{argmin}_\delta R(\theta, \delta),$$

and let this be the optimal decision for the problem.

As an example let $X_1, \ldots, X_n$ be a sample from a one parameter family of distributions $\{P_\theta : \theta \in \Omega \subset \mathbb{R}\}$. The two main classes of decision problems in inference theory are that of point estimation and hypothesis testing:

1. *Point estimation.* Assume that we want to make a decision on the value of the true unknown parameter θ. Then the decision rule should be a real valued function δ defined on the outcome space $\mathcal{X}$. A feasible loss function could take it's minimum value at the true value i.e. when $d = \theta$, for convenience we can let this loss value be 0, and have larger loss the farther from θ we are, for instance by letting $L(\theta, d)$ be a function of $|d - \theta|$.

2. *Hypothesis testing.* Assume that θ_0 is an interesting value for the application at hand, for instance it can be a critical value of some sort or it can measure some causal effect as in regression problems. Assume that we want to make a decision on whether θ is bigger than or smaller than θ_0. Then the decision function δ is a function defined on $\mathcal{X}$ that takes it's values in the set $\{d_0, d_1\}$ where d_0 means that we decide that $\theta > \theta_0$ and d_1 means that we decide that $\theta \leq \theta_0$.

3. *Multiple decision problems.* Assume two values $\theta_0 \leq \theta_1$ are given, and assume there are three possible decisions: $d_0 : \theta \leq \theta_0, d_1 : \theta_0 \leq \theta \leq \theta_1$ and $d_2 : \theta > \theta_1$.

Thus it seems that a statistical problem can be specified using three elements:

1. The family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ needs to be specified. On one hand this is a modeling problem, in which one needs to use a model that is appropriate for the application at hand. On the other hand one also needs to consider the tractability of the model, i.e. we need to be able to work out solutions for equations in the chosen model.

2. The set D of possible decisions. For instance in point estimation problems it can be a real number, in hypothesis testing problems it can be modeled as $\{0, 1\}$.

3. The form of the loss function L.

Example 39 *Assume we have s Normal distribution $N(\xi_i, \sigma^2), i = 1, \ldots, s$ with observations $X_{ij}, i = 1, \ldots, s, j = 1, \ldots, n_i$. We would like to investigate whether the means are the same.*

When $s = 2$ the possible decisions are

$$\begin{aligned} d_0 : &\quad |\xi_1 - \xi_2| \leq \Delta, \\ d_1 : &\quad \xi_2 > \xi_1 + \Delta, \\ d_2 : &\quad \xi_2 < \xi_1 - \Delta, \end{aligned}$$

for some fixed Δ, chosen in an appropriate way.

For general s possible decisions are $d_0, d_1, \ldots, d_s$ where

$$\begin{aligned} d_0 : &\quad \max_{i,j} |\xi_i - \xi_j| \leq \Delta, \\ d_k : &\quad \max_{i,j} |\xi_i - \xi_j| > \Delta \textit{ with } \max_i \xi_i = \xi_k, \end{aligned}$$

for $k = 1, \ldots, s$. □

When the loss function satisfies that for a fixed θ there is only one value d for which $L(\theta, d) = 0$, we can call the problem an *action problem.* This is however not always the case.

Example 40 *Assume $X_1, \ldots, X_n$ are i.i.d data from $N(\xi, \sigma^2)$. Assume we want to make a interval estimate of the mean ξ. Then the decision function has as it's values intervals $\delta(X) = (l(X), u(X))$. An appropriate loss function could take the value 0 if $\xi \in \delta(X)$ and otherwise could depend on the distance from ξ to the the interval $\delta(X)$, so for $d \in D$ an interval*

$$L(\xi, d) \quad = \quad \rho(||\xi - d||)1\{\xi \notin d\},$$

for some monotone function ρ, where $|| \cdot ||$ denotes the smallest distance from ξ to d. Then given ξ there are several intervals d that give $L(\xi, d) = 0$, i.e. there are several decisions d that have loss zero or are correct. □

Examples of loss functions are:

(i) For point estimation of the estimand $g(\theta)$ one can e.g. use weighted quadratic loss

$$L(\theta, d) \quad = \quad w(\theta)(g(\theta) - d)^2,$$

with some specified weight function w, or some other convex loss functions.

(ii) For hypothesis testing problems when one makes a decision on d_0 or d_1 using a decision function δ, let $\Omega = \omega_0 \cup \omega_1$ be the partition of the parameter space defining the two hypothesis we choose between. So d_0 is the correct decision when $\theta \in \omega_0$ and d_1 is the correct decision when $\theta \in \omega_1$. Then the loss function

$$L(\theta, d) \quad = \quad a_0 1\{d = d_0\}1\{\theta \in \omega_1\} + a_1 1\{d = d_1\}1\{\theta \in \omega_0\},$$

is appropriate. The risk becomes

$$R(\theta, \delta) \quad = \quad a_0 P_\theta(\delta(X) = d_0)1\{\theta \in \omega_1\} + a_1 P_\theta(\delta(X) = d_1)1\{\theta \in \omega_0\}.$$

(iii) For the interval testing problem of a real valued parameter θ the loss was chosen as

$$L(\xi, d) \quad = \quad \rho(||\xi - d||)1\{\xi \notin d\},$$

where d is an interval in $\mathbb{R}$.

There are possible generalizations of the definition of decision rules that we have made so far. One such generalization is to allow for decision rules that are random, so that are chosen according to some probability measure. Then for each outcome x, the decision $\delta(x)$ is a random element in D, under some probability measure on D. An example of this is for point estimation problems when one allows for randomized estimators. Another example is for randomized tests, which we will treat in the sequel. Another generalization is for sequential decisions, when one allows the decision rule to depend on the sample size explicitly. This typically comes up when one wants to control the risk and would like to know the required number of data points n to obtain a risk within the limits. However, the number n may depend on the parameters in the distribution and is therefore not possible to calculate. One can then use sequential decisions on whether to continue with the experiment or not, since one can iteratively estimate the unknown parameters and calculate n, in some smart way. Sequential decision rules for optimal stopping will not be treated.

8.1 Optimal procedures

We stated that the optimal decision should be defined as

$$\hat{\delta} \quad = \quad \text{argmin}_\delta R(\theta, \delta).$$

However, the optimal decision might not exist, since two different decisions δ_1 and δ_2 could have risk functions, seen as functions of θ, that intersect.

To remedy this one can restrict the possible decision rules to satisfy certain impartialy conditions such as unbiasedness or invariance/equivariance; these are appropriate for different classes of distributions, the first for exponential familes, the second for transformation group families.

$a)$ For invariance decision problems let $(G, \bar{G}, G^*)$ be groups acting on $\mathcal{X}, \mathcal{P}, D$ respectively, and satisfying the usual invariance assumptions. Let L be an invariant measure. Then a decision rule is called equivariant if

$$g^*\delta(x) \quad = \quad \delta(gx),$$

and invariant if

$$\delta(x) \quad = \quad \delta(gx).$$

for every $g \in G$ and corresponding $g^* \in G^*$. In the latter case the assumption of invariance on the loss function is changed to $L(\bar{g}\theta, d) = L(\theta, d)$ (which formally is equivalent to the previous definition $L(\bar{g}\theta, g^*d) = L(\theta, d)$ if we let $g^* = id$ be the identity map, for every $g \in G$). We have seen that equivariant decision rules come up in point estimation problems, and also in interval estimation problems. Invariant estimation problems occur in hypothesis testing: As an example consider a two sample test for the means ξ_1, ξ_2 in a location family. Making the transformations $\xi_i' = \xi_i + g$, with $g \in \mathbb{R}$, should leave a reasonable decision rule δ unchanged, and a reasonable loss function L should satisfy $L(\bar{g}\theta, d) = L(\theta, d)$.

b) For unbiasedness assume that for each θ there is a unique correct decision d and that each decision d is correct for some θ. We make also the assumption that if a decision d is correct for two different values θ_1, θ_2 then $L(\theta_1, d) = L(\theta_2, d)$, so that then loss is the same. That means the loss will only depend on the actual decision taken d' and on the correct decision d, so then $L(\theta, d') =: \tilde{L}(d, d')$, introducing $\tilde{L}$ to avoid abuse of nation. One can then call a decision rule δ L-unbiased if

$$E_\theta(\tilde{L}(d, \delta(X))) \quad \leq \quad E_\theta(\tilde{L}(d', \delta(X))).$$

We can state this a bit more generally, defining δ to be L-unbiased if

$$E_\theta L(\theta, \delta(X)) \quad \leq \quad E_\theta L(\theta', \delta(X)),$$

for every θ, θ'.

Example 41 *For an interval estimation problem the decision rule $\delta(X) = (l(X), u(X))$ is L-unbiased for*

$$L(\theta, d) \quad = \quad 1\{\theta \notin d\},$$

if $P_\theta(\theta \in \delta(X)) \geq P_\theta(\theta' \in \delta(X))$. □

Example 42 *For a hypothesis testing problem assume d_0, d_1 are possible decisions and ω_0 and ω_1 are the corresponding set of θ values (i.e. the values that make the decision correct). Let the loss be zero for a correct decision, and a_0 when wrong decision is taken and true parameter is in ω_0 and a_1 for a wrong decision and true parameter in ω_1. Then*

$$E_\theta L(\theta', \delta(X)) \quad = \quad a_0 P_\theta(\delta(X) = d_0)\, 1\{\theta' \in \omega_1\} + a_1 P_\theta(\delta(X) = d_1)\, 1\{\theta' \in \omega_0\}.$$

Then L-unbiasedness of a decision δ means

$$\begin{aligned} & a_0 P_\theta(\delta(X) = d_0)\, 1\{\theta \in \omega_1\} + a_1 P_\theta(\delta(X) = d_1)\, 1\{\theta \in \omega_0\} \\ \leq \quad & a_0 P_\theta(\delta(X) = d_0)\, 1\{\theta' \in \omega_1\} + a_1 P_\theta(\delta(X) = d_1)\, 1\{\theta' \in \omega_0\}. \end{aligned}$$

This translates to

$$\begin{aligned} a_0 P_\theta(\delta(X)=d_0) &\geq a_1 P_\theta(\delta(X)=d_1) \quad \textit{if } \theta' \in \omega_1, \\ a_0 P_\theta(\delta(X)=d_0) &\leq a_1 P_\theta(\delta(X)=d_1) \quad \textit{if } \theta' \in \omega_0. \end{aligned}$$

Using $P_\theta(\delta(X)=d_0)+P_\theta(\delta(X)=d_1)=1$, *this becomes*

$$\begin{aligned} P_\theta(\delta(X)=d_1) &\leq \frac{a_0}{a_0+a_1} \quad \textit{if } \theta' \in \omega_1, \\ P_\theta(\delta(X)=d_1) &\geq \frac{a_0}{a_0+a_1} \quad \textit{if } \theta' \in \omega_0. \end{aligned}$$

□

An alternative approach is to use overall measures over the parameters space, of the risk for a decision, such as the Bayes risk

$$\int R(\theta,\delta)\rho(\delta)\,d\theta,$$

where ρ is some prior density. A decision rule δ_ρ that minimizes this is called a Bayes solution. Another overall risk is the maximum risk

$$\max_{\theta\in\Omega} R(\theta,\delta)$$

A decision rule that minimizes this is called minimax.

8.2 Likelihood based decision rules

Assume that X takes a countable number of values $x_1, x_2, \ldots$ with probability $P_\theta(x) = P_\theta(X=x)$, and $\theta \in \Omega$ unknown. The likelihood is the function

$$L_x(\theta) \;=\; P_\theta(x)$$

defined on Ω. One can now base a decision rule δ on the outcomes of the likelihood. Since a large value on the likelihood makes the observed value more probable and one therefore is interested in maximizing the likelihood over Ω one talks of gain functions instead of loss functions.

Assume the decision function δ takes it's values in a countable set $D = \{d_1, d_2, \ldots\}$, where each decision can be stated as $d_k : \theta \in A_k$, for some sets A_k. A reasonable gain function g should then satisfy

$$g(\theta,d) \;=\; a(\theta)1\{d \text{ is a correct decision}\}$$

where a is some positive function, and d is the correct decision. One can then weight the likelihood $L_x(\theta)$ with the gain $g(\theta)$ when θ is the true value. Then one maximizes

$a(\theta)L_x(\theta)$ and selects a decision that would be true if the maximizing value would be the true value of θ.

In point estimation one assumes that the gain $a(\theta)$ does not depend on θ and then one maximizes $L_x(\theta)$ over Ω, which leads to the maximum likelihood estimation. For hypothesis testing assume d_0 and d_1 are the possible decisions and ω_0 and ω_1 are the sets of θ values that define d_0 and d_1 respectively. Assume that the gain is a_0 when $\theta \in \omega_0$ and the decision is correct and the gain is a_1 when $\theta \in \omega_1$ and the decision is correct. Then one can take the decision d_0 if

$$a_0 \sup_{\theta\in\omega_0} L_x(\theta) \;>\; a_1 \sup_{\theta\in\omega_1} L_x(\theta),$$

and decision d_1 if

$$a_0 \sup_{\theta\in\omega_0} L_x(\theta) \;<\; a_1 \sup_{\theta\in\omega_1} L_x(\theta).$$

Or, equivalently, make the decision d_0 if

$$\frac{\sup_{\theta\in\omega_0} L_x(\theta)}{\sup_{\theta\in\omega_1} L_x(\theta)} \;>\; \frac{a_1}{a_0},$$

and d_1 for the opposite inequality. This leads to likelihood ratio tests.

8.3 Admissible decisions and complete classes

Sometimes the decision rule obtained under some impartiality rule such as unbiasedness or equivariance can be less than satisfactory, for the reason that there might exist an estimator that does not satisfy the impartiality rule but that is preferable nevertheless.

If a decision procedure δ is dominated by another decision procedure δ', in the sense that

$$R(\theta,\delta') \;\leq\; R(\theta,\delta),$$

for all θ, with strict inequality for at least one θ, then δ is called inadmissible. If there is no dominating decision rule δ is called admissible.

A class $\mathcal{C} = \{\delta\}$ of decision rules is called complete if for every $\delta' \notin \mathcal{C}$ there is a $\delta \in \mathcal{C}$ that dominates δ'. This means that $\mathcal{C}$ is a complete class if the decision rules in $\mathcal{C}$ dominate all decision rules outside of $\mathcal{C}$. A complete class is called minimal if it does not contain a complete (proper) subclass.

A class $\mathcal{C}$ is called essentially complete if for every decision procedure δ there is a $\delta' \in \mathcal{C}$ such that $R(\theta,\delta') \leq R(\theta,\delta)$ for all θ, so without the assumption of strict inequality for some θ. Clearly a complete class is essentially complete. For an essentially complete class one could have the situation that $\delta_1 \in \mathcal{C}, \delta_2 \notin \mathcal{C}$ and $R(\theta,\delta_1) = R(\theta,\delta_2)$ for all θ; this is not possible for a complete class.

Completeness and essential completeness differ for decision rules that are equivalent (that have the same risk functions): If δ is a decision rule in a minimal complete class $\mathcal{C}$, then any equivalent decision rule must also be in $\mathcal{C}$. If $\mathcal{C}$ is a minimal essentially complete class, then it contains only one representative out of each set of equivalent decision rules.

Minimal essentially complete classes provide the greatest possible reduction of a decision problem: If we take two arbitrary decision rules $\delta_1, \delta_2 \in \mathcal{C}$ neither is uniformly better that the other, each is better than the other on some parts of Ω.

8.4 Uniformly most powerful tests

Assume we have the problem of hypothesis testing. Let $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ be a family of distributions for the random variable X. Divide the parameter set into two disjoint sets $\Omega = \Omega_H \cup \Omega_K$ with corresponding partition $\mathcal{P} = H \cup K$ of the set of probability measures.

Then we have two possible decisions: d_0 which states that H is true and d_1 stating that K is true.

The test is performed with the help of a decision function δ defined on the value space $\mathcal{X}$ of X, and with two possible values $\{d_0, d_1\}$. We can make this specific by setting $d_0 = 0$ and $d_1 = 1$; this is of course completely arbitrary and also leads to no loss of generality.

Let S_0 be the part of $\mathcal{X}$ for which $\delta(x) = d_0$ and let S_1 be the part where $\delta(x) = d_1$. Then S_0 is called the region of acceptance and S_1 the critical region.

The significance level $\alpha \in (0, 1)$ is chosen and is used to find a test procedure, i.e. a critical region S_1, such that

$$P_\theta(\delta(X) = d_1) \;=\; P_\theta(X \in S_1) \leq \alpha, \tag{7}$$

for every $\theta \in \Omega_H$. Subject to this condition we want make the power function

$$\beta(\theta) \;=\; P_\theta(X \in S_1), \tag{8}$$

as large as possible when $\theta \in \Omega_K$.

So the problem of finding the optimal test can be stated as finding the critical region S_1 that maximizes the power function (8) when $\theta \in \Omega_K$ subject to the constraint (7) for $\theta \in \Omega_H$. Such a critical region defines a test that is called most powerful at level α.

So far the procedure for testing has been completely deterministic, i.e. if we have an outcome x of the random variable X, we calculate the decision function $\delta(x)$ to obtain the value $d_0 = 0$ or $d_1 = 1$, and thus reject the hypothesis H or accept it. Next we introduce randomized tests as follows: For every outcome x of X we draw a random variable $R = R(x)$ with two possible outcomes r_0 and r_1. If we obtain the value r_1 we reject the hypothesis H otherwise we accept it. The probability of $R = r_1$ depends on x, and is assumed to not depend on θ; otherwise we would not be able to draw a (random) conclusion on whether we can reject or accept the hypothesis and we would not have a true test. The probability of rejection can modeled with the function $\phi(x)$, thus for every outcome x of X

$$P_\theta(R(x) = r_1) \quad = \quad \phi(x);$$

the function ϕ is called the critical function. Note that the function ϕ is neither a density function not a probability mass function; it is a function $\mathcal{X} \to [0,1]$. Thus for each outcome x of X we reject the hypothesis H with a probability $\phi(x)$ and accept H with probability $1 - \phi(x)$.

Note that in a randomized test, if the critical function ϕ takes only the values 0 and 1, we really have a nonrandomized test. In those cases choosing a critical function is the same as choosing a critical region. Thus both the randomized and nonrandomized tests can be treated is a unified setting, with the nonrandomized tests being a special case of the randomized ones.

Now assume we have a randomized test with critical function ϕ. Then the probability of rejection is

$$\begin{aligned} P(R(X) = r_1) \quad &= \quad \int P(R(x) = d_1)\, dP_\theta(x) = \int \phi(x)\, dP_\theta(x) \\ &= \quad E_\theta \phi(X) \end{aligned}$$

The problem of finding the optimal test can now be formulated as the problem of finding the critical function ϕ that maximizes the power

$$\beta_\phi(\theta) \quad = \quad E_\theta \phi(X),$$

for $\theta \in \Omega_K$, subject to the condition that the level stays below α i.e. that

$$E_\theta \phi(X) \quad \leq \quad \alpha,$$

for all $\theta \in \Omega_H$.

a) Now if the K consists of several points, or equivalently if Ω_K contains more than one θ value then the procedure that maximizes the power function $\beta(\theta)$ typically will depend on the value of $\theta \in K$ we look at. Then one may need other conditions to find a unique optimal test.

b) In the case when K consists of only one distribution, i.e. when Ω_K contains only one parameter, the optimization problem consists of maximizing one integral subject to some inequality constraints, and then there is a unique solution.

Even when K consists of several points, there might be one unique solution to the optimization problem, i.e. one unique test. When this occurs for K that consists of several points, such a solution to the optimization problem is called a uniformly most powerful (UMP) test.

It is possible to formalize the test procedure using loss functions; we refrain from this and use the notions of errors instead. This is a simpler approach than using loss functions, in fact one the approach with errors is equivalent to a loss formulation with the values of the loss functions being 0 or 1, cf. Lehmann.

8.5 Neyman-Pearson's Lemma

In order to introduce the Neyman-Pearson lemma, we study a simple case when both K and H consist of only one probability distribution.

Assume the hypothesis K and H are simple classes i.e. they consist of one probability measure each so $K = \{P_1\}$ and $H = \{P_0\}$. Assume the distributions are discrete so that $\mathcal{X} = \{x_1, x_2, \ldots\}$ is a countable set and $P_0(X = x) = P_0(x), P_1(X = x) = P_1(x)$ are probability mass functions. Let us first look at nonrandomized tests. With S denoting the critical region, given $\alpha \in (0,1)$ the optimal level α test is given by the critical region S that maximizes

$$\sum_{x \in S} P_1(x),$$

under the constraint

$$\sum_{x \in S} P_0(x) \quad \leq \quad \alpha.$$

To find the optimal critical region S, for each $x \in \mathcal{X}$ there are the two values $P_0(x)$ and $P_1(x)$, and we would like to put x in the critical region if it makes the contribution to $\sum_{x \in S} P_1(x)$ as large as possible for each contribution to $\sum_{x \in S} P_0(x)$. Thus it seems that we should pick an x which makes

$$r(x) \quad = \quad \frac{P_1(x)}{P_0(x)}$$

as large as possible. So we put points x into S according to how large a value they give to $r(x)$, in the order from the one giving the largest value and downwards, and keep doing it until we reach some point where we can not add any more points if we are to keep the level at α. We can therefore state the solution S as the set of points x such that $r(x) > c$ where c is chosen so that level of the test is α, i.e.

$$P_0(X \in S) \quad = \quad \sum_{x \in S} P_0(x) = \sum_{x : r(x) > c} P_0(x) = \alpha.$$

There is difficulty in this in that, having chosen α it might happen that the last point that we can add to S gives $P_0(X \in S) < \alpha$ and adding one more point would give $P_0(X \in S) > \alpha$. The technical way to deal with this is to allow for adding fractions of the points x to S, which means randomization; one allows for adding the part $\phi(x)$ of the point to S. The nontechnical and practical way to deal with this is to change the level α to the level obtained by the last point included (or the point after that).

Theorem 15 *(Neyman-Pearson) Let P_0, P_1 be probabilities with densities p_0, p_1 w.r.t. measure μ. Consider the hypothesis testing problem*

$$\begin{aligned} H: &\quad p_0 \\ K: &\quad p_1 \end{aligned}$$

at level α. Then:

(i). (Existence of candidate.) There is a test ϕ and a constant k so that

$$E_0\phi(X) \;=\; \alpha, \tag{9}$$

and

$$\phi(x) = \begin{cases} 1 & \textit{if } p_1(x) > kp_0(x), \\ 0 & \textit{if } p_1(x) < kp_0(x). \end{cases} \tag{10}$$

(ii). (Sufficiency.) If a test ϕ satisfies (9) *and* (10) *for some k then it is most powerful at level α.*

(iii). (Necessity.) If ϕ is most powerful at level α, then (10) *holds for some k, μ-a.s.. It satisfies also* (9)*, unless there is a test of size strictly less than α with power 1.*

Proof. We assume $\alpha \in (0,1)$ (the theorem is true for $\alpha = 0$ and 1 also but this is not interesting).

(i). Define

$$\begin{aligned} \alpha(c) &= P_0(p_1(X) > cp_0(X)) \\ &= \int 1\{p_1(x) > cp_0(x)\}\, dP_0(x) \\ &= \int 1\{p_1(x) > cp_0(x)\}\, p_0(x) d\mu(x). \end{aligned}$$

Then the points x where $p_0(x) = 0$ give a contribution to this integral which is zero, so we only need to consider the points where $p_0(x) > 0$. Thus $\alpha(c) = P(p_1(X)/p_0(X) > c)$ and so $1-\alpha(c)$ is a distribution function, so $\alpha(c)$ is decreasing and right continuous. Also $P_0(p_1(X)/p_0(X) = c) = \alpha(c-) - \alpha(c)$ and $\alpha(-\infty) = 1$.

Now let c_0 be such that $\alpha(c_0) \le \alpha \le \alpha(c_0-)$, and define

$$\phi(x) = 1\{p_1(x) > c_0p_0(x)\} + \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} 1\{p_1(x) = c_0p_0(x)\}$$

Assume first that $\alpha(c_0-) < \alpha(c_0)$. The size of the test defined by ϕ is

$$\begin{aligned} E_0\phi(X) &= P_0(p_1(X) > c_0p_0(X)) + \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} P_0(p_1(X) = c_0p_0(X))) \\ &= \alpha(c_0) + \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} (\alpha(c_0-) - \alpha(c_0)) \\ &= \alpha. \end{aligned}$$

So if we let $k = c_0$, (i) follows. We note that the second term does not make sense if $\alpha(c_0-) = \alpha(c_0)$, but then also $E_0 1\{p_1(X) = c_0p_0(X)\} = 0$ and letting $\infty \cdot 0 = 0$, ϕ becomes well defined a.e., and again we get (i).

(ii). Assume ϕ satisfies (9), (10) and let ϕ^* be another test with $E_0\phi^*(X) \leq \alpha$. Let

$$\begin{aligned} S^+ &= \{x : \phi(x) > \phi^*(x)\}, \\ S^- &= \{x : \phi(x) < \phi^*(x)\}. \end{aligned}$$

If $x \in S^+$ we must have $\phi(x) > 0$ (since $\phi^* \geq 0$), and thus $\phi(x) = 1$ and $p_1(x) > kp_0(x)$. Similarly when $x \in S^-$ we have $\phi(x) < 1$ and thus $\phi(x) = 0$ and $p_1(x) < kp_0(x)$. Therefore

$$\begin{aligned} \int (\phi - \phi^*)(p_1 - kp_0)\, d\mu &= \int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - kp_0)\, d\mu \\ &\geq 0, \end{aligned}$$

which implies

$$\begin{aligned} E_1\phi(X) - E_1\phi^*(X) &= \int (\phi - \phi^*)p_1\, d\mu \geq k \int (\phi - \phi^*)p_0\, d\mu \\ &\geq 0, \end{aligned}$$

and thus ϕ is most powerful.

(iii). Assume ϕ^* is most powerful and let ϕ be another test that satisfies (9), (10). Let S^+, S^- be as defined in (ii), and let

$$S = (S^+ \cup S^-) \cap \{x : p_1(x) \neq kp_0(x)\}.$$

Then, reasoning similarly to (ii), $(\phi - \phi^*)(p_1 - kp_0) > 0$ on S. If we assume that $\mu(S) > 0$ then

$$\int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - kp_0)\, d\mu = \int_S (\phi - \phi^*)(p_1 - kp_0)\, d\mu > 0,$$

which, reasoning similarly to (ii), implies that

$$E_1\phi(X) - E_1\phi^*(X) > 0,$$

so that ϕ is more powerful than ϕ^*, which is a contradiction. Therefore $\mu(S) = 0$, and $\phi = \phi^*$ a.e. μ, and thus ϕ^* satisfies (10) a.e. μ.

To check (9): ϕ^* is a level α test. But if ϕ^* is of size strictly less than α and power strictly less than 1 we could include additional (portions of) points to the rejection region and increase the power until either $E_0\phi^*(X) = \alpha$ or $E_1\phi^*(X) = 1$, which proves the last statement. □

Note that randomization could be necessary on the boundary set $p_1(x) = kp_0(x)$, in order to get the size equal to α.

Corollary 7 *Assume $\alpha \in (0, 1)$ and let β be the power of the most powerful level α test for testing $H : P_0$ against $K : P_1$. Then $\alpha < \beta$ unless $P_0 = P_1$.*

Proof. The constant test $\phi(x) \equiv \alpha$ has both level α and power α, which implies that the power β of the most powerful test is larger so $\alpha \leq \beta$. Now assume that $\alpha = \beta$ (which by assumption also implies $\beta < 1$); then the test $\phi \equiv \alpha$ is most powerful. Then by (*iii*) of the Neyman-Pearson lemma

$$\phi(x) = \begin{cases} 1 & \text{if } p_1(x) > kp_0(x), \\ 0 & \text{if } p_1(x) < kp_0(x). \end{cases}$$

which is only possible if $p_0(x) = kp_1(x)$ a.e. μ for some k, and since both p_0 and p_1 are densities and thus must integrate to one, we must have that $k = 1$ and so $p_0(x) = p_1(x)$ a.e. μ so that $P_0 = P_1$. □

Example 43 *Assume $X \in N(\xi, \sigma^2)$, with σ^2 known. Let $H : \xi = 0$ and $K : \xi = \xi_1$, for fixed $\xi_1 > 0$. Then*

$$\frac{p_1(x)}{p_0(x)} = \frac{e^{-(x-\xi_1)^2/2\sigma^2}}{e^{-x^2/2\sigma^2}} = e^{\frac{\xi_1 x}{\sigma^2} - \frac{\xi_1^2}{2\sigma^2}}.$$

The exponential function is monotone and $\xi_1 > 0$, so the set where $p_1(x)/p_0(x) > k$ is the same as the set where $x > k'$. So k' can be obtained from the restriction

$$P_0(X > k') = \alpha,$$

i.e. $k' = \sigma z_{1-\alpha}$. □

8.6 p-values

Instead of setting the level before hand to be α one could ask for the corresponding level at which one would reject the hypothesis based on the observed data.

Assume that the distribution of $p_1(X)/p_0(X)$ is continuous under P_0. Then the characterization given by the Neyman-Pearson lemma say that the most powerful level α test is given by a critical region $S_\alpha = \{x : p_1(x)/p_0(x) > k\}$ for a $k = k(\alpha)$; since the border set $\{p_1(x) = kp_0(x)\}$ is a P_0-null set there is no need for a randomization. Now $k(\alpha)$ is chosen so that

$$E_0(\phi(X)) = P_0(p_1(X) > kp_0(X)) = \alpha.$$

Instead one could ask for

$$\hat{p} = \inf\{\alpha : p_1(X) > k(\alpha)p_0(X)\} = \inf\{\alpha : X \in S_\alpha\}.$$

For this we have to assume that the critical regions are nested , i.e. that they satisfy

$$S_\alpha \subset S_{\alpha'} \quad \text{if } \alpha < \alpha'.$$

Then it is always possible to define the smallest possible significance level

$$\hat{p}(X) = \inf\{\alpha : X \in S_\alpha\}.$$

Example 44 *Let $X \in N(\theta, \sigma^2)$, with σ^2 known, and let Φ be the c.d.f. of $N(0,1)$. Assume we want to test $H : \theta = 0, K : \theta > 0$. We have established the critical regions for the most powerful level α test as*

$$S_\alpha \;=\; \{X : X > \sigma z_{1-\alpha}\} = \{X : \Phi(\frac{X}{\sigma}) > 1-\alpha\} = \{X : 1-\Phi(\frac{X}{\sigma}) < \alpha\}.$$

Since Φ is continuous we have

$$\hat{p} \;=\; 1 - \Phi(\frac{X}{\sigma}).$$

We see that the distribution of $\hat{p}$ under P_0 is

$$\begin{aligned} P_0(\hat{p} \le u) &= P_0(1 - \Phi(\frac{X}{\sigma}) \le u) \\ &= P_0(\Phi(\frac{X}{\sigma}) \ge 1-u) \\ &= P_0(\Phi(\frac{X}{\sigma}) \le u) \\ &= P_0(\frac{X}{\sigma} \le \Phi^{-1}(u)) \\ &= u, \end{aligned}$$

so $\hat{p}$ is uniformly distributed on $(0,1)$. □

The next lemma gives a result analogous to the one in the example for level α tests for composite null hypothesis where the critical regions are nested.

Lemma 14 *Assume $X \in P_\theta, \theta \in \Omega$, and we want to test the hypothesis*

$$H : \quad \theta \in \Omega_H,$$

for Ω_H a subset in Ω. Assume the test is defined by critical regions that satisfy $S_\alpha \subset S_{\alpha'}$ if $\alpha < \alpha'$. Let $\hat{p}$ be the p-value defined above.

(i). If

$$\sup_{\theta \in \Omega_H} P_\theta(X \in S_\alpha) \;\le\; \alpha,$$

for all $\alpha \in (0,1)$, then

$$P_\theta(\hat{p} \le u) \;\le\; u,$$

for $u \in [0,1]$, for all $\theta \in \Omega_H$.

(ii). If for every $\theta \in \Omega_H$

$$P_\theta(X \in S_\alpha) \;=\; \alpha,$$

for all $\alpha \in (0,1)$, then

$$P_\theta(\hat{p} \le u) \;=\; u,$$

for $u \in [0,1]$, for all $\theta \in \Omega_H$, so that $\hat{p}$ is uniformly distributed on $[0,1]$. □

Note that (ii) says that $\hat{p}$ is $Un(0,1)$-distributed, and that (i) says that $\hat{p} \geq U$ with $U \in Un(0,1)$ where $\geq$ is the stochastic order defined by $F_{\hat{p}} \leq F_U$.
Proof. (i). Let $\theta \in \Omega_H$. Then if $v < u$, we have $S_v \subset S_u$ so that

$$\{\hat{p} \leq u\} \quad = \quad \{\inf\{\alpha : X \in S_\alpha\} \leq u\} \subset \{X \in S_v\},$$

when $v < u$. Taking probabilities P_θ of both sides

$$P_\theta(\hat{p} \leq u) \quad \leq \quad P_\theta(X \in S_v) = v,$$

and letting $v \downarrow u$, (using the continuity of the probability measure), implies

$$P_\theta(\hat{p} \leq u) \quad \leq \quad u,$$

which proves (i).

(ii). We have

$$\{X \in S_\alpha\} \quad \subset \quad \{\hat{p} \leq u\},$$

so that

$$P_\theta(\hat{p} \leq u) \quad \geq \quad P_\theta(X \in S_u) = u,$$

with the equality following by the assumption in (ii). Thus the statement in (ii) follows from (i). □

8.7 Distributions with Monotone Likelihood Ratio

Now assume that $\theta \in \Omega \subset \mathbb{R}$ and let us study composite hypothesis

$$\begin{aligned} H: &\quad \theta \leq \theta_0, \\ K: &\quad \theta > \theta_0, \end{aligned}$$

where θ_0 is a fixed value. As we have already noted the restricted optimization giving a most powerful level α test against the fixed alternative $\theta_1 \in K$ will typically depend on θ_1, and will therefore not be UMP.

Recalling that the most powerful test for simple hypothesis given by the Neyman-Pearson lemma was a function of the likelihood ratio $p_{\theta_1}/p_{\theta_0}$, it seems that imposing a monotonicity restriction of this ratio could be a feasible approach.

The set of densities $\{p_\theta : \theta \in \mathbb{R}\}$ is said to have a monotone likelihood ratio if $\theta \neq \theta'$ implies that $p_\theta \neq p_{\theta'}$ and that for any $\theta < \theta'$

$$\frac{p_{\theta'}(x)}{p_\theta(x)},$$

is an increasing function of $T(x)$ for some real-valued function T.

Theorem 16 *Assume $\theta \in \mathbb{R}$ and $X \sim p_\theta$ with monotone likelihood ratio in T.*

(i). For testing $H : \theta \leq \theta_0$ against $K : \theta > \theta_0$ there is a UMP test given by

$$\phi(x) = \begin{cases} 1 \text{ if } T(x) > C, \\ \gamma \text{ if } T(x) = C, \\ 0 \text{ if } T(x) < C \end{cases}$$

with C, γ determined by

$$E_{\theta_0}\phi(X) = \alpha.$$

(ii). The power function

$$\beta(\theta) = E_\theta \phi(X)$$

is strictly increasing for $0 < \theta < 1$.

(iii). For every θ' the test ϕ above is UMP for testing

$$\begin{aligned} H : &\quad \theta = \theta' \\ K : &\quad \theta > \theta', \end{aligned}$$

at level $\alpha' = \beta(\theta')$.

(iv). For any $\theta < \theta_0$ the test ϕ minimizes the probability of an error of the first kind $\beta(\theta)$ among all tests that satisfy $E_{\theta_0}\phi(X) = \alpha$.

Proof. (i) and (ii): Assume first that we have simple hypothesis testing $H_0 : \theta = \theta_0, K : \theta = \theta_1$ for $\theta_1 > \theta_0$. Neyman-Pearson's lemma says that we should reject for large values of $r(x) = p_{\theta_1}(x)/p_{\theta_0}(x) = g(T(x))$, with g increasing by the assumption of monotone likelihood ratio. Since g is increasing we could equivalently base the decision to reject on the values of $T(x)$. Copying the proof (excercise) of part (i) of the Neyman-Pearson lemma shows that there is a test ϕ such that

$$\phi(x) = \begin{cases} 1 \text{ if } T(x) > C, \\ \gamma \text{ if } T(x) = C, \\ 0 \text{ if } T(x) < C \end{cases}$$

with C, γ determined by

$$E_{\theta_0}\phi(X) = \alpha,$$

and that this test is most powerful for testing H_0, K.

Now take $\theta' < \theta''$ arbitrary and test the single hypothesis $P_{\theta'}$ against $P_{\theta''}$; the resulting test is most powerful at level $\alpha' = \beta(\theta')$ from the Neyman-Pearson lemma. Then the corollary after the Neyman-Pearson lemma says that the power (at θ'') is larger than the level α', i.e. that $\beta(\theta'') > \beta(\theta')$, and thus β is strictly increasing, which proves part (ii). To continue with the proof of (i): the power is increasing so this implies that

$$E_\theta \phi(X) \leq \alpha \qquad \text{if } \theta \leq \theta_0.$$

But

$$\{\phi : E_\theta \phi(X) \leq \alpha \text{ for all } \theta \leq \theta_0\} \quad \subset \quad \{\phi : E_{\theta_0}\phi(X) \leq \alpha\},$$

so that maximizing the power $\beta(\theta_1) = E_{\theta_1}\phi(X)$ over the the ϕ's in set on the right also maximizes the power over the ϕ's in the set on the left. Therefore the test is most powerful for testing $H : \theta \leq \theta_0$ against $K : \theta_1$. But the test is independent of the $\theta_1 > \theta_0$ we used; it is therefore most powerful for testing $H : \theta \leq \theta_0$ against $K : \theta_1 > \theta_0$, and UMP.

(iii). Follows trivially: The test should be of the above form, and the level $\alpha' = \beta(\theta')$ is the right level for the test.

(iv). For any $\theta < \theta_0$ the test that *minimizes* the power for testing $H : \theta_0$ against $K : \theta$ is given, by the Neyman-Pearson lemma, as a test ϕ of the above form with $E_{\theta_0}\phi(X) = \alpha$. □

A class of distributions that satisfy the assumption of monotone likelihood ratio is the one-parameter exponential family.

Corollary 8 *Assume X has density*

$$p_\theta(x) \quad = \quad C(\theta)e^{Q(\theta)T(x)}h(x)$$

with $\theta \in \mathbb{R}$ and Q strictly monotone. Then there is UMP test ϕ for testing $H : \theta \leq \theta_0$ against $K : \theta > \theta_0$. If Q is strictly increasing then

$$\phi(x) \quad = \quad \begin{cases} 1 \text{ if } T(x) > C, \\ \gamma \text{ if } T(x) = C, \\ 0 \text{ if } T(x) > 1, \end{cases}$$

with C, γ given by $E_{\theta_0}\phi(X) = \alpha$. If Q is strictly decreasing the test is given the same expression with inequalities reversed.

Proof. The likelihood ratio is

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \quad = \quad e^{(Q(\theta_1)) - Q(\theta_0))T(x)},$$

and if Q is strictly increasing $Q(\theta_1)) - Q(\theta_0) > 0$ and thus the LR is monotone in T. If Q is strictly decreasing the LR is monotone in $-T$. □

Example 45 *Assume $X \sim Bin(n, p)$ so that*

$$P_p(x) = \binom{n}{p} p^x (1-p)^{n-x},$$

which is one-parameter exponential with $T(x) = x, \theta = p, Q(p) = \log(p/(1-p))$. Then Q is strictly increasing on $(0, 1)$, so there is a UMP test ϕ for testing $H : p \leq p_0$ against $K : p > p_0$. The test rejects H when x is large enough.

Example 46 *Assume $X_1, \ldots, X_n$ are independent Poisson distributed random variables with expectation λ so that*

$$P_\lambda(x_1, \ldots, x_n) = \frac{\lambda^{x_1+\ldots+x_n}}{x_1! \cdots x_n!} e^{-n\lambda}.$$

This is a one-parameter exponential family with $T(x) = x_1 + \ldots + x_n$ and $Q(\lambda) = \log \lambda$. Q is structly increasing on $(0, \infty)$ so there is a UMP test ϕ for testing $H : \lambda \leq \lambda_0$ against $K : \lambda > \lambda_0$; the test will reject the null hypothesis for large values of $T(x)$.

We started the part on testing with a review of different types of losses, using loss functions, that encompassed point estimation, interval estimation and testing problems. However, so far, we have discussed the "loss" in testing problems rather informally only using two types of errors. This can more formally be modeled using two loss functions L_1 and L_2 for the two types of losses. Recall that testing was modeled in decision theory with two possible decisions $D = \{d_0, d_1\}$ with d_1 meaning reject the null hypothesis $H : \Omega_H$ in favor of $K : \Omega_K$ and d_0 meaning accept the null hypothesis.

Now define the two loss functions L_0, L_1 by

$$\begin{aligned} L_0(\theta, d_1) &= 1, \text{ for } \theta \in \Omega_H, \\ L_0(\theta, d_1) &= 0, \text{ for } \theta \in \Omega_K, \\ L_0(\theta, d_0) &= 0, \text{ for all } \theta, \end{aligned}$$

and

$$\begin{aligned} L_1(\theta, d_0) &= 0, \text{ for } \theta \in \Omega_H, \\ L_1(\theta, d_0) &= 1, \text{ for } \theta \in \Omega_K, \\ L_1(\theta, d_1) &= 0, \text{ for all } \theta. \end{aligned}$$

Then to minimize $EL_1(\theta, \delta(X))$ subject to $EL_0(\theta, \delta(X)) \leq \alpha$, over the set of decision functions δ, is equivalent to:

(i). Minimize $P_\theta(\delta(X) = d_0)$ when $\theta \in \Omega_K$ under the assumption that $P_\theta(\delta(X) = d_1) \leq \alpha$ when $\theta \in \Omega_H$, or equivalently,

(ii). Maximize $P_\theta(\delta(X) = d_1) = \beta(\theta)$ when $\theta \in \Omega_K$ under the assumption that $P_\theta(\delta(X) = d_1) \leq \alpha$ when $\theta \in \Omega_H$.

If we disregard the randomization, which might be necessary to obtain an exact level α test, this is exactly what we have been doing for hypothesis testing so far.

Now let us introduce a bit more general loss functions. Again let us study the testing problem $H : \theta \leq \theta_0$ against $K : \theta > \theta_0$ with possible decisions d_0 to accept H and d_1 to reject H. Let $L(\theta, d)$ be a loss function and define

$$\begin{aligned} L_0(\theta) &= L(\theta, d_0), \\ L_1(\theta) &= L(\theta, d_1), \end{aligned}$$

the losses for making decisions d_0 and d_1 respectively. Assume that

$$\begin{aligned} L_0(\theta) &= \begin{cases} 0 \text{ when } \theta \leq \theta_0, \\ \text{ strictly increasing for } \theta > \theta_0, \end{cases} \\ L_1(\theta) &= \begin{cases} 0 \text{ when } \theta > \theta_0, \\ \text{ strictly decreasing for } \theta \leq \theta_0. \end{cases} \end{aligned}$$

Then

$$L_1(\theta) - L_0(\theta) \quad \begin{cases} > 0 \text{ if } \theta < \theta_0, \\ < 0 \text{ if } \theta > \theta_0. \end{cases}$$

Theorem 17 *Assume X has density p_θ with $\theta \in \mathbb{R}$ and monotone likelihood ratio in $T(x)$, and assume the loss function for testing $H : \theta \leq \theta_0, K : \theta > \theta_0$ satisfies*

$$L_1(\theta) - L_0(\theta) \quad \begin{cases} > 0 \textit{ if } \theta < \theta_0, \\ < 0 \textit{ if } \theta > \theta_0. \end{cases} \tag{11}$$

(i). The family $\mathcal{C} = \{\phi_\alpha : 0 < \alpha < 1\}$ of tests ϕ_α given by

$$\phi(x) = \begin{cases} 1 \textit{ if } T(x) > C, \\ \gamma \textit{ if } T(x) = C, \\ 0 \textit{ if } T(x) < C \end{cases} \tag{12}$$

and

$$E_{\theta_0}\phi(X) = \alpha.$$

is essentially complete.

(ii). If in addition $\{x : p_\theta(x) > 0\}$ is independent of θ, the family is minimal essentially complete.

Proof. (i). Let ϕ be an arbitrary test, so that ϕ takes the value 1 for rejecting H i.e. for making decision d_1 and it takes the value 0 for accepting H i.e. for making the decision d_0, and possibly another value in the interval $(0, 1)$; the probability of rejecting H. The loss function is given by

$$L(\theta, \phi(x)) = \phi(x)L_1(\theta) + (1 - \phi(x))L_0(\theta).$$

The risk function is therefore

$$\begin{aligned} R(\theta, \phi) &= EL(\theta, \phi(X)) = \int p_\theta(x) \{\phi(x)L_1(\theta) + (1 - \phi(x))L_0(\theta)\}\, d\mu(x) \\ &= \int p_\theta(x) \{L_0(\theta) + (L_1(\theta) - L_0(\theta))\phi(x)\}\, d\mu(x). \end{aligned}$$

Thus the difference in risk between two such tests ϕ, ϕ' is

$$R(\theta, \phi') - R(\theta, \phi) = (L_1(\theta) - L_0(\theta)) \int (\phi'(x) - \phi(x))p_\theta(x)\, d\mu(x).$$

Using (11) and the definition of the two loss functions L_0, L_1, we see that $R(\theta, \phi') - R(\theta, \phi) \leq 0$ if

$$\begin{aligned} \beta_{\phi'}(\theta) - \beta_\phi(\theta) &= E_\theta \phi'(X) - E_\theta \phi(X) && (13) \\ &= \int (\phi' - \phi) p_\theta \, d\mu \begin{cases} > 0 \text{ for } \theta > \theta_0, \\ = 0 \text{ for } \theta = \theta_0, \\ < 0 \text{ for } \theta < \theta_0. \end{cases} && (14) \end{aligned}$$

Now let $\phi \notin \mathcal{C}$ be any test with level α so that $E_{\theta_0}\phi(X) = \alpha$. Then there is a UMP level α test $\phi' \in \mathcal{C}$ (i.e. that satisfy the above conditions), for testing $\theta = \theta_0$ against $\theta > \theta_0$, so then $\beta_{\phi'}(\theta) - \beta_\phi(\theta) > 0$. Furthermore ϕ' minimizes the power for $\theta < \theta_0$, so then $\beta_{\phi'}(\theta) - \beta_\phi(\theta) < 0$. This implies that(14) is satisfied and therefore $R(\theta, \phi') \leq R(\theta, \phi)$, and thus any test $\phi \notin \mathcal{C}$ is dominated by a test $\phi' \in \mathcal{C}$ in the family, so the family $\mathcal{C}$ is essentially complete.

(ii). (Not included) □

This result implies that the UMP tests obtained previously give rise to an essentially complete class for general decision problems, where the loss functions satisfy the above conditions. Thus one can see UMP tests at given significance levels as a selection of particular procedure from an essentially complete class.

One can weaken the assumption of monotone likelihood ratio. The family of distributions $\{F_\theta : \theta \in \Omega\}$ is said to be stochastically increasing if the distributions are distinct and if $\theta < \theta'$ implies $F_\theta(x) \geq F_{\theta'}(x)$. If $X \sim F_\theta$ and $X' \sim F_{\theta'}$ then we say that X' is stochastically larger than X.

8.8 Confidence sets

Recall that one type of decision problems was the problem of making confidence intervals, i.e. the problem of giving an interval $\delta(X) = (l(X), u(X))$ based on the observation X that contains the unknown parameter value θ.

Let us first consider the problem of giving confidence intervals or confidence set of the form $S(X) = (l(X), \infty)$, so that we are interested in giving a lower bound for the unknown parameter. Then decisions that cover the unknown parameter with a high probability are preferred so that we should have

$$P_\theta(l(X) \leq \theta) \geq 1 - \alpha, \tag{15}$$

with $1 - \alpha$, the confidence level, large. Subject to (15), we would like to have a decision that gives as high precision as possible, i.e. we would like to have the bound $l(X)$ as close to the true value θ as possible. We could formulate this as that for any $\theta' < \theta$ we want the probability

$$P_\theta(l(X) \leq \theta')$$

to be as small as possible.

Define (if it exists)

$$\hat{l} \;=\; \text{argmin}_l P_\theta(l(X) \leq \theta')$$

for every $\theta' < \theta$ subject to (15). Then $\hat{l}$ is called a uniformly most accurate lower confidence bound for θ at level $1-\alpha$.

The problem can be stated a bit more generally using loss functions, so let $L(\theta, l)$ be the loss for giving the lower bound l for the parameter θ (equivalently we could use the formulation $L(\theta, \delta)$ with δ a decision rule as above); a sensible such loss function satisfies that for a fixed θ' and as a function of l

$$L(\theta, l) \;=\; \begin{cases} 0 \text{ if } l > \theta, \\ \text{positive and decreasing if } l \leq \theta. \end{cases} \qquad (16)$$

The first condition is a matter of convention, since we are interested in the loss only for lower bounds, the second states that the farther from the true value we are the higher the loss.

Then we could look for the lower bound l that minimizes

$$E_\theta L(\theta, l(X)) \qquad (17)$$

over the set of functions l subject, to (15).

Lemma 15 *Assume the loss function satisfies* (16). *Then a solution to the optimization problem* (17) *is given by the uniformly most accurate lower bound for* θ.

Proof. Not included □

Now let us return to the confidence set formulation $S(x)$. Then a family of subsets $\{S(x)\}$ of the parameter space is called a family of confidence sets at level $1-\alpha$ if

$$P_\theta(\theta \in S(X)) \;\geq\; 1-\alpha.$$

The next result gives an algorithm for obtaining uniformly most accurate confidence sets, via UMP tests.

Theorem 18 *(i). Consider the testing problem* $H: \theta = \theta_0$ *for* $\theta_0 \in \Omega$ *fixed. Let* $A(\theta)$ *be the acceptance region of a level* α, *test and for each outcome* x *let*

$$S(x) \;=\; \{\theta : x \in A(\theta), \theta \in \Omega\}.$$

The $S(x)$ *is a family of confidence sets for* θ *at level* $1-\alpha$.
(ii). If for each θ_0, $A(\theta_0)$ *is the acceptance region for a UMP test for testing* $H: \theta = \theta_0$ *against alternative* $K(\theta_0)$. *Then* $S(X)$ *minimizes*

$$P_\theta(\theta_0 \in S(X)) \qquad \textit{for all } \theta \in K(\theta_0),$$

among all level $1-\alpha$ *families of confidence sets for* θ.

Proof. (i). By the definition of $S(x)$

$$P_\theta(\theta \in S(X)) \;=\; P_\theta(X \in A(\theta)) \geq 1 - \alpha.$$

(ii). Let $S^*(x)$ be another family of level $1-\alpha$ confidence sets and

$$A^*(\theta) \;=\; \{x : \theta \in S^*(x)\},$$

so that

$$P_\theta(X \in A^*(\theta)) \;=\; P_\theta(\theta \in S^*(X)) \geq 1 - \alpha,$$

and thus $A^*(\theta_0)$ is the acceptance region of a level α test for testing $H : \theta = \theta_0$. Since $A(\theta_0)$ is the acceptance region (the complement of the critical region) for a UMP level α test

$$P_\theta(X \in A^*(\theta_0)) \;\geq\; P_\theta(X \in A(\theta_0)),$$

so that

$$P_\theta(\theta_0 \in S^*(X)) \;\geq\; P_\theta(\theta_0 \in S(X)),$$

which is what we wanted to prove. □

Having established the equivalence between UMP tests and most accurate confidence sets, we apply the results to.

Corollary 9 *Assume the densities $\{p_\theta\}$ have a monotone LR in $T(x)$. Assume the distribution function $F_\theta(t)$ of $T(X)$ is separately continuous in t and θ.*

(i). There is a uniformly most accurate confidence bound l for θ at each confidence level $1-\alpha$.

(ii). If the equation

$$F_\theta(t) \;=\; 1 - \alpha, \tag{18}$$

has a solution $\theta = \hat{\theta}$ in Ω then the solution is unique and $l(x) = \hat{\theta}$.

Proof. (i). Let α be given. Since $\{p_\theta\}$ has a monotone LR in T there is for each θ_0 a constant $C(\theta_0)$ so that $\{T > C(\theta_0)\}$ is a UMP level α rejection region for testing $\theta = \theta_0$ against $\theta > \theta_0$. Let ϕ_{θ_0} denote the critival function for this (non-randomized) test. The power function of this test is

$$\begin{aligned} \beta_{\phi_{\theta_0}}(\theta) &= E_\theta \phi_{\theta_0}(X) \\ &= P_\theta(T(X) > C(\theta_0)). \end{aligned}$$

We have that for any $\theta_1 > \theta_0$

$$\begin{aligned} \alpha &= P_{\theta_1}(T > C(\theta_1)) \\ &= \beta_{\phi_{\theta_1}}(\theta_1) \\ &> \beta_{\phi_{\theta_1}}(\theta_0) \\ &= P_{\theta_1}(T > C(\theta_0)). \end{aligned}$$

Since the d.f. F_θ of $T(X)$ is continuous this implies that $C(\theta_0) < C(\theta_1)$, and thus the function C is strictly increasing. Let $A(\theta) = \{T \leq C(\theta)\}$ be the acceptance region. Then

$$\begin{aligned} S(x) &= \{\theta : x \in A(\theta)\} \\ &= \{\theta : T(x) \leq C(\theta)\} \\ &= \{\theta : \inf\{\eta : T(x) \leq C(\eta)\} \leq \theta\} \\ &= \{\theta : l(x) \leq \theta\}, \end{aligned}$$

since C is increasing, where

$$l(x) = \inf\{\theta : T(x) \leq C(\theta)\}.$$

The previous theorem says that $\{\theta : l(x) \leq \theta\}$ is a family of confidence sets at level $1-\alpha$ that minimizes $P_\theta(l(X) \leq \theta')$ for all $\theta > \theta'$. Thus l is a uniformly most accurate confidence bound (by definition of a such).

(ii). The critical regions are given by sets $\{T > C(\theta)\}$. The power is strictly increasing in θ when it is strictly between zero and one (by the corollary to the Neyman-Pearson lemma). Since C is strictly increasing and F_θ is continuous this implies that the d.f. $F_\theta(t)$ of $T(X)$ is strictly decreasing in θ for all t for which $0 < F_\theta(t) < 1$. Therefore the equation (18) at most one solution. Now assume it has solution $\hat{\theta}$ so that

$$F_{\hat{\theta}}(t) = 1 - \alpha,$$

and so $C(\hat{\theta}) = t$. Thus $t \leq C(\theta)$ is equivalent to $C(\hat{\theta}) \leq C(\theta)$ which is equivalent to $\hat{\theta} \leq \theta$. Thus $l(x) = \hat{\theta}$. □

When either of the random variables X or T are discrete the distribution function F_θ will not be continuous and then the previous corollary can not be applied directly. Then also the optimal test for testing $\theta = \theta_0$ is most often randomized.

But: Let U be a $Un(0,1)$ random variable independent of X, and let ϕ be a randomized test based on X. Then a randomized test can be obtained by providing a (critical set) R for the pair (X, U) that determines when to reject the null hypothesis: so given the outcome $X = x$ it will reject the null hypothesis when $(x, U) \in R$ i.e. with probability $P((x, U) \in R)$: Define the set

$$R = \{(x, u) : u \leq \phi(x)\}.$$

Then

$$\begin{aligned} P((X,U) \in R) &= P(U \leq \phi(X)) \\ &= \int \phi(x)\, dP_X(x) \\ &= E\phi(X) \end{aligned}$$

Thus if we let R above be the rejection region for a test based on the pair (X, U) this will imply that the resulting randomized test has the same power function as the original randomized test ϕ and so the two tests are equivalent.

Furthermore: If X is integer valued (or more generally lattice) we can use the statistic

$$T = X + U,$$

with $U \in Un(0,1)$, for defining a randomized test instead of (X, U), since then $X = [T]$ and $U = T - [T]$ almost surely, and thus T is equivalent to (X, U). Since the distribution of T is continuous the previous corollary can be used.

Thus if X is discrete and integer valued (or lattice) we can obtain a continuous distribution function for a statistics T that is equivalent to a randomized test based on (X, U), with $U \in Un(0,1)$ independent of X, that has the same power function as (so is equivalent to) any randomized test based on X.

Now let l, u be lower and upper bounds for θ with respective levels $1-\alpha_1, 1-\alpha_2$, and assume that $l(x) < u(x)$ for all x. (This will happen for instance when $\alpha_1+\alpha_2 < 1$ and there is a monotone LR in T which has a distribution separately continuous in θ, t.) Then

$$P_\theta(l(X) \leq \theta \leq u(X)) = 1 - \alpha_1 - \alpha_2,$$

for all θ.

Now assume that $L_1(\theta, l)$ is decreasing in l on $l \leq \theta$ and zero on $l > \theta$, and $L_2(\theta, u)$ is increasing in u on $u \geq \theta$ and zero on $u < \theta$. Then if $\hat{l}$ and $\hat{u}$ are uniformly most accurate, at levels $1-\alpha_1$ and $1-\alpha_2$, they minimize $E_\theta L_1(\theta, l(X))$ and $E_\theta L_2(\theta, u(X))$ at respective levels $1-\alpha_1$ and $1-\alpha_2$. Let

$$L(\theta, l, u) = L_1(\theta, l) + L_2(\theta, u).$$

Then it follows that the interval function $(\hat{l}, \hat{u})$ miminizes

$$E_\theta L(\theta, l(X), u(X)),$$

under the constraints

$$\begin{aligned} P_\theta(l(X) > \theta) &\leq \alpha_1, \\ P_\theta(u(X) < \theta) &\leq \alpha_2. \end{aligned}$$

Examples on loss functions that satisfy the above restrictions are

$$\begin{aligned} L(\theta, l, u) &= \begin{cases} u - l \text{ for } l \leq \theta \leq u, \\ u - \theta \text{ for } \theta < u, \\ \theta - l \text{ for } \theta > l, \end{cases} \\ L_{a,b}(\theta, l, u) &= a(\theta - l)^2 + b(\theta - u)^2. \end{aligned}$$