Statistical Research Report
Institute of Mathematics
University of Oslo

No. 8
July 1969.

A COMMENT ON THE COMPOUND DECISION THEORY

By

Else Sandved
University of Oslo

A COMMENT ON THE COMPOUND DECISION THEORY.

By Else Sandved
University of Oslo.

1. An example by Robbins.

The Compound Decision Theory was introduced by Robbins [1] and has been developed particularly by him and Ester Samuel in several papers (see for instance [3]). To get an illustration of the concept, we shall consider the following simple example, given by Robbins in [1] .

Let $x_1,\ldots,x_n$ be independent random variables, each normally distributed with variance 1 and with means $\theta_1,\ldots,\theta_n$, respectively, where $\theta_i = +1$ or -1. On the basis of $x_1,\ldots x_n$ we are to decide, for every i, whether the true value of θ_i is 1 or -1. Let Ω denote the set of all 2^n possible parameter points $\underline{\theta} = (\theta_1,\ldots,\theta_n)$ and let $w(\underline{\theta}',\underline{\theta}) = \frac{1}{n}$ (no. of i for which $\theta_i' \neq \theta_i$) be the loss involved when the true parameter point is $\underline{\theta}$ and the decision $(\underline{\theta}=\underline{\theta}')$ is taken.

A simple and reasonable decision rule, when the loss function is as above, seems to be the rule

$\tilde{R}$: estimate θ_i by $\operatorname{sgn}(x_i)$; $i = 1,\ldots,n$.

The corresponding risk function $L(\tilde{R},\underline{\theta}) = Ew(\underline{\theta}',\underline{\theta})$ equals $F(-1) = 0.1587$ for all $\underline{\theta}$, where F is the cumulative normal distribution function. $\tilde{R}$ is the maximum likelihood estimator of $\underline{\theta}$, and Robbins shows that $\tilde{R}$ is the unique minimax decision rule.

2. The Bayes Case.

Suppose that in the example above the θ_i' s are independent random variables taking the values 1 and -1 with probabilities p and 1-p, respectively, where p is known. Let $u(x_i)$ be the conditional probability of estimating θ_i to be 1, given x_i. The corresponding risk

$$p\int f(x-1)(1-u(x))dx + (1-p)\int f(x+1)u(x)dx,$$

where f is the normal density function, is minimized by the rule

R_p : estimate θ_i by $\operatorname{sgn}(x_i - \tfrac{1}{2}\ln((1-p)/p))$; $i = 1, \ldots, n$

which has the risk

$$h(p) = pF(-1 + \tfrac{1}{2}\ln((1-p)/p)) + (1-p)F(-1-\tfrac{1}{2}\ln((1-p)/p)).$$

$h(p)$ is less than $F(-1)$ for $p \neq 0.5$ and equal to $F(-1)$ for $p = 0.5$, and R_p will therefore be preferable to $\tilde{R}$ in this case, unless $p = 0.5$.

3. The Empirical Bayes Case.

If in the Bayes Case above p is unknown, and the n x_i's are used to estimate p, then a decision rule corresponding to R_p, with p substituted by the estimate of p, could be used. This would be an example of an Empirical Bayes Case. See Robbins [2].

4. The case where the frequency of θ_i's equal to 1 is known.

Suppose that the situation is as in section 1, except that the frequency $p = \frac{1}{n}$ (no. of θ_i's equal to 1) is known. Then the rule R_p in section 2 minimizes the risk among all simple rules, that is rules where the estimate of θ_i depends on x_i only, and the risk of this R_p is also $h(p)$.

5. The Compound Decision Case.

Let us denote the problem in section 1 of the present paper as a Compound Decision Problem if it satisfies the describtion of Robbins in [1]: No relation whatever is assumed to hold amongst the unknown parameters θ_i. Then the frequency p in section 4 is completely unknown, but may be estimated by means of $x_1, \ldots, x_n$. The estimator $v = \frac{1}{2}(1+\bar{x})$, where $\bar{x} = \frac{1}{n}\sum_1^n x_i$, is unbiased for p. As v can take on values outside $[0,1]$, it is truncated at 0 and 1, and the resulting estimator

$$v' = \begin{cases} 0 & \text{if } v \leq 0 \\ v & \text{if } 0 < v < 1 \\ 1 & \text{if } v \geq 1 \end{cases}$$

is substituted for p in R_p. Hence one gets the decision rule

$$R^*: \text{ estimate } \theta_i \text{ by } \begin{cases} -1 & \text{if } \bar{x} \leq -1 \\ \operatorname{sgn}(x_i - \frac{1}{2}\ln((1-\bar{x})/(1+\bar{x}))) & \text{if } -1<\bar{x}<1 \\ 1 & \text{if } \bar{x} \geq 1 . \end{cases}$$

Let h(p,n) denote the risk function of R^*, where p equals the frequency in section 4. This risk function and the risk function $h(p) = \lim_{n\to\infty} h(p,n)$ for R_p and the risk function F(-1) for $\tilde{R}$ are compared in the following table (see [1]).

p	F(-1)	h(p)	h(p,100)	h(p,1000)
0.0 or 1.0	0.1587	0	0.0041	
0.1 or 0.9	0.1587	0.0691	0.0763	
0.2 or 0.8	0.1587	0.1121	0.1174	
0.3 or 0.7	0.1587	0.1387	0.1439	
0.4 or 0.6	0.1587	0.1538	0.1591	
0.5	0.1587	0.1587	0.1628	0.1591

Table 1.

For p = 0.5, h(p,n) is always greater than F(-1), though the difference is very small for large n. For any $p \neq 0.5$, h(p,n) is less than F(-1) for large enough n. For p near 0 or 1, h(p,n) is much less than F(-1), at least for n as large as 100.

If the case is as in sections 2 or 4, and $p \neq 0.5$, then the rule R_p is obviously preferable to $\tilde{R}$. If the case is as in section 5, then Table 1 apparently shows that there are strong reasons for applying R^* instead of $\tilde{R}$. Intuitively, however, it seems very unreasonable to mix the n problems together in the way that is done in R^*, since the θ_i's have nothing to do with each other. Below we shall give some further arguments for not preferring R^* to $\tilde{R}$. These arguments against R^* are not applicable to the rule suggested in the Empirical Bayes Case, that is, when the problems are presented to the statistician in random order.

6. Arguments for not applying R^*

Consider first the asymptotic case where we assume that any sequence $\frac{1}{n}(\theta_{i_1}+\theta_{i_2}+\ldots+\theta_{i_n})$, where $i_1 < i_2 < \ldots < i_n$, has a limit as $n \to \infty$. Then the asymptotic risk of R^* is $h(p)$, where $p = \lim_{n\to\infty}\frac{1}{n}(\theta_1+\ldots+\theta_n)$. Now it is possible to find a sequence of methods, say R_1^*, R_2^*, ...etc, where R_1^* is asymptotically uniformly at least as good as R^*, and where R_{i+1}^* is asymptotically uniformly at least as good as R_i^*, $i = 1,2,\ldots$etc. This sequence runs as follows: Denote the original sequence of problems by (θ_1,x_1), (θ_2,x_2),...etc. Then R_1^* consists in applying R^* separately on the two subsequences of problems

$$(\theta_1,x_1),(\theta_3,x_3),(\theta_5,x_5),\ldots$$

and

$$(\theta_2,x_2),(\theta_4,x_4),(\theta_6,x_6),\ldots.$$

Let $p_1 = \lim_{k\to\infty}\frac{1}{k}(\theta_1+\theta_3+\ldots+\theta_{2k-1})$ and $p_2 = \lim_{k\to\infty}\frac{1}{k}(\theta_2+\theta_4+\ldots+\theta_{2k})$. Then $p = \frac{1}{2}(p_1+p_2)$, and because of the concavity of $h(p)$, the asymptotic risk of R_1^*, namely $\frac{1}{2}(h(p_1)+h(p_2))$, is less than $h(p)$, unless $p_1=p_2$, in which case $\frac{1}{2}(h(p_1)+(h(p_2)) = h(p)$. Hence, if $p_1 \neq p_2$, then R_1^* is asymptotically uniformly better than R^*.

The construction of R_1^*, R_2^*,...etc. is obvious: The relation between R_{i+1}^* and R_i^* is the same as the relation between R_1^* and R^*, $i = 1,2,\ldots$etc.

Let us now consider the more interesting case where $n = 2k$ is large but fixed. If the problems are presented to the statistician in random order, then this is not a Compound Decision Problem according to the describtion of Robbins:"No relation whatever is assumed to hold amongst the unknown parameters θ_i", because randomization creates relations between the θ_i's, for instance

the relation that $p_1 \approx p_2$ [illegible] with high probability, where $p_1 = \frac{1}{2k}(\theta_1+\theta_3+\ldots+\theta_{2k-1})$ and $p_2 = \frac{1}{2k}(\theta_2+\theta_4+\ldots+\theta_{2k})$. Hence this situation, where the θ_i' s are presented in random order, should rather be called an Empirical Bayes Case.

Now consider the case where the problems are not presented in random order, but according to something else, for instance according to time order. If we do not believe that Nature randomizes the problems for us, then there is no reason why p_1 should be near p_2, and if p_1 and p_2 are not close together, then the rule R_1^* is better than R^*, because if $n = 2k$ is large, then k is also large. Hence the very same sort of argument for preferring R^* to $\tilde{R}$ applies for preferring R_1^* to R^* and for preferring R_2^* to R_1^*, where R_2^* consists in applying the rule R^* separately on each of the four subsequences of problems

$$(\theta_1,x_1),(\theta_5,x_5),(\theta_9,x_9),\ldots$$
$$(\theta_2,x_2),(\theta_6,x_6),(\theta_{10},x_{10}),\ldots$$
$$(\theta_3,x_3),(\theta_7,x_7),(\theta_{11},x_{11}),\ldots$$
$$(\theta_4,x_4),(\theta_8,x_8),(\theta_{12},x_{12}),\ldots.$$

Continuing in this way, it is seen that there are always strong reasons for preferring R_{i+1}^* to R_i^* for any i, and finally one gets that the perhaps most preferable rule is to apply the rule R^* separately on the n "subsequences" consisting of one problem each, but that amounts to apply the rule $\tilde{R}$!

Fortunately, in practise there are often reasons to believe some relations to hold between the unknown parameters θ_i. In the example above it may for instance be possible to stratify the n original problems into strata where the frequency of θ_i' s equal

to 1 differ considerably from strata to strata. Then an Empirical Bayes rule, constructed for each stratum separately, will probably have lower risk than any of the rules $\tilde{R}$, R^* or R_i^*.

REFERENCES.

[1] H. Robbins, "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems", Proc. Second Berk. Symp. 1951 pp 131-148.

[2] H. Robbins,"An Empirical Bayes Approach to Statistics," Proc. Third Berk. Symp. 1956 pp 157-164.

[3] Ester Samuel, "The compound statistical decision problem," Sankhyā Vol 29 Part 2 pp 123-139.