Statistical Research Report
Institute of Mathematics
University of Oslo

No. 3
September 1966.

THE PRESENT STATE OF THE DECISION THEORY AND THE NEYMAN-PEARSON THEORY

by

Erling Sverdrup

0. INTRODUCTION.

A. Some general views about statistical inference.

The principles of statistical induction will quite naturally deal with two questions. First, what kind of general rules should be used when formulating the a priori assumptions (the model) and the purpose (the decision situation) of a specific statistical investigation ? Second, with a given model and purpose what kind of general principles should be applied to derive the correct procedure for making a decision ? Any statistician working with practical statistical investigations has consciously or subconsciously adopted some kind of attitude to these questions. Not least for this reason is it important explicitely to be concerned with them.

The two questions are closely related. Nevertheless it is important to keep them separated. Thus a general principle of statistical inference should not be motivated by arguing that only models of certain types can occur in practice, i.e. by referring to certain empirical results. The principles of inference and the empirical results must not be confused.

The formulation of the model is of course a very important, and sometimes a very difficult, part of the process of statistical inference. Important contributions to the advancement of statistical science have consisted in demonstrating how problems and situations from certain fields of applications can be formulated stochastically. It is not only politics which is the art of that which is feasible. It is to a high degree the case of statistics. The model ought to be realistic, but it has to be admitted that the choice of the model is often made with a view to what could be done with it.

We shall be concerned with the second of the two questions which has been outlined above, and since we shall not be concerned with economics, biology etc, our aim is to formulate principles which are meaningful relatively to "arbitrary" models and decision problems.

This question has attracted the attention of the

statisticians at least since the beginning of the 1920-th under the leadership of among others R.A. Fisher [7] and [8], Jerzy Neyman and E.S. Pearson [23], Abraham Wald [29] and [30] and E.L. Lehmann in a number of papers, see in particular his monograph [22]. (R.A. Fisher himself would perhaps reject the present approach to the problem, as a "wooden attitude".)

Our starting point is a model which is defined stochastically and a situation which is non-Bayes. In the now "classical" statistics it is perhaps the following principles which have attained prominence: a. The principle of sufficiency. b. Appraisal of a test by means of the power function, in particular the principle of unbiasedness in the power. c. The principle of invariance. d. Markov-estimation (minimal variance subject to mean unbiasedness). That the principle of unbiasedness in the power is considered so prominent may perhaps surprise somebody. It will be substantiated below.

In the situations where these principles are successful, it is in some cases not so essential whether the situation is Bayes or non-Bayes, the results can be reformulated for Bayesian situations. However, by an extention of the principles to more general decision situations, difficulties arise which are the direct reasons for the revival of the Bayesian idea. It is outside the assignement alloted to me for this conference to deal with this idea. There is, however, a particular class of decision problems where the direct continuation of the non-Bayesian idea has been relatively successful, viz. the socalled "multiple test situations". Their special structure will be defined below.

In order to limit the scope of these lectures, the title of which is indeed very ambitious, I shall treat the following subjects

I. Sufficient statistics.
II. Power unbiasedness.
III. Multiple testing of hypothesis.

Markov estimation and point estimation in general will only be treated occasionally. Non-parametric methods will only be touched upon to the extent to which they illustrate the principles. In the last years there has taken place an important development in this field which could be said to be a direct continuation of the "classical" statistical ideas. I am also aware of the

interesting discussion about the minimax principle which took place only a few years ago and the important contributions which were then made to the understanding of this principle. The necessity of limiting the subjects treated in these lectures is the only reason for not dealing with it.

Both Neyman-Pearson and Fisher were hampered by the mathematical difficulties present at that time and it led them in some instances to be preoccupied with mathematical concepts which were inessential from a statistical point of view. In particular Neyman and Pearson insisted on mathematical rigor, which was not easy to live up to with the commonly accepted mathematical tools among statisticians. The general acceptance of measure theory was a relief in that respect. It made it possible for the statisticians to free themselves from preoccupation with mathematics and concentrate on statistical ideas. Perhaps in no other field of statistics is this more pronounced than in the theory of sufficiency.

B Conditional probability.

After the clarification of the concept of conditional probability by A. Kolmogorov [13] some years passed before the statisticians realized to which extent it concerned them. Perhaps the first one was D. Blackwell [4]. We shall very briefly review some main results which we shall need.

A statistician may indiscriminately undertake a very detailed description of the statistical data. On the other hand he may bring forth that only statements about the data of a certain type are of interest. Thus the starting point is a sigmafield $\mathcal{A}$ of subsets (statements) in the sample space $\mathcal{X}$ of sample points x. A subsigmafield $\mathcal{A}_0$ of $\mathcal{A}$ may be of special interest. Let $P(A)$ be a probability measure over $\mathcal{A}$. Furthermore $P(A|\mathcal{A}_0,x)$ for $A \in \mathcal{A}$ and $x \in \mathcal{X}$ is the conditional probability of A given "the most accurate description of x by means of statements from $\mathcal{A}_0$". It is defined as the almost unique $\mathcal{A}_0$-measurable function of x which satisfies

$$P(A \cap B) = \int_B P(A|\mathcal{A}_0,x)dP \qquad (1)$$

for all $B \in \mathcal{A}_0$. $P(A|\mathcal{A}_0,x)$ does always exist and is almost uniquely defined by (1).

Let $y = Y(x)$ be a measurable function from $(\mathcal{X},\mathcal{A})$ into $(\mathcal{Y},\mathcal{B})$ and let $\mathcal{A}_0$ be generated by $Y(x)$, i.e. $\mathcal{A}_0 = Y^{-1}(\mathcal{B})$ is the set of all $A \in \mathcal{A}$ of the form $Y^{-1}(B)$ with $B \in \mathcal{B}$. Then $P(A|\mathcal{A}_0,x)$ is a function of x only through $Y(x)$ and is called the conditional probability of A given $Y(x)$.

$P(A|\mathcal{A}_0,x)$ will in general almost satisfy the fundamental relations of a probability measure. But the null-set of all x for which a relation does not hold will in general depend on the special sets which enter into the relation. If, however, $\mathcal{X}$ is Euclidian and $\mathcal{A}$ the Borel-class, then $P(A|\mathcal{A}_0,x)$ can always be specified such that for any fixed x it is a probability measure over $\mathcal{A}$. See Lehmann [22] page 44.

Let $f(x)$ be integrable $(\mathcal{X},\mathcal{A},P)$, i.e. $Ef(X) = \int f(x)dP$ exists. The conditional expectation $E[f(x)|\mathcal{A}_0,x]$ relatively to $\mathcal{A}_0$ is the almost unique $\mathcal{A}_0$ measurable function of x which satisfies

$$\int_B f(x)dP = \int_B E[f(x)|\mathcal{A}_0,x]\,dP \tag{2}$$

for all $B \in \mathcal{A}_0$. The conditional expectation of any integrable f does always exist and satisfies "almost" the usual rules of operations for unconditional expectations. Furthermore

$$Ef(X) = EE[f(X)|\mathcal{A}_0,X] \tag{3}$$

In connection with a rigorous treatment of "conditioning" of tests and unbiased tests, the following results are important. If $\mathcal{X}$ is Euclidian and $\mathcal{A}$ the Borel class, then the conditional expectation can be specified as a proper expectation relatively to the conditional probability measure, such that

$$E[f(X)|\mathcal{A}_0,x] = \int f(\cdot)dP(\cdot\,|\mathcal{A}_0,x) \tag{4}$$

If in addition $\mathcal{A}_0$ is generated by a statistic $Y(x)$ into a Euclidian space, then there exists a null-set N (independent of f and g) such that for all $x \notin N$, we have

$$E\left[g(Y(X))f(X)\,|\,Y(X)\right] = g(Y(X))E\left[f(X)\,|\,Y(X)\right] \tag{5}$$

for all f and g. Under the same condition we can introduce a conditional sample space for almost all y, i.e.

$$P(Y^{-1}(\{y\})\,|\,y) = 1 \quad \text{a.e.} \tag{6}$$

I. SUFFICIENCY.

A. Introductory survey.

What kind of informations can we extract from the statistical data X and what is relevant about X in view of the purpose of the investigation ? It is natural to say that X (alone) gives no information about a parameter θ if the probability distribution of X is independent of θ. This leads us to say that $Z(x)$ gives no information about θ in addition to what is contained in $Y(x)$, if the conditional distribution of $Z(x)$ given $Y(x)$ is independent of θ. If no $Z(x)$ gives additional information about θ when $Y(x)$ is known, then $Y(x)$ obviously contains all information about θ, and is called sufficient. If Y is sufficient, then the decision making can be based on Y alone. If a statistician prefers some $Z(x)$ anyhow, simply because he "likes" the operating characteristic, he may get his way. A suitable randomized experiment will bring him from Y to Z. (In the case of point estimation the operating characteristic is the sampling distribution of the estimate. The statistician does not want to commit himself to e.g. minimizing the variance. Example: $X_1,\dots,X_n$ are independent normal $(0,\sigma)$; σ unknown, $Y = \sum X_j^2$, $Z = \sum |X_j|$). This is the classical definition of sufficiency.

But the motivation above shows that also a decision theoretical definition of sufficiency is justified. If to any decision situation about θ and any decision procedure δ, there exists a procedure δ_0 which depends only on Y and have the same operating characteristic as δ, then Y is said to be sufficient.

The two definitions above are roughly equivalent. According to L. le Cam [16], Kolmogorov [14] has called attention to a third definition based on the Bayesian point of view. Y is sufficient if for any a priori probability distribution for θ the a posteriori probability distribution of θ is the same relatively to X as relatively to Y.

Both Fisher and Neyman had in mind minimal sufficiency when they talked about sufficiency. Assume that $X_1,\ldots,X_n$ are Bernoulli variables, i.e. independent and such that $Pr(X_i = 1) = p$, $Pr(X_i = 0) = 1-p$. Then $Y = \sum X_i$ is minimal sufficient in the sense that any $g(Y)$ which does not have a unique inverse, is not sufficient.

Neyman-Pearson [24] and Fisher were talking about specific sufficiency relatively to a certain parameter. Let $\theta = (\rho,\tau)$, where ρ is the decision parameter, i.e. the parameter which the decision situation is concerned with, whereas τ is the nuisance parameter. If $R(x)$ for any given τ is minimal sufficient, then $R(x)$ is specifically sufficient for ρ. Suppose e.g. that $X_1,\ldots,X_n$ are independent normal (ξ,σ). Then $\bar{X}$ is specifically sufficient for ξ. Whereas minimal sufficient statistics exist under very general conditions, this is far from being true of specific sufficient statistics. Thus in the example just given no specific sufficient statistic exists for σ. Because for any given ξ it is $\sum(X_j-\xi)^2$ which is the minimal sufficient statistic, but this is not a statistic if ξ is unknown.

It seems doubtful whether specific sufficiency in the sense taken above is an important concept in the decision theory. It is difficult to find any direct connection between this concept and decision functions. Consider e.g. Student's situation with testing of $\xi = 0$ (or constructing confidence interval for ξ). Then $\bar{X}$ is specifically sufficient, but in order to perform the testing we have to consider $\sum(X_j-\bar{X})^2$. It is of course $(\bar{X}, \sum(X_j-\bar{X})^2)$ which is the minimal sufficient statistic for the model. Fisher [9] was aware of the difficulty and introduced the concept of ancillary statistic. $T(x)$ is ancillary if it jointly with the specific sufficient statistic $R(x)$ is minimal sufficient and the probability distribution of $T(x)$ only depends on the nuisance parameter τ. Rao [25] and Basu [3] have proved some interesting mathematical properties about ancillary statistics

A very interesting approach from a statistical point of view, is due to D.A.S. Fraser [10]. He does not need the concept of ancillary statistic, instead he adds to the above definition of spesific sufficiency the property that the distribution of $R(x)$ shall be independent of the nuisance parameter. This is a rather restrictive property ($\bar{X}$ in the example above is then not

specifically sufficient). On the other hand he is then able to establish links with decision problems. He shows that by testings and point estimations concerning ϱ, the statistician may limit himself to procedures depending on $R(x)$ without loosing power or efficiency.

Below we shall expand upon some, but not all, of the ideas which we have sketched above.

B. The basic theory of sufficiency.

We shall briefly outline the theory and we base our presentation on the paper by Halmos and Savage [11]. Let $\mathcal{P}$ be a family of probability measures P for a random variable X over $(\mathcal{X},\mathcal{A})$. A subsigmafield $\mathcal{A}_0$ of $\mathcal{A}$ is sufficient for the family $\mathcal{P}$ if for all $A \in \mathcal{A}$ there exists a $\mathcal{A}_0$-measurable function $\varphi_A(x)$ which for all P is the conditional probability relatively to $\mathcal{A}_0$, i.e. for which

$$\varphi_A(x) = P(A|\mathcal{A}_0,x) \quad \text{a.e.} \quad (\mathcal{A}_0,P) \tag{1}$$

for all $P \in \mathcal{P}$.

This definition is equivalent to a corresponding definition expressed by means of the conditional expectation of a function f. "For all $A \in \mathcal{A}$" is replaced by "for all $(\mathcal{A},P)$ - integrable f".

We shall limit ourself to families $\mathcal{P}$ which are dominated, by which we mean that there exists a sigmafinite measure μ such that all $P \in \mathcal{P}$ are absolutely continuous with respect to μ. It is obvious that if such a μ exists, it can be chosen finite.

From a practical point of view the limitation to dominated families seems to be unimportant. It can be proved (with some difficulty) that the following holds.

Theorem 1. If the family of probability measures $\mathcal{P}$ is dominated, then there exists a (finite or countable) subfamily $\mathcal{P}_0 = \{P_1,P_2,\dots\}$ of $\mathcal{P}$ such that for any A for which $P_i(A) = 0$; $i = 1,2,\dots$; we also have $P(A) = 0$ for all $P \in \mathcal{P}$.

Such a $\mathcal{P}_0$ will be called a dense subfamily in $\mathcal{P}$. It is seen that if $a_i > 0$, $\sum a_i = 1$, then

$$\Pi = \sum a_i P_i \tag{2}$$

is a probability measure which is 0 if and only if P is 0 for all $P \in \mathcal{P}$. Π will be called a dense measure for $\mathcal{P}$. Π need not belong to $\mathcal{P}$.

We have the following fundamental results about sufficient subsigmafields.

Theorem 2. Let $\mathcal{P}$ be a family of dominated probability measures over $(\mathcal{X},\mathcal{A})$ and denote by Π a dense measure for $\mathcal{P}$. A subsigmafield $\mathcal{A}_0$ is sufficient if and only if for any $P \in \mathcal{P}$ there exists a $\mathcal{A}_0$-measurable function $g_P(x)$ such that

$$dP = g_P(x)d\Pi \tag{3}$$

Proof: By theorem 1 (3) is true with g_P $\mathcal{A}$-measurable. We have to prove that it is necessary and sufficient that g_P has an $\mathcal{A}_0$-measurable version. Suppose $\mathcal{A}_0$ suffisient. We can then find a $\varphi_A(x)$ as in (1). We now have with $A \in \mathcal{A}$ and $B \in \mathcal{A}_0$;

$\Pi(A \cap B) = \sum a_i P_i(A \cap B) = \sum a_i \int_B \varphi_A dP_i = \int_B \varphi_A d \sum a_i P_i =$

$= \int_B \varphi_A(x)d\Pi$, showing that φ_A is the conditional probability also relatively to Π. Let E denote expectation relatively to Π and write $\bar{g}_P = E(g_P|\mathcal{A}_0)$. Then we have for $A \in \mathcal{A}$, remembering that $\bar{g}_P$ and φ_A are $\mathcal{A}_0$ measurable,

$$\int_A g_P d\Pi = P(A) = \int \varphi_A dP = \int \varphi_A g_P d\Pi = E\,\varphi_A g_P = EE[\varphi_A g_P|\mathcal{A}_0] =$$

$$= E\varphi_A E[g_P|\mathcal{A}_0] = E\varphi_A \bar{g}_P = E\bar{g}_P E(I_A|\mathcal{A}_0) = EE(\bar{g}_P I_A|\mathcal{A}_0) = E\bar{g}_P I_A =$$

$= \int_A \bar{g}_P d\Pi$, where I_A denotes the indicator function for A.

We have proved that $\int_A g_P d\Pi = \int_A \bar{g}_P d\Pi$ for all A. Hence $g_P = \bar{g}_P$ a.e., which shows that g_P can be specified as an $\mathcal{A}_0$-

measurable function. - Suppose now that g_P in (3) is $\mathcal{A}_0$-measurable. We then have for $A \in \mathcal{A}$ and $B \in \mathcal{A}_0$, $P(A \cap B) = \int_B P(A|\mathcal{A}_0)dP$. Let now on the other hand $\Pi(A|\mathcal{A}_0,x)$ denote the conditional probability relatively to Π, then $P(A \cap B) = EI_A I_B g_P = Eg_P I_B E(I_A|\mathcal{A}_0) = Eg_P I_B \Pi(A|\mathcal{A}_0) = \int_B \Pi(A|\mathcal{A}_0) g_P d\Pi = \int_B \Pi(A|\mathcal{A}_0)dP$. Comparing the two expressions for $P(A \cap B)$ we see that $P(A|\mathcal{A}_0,x) = \Pi(A|\mathcal{A}_0,x)$ a.e. $(\mathcal{A}_0,P)$. Hence we have obtained (1) with $\varphi_A = \Pi(A|\mathcal{A}_0)$. Thus $\mathcal{A}_0$ is suffisient Q.E.D.

Theorem 3. Let $\mathcal{P}$ be a family of probability measures over $(\mathcal{X},\mathcal{A})$ dominated by a sigmafinite measure μ. A subsigmafield $\mathcal{A}_0$ is sufficient for $\mathcal{P}$ if and only if there exists an $\mathcal{A}$-measurable non-negative function $h(x)$ and for each $P \in \mathcal{P}$ an $\mathcal{A}_0$-measurable function g_P such that

$$dP = g_P(x)h(x)d\mu \tag{4}$$

If this is true, then $h(x)$ can always be chosen integrable.

Proof: The necessity of (4) follows immediately from theorem 2 equation (3) and the fact that we can write $d\Pi = h(x)d\mu$. $h(x)$ is integrable since Π is finite. - Suppose now that (4) holds. With Π dense for $\mathcal{P}$ we have $d\Pi = \sum a_i dP_i = h \sum a_i g_{P_i} d\mu = hkd\mu$ where $k(x) \geq 0$ and $\mathcal{A}_0$-measurable. Let $N = \{x|k(x) = 0\}$. Then for $x \in N$ we have $\sum a_i g_{P_i} = 0$, hence $g_{P_i} = 0$, hence $P_i(N) = 0$, hence $P(N) = 0$ and $k(x) > 0$ a.e. $\mathcal{P}$. Thus we may write $dP = g_P h d\mu = \frac{g_P}{k} khd\mu = \bar{g}_P d\Pi$ where $\bar{g}_P(x) = g_P(x)/k(x)$ for $x \notin N$. Hence $\mathcal{A}_0$ is sufficient by theorem 2.

Suppose now that $\mathcal{A}_0$ is generated by a statistic $Y(x)$ into $(\mathcal{Y},\mathcal{B})$, i.e. $\mathcal{A}_0 = Y^{-1}(\mathcal{B}) = \{Y^{-1}(B)|B \in \mathcal{B}\}$. It is then known that a real $\mathcal{A}$-measurable function of x is $\mathcal{A}_0$ measurable if and only if f depends measurably on x only through $Y(x)$. Hence a necessary and sufficient condition for $\mathcal{A}_0$ to be sufficient is that

$$dP = f_P(Y(x))h(x)d\mu \tag{5}$$

where f_P is $\mathcal{B}$ measurable and $h(x)$ is $\mathcal{A}$ -measurable. This is the famous factorization theorem about sufficiency. If $\mathcal{A}_0 = Y^{-1}(\mathcal{B})$ is suffisient we say briefly that the statistic $Y(x)$ is sufficient.

Returning to the general case of any subsigmafield $\mathcal{A}_0$ we have, regardless of whether $\mathcal{P}$ is dominated or not,

Theorem 4. Let $\mathcal{A}_0$ be sufficient for a family $\mathcal{P}$ of probability measures over the Borel class $\mathcal{A}$ in the Euclidian space $\mathcal{X}$. Then there exists a function $\Pi(A|\mathcal{A}_0,x)$ of $A\in\mathcal{A}$ and $x\in\mathcal{X}$, which is $\mathcal{A}_0$-measurable for all A and a probability measure for all x, and which is a conditional probability for A relatively to $\mathcal{A}_0$ for all $P\in\mathcal{P}$.

The truth of this theorem follows by going through the usual proof of the result that a conditional probability can always be specified as a probability measure in the Euclidian case. It is also a special case of theorem 11 below.

C. Minimal sufficiency. Complete families of distributions.

We shall go into more details in this section since the theory seems not to be so generally known. We refer to Bahadur [1].

Let $\mathcal{P}$ be a family of probability measures over $(\mathcal{X},\mathcal{A})$ and let $\mathcal{A}_1$ and $\mathcal{A}_2$ be two subsigmafields. $\mathcal{A}_1$ is said to be almost a subsigmafield of $\mathcal{A}_2$; $\mathcal{A}_1 \underset{\text{a.e.}}{\subset} \mathcal{A}_2$; if to any $A_1\in\mathcal{A}_1$ there corresponds an $A_2\in\mathcal{A}_2$ such that $P[(A_1-A_2)\cup(A_2-A_1)] = 0$ for all $P\in\mathcal{P}$. The subsigmafields are equivalent if they are almost subsigmafields of each other; $\mathcal{A}_1 \underset{\text{a.e.}}{=} \mathcal{A}_2$.

A sufficient subsigmafield $\mathcal{A}_0$ is said to be the most summary, or the minimal sufficient subsigmafield if it is almost a subsigmafield of any sufficient subsigmafield. If a statistic $Y(x)$ generates such a subsigmafield, then $Y(x)$ is said to be a minimal sufficient subsigmafield.

Theorem 5. Let $\mathcal{P}$ be a dominated family of probability measures, let Π be dense for $\mathcal{P}$ and $dP = g_P(x)d\Pi$. Furthermore let $\mathcal{T}$ be the class of all sets of the form $A_P(r) = \{x | g_P(x) \leq r\}$ where $P \in \mathcal{P}$ and $0 \leq r < \infty$. Then the least sigmafield $\mathcal{A}_0$ over $\mathcal{T}$ is the most summary subsigmafield for $\mathcal{P}$.

Proof: From the construction of $\mathcal{A}_0$ it is seen that $g_P(x)$ is $\mathcal{A}_0$-measurable. Hence by theorem 2 $\mathcal{A}_0$ is sufficient.

Suppose now that $\mathcal{A}_1$ is an arbitrary sufficient sigmafield for $\mathcal{P}$. Then by theorem 2 there exists an $h_P(x)$ which is $\mathcal{A}_1$-measurable and such that $dP = h_P(x)d\Pi$ and thus $h_P = g_P$ a.e. (Π).

Define now K as the class of all $A \in \mathcal{A}$ such that for some $B \in \mathcal{A}_1$, $P[(A-B) \cup (B-A)] = 0$ for all $P \in \mathcal{P}$. It is seen that K is a sigmafield. But with $B_P(r) = \{x | h_P(x) \leq r\}$ we have, since $h_P = g_P$ a.e. (Π), that

$$[A_P(r)-B_P(r)] \cup [B_P(r)-A_P(r)]$$

has Π-measure 0 and hence P'-measure 0 for any $P' \in \mathcal{P}$. Hence $A_P(r) \in K$, i.e. $\mathcal{A}_0 \subset K$ since $\mathcal{A}_0$ is the least sigmafield over all $A_P(r)$. But from the construction of K it follows that K is almost a subsigmafield of $\mathcal{A}_1$. Hence $\mathcal{A}_0 \underset{\text{a.e.}}{\subset} \mathcal{A}_1$. Q.E.D.

It is seen that in the first place theorem 5 says that for any dominated family of distributions there always exists a minimal sufficient subsigmafield. In the second place it gives a manner of constructing such fields and thus (if possible) minimal sufficient statistics.

From theorem 5 it now easily follows

Theorem 6. Let $\mathcal{P} = \{P_\tau | \tau \in \Omega\}$ be a Darmois-Koopman family with

$$dP_\tau = A(\tau)\, e^{\sum_{j=1}^{s} \tau_j Y_j(x)}\, h(x)d\mu$$

where μ is sigmafinite. If Ω contains s linearly independent vectors then $Y(x) = \{Y_1(x),\ldots,Y_s(x)\}$ is a minimal sufficient

statistic for $\mathcal{P}$, i.e. for τ.

Example 1. In Student's (a priori) situation the first two moments, taken together, is a minimal sufficient statistic.

Example 2. Under the Fisher-Behrens' null-nypothesis the components of $X = (V_1,\ldots,V_p,W_1,\ldots,W_q)$ are independent normal, $EV_i = EW_j = \xi$, var $V_i = \sigma_1^2$, var $W_j = \sigma_2^2$. The four statistics $\sum V_i$, $\sum W_j$, $\sum V_i^2$, $\sum W_j^2$ taken together is then a minimal sufficient statistic for the three parameters $\xi, \sigma_1^2, \sigma_2^2$.

The construction of minimal sufficient statistics based on theorem 6 is rather inconvenient. The following result due to E.L. Lehmann and H. Scheffé [18] leads to a much easier manner of constructing minimal sufficient statistics. In this context we shall call any function $Y(x)$ from $(\mathcal{X},\mathcal{A})$ into a space (possibly abstract) $\mathcal{Y}$ a statistic. No sigmafield is defined in $\mathcal{Y}$ and no requirement about measurability of Y is imposed. Let $\mathcal{A}_Y$ be the class (sigmafield) of all sets in $\mathcal{A}$ of the form $Y^{-1}(B)$ where $B \subset \mathcal{Y}$. Then $\mathcal{A}_Y$ is said to be generated by Y. The contours of $Y(x)$ is the sets of the form $\{x|Y(x) = y\}$. To any partitioning of the space $\mathcal{X}$ there corresponds a statistic $Y(x)$ such that $Y(x)$ has the sets in the partitioning as contours.

Assume now that the family $\mathcal{P}$ of measures P over $(\mathcal{X},\mathcal{A})$ is dominated by a sigmafinite measure μ, i.e. $dP = f_P(x)d\mu$. A certain statistic Y_o is defined by defining its contours. A contour through $x_o \in \mathcal{X}$ is the set of all x for which there exists an $h(x,x_o) \neq 0$ such that

$$f_P(x) = h(x,x_o)f_P(x_o) \tag{6}$$

for all $P \in \mathcal{P}$. Then Y_o is said to be contour constructed by means of $\mathcal{P}$.

Theorem 7. Let $\mathcal{P}$ be a dominated family of probability measure P over a Borel-class $\mathcal{A}$ in the Euclidian space $\mathcal{X}$ and let Y_o be contour constructed by means of $\mathcal{P}$. We then have: (i), Y_o is sufficient; (ii), Y_o is minimal sufficient

in the sense that given any sufficient statistic $Y(x)$ there exists a function $t(\ )$ such that $Y_0(x) = t(Y(x))$ almost everywhere (μ) and hence almost everywhere $(\mathcal{P})$; (iii), $\mathcal{A}_{Y_0}$ is minimal sufficient in the sense defined earlier.

(i) follows easily from the factorization condition (4). For the proof of (ii) the reader is refered to [18]. It was shown by R.R. Bahadur [2] that two statistics $Z(x)$ and $Y(x)$ are such that $Z(x) = t(Y(x))$ almost everywhere (μ) if and only if $\mathcal{A}_Z$ is almost a subsigmafield of $\mathcal{A}_Y$. Furthermore it was shown that any subsigmafield is equivalent to a subsigmafield generated by a statistic. Hence the result (iii) says neither more nor less than (ii). (Both results of Bahadur assumes that $(\mathcal{X},\mathcal{A})$ is Euclidian-Borel.)

The contour construction given above is very convenient to apply. As a matter of fact it amounts to "looking at the mathematical form of the probability density" and identifying the minimal sufficient statistic almost immediately. Thus theorem 6 could have been proved by the contour construction method.

We shall now introduce the important concept of completeness which is borrowed from the functional analysis. This is a was first introduced by E.L.Lehmann and H.Scheffé in 1947 [16a]. See also [18] convenient mathematical property of families of probability measures with which the statisticians are frequently dealing. These families are perhaps often preferred for just that reason. But completeness in itself cannot be said to constitute a fundamental idea in statistical inference. A family is said to be <u>(boundedly) complete</u> over a sigmafield $\mathcal{A}_0$ if for any (bounded) $\mathcal{A}_0$-measurable function $f(x)$ for which $\int f(x)dP = 0$ for all $P \in \mathcal{P}$, we have $f(x) = 0$ a.e. $(\mathcal{P})$. Note that it is usually not relatively to the original observations $(\mathcal{A})$ that a family is complete, but relatively to some statistic $(\mathcal{A}_0)$. Thus if in theorem 6, Ω contains an open subset in the s-dimensional space, then the family of probability measures for $Y(x)$ is complete; i.e. the family of probability measures is complete over the subsigmafield $\mathcal{A}_0$ generated by $Y(x)$. In this example $Y(x)$ was also sufficient. In fact, there is an important connection between sufficiency and completeness (theorems 7 and 8 below) which we shall now prove through some lemmas.

We shall call $f(x)$ almost $\mathcal{A}_1$-measurable if there exists a sigmafield $\mathcal{A}_2$ equivalent to $\mathcal{A}_1$ such that $f(x)$ is $\mathcal{A}_2$-measurable.

Lemma 1. If $\mathcal{P}$ is (boundedly) complete over a subsigmafield $\mathcal{A}_0$, $g(x)$ is (bounded and) almost $\mathcal{A}_0$-measurable, and furthermore $\int g(x)dP = 0$ for all $P \in \mathcal{P}$, then $g(x) = 0$ a.e. $(\mathcal{P})$.

Proof: Let $\mathcal{A}_1$ be equivalent to $\mathcal{A}_0$ and $g(x)$ $\mathcal{A}_1$-measurable. Then there exists an $h(x)$ which is $\mathcal{A}_0$-measurable and equal to $g(x)$ a.e. This is seen by first assuming that g is an indicator function and then extend to any integrable function g. Hence $\int h(x)dP = 0$ for all P. Hence $g = h = 0$ a.e.

Lemma 2. If $\mathcal{A}_1$ and $\mathcal{A}_2$ are sufficient subsigmafields such that for all $A \in \mathcal{A}$

$$P(A|\mathcal{A}_1,x) = P(A|\mathcal{A}_2,x)$$

a.e. (P) for all $P \in \mathcal{P}$, then $\mathcal{A}_1$ and $\mathcal{A}_2$ are equivalent.

Proof: We have in particular for $A \in \mathcal{A}_1$

$$P(A|\mathcal{A}_2,x) = P(A|\mathcal{A}_1,x) = I_A(x) \quad \text{a.e.} \tag{7}$$

hence $P(A|\mathcal{A}_2,x) = 0$ or 1 a.e. Let now

$$B = \{x|P(A|\mathcal{A}_2,x) = 1\} \tag{8}$$

where $P(A|\mathcal{A}_2,x)$ and hence B can be specified to be independent of P. Then $B \in \mathcal{A}_2$ and $I_B(x) = I_A(x)$ a.e. or $P((A-B) \cup (B-A)) = 0$.

Theorem 8. Suppose that there exists a minimal sufficient subsigmafield for $\mathcal{P}$ and that $\mathcal{P}$ is boundedly complete over a sufficient subsigmafield $\mathcal{A}_0$. Then $\mathcal{A}_0$ is minimal sufficient for $\mathcal{P}$.

Theorem 9. If $\mathcal{P}$ is dominated and boundedly complete over a sufficient subsigmafield $\mathcal{A}_0$, then $\mathcal{A}_0$ is minimal sufficient for $\mathcal{P}$.

Proof: Theorem 9 is an immediate consequence of the theorems 5 and 8. Hence it remains to prove theorem 8.

Suppose that $\mathcal{A}_1$ is minimal sufficient and let

$$V(x) = P(A|\mathcal{A}_0,x)-P(A|\mathcal{A}_1,x) \tag{9}$$

V can be specified to be independent of $P\in\mathcal{P}$ since $\mathcal{A}_1$ and $\mathcal{A}_0$ are sufficient. Furthermore $|V(x)| \leq 1$ a.e., hence bounded and

$$EV(x) = \int V(x)dP = 0 \quad \text{for all} \quad P\in\mathcal{P} \tag{10}$$

Define now $\mathcal{A}_0'$ as the class of all $B\in\mathcal{A}$ for which there exists an $A\in\mathcal{A}_0$ such that $P[(A-B)\cup(B-A)] = 0$ for all $P\in\mathcal{P}$. Then $\mathcal{A}_0'$ is a sigmafield and by the definition of $\mathcal{A}_0'$, we have $\mathcal{A}_0 \subset \mathcal{A}_0'$. On the other hand we see that $\mathcal{A}_0' \subset \mathcal{A}_0$ almost by the definition of " $\subset$ almost". Hence $\mathcal{A}_0' = \mathcal{A}_0$. But from the definition of minimal sufficiency $\mathcal{A}_1 \subset \mathcal{A}_0$ almost, hence $\mathcal{A}_1 \subset \mathcal{A}_0'$ (without "almost"). Hence the last term in (9) is $\mathcal{A}_0'$-measurable, hence almost $\mathcal{A}_0$-measurable. Since the first term obviously is $\mathcal{A}_0$-measurable, it follows that V is almost $\mathcal{A}_0$-measurable. Thus by lemma 1, $V(x) = 0$ a.e., and by lemma 2 $\mathcal{A}_0 = \mathcal{A}_1$ almost, Q.E.D.

That vica versa minimal sufficiency does not imply completeness is easily seen by reference to the Behrens-Fisher nullhypothesis, see example 2 above. Here $E(\bar{V}-\bar{W}) = 0$ for all (ξ,σ_1,σ_2) despite the fact that $\bar{V} \neq \bar{W}$ a.e.

It is common practice to say that a statistic Y is complete when the class of sampling distributions for Y, i.e. $\{PY^{-1}|P\in\mathcal{P}\}$, is complete.

The following example shows that minimal sufficiency does not even imply boundedly completeness.

Example 3. We shall utilize this example to illustrate the construction of minimal sufficient sigmafields by means of

theorem 5. The construction could also have been performed by applying theorem 7.

Assume that $X = (X_1, X_2, \ldots, X_n)$ has components which are independent, each uniformly distributed over $(\varrho, \varrho+1)$, where $\varrho > 0$ is unknown. Let $Y_1 = \min X_j$, $Y_2 = \max X_j$. It is seen that the density of X can be written

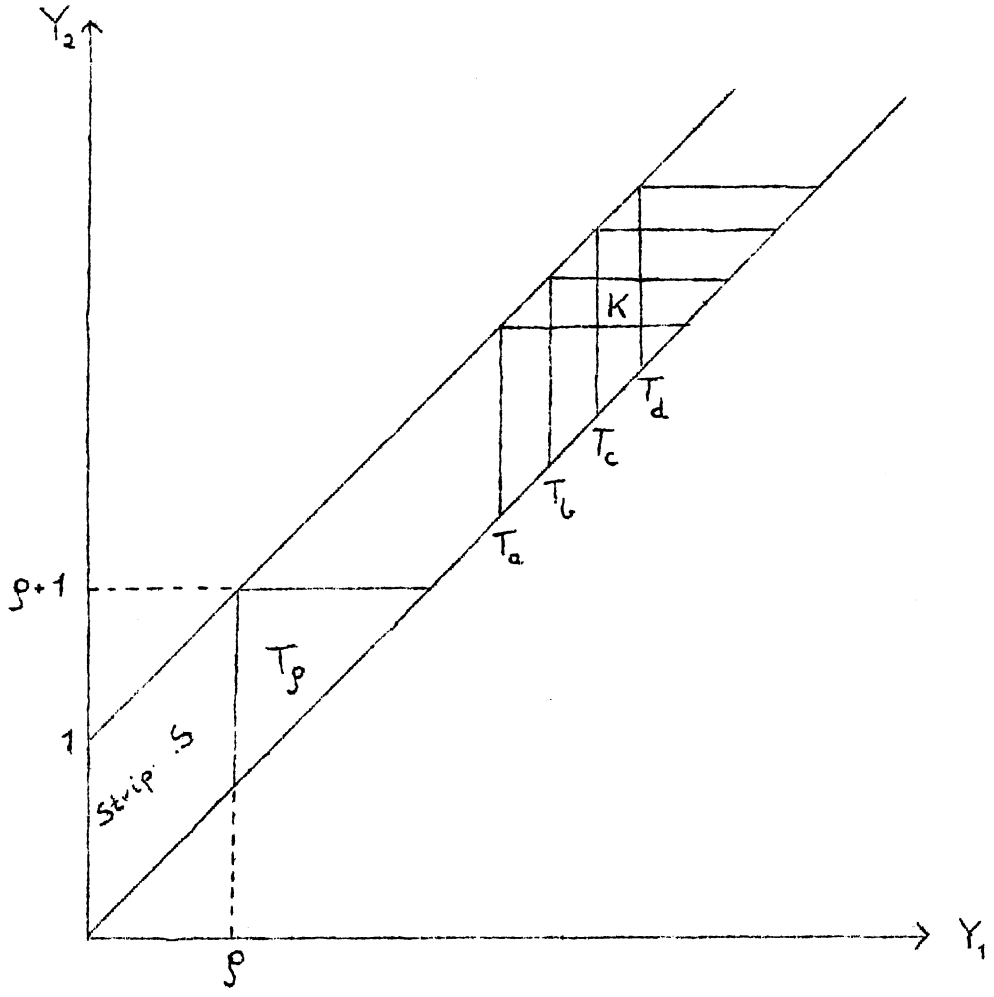
$$f_\varrho = 1 \quad \text{if} \quad \varrho < Y_1 < Y_2 < \varrho+1 \qquad (11)$$
$$= 0 \quad \text{otherwise.}$$

From theorem 3 (see equation (4)) it follows that $Y = (Y_1, Y_2)$ is sufficient. It is not boundedly complete, since

$$E_\varrho(Y_2 - Y_1 - \tfrac{n-1}{n+1}) = 0 \quad \text{for all } \varrho \qquad (12)$$

and $\left| Y_2 - Y_1 - \frac{n-1}{n+1} \right| < 1$ with probability 1 for all ϱ. Let us introduce $P_\varrho(A) = \int_A f_\varrho \, dx_1, \ldots, dx_n$ and let $\varrho_1, \varrho_2, \ldots$ be the set of all positive rational numbers. Then $\{P_{\varrho_j} \mid j = 1,2,\ldots\}$ is dense in $\mathcal{P} = \{P_\varrho \mid \varrho > 0\}$ in the sense of absolute continuity (see remark in connection with theorem 1). Then $\Pi(A) = \int_A h \, dx_1, \ldots, dx_n$ where $h = \sum f_{\varrho_j} 2^{-j}$, is dense for $\mathcal{P}$ and h is a probability density which is dense on the strip S (see figure). Obviously $S \cap \bar{T}_\varrho = \{x \mid g_\varrho < \frac{1}{2}\} \in \mathcal{T}$ (where bar denotes complimentation and the letters on the figure represent inverse images in the $\mathcal{X}$-space of the triangles etc. shown on the figure). Hence S = the strip $= \bigcup_j (\bar{T}_{\varrho_j} \cap S) \in \mathcal{A}_0$, i.e. $T_\varrho = S - S \cap \bar{T}_\varrho \in \mathcal{A}_0$, hence any rectangle $K \subset S$ can be expressed as $\bar{T}_a \cap T_b \cap T_c \cap \bar{T}_d \in \mathcal{A}_0$ (see figure). This shows that $\mathcal{A}_0$ consists of all sets $\{x \mid (y_1, y_2) \in B\}$ where B is a Borel-set in S. Thus (Y_1, Y_2) <u>is minimal sufficient for ϱ</u>.

In the case $n = 1$ it is easily seen that X is minimal sufficient whereas $E_\varrho f(X) = 0$ for all ϱ if $f(x)$ is bounded and periodic with period 1 and such that $\int_0^1 f \, dx = 0$. This is a very simple example of the fact that minimal sufficiency does not imply bounded completeness.



In order to throw some light on the concept of ancillary statistics we shall give some theorems, mainly due to Basu [3]. A subsigmafield $\mathcal{A}_1$ is called ancillary for $\mathcal{P}$ if $P(A)$ is independent of $P \in \mathcal{P}$ for all $A \in \mathcal{A}_1$. (We can, if we like, think of $\mathcal{P}$ as being generated by varying the decision parameter, keeping the nuisance parameter fixed.) We omit the proofs, which are quite simple.

Theorem 10. Assume that $\mathcal{P}$ is boundedly complete over a sufficient sigmafield $\mathcal{A}_0$ and that $\mathcal{A}_1$ is ancillary for $\mathcal{P}$. Then $\mathcal{A}_1$ and $\mathcal{A}_0$ are independent. (If $P(A_0 \cap A_1) = P(A_0)P(A_1)$ for all $A_0 \in \mathcal{A}_0$ and $A_1 \in \mathcal{A}_1$, then $\mathcal{A}_0$ and $\mathcal{A}_1$ are said to be independent.)

Examples: $X_1,\ldots,X_n$ are independent normal (ξ,σ). The sample mean $\bar{X}$ is sufficient and complete for fixed σ, whereas the distribution of $Z = \sum(X_j-\bar{X})^2$ is independent of ξ. Hence Z and $\bar{X}$ are independent. We have another example if $\xi = 0$. Then $V = \sum X_i^2$ is sufficient and complete whereas the distribution of the Student statistic T is independent of σ. Hence V and T are stochastically independent when $\xi = 0$.

The following theorem is a partial converse of theorem 10. In this connection we shall call two distributions P_1 and P_2 singular if there exists a set $N \in \mathcal{A}$ such that $P_1(N) = 0$, $P_2(N) = 1$.

Theorem 11. Assume that no two distributions of $\mathcal{P}$ are singular, that $\mathcal{A}_0$ is a sufficient sigma-field and that $\mathcal{A}_1$ is independent of $\mathcal{A}_0$. Then $\mathcal{A}_1$ is ancillary for $\mathcal{P}$.

D. General decision theory and sufficiency.

We shall first give a description of what is meant by a statistical decision problem. As usual $(\mathcal{X},\mathcal{A})$ will be the sample space for the observations X and $\mathcal{P}$ is the model, i.e. the family of probability distributions over the sample space. $\mathcal{P}$ expresses the a priori knowledge of the statistician.

The statistical investigation shall result in a decision d. The purpose of the investigation is defined by specifying a class R_d of decisions which a priori is feasible and of interest. We define a sigmafield $\mathcal{D}$ in R_d which contains all one-point sets.

A non-randomized decision function is a measurable function from the sample space to the decision space $(R_d,\mathcal{D})$ which for all $X \in \mathcal{X}$ gives the decision $d \in R_d$ to be taken. More generally we shall define a randomized decision function $\delta(D|x)$ as a function from $\mathcal{D} \times \mathcal{X}$ to the interval $[0,1]$. For each x it specifies the random mechanism according to which for any x we make a decision in R_d. It can be considered as the conditional probability distribution for d given X. Hence the unconditional distribution of d when $P \in \mathcal{P}$ is true, is given

by

$$\beta_\delta(D,P) = \Pr(d \in D) = \int \delta(D|x)\,dP \tag{13}$$

This is the operating charactheristic for the decision function δ. By studying this characteristic as a function of D and P we get an impression of how good δ is. When we are studying δ in this manner we take into consideration that for each P there are some decisions d which are desirable and some not desirable. We could express this circumstance by introducing a loss function, but at present we shall refrain from doing that, since we shall obtain certain results which are independent of the loss function.

By testing hypothesis the operating characteristic is given by the power function, by point estimation it is simply given by the sampling distribution of the point estimator.

Two decision functionsare said to be equivalent if they have the same operating characteristic.

The principle of sufficiency is now to the effect that if $\mathcal{A}_0$ is sufficient then $\delta(D|x)$ should be $\mathcal{A}_0$-measurable for each D.

We have now the following useful result proved by Bahadur [1].

Theorem 12. Suppose R_d Euclidian, $\mathcal{D}$ the Borel class and let $\mathcal{P}$ be an arbitrary class of probability measures P over $\mathcal{A}$. The following statements are the equivalent

(i) $\mathcal{A}_0$ is sufficient.

(ii) For any $\mu(D,x)$ defined for $D \in \mathcal{D}$ and $x \in \mathcal{X}$ which is a measure as a function of D and $(\mathcal{A},\mathcal{P})$-integrable as a function of x; there exists a $\mu_0(D,x)$ which is a measure as a function of D and $(\mathcal{A}_0,\mathcal{P})$-integrable as a function of x and which is such that

$$\mu_0(D,x) = E_P[\mu(D,X)|\mathcal{A}_0,x] \quad \text{a.e.} \quad (\mathcal{P}) \tag{14}$$

for all $D \in \mathcal{D}$ and $P \in \mathcal{P}$.

The proof of this theorem is quite similar to Doob's proof of the existence of a conditional probability measure, see section

O. B above. As a matter of fact this result appears as a special case of theorem 12 by letting $\mathcal{P}$ consist of one element P. Then every subsigmafield $\mathcal{A}_0$ is sufficient and the result follows immediately by letting $\mu = I_D(x)$, $(R_d, \mathcal{D}) = (\mathcal{X}, \mathcal{A})$ and $D = A \in \mathcal{A}$.

Furthermore by letting $\mu(D,x)$ be a decision function $\delta(D|x)$ we get

Theorem 13. Suppose that R_d is Euclidian and $\mathcal{D}$ the Borel class. If and only if $\mathcal{A}_0$ is sufficient is it possible to replace an arbitrary decision function $\delta(D|x)$ by an $\mathcal{A}_0$-measurable decision function $\delta_0(D|x)$ which is equivalent to δ. δ_0 and δ are related as μ_0 and μ in (14).

We now have for an arbitrary decision space

Theorem 14. Let $\mathcal{X}$ be Euclidian and $\mathcal{A}$ the Borel class. If and only if $\mathcal{A}_0$ is sufficient is it possible to replace an arbitrary decision function $\delta(D|x)$ by an $\mathcal{A}_0$-measurable decision function $\delta_0(D|x)$ which is equivalent to δ.

Proof: We have for the operating characteristic, since $\mathcal{A}_0$ is sufficient

$$E_P\delta(D|X) = E_P E_P[\delta(D|X)|\mathcal{A}_0, X] = E_P E_\Pi[\delta(D|X)|\mathcal{A}_0, X] \qquad (15)$$

for $\Pi \in \mathcal{P}$. But from O. B (4) it now follows that the inner expectation can be considered as a proper expectation relatively to a conditional probability measure $\Pi(A|\mathcal{A}_0, x)$. We denote this conditional expectation by $\delta_0(D|x)$ and have then

$$\delta_0(D|x) = E_\Pi[\delta(D|X)|\mathcal{A}_0, x] = \int \delta(D|x')\, d\Pi(x'|\mathcal{A}_0, x) \qquad (16)$$

We now easily verify that δ_0 for all x is a probability measure as a function of D and consequently a decision function based on the principle of sufficiency. Its equivalence with δ follows from (15).

It is seen from theorems 12 and 13 that <u>if either the decision space or the sample space is Euclidian, then nothing</u>

is lost by using the principle of sufficiency, regardless of how the loss function in the decision problem is defined.

II. TESTS WITH OPTIMAL POWER.

A. Introduction.

Basic for the "classical" theory of testing hypothesis is the idea of fixing a certain level of significance ε, which is the maximal (or upper bound of the) probability of falsely rejecting the hypothesis. The idea is well established in statistical applications, but should hardly ever be applied with strict consistency. The level is a price to be paid for high sensitivity (power). If the statistical situation is such that the "market" is deluged by sensitive tests, then the "buyer" (the statistician) will lower the price which he is willing to pay. The idea is formalized by Lehmann [21]. We shall, however, in these lectures adapt the attitude that we wish to maximize the power for a given level.

There are certain simple situations, mainly where one wants to test a "simple" (completely specified) hypothesis against one-parametric one-sided alternatives, where a uniformly most powerful test at a given level exists. If one or more of the assumptions just mentioned are not fulfilled, then a uniformly most powerful test will usual not exist (but this is not a general rule), and in that case we shall be at loss what to do. In many cases we can, however, reduce the more complicated situation to a simpler situation by introducing some additional assumptions on the test methods. One resorts to one or more of the following devices: (i) the test is conditioned, (ii) it is made unbiased in the power, (iii) it is made invariant. We shall below treat the two first possibilities. But first we are going to say something about the technique of constructing tests with optimal power.

B. Neyman-Pearson constructed tests.

Let $(\mathfrak{X},\mathcal{A},\mu)$ be a measure space, let $f_1(x),\ldots,f_m(x)$, $f(x)$ $(= f_{m+1}(x))$ be given real integrable functions over this space and let $c_1,\ldots,c_m$ be given constants. We denote by Δ

the class of all test functions $\delta(x)$ over the measure space, i.e. measurable functions such that $0 \leq \delta(x) \leq 1$. Furthermore C is the set of all $\delta \in \Delta$ such that

$$\int \delta f_i d\mu = c_i; \quad i = 1,2,\ldots,m \tag{1}$$

We have the following classical result.

Theorem 1. Suppose that $\delta_0 \in C$ has the following property: There exist numbers $k_1,\ldots,k_m$ such that

$$\begin{aligned} \delta_0(x) &= 1 \quad \text{for all} \quad x \quad \text{with} \quad f(x) > \sum_{i=1}^{m} k_i f_i(x) \\ \delta_0(x) &= 0 \quad \text{for all} \quad x \quad \text{with} \quad f(x) < \sum_{i=1}^{m} k_i f_i(x) \end{aligned} \tag{2}$$

Then

$$\int \delta_0 f d\mu \geq \int \delta f d\mu \tag{3}$$

for all $\delta \in C$. For those i for which $k_i \geq 0$, replace "=" in (1) by "$\leq$" and call the new set of equations (1)'. Then (3) is true for any $\delta \in \Delta$ satisfying (1)'.

Suppose now that $\mathfrak{X}$ is Euclidian. About the existence of a δ_0 the following could be said. The class of all test functions is "power-compact" in the sense that to any sequence $\delta_1,\delta_2,\ldots \in \Delta$ there exists a subsequence $\delta_{n_1},\delta_{n_2},\ldots \in \Delta$ and a $\delta_0 \in \Delta$ such that

$$\int \delta_{n_j} g d\mu \to \int \delta_0 g d\mu \tag{4}$$

for any integrable g, hence in particular for $g = f_1,\ldots,f_{m+1}$. Thus if C is non-empty, there exists a maximizing $\delta_0 \in C$.

As a matter of fact we can under general assumptions about $c_1,\ldots,c_m$ say something more, viz. that δ_0 must be "almost" Neyman-Pearson constructed. We denote by M the class of all points $(\int \delta f_1 d\mu,\ldots,\int \delta f_m d\mu)$ which is generated by varying

δ in Δ. It is easily seen that M is convex. From the compactness property just mentioned it now follows that M is closed.

The property about δ_0 which we have announced can now be formulated as follows. (See Lehmann [22]).

Theorem 2. Suppose that $\mathcal{X}$ is Euclidian and that $(c_1,\ldots,c_m)$ is an inner point in M. Then there exists a $\delta_0' \in C$ which maximizes $\int \delta f d\mu$. To any such δ_0' there exists a $\delta_0 \in C$ which has the properties given in theorem 1 and such that it equals δ_0' almost everywhere (μ).

By means of this theorem we can in certain situations prove the uniqueness of the Neyman-Pearson constructed tests, in other cases it can be used to ascertain that tests which are unbiased with a certain level can always be constructed. From our point of view it is perhaps more important that the theorem can be used to prove the non-existence of uniformly most powerful tests.

C. Conditioning of tests.

We shall later discuss the possibility of a general formulation of conditioning of tests. At present we shall take as a starting point that conditioning in certain situations can be justified intuitively.

An important class of test situations is defined by the Darmois-Koopman class of distributions (see [6] and [15]). $P_{\tau,\varrho}$ is given by

$$dP_{\tau,\varrho} = A(\tau,\varrho)e^{\sum_{j=1}^{s}\tau_j Y_j(x)+\varrho V(x)}\, dP_0 \qquad (5)$$

$\tau = (\tau_1,\tau_2,\ldots,\tau_s)$ and ϱ vary independently. Without impairing generality we assume that $\tau = 0$ and $\varrho = 0$ are a priori admissible and that $A(0,0) = 1$. The sample space $(\mathcal{X},\mathcal{A})$ for the observations X is Euclidian. The functions Y_j and V are known a priori. We shall test the hypothesis

$\varrho = 0$ against $\varrho > 0$. Thus ϱ is the decision parameter whereas τ is the nuisance parameter. In certain special cases of Darmois-Koopman classes it is natural to use conditional test given $Y(x) = (Y_1(x),\ldots,Y_s(x))$ (see examples below).

It is seen that $(Y(x),V(x))$ is a sufficient statistic relatively to the a priori situation. Hence without impairing the power we can limit ourself to consider tests which depend on the observations only through (Y,V) (see theorem 13 in I. D). Denote by $F_o(y,v)$ the cumulative distribution function for (Y,V) when $\tau = 0$, $\varrho = 0$. It then follows from (5) that the cumulative sampling function $F_{\tau,\varrho}(y,v)$ for an arbitrary (τ,ϱ) is given by

$$dF_{\tau,\varrho} = A(\tau,\varrho)e^{\sum_{j=1}^{s}\tau_j y_j + \varrho v}\, dF_o \tag{6}$$

Furthermore, denote by $F_{\tau,\varrho}(v|y)$ and $F_o(v|y)$ the conditional cumulative functions given $Y = y$. We then get

$$dF_{\tau,\varrho}(v|y) = A_y(\varrho)e^{\varrho v}dF_o(v|y) \tag{7}$$

where

$$\left[A_y(\varrho)\right]^{-1} = \int_{-\infty}^{\infty} e^{\varrho v}dF_o(v|y)$$

Thus we have obtained a class of <u>one-parametric alternatives</u> and we shall test the <u>simple</u> (completely specified) hypothesis $\varrho = 0$.

<u>Theorem 3</u>. The a priori distribution of X is given by (5) and we want to test $\varrho \leq 0$ against $\varrho > 0$. There exists a uniformly most powerful conditional test with level ε. It consists in rejecting the null hypothesis when

$$V(X) > c(Y(X)) \tag{8}$$

and reject with probability $\pi(Y(X))$ when

$$V(X) = c(Y(X)) \tag{9}$$

and accept otherwise. $c(y)$ and $\pi(y)$ are given by

$$\varepsilon = 1-F_0(c(y)|y)+\pi(c(y))\left[F_0(c(y)|y)-F_0(c(y)-|y)\right] \tag{10}$$

It is obvious that this test also unconditionally has level ε, but unconditionally nothing is said directly about its optimal properties.

Example 1. X_1, X_2 are independent and Poisson distributed with parameters λ_1 and λ_2, and the null hypothesis is $\lambda_2 \leq a\lambda_1$. By conditioning w.r.t. $X = X_1+X_2$ it is seen that the testing can be carried out on a binomial distribution and that the test is conditionally optimal.

Example 2. Consider a double dichotomic frequency table and testing of independence both under the assumption of multinomial distribution and under assumptions of two binomial distributions. By conditioning given the marginals it is seen from theorem 3, that the well-known hypergeometric testing is obtained.

By double classification with more than two levels for at least one of the classifications we have a priori a Darmois Koopman family of distribution of type (5) with $\varrho V(x)$ replaced by $\sum_{i=1}^{t} \varrho_i V_i(x)$; $t \geq 2$. Independence corresponds to $\varrho_1 = \ldots = \varrho_t = 0$. Conditional testing, given the marginals, leads to testing on the basis of the generalized hypergeometric distribution. The Neyman-Pearson constructed test will depend on the alternative $(\varrho_1,\ldots,\varrho_t)$ and it follows from theorem 2 that no uniformly most powerful conditional test exists. Hence we have to be content with a compromise test, such as the chi-square goodness of fit test.

In the two examples given above it is perhaps felt that conditional testing is intuitively reasonable. The statistics on which the conditioning is based have precisely the property required of the "ancillary" statistic, viz. that they are important data to be taken into account when judging whether

there are significance, but that their distribution does not (or only very slightly) depend on the decision parameters.

The following example is somewhat different and the intuitive feeling in favour of conditioning is perhaps missing.

Example 3. $X_1,\ldots,X_n$ are independent normal (ξ,σ) and the null hypothesis is $\xi \leq K\sigma^2$ against $\xi > K\sigma^2$. Then $V = \sum_{i=1}^{n} X_i$, $Y = \sum_{i=1}^{n} X_i^2$ form a set of sufficient statistics for (ξ,σ). Y is not ancillary for ξ/σ^2 in this case since Y/σ^2 is eccentric chi-square distributed with n degrees of freedom and eccentricity $\lambda = n\xi^2/\sigma^2$. We consider conditional testing anyhow. We then obtain again a one-parametric situation, since the conditional distribution of V depends only on ξ/σ^2. Theorem 3 can be applied and we are led to a kind of conditional Student testing given $\sum X_i^2$. This test is the uniformly most powerful test among all conditional tests. For $K = 0$ we have the usual Student hypothesis. In that case the Student statistic happens to be independent of Y and we have the ordinary unconditional Student test, which therefore is uniformly most powerful among all <u>conditional</u> tests.

We have above discussed the situation in the case of some Darmois-Koopman families of distributions. We shall now consider the following <u>non-parametric situation</u>.

We assume that it is a priori known that $X_1,\ldots,X_n$ are independent with probability density respectively $f(x-t_i\varrho)$; $i = 1,2,\ldots,n$. The numbers $t_1,t_2,\ldots,t_n$ are a priori known and they are not all equal, but the functional form f and the scalar ϱ are unknown. We shall test that $X_1,\ldots,X_n$ are identically distributed, i.e. that $\varrho = 0$. Thus ϱ is the decision parameter, whereas f is a nuisance "parameter". We see that if $t_1 = t_2 = \ldots = t_m = 0$, $t_{m+1} = \ldots = t_n = 1$, then we have a non-parametric two sample situation, where we want to test if the two samples are from the same population.

Let us consider the order statistic $Y(X) = (Y_1(X),\ldots,Y_n(X))$, where $X = (X_1,\ldots,X_n)$ and $Y_1(X),\ldots,Y_n(X)$ are $X_1,\ldots,X_n$ ordered in a non-decreasing sequence. The distribution of Y is not independent of ϱ. Nevertheless, it is customary to recommend conditional testing given Y, i.e. "combinatorial" or "non-parametric" testing.

For the density of $X_1,\ldots,X_n$ we write for short

$$f(x_1-t_1\varrho)\ \ldots\ f(x_n-t_n\varrho) = p(x-t\varrho;f) \tag{11}$$

Let $R(y)$ be the set of all $n!$ permutations of $y = (y_1,\ldots,y_n)$. We then have

$$\Pr[X = x \mid Y = y] = p(x-t\varrho;f) / \sum_{x' \in R(y)} p(x'-t\varrho;f) \tag{12}$$

provided $x \in R(y)$. Otherwise the same probability is 0. (12) reduces to $1/n!$ when $\varrho = 0$. The conditionally most powerful test relatively to a given alternative (f,ϱ) can now be obtained by a Neyman-Pearson construction. We get the following test. For given y consider the $n!$ quantities $p(x-t\varrho;f)$ obtained by varying x in $R(y)$. We arrange them in a non-increasing sequence and denote them by $p^{(1)},\ldots,p^{(n!)}$. The corresponding values of x are denoted by $x^{(1)},\ldots,x^{(n!)}$. Determine k and $0 \leq \gamma < 1$ such that $k+\gamma = n!\,\varepsilon$. For any y the hypothesis is rejected if $Y(X) = y$ and X is one of the points $x^{(1)},\ldots,x^{(k)}$. If $Y(X) = y$ and $X = x^{(k+1)}$ the hypothesis is rejected with probability γ. Otherwise the hypothesis is accepted.

It is seen that the test depends on the special alternative (f,ϱ). Hence by theorem 2 there exists no uniformly most powerful non-parametric test (i.e. conditional test given Y). Thus also in this case we have to be content with some compromise test, such as e.g. the Wilcoxon test.

D. Unbiased tests.

As above we shall denote the sample point by X, the

sample space by $(\mathcal{X},\mathcal{A})$ and the a priori family of distributions P over $(\mathcal{X},\mathcal{A})$ by $\mathcal{P}$. Let $\mathcal{P}_H$ be a proper subfamily of $\mathcal{P}$. A test of the hypothesis $P \in \mathcal{P}_H$ is defined by $\delta(x) = \Pr(\text{rejection} \mid X = x)$. The power function is then $\beta(P,\delta) = \Pr(\text{rejection}) = \int \delta(x)dP$. The level is ε if $\beta(P,\delta) \leq \varepsilon$ for $P \in \mathcal{P}_H$. The test is unbiased if in addition $\beta(P,\delta) \geq \varepsilon$ for $P \in \mathcal{P} - \mathcal{P}_H$. We introduce some limit concept, $\lim P_n = P$, in $\mathcal{P}$ and denote by $\mathcal{P}_H'$ the boundary points of $\mathcal{P}_H$. Suppose now that for any δ the power of δ is a continuous function for all $P \in \mathcal{P}_H'$. Then unbiasedness implies similarity, i.e. $\beta(P,\delta) = \varepsilon$ for $P \in \mathcal{P}_H'$.

Consider in particular the Darmois-Koopman situation.

Theorem 4. Assume that the situation is as has just been described and that $\mathcal{P}_H'$ is a Darmois-Koopman family of distributions

$$dP_\tau = A(\tau)e^{\sum_{j=1}^{r} \tau_j Y_j(x)} dP_0 \tag{13}$$

obtained by varying τ in a set ω which contains an r-dimensional "box" (containing $\tau = 0$). A test is then unbiased if and only if it is a conditional test relatively to $Y(x)$, i.e.

$$E_0(\delta(X) \mid y) = \varepsilon \tag{14}$$

or equivalently

$$E_\tau(\delta(X) \mid y) = \varepsilon \tag{15}$$

where E_τ denotes expectation relatively to P_τ.

Thus it is seen that conditioning of tests in the examples 1-3 (and in many other situations) is "justified" by the requirement of unbiasedness. This also applies to tests which cannot be justified by Fisher's ancillary principle, such as the test in example 3.

We have a similar result about non-parametric situations.

Theorem 5. Consider the non-parametric situation described in the last part of section C. f is known to be continuous almost everywhere, otherwise f is unknown. Then a

test is unbiased if and only if it is a combinatorial test, i.e. a conditional test relatively to the order statistics, i.e.

$$\sum \delta(x) = \varepsilon n! \qquad (16)$$

where the sum is over all permutations of the order statistic. The most powerful test relatively to a given alternative is the one given at the end of section C.

By means of theorem 2 it now follows that <u>in the non-parametric situation there exists no uniformly most powerful unbiased test.</u>

We return to the Darmois-Koopman situation. We find from the theorems 3 and 4

<u>Theorem 6</u>. Suppose that the distribution of X is a priori given by (5) and that we want to test $\varrho \leq 0$ against $\varrho > 0$. The test given in theorem 3 is the uniformly most powerful unbiased test.

From this theorem it follows that the tests in examples 1, 2 (first part), 3 are uniformly most powerful unbiased. By application of the theorems 4 and 2 it follows that by doubly classified frequency tables with more than two levels for at least one of the classifications, there exists no uniformly most powerful test for the hypothesis of independence.

The main content of the famous paper by Neyman and Pearson [23] is given in a somewhat generalized and modernized version in the theorems 4 and 6 above. But Neyman and Pearson didn't make use of the concept of unbiasednes (only similarity) and the connection with conditional testing was clarified later (see [27] and [19]). Furthermore Neyman and Pearson didn't expressedly deal with the Darmois-Koopman family, but this was implicit in their assumptions.

E. <u>Justification of conditioning</u>.

Conditioning of tests has worried statisticians through many years. It has been felt that tests cannot be conditioned

arbitrarily. Certain principles are needed. The example 2 above is a classical example. Can we, without further ado, assume that the marginals are given non-random variables, regardless of whether they were chosen in advance in the statistical experiment or not ? (Usually they are not chosen in advance.) Another classical example is regression analysis (see example 5 below). Should the independent variables be taken as "given" ("fixed") variables or as stochastic variables ?

The problem can also be formulated as the problem of what is the "correct" sample space or as the problem of what is the "hypothetical repetitions".

If we are just looking for a definite rule, then we have got it in the principle of unbiasedness. It justifies in an elegant manner the tests in the examples 1,2 and 3 and in many other Poisson, multinomial and linear-normal situations. It also justifies combinatorial tests in non-parametric situations. However, in the cases where conditional tests seems intuitively reasonable (examples 1 and 2), one is not always convinced that the real motive is unbiasedness. Furthermore, in some cases one feels that the consequences of unbiasedness is not supported by intuition (example 3 above).

The following example illustrates the situation.

Example 4. The quantity ξ is to be measured in order to find out if $\xi = 0$ or $\xi > 0$. The actual measurement is denoted by X. There are two instruments available. With instrument I X is normal (ξ, σ_1) and with instrument II X is normal (ξ, σ_2). Let Y(= I or II) be the brand of the instrument. One intends to call on an institution which is known to have one of the instruments and perform the measurement there. It seems obvious that one should assert that $\xi > 0$ if the instrument is of brand I and $X > 1.64\sigma_1$ or if the instrument is of brand II and $X > 1.64\sigma_2$ (5 % level).

This is conditional testing given Y and it is clearly very reasonable. However, it could be objected that we are in the "wrong" sample space. We should really consider the space for (X,Y). Suppose now that $Pr(Y = I) = p$ and $Pr(Y = II) = 1-p$, where p is unknown. Thus we have a nuisance

parameter p in addition to the decision parameter ξ. By applying the principle of unbiasedness we are now easily led to the test given above.

But suppose now that we obtained the additional information that when the institution purchased the instrument a coin was tossed and that the outcome determined which instrument to buy. Hence $p = \frac{1}{2}$. This information ought to be quite irrelevant, since we are going to look at the instrument and observe of which brand it is. Thus conditional test should still be reasonable.

But now this result could not be obtained from the principle of unbiasedness. This principle was only helpful when p was unknown. It is easily seen that the most powerful test relatively to the alternative ξ is to reject the hypothesis if and only if

$$X > \frac{\xi}{2} + \frac{\sigma_Y^2}{\xi} K \tag{17}$$

where K is determined from

$$\frac{1}{2}\left[1-G\left(\frac{\xi}{2\sigma_1}+\frac{\sigma_1}{\xi}K\right)\right] + \frac{1}{2}\left[1-G\left(\frac{\xi}{2\sigma_2}+\frac{\sigma_2}{\xi}K\right)\right] = 0.05 \tag{18}$$

and where G is the gaussian integral. The test depends on ξ and there is no uniformly most powerful test. (See also Cox [5]).

Example 5. Suppose that in the conditional distribution given $V_1,\ldots,V_n$ we have that $X_1,\ldots,X_n$ are independent normal with variance σ^2 and expectations $\alpha+\beta V_i$; $i = 1,2,\ldots,n$, respectively. We want to test something about α,β,σ. If either (i), nothing is known about the distribution of $V_1,\ldots,V_n$ or (ii) they are independent and identically distributed, or (iii) they are independent normal (ν,τ) where ν and τ are unknown; then the situation is clear. Unbiasedness implies conditional testing given $V_1,\ldots,V_n$. If, however, (iv) $V_1,\ldots,V_n$ are independent normal $(0,1)$, then conditional testing can not be justified by means of the principle of unbiasedness. There are obviously other reasons for conditioning. The distribution of $V_1,\ldots,V_n$ depends in

neither of the situations (i), (ii), (iii) or (iv) on α, β, σ; but $V_1,\ldots,V_n$ are nevertheless important when judging significance. They play, according to Fisher, about the same role as the size n of the sample, and can therefore be taken as given.

A precise formulation of a principle of conditioning has been suggested by Cox [5] who tried to connect it with the principle of sufficiency.

We shall proceed in a different manner.

First, it should be remarked that the formulations such as "the distribution of the variables shall be independent of the decision parameter" is not precise. In example 1 it might perhaps be said that the distribution of $X = X_1+X_2$ is dependent of λ_1/λ_2 through $\lambda_1+\lambda_2 = \lambda_2(1+\lambda_1/\lambda_2)$. But the family of distributions of X is the same regardless of how λ_1/λ_2 is specified, or whether it is specified at all. Furthermore the notiens "decision parameter" and "nuisance parameter" are diffuse. In example 3 it is not the "whole" parameter $\frac{\xi}{\sigma^2} - K$ which is the decision parameter, only its sign. $\left|\frac{\xi}{\sigma^2} - K\right|$ is nuisance.

In order to condition given a statistic it seems reasonable to require that this statistic shall give us no intimation about whether the hypothesis is wrong or right. This requirement can be formulated as follows. As usual $\mathcal{P}$ denotesthe a priori family of distributions, $\mathcal{P}_H$, which is a proper subfamily of $\mathcal{P}$, denotes the family of distributions under the hypothesis and $\mathcal{P}_{alt}$ the family of distributions under the alternative. Let $\mathcal{A}_o$ be a subsigmafield in the sample space $(\mathcal{X},\mathcal{A})$. P^o is the measure P confined to $\mathcal{A}_o$ and $\mathcal{P}_H^o$ and $\mathcal{P}_{alt}^o$ are the families of all P^o generated by varying P in $\mathcal{P}_H$ and $\mathcal{P}_{alt}$ respectively. Now, one condition for conditioning on a statistic Y, which generates $\mathcal{A}_o$, should be that $\mathcal{P}_H^o = \mathcal{P}_{alt}^o$.

The merits of the principle of conditioning by ancillary statistics on the one hand side and of the principle unbiasedness in the power on the other, are compared in the following table of conditional tests which are and are not justified by the two principles.

Anc.princ.	Ex.1				Ex.4	Ex.5(iii)	Ex.5(iv)
Unb.	Ex.1	Ex.2	Ex.3	Non.par.		Ex.5(iii)	

[That the definition above is not complete is seen from the following example due to Else Sandved [26]. $X_1,\ldots,X_n$ are independent normal with unknown expectation ξ and unknown variance σ^2. The hypothesis is $\xi < 0$ against $\xi > 0$. Let $\bar{X}$ be the sample mean. Then $(|\bar{X}|, \sum (X_i-\bar{X})^2)$ is ancillary according to the definition above. Conditioning with respect to this statistic is seen to lead to absurd tests.]

In a general decision problem with an a priori family of distributions $\mathcal{P}$, a decision space R_d and a loss function $L(P,d)$, the requirement of an ancillary subsigmafield $\mathcal{A}_0$ is as follows. There exists a version $P(A|\mathcal{A}_0,x)$ of the conditional probability relatively to $\mathcal{A}_0$, such that $L(P,d)$ depends on P only through $P(\cdot|\mathcal{A}_0,x)$. Expressed differently, let P^0 denote P restricted to $\mathcal{A}_0$ and $\mathcal{P}^0$ the family of all P^0 generated by varying P in $\mathcal{P}$. If now $P(\cdot|\mathcal{A}_0,x)$ is kept fixed and P^0 varies in $\mathcal{P}^0$, then $L(P,d)$ should be constant for all

$$P(\cdot) = \int P(\cdot|\mathcal{A}_0,x)\,dP_0.$$

III. MULTIPLE DECISION THEORY

A. Introduction.

Statistical theory has predominantly been preoccupied with the following two types of situations.

(i). Testing hypothesis, i.e. choice between two decisions which can be either,
   a. Rejection or acceptance of the hypothesis H,
   b. Rejection or not rejection of H.

(ii). Point-estimation.

(iii). Interval-estimation (of one parameter).

Many methods with nice optimum properties have been developed for these situations.

In practical statistics the situation is often more complicated. In connection with a specific statistical investigation it may be necessary to perform several tests and point estimations. (We shall below only deal with testings and point estimations, not interval estimations.) We then have a multiple decision problem. If in such situations we are combining well-known tests and estimation methods, we are loosing control with what we are really doing. We don't know in which sense the combined method is good. (A more elementary error is to betray oneself with regard to the level of the combined test.) One would perhaps be inclined to think that such a problem should be reconsidered from the very beginning, independently of which method would have been used isolated for each component problem.

However, sometimes one feels that combining the recognized methods for the component problems is in some sense good, if there is a certain connection between the optimum properties which are required of the component methods on the one hand side and the multiple decision procedure on the other. This idea has been developed by E.L. Lehmann [20] and will be considered below.

In classical statistics there has been a tendency to press any situation into a two-decision problem. The excuse for

doing so has been that it simplifies matter. But that is not always the case. Bartlett's test for the equality of variances in several groups is relatively complicated and of doubtful value. Hartley's maximum F-test for pairwise comparison of variances is both simpler and seems to be based on a more reasonable way of posing the problem. It is rather peculiar that the statisticians have often had qualms of conscience when applying such methods. Thus H.O. Hartley [12] calls his method a "short-cut" test and J.W. Tukey [28] talks about "quick and dirty methods". The practical intuition of the statistician leads him to feel that such methods are to be preferred and he attempts to justify this preference with the amount of computational work, despite the fact that this could obviously not be the motive. We shall below try to find the motive.

B. Unbiasedness in the risk.

We must first define the concept of unbiasedness in the risk. This is a concept which includes unbiasedness in the expectation (for a point estimate) and unbiasedness in the power (of a test) as special cases. (See Lehmann [17].)

We assume that the decision situation is as described in section I. D. Over a sample space $(\mathcal{X},\mathcal{A})$ is defined an a priori family Ω (above denoted by $\mathcal{P}$) of probability measures P. A decision space $(R_d,\mathcal{D})$ is given. The decision function $\delta(D|x)$ has $\mathcal{X}\times\mathcal{D}$ as domain. The operating characteristic of δ is $E_P\delta(D|X) = \beta_\delta(D,P)$.

Suppose now that $L(P,d)$ is the loss inflicted by making a decision d when P is the true distribution of the sample point in the experiment. L could be said to measure the distance between the "true state of nature" P and the decision d. The risk is the expected distance when δ is applied,

$$r(P,\delta) = \int L(P,\cdot)\,d\beta_\delta(\cdot,P) \qquad (1)$$

The distance to a "wrong" distribution P' is obviously

$$r(P',P,\delta) = \int L(P',\cdot)\,d\beta_\delta(\cdot,P) \tag{2}$$

It is reasonable <u>to require of δ that it should be such that expected distance to a wrong distribution is at least equal to expected distance to a true distribution</u>, i.e.

$$r(P',P,\delta) \geq r(P,\delta) \tag{3}$$

for all P' and $P \in \Omega$. In that case δ is called <u>unbiased in the risk</u>.

<u>Example 1</u>. <u>Point estimation</u>. Let $\theta = g(P)$ be a parameter, $R_d = \{g(P) \mid P \in \Omega\}$,

$$L(P,d) = L(P,\hat{\theta}) = (g(P)-\hat{\theta})^2 \tag{4}$$

and let $\delta(D|x) = 1$ if D contains $\hat{\theta}(x)$ and 0 otherwise. The requirement of unbiasedness in the risk of δ then reduces to unbiasedness in the expectation of the estimate $\hat{\theta}(x)$ of θ.

<u>Example 2</u>. <u>Testing hypothesis</u>. Given $\omega \subset \Omega$. Let d_o = "do not reject $P \in \omega$", d_1 = "$P \in \Omega - \omega$" and $R_d = \{d_o, d_1\}$. Furthermore

$$L(P,d_o) = \begin{cases} b & \text{if} \quad P \in \Omega - \omega \\ 0 & \text{if} \quad P \in \omega \end{cases}$$

$$L(P,d_1) = \begin{cases} 0 & \text{if} \quad P \in \Omega - \omega \\ a & \text{if} \quad P \in \omega \end{cases} \tag{5}$$

Then unbiasedness in the risk of δ reduces to unbiasedness in the power with level $\varepsilon = b/(a+b)$.

<u>Example 3</u>. <u>Interval estimation</u>. An interval estimate is unbiased if the probability that it covers the true estimand is at least as great as the probability that it covers a wrong estimand. This can also be shown to be a special case of unbiasedness in the risk.

C. Optimality of combined use of statistical methods.

For each γ (in some indicator space) there is defined a test situation to the effect that $P \in \omega_\gamma$ should be tested against $P \in \Omega - \omega_\gamma = \omega_\gamma^{-1}$. The loss inflicted by erroneous rejections are a_γ and b_γ respectively. We shall jointly make a decision for all γ, i.e. make a choice between sets of the type

$$\Omega_i = \bigcap_\gamma \omega_\gamma^{\mathcal{H}_{i\gamma}} \tag{6}$$

where $\mathcal{H}_{i\gamma}$ is either -1 or 1. For fixed i the corresponding sequence of $\mathcal{H}_{i\gamma}$ may result in Ω_i being empty; i.e. a contradictory decision. We exclude those sequences.

The loss by making the decision $P \in \Omega_k$ when $P \in \Omega_i$ is equal to the sum of the losses for each γ. This can formally be written as

$$L_{ik} = \frac{1}{4} \sum_\gamma \left[(1 + \mathcal{H}_{i\gamma})(1 - \mathcal{H}_{k\gamma}) a_\gamma + (1 - \mathcal{H}_{i\gamma})(1 + \mathcal{H}_{k\gamma}) b_\gamma \right] \tag{7}$$

More generally we could give each term in (7) a weight and write $\int d\mu$ instead of $\sum$.

Suppose now that there exist tests $\delta_\gamma(x)$ for all γ and that we independently use these tests for all γ. This would result in a multiple decision procedure

$$\psi_i(x) = \prod_\gamma \frac{1}{2} \left[(1 + \mathcal{H}_{i\gamma})(1 - \delta_\gamma(x)) + (1 - \mathcal{H}_{i\gamma}) \delta_\gamma(x) \right] \tag{8}$$

which is the probability of stating that $P \in \Omega_i$ when $X = x$. We now assume that

$$\sum_i \psi_i(x) = 1 \tag{9}$$

with probability 1 for all P; where the sum is taken over all i for which Ω_i is non-empty. From this assumption we

find that (8) is true if and only if

$$\delta_\gamma(x) = \frac{1}{2} \sum_i (1 - \alpha_{i\gamma}) \psi_i(x) \tag{10}$$

almost everywhere. Thus it is seen <u>that any arbitrary decision procedure can be considered as constructed from a sequence of procedures</u> δ_γ, <u>i.e. as a multiple decision procedure</u>.

This result and (7) - (10) follow if the set of all γ is countable. However, with the limitation to non-randomized ψ_i and δ_γ, the same result holds when the set of γ is non-countable. The relations (8), (9) and (10) will have to be replaced by the corresponding set relations. The limitation to non-randomized procedures is not serious since we may replace a randomized test with a non-randomized by supplementing the sample with an additional "observation", independent of the sample and with distribution independent of $P \in \Omega$.

We find from (7) for the risk

$$r(P,\psi) = EL_{ik} = \frac{1}{2} \sum_\gamma \left[(1+\alpha_{i\gamma}) \beta_\gamma(P,\delta_\gamma) a_\gamma + (1-\alpha_{i\gamma})(1-\beta_\gamma(P,\delta_\gamma)) b_\gamma \right]$$

$$= \sum r_\gamma(P,\delta_\gamma), \tag{11}$$

where β_γ is the power function and r_γ the risk function for the component test. (Note that k is the random variable in EL_{ik}.)

We now imagine that there is introduced a certain limit concept in Ω and we denote by $\bar{\omega}$ the set of all boundary points for the set ω. We have the following important <u>connection between the optimality of each</u> δ_γ <u>and the corresponding</u> ψ. (Lehmann [20])

<u>Theorem 1</u>. Suppose that

(i) $\bigcup \Omega_i = \Omega$, $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$

and that the decision space consists of decisions d_i = "$P \in \Omega_i$". The decision problem can be decomposed in a finite number of test problems $(\omega_\gamma, \Omega - \omega_\gamma)$ with losses (a_γ, b_γ).

(ii). For all γ, $\delta_{\gamma 0}$ is uniformly most powerful for $P \in \Omega - \omega_\gamma$ and uniformly least powerful for $P \in \omega_\gamma$ among all δ_γ which are similar with level $\varepsilon_\gamma = b_\gamma/(a_\gamma+b_\gamma)$ for $P \in \bar{\omega}_\gamma$.

(iii). $\delta_{\gamma 0}$ is such that the corresponding ψ_{i0} given by (8) satisfies (9).

(iv). For all δ_γ the power $E\delta_\gamma(X)$ is continuous for $P \in \bar{\omega}_\gamma$.

(v). For any γ_0 and $P_0 \in \bar{\omega}_{\gamma 0}$ there exist an Ω_i and an Ω_k such that $P_0 \in \bar{\Omega}_i \cap \bar{\Omega}_k$ and such that $\mathcal{H}_{i\gamma} \neq \mathcal{H}_{k\gamma}$ if and only if $\gamma = \gamma_0$.

Then ψ_{i0} is the uniformly least risky procedure among all procedures which are unbiased in the risk.

The proof is relatively simple. It is obvious from (ii) and (11) that ψ_{i0} is at most as risky as any other procedure the component tests of which are similar on $\bar{\omega}_\gamma$ for all γ. It then remains to prove that unbiasedness in the risk for the multiple procedure implies similarity of the component procedures, and this is a consequence of (v) and (11).

The finiteness of the number of components is needed to prove the last implication. This is, however, not so serious as it might seem, since it is only needed for that purpose and it is possible to prove the same implication in many important situations where the set of γ is countable or a continuum.

We have defined the problem as a choice between Ω_i defined by (6), hence a choice between strong statements. There exists no possibility of concluding the investigation with a more cautious statement if X should assume a value which makes a strong statement reckless.

In order to make it possible to have a choice also between more or less strong statements, we can let the component test situations consist in a choice between rejecting a hypothesis and making no statement at all, i.e. a choice between $P \in \Omega - \omega_\gamma = \omega_\gamma^{-1}$ and $P \in \Omega = \omega_\gamma^{0}$. Thus we have to replace Ω_i by

$$\Omega_i' = \bigcap_\gamma \omega_\gamma^{\frac{1}{2}(\alpha_{i\gamma}-1)} \tag{12}$$

However, in this case Ω_i' may not define the sequence $\alpha_i = \{\alpha_{i\gamma}\}$ uniquely. Suppose e.g. that $\omega_\gamma^{-1} \subset \omega_\varrho^{-1}$. Then the intersections $\omega_\gamma^{-1} \cap \omega_\varrho^{-1}$ and $\omega_\gamma^{-1} \cap \omega_\varrho^{0}$ are identical. Hence if $\alpha_{i\gamma} = -1$, we can choose $\alpha_{i\varrho}$ either 1 or -1. Hence it follows that L_{ik} given by (7) is not defined. In that case we have however $\omega_\varrho \subset \omega_\gamma$. A rejection of $P \in \omega_\gamma$ should therefore lead to rejection of $P \in \omega_\varrho$, i.e. $\alpha_{i\varrho} = -1$. We make it a general rule to eliminate from the list of sequences $\alpha_i = \{\alpha_{i\gamma}\}$ those for which there exists another sequence $\alpha_j = \{\alpha_{j\gamma}\}$ with $\alpha_{j\gamma} \leq \alpha_{i\gamma}$ for all γ. In addition we eliminate all sequences leading to an empty Ω_i'.

A theorem corresponding to theorem 1 can now be proved when Ω_i is replaced by Ω_i', i.e. <u>joint and independent use of uniformly most powerful similar tests leads to a procedure which is least risky among all procedures which are unbiased in the risk</u>. (See Lehmann [20]).

<u>Example 4</u>. $X_1, \ldots, X_n$ are independent normal (ξ, σ). We want to decide whether $\sigma <, =, > 1$. Denote by ω_1 the set of all P for which $\sigma \geq 1$ and by ω_2 the set of all P for which $\sigma \leq 1$. Then

$$\Omega_1 = \omega_1 \cap \omega_2 = \{P|\sigma = 1\}, \quad \Omega_1' = \Omega,$$
$$\Omega_2 = \omega_1 \cap \omega_2^{-1} = \{P|\sigma > 1\} = \Omega_2'$$
$$\Omega_3 = \omega_1^{-1} \cap \omega_2 = \{P|\sigma < 1\} = \Omega_3'$$

We then get for the loss

i \ k		1	2	3
		Ω_1	Ω_2	Ω_3
1	$\sigma = 1$	0	a_2	a_1
2	$\sigma > 1$	b_2	0	a_1+b_2
3	$\sigma < 1$	b_1	a_2+b_1	0

(13)

where Ω_1 may have accent.

The uniformly most powerful tests for ω_1 and ω_2 with levels $b_1/(a_1+b_1)$ and $b_2/(a_2+b_2)$ are well-known. Let c_1 and c_2 be the $b_1/(a_1+b_1)$ and $a_2/(a_2+b_2)$ fractiles for the chi-square distribution with $n-1$ degrees of freedom. We are led to state that $\sigma < 1$ if $Z = \sum(X_i-\bar{X})^2 < c_1$ and $\sigma > 1$ if $Z > c_2$. If $c_1 < Z < c_2$ we say nothing (Ω_1') alternatively that $\sigma = 1$ (Ω_1). a_i and b_i are assumed such that $c_1 < c_2$, otherwise assumption (9) would not be fulfilled. The procedure is the uniformly least risky among all procedures which are unbiased in the risk.

Example 5. Suppose that $X_{i\alpha}; \alpha = 1,2,\ldots,m;$ $i = 1,2,\ldots,q;$ are independent normal, $EX_{i\alpha} = \xi_i$, var $X_{i\alpha} = \sigma^2$. For each pair (i,j) we want to decide whether $\xi_i > \xi_j$, $\xi_i < \xi_j$ or to make no statement, i.e. as far as it can be justified with the limited amount of information in the sample we want to order $\xi_1,\ldots,\xi_q$ in an increasing sequence. Thus we don't want to commit ourself in advance to undertake a complete ordering of the means. We denote by $\bar{X}_i$ the sample mean in group i and write $S^2 = \frac{1}{q(m-1)}\sum(X_{i\alpha}-\bar{X}_i)^2$. Furthermore let $v_{1-\varepsilon}$ denote the $1-\varepsilon$ fractile for the distribution of the Studentized range with q groups and $n-q$ degrees of freedom. It is then recommended to make the three decisions mentioned above according as $\bar{X}_i-\bar{X}_j > \frac{S}{\sqrt{m}} v_{1-\varepsilon}$, $\bar{X}_i-\bar{X}_j < -\frac{S}{\sqrt{m}} v_{1-\varepsilon}$, $|\bar{X}_i-\bar{X}_j| \leq \frac{S}{\sqrt{m}} v_{1-\varepsilon}$. If this is done then the probability of making a wrong statement when $\xi_1 = \xi_2 = \ldots = \xi_q$ is equal to ε. This also controls the probability of making an error in all other situations, since the probability of making a wrong statement when not all ξ_i are equal is at most ε.

It should be noted that to the given level ε for total control of error there corresponds a certain level ε' in the Student's test for deciding whether $\xi_i > \xi_j$, where i and j are selected in advance. According to this test it is stated that $\xi_i > \xi_j$ if $\bar{X}_i-\bar{X}_j > t_{1-\varepsilon}\sqrt{\frac{2}{m}}\,S$, where $t_{1-\varepsilon}$ is the $1-\varepsilon$ fractile of Student's / distribution with $n-q$ degrees of freedom. Hence ε' is given by $t_{1-\varepsilon'} = \frac{1}{\sqrt{2}} v_{1-\varepsilon}$. Hence the method of

Studentized range amounts to a joint use of Student's tests and these tests are for each i and j uniformly most powerful among all similar (and hence also unbiased) tests for the null hypothesis $\xi_i \leq \xi_j$.

Let us now assume that for each (i,j) the loss by stating that $\xi_i > \xi_j$ when indeed $\xi_i \leq \xi_j$, is a_{ij} and that the loss by making no statement when $\xi_i > \xi_j$ is b_{ij}. Assume furthermore that $b_{ij} = \varepsilon'(a_{ij}+b_{ij})$. For any true ordering of $\xi_1,\ldots,\xi_q$ and any of the possible outcomes by the use of the Studentized range method, we can now compute the total loss inflicted by making a wrong statement or refraining from making a statement. The risk is the expected value of this loss. By the optimum property just mentioned and by theorem 1 it now follows that the method of Studentized range is the uniformly least risky among all methods which are unbiased in the risk.

Example 6. By point estimation of a scalar parameter $\theta = g(P)$ we have a choice between sets of the form $\Omega_i = \{P|\theta = i\}$. Also this problem can be decomposed in classical test problems. As a matter of fact we only have to define $\omega_\gamma = \{P|\theta \leq \gamma\}$. It is then seen that

$$\Omega_i = \bigcap_{\gamma \geq i} \omega_\gamma \cap \bigcap_{\gamma < i} \omega_\gamma^{-1} \tag{14}$$

Let A_γ be the region of acceptance of ω_γ. We find that joint use of A_γ for all γ leads to the estimate

$$\hat{\theta}(x) = \inf_{x \in A_\gamma} \gamma \tag{15}$$

The loss by stating that $\theta = k$ when $\theta = i$ is

$$L_{ik} = \int_i^k a_\gamma d\mu, \qquad L_{ik} = \int_k^i b_\gamma d\mu \tag{16}$$

for $k < i$ and $k > i$ respectively. μ is an arbitrary measure.

The implication mentioned after theorem 1 can now be proved in this example and it follows that $\hat{\theta}$ is the least risky of all point estimates which are unbiased in the risk.

If in particular $a_\gamma = b_\gamma = 1$, μ = Lebesque-measure, we get $L_{ik} = |i-k|$. In that case the requirement of unbiasedness in the risk means unbiasedness in the median, i.e. the (population) median of $\hat{\theta}$ is θ. Thus $\hat{\theta}$ uniformly minimizes the expected absolute error among all estimates which are unbiased in median.

If in particular $X_1,\ldots,X_n$ are independent normal (ξ,σ) the point estimate $\hat{\xi}$ for ξ is given by

$$\hat{\xi} = \overline{X} - t_{\hat{\xi}} \frac{S}{\sqrt{n}} \qquad (17)$$

with the same notations as in example 5 and with t_γ equal to the $a_\gamma/(a_\gamma+b_\gamma)$ fractile of Student's distribution. Certain restrictions on a_γ and b_γ are needed which are fulfilled if they are independent of γ. We then see that (17) gives $\hat{\xi}$ explicitely, and if $a_\gamma = b_\gamma = 1$, we get $\hat{\xi} = \overline{X}$.

Example 7. Assume in example 6 above that we want to test $\xi \leq \xi_0$. If $\xi > \xi_0$ we want a point estimate $\hat{\xi}$. Assume also that a_γ and b_γ are independent of γ. By proceeding as in the examples 4 -6 we are led to Student's test, i.e. to reject the hypothesis if $\hat{\xi}$, given by (17), is greater than ξ_0, in which case $\hat{\xi}$ should be used as estimate for ξ. If a_γ and b_γ depends on γ, (9) will not always be satisfied.

References.

[1] R.R. Bahadur: Sufficiency and statistical decision functions. Ann.Math.Stat. Vol. 25(1954).

[2] R.R. Bahadur: Statistics and subfields. Ann.Math.Stat. Vol. 26(1955).

[3] D. Basu: On statistics independent of a complete sufficient statistic. Sankhya. Vol. 15(1955).

[4] D. Blackwell: Conditional expectation and unbiased sequential estimation. Ann.Math.Stat. 1947.

[5] D.R. Cox: Some problems connected with statistical inference. Ann.Math.Stat. Vol. 29(1958).

[6] G. Darmois: Sur les lois de probabilité à estimation exhaustive. Compt.Rendu Acad.Sci. Paris, Vol.260(1935).

[7] R.A. Fisher: On the mathematical foundation of theoretical statistics. Phil.Trans.Roy.Soc. Series A 222(1921).

[8] R.A. Fisher: Theory of statistical estimation. Proc.Camb.Phil.Soc. 22(1925).

[9] R.A. Fisher: Probability, likelihood and quantity of information in the logic of uncertain inference. Proc.Roy.Soc. Series A 146(1934).

[10] D.A.S. Fraser: "Sufficient statistics with nuisance parameters." Ann.Math.Stat. Vol. 27(1956).

[11] Paul R. Halmos and L.J. Savage: Application of the Radon-Nikodym theorem to the theory of sufficient statistics. Ann.Math.Stat. Vol. 10(1949).

[12] H.O. Hartley: Maximum F ratio as a short-cut test for heterogeneity of variances. Biometrika, Vol. 37(1950).

[13] A.N. Kolmogorov: Grundbegriffe der Wahrscheinlichkeitsrechnung. Berlin 1933.

[14] A.N. Kolmogorov: Sur l'estimation statistique des paramètres de la loi de Gauss. Izv.Akad.Nauk SSSR, Ser. Math. 6(1942).

[15] B.O. Koopman: On distributions admitting a sufficient statistic. Trans.Am.Math.Soc. Vol. 39(1936).

[16] L. Le Cam: Sufficiency and approximate sufficiency. Ann.Math.Stat. 35(1964).

[17] E.L. Lehmann: A general concept of unbiasedness. Ann.Math.Stat. Vol. 22(1951).

[18] E.L. Lehmann and H. Scheffé: Completeness, similar regions and unbiased estimation. Part I. Sankhya. 1950.

[19] E.L. Lehmann and Henry Scheffè: Completeness, similar regions and unbiased estimation. Part II. Sankhya. 1955.

[20] E.L. Lehmann: A theory of multiple decision problems. I and II. Ann.Math.Stat. Vol. 28(1957).

[21] E.L. Lehmann: Significance level and power. Ann.Math.Stat. Vol. 29(1958).

[22] E.L. Lehmann: Testing statistical hypothesis. New York 1959.

[23] J. Neyman and E.S. Pearson: On the problem of the most efficient test of statistical hypothesis. Phil.Trans. Series A 231(1933).

[24] J. Neyman and E.S. Pearson: Sufficient statistics and uniformly most powerful tests of statistical hypothesis. Stat.Research Memoirs, Vol. I(1936).

[25] C.R. Rao: Minimum variance estimation in distributions admitting ancillary statistics. Sankhya. Vol. 12(1952).

[16 a] E.L. Lehmann and H. Scheffé: On the problem of similar regions. Mat.Ac.Sciences. 1947.

[26] Else Sandved: A principle for conditioning on an ancillary statistic. To be published in Skand.Akt. (1965).

[27] Erling Sverdrup: Similarity, unbiasedness, minimaxibility and admissibility of statistical test procedures. Skand.Akt. 1953.

[28] J.W. Tukey: Quick and dirty methods in statistics. Proc. Fifth Annual Convention, Am.Soc.for Quality Control (1951).

[29] A. Wald: Contribution to the theory of statistical estimation and testing of hypothesis. Ann.Math.Stat. Vol. 10(1939).

[30] A. Wald: Statistical decision functions. New York 1959.