# A Remark to the Wald's Theory of Statistical Inference

By Kameo MATUSITA

(Received July 18, 1948)

Recently A. Wald has developed an interesting theory on the statistical inference.(1) He remarks that the underlying assumptions of his theory can be weakened in various ways though he did not do so for the sake of simplicity. In this note I shall point out that the assumptions of his theory can, from the mathematical point of view, be so generalized that the main results of his theory remain true and are obtainable without increasing complexity. The theorems and lemmas as in S. D. F. which can be proved just in the same way, will be merely stated.

First, we state our underlying assumptions.

*A. I. Let a sample space* $\mathfrak{R}$ *be a topological space satisfying the following conditions:*

i) *Let* $\mathfrak{S}$ *be the* $\sigma$*-system (i. e., the completely additive system) generated by all the open sets in* $\mathfrak{R}$*. There is defined a completely additive measure* $m(E)$ *of* $\mathfrak{S}$*.*

ii) *There exists a monotone-ascending sequence of bicompact subsets* $\{\mathfrak{R}_n\}$*, which converges to* $\mathfrak{R}$*, each belonging to* $\mathfrak{S}$*.*

*A. II. The set* $\Omega$ *of distribution funcions on* $\mathfrak{R}$*, i. e., measures of* $\mathfrak{S}$ *with the total measure 1, makes a compact regular topological space satisfying the second axiom of countability, and each element* $x(E)$ *of* $\Omega$ *has the expression*

$$x(E) = \int_E p(x, t)\, dm, \qquad E \in \mathfrak{S}$$

*where we assume* $p(x, t)$ *is continuous on the product space* $\Omega \times \mathfrak{R}$*.*(2)

(1) A. Wald, *Statistical Decision Functions which minimize the Maximum Risk*, Ann. Math. 1945. This paper will be referred to as S. D. F.

(2) I shall give an example, where $\Omega$ is not a closed subset in the (finite dimensional) Euclidean space.

Let $\mathfrak{R}$ be the closed interval $[0, 1]$ of real numbers. We consider a class $\Omega_0$ of distribution functions whose density functions belong to the $L_2$-class in $[0, 1]$ and introduce in $\Omega_0$ a topology in the following way. If $\mu_n(E) = \int_E f_n(x)\, dm$ $(n = 1, 2, \cdots)$, where $m$ means the Lebesgue-measure, we define a convergence $\mu_n \to \mu$ $(n \to \infty)$ by the mean convergence $f_n \to f$ $(n \to \infty)$. Then, because of $m(\mathfrak{R}) = 1$,

$$|\mu_n(E) - \mu(E)| = \left|\int_E (f_n(x) - f(x))\, dm\right| \leqq \left|\int_E (f_n(x) - f(x))^2\, dm\right|^{\frac{1}{2}},$$

therefore, for any (Lebesgue-measurable) set $E$ in $\mathfrak{R}$ $\mu_n(E) \to \mu(E)$ when $f_n \to f$. Now, let $\Omega$ be a compact subset of $\Omega_0$ with respect to this topology. Then $\Omega$ satisfies clearly our assumptions.

We denote by $\mathfrak{T}$ the $\sigma$-system generated by all the open sets in $\Omega$.

*A. III. There is defined a weight function $W(x, x')$ on $\Omega \times \Omega$, which satisfies the following conditions:*

i) *$W(x, x')$ is real-valued and*

$$1 \geqq W(x, x') \geqq 0, \qquad W(x, x) = 0.$$

ii) *$\Omega$ can be decomposed into the finite number of mutually disjoint subsets, each belonging to $\mathfrak{T}$, say*

$$\Omega = \Omega_1 + \cdots + \Omega_k, \qquad \Omega_i \in \mathfrak{T},$$

*such that in each $\Omega_i \times \Omega_j$ $(i, j = 1, 2, \cdots, k)$ $W(x, x')$ is continuous and $W(x, x')$ has a continuous extension $W_{i,j}(x, x')$ on $\overline{\Omega}_i \times \overline{\Omega}_j$ with the property that: if $x' \in \overline{\Omega}_j - \Omega_j$, then there exists $x''$ in $\Omega_j$, such that $W_{ij}(x, x') = W_{ij}(x, x'') = W(x, x'')$ holds for all $x$ in $\Omega$, where $i$ is the suffix of $\Omega_i$ containing $x$.*

$W(x, x')$ is, then, clearly measurable with respect to $\mathfrak{T} \times \mathfrak{T}$.

*A. IV. If the integral*

$$\int_\Omega W(x, x')\, p(x, t)\, d\mu$$

*with respect to a distribution function $\mu$ defined on $\mathfrak{T}$ and the integral variable $x$ has the minimum value at $x = x_1$ and $x_2$ for a fixed $t$, then $x_1$ and $x_2$ are interchangeable with respect to $W(x, x')$, i.e., $W(x, x_1) = W(x, x_2)$ for all $x$.*

On these assumptions the following lemmas and theorems are deduced step by step as in S.D.F.

*Lemma 1.* For a fixed $t$ the integral $\int_\Omega W(x, \bar{x})\, p(x, t)\, d\mu$ with respect to $\mu$ and $x$ is a continuous function of a variable $\bar{x}$ on each $\Omega_i$ and has the maximum and minimum values in each $\Omega_i$ $(i = 1, 2, \cdots, k)$.

A mapping $x = \varphi(t)$ from $\mathfrak{R}$ into $\Omega$ is called a statistical decision function or decision function (*s. d. f.* or *d. f.*). By associating a point $t$ in $\mathfrak{R}$ to a point $\bar{x}$ which minimizes $\int_\Omega W(x, \bar{x})\, p(x, t)\, d\mu$, we have a *d. f.* This *d. f.* depends of course on a distribution $\mu$ and will be denoted by $x = \varphi_\mu(t)$.

*Lemma 2.* Let $\{\mu_n\}$ be a sequence of distribution functions on $\mathfrak{T}$, which converges to a distribution function $\mu$, i.e., for any set $E$ with $\mu(\overline{E} - E^0) = 0$(3) it holds $\mu_n(E) \to \mu(E)$ $(n \to \infty)$. Then it holds for any continuous function $f(x)$ on $\Omega$ and any set $E$ in $\Omega$ with $\mu(\overline{E} - E^0) = 0$

3) $\overline{E}$ and $E^0$ represent the closure and the kernel of $E$ respectively.

$$\int_E f(x)\, d\mu_n \to \int_E f(x)\, d\mu, \qquad (n \to \infty).$$

This is a "Helly's Theorem" and is proved as follows.(4)

First, we notice that $\Omega$ is metrizable, for $\Omega$ is a (compact) regular topological space satisfying the second axiom of countability. We denote the distance of every two points $x$, $x'$ in $\Omega$ concerning this metric by $\| x - x' \|$. Now, for any arbitrary positive number $\alpha$ and any point $x_0$ in $\Omega$ set

$$E_\alpha = \{x ; \| x - x_0 \| < \alpha\}.$$

Then, there are, in general, at most countably many positive numbers $\alpha$ such that $\mu(\bar{E}_\alpha - E_\alpha) > 0$. For, if $F(\alpha) = \mu(E_\alpha)$, $F(\alpha)$ is a monotone non-decreasing function of $\alpha$, therefore, $F(\alpha)$ has at most countably many discontinuous points, that is, the number of the points such that $F(\alpha) < F(\alpha + 0)$, or $F(\alpha + 0) - F(\alpha) = \mu(\bar{E}_\alpha) - \mu(E_\alpha) = \mu(\bar{E}_\alpha - E_\alpha) > 0$ is at most countably infinite. Hence, the totality $\Sigma$ of the sets $E_\alpha$ with $\mu(\bar{E}_\alpha - E_\alpha) = 0$ forms a basis of $\Omega$.

Now, as $f(x)$ is continuous, for any positive number $\varepsilon$ and any point $x$ there is a neighborhood of $x$, U, in $\Sigma$ such that the oscillation of $f(x)$ in U is less than $\varepsilon$, that is,

$$\sup_{x', x'' \in U} | f(x') - f(x'') | < \varepsilon.$$

Since $\Omega$ is bicompact because of its compactness and satisfaction of the second axiom of countability, we can make a finite covering of $\Omega$ by such neighborhoods, $\Omega \subset E_1 \cup \cdots \cup E_q$. Putting

$$\begin{aligned} e_1 &= E \frown E_1, \\ e_2 &= E_2 - (e_1), \\ e_3 &= E_3 - (e_1 + e_2), \\ &\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots \\ e_q &= E_q - (e_1 + e_2 + \cdots + e_{q-1}), \end{aligned}$$

we have

$$\mu(\bar{e}_i - \bar{e}_i^0) = 0, \qquad (i = 1, \cdots q)$$

and

$$E = e_1 + e_2 + \cdots + e_q, \qquad e_i \frown e_j = 0 \ (i \neq j).$$

In each $e_i$ the oscillation of $f(x)$ is, of course, less than $\varepsilon$. Now, denoting arbitrary points of $e_1, e_2, \cdots, e_q$ by $x_1, x_2, \cdots, x_q$ respectively, we have

(4) As to this proof the author owes much to O. Takenouchi who debated the matter with him.

$$\left|\int_E f(x)\,d\mu - \sum_{i=1}^{q} f(x_i)\ \mu(e_i)\right|$$
$$\leqq \sum_{i=1}^{q}\left|\int_{e_i} f(x)\,d\mu - f(x_i)\ \mu(e_i)\right|$$
$$\leqq \sum_{i=1}^{q}\int_{e_i} |f(x)-f(x_i)|\,d\mu$$
$$< \varepsilon\mu(E) \leqq \varepsilon,$$

and similarly,

$$\left|\int_E f(x)\,d\mu_n - \sum_{i=1}^{q} f(x_i)\,\mu_n(e_i)\right| < \varepsilon.$$

Hence, it follows

$$\left|\int_E f(x)\,d\mu_n - \int_E f(x)\,d\mu\right|$$
$$< \left|\sum_{i=1}^{q} f(x_i)\,(\mu_n(e_i) - \mu(e_i))\right| + 2\varepsilon$$
$$< C\cdot\sum_{i=1}^{q} |\mu_n(e_i) - \mu(e_i)| + 2\varepsilon,$$

where, $C$ denotes an upper bound of $|f(x)|$ in E. Since $\mu_n(e_i)\to\mu(e_i)$ $(n\to\infty)$ and we can take $\varepsilon$ arbitrarily small, we have

$$\int_E f(x)\,d\mu_n \to \int_E f(x)\,d\mu \qquad (n\to\infty).$$

From this lemma it follows

*Lemma 3.* Let $\{f_n(x)\}$ converge uniformly to $f(x)$, where each $f_n(x)$ is measurable with respect to $\mathfrak{T}$, and $f(x)$ is bounded and continuous on each $\Omega_i$. Furthermore, let $\{\mu_n\}$ be a convergent sequence of distribution functions on $\mathfrak{T}$, whose limit distribution function $\mu$ has the property $\mu(\overline{\Omega}_i - \Omega_i^0) = 0$, $(i = 1, 2, \cdots, k)$. Then it holds

$$\int_\Omega f_n(x)\,d\mu_n - \int_\Omega f_n(x)\,d\mu \to 0 \qquad (n\to\infty).$$

Next we have

*Lemma 4.* Let $\{\mu_n\}$ be a convergent sequence of distribution functions on $\mathfrak{T}$, whose limit distribution $\mu$ has the property $\mu(\overline{\Omega}_i - \Omega_i^0) = 0$ $(i = 1, 2, \cdots, k)$, and let $\{t_n\}$ be a convergent sequence of points in $\mathfrak{R}$, whose limit point is $t_0$. Then it holds

$$W\big(x, \varphi_{\mu_n}(t_n)\big) \to W\big(x, \varphi_\mu(t_0)\big), \qquad (n\to\infty),$$

uniformly in $x$.

From this lemma we see at once that $W(x, \varphi_\mu(t))$ is continuous on $\Omega_i \times \mathfrak{R}$ $(i = 1, \cdots, k)$, therefore measurable with respect to $\mathfrak{T} \times \mathfrak{S}$.

*Theorem I.* The risk function

$$r(x, \varphi_\mu) = \int_{\mathfrak{R}} W(x, \varphi_\mu(t)) p(x, t)\, dm$$

with respect to a distribution $\mu$ in $\Omega$ is continuous on each $\Omega_i$ and its induced function in $\Omega_i$ has a continuous extention on $\overline{\Omega}_i$. Consequently, $r(x, \varphi_\mu)$ is, of course, measurable with respect to $\mathfrak{T}$.

The proof of this theorem follows from the

*Lemma 5.* For any positive $\varepsilon$ there exists a bicompact and closed subset $\mathfrak{R}_\varepsilon$ in $\mathfrak{R}$ such that

$$\int_{\mathfrak{R}_\varepsilon} p(x, t)\, dm \geqq 1 - \varepsilon$$

holds for all points $x$ in $\Omega$.

*Theorem II.* For a distribution function $\mu$ in $\Omega$ there exists a *d. f.* $x = \varphi(t)$ which minimizes the value of the average risk, $\int_\Omega r(x, \varphi)\, d\mu$. If both *d. f.*'s $\varphi^*(t)$ and $\varphi^{**}(t)$ minimize the average risk, then $r(x, \varphi^*) \equiv r(x, \varphi^{**})$ identically in $x$.

In the following we denote by $r_\mu(x)$ the risk function generated by a *d. f.* $\varphi_\mu(t)$ and put $r_\mu = \int_\Omega r_\mu(x)\, d\mu$.

*Theorem III.* If $\mu_n \to \mu (n \to \infty)$ and $\mu(\overline{\Omega}_i - \Omega_i^0) = 0$ $(i = 1, \cdots, k)$, then $r_{\mu_n}(x) \to r_\mu(x)$ $(n \to \infty)$ uniformly in $x$.

This is an immediate consequence of Lemma 4.

We call a least favorable distribution, (*l. f. d*), a distribution function $\lambda$ such that for any distribution $\mu$, $r_\lambda \geqq r_\mu$ hold.

*Theorem IV.* There exists a *l. f. d.*

The proof of this theorem follows from the

*Lemma 6.* The set of all distributions on $\mathfrak{T}$ is compact. This also holds, if we impose upon the distributions the condition $\mu(\overline{\Omega}_i - \Omega_i^0) = 0$ $(i = 1, \cdots, k)$.

*Proof.* Let $\{\mu_n\}$ be a sequence of distributions, and let $\{e_1, e_2, \cdots\}$ be a basis of $\Omega$ (existing by a second axiom of countability). Now we add to this set $\{e_1, e_2, \cdots\}$ the whole set $\Omega$ (in the case when we consider the condition $\mu(\overline{\Omega}_i - \Omega_i^0) = 0$, we add furthermore $\overline{\Omega}_i - \Omega_i^0$ $(i = 1, \cdots, k)$) and we denote the so enlarged set by $\{A_1, A_2, \cdots\}$. Then we select out of a sequence $\{\mu_n(A_1)\}$ a convergent subsequence $\{\mu_{n(1)}(A_1)\}$. This can be done because $\{\mu_n(A_1)\}$ is bounded. Next, we select out of $\{\mu_{n(1)}(A_2)\}$ a convergent subsequence $\{\mu_{n(2)}(A_2)\}$, and so on. Now, put $\mu_{n'}=\mu_{n(n)}$. Then $\{\mu_{n'}\}$ is obviously a subsequence of $\{\mu_n\}$ and is convergent on any $A_i$ $(i=1, 2, \cdots)$. So, let $\mu(A_i)=\lim_{n'\to\infty}\mu_{n'}(A_i)$ $(i=1, 2, \cdots)$. Let $E$ be an arbitrary subset of $\Omega$. We take finite or countable subset $\{e_1, e_2, \cdots\}$, so that $E\subset \bigcup e_i$, and put

$$\mu(E)=\inf\sum_i \mu(e_i)$$

where inf is taken with respect to all possible converings of $E$ by $\{e_i\}$. Defining a measure $\mu$ in this way, we call a set $E$ in $\Omega$ $\mu$-measurable, if for any set $X$ it holds

$$\mu(X)=\mu(X\cdot E)+\mu\big(X\cdot(\Omega-E)\big).$$

Then the class of $\mu$-measurable sets in $\Omega$ makes a $\sigma$-system $\mathfrak{T}^*$. Since $\Omega$ is metrizable, each $e_i$, and consequently, each Borel set generated by $\{e_i\}$ belongs to $\mathfrak{T}^*$. Hence $\mathfrak{T}\subset\mathfrak{T}^*$. Of course, $\Omega$ belongs to $\mathfrak{T}^*$, and $\mu$ is a distribution on $\mathfrak{T}$. In the case where we consider the additional condition, $\mu(\overline{\Omega}_i-\Omega_i^0)=0$ $(i=1, \cdots, k)$ holds. Now, let $F$ be a closed set in $\Omega$. Then $F$ belongs to $\mathfrak{T}$, while it has always a finite covering $\bigcup_{i=1}^{q} e_{\alpha(i)}$, $\alpha(i)$, $q$ being the integers determined by $F$. Since

$$\mu(F)=\inf\sum_{i=1}^{q}\mu(e_{\alpha(i)}),$$

There exists for any $\varepsilon>0$, a finite covering $\bigcup_{i=1}^{q'} e_{\alpha'(i)}$ of $F$ such that

$$\mu(F)+\varepsilon>\sum_{i=1}^{q'}\mu(e_{\alpha'(i)}).$$

On the other hand we have

$$\sum_{i=1}^{q'}\mu_n(e_{\alpha'(i)})\geqq\mu_n(F).$$

Hence

$$\mu(F)+\varepsilon>\sum_{i=1}^{q'}\mu(e_{\alpha'(i)})=\sum_{i=1}^{q'}\lim_{n'\to\infty}\mu_{n'}(e_{\alpha'(i)})$$

$$=\lim_{n'\to\infty}\sum_{i=1}^{q'}\mu_{n'}(e_{\alpha'(i)})\geqq\overline{\lim_{n'\to\infty}}\,\mu_{n'}(F).$$

As $\varepsilon$ can be taken arbitrarily small, we have

$$\mu(F) \geqq \overline{\lim_{n'\to\infty}} \mu_{n'}(F).$$

Next, let $G$ be an open set in $\Omega$. Then $\Omega - G$ is closed, therefore $\mu(\Omega - G) \geqq \overline{\lim_{n'\to\infty}} \mu_{n'}(\Omega - G)$. Hence it follows at once

$$\mu(G) \leqq \underline{\lim_{n'\to\infty}} \mu_{n'}(G).$$

Now, let $E$ be a set in $\mathfrak{T}$ with $\mu(\bar{E} - E^0) = 0$. Then

$$\mu(\bar{E}) = \mu(E) = \mu(E^0), \qquad \mu_{n'}(\bar{E}) \geqq \mu_{n'}(E) \geqq \mu_{n'}(E^0),$$

and we see

$$\mu(\bar{E}) \geqq \overline{\lim} \mu_{n'}(\bar{E}) \geqq \underline{\lim} \mu_{n'}(E^0) \geqq \mu(E^0).$$

Hence there exist $\lim_{n'\to\infty} \mu_{n'}(\bar{E})$ and $\lim_{n'\to\infty} \mu_{n'}(E^0)$, and

$$\lim_{n'\to\infty} \mu_{n'}(\bar{E}) = \lim_{n'\to\infty} \mu_{n'}(E^0) = \mu(E).$$

Therefore

$$\lim_{n'\to\infty} \mu_{n'}(E) = \mu(E),$$

which proves the lemma.

*Theorem V.* If $\lambda$ is a *l. f. d.*, then we have

$$r_\lambda(x) \leqq r_\lambda$$

for all points $x$ in $\Omega$.

For any distribution $\mu$ on $\Omega$, we denote by $\Omega_\mu$ the set $\left\{x\,;\ \text{open } \omega \ni x \to \int_\omega d\mu > 0\right\}$.

*Theorem VI.* If $\lambda$ is a *l. f. d.*, then we have

$$r_\lambda(x) = r_\lambda$$

for all $x$ in $\Omega_\lambda$ possibly except the boundary points of $\Omega_i$ $(i = 1, \cdots, k)$.

*Corollary.* If $\bar{\Omega}_\lambda = \Omega$, it is $r_\lambda(x) = r_\lambda$ for all $x$ in $\Omega$.

*Theorem VII.* If both $\lambda$ and $\mu$ are *l. f. d.*, then $r_\lambda(x) = r_\mu(x)$ for all $x$ in $\Omega$.

*Theorem VIII.* If $\operatorname{Max}_x r_\lambda(x) = r_\lambda$ for a distribution $\lambda$, then $\lambda$ is a *l. f. d.*

*Remark:* This is the inverse of Theorem VI. Therefore we have the logical equivalence:

$$\lambda = l.f.d. \rightleftarrows r_\lambda(x) \leqq r_\lambda.$$

*Theorem IX.* If a *d.f.* $x = \varphi(t)$ minimizes the average risk with respect to a *l. f. d.* $\lambda$, then $\varphi(t)$ minimizes also the maximum risk, i.e., the maximum of the risk function.

*Theorem X.* There exists a *d.f.* which minimizes the maximum risk.

*Theorem XI.* If a *d.f.* $\varphi(t)$ minimizes the maximum risk, and if $\lambda$ is a *l.f.d.*, then $\varphi(t)$ minimizes the average risk with respect to $\lambda$.

*Remark:* This is the inverse of Theorem XI. Therefore,

$$\underset{\varphi}{\text{Min.}}\,\underset{x}{\text{Max}}\, r(x, \varphi) = r(x, \varphi_0) \rightleftarrows \int_{\Omega} r(x, \varphi_0)\, d\lambda = r_\lambda.$$

*Theorem XII.* If both *d. f.* $\varphi^*(t)$ and $\varphi^{**}(t)$ minimize the maximum risk, then $r(x, \varphi^*) = r(x, \varphi^{**})$ for all $x$ in $\Omega$.

A *d.f.* $\varphi(t)$ is called admissible, when there exists no *d.f.* $\psi(t)$ such that $r(x, \psi) \leqq r(x, \varphi)$ for all $x$ in $\Omega$ and that $r(x, \psi) < r(x, \varphi)$ for at least one point $x_0$ in $\Omega$. Further, a *d.f.* is called an optimum *d.f.*, when it is adimissible and minimizes the maximum risk.

*Theorem XIII.* If a *d.f.* $\varphi(t)$ minimizes the maximum risk, then $\varphi(t)$ is admissible.

*Corollary.* There exists an optimum *d.f.*

*Theorem XIV.* If $\lambda$ is a *l.f.d.* and if a *d.f.* $\varphi(t)$ minimizes the maximum risk, then $r(x, \varphi)$ takes a constant value $r_\lambda$ over $\Omega_\lambda$ possibly except at the boundary points of $\Omega_i$ $(i = 1, 2, \cdots, k)$. Moreover, if $\bar{\Omega}_\lambda = \Omega$, then $r(x, \varphi)$ takes a constant value over $\Omega$.

*Institute of Statistical Mathematics, Tokyo.*