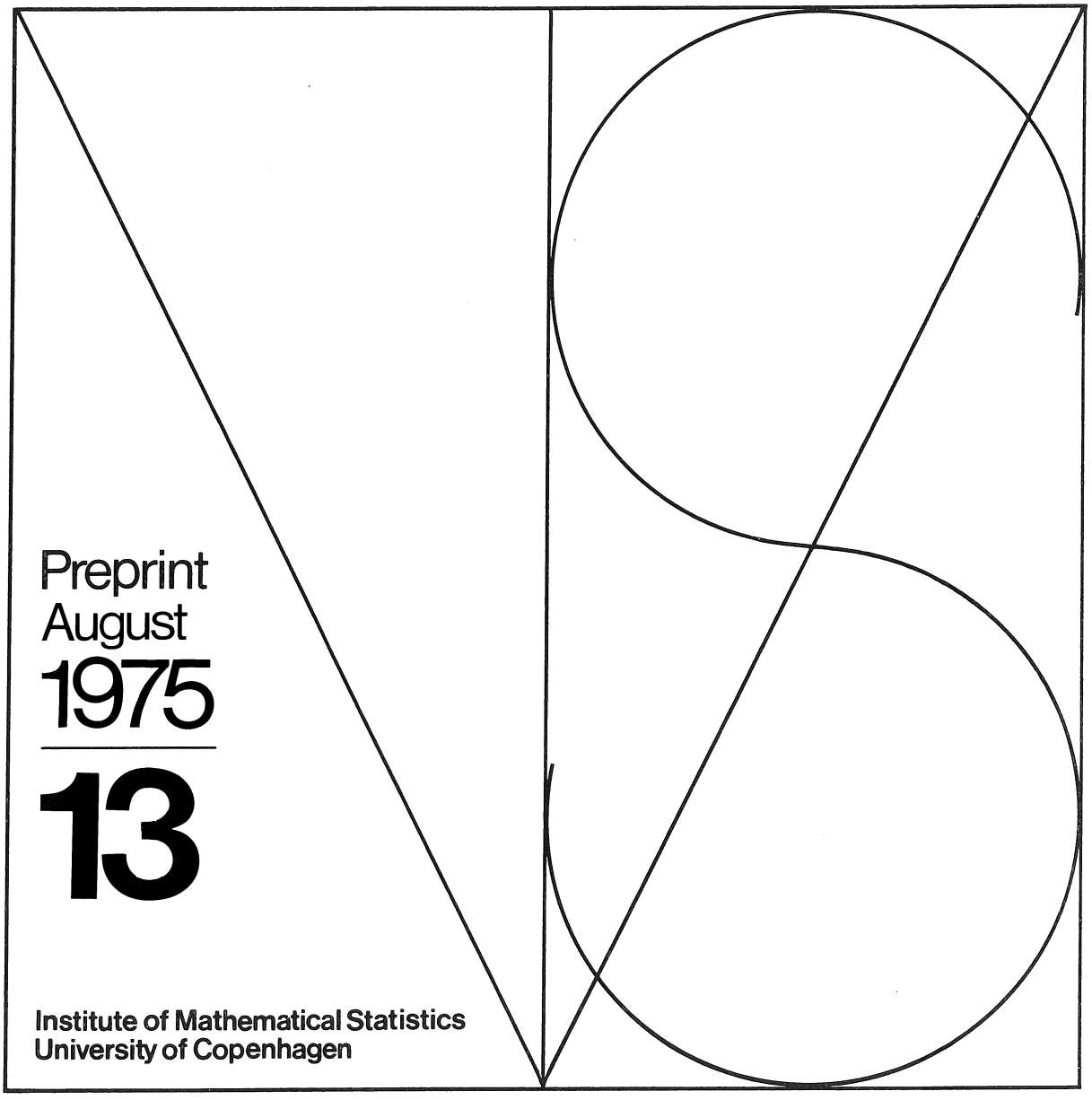
Tue Tjur

# A Constructive Definition of Conditional Distributions



Preprint
August
1975

13

Institute of Mathematical Statistics
University of Copenhagen

Tue Tjur

# A CONSTRUCTIVE DEFINITION OF CONDITIONAL DISTRIBUTIONS

Preprint 1975 No. 13

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

August 1975

# A CONSTRUCTIVE DEFINITION OF CONDITIONAL DISTRIBUTIONS

by

Tue Tjur,

Institute of Mathematical Statistics,

University of Copenhagen.

*SUMMARY*. For Radon probability measures on locally compact and $\sigma$-compact spaces a constructive, pointwise definition of conditional distributions is given, based on a well-known idea of "differentiation". It is proved that such conditional distributions, when almost everywhere defined, have the properties of conditional distributions in the classical sense. Existence of such conditional distributions is discussed. It is concluded that the constructive definition applies to (almost?) all relevant situations as long as the variable conditioned upon is finite dimensional. For wellbehaved distributions on Euclidean spaces it is shown how to express the conditional distribution by its density with respect to "area measure" on the level surface of the transformation conditioned upon. For conditioning in stochastic processes it is proved that the conditional distribution of the process is defined if and only if the corresponding finite dimensional conditional distributions are defined.

*AMS 1970 subject classification.*

Primary 60A05; Secondary 28A15, 28A30, 28A75, 60B05.

*Key words and phrases.* Conditional distribution.

*1. INTRODUCTION.* In the classical formulation of probability theory, the definition of conditional distributions creates some difficulties. Within the framework of abstract measure theory, it is impossible to give a definition of a conditional distribution given a *fixed* value of a stochastic variable. One has to deal with the *family* of conditional distributions, given all possible values of the conditioning variable. Moreover, this family is not uniquely determined. The conditional distributions can be changed on a null set, without affecting the defining properties. This is unsatisfactory in most concrete situations, where an obvious "canonical" family of conditional distributions exists.

It has been noticed by many authors that such a family of conditional distributions can very often be constructed as follows: For any fixed value of the variable we are to condition upon, consider the conditional distribution, given that the variable takes its value in a small neighbourhood. This conditional distribution is welldefined in the elementary sense, as long as the conditioning event is of positive probability. As the neighbourhood becomes smaller it will tend to a limiting distribution, and this is the conditional distribution, given that the variable takes the prescribed value.

It is our aim to show that this intuitive procedure can be given a more explicit formulation as a proper definition of a conditional distribution. This definition seems to be applicable to all finite dimensional problems of any interest, and even for stochastic processes it works fairly well in some cases.

*2. RADON MEASURES.* In order to introduce conditional distributions as indicated, we have to make some topological assumptions. We assume that the spaces X,Y,... considered in the following are *locally compact* and $\sigma$-*compact* (see e.g. Kelley (1955)), and the measures considered will be *Radon measures* (i.e. measures in the sense of Bourbaki (1965a)). These assumptions are not restrictive in

practice. In finite dimensional probabilistic situations they are always satisfied, and for stochastic processes they will be satisfied if the state space is compactified and only the regular version of the process is considered, cf. Nelson (1959) or Tjur (1972). Our main reference concerning measure theory is Bourbaki (1965a), (1965b) and (1959).

Bounded Radon measures can be regarded as abstract measures (i.e. measures in the usual sense) which are *regular* with respect to the class of compact sets (see e.g. Halmos (1950)). But a Radon measure can also be regarded as a positive linear functional on the space $\mathcal{K}(X)$ of continuous realvalued functions with compact support, and this will be our basic point of view. Hence, by a *measure* (we omit the word "Radon") we mean a positive linear mapping $\mu: \mathcal{K}(X) \to \mathbb{R}$. We write $\mu(f)$ or $\int f(x)\, \mu(dx)$ for the value of $\mu$ at $f \in \mathcal{K}(X)$. The short notation $\mu(f)$ is also applied for $\mu$-integrable functions $f$. We write $\mu(B)$ instead of $\mu(1_B)$, regarding $\mu$ as a set function, whenever convenient ($1_B$ denotes the indicator function of $B \subseteq X$).

By $\mathcal{M}(X)$ and $\mathcal{P}(X)$ we denote the set of (Radon) measures and (Radon) probability measures, respectively. These sets of measures are regarded as topological spaces in the topology induced by the mappings $\mu \to \mu(f)$, $f \in \mathcal{K}(X)$. This topology is called the weak topology. On $\mathcal{P}(X)$, the weak topology coincides with the topology induced by the richer family of mappings $\mu \to \mu(f)$ where $f$ is bounded and continuous.

Let $\mu$ denote a probability measure on $X$. A mapping $t: X \to Y$ is called $\mu$-*measurable* (or *Lusin measurable* with respect to $\mu$) if for any $\varepsilon > 0$ there exists a compact set $K \subseteq X$ with $\mu(X \setminus K) \leq \varepsilon$ such that $t$ restricted to $K$ is continuous. This concept of measurability turns out to be more convenient than that of Borel

measurability, when Radon measures are considered. As to the connection between the two notions of measurability, by a generalization of a theorem due to Lusin (1912) any Borel measurable mapping of X into a Euclidean space (or into a locally compact space Y admitting a denumerable base) is Lusin measurable with respect to any Radon measure on X.

A set $A \subseteq X$ is called $\mu$-measurable if $1_A: X \to \mathbb{R}$ is $\mu$-measurable. For probability measures (and for bounded measures) a set A is $\mu$-measurable if and only if it is $\mu$-integrable.

*3. THE DEFINITION.* For $\mu \in \mathcal{P}(X)$, let $t: X \to Y$ be $\mu$-measurable. Then $g \circ t$ is $\mu$-integrable for $g \in \mathcal{K}(Y)$, and so the transformed measure $\nu = t(\mu)$, given by $\nu(g) = \mu(g \circ t)$, is welldefined. It can be proved (see Bourbaki (1965b)) that a function g on Y (or a subset $B \subseteq Y$) is $\nu$-integrable if and only if $g \circ t$ (or $t^{-1}(B)$) is $\mu$-integrable, and in this case, $\nu(g) = \mu(g \circ t)$ (or $\nu(B) = \mu(t^{-1}(B))$).

Let x denote the random element of the probability field $(X,\mu)$ and consider the stochastic variable t(x). For a fixed point $y_0$ in the support of $\nu$ we consider the problem of defining the *conditional distribution of* x, *given* $t(x) = y_0$.

For a $\nu$-measurable set B with $\nu(B) > 0$, denote by $\mu^B$ the conditional distribution of x, given that $t(x) \in B$, i.e.

$$\mu^B(f) = \frac{1}{\nu(B)} \int_{t^{-1}(B)} f(x)\, \mu(dx)$$

for $f \in \mathcal{K}(X)$. We shall introduce a suitable notion of "letting B tend to $y_0$", and then define the conditional distribution of x, given $t(x) = y_0$, as the limit (in the weak topology) of $\mu^B$ when B tends to $y_0$.

In order to specify this limiting procedure we shall need the concept of a

*generalized sequence* (or a *net*). This is simply a sequence where the set of positive integers is replaced by an arbitrary partially ordered and upwards directed set. As to the present situation the directed set is defined as follows:

Let $D_{y_0}$ denote the set of pairs (V,B), where V is an *open neighbourhood* of $y_0$ and B is a $\nu$-*integrable subset of* V *with* $\nu(B) > 0$. We say that (V,B) *is closer to* $y_0$ *than* (V',B') if $V' \supseteq V$. This relation is a partial ordering of $D_{y_0}$, making $D_{y_0}$ a *directed set*, i.e. for any two elements of $D_{y_0}$ there exists a third element closer to $y_0$ than both of them. The proofs of these facts are straight forward. Thus the probability measures $\mu^B$, $(V,B) \in D_{y_0}$, constitute a generalized sequence.

3.1. *Definition: Suppose that the generalized sequence* $(\mu^B)$ *converges to a probability measure on* X. *This probability measure is then denoted* $\mu^{y_0}$ *and called the conditional distribution of* $x \in (X,\mu)$, *given* $t(x) = y_0$.

The intuitive idea behind the definition may have escaped the readers attention during the rather technical constructions. However, the introduction of the generalized sequence is just a way of formalizing the very simple idea of drawing a set B closer and closer to $y_0$. As it makes no sense to talk about a set B being "closer to" $y_0$ than another, the *neighbourhood* V is introduced as an auxiliary variable, to take care of the ordering.

It may happen that $\mu^B$ converges to a measure of mass < 1, for example to the null-measure. Notice that such a "defective conditional distribution" is not accepted by the definition.

For the sake of completeness, it should be mentioned that the concept of a generalized sequence can be avoided in the definition. The conditional distribution can be characterized as the probability measure $\mu^{y_0}$ - unique, if it exists - with the property that for any $\varepsilon > 0$ and any $f \in \mathcal{K}(X)$ there exists an open neighbourhood V of $y_0$ such that for any $B \subseteq V$ with $\nu(B) > 0$ we have

$|\mu^{y_0}(f) - \mu^B(f)| \leqq \varepsilon$. This characterization follows immediately from the definition.

*4. A CONTINUITY PROPERTY.* Consider the situation $t: X \to Y$, $\nu = t(\mu)$ of the preceding section, and let $Y_0$ denote the set of points $y \in Y$ such that the conditional distribution $\mu^y$ is defined.

4.1. *Theorem: The mapping from* $Y_0$ *to* $\mathcal{P}(X)$, *taking* $y$ *into* $\mu^y$, *is continuous.*

*Proof:* For a fixed point $y_0 \in Y_0$ let $W \subseteqq \mathcal{P}(X)$ denote a *closed* neighbourhood of $\mu^{y_0}$ (it suffices to consider closed neighbourhoods, since $\mathcal{P}(X)$ is a completely regular space). Let $V$ denote an open neighbourhood of $y_0$ such that $\mu^B \in W$ for $B \subseteqq V$, $\nu(B) > 0$. For a point $y \in V \cap Y_0$, consider the distribution $\mu^B$ as $B$ approaches $y$ in the sense of section 3. From a certain stage of this limiting procedure we have $B \subseteqq V$, and from then on, $\mu^B \in W$. Since $W$ is closed, the limit $\mu^y = \lim \mu^B$ is also an element of $W$, and this proves the theorem.

4.2. *Corollary: Suppose that the conditional distribution* $\mu^y$ *is defined for* $\nu$ *- almost all* $y$. *Then, the mapping* $y \to \mu^y$ *(or any extension to* $Y$ *of this mapping) is* $\nu$*-measurable.*

This is an immediate consequence of the regularity of $\nu$ and the definition of Lusin measurability.

*5. THE MAIN THEOREM.* We shall prove that conditional distributions in our sense, when almost everywhere defined, are also conditional distributions in the classical sense (Doob (1953)):

5.1. *Theorem: Let* $t: X \to Y$, $\nu = t(\mu)$ *be as in section 3, and suppose that the conditional distribution* $\mu^y$ *is defined for* $\nu$ *- almost all* $y$. *Let* $f: X \to \mathbb{R}$ *be* $\mu$*-integrable. Then* $f$ *is* $\mu^y$*-integrable for* $\nu$ *- almost all* $y$, *and the (almost*

*everywhere defined) function* $y \to \mu^y(f)$ *is* $\nu$*-integrable. For any bounded* $\nu$*-measurable function* $g: Y \to \mathbb{R}$ *we have*

$$\int \mu^y(f)\, g(y)\, \nu(dy) \tag{5.2}$$

$$= \int f(x)\, g(t(x))\, \mu(dx).$$

*In particular, for any* $\mu$*-measurable set* $A \subseteq X$ *and any* $\nu$*-measurable set* $B \subseteq Y$ *we have (putting* $f = 1_A$ *and* $g = 1_B$ *in* (5.2))

$$\int_B \mu^y(A)\, \nu(dy) \tag{5.3}$$

$$= \mu(t^{-1}(B) \cap A).$$

*Remark:* (5.3) is valid for Borel sets A and B (since Borel sets are Lusin measurable with respect to any measures), and this constitutes Doob's definition of a family $(\mu^y)$ of conditional distributions.

*Proof:* To begin with, let f be a $\mathcal{K}(X)$-function and let $f_0: Y \to \mathbb{R}$ denote (a version of) the conditional expectation of f given t in the sense of Kolmogorov (1933). That is, $f_0$ is the unique (up to equivalence) $\nu$-integrable function satisfying the equation

$$\int f_0(y)\, g(y)\, \nu(dy) = \int f(x)\, g(t(x))\, \mu(dx) \tag{5.4}$$

for any bounded $\nu$-measurable function g. Let $y_0$ denote a point such that the conditional distribution $\mu^{y_0}$ is defined. Let $B \subseteq Y$ be $\nu$-integrable with $\nu(B) > 0$. For B tending to $y_0$ in the sense of section 3 we have (inserting $g = \nu(B)^{-1} 1_B$ in (5.4))

(5.5)
$$\frac{1}{\nu(B)} \int_B f_0(y)\ \nu(dy)$$

$$= \frac{1}{\nu(B)} \int_{t^{-1}(B)} f(x)\ \mu(dx)$$

$$= \mu^B(f) \to \mu^{y_0}(f).$$

According to theorem 4.1, the mapping $y \to \mu^y(f)$ is continuous on its domain $Y_0$. Hence, for any $\varepsilon > 0$ we can choose an open neighbourhood V of $y_0$ such that $|\mu^y(f) - \mu^{y_0}(f)| \leqq \varepsilon$ for $y \in V \cap Y_0$. For $B \subseteq V$ we have then (noticing that the function $y \to \mu^y(f)$ is bounded and measurable, thus integrable)

$$\left|\frac{1}{\nu(B)} \int_B \mu^y(f)\ \nu(dy) - \mu^{y_0}(f)\right|$$

$$= \left|\frac{1}{\nu(B\cap Y_0)} \int_{B\cap Y_0} \mu^y(f)\ \nu(dy) - \mu^{y_0}(f)\right|$$

$$= \left|\frac{1}{\nu(B\cap Y_0)} \int_{B\cap Y_0} \left(\mu^y(f) - \mu^{y_0}(f)\right)\ \nu(dy)\right|$$

$$\leqq \frac{1}{\nu(B\cap Y_0)} \int_{B\cap Y_0} \left|\mu^y(f) - \mu^{y_0}(f)\right|\ \nu(dy)$$

$$\leqq \frac{1}{\nu(B\cap Y_0)} \int_{B\cap Y_0} \varepsilon\ \nu(dy) = \varepsilon .$$

From this we conclude that

(5.6)
$$\frac{1}{\nu(B)} \int_B \mu^y(f)\ \nu(dy) \to \mu^{y_0}(f)$$

for B tending to $y_0$. Now put $h(y) = f_0(y) - \mu^y(f)$. Subtracting (5.6) from (5.5) we get

$$\frac{1}{\nu(B)} \int_B h(y)\ \nu(dy) \to 0$$

for B tending to $y_0$. This holds for $\nu$ - almost all $y_0$. Intuitively, it is obvious that a function h with this property must be a $\nu$ - null - function. The proof of this is postponed until later (lemma 5.8). We conclude that $\mu^y(f) = f_0(y)$ for almost all y. Hence, we can replace $f_0(y)$ by $\mu^y(f)$ in equation (5.4). Doing so, we get equation (5.2) of the theorem (but, as yet, only for $f \in \mathcal{K}(X)$). For $g = 1$ this equation takes the form

$$\int \mu^y(f)\ \nu(dy) = \mu(f). \tag{5.7}$$

Since the mapping $y \to \mu^y$ is $\nu$-measurable, the *mixture* (or the *integral*, see Bourbaki (1965b)) of the measures $\mu^y$ with respect to $\nu$ is welldefined, and (5.7) shows that this measure is equal to $\mu$. It follows from basic results on integration with respect to a mixture (Bourbaki (1965b)) that (5.7) is valid for any $\mu$-integrable function f, in the sense that $\mu^y(f)$ is welldefined for $\nu$ - almost all y, and the integral on the left is welldefined. Correspondingly, equation (5.2) (which has already been proved for $f \in \mathcal{K}(X)$) can be extended to the case where f is $\mu$-integrable: For fixed g (bounded and $\nu$-integrable) both sides of (5.2) are welldefined for any $\mu$-integrable function f, and both sides depend continuously upon f as f varies in the normed function space $L^1(\mu)$ of $\mu$-integrable functions (equipped with the topology induced by the 1-norm). Since (5.2) is valid for f in the dense subspace $\mathcal{K}(X)$ of $L^1(\mu)$, it must be valid for all $f \in L^1(\mu)$.

It remains to prove

5.8. *Lemma: A $\nu$-integrable function* $h: Y \to \mathbb{R}$ *is a $\nu$ - null - function if (and only if) it has for $\nu$ - almost all* $y_0$ *the property that*

$$\frac{1}{\nu(B)} \int_B h(y)\ \nu(dy) \to 0$$

*as* B *tends to* $y_0$ *in the sense of section 3.*

*Proof:* For $\varepsilon > 0$ let H denote a compact set such that $\nu(H) > 1 - \varepsilon$, and such that the condition of the lemma is satisfied for *all* $y_0 \in H$. Let n be a positive integer. For $y_0 \in H$ we can choose an open neighbourhood V of $y_0$ such that

$$\left|\frac{1}{\nu(B)} \int_B h(y)\ \nu(dy)\right| \leqq \frac{1}{n} \tag{5.9}$$

for $B \subseteqq V$, $\nu(B) > 0$. Then, by an indirect argument, we conclude that $|h| \leqq 1/n$ almost sure on V : If this was not the case, one of the two sets $B_+ = \{y \in V \mid h(y) > 1/n\}$ and $B_- = \{y \in V \mid h(y) < -1/n\}$ would lead us to a contradiction, when inserted for B in (5.9). Thus any point $y_0 \in H$ has an open neighbourhood V such that $|h| \leqq 1/n$ almost sure on V. Covering H with a finite number of such neighbourhoods, we conclude that $|h| \leqq 1/n$ almost sure on H. Letting $n \to \infty$ we conclude that $h = 0$ almost sure on H, and obviously (for $\varepsilon \to 0$) this proves that h is a $\nu$ - null - function.

## *6. DECOMPOSITIONS.*

6.1. *Definition: Let* $t: X \to Y$ *be a continuous transformation and let* $\lambda$ *denote a measure on* X. *A family* $(\lambda_y)_{y \in Y}$ *of measures on* X *together with a measure* $\lambda'$ *on* Y *is called a decomposition of* $\lambda$ *with respect to* t, *if the following conditions are satisfied:*

(6.2) *The mapping from* Y *to* $\mathcal{M}(X)$, *taking* y *into* $\lambda_y$, *is continuous.*

(6.3) *The support of* $\lambda_y$ *is contained in* $t^{-1}(y)$.

(6.4) $\int \lambda_y(f)\ \lambda'(dy) = \lambda(f)$ *for* $f \in \mathcal{K}(X)$.

*Remark:* It follows from (6.2) and (6.3) that the left of (6.4) is welldefined, the function $y \to \lambda_y(f)$ being a $\mathcal{K}(Y)$ - function.

In short, a decomposition of a measure with respect to a transformation is a representation of the measure as a mixture of measures on the level surfaces of the transformation. Notice that our regularity conditions are very restrictive (t should be continuous etc.). More general definitions of decompositions (or *disintegrations*), based on the same intuitive idea, can be found in Halmos (1941) and Bourbaki (1959).

6.5. *Example:* Let X and Y be open subsets of Euclidean spaces $\mathbb{R}^n$ and $\mathbb{R}^k$, $n \geq k$, and let $\lambda$ denote Lebesgue measure on X. Suppose that $t: X \to Y$ is *surjective* and *continuously differentiable with differential of maximal rank* (i.e. the rows of the $k \times n$ - matrix $Dt(x)$ are linearly independent for any $x \in X$). Then there exists a unique decomposition of $\lambda$ with respect to t such that the measure $\lambda'$ of the definition is Lebesgue measure on Y. The measures $\lambda_y$ on the level surfaces are given by

$$\lambda_y(f) = \int_{t^{-1}(y)} \frac{f(x)}{\sqrt{\det(Dt(x)\ Dt(x)^*)}}\ dx\ , \tag{6.6}$$

where $\int_{t^{-1}(y)} \ldots\ dx$ denotes (here and in the following) integration with respect to *geometric measure* on the $n - k$ - dimensional manifold $t^{-1}(y)$. By geometric measure we mean the multidimensional analogue of *area measure* on a twodimensional manifold in $\mathbb{R}^3$. See for example Hicks (1965) for the definition. Integrals with respect to geometric measures are called *surface integrals*.

Notice that the determinant in formula (6.6) is positive (the asterisk denotes transposition) and depends continuously upon x. Thus the right hand side of (6.6) is welldefined.

We shall not show in detail how to establish this decomposition of Lebesgue measure on X. A similar result under less restrictive conditions can be found in Federer (1969). The formula

$$(6.7) \qquad \int_X f(x)\,dx = \int_Y \left( \int_{t^{-1}(y)} \frac{f(x)}{\sqrt{\det(Dt(x)\ Dt(x)^*)}}\,dx \right) dy$$

(which is just (6.4), written out for this special case) can be proved by more or less heuristic differential geometric arguments concerning the infinitesimal elements of the integrals involved. However, the proper way of establishing the decomposition is by a *local reparametrization* of the problem. A small neighbourhood U of $x_0 \in X$ can be transformed by a one to one differentiable transformation $s: U \to \mathbb{R}^n$ such that the first k coordinate functions of s coincide with those of t. Thus the level surfaces of t are mapped into parallel pieces of n - k - dimensional affine subspaces, and, via the formula for change of variables in a multiple integral, the whole problem is reduced to the case where t is linear. For a full proof along these lines, see Tjur (1974).

*7. DECOMPOSITIONS AND CONDITIONAL DISTRIBUTIONS.* Obviously the concept of a decomposition is closely related to that of conditioning. The definitions of decompositions in Halmos (1941) and Bourbaki (1959) are sufficiently general to allow for *existence* of decompositions of a given measure with respect to a given transformation under certain regularity conditions. As a special case existence of conditional distributions in the sense of Doob (1953) comes out. Conversely, a family of conditional distributions in Doob's sense has (still under some regularity assumptions) the properties of a decomposition in the wide sense (see e.g. Blackwell (1956)).

In case t is continuous there is a similar connection between our (more exclusive) concepts of decompositions and conditional distributions:

7.1. *Theorem:* *Let* $t: X \to Y$ *and* $\nu = t(\mu)$ *be as in section 3, and assume that* $t$ *is continuous and* $Y$ *is the support of* $\nu$. *For a family* $(\xi_y)_{y \in Y}$ *of probability measures on* $X$, *the following two conditions are equivalent:*

(7.2) *The measures* $\xi_y$ $(y \in Y)$ *and* $\nu$ *constitute a decomposition of* $\mu$ *with respect to* $t$.

(7.3) *The conditional distribution* $\mu^y$ *is defined for all* $y \in Y$, *and* $\mu^y = \xi_y$.

*Proof:* Assume (7.2). For $\nu(B) > 0$, $f \in \mathcal{K}(X)$, we have (by the rule of integration with respect to a mixture, cf. Bourbaki (1965b))

$$\mu^B(f) = \frac{1}{\nu(B)} \int_{t^{-1}(B)} f(x)\, \mu(dx)$$

$$= \frac{1}{\nu(B)} \int_B \xi_y(f)\, \nu(dy).$$

The function $y \to \xi_y(f)$ being continuous, a standard argument (similar to that preceding formula (5.6)) shows that this expression converges to $\xi_{y_0}(f)$ as $B$ tends to $y_0$ in the sense of section 3.

Conversely, suppose that (7.3) is satisfied. By theorem 4.1 the mapping $y \to \xi_y$ is then continuous, and theorem 5.1 shows (for $g = 1$ in (5.2)) that $\mu$ is the mixture of the measures $\xi_y$ with respect to $\nu$. It remains to prove that the support of $\xi_{y_0} = \mu^{y_0}$ is contained in $t^{-1}(y_0)$.

For $g \in \mathcal{K}(Y)$, $\nu(B) > 0$, we have (since $\nu = t(\mu)$)

$$\frac{1}{\nu(B)} \int_{t^{-1}(B)} g(t(x))\, \mu(dx) = \frac{1}{\nu(B)} \int_B g(y)\, \nu(dy).$$

For B tending to $y_0$ this equation yields

$$\mu^{y_0}(g \circ t) = g(y_0).$$

Hence, the transformed measure $t(\mu^{y_0})$ (defined by $t(\mu^{y_0})(g) = \mu^{y_0}(g \circ t)$) equals the one point measure at $y_0$. From this it follows immediately that $\mu^{y_0}$ is concentrated on $t^{-1}(y_0)$.

*8. CONDITIONAL DISTRIBUTIONS ON EUCLIDEAN SPACES.*

8.1. *Theorem: Let* X *and* Y *be open subsets of* $\mathbb{R}^n$ *and* $\mathbb{R}^k$, $n \geq k$. *Let* $\mu$ *denote a probability measure on* X *given by a positive and continuous density* p *with respect to Lebesgue measure. Let* $t: X \to Y$ *be as in example 6.5. Then the transformed measure* $\nu = t(\mu)$ *has a density* q *with respect to Lebesgue measure on* Y, *given by the surface integral*

$$q(y) = \int_{t^{-1}(y)} \frac{p(x)}{\sqrt{\det(Dt(x)\ Dt(x)^*)}}\, dx. \tag{8.2}$$

*Assume, in addition, that this function* q *is positive and continuous. Then the conditional distribution* $\mu^y$ *is defined for all* $y \in Y$ *and given by the density*

$$p^y(x) = \frac{p(x)}{q(y)\ \sqrt{\det(Dt(x)\ Dt(x)^*)}} \tag{8.3}$$

*with respect to the geometric measure on the* n - k - *dimensional surface* $t^{-1}(y)$.

*Remark:* In case n = k, the "level surfaces" $t^{-1}(y)$ are discrete subsets of X, and "geometric measure" should be interpreted as *counting measure*. The decomposition of example 6.5 is valid in this case also.

*Proof:* For shortness, we introduce the function $F_t(x) = 1/\sqrt{\det(Dt(x)\ Dt(x)^*)}$. For $g \in \mathcal{K}(Y)$ we have, by formula (6.7),

$$\nu(g) = \mu(g \circ t) = \int_X g(t(x))\ p(x)dx$$

$$= \int_Y \left( \int_{t^{-1}(y)} g(t(x))\ p(x)\ F_t(x)dx \right) dy$$

$$= \int_Y g(y) \left( \int_{t^{-1}(y)} p(x)\ F_t(x)dx \right) dy \quad ,$$

i.e., $\nu$ has the density

$$q(y) = \int_{t^{-1}(y)} p(x)\ F_t(x)dx$$

as stated in the theorem. Now assume that the function q, given by this formula, is positive and continuous (actually it *is* positive, but not necessarily continuous). Denote by $\xi_y$ the measure on X given by the density

$$p^y(x) = \frac{p(x)\ F_t(x)}{q(y)}$$

with respect to the geometric measure on $t^{-1}(y)$. Then, by straight forward arguments, the $\xi_y$ 's together with $\nu$ constitute a decompositon of $\mu$ with respect to t. An application of theorem 7.1 completes the proof.

A "pointwise" version of theorem 8.1 can be proved under weaker regularity conditions than assumed here. Rather than going into details with this we shall indicate how one can reformulate almost any finite dimensional problem in such a way that the strong regularity assumptions of the theorem are satisfied:

First of all, the dimensions n and k should be the "intrinsic" dimensions of the problem. Therefore, if $\mu$ or $\nu$ (or both) is concentrated on a submanifold of a Euclidean space, this manifold must be parametrized and the problem of conditioning must be stated in terms of the distribution of the parameter. We have disregarded the case where either $\mu$ or $\nu$ is of "mixed dimension", a rarely occuring case which can be handled by a "piecewise" technique.

After this "reparametrization" the measures $\mu$ and $\nu$ will be measures on proper domains (i.e. nice sets with non-empty interiors) in Euclidean spaces, and in practice such measures will always be given by densities with respect to Lebesgue measure. We claim that the regularity conditions of theorem 8.1 are now *essentially* satisfied. By this we mean that they will be satisfied if certain null-sets are removed from X and Y. For example, if p is not continuous, removal of some closed manifolds of lower dimensions, containing the discontinuity points of p, will solve the problem. The points x with $p(x) = 0$ can obviously be excluded. In a similar manner the "singularity manifolds" of q and t are removed. In order to make X and Y open, the boundaries (which are manifolds of lower dimensions, thus null-sets) are removed. To make t surjective, Y is replaced by $t(X)$.

Obviously, these modifications are not possible in arbitrary situations. Loosely speaking, we have assumed that p, q and t are "piecewise nice", where "niceness" refers to the regularity conditions of theorem 8.1. However, in practice this means that t, p and q should be given in an explicit, nonpathological manner; and this is necessary anyway, if we are to "solve" the conditioning problem in some sense. Examples can be given where the conditional distribution is nowhere defined, but such examples posess the unmistakeable features of "proffesional counterexamples", involving dense denumerable subsets etc..

The modification described above may be inconvenient for other reasons. However, from the mere *existence* of such a modification we conclude that the conditional

distribution $\mu^y$ is *almost* everywhere defined, and this suffices for most purposes (cf. section 5).

It should be clear from the above considerations that our definition of a conditional distribution does not suffer from existence problems in the finite dimensional case. Moreover, our definition can obviously be applied to the discrete case (i.e. when X and Y are denumerable sets). Hence, in order to cover the classical fields of probability theory, it remains to apply the definition to stochastic processes.

*9. CONDITIONAL DISTRIBUTIONS OF A STOCHASTIC PROCESS.* We shall take our starting point in the simplest possible definition of a stochastic process as a measure on a topological space, namely that of Nelson (1959), see also Bourbaki (1969) or Tjur (1972):

Let X be a *compact* space (the state space) and let I denote an arbitrary set (the time scale). A stochastic process is determined by a (Radon) probability measure $\mu$ on the compact space $X^I$ or, equivalently, by a consistent family $(\mu_{I_0})$ of probability measures on the "finite dimensional" spaces $X^{I_0}$, where $I_0$ denotes a finite subset of I.

Let $t: X^I \to Y$ (Y locally compact and $\sigma$-compact) be $\mu$-measurable and denote by $\nu \in \mathcal{P}(Y)$ the transformed measure $t(\mu)$.

9.1. *Theorem: For a point* $y_0 \in Y$ *the following two conditions are equivalent:*

(9.2) *The conditional distribution* $\mu^{y_0}$ *of the sample function* $(x_i) \in (X^I,\mu)$, *given* $t((x_i)) = y_0$, *is defined.*

(9.3) *For any finite subset* $I_0 = \{i_1,\ldots,i_n\}$ *of* I *the conditional distribution of* $(x_{i_1},\ldots,x_{i_n})$, *given* $t((x_i)) = y_0$, *is defined.*

*In case of existence, the conditional distributions in* (9.3) *constitute the consistent family of the conditional distribution in* (9.2).

*Remark:* The conditional distributions in (9.3) should be interpreted in the following (not immediate) manner: For $I_0 = \{i_1,\ldots,i_n\}$ consider the distribution of $(x_{i_1},\ldots,x_{i_n}, t((x_i)))$. This is a probability measure $\pi$ on $X^{I_0} \times Y$. In *this* distribution, condition on the last coordinate t(x). If defined, this conditional distribution $\pi^{y_0}$ can be transformed by the projection $p: X^{I_0} \times Y \to X^{I_0}$. The transformed measure $p(\pi^{y_0})$ is what we call the conditional distribution of $(x_{i_1},\ldots,x_{i_n})$, given $t((x_i)) = y_0$.

*Proof:* It is easy to see that condition (9.3) (as specified by the above remark) is equivalent to the following: The conditional distribution of $(x_{i_1},\ldots,x_{i_n})$, given $t(x) \in B$, converges for B tending to $y_0$ in the sense of section 3. Since weak convergence of distributions on $X^I$ is equivalent to weak convergence of the finite dimensional marginal distributions (this is easy to prove by a standard compactness argument), the theorem follows.

The theorem shows that conditioning on a finite dimensional variable t(x), derived from a stochastic process, can be traced back to finite dimensional conditioning problems.

A more intricate problem arises when the variable t(x) is infinite dimensional. This problem of conditioning *on* a stochastic process is too complicated to be dealt with here. See Tjur (1974) for some (perhaps rather useless) results on reduction to finite dimensional problems. It is known that the conditional distribution is not always defined, even in nice problems of this type. It may be possible to overcome this difficulty by introduction of finer topologies than the product topology on the space of sample functions. This means that the sample function properties must be considered, and a more general measure theory may be convenient in order to avoid compactifications (cf. Bourbaki (1969)).

Just to indicate that our definition is applicable to some problems of this type it can be mentioned that the strong Markov property for Feller processes on a compact state space (see Tjur (1972)) can be stated in terms of conditional distributions as defined here. As the conditioning variable one should take the *stopped sample function*, i.e. the behavoiur of the process up to the stopping time, including information about a possible jump at the stopping time. To the author's opinion, this gives a more satisfactory formulation than can be obtained in terms of the (somewhat obscure) concept of a stopping time $\sigma$-algebra.

*References.*

Blackwell, D. (1956). On a class of probability spaces. *Proc. Third Berkeley Symp. Math. Statist. Prob. vol. 2*, 1 - 6. University of California Press, Berkeley.

Bourbaki, N. (1959). Intégration, chapitre 6. *Paris, Herman.*

Bourbaki, N. (1965a). Intégration, chapitre 1 - 4. *Paris, Herman.*

Bourbaki, N. (1965b). Intégration, chapitre 5. *Paris, Herman.*

Bourbaki, N. (1969). Intégration, chapitre 9. *Paris, Herman.*

Doob, J.L. (1953). *Stochastic Processes.* Wiley, New York.

Federer, H. (1969). *Geometric Measure Theory.* Springer, Berlin.

Halmos, P.R. (1941). The decompositions of measures. *Duke Math. J.* 8, 386-392.

Halmos, P.R. (1950). *Measure Theory.* Van Nostrand, New York.

Hicks, N.J. (1965). *Notes on Differential Geometry.* Van Nostrand, Princeton.

Kelley, J.L. (1955). *General Topology*. Van Nostrand, Princeton.

Kolmogorov, A.N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin.

Lusin, N.(1912). Sur les propriétés des fonctions mesurable. *C.R. Acad. Sci. Paris* 154, 1688-1690.

Nelson, E. (1959). Regular measures on function spaces. *Ann. Math.* 69, 630-643.

Tjur, T. (1972). *On the Mathematical Foundations of Probability*. Inst. Math. Statist. Univ. Copenhagen.

Tjur, T. (1974). *Conditional Probability Distributions*. Inst. Math. Statist. Univ. Copenhagen.