

II

Probability, Information, and Correlation

§1

Joint distributions, marginal & conditionals - (notation) (of countable denumerable)

We assume that we have a collection of sets, X, Y, \dots, Z whose elements we denote by $x_i \in X, y_j \in Y, \dots, z_k \in Z$ and that we have a joint probability distribution,

$P(x_i, y_j, \dots, z_k)$ which represents the occurrence of the combined event $x_i, y_j, \dots \text{ and } z_k$. We write this, for brevity, $P_{(x_i, y_j, \dots, z_k)} = P(x_i, y_j, \dots, z_k)$. We may then think of X, Y, \dots, Z as random variables whose values are the elements of their sets.

For any subset, Y, \dots, Z of a set of random variables W, X, Y, \dots, Z with joint distribution $P_{(w_i, x_i, y_i, \dots, z_i)}$ we define the Marginal distribution, $P_{(y_j, \dots, z_k)}$:

$$(1) \quad P(y_k \dots z_e) = \sum_{i_1 \dots i_l} P(w_i \dots x_j, y_k \dots z_e)$$

which represents the probability of the joint occurrence of $y_k \dots$ and z_e , with no restriction upon the remaining variables.

Next, for any subset $Y \dots Z$ of the random variables we define the Conditional distribution, $P^{w_i \dots x_j}(y_k \dots z_e)$, conditioned upon the values $W=w_i \dots X=x_j$, to be:

$$(2) \quad P^{w_i \dots x_j}(y_k \dots z_e) = \frac{P(w_i \dots x_j, y_k \dots z_e)}{P(w_i \dots x_j)}$$

which represents the probability of the joint event $Y=y_k \dots Z=z_e$, conditioned by the fact that $W \dots X$ are known to have taken the values $W=w_i \dots X=x_j$.

(2)

Finally, for any ^{numerical valued} function $F(y_1, \dots, z_e)$ we define the expectation, $\text{Exp}[F]$ as:

$$(1e(3)) \quad \text{Exp}[F] = \sum_{k=1}^l P(y_k, \dots, z_e) F(y_k, \dots, z_e)$$

and we note that if $P(y_k, \dots, z_e)$ is a marginal distribution of some larger distribution $P(w_i, x_j; y_k, \dots, z_e)$ that

$$\begin{aligned} \text{Exp}[F] &= \sum_{k=1}^l \left(\sum_{i,j} P(w_i, x_j; y_k, \dots, z_e) \right) F(y_k, \dots, z_e) \\ &= \sum_{i,j,k=1}^l P(w_i, x_j, y_k, \dots, z_e) F(y_k, \dots, z_e) \end{aligned}$$

so that if we wish to compute $\text{Exp}[F]$ with respect to some joint distribution, it suffices to use any marginal distribution of the original distribution which contains at least those variables which occur in F .

We may also be interested in Conditional Expectation, $\text{Exp}_{w_i, \dots, x_j}^{\text{w}, \dots, \text{x}}[F]$ which we define as:

$$(1e(4)) \quad \text{Exp}_{w_i, \dots, x_j}^{\text{w}, \dots, \text{x}}[F] = \sum_{k=1}^l P^{w, \dots, x}(y_k, \dots, z_e) F(y_k, \dots, z_e)$$

and we note the following rules: $\text{Exp}[\text{Exp}(E)] = \text{Exp}[E]$

(Def of Independence) $\text{Exp}[\text{Exp}^{w, \dots, w}[F]] = \text{Exp}[F]$

$$\text{Exp}[F+G] = \text{Exp}[F] + \text{Exp}[G]$$

§2 Information:

Suppose that we have a single random X , with distribution $P(X_i)$. Then we define a number, I_X , called the information of X to be:

$$(2.1) \quad I_X = \sum_i P(X_i) \ln P(X_i) = \text{Exp} [\ln P(X_i)]$$

The information is essentially a measure of the spread of a probability distribution as will be made more clear shortly. That is, any change in the distribution $P(X_i)$ which "levels out" the probabilities decreases the information. It is zero for "perfectly sharp" distributions, of form $P(X_i) = \delta_{ij}$, and strictly negative for all others. Many arguments can be given to show that this definition corresponds closely to our intuitive notions of what constitutes information.

In a similar fashion, we define the information of a group of random variables X, Y, \dots, Z , with joint distribution $P(X_i, Y_j, \dots, Z_k)$, I_{XYZ} , to be:

$$(2.2) \quad I_{XYZ} = \sum_{ijk} P(X_i, Y_j, Z_k) \ln P(X_i, Y_j, Z_k) = \text{Exp} [\ln P(X_i, Y_j, Z_k)]$$

Finally, we define a conditional information $I_{XY \dots Z}^{v_1 \dots w_m}$

$$(2.3) \quad I_{XY \dots Z}^{v_1 \dots w_m} = \sum_{ijk} P^{v_1 \dots w_m}(X_i, Y_j, Z_k) \ln P^{v_1 \dots w_m}(X_i, Y_j, Z_k) \\ = \text{Exp}^{v_1 \dots w_m} [\ln P^{v_1 \dots w_m}]$$

a quantity which measures our information about $XY \dots Z$, given that we know that $V=v_1, \dots, W=w_m$.

Some further properties of information are:

(1) Information is a function only of the probabilities themselves, and in no way depends upon numerical values of the random variables, so that it is defined for probability distributions over arbitrary sets, and not restricted to distributions over numerical values, as are the usual measures of "spread" such as variance, etc.

(2) For independent random variables X, Y, \dots, Z ,
 $I_{XY\dots Z} = I_X + I_Y + \dots + I_Z$ so that the information about X, Y, \dots, Z is the sum of the individual informations, which is in accord with our intuitive feeling that if we are given information about independent events then our total information is simply the sum of the individual amounts of information. Also, this requirement of additivity for independent events limits the possible definitions of information to essentially only the one given here, a fact that lends great plausibility to our definition [see Shannon].

Finally, we shall list some useful Factor Theorems:

a) f is convex, G function of random variables,

$$\Rightarrow f[\text{Exp}[G]] \leq \text{Exp}[f(G)]$$

b) $X \ln X$ is convex $(\ln x$ is concave)

c) Theorem $P'_i = \sum_j Q_{ij} P_j$, $\sum_j Q_{ij} = \sum_i Q_{ij} = 1$, $Q_{ij} \leq 1$

$$\Rightarrow I' \leq I \quad (\text{so that any leveling out decreases info})$$

(d) $I_x \geq I_{xy}$, $I_x \leq I_{xT} - I_{yT} \leq 0$

§3 Correlation

Suppose that we have a pair of discrete random variables, X and Y , with joint distribution $P(X_i, Y_j)$. If one makes the statement that X and Y are correlated, what is basically meant is that one learns something about one variable when he is told the value of the other. Let us focus our attention upon the variable X . If we are not informed of the value of Y , then our information about X , I_X , is given by the marginal distribution $P(X_i)$. However, if we are now told that Y has the value y_j , then our information about X changes to the information of the conditional distribution $P^{y_j}(X_i)$, $I_X^{y_j}$. According to what has been said, we wish the degree of correlation to measure how much we learn about X by being informed of the value of Y . However, since this change of information, $I_X^{y_j} - I_X$ may depend upon the particular value, y_j , of Y which we are told, the natural thing to do, in order to arrive at a single number for a measure of the degree of correlation, is to consider the expected change in information about X , given that we are to be told the value of Y . This quantity we shall call the Correlation information, or correlation quantity, and denote by $\{X, Y\}_j$; thus:

$$(3.1) \quad \{X, Y\}_j = \text{Exp}[I_X^{y_j} - I_X] = \text{Exp}[I_X^{y_j}] - I_X$$

Expanding the quantity $\text{Exp}[\mathcal{I}_X^{y_i}]$; using rules (1,4)

$$\begin{aligned}
 (3.2) \quad \text{Exp}[\mathcal{I}_X^{y_i}] &= \text{Exp}\left[\text{Exp}^j\left[\ln P^{y_i}(x_i)\right]\right] \\
 &= \text{Exp}\left[\ln \frac{P(x_i, y_i)}{P(y_i)}\right] = \text{Exp}\left[\ln P(x_i, y_i) - \ln P(y_i)\right] \\
 &= \text{Exp}\left[\ln P(x_i, y_i)\right] - \text{Exp}\left[\ln P(y_i)\right] = \mathcal{I}_{XY} - \mathcal{I}_Y
 \end{aligned}$$

and combining with (3.1) we have:

$$(3.3) \quad \{X, Y\} = \mathcal{I}_{XY} - \mathcal{I}_X - \mathcal{I}_Y$$

So that it is symmetric in X and Y , and hence equal to the expected change in information about Y , given that we will be told the value of X . Furthermore, by (3.3) corresponds precisely to the amount of "missing information" if we possess only the marginal distribution for X and Y .

Theorem 1 $\{X, Y\} = 0$ if and only if X and Y are independent, and is otherwise strictly positive.

Proof: write $P_{ij} = P(x_i, y_j)$, $P(x_i) = P_i$, $P(y_j) = P_j$

and let $Q_{ij} = \frac{P_{ij}}{P_i P_j}$ ($= 1$ if $P_{ij} = 0$) so that $P_{ij} = Q_{ij} P_i P_j$

$$\text{then } \{X, Y\} = \text{Exp}\left[\ln P_{ij}\right] - \text{Exp}\left[\ln P_i\right] - \text{Exp}\left[\ln P_j\right]$$

$$= \text{Exp}\left[\ln P_{ij}/P_i P_j\right] = \text{Exp}\left[\ln Q_{ij}\right] = \sum_{ij} P_i P_j Q_{ij} \ln Q_{ij}$$

making use of the inequality $x \ln x \geq 1 - x$ unless $x = 1$ we have

Part II
inference

$$P_i P_j Q_{ij} \ln Q_{ij} > P_i P_j (1 - Q_{ij}) \quad \text{unless } Q_{ij} = 1 \text{ or } P_i P_j = 0$$

$$\Rightarrow \sum_{ij} P_i P_j Q_{ij} \ln Q_{ij} > \sum_{ij} P_i P_j - \sum_{ij} Q_{ij} P_i P_j = 0 \quad \text{unless for all } i, j \\ Q_{ij} = 1 \text{ or } P_i P_j = 0$$

$$\Rightarrow \{X, Y\} > 0 \quad \text{unless } P_{ij} = P_i P_j \text{ all } i, j \text{ (independent)}$$

QED.

In this respect the correlation so defined is superior to the usual correlation coefficients of statistics, such as covariance, etc., which can be zero even when the variables are not independent, and which can assume both positive + negative values. A negative correlation is, after all, quite as useful as a positive correlation.

We can generalize (3.3) to define a group correlation for the groups of random variables $(U \dots V), (W \dots X), \dots, (Y \dots Z)$, denoted by $\{U \dots V, W \dots X, Y \dots Z\}$, to be:

$$\underline{3.4} \quad \{U \dots V, W \dots X, \dots, Y \dots Z\} = I_{U \dots V, W \dots X, \dots, Y \dots Z} - I_{U \dots V} - I_{W \dots X} - \dots - I_{Y \dots Z}$$

again measuring the information deficiency of the marginals. Theorem 1 is also satisfied by the group correlation, so that it is 0 if and only if the groups are mutually indep. And, of course, we can define conditional correlations if we wish, in the obvious manner.

we list some easily proved relations:

3.5 Commas removal:

$$\{..., u, v, ...\} = \{..., uv, ...\} + \{u, v\}$$

$$\{..., u, v, ..., w, ...\} = \{..., uv...w, ...\} + \{u, v, ..., w\}$$

3.6 commutator:

$$\{..., u, vw, ...\} - \{..., uv, w, ...\} = \{u, v\} - \{v, w\}$$

3.7 $\{X\} = 0$ (definition of bracket with no commas)

3.8 $\{..., xxv, ...\} = \{..., xv, ...\}$ (variables repeated within commas may be omitted)

3.9 $\{..., uv, vw, ...\} = \{..., uv, w, ...\} - I_v - \{v, w\}$

3.10 $\{x, x\} = -I_x$ (repeated across comma)

3.11 $\{u, vw, x\}^{w_1\dots} = \{u, vx\}^{w_1\dots}$

$\{u, w, x\}^{w_1\dots} = \{u, x\}^{w_1\dots}$ (conditional variables may be removed)

3.12 $\{xy, z\} \geq \{x, z\}, \quad \{x, yz\} \geq \{x, y\} + \{x, z\} - \{y, z\}$

3.13 $\{x, y, z\} \geq \{x, y\} + \{x, z\}$ (3.5-3.11) *variable may be replaced by a group*

Note, in above formulae, any random variable, W may be replaced by a group $XZ\dots Z$ and the relation remains true.

Note that we are now able to compute correlations for distributions which have both discrete and continuous parts, by simply choosing as measure Lebesgue measure for the continuous part and the uniform measure for the discrete part, so that if our distribution is $\rho(x)$ density, with discrete "lumps" $\tilde{\rho}(x_i)$, we define the info to be

$$I_x = \int \rho(x) \ln \frac{\rho(x)}{\rho} dx + \sum_i \tilde{\rho}(x_i) \ln \tilde{\rho}(x_i)$$

or if $P(x, y)$ and lumps $\tilde{P}(x_i, y_j)$

$$\Rightarrow P(x) = \int P(x, y) dy \quad \tilde{P}(x_i) = \sum_j \tilde{P}(x_i, y_j)$$

$$I_{xy} - I_x - I_y = \int P(x, y) \ln P(x, y) dxdy + \sum_{ij} \tilde{P}_{ij} \ln \tilde{P}_{ij}$$

$$- \int P(x) \ln P(x) dx - \sum_i \tilde{P}_i \ln \tilde{P}_i$$

$$= \int P(x, y) \ln \frac{P(x, y)}{P(x)P(y)} dxdy + \sum_{ij} \tilde{P}_{ij} \ln \frac{\tilde{P}_{ij}}{\tilde{P}_i \tilde{P}_j}$$

$$= \underbrace{\{X, Y\}}_{\text{continuous}} + \underbrace{\{X, Y\}}_{\text{discrete}}$$

This result is useful for quantum mechanics where spectra can be mixed discrete and continuous.

Theorem 1. If directed set, if f, g functions
on directed set,

3. $\lim f = a \quad \lim g = b$

$\Rightarrow \lim(f+g) = a+b$

so that if the individual limits exist
then the limit of the sum exists and is
equal to the sum of the limits.

Therefore, Defining Informations

$I_{x_1 \dots z}, I_x \dots I_z$ as directed
set limits

We have that: $I_x \dots I_z \cdot I_{x \dots z}$ exist
(they always do)
by monotonicity

$$\Rightarrow \lim(I_{x \dots z}^P - I_x^P - I_y^P - I_z^P)$$

$$= \lim I_{x \dots z}^P - \lim I_x^P - \dots - \lim I_z^P$$

$$\{x, P\} = I_{x \dots z} - I_x - \dots - I_z$$

so that the formula holds as long as right hand
side not indeterminate

Moreover, Since for each partition invariant
so is rest QED -

We now consider the effects decomposition of the values of random variables. For example, we may discover that the event X_i is actually the disjunction of several ^{exclusive} events $\tilde{X}_i^1 \dots \tilde{X}_i^n$, so that X_i occurs if any of the \tilde{X}_i^u occurs. If the probabilities for the \tilde{X}_i^u are $P(\tilde{X}_i^u)$, then $P(X_i) = \sum_u P(\tilde{X}_i^u)$. Similarly, if we had a joint distribution $P(\tilde{X}_i^u, y_j, z_k)$ then $P(X_i, y_j, z_k) = \sum_u P(\tilde{X}_i^u, y_j, z_k)$. In general, we shall say that a distribution $P' = P'(\tilde{X}_i^u, \dots, \tilde{Y}_j^v)$ is a refinement of the distribution $P(X_i, \dots, Y_j)$ if:

$$P(X_i, \dots, Y_j) = \sum_{u, \dots, v} P'(\tilde{X}_i^u, \dots, \tilde{Y}_j^v)$$

Thus by a refinement of a probability distribution we mean the ability to distinguish between events which were previously considered to be a single event. For example, if we had a continuous probability density $P(X, Y)$, then by division of the axes into ^{a finite number of} intervals we arrive at a finite point probability distributions over the rectangles in $X-Y$ space by integrating $P(X, Y)$ over each rectangle, which represents the probability that X, Y is contained in the rectangle. If we now subdivide the intervals, and hence the rectangles, the new joint distribution is a refinement of the old distrit.

We now state an important theorem concerning the behavior of correlations under a refinement of a joint distribution:

Theorem 2 No correlation product decreases under refinement of a joint probability distribution.

$$P' \text{ refinement of } P \Rightarrow \{X, \dots, Y\}' \geq \{X, \dots, Y\}$$