

Chapter 1

Conditional Independence

Date: September 12, 2004

In this chapter we will work on a general probability space (Ω, \mathbb{F}, P) . All events occurring will silently be assumed to be \mathbb{F} -measurable, all σ -algebras occurring will silently be assumed to be subalgebras of \mathbb{F} , and all stochastic variables $X : (\Omega, \mathbb{F}) \rightarrow (\mathcal{X}, \mathbb{E})$ will silently be assumed to be $\mathbb{F} - \mathbb{E}$ measurable.

The general convention is that stochastic variables with names like X or X_i or variations thereof have values in a generic space $(\mathcal{X}, \mathbb{E})$, unless it is explicitly stated that they are realvalued (or integervalued or whatever). Similarly, variables with names like Y or Z will have values in $(\mathcal{Y}, \mathbb{K})$ and $(\mathcal{Z}, \mathbb{G})$ respectively.

Recall that $(\mathcal{X}, \mathbb{E})$ is a **Borel space** if it is in bijective, bimeasurable correspondence with (\mathbb{R}, \mathbb{B}) . Such a correspondence enables us to replace \mathcal{X} with \mathbb{R} , whenever there is an advantage in that. It turns out that every sensible space has this property, unless it is very, very small (countable) or very, very huge (non-separable metric spaces, with the σ -algebra generated by the open sets, say).

The above generic \mathcal{X} , \mathcal{Y} and \mathcal{Z} -spaces are always assumed to be either Borel spaces or countable spaces. Any countable space can of course be embedded in (\mathbb{R}, \mathbb{B}) - map the points to the integers, say - so for most purposes it is not necessary to make explicit reformulations of the results for the countable case.

1.1 Conditional probabilities - a review

We recall that for a realvalued stochastic variable X satisfying that $E|X| < \infty$ and a σ -algebra \mathbb{H} , the **conditional expectation** $E(X | \mathbb{H})$ of X given \mathbb{H} is any \mathbb{H} -measurable stochastic variable satisfying the integral conditions

$$\int_H E(X | \mathbb{H}) dP = \int_H X dP \quad \text{for all } H \in \mathbb{H}. \quad (1.1)$$

The conditional expectation is guaranteed to exist by the Radon-Nikodym theorem, and it is more or less unique: two \mathbb{H} -measurable variables satisfying (1.1) will be a.e.-identical. If the only nullset in \mathbb{H} is the empty set, then the conditional expectation with respect to \mathbb{H} is in fact unique.

We shall be concerned with the **conditional probability** of an event A given \mathbb{H} . This is simply the conditional expectation of the indicator 1_A , that is

$$P(A | \mathbb{H}) = E(1_A | \mathbb{H}).$$

The integrability condition (1.1) will in this case take the form

$$\int_H P(A | \mathbb{H}) dP = P(A \cap H) \quad \text{for all } H \in \mathbb{H}. \quad (1.2)$$

$P(A | X = x)$. More general: $E(Y | X = x)$. Distributional construct.

We will make frequent use of the monotonicity property of conditional expectations, that make sure that

$$0 \leq P(A | \mathbb{H}) \leq 1 \quad \text{a.e.}$$

and even that

$$A \subset B \quad \Rightarrow \quad P(A | \mathbb{H}) \leq P(B | \mathbb{H}) \quad \text{a.e.}$$

Furthermore, the double conditioning theorem says in this context that

$$E(P(A | \mathbb{H}) | \mathbb{G}) = P(A | \mathbb{G}) \quad \text{a.e.}$$

whenever the two σ -algebras \mathbb{G} and \mathbb{H} satisfies that $\mathbb{G} \subset \mathbb{H}$.

The situation where $\mathbb{H} = \mathbb{F}(X)$ for a stochastic variable with values in a space $(\mathcal{X}, \mathbb{E})$ (which of course is quite usual) calls out for special notation. If Y is another stochastic

variable, with values in a space $(\mathcal{Y}, \mathbb{G})$, then it is $\mathbb{F}(X)$ -measurable if and only if there is a $\mathbb{E}-\mathbb{G}$ measurable map $\phi : \mathcal{X} \rightarrow \mathcal{Y}$ such that $Y = \phi(X)$. In the case of a conditional probability with respect to $\mathbb{F}(X)$ we will typically write this map ϕ as $P(A | X = x)$. Hence the map $x \mapsto P(A | X = x)$ is a $\mathbb{E} - \mathbb{B}$ measurable map $\mathcal{X} \rightarrow \mathbb{R}$ which satisfies that

$$\int_B P(A | X = x) dX(P)(x) = \int_{(X \in B)} P(A | X = \bullet) \circ X dP = P(A \cap (X \in B)) \quad (1.3)$$

for any $B \in \mathbb{E}$ and $A \in \mathbb{F}$. The laque of uniqueness of $P(A | X)$ carries over, and can at most bes sure that any two versions of $P(A | X = x)$ will agree $X(P)$ -almost surely.

More generally, we speak of $E(Y | X = x)$ for any realvalued, integrable stochastic variable, as any map that composed with X gives $E(Y | X)$.

An important point of the 'pointwise conditional means' $E(Y | X = x)$ is that they are in a sense a more solid construct, with at real existence - they do not vapourize into the abstractions of the backgroundspace Ω . They are realvalued maps on \mathcal{X} , and hence something that - in principle - can be computed, not merely shown to exist.

And they are distributional concepts. Suppose (X, Y) has the same joint distribution as (X', Y') , and that $\phi : \mathcal{X} \rightarrow \mathbb{R}$ is a version of $E(Y | X = x)$ - then it is also a version of $E(Y' | X' = x)$. This follows from (1.3).

$P(A | X = x)$. More general: $E(Y | X = x)$. Distributional construct.

We say that two σ -algebras \mathbb{H} and \mathbb{H}^* are **almost surely equal**, in shorthand written $\mathbb{H} = \mathbb{H}^*$ a.e., if every event $H \in \mathbb{H}$ can be matched by an event $H^* \in \mathbb{H}^*$, such that $P(H \Delta H^*) = 0$, and vice versa. A typical situation where this concept is relevant, is when we have two stochastic variables, X and X^* , that are equal almost everywhere. The two σ -algebras $\mathbb{F}(X)$ and $\mathbb{F}(X^*)$ will not be identical, but as

$$(X \in B) = (X^* \in B) \quad \text{a.e.}$$

for any set B , the two σ -algebras will be almost surely equal.

Lemma 1.1 *If the two σ -algebras \mathbb{H} and \mathbb{H}^* are almost surely equal, then*

$$P(A | \mathbb{H}) = P(A | \mathbb{H}^*) \quad \text{a.e.}$$

for every event A .

PROOF: The simple situation arise if $\mathbb{H} \subset \mathbb{H}^*$. In that case we simply prove that $P(A | \mathbb{H})$ - which clearly is \mathbb{H}^* -measurable - satisfies the integral condition for being a version of $P(A | \mathbb{H}^*)$. If we consider $H^* \in \mathbb{H}^*$, we can match it by some event $H \in \mathbb{H}$, and thus we have that

$$\int_{H^*} P(A | \mathbb{H}) dP = \int_H P(A | \mathbb{H}) dP = P(A \cap H) = P(A \cap H^*).$$

If \mathbb{H} and \mathbb{H}^* are not nested, they are certainly both contained in $\mathbb{H} \vee \mathbb{H}^*$. But this larger σ -algebra is seen to be almost surely equal to both \mathbb{H} and \mathbb{H}^* . It is generated by intersections of the form $H \cap H^*$, with $H \in \mathbb{H}$ and $H^* \in \mathbb{H}^*$, and such intersections are easily matched by \mathbb{H} -events or by \mathbb{H}^* -events. And the $\mathbb{H} \vee \mathbb{H}^*$ -events that can be matched, are easily seen to form a σ -algebra.

Hence, in the general case, we can write

$$P(A | \mathbb{H}) = P(A | \mathbb{H} \vee \mathbb{H}^*) = P(A | \mathbb{H}^*) \quad \text{a.e.}$$

□

We will need a ramification of lemma 1.1 to situations where the two σ -algebras are not exactly equal up to nullsets, but where each event in one of the algebras is 'almost matched' in the other. It is rather difficult to formulate the result in the general case, and we will only need it under quite special circumstances, so the following formulation, with σ -algebras generated by stochastic variables, will suffice:

Theorem 1.2 *Let X and Y be two stochastic variables, with values in the same space $(\mathcal{X}, \mathbb{E})$. For any event A and any $\alpha > 0$ it holds that*

$$P\left(\left|P(A | X) - P(A | Y)\right| > \alpha\right) \leq \frac{16 P(X \neq Y)}{\alpha}.$$

PROOF: The key technical result we will have to prove is that

$$\left|\int_D P(A | X) dP - P(A \cap D)\right| \leq 2P(X \neq Y), \quad (1.4)$$

for any event $D \in \mathbb{F}(X, Y)$. If we have this inequality at our disposal, we can use it on the event

$$D^+ = (P(A | X) - P(A | X, Y) > \alpha)$$

which is $\mathbb{F}(X, Y)$ -measurable, to obtain that

$$\begin{aligned} \alpha P(D^+) &\leq \int_{D^+} P(A | X) - P(A | X, Y) dP = \int_{D^+} P(A | X) dP - P(A \cap D^+) \\ &\leq 2P(X \neq Y). \end{aligned}$$

Using a similar argument in the other tail, we obtain that

$$P\left(\left|P(A | X) - P(A | X, Y)\right| > \alpha\right) \leq \frac{4P(X \neq Y)}{\alpha}.$$

And observing that the event

$$\left(\left|P(A | X) - P(A | Y)\right| > \alpha\right)$$

is a subset of

$$\left(\left|P(A | X) - P(A | X, Y)\right| > \frac{\alpha}{2}\right) \cup \left(\left|P(A | Y) - P(A | X, Y)\right| > \frac{\alpha}{2}\right)$$

the theorem is established.

To show (1.4), take $D \in \mathbb{F}(X, Y)$. We can assume that $D = ((X, Y) \in B)$ for some set $B \in \mathbb{E} \otimes \mathbb{E}$. Let

$$D^* = ((X, X) \in B).$$

Clearly D^* is $\mathbb{F}(X)$ -measurable, and

$$(D \Delta D^*) \subset (X \neq Y).$$

Now we have that

$$\begin{aligned} &\left|\int_D P(A | X) dP - P(A \cap D)\right| \\ &\leq \left|\int_D P(A | X) dP - \int_{D^*} P(A | X) dP\right| + |P(A \cap D^*) - P(A \cap D)| \\ &\leq 2P(D \setminus D^*) + 2P(D^* \setminus D) \end{aligned}$$

as desired. Here we have used that the integrand $P(A | X)$ is bounded by 1. □

Corollary 1.3 *Let X, X_1, X_2, \dots be a sequence of stochastic variables. If*

$$P(X_n = X) \rightarrow 1 \quad \text{for } n \rightarrow \infty,$$

it holds for any event A that

$$P(A | X_n) \xrightarrow{P} P(A | X).$$

PROOF: It follows directly from theorem 1.2 that for any $\alpha > 0$,

$$P\left(\left|P(A | X_n) - P(A | X)\right| > \alpha\right) \leq \frac{16 P(X_n \neq X)}{\alpha} \rightarrow 0 \quad \text{for } n \rightarrow \infty,$$

which establishes convergens in probability. □

1.2 Conditional distributions - a review

Recall that a $(\mathcal{X}, \mathbb{E})$ -Markov kernel on $(\mathcal{Y}, \mathbb{K})$ is a collection $(\hat{P}_x)_{x \in \mathcal{X}}$ of probability measures on $(\mathcal{Y}, \mathbb{K})$, indexed by points in \mathcal{X} in such a way that the map $x \mapsto \hat{P}_x(B)$ is \mathbb{E} measurable for any fixed $B \in \mathbb{K}$.

Recall that such a Markovkernel is termed the **conditional distribution of Y given X** if it satisfies that

$$P(X \in A, Y \in B) = \int_A \hat{P}_x(B) dX(P)(x) \quad \text{for all } A \in \mathbb{E}, B \in \mathbb{K}.$$

The main problem with the concept of conditional distributions is that the conditional distribution is not unique. If we have one conditional distributions, we can alter the probability measures in it as we like, as long as we only alter a few of them - to be precise, we can only alter them for x 's in a $X(P)$ -nullset. This non-uniqueness give rise to a multitude of formulational problems, that tends to obscure the theory, but somehow is of no real consequence.

Theorem 1.4 *If Y has values in a Borel space, the conditional distribution of Y given X does exist.*

The idea behind theorem 1.4 is that the conditional probabilities $P(Y \in B \mid X = x)$, which are known to exist by the Randon-Nikodym theorem, can be chosen with proper care and combined into genuine probability measures. The whole thing is a giant nullset circus, and it is not very illuminating. But note that it may be impossible to combine the conditional probabilities correctly without the Borel space assumption - there are explicit counterexamples in Billingsley (1995).

The existence theorem is nice and reassuring to have, but is usually not that important. In most cases we do not have to rely on abstract existence theorems, but have explicit formulas for the conditional distributions:

Conditional distributions in case of independence.

Example 1.5 Suppose $(\mathcal{X}, \mathbb{E})$ is a countable set, equipped with the σ -algebra of all subsets. The conditional distribution of Y given X is thus

$$\hat{P}_x(B) = \frac{P(X = x, Y \in B)}{P(X = x)} \quad \text{for all } B \in \mathbb{K},$$

At least this is true for all the x 's for which $P(X = x) > 0$, and that will suffice: the remaining x 's form a $X(P)$ -nullset.

If \mathcal{Y} is also countable, we usually give the conditional distribution in terms of the conditional probability function,

$$q_x(y) = \frac{P(X = x, Y = y)}{P(X = x)}$$

as this function satisfies

$$\hat{P}_x(B) = \sum_{y \in B} q_x(y).$$

o

Example 1.6 Let $\bar{\mu}$ and $\bar{\nu}$ be σ -finite measures on $(\mathcal{X}, \mathbb{E})$ and $(\mathcal{Y}, \mathbb{K})$. If X and Y has joint density f with respect to $\bar{\mu} \otimes \bar{\nu}$, meaning that

$$P(X \in A, Y \in B) = \int_{A \times B} f(x, y) d\bar{\mu} \otimes \bar{\nu}(x, y) \quad \text{for all } A \in \mathbb{E}, B \in \mathbb{K},$$

then the conditional distribution of Y given X is a family of measures with density with respect to $\bar{\nu}$. To be precise

$$\hat{P}_x(B) = \int_B g_x(y) d\bar{\nu}(y)$$

where the densities are given by

$$g_x(y) = \frac{f(x,y)}{h(x)}, \quad h(x) = \int f(x,y) d\bar{\nu}(y).$$

At least this is true for all x 's for which $0 < h(x) < \infty$, and this will suffice: the remaining x 's form a $X(P)$ -nullset. ◦

There are two virtues of the conditional distributions. One is conceptual: they allow us to think of the combined observation of X and Y in terms of a two-step procedure, where we first observe X , and then observe Y . This two-step mental picture is often the key to reduce difficult questions involving simultaneous distributions to more manageable ones.

The other virtue is computational. If $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is measurable, it holds that

$$E\phi(X, Y) = \iint \phi(x, y) d\hat{P}_x(y) dX(P)(x) \tag{1.5}$$

whenever the left-hand side exist - that is, if ϕ is non-negative or if $E|\phi(XY)| < \infty$. We refer to (1.5) as the extended Tonelli theorem or the extended Fubini theorem, as the case may be.

In our setting, we are concerned about with the somewhat simpler concept of conditional probabilities. Conditional distributions give an explicit way of calculating conditional probabilities, where the scatter of nullsets somehow is gone (it is of course still there, now buried in the non-uniqueness of the conditional distributions). If we compose that map $x \mapsto \hat{P}_x(B)$ with the stochastic variable X , the resulting stochastic variable is obviously measurable with respect to $\mathbb{F}(X)$, and satisfies the integral condition for being a version of $P(Y \in B | X)$.

This should be written out as a theorem

In fact, we can extend this argument somewhat. If $E|\phi(Y)| < \infty$, we may construct the map $x \mapsto \int \phi(y) d\hat{P}_x(y)$. Composing with the stochastic variable X , we get a

version of $E(\phi(Y) | X)$. Litterally speaking, the integrals $\int \phi(y) d\hat{P}_x(y)$ may not exist for every value of x , but x 's for which this goes wrong, form a $X(P)$ -nullset, and so we can replace the meaningless integral by zero or some other suitable value, without interfering with the integral property.

Change the theorem - introduce X' with the same distribution as X .

Theorem 1.7 *Let X and Y be stochastic variables with values in $(\mathcal{X}, \mathbb{B})$ and $(\mathcal{Y}, \mathbb{K})$. There exists a map $\phi : \mathcal{X} \times (0, 1) \rightarrow \mathcal{Y}$, which is $\mathbb{B} \otimes \mathbb{B}_{(0,1)} - \mathbb{K}$ measurable, with the following property: if U is a realvalued stochastic variable, independent of X and uniformly distributed on $(0, 1)$, and if we let*

$$Y' = \phi(X, U)$$

then (X, Y') has the same distribution as (X, Y) .

PROOF: Due to the underlying assumption that the spaces involved are Borel spaces, we may assume that $(\mathcal{Y}, \mathbb{K}) = (\mathbb{R}, \mathbb{B})$. Let $(\hat{P}_x)_{x \in \mathcal{X}}$ be the conditional distribution of Y given X .

We know that the conditional distribution of U given X is very degenerate:

$$\hat{Q}_x = \nu \quad \text{for all } x \in \mathcal{X}$$

where ν is the uniform distribution on $(0, 1)$. By the substitution theorem, the conditional distribution of Y' given X is

$$\hat{R}_x = \phi \circ i_x(\hat{Q}_x) = \phi \circ i_x(\nu).$$

The proof is complete, once we show how to choose ϕ such that $\hat{R}_x = \hat{P}_x$ for every x , as the joint distribution is uniquely determined from one marginal distribution and the conditional distribution of the remaining marginal given the first.

The deep claim is not so much that it is possible to choose ϕ is such a way that

$$\phi \circ i_x(\nu) = \hat{P}_x \quad \text{for all } x \in \mathcal{X}. \tag{1.6}$$

For if we let F_x be the distribution function corresponding to \hat{P}_x , and if we let q_x be a quantile function for F_x , it is well known that $q_x(\nu) = \hat{P}_x$. So we may let

$$\phi(x, u) = q_x(u),$$

and (1.6) will be satisfied bona fide.

What is a deep claim is that the construction can be carried out in a way that guarantees ϕ to be measurable. There is a choice involved, in the sense that quantile functions are not unique, and even though the individual quantile functions are increasing, and thus necessarily measurable, the various choices may destroy joint measurability.

The key is to get rid of the choices, and find an operationally defined quantile function. A nice one is

$$q_x(p) = \inf\{y \in \mathbb{R} \mid F_x(y) > p\} \quad \text{for all } x \in \mathcal{X}, p \in (0, 1).$$

The idea is to single out the largest possible p -quantile whenever there is a choice. Let us prove that this is in fact a quantile function:

For fixed x and p , we have that

$$\{y \in \mathbb{R} \mid F_x(y) > p\} = \begin{cases} (y_0, \infty) \\ [y_0, \infty) \end{cases},$$

for some $y_0 \in \mathbb{R}$. Whether we have the open or the halfclosed interval, depends on the specifics of the situation, but in both cases we see that $q_x(p) = y_0$. For each n we have that $y_0 + \frac{1}{n} > y_0$, and thus

$$F_x\left(y_0 + \frac{1}{n}\right) > p.$$

Using right continuity of F_x , we can conclude that

$$F_x(y_0) \geq p.$$

Similarly, $y_0 - \frac{1}{n} < y_0$, and so

$$F_x\left(y_0 - \frac{1}{n}\right) \leq p.$$

Using monotonicity of F_x , we can conclude that

$$F_x(y_0-) \leq p.$$

Together these inequalities show that y_0 is a p -quantile for F_x .

As for measurability, an elementary argument shows that

$$\{(x, p) \mid q_x(p) < z\} = \bigcup_{w < z, w \in \mathbb{Q}} \{(x, p) \mid F_x(w) > p\}. \quad (1.7)$$

For any fixed w , the map

$$x \mapsto F_x(w) = \hat{P}_x((-\infty, w])$$

is measurable, as $(\hat{P}_x)_{x \in \mathcal{X}}$ is a Markov kernel. Hence $(x, p) \mapsto (F_x(w), p)$ is measurable, and thus

$$\{(x, p) \mid F_x(w) > p\} = \{(x, p) \mid F_x(w) - p > 0\}$$

is measurable set. The fact that the righthand side of (1.7) is a countable union, shows that the lefthand side is a measurable set. □

The point of theorem 1.7 is that we may think of as any pair of variables as generated in a two-step procedure, where the generation of the second variable can be accomplished by mixing the first variable with random noise. It is the way that the mixing is carried out, that determines the joint distribution.

The **update function** ϕ is not at all unique. There are literally uncountably many ways to choose it. In certain cases it matters which one we use, in most cases it is irrelevant. However, in typical applications there is a specific update function that almost forces itself upon us.

Example 1.8 Construct update function for discrete kernel. Simulation.

○

1.3 Conditionally independent events

Definition 1.9 Two events A and B are **conditionally independent** given a σ -algebra \mathbb{H} , if

$$P(A \cap B \mid \mathbb{H}) = P(A \mid \mathbb{H}) P(B \mid \mathbb{H}) \quad a.e. \quad (1.8)$$

Symbolically, we will write $A \perp\!\!\!\perp B \mid \mathbb{H}$ if (1.8) is satisfied.

Speaking colloquially, we will frequently say that A and B are independent given \mathbb{H} if (1.8) is satisfied - repeated use of the word *conditionally* makes the sentences sound tedious.

Please note that conditional independence represents an intricate relation between the two events and the σ -algebra. The σ -algebra \mathbb{H} is really an integral part of the definition. Whether A and B are conditionally independent or not, depends crucially on which σ -algebra we are conditioning.

If $\mathbb{H} \subset \mathbb{G}$ are two σ -algebras, it is completely possible that two events A and B are independent given \mathbb{H} , while they are not independent given the finer σ -algebra \mathbb{G} . But it is equally possible that A and B are independent given \mathbb{G} , while they are not independent given the coarser σ -algebra \mathbb{H} . Changing the σ -algebra on which we are conditioning is usually a very challenging task - and indeed a task which is at the core of Markov Chain Theory.

Example 1.10 Recall that a σ -algebra \mathbb{H} is a **trivial** if every event in \mathbb{H} has probability 0 or 1. The most obvious trivial σ -algebra is

$$\mathbb{H} = \{\emptyset, \Omega\},$$

but there are plenty of other trivial algebras arising all over probability theory - tail algebras, symmetric algebras, invariant σ -algebras in ergodic theory and what not. If \mathbb{H} is trivial, we observe that

$$P(A | \mathbb{H}) = P(A) \quad \text{a.e.}$$

for any event A , since the relation

$$\int_H P(A) dP = P(A \cap H),$$

is satisfied for all \mathbb{H} -sets H , both those of probability 0 (where there is nothing to prove) and those of probability 1 (where there is also nothing to prove). Hence (1.8) translates to

$$P(A \cap B) = P(A)P(B). \quad (1.9)$$

A priori the formula has an a.e.-qualifier, but as it is a relation between deterministic numbers, it is either true or false, with no probability involved.

Hence we see that conditional independence of two events given a trivial σ -algebra is simply classical independence of the events.

◦

Example 1.11 If C is yet another event, and if \mathbb{H} is the σ -algebra generated by that event,

$$\mathbb{H} = \{\emptyset, C, C^c, \Omega\},$$

then it is readily checked that

$$P(A | \mathbb{H}) = \begin{cases} \frac{P(A \cap C)}{P(C)} & \text{on } C \\ \frac{P(A \cap C^c)}{P(C^c)} & \text{on } C^c \end{cases} \quad \text{a.e.}$$

for any event A . If we suppose that \mathbb{H} is non-trivial, meaning that $P(C) \in (0, 1)$, we see that (1.8) translates to the two conditions

$$\frac{P(A \cap B \cap C)}{P(C)} = \frac{P(A \cap C)}{P(C)} \frac{P(B \cap C)}{P(C)},$$

$$\frac{P(A \cap B \cap C^c)}{P(C^c)} = \frac{P(A \cap C^c)}{P(C^c)} \frac{P(B \cap C^c)}{P(C^c)}.$$

These two conditions cannot be deduced from each other, and they are not related to (1.9). For instance, the probability table

	C			C^c	
	B	B^c		B	B^c
A	$\frac{2}{18}$	$\frac{1}{18}$	A	$\frac{2}{18}$	$\frac{4}{18}$
A^c	$\frac{4}{18}$	$\frac{2}{18}$	A^c	$\frac{1}{18}$	$\frac{2}{18}$

corresponds to a situation where $A \perp\!\!\!\perp B | \mathbb{H}$ but where A and B are dependent, as can readily be checked.

On the other hand, the probability table

	C			C^c	
	B	B^c		B	B^c
A	$\frac{1}{12}$	$\frac{2}{12}$	A	$\frac{2}{12}$	$\frac{1}{12}$
A^c	$\frac{2}{12}$	$\frac{1}{12}$	A^c	$\frac{1}{12}$	$\frac{2}{12}$

corresponds to a situation where A and B are independent, but where they are **not** independent given \mathbb{H} .

◦

Example 1.12 If we have a finite partition \mathbb{D} of Ω ,

$$\mathbb{D} = \{D_1, \dots, D_n\}$$

where the **atoms** of \mathbb{D} (the D_i -sets) are pairwise disjoint and unite to the whole of Ω , the σ -algebra generated by \mathbb{D} is the family of all unions,

$$\mathbb{H} = \left\{ \bigcup_{i \in I} D_i \mid I \subset \{1, \dots, n\} \right\}.$$

If we let

$$\mathbb{D}^* = \{D \in \mathbb{D} \mid P(D) > 0\},$$

it is easily checked that

$$P(A \mid \mathbb{H}) = \sum_{D \in \mathbb{D}^*} \frac{P(A \cap D)}{P(D)} 1_D \quad \text{a.e.}$$

for any event A . In this setting, condition (1.8) translates into

$$\frac{P(A \cap B \cap D)}{P(D)} = \frac{P(A \cap D)}{P(D)} \frac{P(B \cap D)}{P(D)} \quad \text{for all } D \in \mathbb{D}^*.$$

Again, whether this holds or not is very sensitive to the specific atoms. If an atom is divided into two, there is no telling if A and B are independent on each of the two subatoms, just because we know if they are independent on the original atom. And similarly, if two atoms are coalesced, we may loose or create conditional independence, as the case may be.

◦

1.4 Conditionally independent σ -algebras

Definition 1.13 Two classes of events, \mathcal{A} and \mathcal{B} , are conditionally independent given a σ -algebra \mathbb{H} if

$$A \perp\!\!\!\perp B \mid \mathbb{H} \quad \text{for all } A \in \mathcal{A}, B \in \mathcal{B}. \quad (1.10)$$

Symbolically, we will write $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathbb{H}$ if (1.10) is satisfied.

We will almost exclusively use this concept in situations where the two classes of events are σ -algebras, but it is nice to be allowed to formulate things in a slightly broader fashion. We may for instance see that it typically is enough to check (1.10) on two generators of the σ -algebras under consideration:

Lemma 1.14 *Let \mathcal{A} and \mathcal{B} be two classes of events, both stable under formation of intersections. Then*

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathbb{H} \quad \Rightarrow \quad \sigma(\mathcal{A}) \perp\!\!\!\perp \sigma(\mathcal{B}) \mid \mathbb{H}.$$

PROOF: A prototypical application of Dynkin's lemma. For each set $F \in \mathcal{F}$ we consider the class

$$C_F = \{E \in \mathcal{F} \mid F \perp\!\!\!\perp H \mid \mathbb{H}\},$$

and we observe that this is a Dynkin class. If we take $A \in \mathcal{A}$, we know that $\mathcal{B} \subset C_A$. Using Dynkin's lemma, we see that $\sigma(\mathcal{B}) \subset C_A$. On the other hand, conditional independence of two events is a property that is symmetric in the two events, so we can reformulate this fact as $\mathcal{A} \subset C_B$ for any set $B \in \sigma(\mathcal{B})$. Using Dynkin's lemma again establishes that $\sigma(\mathcal{A}) \subset C_B$ for any set $B \in \sigma(\mathcal{B})$. And though this may look awkward, it is in fact the property we are after. □

Conditional independence of classes of events is of course just as sensitive to the exact choice of the σ -algebra on which we are conditioning, as conditional independence of events were. In fact, if

$$\mathbb{A} = \{\emptyset, A, A^c, \Omega\}, \quad \mathbb{B} = \{\emptyset, B, B^c, \Omega\},$$

then $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$ if and only if $A \perp\!\!\!\perp B \mid \mathbb{H}$, as is readily seen from lemma 1.14. So the counterexamples to any kind of simple behaviour under change of the conditioning algebra given in section 1.3 also apply in this setting.

Example 1.15 Blow this up? Furthermore, check that \mathbb{A} and \mathbb{B} are independent given any trivial σ -algebra. Check that if \mathbb{A} , \mathbb{B} and \mathbb{H} are independent, the $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$. *****

○

Lemma 1.16 (Reduction) *Let \mathcal{A} and \mathcal{B} be two classes of events, and let $\mathcal{A}' \subset \mathcal{A}$ be a subclass. Then*

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathbb{H} \quad \Rightarrow \quad \mathcal{A}' \perp\!\!\!\perp \mathcal{B} \mid \mathbb{H}.$$

PROOF: This is a quite trivial observation, which hardly deserves to be called a lemma. The statement on the righthand side involves fewer events than the statement on the lefthand side, so the implication is obvious. \square

Theorem 1.17 *Let \mathbb{A}, \mathbb{B} and \mathbb{H} be three σ -algebras. Suppose that $\mathbb{A} \perp \mathbb{B} \mid \mathbb{H}$. If X is an \mathbb{A} -measurable realvalued random variable, and if Y is a \mathbb{B} -measurable realvalued random variable, such that $E|X| < \infty$, $E|Y| < \infty$ and $E|XY| < \infty$, then it holds that*

$$E(XY \mid \mathbb{H}) = E(X \mid \mathbb{H}) E(Y \mid \mathbb{H}) \quad a.e.$$

PROOF: A prototypical extension result. We know the theorem to be true for indicator variables. Hence it is true for simple variables. The monotone convergence theorem for conditional expectations will show it is true for non-negative variables, and a final handwaving will dismiss the problems of positive and negative parts. \square

Conditional independence is by its very definition symmetric in the two events, or more general, in the two classes of events. Rather surprisingly, it turns out that the most fruitfull way of working the the concept is through an asymmetric formulation:

Theorem 1.18 *Let \mathbb{A}, \mathbb{B} and \mathbb{H} be σ -algebras. It holds that $\mathbb{A} \perp \mathbb{B} \mid \mathbb{H}$ if and only if*

$$P(A \mid \mathbb{B} \vee \mathbb{H}) = P(A \mid \mathbb{H}) \quad a.e \quad (1.11)$$

for every event $A \in \mathbb{A}$.

PROOF: Notice that for any three events $A \in \mathbb{A}$, $B \in \mathbb{B}$ and $H \in \mathbb{H}$ we have that

$$\begin{aligned} \int_{B \cap H} P(A \mid \mathbb{H}) dP &= \int_H 1_B P(A \mid \mathbb{H}) dP = \int_H E(1_B P(A \mid \mathbb{H}) \mid \mathbb{H}) dP \\ &= \int_H P(A \mid \mathbb{H}) P(B \mid \mathbb{H}) dP \end{aligned} \quad (1.12)$$

Suppose that \mathbb{A} and \mathbb{B} are conditionally independent given \mathbb{H} . Then we can work the above line of equations one step further to see that

$$\int_{B \cap H} P(A \mid \mathbb{H}) dP = \int_H P(A \cap B \mid \mathbb{H}) dP = P(A \cap B \cap H).$$

The events of the form $B \cap H$ is a generator for the σ -algebra $\mathbb{B} \vee \mathbb{H}$, stable under formation of intersections, and as $P(A | \mathbb{H})$ is \mathbb{H} -measurable, and hence a fortiori $\mathbb{B} \vee \mathbb{H}$ -measurable, we conclude that $P(A | \mathbb{H})$ indeed does satisfy all conditions for being the conditional probability of A given $\mathbb{B} \vee \mathbb{H}$. And hence (1.11) holds.

For the opposite implication, we may utilize (1.11) on the starting end of (1.12), and obtain that

$$\int_H P(A | \mathbb{H}) P(B | \mathbb{H}) dP = \int_{H \cap B} P(A | \mathbb{B} \vee \mathbb{H}) dP = P(A \cap B \cap H).$$

As $P(A | \mathbb{H}) P(B | \mathbb{H})$ is indeed \mathbb{H} -measurable, we see that it satisfies all conditions for being the conditional probability of $A \cap B$ given \mathbb{H} . And hence A and B are conditionally independent given \mathbb{H} . □

The asymmetric condition (1.11) is usually paraphrased by saying that there is no extra information in \mathbb{B} for making predictions on the occurrence of an \mathbb{A} -set, when we already have access to the information in \mathbb{H} . All the information in \mathbb{B} , useful for that prediction, is already contained in \mathbb{H} . The symmetry between \mathbb{A} and \mathbb{B} is not clearly visible here, but somehow it is still there.

Corollary 1.19 *Let \mathbb{A} , \mathbb{B} and \mathbb{H} be σ -algebras. If $\mathbb{A} \perp \mathbb{B} | \mathbb{H}$ then it holds for any \mathbb{A} -measurable real random variable X such that $E|X| < \infty$ that*

$$E(X | \mathbb{B} \vee \mathbb{H}) = E(X | \mathbb{H}) \quad a.e \tag{1.13}$$

PROOF: Follows from 1.18 by the same extension technique, that was used to prove theorem 1.17. □

Example 1.20 Show that $\mathbb{A} \perp \mathbb{H} | \mathbb{H}$. And thus $\mathbb{A} \perp \mathbb{B} | \mathbb{H}$ whenever $\mathbb{B} \subset \mathbb{H}$. *****

○

Lemma 1.21 *Let \mathbb{A} , \mathbb{B} , \mathbb{H} and \mathbb{H}^* be σ -algebras. Suppose that \mathbb{H} and \mathbb{H}^* are almost surely equal. Then*

$$\mathbb{A} \perp \mathbb{B} | \mathbb{H} \quad \Leftrightarrow \quad \mathbb{A} \perp \mathbb{B} | \mathbb{H}^* .$$

PROOF: This result is a pretty trivial consequence of lemma 1.1. If $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}$, and if $A \in \mathbb{A}$ and $B \in \mathbb{B}$, then we have that

$$P(A \cap B \mid \mathbb{H}^*) = P(A \cap B \mid \mathbb{H}) = P(A \mid \mathbb{H})P(B \mid \mathbb{H}) = P(A \mid \mathbb{H}^*)P(B \mid \mathbb{H}^*) \quad \text{a.e}$$

and so $\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H}^*$

□

1.5 Shifting information around

Insert some text

Theorem 1.22 *Let \mathbb{A} , \mathbb{B} and \mathbb{H} be σ -algebras.*

$$\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H} \quad \Rightarrow \quad \mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{H}) \mid \mathbb{H}.$$

PROOF: Take $A \in \mathbb{A}$. We have that

$$P(A \mid (\mathbb{B} \vee \mathbb{H}) \vee \mathbb{H}) = P(A \mid \mathbb{B} \vee \mathbb{H}) = P(A \mid \mathbb{H}),$$

where the first equality is true for trivial reason (we are conditioning on the same σ -algebra), and the second equality is true due to the conditional independence of \mathbb{A} and \mathbb{B} given \mathbb{H} . But now conditional independence of \mathbb{A} and $\mathbb{B} \vee \mathbb{H}$ given \mathbb{H} follows from theorem 1.18.

□

Theorem 1.23 *Let \mathbb{A} , \mathbb{B} and \mathbb{H} be σ -algebras. Suppose that \mathbb{G} is yet another σ -algebra, satisfying that $\mathbb{H} \subset \mathbb{G} \subset \mathbb{H} \vee \mathbb{B}$. Then it holds that*

$$\mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{H} \quad \Rightarrow \quad \mathbb{A} \perp\!\!\!\perp \mathbb{B} \mid \mathbb{G}.$$

PROOF: Take $A \in \mathbb{A}$. By repeated conditioning we have that

$$\begin{aligned} P(A | \mathbb{B} \vee \mathbb{G}) &= E\left(P(A | \mathbb{B} \vee \mathbb{G} \vee \mathbb{H}) | \mathbb{B} \vee \mathbb{G}\right) = E\left(P(A | \mathbb{B} \vee \mathbb{H}) | \mathbb{B} \vee \mathbb{G}\right) \\ &= E\left(P(A | \mathbb{H}) | \mathbb{B} \vee \mathbb{G}\right) = P(A | \mathbb{H}) \end{aligned}$$

as $P(A | \mathbb{H})$ is itself \mathbb{H} -measurable, and thus \mathbb{G} -measurable, and in particular $\mathbb{B} \vee \mathbb{G}$ -measurable. But by the exact same argument we have that

$$P(A | \mathbb{G}) = E\left(P(A | \mathbb{H} \vee \mathbb{B}) | \mathbb{G}\right) = E\left(P(A | \mathbb{H}) | \mathbb{G}\right) = P(A | \mathbb{H}).$$

And thus in particular $P(A | \mathbb{B} \vee \mathbb{G}) = P(A | \mathbb{G})$, which establishes conditional independence of \mathbb{A} and \mathbb{B} given \mathbb{G} . □

Theorem 1.24 *Let \mathbb{A} , \mathbb{B} , \mathbb{G} and \mathbb{H} be σ -algebras. It holds that*

$$\mathbb{A} \perp\!\!\!\perp \mathbb{B} | \mathbb{H} \text{ and } \mathbb{A} \perp\!\!\!\perp \mathbb{G} | \mathbb{B} \vee \mathbb{H} \quad \Rightarrow \quad \mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{G}) | \mathbb{H}.$$

PROOF: Take $A \in \mathbb{A}$. It holds that

$$P(A | (\mathbb{B} \vee \mathbb{G}) \vee \mathbb{H}) = P(A | \mathbb{B} \vee \mathbb{H}) = P(A | \mathbb{H}).$$

The first equality is due to conditional independence of \mathbb{A} and \mathbb{G} given $\mathbb{B} \vee \mathbb{H}$, the second is due to conditional independence of \mathbb{A} and \mathbb{B} given \mathbb{H} . The combination of course gives that \mathbb{A} and $\mathbb{B} \vee \mathbb{G}$ are independent given \mathbb{H} . □

Example 1.25 None of the theorems so far will tell us how to throw information away in the conditioning algebra, while retaining conditional independence. But the theorems can in certain situations be combined to that effect.

Suppose that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} | \mathbb{G} \vee \mathbb{H}$. Theorem 1.24 tells us that if we furthermore know that

$$\mathbb{A} \perp\!\!\!\perp \mathbb{G} | \mathbb{H}$$

then $\mathbb{A} \perp\!\!\!\perp (\mathbb{B} \vee \mathbb{G}) | \mathbb{H}$. But we can throw events away in the classes that are conditionally independent for free, so it acutally follows that $\mathbb{A} \perp\!\!\!\perp \mathbb{B} | \mathbb{H}$. By symmetry, we can also get rid of \mathbb{G} if we know it is conditionally independent of \mathbb{B} given \mathbb{H} . ○

Theorem 1.26 Let X, X_1, X_2, \dots be a sequence of stochastic variables. Suppose that

$$P(X_n = X) \rightarrow 1 \quad \text{for } n \rightarrow \infty.$$

Then it holds for any two events A and B that

$$A \perp\!\!\!\perp B \mid X_n \quad \text{for all } n \quad \Rightarrow \quad A \perp\!\!\!\perp B \mid X.$$

PROOF: From corollary 1.3 it follows that

$$P(A \mid X_n) \xrightarrow{P} P(A \mid X), \quad P(B \mid X_n) \xrightarrow{P} P(B \mid X), \quad P(A \cap B \mid X_n) \xrightarrow{P} P(A \cap B \mid X),$$

for $n \rightarrow \infty$. As

$$P(A \cap B \mid X_n) = P(A \mid X_n) P(B \mid X_n) \quad \text{for all } n,$$

uniqueness of the limit for convergence in probability guarantees that

$$P(A \cap B \mid X) = P(A \mid X) P(B \mid X) \quad \text{for all } n,$$

as desired. □

1.6 Conditionally independent stochastic variables

In many cases we have σ -algebras generated by stochastic variables. We will make no distinction between the stochastic variable X and the σ -algebra $\mathbb{F}(X)$ generated by X , and we will write things like

$$X \perp\!\!\!\perp Y \mid Z \quad \text{instead of} \quad \mathbb{F}(X) \perp\!\!\!\perp \mathbb{F}(Y) \mid \mathbb{F}(Z)$$

without notification.

Example 1.27 X, Y and Z are truly independent. Then $X \perp\!\!\!\perp Y \mid Z$. ○

Example 1.28 Consider a normal distribution in three dimensions, where the one-dimensional marginals are standard normals,

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \beta \\ \rho & 1 & \beta \\ \beta & \beta & 1 \end{pmatrix}\right). \quad (1.14)$$

Here we have taken the two correlations involving Z to be identical, to keep the problem simple.

Independence of X and Y is controlled by the parameter ρ . If $\rho = 0$, if $\rho > 0$ they are positively correlated and if $\rho < 0$ they are negatively correlated.

The conditional distribution of X and Y given $Z = z$ is

$$\begin{aligned} & \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \beta \\ \beta \end{pmatrix}(z-0), \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \begin{pmatrix} \beta \\ \beta \end{pmatrix}(\beta\beta)\right) \\ & = \mathcal{N}\left(\begin{pmatrix} \beta z \\ \beta z \end{pmatrix}, \begin{pmatrix} 1-\beta^2 & \rho-\beta^2 \\ \rho-\beta^2 & 1-\beta^2 \end{pmatrix}\right). \end{aligned}$$

Note that the variance does not depend on the specific value of z . Hence we can conclude that X and Y are conditionally independent given Z if

$$\rho - \beta^2 = 0.$$

More precisely, the sign of $\rho - \beta^2$ controls the direction of the conditional correlation between X and Y given Z .

In figure 1.1 we have illustrated this phenomenon. In the (ρ, β) -plane we have found the domain which corresponds to legal covariance-matrices (all three eigenvalues being non-negative). It is seen that this domain is divided into three: a part which corresponds to negative marginal correlation **and** negative conditional correlation between X and Y . A part which corresponds to positive marginal correlation but negative conditional correlation. And a third part which corresponds to positive marginal and conditional correlation. If we did not employ the restriction that the two Z -correlations should be equal, we could of course have a fourth domain, corresponding to negative marginal but positive conditional correlation.

In this context, the message is that marginal correlations and conditional correlations are two very different things, and in particular that marginal independence and conditional independence are unrelated phenomena.

◦

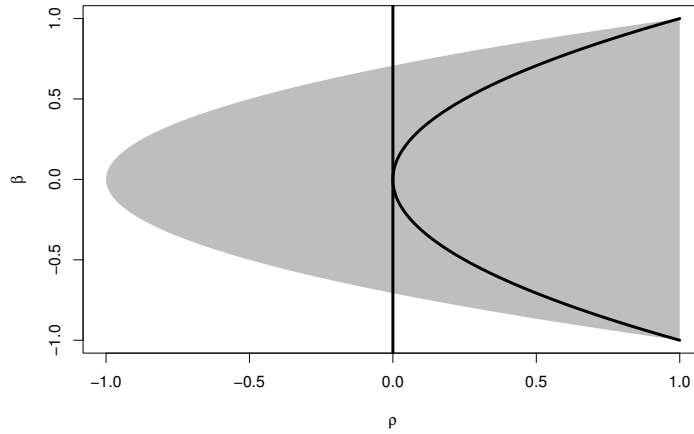


Figure 1.1: Marginal independence and conditional independence in normal distributions of type (1.14). The shaded area contains the (ρ, β) -values for which the normal distribution exists. The vertical line corresponds to marginal independence of X and Y (positive correlation is on the right hand side). The parabolic curve corresponds to conditional independence of X and Y given Z (positive conditional correlation is in the interior of the parabola). Note the domain where there is positive marginal correlation but negative conditional correlation.

Theorem 1.29 Let the conditional distributions of Y and Z given X be respectively $(\hat{P}_x)_{x \in \mathcal{X}}$ and $(\hat{Q}_x)_{x \in \mathcal{X}}$. Define

$$\hat{R}_x = \hat{P}_x \otimes \hat{Q}_x, \hat{L}_{x,z} = \hat{P}_x.$$

If $Y \perp\!\!\!\perp Z \mid X$, then $(\hat{R}_x)_{x \in \mathcal{X}}$ is the conditional distribution of (Y, Z) given X , and $(\hat{L}_{x,z})_{(x,z) \in \mathcal{X} \times \mathcal{Z}}$ is the conditional distribution of Y given (X, Z) .

PROOF: It is easily checked that $(\hat{R}_x)_{x \in \mathcal{X}}$ is a \mathcal{X} -kernel on $\mathcal{Y} \times \mathcal{Z}$. To check the integral condition, we write

$$\begin{aligned} \int_A \hat{R}_x(B \times C) dX(P)(x) &= \int_A \hat{P}_x(B) \hat{Q}_x(C) dX(P)(x) \\ &= \int_{(X \in A)} P(Y \in B \mid X) P(Z \in C \mid X) dP \\ &= \int_{(X \in A)} P(Y \in B, Z \in C \mid X) dP \\ &= P(X \in A, Y \in B, Z \in C) \end{aligned}$$

The second equality here is an application of the remarks on page 8, connecting the various concepts of conditioning. The third equality is an application of conditional independence, and the fourth is simply the definition of conditional probabilities. Standard arguments extend these computations from product sets $B \times C$ to general measurable subsets of $\mathcal{Y} \times \mathcal{Z}$. Hence $(\hat{R}_x)_{x \in \mathcal{X}}$ is the conditional distribution of (Y, Z) given X .

The proof of second half of the theorem proceeds in exactly the same way, utilizing the asymmetric formulation of conditional independence instead of the definition. \square

There are converses of both halves of this theorem. We choose to formulate them separately. They may come in handy under various circumstances, but in general we will go to quite some length in order to circumvent any use of them. The morale is that conditional independence should be established from first principles, not by computing conditional distributions. * * * * *

Theorem 1.30 *Suppose that the conditional distribution $(\hat{R}_x)_{x \in \mathcal{X}}$ of (Y, Z) given X has product structure of the form*

$$\hat{R}_x = \hat{P}_x \otimes \hat{Q}_x \quad \text{for all } x \in \mathcal{X}$$

for two families $(\hat{P}_x)_{x \in \mathcal{X}}$ and $(\hat{Q}_x)_{x \in \mathcal{X}}$ of probability measures on \mathcal{Y} and \mathcal{Z} respectively. Then both these families are Markov kernels, they are the conditional distributions of Y given X and of Z given X respectively, and it holds that $Y \perp\!\!\!\perp Z \mid X$.

PROOF: The first two statements are trivially checked. The statement about conditional independence follows from the product structure, when combined with the remarks on page 8, connecting the various concepts of conditioning. With this particular

choice of conditional probabilities the a.e.-quantifier in the definition of conditional probability is superfluous, as the factorization will hold deterministically. \square

Theorem 1.31 *Suppose that the conditional distribution $(\hat{L}_{x,z})_{(x,z) \in \mathcal{X} \times \mathcal{Z}}$ of Y given (X, Z) has the structure*

$$\hat{L}_{x,z} = \hat{P}_x \quad \text{for all } x \in \mathcal{X}, z \in \mathcal{Z}$$

for some family $(\hat{P}_x)_{x \in \mathcal{X}}$ of probability measures on \mathcal{Y} . Then this family is a Markov kernel, it is the conditional distribution of Y given X , and it holds that $Y \perp\!\!\!\perp Z \mid X$.

PROOF: The first two statements are trivially checked. The statement about conditional independence follows from the asymmetric characterization. \square

Application: The statement that $Y \perp\!\!\!\perp Z \mid X$ is a statement about the joint distribution of (X, Y, Z) . In the sense that three other variables X', Y' and Z' that happen to have the same joint distribution, will satisfy the same conditional independence relation as the original triple.

* * * * *