01.001 - Introduction to Probability and Statistics

Alexander Binder

Week 11: First lecture on 27th of March Second Lecture on 29th of March

[The following notes are compiled from various sources such as textbooks, lecture materials, Web resources and are shared for academic purposes only, intended for use by students registered for a specific course. In the interest of brevity. every source is not cited. The compiler of these notes gratefully acknowledges all such sources.]

Summary of key ideas from week's second lecture

In this session we will obtain all the tools to analyze probability distributions in n-dimensions.

Independence: Two variables

Definition:

Two random variables X and Y are said to be **independent** if for every pair of x and y values

$$p(x,y) = p_X(x) \cdot p_Y(y)$$

when X and Y are discrete, or

$$f(x,y) = f_X(x) \cdot f_Y(y)$$

when X and Y are continuous.

If the above relation is **not** satisfied for all (x,y), then X and Y are said to be **dependent**.

The above definition implies that two variables are independent if their joint pmf or pdf is the product of the two marginal pmf's or pdf's.

Independence - an overview for N variables

For 2 variables it is clear: one variable can be independent the other. No more design choices left.

For n variables are multiple types of independences possible! One can ask if all N variables are independent from each other. For i.i.d. random variables, all N variables are independent from each other.

However there are other types of independence: two sets of variables can be independent, even if all variables are not independent.

Independence: N variables (from each other)

Definition:

N random variables $X_1, \ldots, X_n$ are said to be **independent** (from each other) if for every n-tuple of $x_1, \ldots, x_n$ values

$$p(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \ldots \cdot p_{X_n}(x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$$

when X_i are discrete, or

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \ldots \cdot f_{X_n}(x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

when X_i are continuous.

If the above relation is **not** satisfied for all $(x_1, \ldots, x_n)$, then the X_i are said to be **dependent**.

If N variables are independent (from each other), then all subsets of K variables ($K < N$) are also independent from each other.

WARNING!

In general: For n variables, independence of all pairs as in the case of 2 variables does not imply independence of n variables.

If all pairs of two variables are independent, it does NOT imply that n variables would be independent. Consider this counterexample:

$$(X, Y, Z) = \begin{cases} (0,0,0) & \text{with probability } 0.25 \\ (0,1,1) & \text{with probability } 0.25 \\ (1,0,1) & \text{with probability } 0.25 \\ (1,1,0) & \text{with probability } 0.25 \end{cases}$$

Compute $P(X = 0), P(Y = 0), P(Z = 0)$ and $P(X = 0)P(Y = 0)P(Z = 0)$. Verify that $P(Y = 0, Z = 0)$ is a product of the one-dimensional marginals.

*Independence: **between two subsets of** N variables*

Suppose we want to model 5 variables $(X_1, X_2, X_3, X_4, X_5)$. Suppose we have a density for the first three $f_{X_1,X_2,X_3}(x_1, x_2, x_3)$ and a density for the last two $f_{X_4,X_5}(x_4,x_5)$. You can design a density for 5 variables from it by defining:

$$f(x_1,x_2,x_3,x_4,x_5) = f_{X_1,X_2,X_3}(x_1,x_2,x_3)\cdot f_{X_4,X_5}(x_4,x_5)$$

This is a density: products of nonnegative factors are nonnegative and it holds:

$$\begin{aligned}&\int_{X_1,X_2,X_3,X_4,X_5} f_{X_1,X_2,X_3}(x_1,x_2,x_3)\cdot f_{X_4,X_5}(x_4,x_5)dx_1dx_2dx_3dx_4dx_5\\ =&\int_{X_1,X_2,X_3} f_{X_1,X_2,X_3}(x_1,x_2,x_3)dx_1dx_2dx_3\cdot\int_{X_4,X_5} f_{X_4,X_5}(x_4,x_5)dx_4dx_5\\ =&1\cdot 1 = 1\end{aligned}$$

It does **not mean that** these 5 variables are *independent from each other*. If $f_{X_4,X_5}(x_4,x_5)\neq f_{X_4}(x_4)\cdot f_{X_5}(x_5)$, then independence of each other will not be true.

The important thing to see is: in this case the set of variables (X_1,X_2,X_3) is independent from the set (X_4,X_5).

Lets formalize this for two sets:

Lets denote X_{Set1}, X_{Set2} to be an **ordered** subset of $(X_1,\ldots,X_n)$. For example:

$n = 10$, $X_{Set1} = (X_2,X_6,X_9)$, $X_{Set2} = (X_1,X_3,X_4,X_5,X_7,X_8,X_{10})$ (in this case we have covered all 10 variables in Set1 and Set2)

or $X_{Set1} = (X_3,X_4,X_5)$, $X_{Set2} = (X_1,X_2,X_7)$ (in this case both sets cover only 6 out of 10 variables)

We define $X_{Set1\cup Set2}$ to be the ordered set union of these two sets.

Example:

$$\begin{aligned}X_{Set1} = (X_3,X_4,X_5,X_9), X_{Set2} &= (X_1,X_2,X_7)\\ X_{Set1\cup Set2} &= (X_1,X_2,X_3,X_4,X_5,X_7,X_9)\end{aligned}$$

Definition: Independence: between two subsets of N variables

The two sets X_{Set1}, X_{Set2} of variables are independent, if

$$f_{X_{Set1\cup Set2}}(x_{Set1\cup Set2}) = f_{X_{Set1}}(x_{Set1})\cdot f_{X_{Set2}}(x_{Set2})$$

Relationship between Independence Categories

Independence of N variables is stronger than independence between two sets: if N variables are independent, then all subsets are independent, too.

*Independence: **between more than two subsets of** N variables*

Of course a density $f(x_1,\ldots,x_n)$ can be a product of more than two factors. If it factors for example in three factors, then one would have three sets of variables $X_{set1}, X_{set2}, X_{set3}$ and

$$f(x_1, \ldots, x_n) = f_{X_{set1}}(x_{set1}) f_{X_{set2}}(x_{set2}) f_{X_{set3}}(x_{set3})$$

The important thing is to understand that independence in the case of N variables can have many forms. It is not as simple as independence between 2 variables. The simplest form of independence for N variables is that the N variables are independent from each other. This simplest form is also the strongest: It implies independence between all possible sets of variables (no matter 2,3 or any other number of sets).

Independence and Its Impact on Modeling Data

One can see in the AI class: independence assumptions make models simpler, having less parameters to be fitted, and easier to learn. However independence assumptions have two drawbacks:

1. **If two variables are assumed to be independent in your model, then you are unable to learn a relationship between these variables from data.** If the data has a relationship between two variables, then this relationship will be lost in the learnt model with such an independence assumption.

2. Independence assumptions need to be checked if they are appropriate for the data that you deal with.

Conditional Probability: Two Variables

The most important thing to understand that a conditional density for variable X_1 conditioned on variable X_2 must integrate to 1 over the variable X_1.

Definition: Let X_1 and X_2 be two continuous random variables with joint pdf $f(x_1, x_2)$ and marginal X_i pdf $f_i(x_i)$. Then for any X_1 value of x_1 for which $f_{X_1}(x_1) > 0$, the **conditional probability density function of** X_2 **given that** $X_1 = x_1$ is[1]:

$$f_{X_2|X_1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)} \qquad \text{where } x_2 \text{ is defined}$$

analogously, if $f_{X_2}(x_2) > 0$, then the **conditional probability density function of** X_1 **given that** $X_2 = x_2$ is

$$f_{X_1|X_2}(x_1|x_2) = \frac{f(x_1, x_2)}{f_{X_2}(x_2)} \qquad \text{where } x_1 \text{ is defined}$$

[1] Hopefully you are reminded of the conditional probability formula for $P(B|A)$

If X_1 and X_2 are discrete, replacing pdf's by pmf's in this definition gives the **conditional probability mass function of Y given that** $X_1 = x_1$.

Compare this to the definition of conditional probability:

$$\begin{aligned} P(A_1|A_2) &= \frac{P(A_1, A_2)}{P(A_2)} \\ f_{X_1|X_2}(x_1|x_2) &= \frac{f(x_1, x_2)}{f_{X_2}(x_2)} \end{aligned}$$

$$-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-$$

$$\begin{aligned} P(A_2|A_1) &= \frac{P(A_1, A_2)}{P(A_1)} \\ f_{X_2|X_1}(x_2|x_1) &= \frac{f(x_1, x_2)}{f_{X_1}(x_1)} \end{aligned}$$

These definitions are not a surprise and can be remembered in the following way. For a conditional probability $f_{X_1|X_2}(x_1|x_2)$ it must hold that: integrating it over x_1 (the variable that is not used for conditioning) must yield 1.

If we assume that the conditional density $f_{X_1|X_2}(x_1|x_2)$ is proportional to the joint density $f(x_1, x_2)$, that is $f_{X_1|X_2}(x_1|x_2) = cf(x_1, x_2)$, then we can derive the formula of conditional density by solving for c:

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) &= cf(x_1, x_2) \\ 1 &= \int_{x_1} f_{X_1|X_2}(x_1|x_2) dx_1 \\ \text{plug in the proportionality relationship } \Rightarrow 1 &= \int_{x_1} cf(x_1, x_2) dx_1 \\ \Rightarrow 1 &= cf_{X_2}(x_2) \\ \Rightarrow c &= \frac{1}{f_{X_2}(x_2)} \end{aligned}$$

The important point is to remember: the conditional density is proportional to the joint density, but must integrate (discrete case: sum up) to 1 over the variables that were not fixed by conditioning. Then one can immediately recover the formulas above.

As an important consequence/application of conditional densities:

$$P(X_1 \in A|X_2 = b) = \int_{x_1 \in A} f_{X_1|X_2}(x_1|b) dx_1 = \frac{1}{f_{X_2}(b)} \int_{x_1 \in A} f(x_1, b) dx_1$$

This links conditional densities to the corresponding conditional probabilities. Final remark: conditional densities are only defined where the marginal used to divide (e.g. $f_{X_2}(x_2)$) is non-zero.

Conditional Probability: n Variables

We have the same effect as for marginals. We can have many different conditional probabilities and densities.

For example, consider 3 variables X_1, X_2, X_3. Then we can have
$f_{X_1,X_3|X_2}$
$f_{X_1,X_2|X_3}$
$f_{X_2,X_3|X_1}$
$f_{X_1|X_2,X_3}$
$f_{X_2|X_1,X_3}$
$f_{X_3|X_1,X_2}$
but we can also have
$f_{X_3|X_1}$
How many conditional probabilities do exist?

We can write down intuitively

$$\begin{aligned}
f_{X_1,X_3|X_2}(x_1, x_3 \mid x_2) &= \frac{f(x_1, x_2, x_3)}{f_{X_2}(x_2)} \\
f_{X_1,X_2|X_3}(x_1, x_2 \mid x_3) &= \frac{f(x_1, x_2, x_3)}{f_{X_3}(x_3)} \\
f_{X_2|X_1,X_3}(x_2 \mid x_1, x_3) &= \frac{f(x_1, x_2, x_3)}{f_{X_1,X_3}(x_1, x_3)} \\
f_{X_3|X_1,X_2}(x_3 \mid x_1, x_2) &= \frac{f(x_1, x_2, x_3)}{f_{X_1,X_2}(x_1, x_2)} \\
f_{X_3|X_1}(x_3 \mid x_1) &= \frac{f_{X_1,X_3}(x_1, x_3)}{f_{X_1}(x_1)} \\
f_{X_1|X_2}(x_1 \mid x_2) &= \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}
\end{aligned}$$

Note that in the last two terms, the density in the enumerator (above the /) is a marginal density, too, as $X_1|X_2$ is only a subset of (X_1, X_2, X_3), so $f(x_1, x_2)$ is a marginal density

How to define all these conditional densities for n variables?

Lets denote X_{Set1}, X_{Set2} to be an **ordered** subset of $(X_1, \ldots, X_n)$.
We define $X_{Set1 \cup Set2}$ to be the ordered set union of these two sets.
Definition of conditional density:

$$f_{X_{Set1}|X_{Set2}}(x_{Set1 \cup Set2}) = \frac{f_{X_{Set1 \cup Set2}}(x_{Set1 \cup Set2})}{f_{X_{Set2}}(x_{Set2})}$$

If $X_{Set1 \cup Set2}$ covers all variables in $(X_1, \ldots, X_n)$, then we have

$$f_{X_{Set1}|X_{Set2}}(x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n)}{f_{X_{Set2}}(x_{Set2})}$$

Again, this formula can be easily derived and remembered by noting: $f_{X_{Set1}|X_{Set2}}(x_{Set1 \cup Set2}) = c f_{X_{Set1 \cup Set2}}(x_{Set1 \cup Set2})$, and the fact that integrating this over X_{Set1} must yield 1.

WARNING: A conditional probability integrates up to one if you integrate it over all its non-conditioning variables

$$\int_{X_{Set1}} f_{X_{Set1}|X_{Set2}}(x_{Set1 \cup Set2}) dX_{Set1} = 1$$

If you compute an integral that includes at least one of the variables in the set used for conditioning, then it does **NOT** integrate up to one anymore. It will be a funny number without any meaning about probabilities or densities.

$$\int_{X_{Set1}} \int_{X_s} f_{X_{Set1}|X_{Set2}}(x_{Set1 \cup Set2}) dX_{Set1} dX_s = ??\text{ Whatever! },\ X_s \in X_{Set2}$$
$$\int_{X_s} f_{X_{Set1}|X_{Set2}}(x_{Set1 \cup Set2}) dX_s = ??\text{ Whatever! },\ X_s \in X_{Set2}$$

Conditional Probabilities and their Use for Creating models I

Example 1: Modelling of Task Times

Suppose you own factories. you want to estimate probabilities how long it takes to finish a task in these factories. You decide to model task times by a density

$$f(T = y)$$

is the density for "the production task is finished after y time". You know then

$$P(\text{ the production task is finished after } x \text{ time}) = \int_0^x f(y) dy$$

You can take 100 workers in factory site 1, measure their finishing time, then you have a model. Suppose you want to use this model for your second factory site – factory site 2. Will it be a good model?

Maybe not. If you think longer, the time to finish the task might depend on the experience level of your employees.

Suppose you have a discrete variable measuring experience level $E \in \{1, \dots 5\}$. It might be true, that $f(T = y \mid E = 1) \neq f(T = y \mid E = 2)$ – the density to finish the task might depend on the experience level.

In factory site 1, you had a certain probability for experience levels of your employees $P(E = l \mid Site = 1)$. Your model $f(T = y \mid$

$Site = 1)$ is a mixture of densities $f(T = y \mid E = l)$ depending on the experience level of your workers. This model is weighted by the probabilities of experience at your Site 1:

$$f(T = y \mid Site = 1) = \sum_{l=1}^{5} f(T = y \mid E = l) P(E = l \mid Site = 1)$$

If the probability of experience levels is different at at your site 2: $P(E = l \mid Site = 1) \neq P(E = l \mid Site = 2)$, then your model from Site 1 will cause errors in estimating the finishing times for Site 2.

What you can do however is, estimate densities depending on the experience levels $f(T = y \mid E = l)$, and measure experience levels for each site $P(E = l \mid Site)$.

Then, you can use the densities depending on the experience levels obtained at Site 1 $f(T = y \mid E = l)$, to build a model for Site 2:

$$f(T = y \mid Site = 2) = \sum_{l=1}^{5} f(T = y \mid E = l) P(E = l \mid Site = 2)$$

This is an application of the conditional probability rule

$$P(X_1) = \sum_{x_2} P(X_1, X_2 = x_2) = \sum_{x_2} P(X_1 | X_2 = x_2) P(X_2 = x_2)$$

What can you learn from that: conditional probabilities can be used to adapt a model that was created for one cohort of measurements, to another cohort. Here the cohorts are your workers that work on the tasks. The assumption is: the model might depend on properties of your cohort, and the probability of these properties might be different between cohorts.

I made here one assumption: the density for task finishing time does depend on experience level, but there is no hidden influence of the Site itself (could be: e.g. different ages/generations of machinery used): $f(T = y \mid E = l, Site) = f(T = y \mid E = l)$

Conditional Probability and its relation to independence

If two sets of variables X_{Set1}, X_{Set2} are independent, then we have for all values $x_{Set1 \cup Set2}$:

$$f_{X_{Set1}|X_{Set2}}(x_{Set1} \mid x_{Set2}) = f_{X_{Set1}}(x_{Set1})$$
$$f_{X_{Set2}|X_{Set1}}(x_{Set2} \mid x_{Set1}) = f_{X_{Set2}}(x_{Set2})$$

Why does that hold?

by independence we have

$$\begin{aligned} f_{X_{Set1 \cup Set2}}(x_{Set1 \cup Set2}) &= f_{X_{Set1}}(x_{Set1}) \cdot f_{X_{Set2}}(x_{Set2}) \\ \Leftrightarrow \frac{f_{X_{Set1 \cup Set2}}(x_{Set1 \cup Set2})}{f_{X_{Set1}}(x_{Set1})} &= f_{X_{Set2}}(x_{Set2}) \\ \text{definition of conditional: } \Leftrightarrow f_{X_{Set2}|X_{Set1}}(x_{Set2} \mid x_{Set1}) &= f_{X_{Set2}}(x_{Set2}) \end{aligned}$$

What does that mean? Consider the first equation.

$$f_{X_{Set1}|X_{Set2}}(x_{Set1} \mid x_{Set2}) = f_{X_{Set1}}(x_{Set1})$$

That means:

1. the conditional density function with respect to X_{Set1} does not change when knowing the value x_{Set2} of the variables in the set X_{Set2} used for conditioning on.
2. That conditional density with respect to X_{Set1} is same as the marginal density over the variables X_{Set1}.

In short, knowing the value of the variable set used for conditioning, gives you no extra information about the conditional density.

Conditional Probabilities and their Use for Creating models II

Example 2: Modelling of awakeness levels after drink consumptions

Disclaimer: This example is a bit more involved. Try to understand the thinking. Try to understand the usage of the basic idea:

$P(X_1) = \sum_{x_3} P(X_1, X_3 = x_3) = \sum_{x_3} P(X_1|X_3 = x_3)P(X_3 = x_3)$. Note that this equation also holds, when we would apply a conditioning with the condition $X_2 = x_2$ to all probability terms in above equation: $P(A) \to P(A|X_2 = x_2)$, $P(A|B) \to P(A|B, X_2 = x_2)$. The example itself is not typical exam stuff, the basic idea could be.

You can measure levels of awakeness. You want to model the awakeness of a person after drinking a drink, as a function of the caffeine content in the drink. You want to make a study with 100 participants. You measure their reaction with a probstat integral task, after the drink.

Your goal is to model $P(\text{awake level after} = l \mid \text{caffeine})$.

When thinking about the model, you can come up with further dependencies. The effect might depend on how awake one was before gulping down your newest drink creation.

So maybe you want to model

$$P(\text{awake level after} = l \mid \text{caffeine}, \text{AWLVL before} = b)$$

The story continues here. The impact of caffeine (and sugar) depends on age of the drinker. Therefore, you may want to record the age of study participants and include age as model variable:

$$P(\text{awake level after} = l \mid \ \text{caffeine level}, \text{AWLVL before} = b, \text{age})$$

Suppose, you have run the study, and fitted a model based on your study data:

$$P(\text{awake level after} = l \mid \ \text{caffeine level}, \text{AWLVL before} = b, \text{age})$$

Now you want to present the model, but **without age dependency**. How to do that ?? You can consider a model which is an **average over the age distribution of Singaporean people**. You may want to calculate a model which is

- does not depend the age of the person who drinks.
- is an average over **the typical age distribution** of Singaporeans, rather than your study participants' age distribution.

So your final model of awakeness levels should have no age parameter and look like this:

$$P(\text{awake level after} = l \mid \text{AWLVL before} = b,\ \text{caffeine level})$$

Note: Your study participants age might be biased: Students who want FREE DRINKS, and pensioned people who want to kill time, might be overrepresented in your study.

Suppose, the age distribution of Singaporeans is given by

$$P(age = k)$$

How can you make use of that for your model ?

$$\begin{aligned} &P(\text{awake level after} = l \mid \text{AWLVL before} = b,\ \text{caffeine level}) \\ = \sum_k &P(\text{awake level after} = l \mid \text{AWLVL before} = b,\ \text{caffeine level}, \textbf{age} = k) \\ &\cdot P(\textbf{age} = k \mid \text{AWLVL before} = b,\ \text{caffeine level}) \end{aligned}$$

This uses your fitted model (line with the $\sum_k$) and a conditional probability related to $P(age = k)$.This formula can be simplified to (I explain this in a moment)

$$\begin{aligned} &P(\text{awake level after} = l \mid \text{AWLVL before} = b,\ \text{caffeine level}) \\ = &\sum_k P(\text{awake level after} = l \mid \text{AWLVL before} = b,\ \text{caffeine level}, \textbf{age} = k) P(\textbf{age} = k) \end{aligned}$$

This now uses the age distributions $P(age = k)$ of Singaporeans to compute an average.

You can adapt this model to be an average over a different age group than the average Singaporeans, e.g. students in a university.

Lets show how to arrive at this formula. We have

$$P(\text{awake level after} = l, \textbf{age=k}) = P(\text{awake level after} = l \mid \textbf{age} = k) P(\textbf{age} = k)$$

Summing above over all ages k, gives the marginal probability of awakeness level being at l: $P(\text{awake level after} = l)$

$$P(\text{awake level after} = l) = \sum_k P(\text{awake level after} = l \mid \textbf{age} = k) P(\textbf{age} = k)$$

This equation also holds, when conditioned on both sides (with awakeness level before and caffeine)

$$\begin{aligned} &P(\text{awake level after} = l \mid \text{AWLVL before} = b,\ \text{caffeine level}) \\ = &\sum_k P(\text{awake level after} = l \mid \textbf{age} = k, \text{AWLVL before} = b,\ \text{caffeine level}) \\ &\cdot P(\textbf{age} = k \mid \text{AWLVL before} = b,\ \text{caffeine level}) \end{aligned}$$

Now we make one assumption: the study participants age does not depend on the other entries (see above $P(X|Z) = P(X)$ in case of independence), namely: 1. awakeness level before drinking and 2. the caffeine level of the drink that they will be tested with.

$$P(\textbf{age} = k \mid \text{AWLVL before} = b,\ \text{caffeine level}) = P(\textbf{age} = k)$$

Plugging this in gives what we wanted to prove:

$$\begin{aligned} &P(\text{awake level after} = l \mid \text{AWLVL before} = b,\ \text{caffeine level}) \\ = &\sum_k P(\text{awake level after} = l \mid \text{AWLVL before} = b,\ \text{caffeine level}, \textbf{age} = k) P(\textbf{age} = k) \end{aligned}$$

Next problem: If you want to use this model later on people outside of the initial study and cannot measure their awakeness level before drinking, then you can plug in a guessed probability distribution $P(\text{AWLVL before} = b)$ over awakeness levels

$$\begin{aligned} &P(\text{awake level after} = l \mid\ \text{caffeine level}) \\ = &\sum_b P(\text{awake level after} = l \mid \text{AWLVL before} = b,\ \text{caffeine level}) P(\text{AWLVL before} = b) \end{aligned}$$

Now you are back to a simple model: awakeness level as a function of thr caffeine level.[2]

The moral here is:

- again an example for adapting a model to a cohort with different properties - here: age distribution (study vs general population)
- conditional probabilities can be used to model dependencies between variables that do not hold deterministically but with a certain probability - like awakeness level as a function of the caffeine content. $P(\text{awake level after} = l \mid \text{caffeine})$.
- Conditioning allows to build models that can be used depending on what information you have when using the model. Example: the awakeness before drinking is available during study time, but not later when the model is applied? No problem, then average it out with a probability $P(\text{AWLVL before} = b)$ – but take note of the footnote requiring a certain independence.

The coarse idea is to model

$$P(effect\ variables \mid cause\ variables)$$

You may have some hidden effect variables that vary

between 1. the study population on which you fit your model and 2. the population that you want to use your model on (worker experience across sites, age):

$$\begin{aligned} & P(effect\ variables \mid cause\ variables) \\ = \sum_{hidden\ variables} & P(effect\ variables \mid cause\ variables, hidden\ variables) P(hidden\ variables) \end{aligned}$$

I hope you can see from these examples the power of describing guessed dependencies between variables by conditional distributions. Its more than just a math formalism.

Conditional Probabilities and Bayes rule - out of exams

There is a way to make a meaningful integration over the conditioning variables by using Bayes rule and multiplying it with the right marginal or conditional probabilities. We know that by definition:

$$f_{X_1|X_2}(x_1, x_2) \cdot f_{X_2}(x_2) = f_{X_1,X_2}(x_1, x_2)$$

Integrating over this is meaningful again. Similarly:

$$f_{X_1,X_2|X_3,X_4}(x_1, x_2, x_3, x_4) \cdot f_{X_4|X_3}(x_3, x_4) = f_{X_1,X_2,X_4|X_3}(x_1, x_2, x_3, x_4)$$

[2] You have to use here again one assumption: the awakeness level before does not depend on the caffeine level that gets served: $P(\text{AWLVL before}) = P(\text{AWLVL before} \mid \text{caffeine level})$

The last conditional probability integrates over X_1, X_2, X_4 up to one. That looks confusing? Surprising? Multiplying a variable that is conditioned (X_4 in above example) with the corresponding marginal or conditional probability turns the variable into a non-conditioned variable.

We can write is as a general rule, with proof to make it more clear:

$$\begin{aligned} f_{X_{Set1}|X_{Set2\cup Set3}}(x_{Set1\cup Set2\cup Set3}) \cdot f_{X_{Set3}|X_{Set2}}(x_{Set2\cup Set3}) = \\ = f_{X_{Set1\cup Set3}|X_{Set2}}(x_{Set1\cup Set2\cup Set3}) \end{aligned}$$

The rule for getting a meaningful conditional probability is: you have to multiply the conditional probability $f_{X_{Set1}|X_{Set2\cup Set3}}$ with another conditional probability which is defined over $Set2 \cup Set3$ – not more and not less random variables.

Proof:

$$\begin{aligned} \text{by definition } f_{X_{Set1}|X_{Set2\cup Set3}} &= \frac{f_{X_{Set1\cup Set2\cup Set3}}}{f_{X_{Set2\cup Set3}}} \\ \text{by definition } f_{X_{Set3}|X_{Set2}} &= \frac{f_{X_{Set2\cup Set3}}}{f_{X_{Set2}}} \\ f_{X_{Set1}|X_{Set2\cup Set3}} \cdot f_{X_{Set3}|X_{Set2}} &= \frac{f_{X_{Set1\cup Set2\cup Set3}}}{f_{X_{Set2\cup Set3}}} \frac{f_{X_{Set2\cup Set3}}}{f_{X_{Set2}}} \\ &= \frac{f_{X_{Set1\cup Set2\cup Set3}}}{f_{X_{Set2}}} \\ \text{by definition } &= f_{X_{Set1\cup Set3}|X_{Set2}} \end{aligned}$$

Worked-out examples related to week's second lecture

1.

2. Using the joint probability density function of (X, Y) (given below) of the earlier example involving a start-up company operating a telephone helpline (X=fraction of time the telephone helpline is busy), as well as an online chat channel (Y=fraction of time the online chat channel is busy),

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{otherwsie} \end{cases}$$

   (a) What is the conditional pdf of Y given that $X = 0.8$?

   (b) What is the probability that the online chat channel is busy at most half of the time given that $X = 0.8$?

3. In your internship, you are asked to analyze a large hardware system. You find that the system consists of two components, which do not depend on one another. From past data you calculate the expected lifetimes of the components as 1000 hours and 1200 hours respectively. What is the probability that both component lifetimes are at least 1500 hours?

4. you have a conditional density $f_{X_1,X_2|X_3,X_4}(x_1, x_2, x_3, x_4)$.
   (a) For what set of variables this is a joint density ?
   (b) For what set of variables this is a marginal density ?
   (c) For what set of variables this is a conditional density?
   (d) Are you surprised that this question can be asked? :P

5. The moral of the following task is: is: Conditional independence allows to build more complex models, when the conditioning variable is marginalized out.

   Consider the following probability distribution over three variables X_1, X_2, X_3. Each of them takes only two values: $X_i \in \{0, 1\}$.

   We assume conditional independence when constructing $P(X_1, X_2 \mid X_3)$, that is:

$$P(X_1, X_2 \mid X_3) = P(X_1 \mid X_3) P(X_2 \mid X_3)$$

   The values for all probabilities are given as follows:

$P(X_1 \mid X_3) =$

	$x_3 = 0$	$x_3 = 1$
$x_1 = 0$	0.7	0.3
$x_1 = 1$	0.3	0.7

$P(X_2 \mid X_3) =$

	$x_3 = 0$	$x_3 = 1$
$x_2 = 0$	0.9	0.1
$x_2 = 1$	0.1	0.9

   We assume $P(X_3 = 0) = 0.5$.

   (a) Are X_1 and X_2 independent for their marginal probability $P(X_1, X_2)$?

6. Consider the following probability distribution over three variables X_1, X_2, X_3. Each of them takes only two values: $X_i \in \{0, 1\}$.

$P(x_1, x_2, x_3 = 0) =$

	$x_2 = 0$	$x_2 = 1$
$x_1 = 0$	0.07	0.03
$x_1 = 1$	0.105	0.045

$P(x_1, x_2, x_3 = 1) =$

	$x_2 = 0$	$x_2 = 1$
$x_1 = 0$	0.35	0.15
$x_1 = 1$	0.175	0.075

(a) What is the conditional probability $P_{X_1, X_3 | X_2}(x_1, x_2, x_3)$?

(b) Is the set X_1, X_3 independent of the set X_2 ?

7. Suppose we have a density defined by

$$f(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9) = f_1(x_1, x_2, x_3) \cdot f_2(x_4, x_5) \cdot f_3(x_6, x_7) \cdot f_4(x_8, x_9)$$

   (a) You can see an independence between how many sets of variables?

   (b) What are the *triples* of sets of variables that are independent? Do not count any marginals of above probabilities. List them all.

   (c) How many *pairs* of sets of variables exist that are independent? Do not count any marginals of above probabilities.

8. Consider the following probability distribution over two variables X_1, X_2. Each of them takes three values: $X_i \in \{0, 1, 2\}$.

$P(x_1, x_2) =$

	$x_2 = 0$	$x_2 = 1$	$x_2 = 2$
$x_1 = 0$	0.1	0.05	0.2
$x_1 = 1$	0.1	0.1	0.1
$x_1 = 2$	0.2	0.1	0.05

(a) What is the conditional probability $P_{X_2|X_1}(x_1, x_2)$?