

Bayesian probability theory - Lesson 3: Multivariate distributions - more Bayesics

Wolfgang von der Linden, Gerhard Dorn, Johanna Moser

Transcript

Welcome to the third unit of the course Bayesian probability theory. My name is Wolfgang von der Linden and I will enable you to help Captain Bayes and her crew to investigate probabilistic problems involving more than one random variable. We will come to discuss **joint**, **conditional** and **marginal probabilities** and **in turn correlations**. The aim of this unit is to discuss probabilistic problems from an enhanced perspective. We will learn how to include additional features into the description of an inference problem. That will allow us to update our state of knowledge based on additional information or experiments. The introduction of several random variables will seamlessly lead to the second fundamental rule in probability theory: **the product rule**. The question of Captain Bayes about the distribution of a sum of random variables will lead us to the **central limit theorem**. Finally we will identify the rules of probability theory as the general calculus of propositions.

In the previous lesson we have discussed characteristics like mean and variance of probability distributions, that rely on one variable. In this session we extend our language to problems where more than one random variable is concerned. We combine them into a **tuple** $P(X, Y)$ which often has the properties of a vector. Each of the components of the tuple or vector describes a different feature of the experiment or generally of the **inference problem**.

Joint probabilities

The strange die

- Number of pips on **Top**
- Color facing the gambler



$$T \in \{\square, \begin{smallmatrix} \square \\ \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square \\ \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square \\ \square & \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square \end{smallmatrix}\}$$

$$C \in \{\text{green}, \text{red}\}$$

$$S = T \times 2$$
$$S = T - 1$$

For example, the strange die has two features that are important for the final score, namely the number of pips on top and the color of the front side, the one facing the gambler. The front color is used as a modifier for the number of top pips, with green doubling the number and red reducing it by one.

Question 1. What is the score S of the feature pair $(\begin{smallmatrix} \square \\ \square \end{smallmatrix}, \text{green})$?

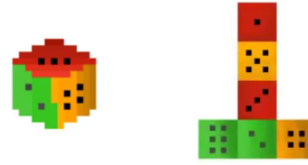
- a) 3 points
- b) 5 points
- c) 2 points
- d) 1 point
- e) 4 points

Now we can describe the outcomes of the strange die by the tuple of these two features: top pips and front color. For instance $\begin{smallmatrix} \square \\ \square \end{smallmatrix}$ and *red*. In the latter case, the score is 4 points. Note that the sample space consists of these feature pairs and they now *form the elementary events*. It is a matter of perspective! All in all there are 6 times 2, or 12 possible such feature pairs. Note also that the score is a random variable that ranges from 0 to 12, but not all integers in this range are realized. Since a die has four side faces, you could also use more colors and different rules to invent your own strange die game.

The strange die

- Number of pips on **T**op
- **C**olor facing the gambler
- **S**core of total points

$$S \in \{0, 1, 2, 3, 4, 5, 6, 8, 10, 12\}$$

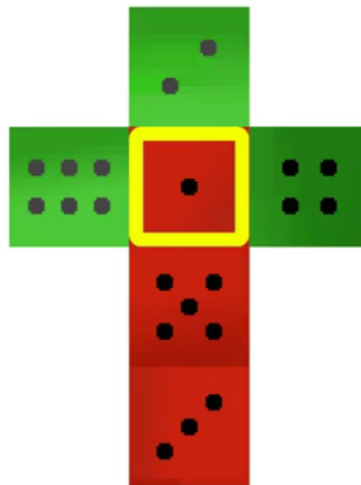


red: $T - 2$
green: $T * 2$
orange: $T +$ the hidden side on the table

Question 2. Look at the cube net of the three colored strange die. Which of the following statements are true?

- When having four pips on tom ($T = \text{☐}$), I get the highest score S when an orange face is pointing at me (i.e. the gambler).
- The combination (☐ , red) is not possible.
- It is more likely to have a green or orange side facing the gambler than having a red one.
- The lowest possible score is $S = -1$.


Next we want to assign probabilities to the scores: Laplace colored the faces of the die *green for an even* and *red for an odd number of pips*. The die Bernoulli and Laplace are using for the trinautic tournament is strange because the pips are arranged in an unusual way. We illustrate this by a net of the cube.



The yellow framed face represents the top of the die, showing the number ☐ in this

example. The four faces surrounding it are the possible candidates for the sides pointing towards the gambler. Note that there are in total 6 times 4, or 24 possible geometric realizations of the die experiment if you take the side face orientation also into account. In order to assign probabilities we count the geometric realizations that lead to a defined feature pair of top pips and front color. In this example with one pip (\square) on top we count 3 realizations for the green and 1 realization for the red modifier. Similarly we proceed with the other values for the top pips: \square , \square , \square , \square and \square and obtain the following table.

Assigning probabilities



Principle of indifference

Geometric realizations		
$N(T, C)$	red	green
\square	1	3
\square	2	2
\square	1	3
\square	3	1
\square	2	2
\square	3	1







According to the **principle of indifference** all geometric realizations are equally probable and therefore the classical assignment of probability for a pair (top pips, front color) is given by the *number of favorable geometric realizations* - which are the numbers in the table - *divided by total number of geometric realizations*, which is 24. The probability for the pair \square and red is therefore 3 divided by 24 resulting in 12.5 %.

Question 3. What is the probability of the feature pair (\square , green)?

- a) 25%
- b) 1/24
- c) 8.3%
- d) 12.5%

That leads to the following table of **joint probabilities**.

Assigning probabilities

Joint probabilities			(S, C)	
	red	green	red	green
	4.2%	12.5%	0	4.2%
	8.3%	8.3%	1	8.3%
	4.2%	12.5%	2	4.2%
	12.5%	4.2%	3	12.5%
	8.3%	8.3%	4	8.3%
	12.5%	4.2%	5	12.5%
			6	12.5%
			8	4.2%
			10	8.3%
			12	4.2%

From the table of joint probabilities for the feature pairs we can easily construct another representation, namely the pair “score and front color”. If the font color is green, the score is obtained by multiplying the top pips by 2. Likewise, if the front color is red, the score is obtained by reducing the top pips by one. These tables represent different joint probabilities. Needless to say that joint probabilities can also be defined for more than two propositions or variables $P(A, B, C, \dots)$.

Conditional probabilities

Now that we know the joint probabilities, there are different aspects of the inference problem that we can be address. In case of the strange die, we could ask the question: “*What is the probability for a particular score, for example 6 points, given the front side is green?*”. Such an object is called **conditional probability**, since one feature - here the color - is *fixed*. We use a vertical line to separate the random variables or events of interest from the fixed conditions. In terms of propositions the notation is shorter, $P(A|B)$ describes the probability for A given B is true for sure. When we examine the probabilities for all values of the random variable score, given that the color is green, we obtain the **conditional probability distribution**.

Question 4. Fill in the gaps correctly (“a variable”, “the conditional bar”, “a conditional probability”, “a conditional probability distribution”, “an outcome”)

The term $P(X_i|\text{red})$ is a whereas $P(X|\text{red})$ is since the first argument is in the first place and in the second.

Note that you always find an event, a proposition, or an outcome behind

We want to illustrate this concept on the strange die and compute the conditional probability $P(S|\text{green})$. Note that the sample space has shrunk as only the elements with green front face are considered. The elements that we consider are the same as those in the joint probability. Therefore the probability ratios between them have to be the same. Consequently, the *conditional probability is proportional to the joint one* in other words, only the **normalization** Z has changed which is now given by *summing the joint probability* of the score values resulting in the depicted table.

Question 5. What is the value of the normalization Z in this case:

$$Z = \sum_i P(S_i|C = g)$$

- a) 0.5
- b) 1/24
- c) 25%
- d) 12.5%

Similarly, we obtain the conditional probabilities for the score given the front color is red.

Conditional probability distribution

	$P(S \text{g})$		$P(S \text{r})$
		0	8,3%
		1	16,7%
	25,0%	2	8,3%
		3	25,0%
	16,7%	4	16,7%
		5	25,0%
	25,0%	6	
	8,3%	8	
	16,7%	10	
	8,3%	12	

$$P(S|\text{g}) = \frac{P(S, C = \text{g})}{\sum_i P(S_i, C = \text{g})}$$

Now we change the perspective into the opposite. Instead of fixing a feature we simply don't care about it. For example we might want to know the probability for the score

being 4, say, irrespective of the front color or we are interested in the probability that the front color is green, irrespective of the number of scores. We use again the table of joint probabilities and consider the joint feature pair (score and color) as elementary event. For *fixed score* S , the events with the same score but different colors are exclusive and therefore the **simplified sum rule** $P(S_1 \text{ or } S_2) = P(S_1) + P(S_2)$ applies. On the other hand, from **boolean algebra** we know that the argument of the left hand side turns out to be the score. In our example, we obtain the probability of a certain score just by *summing the rows* and noting the resulting probability on the right margin. Similarly, we obtain the marginal probabilities for the colors by *summing the columns* of the table. Thereby, we find $P(\text{red})$ is equal to $P(\text{green})$ is equal to 0.5.

Marginal probability distribution

$P(S, C)$	red	green	$P(S)$
0	4.2%		4.2%
1	8.3%		8.3%
2	4.2%	12.5%	16.7%
3	12.5%		12.5%
4	8.3%	8.3%	16.7%
5	12.5%		12.5%
6		12.5%	12.5%
8		4.2%	4.2%
10		8.3%	8.3%
12		4.2%	4.2%
	50.0%	50.0%	

Question 6. Do you see an easy explanation for this result?

- a) The dice has 3 green and 3 red faces so the probability for green, as well as the probability for red, is 0.5.
- b) The cube net of the problem was chosen in such a way that the probability for green adds up to 50%. The probability for red follows from the sum rule.
- c) The four side faces can have two different colors. In total, we find green faces and two red ones on average.

Marginal probability

This brings us to the **marginalization rule** in its most general form: We consider any set of **complete** and **exclusive** propositions B_i , which means they form a *non-*

overlapping partition of the sample space. Then we obtain the generalized marginalization rule.

$$P(A) = \sum_i P(A, B_i)$$

The importance of this rule cannot be overestimated. Again, a generalization to more than two variables is given by multiple sums.

Question 7. Let's try to apply the marginalization rule. Which example applies it correctly?

- a) $P(g) = P(g, \text{even number of pips}) + P(g, \text{uneven number of pips})$
- b) $P(\boxplus) = P(\boxplus, g) + P(\boxplus, r)$
- c) $P(g) = P(g, > \boxplus) + P(g, \leq \boxplus)$
- d) $P(\boxtimes) = P(\boxtimes | g) + P(\boxtimes | r)$
- e) $P(g) = P(g, \text{odd number of pips}) + P(g, \geq \boxplus)$
- f) $P(g) = P(g, \boxplus) + P(g, \neg \boxplus)$

The product rule

Now we come to the second fundamental rule of probability theory, the **product rule**. We combine the formula we obtained for the conditional probability and the formula for the marginal probability and obtain $P(S_j|g) = \frac{P(S_j, g)}{P(g)}$. Written differently, we obtain the famous product rule, which has the general form.

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

Note that A and B do not have to be elementary events or be assigned to random variables but can be *arbitrary propositions*. The product rule enables us to change the perspective of an inference problem, since it is the formula that links **assumptions** - propositions behind the conditional bar - with **open questions** - propositions before the conditional bar. We will see in unit 6 how this product rule enables us to swap assumption and open question to tackle inverse problems. Note, that for *uncorrelated* propositions A and B the product rule simplifies to $P(A, B) = P(A)P(B)$, like the probability for rain and rolling six pips on a die is equal to the product of the two probabilities.

Question 8. Let's apply the product rule! Which of the following equations describe the application of the product rule?

- a) $P(\text{f}, g) = P(\text{f} | g) \cdot P(\text{f})$
- b) $P(\text{f}) = P(\text{f} | > \text{f}) \cdot P(> \text{f})$
- c) $P(r, \text{f}) = \frac{P(r, \text{f})}{P(\text{f})}$
- d) $P(\text{f} | r) \cdot P(r) = P(r | \text{f}) \cdot P(\text{f})$
- e) $P(\text{f}, g) = P(g | \text{f}) \cdot P(g)$

It was a long and winding road in the history of science until this general form could be proven rigorously. It is applicable to the most general definition of probability as a *measure for the truth of a proposition*. Then probability theory can be considered as calculus of propositions and is a seamless generalization of Boolean algebra to partial truth. The only axioms this calculus is based on are **consistency**, which means, one gets the same result no matter along which path a calculation is performed, and **continuity** with respect to small changes of parameters. If there are more than two propositions, there are several choices of propositions for which we can compute the probabilities conditional on the others - corresponding to different perspectives of the problem.

The conditional complex

Note: *There is no such thing as an unconditional probability, it always relies on assumptions.* Like the die is perfectly symmetric, the throw guarantees a chaotic trajectory, the players are not cheating, the die has six faces. It will never land on an edge or corner and so on. We combine all these assumptions in the so called **background information**. We do not always write the background information explicitly, but keep in mind that it *has to be fixed and cannot be changed* during the calculations to remain consistent since you cannot expect the same outcome of an experiment when you alter its setup. If some assumptions may be altered (new information, looking at different aspects) then we will write them explicitly behind the conditional bar. As a matter of fact, the background information is nothing but a proposition, and one is immediately prompted to move it in front of the conditional bar. That allows to check our assumptions, like what is the probability that a coin is fair, if it lands heads up ten times in a row. $P(10 \text{ heads in a row} | \text{coin is fair}) \Rightarrow P(\text{coin is fair} | 10 \text{ heads in a row}) = ?$ But you'll have to be patient - we will cover this topic only in a later unit.

So far, we have discussed three different aspects of probability distributions, joint, marginal and the conditional probability. Now we want to turn to the computation of **mean values** for these cases. In the adventure of Captain Bayes the question was raised about the average score obtained in the strange die game. One way to compute it, is to define the function $S(T, C)$ that evaluates the score from the number of top pips and the front color. Then the average score is given by the *sum of this function times the joint probability*.

$$\langle S(T, C) \rangle = \sum_{ij} S(T_i, C_j) P(T_i, C_j)$$

$P(A, B)$				
$A \backslash B$	5	8	$P(A)$	
2	0.2	0.30	0.5	
3	0.15	0.06	0.21	
5	0.05	0.24	0.29	
$P(B)$	0.4	0.6		

$S = A^2 + B$			
$A \backslash B$	5	8	
2	9	12	
3	14	17	
5	30	33	

Question 9. Calculate the mean value of the score.
 The mean value of the score $S = A^2 + B$ given by the probability distribution $P(A, B)$ in the two tables above is $\langle S \rangle = \dots\dots\dots$

Alternatively, we could use the marginal probability mass function for the score $\sum_i S_i P(S_i)$, which we computed already, resulting in 4.5. Finally, we can address Laplace’s question concerning the average score he gets, assuming he always sees the green front face. In this case the answer is given by the mean value of the conditional probability distribution $\langle S \rangle_{|g} = \sum_i S_i P(S_i|g)$ resulting in six points on average. Quite similarly we obtain the averaged conditional on red to be three.

$P(A, B)$				
$A \backslash B$	5	8	$P(A)$	
2	0.2	0.30	0.5	
3	0.15	0.06	0.21	
5	0.05	0.24	0.29	
$P(B)$	0.4	0.6		

Question 10. Calculate the requested values using the table above.

The normalization needed to get the conditional probability distribution $P(A|B = 5)$ is $Z_{B=5} = \dots\dots\dots$

The conditional mean value $\langle A \rangle_{B=5} = \dots\dots\dots$

The conditional mean value $\langle B \rangle_{A=2} = \dots\dots\dots$

Often times, the conditional probabilities and their means are more easily accessible than the joint probabilities, in which case we can employ a useful expression for the total mean.

$$\langle S \rangle = \sum_i \langle S \rangle_{C_i} P(C_i)$$

Applied to the strange die, we then find the result in agreement with the previous result. All of the above considerations can readily be generalized to other problems with two random variables.

Characterizing joint distributions

$$\text{joint:} \quad \langle f(A, B) \rangle = \sum_{ij} f(A_i, B_j) \cdot P(A_i, B_j)$$

$$\text{conditional:} \quad \langle f(A) \rangle_{|B_j} = \sum_i f(A_i, B_j) \cdot P(A_i|B_j)$$

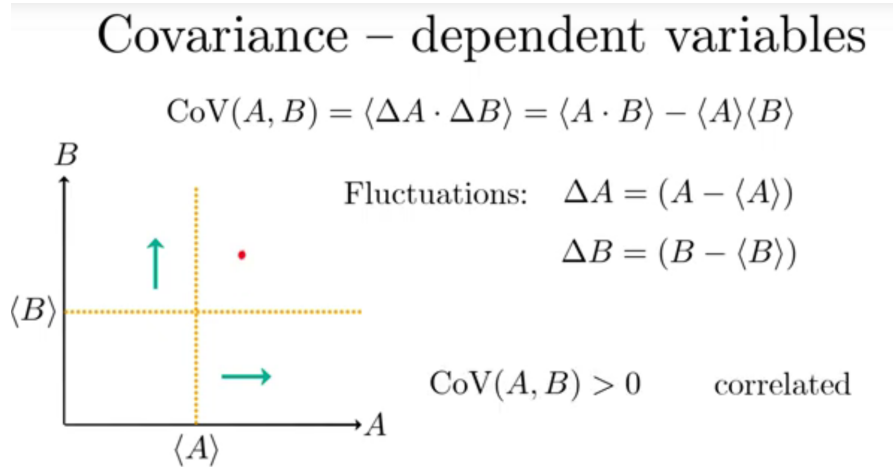
$$\text{marginal:} \quad \langle f(A) \rangle = \sum_j \langle f(A) \rangle_{|B_j} \cdot P(B_j)$$

The **variance** is in principle just a special case of the previous discussion. But due to its importance we will give the result explicitly. We compute it for the score of the strange

die for which we need $\langle S^2 \rangle - \langle S \rangle^2 = 8.9$.

Covariance

Another important object in the analysis of random variables or generally in inference problems is the **covariance** which allows to identify correlations between different features A and B .



It shows whether on average an *increase of A (above its mean) is associated with an increase of B (above its mean) or whether it is associated with a decrease of B (below its mean)*. For example, we could test whether, in a population, a person's weight is correlated or anti-correlated with his or her height. If *fluctuations of one feature are independent of fluctuations of the other feature* then the covariance is **zero**. That implies according to its definition that the *mean of a product is the product of the means* and, therefore, the joint probability factorizes into the product of the marginal probabilities.

$$\boxed{\text{CoV}(A, B) = 0 \Rightarrow \langle AB \rangle = \langle A \rangle \langle B \rangle \Rightarrow P(A, B) = P(A)P(B)}$$

Now we can apply the covariance to examine Pascal's notes on the sailing deviations. Since we have measured data, we can estimate the joint probabilities by the relative frequencies. As discussed along with the sample variance we have to correct for a bias due the use of the sample mean value.

Covariance – dependent variables

$$\text{CoV}(A, B) \approx \frac{\mathcal{N}}{\mathcal{N} - 1} (\overline{A \cdot B} - \overline{A} \cdot \overline{B})$$

$$\text{CoV}(A, B) \approx \frac{1000}{999} (0.082 - 2.21 \cdot (-0.016)) \approx 0.13$$

$P(A, B)$	-1	0	1	$P(A)$
0	1.7%	25.5%	1.1%	28.3%
$\frac{\pi}{2}$	6.2%	17.8%	1.0%	25.0%
π	1.1%	21.9%	1.2%	24.2%
$\frac{3\pi}{2}$	1.1%	16.9%	4.5%	22.5%
$P(B)$	10.1%	82.1%	7.8%	

$A \cdot B$	-1	0	1
0	0	0	0
$\frac{\pi}{2}$	$\frac{\pi}{2}$	0	$\frac{\pi}{2}$
π	π	0	π
$\frac{3\pi}{2}$	$\frac{3\pi}{2}$	0	$\frac{3\pi}{2}$

The covariance of the sailing direction and the sailing deviation of Pascal's notes is 0.13, indicating a correlation between sailing direction and deviation, which can be considered as a hint for the presence of an ocean drift. All what we have discussed so far can readily be generalized to more than 2 features, variables or rather propositions.

The central limit theorem

In the adventure Bayes was wondering how the sum of the strange-die-scores of her crew might be distributed. This brings us to the **central limit theorem (CLT)**. The scores of the crew-members are actually random variables. Let's be flexible as to the size of the crew and just denote it by N . The individual random variables have special properties: They are **Independently and Identically Distributed** which is abbreviated by "i.i.d.". What does that mean?

Identical means that *all random variables have the same probability mass function*, characterized by its mean and variance. **Independent** means uncorrelated and implies that the *covariance between different variables is zero*. Now we turn to the averaged total score, as that is what we are really interested in. We call it **arithmetic mean**, although the objects are random variables.

$$T = \frac{1}{N} \sum_i S^{(i)}$$

It should be obvious that T is also a random variable. For the mean of T we obtain from the previous slide and the linearity property that it is identical to the mean of the

distribution of the individual terms.

$$\langle S^{(i)} \rangle = \mu \Rightarrow \langle T \rangle = \mu$$

The variance of T is given by variance of the distribution of the individual terms divided by N .

$$\text{Var}(T) = \frac{\sigma^2}{N}$$

Even more thrilling is what Captain Bayes wanted to know, namely the detailed form of the distribution of T . The central limit theorem states that the probability distribution of T , being the mean of N Independently and Identically Distributed random variables approaches the **Gaussian distribution** for N going to infinity with mean and variance given before.

$$\lim_{N \rightarrow \infty} T \approx \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{N}}\right)$$

So the distribution of T universally approaches a distribution called "**Gaussian**" irrespective of the shape of the probability mass function of the individual terms.

What is a Gaussian? The Gaussian distribution is one of the most important and ubiquitous distributions in probability theory. It is given by the following formula and parameterized by mean and variance.

$$\mathcal{N}(X|\mu, \frac{\sigma}{\sqrt{N}}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$

It has the shape of a bell and is thus often called **bell curve** or **Normal distribution**. The limit N going to infinity itself is actually not very relevant for most real world applications, but already for moderately large N ($N \geq 30$) the distribution is confusingly similar to a Gaussian. So Captain Bayes will find that the distribution of the averaged score of her crew in the strange die game will look like a Gaussian.

This concludes the third unit. Investigate the sums of random variables and the central limit theorem using the interactive simulations, feel free to ask questions in the forum and feel encouraged to test your knowledge in the quiz!



ITPCP, TU Graz

<https://creativecommons.org/licenses/by/4.0/legalcode>