

INFO-F-422: Statistical foundations of machine learning

Foundations of probability

Gianluca Bontempi

Machine Learning Group
Computer Science Department
`mlg.ulb.ac.be`

Why probability and machine learning?

Probability is a convenient way of

1. representing variability and uncertainty,
2. modelling dependencies in a multivariate setting,
3. representing data as realisations of a random experiment (probabilistic *data generating process*).

...probability theory is an exact discipline developed logically from clearly defined axioms, and when it is applied to real problems it works.. (A. Papoulis)

Warning: probability can be highly counterintuitive: human intuition often does not help...

Random experiment in practice

- ▶ Modelling a random phenomenon requires the definition of a set of **variables** of interest.
- ▶ A variable is any property or descriptor (categorical or numerical) of a phenomenon that can take multiple values (also called *outcomes*).
- ▶ We will **assume** that the variability of measurements can be **represented** by the probability formalism.
- ▶ A random experiment is a **compact (and approximate)** way of modeling the disparate set of causes which led to variability.

Weather model

Consider a probabilistic model of the day's weather based on three **categorical random variables** where

1. the first represents the **sky condition** and takes value in the finite set {CLEAR, CLOUDY}.
2. the second represents the **barometer trend** and takes value in the finite set {RISING, FALLING},
3. the third represents the **humidity** and takes value in {DRY, WET}.

Weather example: sample space

The associated sample space¹ is made of 8 *outcomes*

	Sky	Barometer	Humidity
ω_1	CLEAR	RISING	DRY
ω_2	CLEAR	RISING	WET
ω_3	CLEAR	FALLING	DRY
ω_4	CLEAR	FALLING	WET
ω_5	CLOUDY	RISING	DRY
ω_6	CLOUDY	RISING	WET
ω_7	CLOUDY	FALLING	DRY
ω_8	CLOUDY	FALLING	WET

Observing the weather boils down to running a random experiment, which returns one of those outcomes.

¹A sample space can be finite, countably infinite or uncountably infinite.

- ▶ Event is:
 - ▶ A subset of experimental outcomes is called *event*.
 - ▶ Any declarative statement made using variables is an event.
- ▶ Example: *the sky is clear or the weather is dry and the barometer is falling*.
- ▶ An **event** $\mathcal{E}$ is the assignment of a set of values to a variable or a set of variables.
- ▶ Since events are sets we can use the set theory terminology (universal set, intersection, union, complement, difference, disjoint, partition).

Events and set theory

Since events $\mathcal{E}$ are subsets, we can apply to them the terminology of the set theory:

- ▶ $\mathcal{E}^c = \{\omega \in \Omega : \omega \notin \mathcal{E}\}$ denotes the *complement* of $\mathcal{E}$.
- ▶ $\mathcal{E}_1 \cup \mathcal{E}_2 = \mathcal{E}_1 + \mathcal{E}_2 = \{\omega \in \Omega : \omega \in \mathcal{E}_1 \text{ OR } \omega \in \mathcal{E}_2\}$ refers to the event that occurs when *at least one* of $\mathcal{E}_1$ and $\mathcal{E}_2$ occurs.
- ▶ $\mathcal{E}_1 \cap \mathcal{E}_2 = \{\omega \in \Omega : \omega \in \mathcal{E}_1 \text{ AND } \omega \in \mathcal{E}_2\}$ refers to the event that occurs when *both* $\mathcal{E}_1$ and $\mathcal{E}_2$ occur.
- ▶ two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are *mutually exclusive* or *disjoint* if

$$\mathcal{E}_1 \cap \mathcal{E}_2 = \emptyset$$

that is each time that $\mathcal{E}_1$ occurs, $\mathcal{E}_2$ does not occur.

- ▶ a *partition* of Ω is a set of disjoint sets $\mathcal{E}_j$, $j = 1, \dots, n$ such that

$$\cup_{j=1}^n \mathcal{E}_j = \Omega$$

Axiomatic definition of probability

- ▶ Axiomatic approach to probability assigns to each event $\mathcal{E}$ a real number $\text{Prob}\{\mathcal{E}\}$ (*probability of the event $\mathcal{E}$*).
- ▶ Any function $\text{Prob}\{\cdot\}$ or measure that satisfies
 1. $\text{Prob}\{\mathcal{E}\} \geq 0$ for any $\mathcal{E}$.
 2. $\text{Prob}\{\Omega\} = 1$
 3. $\text{Prob}\{\mathcal{E}_1 \cup \mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1\} + \text{Prob}\{\mathcal{E}_2\} - \text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\}$is a *probability distribution*.
- ▶ Axioms do not say anything about how probability should be interpreted.
- ▶ It follows that

$$\text{Prob}\{\mathcal{E}\} = \sum_{\omega \in \mathcal{E}} \text{Prob}\{\omega\}$$

and

$$\text{Prob}\{\mathcal{E}_1 \cup \mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1\} + \text{Prob}\{\mathcal{E}_2\}$$

if $\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = 0$ that is $\mathcal{E}_1$ and $\mathcal{E}_2$ are mutually exclusive (or disjoint).

Weather example: probability distribution

	z_1 (Sky)	z_2 (Barom.)	z_3 (Humid)	$P(z_1 = z_1, z_2 = z_2, z_3 = z_3)$
ω_1	CLEAR	RISING	DRY	0.4
ω_2	CLEAR	RISING	WET	0.07
ω_3	CLEAR	FALLING	DRY	0.08
ω_4	CLEAR	FALLING	WET	0.10
ω_5	CLOUDY	RISING	DRY	0.09
ω_6	CLOUDY	RISING	WET	0.11
ω_7	CLOUDY	FALLING	DRY	0.03
ω_8	CLOUDY	FALLING	WET	0.12

NB: this is a **probabilistic model** and NOT a dataset...

Weather example: a dataset

<i>Time</i>	z_1 (Sky)	z_2 (Barometer)	z_3 (Humidity)
Day 1	CLOUDY	RISING	WET
Day 2	CLEAR	RISING	DRY
Day 3	CLEAR	RISING	DRY
Day 4	CLOUDY	FALLING	WET
Day 5	CLEAR	RISING	WET
Day 6	CLEAR	FALLING	DRY
Day 7	CLEAR	RISING	DRY
...	...	...	...
...	CLEAR	FALLING	WET
...	CLOUDY	RISING	DRY
...	CLEAR	RISING	DRY
...	CLEAR	RISING	DRY
...	CLOUDY	FALLING	DRY
...	CLEAR	RISING	DRY
...	...	...	...

Definition (Independent events)

Two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are *independent* if

$$\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1, \mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1\} \text{Prob}\{\mathcal{E}_2\}$$

and we write $\mathcal{E}_1 \perp\!\!\!\perp \mathcal{E}_2$.

Examples:

- ▶ event $\mathcal{E}_1$: your professor is Italian, event $\mathcal{E}_2$: bad weather in Brussels
- ▶ event $\mathcal{E}_1$: commute time ≤ 10 minutes, event $\mathcal{E}_2$: bad weather in Buenos Aires

Homework

1. Are two mutually exclusive events (with non null probability) independent?
2. Are two identical events (with probability smaller than one) independent?
3. If two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent, what about their complements?
Hint: $\text{Prob}\{\mathcal{E}_1^c \cap \mathcal{E}_2^c\} = 1 - \text{Prob}\{\mathcal{E}_1 \cup \mathcal{E}_2\}$
4. Find a pair of independent events in the Weather example.

1. Are two mutually exclusive events (with non null probability) independent?

► No. If $\mathcal{E}_1 \cap \mathcal{E}_2 = \emptyset$ then

$$\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \text{Prob}\{\emptyset\} = 0 \neq \text{Prob}\{\mathcal{E}_1\} \text{Prob}\{\mathcal{E}_2\}$$

2. Are two identical events (with probability smaller than one) independent?

► No. If $\mathcal{E}_1 = \mathcal{E}_2$ then

$$\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \text{Prob}\{\mathcal{E}_1\} \neq \text{Prob}\{\mathcal{E}_1\} \text{Prob}\{\mathcal{E}_2\}$$

3. If two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent, what about their complements?

► They are independent as well, since

$$\begin{aligned} \text{Prob}\{\mathcal{E}_1^c \cap \mathcal{E}_2^c\} &= 1 - \text{Prob}\{\mathcal{E}_1 \cup \mathcal{E}_2\} = \\ &= 1 - \text{Prob}\{\mathcal{E}_1\} - \text{Prob}\{\mathcal{E}_2\} + \text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \\ &= 1 - \text{Prob}\{\mathcal{E}_1\} - \text{Prob}\{\mathcal{E}_2\} + \text{Prob}\{\mathcal{E}_1\} \text{Prob}\{\mathcal{E}_2\} = \\ &= (1 - \text{Prob}\{\mathcal{E}_1\})(1 - \text{Prob}\{\mathcal{E}_2\}) = \text{Prob}\{\mathcal{E}_1^c\} \text{Prob}\{\mathcal{E}_2^c\} \end{aligned}$$

5. Find a pair of independent events in the Weather example.

- ▶ One pair is $\mathcal{E}_1 = \{\omega_1, \omega_4\}$ and $\mathcal{E}_2 = \{\omega_2, \omega_4, \omega_7\}$, since $\mathcal{E}_1 \cap \mathcal{E}_2 = \omega_4$, $\text{Prob}\{\mathcal{E}_1\} = 0.5$, $\text{Prob}\{\mathcal{E}_2\} = 0.2$ and $\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \text{Prob}\{\omega_4\} = 0.1 = \text{Prob}\{\mathcal{E}_1\} \text{Prob}\{\mathcal{E}_2\} = 0.5 * 0.2$

Independent variables

- ▶ Let $\mathbf{x}$ and $\mathbf{y}$ be two random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$.
- ▶ The two variables $\mathbf{x}$ and $\mathbf{y}$ are defined to be *independent* if the joint probability

$$\text{Prob}\{\mathbf{x} = x, \mathbf{y} = y\} = \text{Prob}\{\mathbf{x} = x\} \text{Prob}\{\mathbf{y} = y\}, \quad \forall x \in \mathcal{X}, \quad y \in \mathcal{Y}$$

- ▶ NOT A BENE: **for all** values x of the r.v. $\mathbf{x}$ and all value y of the r.v. $\mathbf{y}$!
- ▶ Independency means that the observation of a certain value of one variable will NOT affect the probability of observing a certain value of the other.
- ▶ Independence is a **symmetric** relation.

Exercise: independent events vs independent variables

Consider two categorical variables.

z_1	z_2	$P(\mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2)$
Neg	Neg	0.1
Neg	Pos	0.2
Zero	Neg	0.2
Zero	Pos	0.1
Pos	Neg	0.2
Pos	Pos	0.2

Are there two independent events (with probability larger than 0 and smaller than 1)?

Yes, e.g. $\mathcal{E}_1 = \{\omega_1, \omega_4\}$ and $\mathcal{E}_2 = \{\omega_1, \omega_2, \omega_3\}$

$$\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \text{Prob}\{\omega_1\} = 0.1 = \text{Prob}\{\mathcal{E}_1\} \text{Prob}\{\mathcal{E}_2\} = 0.2 * 0.5$$

Are the two variables independent? No

$$\text{Prob}\{\mathbf{z}_1 = N, \mathbf{z}_2 = N\} = 0.1 \neq \text{Prob}\{\mathbf{z}_1 = N\} \text{Prob}\{\mathbf{z}_2 = N\} = 0.3 * 0.5$$

Definition (Conditional probability of an event)

If $\text{Prob}\{\mathcal{E}_1\} > 0$ then the conditional probability of $\mathcal{E}_2$ given $\mathcal{E}_1$ is

$$\text{Prob}\{\mathcal{E}_2 \mid \mathcal{E}_1\} = \frac{\text{Prob}\{\mathcal{E}_1, \mathcal{E}_2\}}{\text{Prob}\{\mathcal{E}_1\}}$$

Note that if two events are independent, the conditional probability $\text{Prob}\{\mathcal{E}_2 \mid \mathcal{E}_1\} = \text{Prob}\{\mathcal{E}_2\}$.

Examples of non-independent events:

- event $\mathcal{E}_2$: bad weather in Brussels; event $\mathcal{E}_1$: short commute time,

What about $\text{Prob}\{\mathcal{E}_2 \mid \mathcal{E}_1\}$?

Let us consider two random variables $\mathbf{z}_1$ and $\mathbf{z}_2$.

Definition (Conditional probability of a random variable)

If $\text{Prob}\{\mathbf{z}_1 = z_1\} > 0$ then the conditional probability of $\mathbf{z}_2 = z_2$ given $\mathbf{z}_1 = z_1$ is

$$\text{Prob}\{\mathbf{z}_2 = z_2 \mid \mathbf{z}_1 = z_1\} = \frac{\text{Prob}\{\mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2\}}{\text{Prob}\{\mathbf{z}_1 = z_1\}}$$

Weather example: conditional probability

Let the joint distribution be given by the table

z_1	z_2	z_3	$P(z_1 = z_1, z_2 = z_2, z_3 = z_3)$
CLEAR	RISING	DRY	0.4
CLEAR	RISING	WET	0.07
CLEAR	FALLING	DRY	0.08
CLEAR	FALLING	WET	0.10
CLOUDY	RISING	DRY	0.09
CLOUDY	RISING	WET	0.11
CLOUDY	FALLING	DRY	0.03
CLOUDY	FALLING	WET	0.12

From the joint distribution we can calculate the marginal probabilities $P(CLEAR, RISING) = 0.47$ and $P(CLOUDY) = 0.35$ and the conditional value

$$\begin{aligned} P(DRY|CLEAR, RISING) &= \\ &= \frac{P(DRY, CLEAR, RISING)}{P(CLEAR, RISING)} = \frac{0.40}{0.47} \approx 0.85 \end{aligned}$$

This law links conditional, joint and marginal (i.e. non conditional) probabilities.

Theorem (Law of total probability)

Let us consider a set of mutually exclusive and exhaustive events $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k$

It can be shown that

$$\text{Prob}\{\mathcal{E}\} = \sum_{i=1}^k \text{Prob}\{\mathcal{E}|\mathcal{E}_i\} \text{Prob}\{\mathcal{E}_i\} = \sum_{i=1}^k \text{Prob}\{\mathcal{E}, \mathcal{E}_i\}.$$

Marginal probability decomposed in the weighted sum of (easier) conditional probabilities where the conditional events form a partition of the sample space.

Example: law total probability

z_1	z_2	z_3	$P(z_1 = z_1, z_2 = z_2, z_3 = z_3)$
CLEAR	RISING	DRY	0.4
CLEAR	RISING	WET	0.07
CLEAR	FALLING	DRY	0.08
CLEAR	FALLING	WET	0.10
CLOUDY	RISING	DRY	0.09
CLOUDY	RISING	WET	0.11
CLOUDY	FALLING	DRY	0.03
CLOUDY	FALLING	WET	0.12

$P(\text{CLOUDY}) = 0.35$ can be obtained in several manners:

$$P(\text{CLOUDY}) = P(\text{CLOUDY}, \text{RISING}) + P(\text{CLOUDY}, \text{FALLING}) = 0.2 + 0.15 = 0.35$$

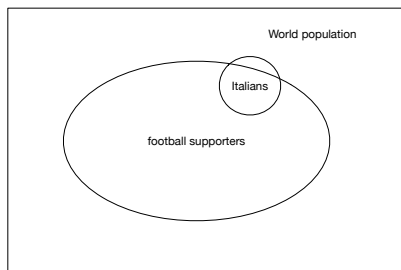
$$P(\text{CLOUDY}) = P(\text{CLOUDY}, \text{DRY}) + P(\text{CLOUDY}, \text{WET}) = 0.12 + 0.23 = 0.35$$

$$P(\text{CLOUDY}) = P(\text{CLOUDY}|\text{DRY})P(\text{DRY}) + P(\text{CLOUDY}|\text{WET})P(\text{WET}) = \\ 0.12/(0.12 + 0.48) * 0.6 + 0.23/(0.23 + 0.17) * 0.4 = 0.2 * 0.6 + 0.575 * 0.4 = 0.35$$

Direct and inverse conditional probability

- ▶ NOTA BENE: Direct and inverse conditional probability are not the same !!
- ▶ it is generally NOT the case that $\text{Prob} \{ \mathcal{E}_2 | \mathcal{E}_1 \} = \text{Prob} \{ \mathcal{E}_1 | \mathcal{E}_2 \}$.
- ▶ Think about the probability that a generic person "likes football given that (s)he is Italian", and the probability of that someone "is Italian given that (s)he likes football"
- ▶ Are they the same? Why?

Direct and inverse conditional probability



Note that:

$P(\text{Italian}|\text{Supporter}) + P(\text{NOT Italian}|\text{Supporter}) = 1$ but ...

$P(\text{Supporter}|\text{Italian}) + P(\text{Supporter}|\text{NOT Italian})$ is not necessarily one

Weather example: conditional and inverse probability

z_1	z_2	z_3	$P(z_1 = z_1, z_2 = z_2, z_3 = z_3)$
CLEAR	RISING	DRY	0.4
CLEAR	RISING	WET	0.07
CLEAR	FALLING	DRY	0.08
CLEAR	FALLING	WET	0.10
CLOUDY	RISING	DRY	0.09
CLOUDY	RISING	WET	0.11
CLOUDY	FALLING	DRY	0.03
CLOUDY	FALLING	WET	0.12

$P(\text{CLEAR} \mid \text{RISING}) = 0.47/0.67 = 0.701$ is different

$P(\text{RISING} \mid \text{CLEAR}) = 0.47/0.65 = 0.723$

$P(\text{CLEAR} \mid \text{FALLING}) = 0.18/0.33 = 0.545$ is different

$P(\text{FALLING} \mid \text{CLEAR}) = 0.18/0.65 = 0.277$

Given $P(\text{RISING} \cup \text{FALLING}) = P(\text{RISING}) + P(\text{FALLING}) = 1$ it follows

$P(\text{RISING} \mid \text{CLEAR}) + P(\text{FALLING} \mid \text{CLEAR}) = 1$ but

$P(\text{CLEAR} \mid \text{RISING}) + P(\text{CLEAR} \mid \text{FALLING}) \neq 1$

Conditional probabilities are probabilities...

- ▶ For any fixed $\mathcal{E}_1$, the quantity $\text{Prob}\{\cdot | \mathcal{E}_1\}$ satisfies the axioms of probability. E.g, if $\mathcal{E}_2$, $\mathcal{E}_3$ and $\mathcal{E}_4$ are disjoint events we have that

$$\text{Prob}\{\mathcal{E}_2 \cup \mathcal{E}_3 \cup \mathcal{E}_4 | \mathcal{E}_1\} = \text{Prob}\{\mathcal{E}_2 | \mathcal{E}_1\} + \text{Prob}\{\mathcal{E}_3 | \mathcal{E}_1\} + \text{Prob}\{\mathcal{E}_4 | \mathcal{E}_1\}$$

- ▶ This does NOT generally hold for $\text{Prob}\{\mathcal{E}_1 | \cdot\}$, that is when we fix the term $\mathcal{E}_1$ on the left of the conditional bar. For two disjoint events $\mathcal{E}_2$ and $\mathcal{E}_3$, in general

$$\text{Prob}\{\mathcal{E}_1 | \mathcal{E}_2 \cup \mathcal{E}_3\} \neq \text{Prob}\{\mathcal{E}_1 | \mathcal{E}_2\} + \text{Prob}\{\mathcal{E}_1 | \mathcal{E}_3\}$$

- ▶ If $\mathcal{E}^c$ denotes the complement of $\mathcal{E}$ then

$$\text{Prob}\{\mathcal{E}_1^c | \mathcal{E}_2\} = 1 - \text{Prob}\{\mathcal{E}_1 | \mathcal{E}_2\}$$

but

$$\text{Prob}\{\mathcal{E}_1 | \mathcal{E}_2^c\} \neq 1 - \text{Prob}\{\mathcal{E}_1 | \mathcal{E}_2\}$$

and all probabilities are conditional ...

- ▶ Whenever we make a probability statement there is always some conditional information that we are conditioning on, though we do not make it explicit.
- ▶ For instance the a priori probability $\text{Prob}\{z = \text{Italian}\}$ that someone has the Italian nationality is implicitly conditioned on the hypothesis that a nationality is defined according to some criterion (e.g. an official passport).
- ▶ By changing the definition of nationality (e.g. by considering that it depends on the parent nationality) would produce a different value
- ▶ The unconditional probability $\text{Prob}\{z = \text{Italian}\}$ should be better written as $\text{Prob}\{z = \text{Italian}|K\}$ where K defines some background knowledge.

Conditional probability and miscommunication

Consider the following statement of a right-wing politician to justify the prohibition of light drugs:

Since most heroin addicts used marijuana, most marijuana users will become heroin addicts

Is it a sound probabilistic argument?

Read also about the "prosecutor fallacy" example in the handbook.

Conditional probability is non monotonic

It does not necessarily increase or decrease by adding conditioning information

Suppose I consider the probability of the event $\mathcal{E}$: "sunny day somewhere" and its evolution

$$P(\mathcal{E})$$

$$P(\mathcal{E} \mid \text{country} = \text{"Italy"})$$

$$P(\mathcal{E} \mid \text{country} = \text{"Italy"}, \text{region} = \text{"Valle d'Aosta"})$$

$$P(\mathcal{E} \mid \text{country} = \text{"Italy"}, \text{region} = \text{"Valle d'Aosta"}, \text{month} = \text{"december"})$$

$$P(\mathcal{E} \mid \text{country} = \text{"Italy"}, \text{region} = \text{"Valle d'Aosta"}, \\ \text{month} = \text{"december"}, \text{daybefore} = \text{"sunny"})$$

Theorem (Bayes' theorem)

Let us consider a set of mutually exclusive and exhaustive events $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k$. The conditional ("inverse") probability of any $\mathcal{E}_i$, $i = 1, \dots, k$ given that $\mathcal{E}$ has occurred is given by

$$\text{Prob}\{\mathcal{E}_i | \mathcal{E}\} = \frac{\text{Prob}\{\mathcal{E} | \mathcal{E}_i\} \text{Prob}\{\mathcal{E}_i\}}{\text{Prob}\{\mathcal{E}\}} = \frac{\text{Prob}\{\mathcal{E} \cap \mathcal{E}_i\}}{\sum_{j=1}^k \text{Prob}\{\mathcal{E} | \mathcal{E}_j\} \text{Prob}\{\mathcal{E}_j\}}$$

Example: how probable was the bad weather last Wednesday when it took a long time going back home by car?

- ▶ event $\mathcal{E}$: commute car time longer than average
- ▶ event $\mathcal{E}_1$: nice weather in Brussels, event $\mathcal{E}_2$: bad weather in Brussels

Weather example: Bayes theorem

z_1	z_2	z_3	$P(z_1 = z_1, z_2 = z_2, z_3 = z_3)$
CLEAR	RISING	DRY	0.4
CLEAR	RISING	WET	0.07
CLEAR	FALLING	DRY	0.08
CLEAR	FALLING	WET	0.10
CLOUDY	RISING	DRY	0.09
CLOUDY	RISING	WET	0.11
CLOUDY	FALLING	DRY	0.03
CLOUDY	FALLING	WET	0.12

$P(CLEAR|FALLING) = 0.545 \neq P(FALLING|CLEAR) = 0.277$ but

$$P(CLEAR|FALLING) = \frac{P(FALLING|CLEAR)P(CLEAR)}{P(FALLING)} = \frac{0.277*0.65}{0.33} = 0.545$$

We have seen that

$$\text{Prob} \{ \mathcal{E}_1, \mathcal{E}_2 \} = \text{Prob} \{ \mathcal{E}_1 \} \text{Prob} \{ \mathcal{E}_2 | \mathcal{E}_1 \}$$

In more general terms

Definition (Chain rule)

For any sequence of events $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_n$,

$$\begin{aligned} \text{Prob} \{ \mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_n \} = \\ \text{Prob} \{ \mathcal{E}_1 \} \text{Prob} \{ \mathcal{E}_2 | \mathcal{E}_1 \} \text{Prob} \{ \mathcal{E}_3 | \mathcal{E}_1, \mathcal{E}_2 \} \dots \text{Prob} \{ \mathcal{E}_n | \mathcal{E}_1, \mathcal{E}_2 \dots \mathcal{E}_{n-1} \} \end{aligned}$$

Weather example: chain rule

z_1	z_2	z_3	$P(z_1 = z_1, z_2 = z_2, z_3 = z_3)$
CLEAR	RISING	DRY	0.4
CLEAR	RISING	WET	0.07
CLEAR	FALLING	DRY	0.08
CLEAR	FALLING	WET	0.10
CLOUDY	RISING	DRY	0.09
CLOUDY	RISING	WET	0.11
CLOUDY	FALLING	DRY	0.03
CLOUDY	FALLING	WET	0.12

$P(\text{CLEAR}, \text{FALLING}, \text{DRY}) = 0.08$ can be obtained by the chain rule

$$\begin{aligned} P(\text{CLEAR}, \text{FALLING}, \text{DRY}) &= \\ P(\text{CLEAR} | \text{FALLING}, \text{DRY}) P(\text{FALLING} | \text{DRY}) P(\text{DRY}) &= \\ 0.727 * 0.183 * 0.6 &= 0.08 \end{aligned}$$

Conditional independence

Independence is not a stable relation. Though $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$, the r.v. $\mathbf{x}$ may become dependent with $\mathbf{y}$ once we observe a third variable $\mathbf{z}$. In the same way, two dependent variables $\mathbf{x}$ and $\mathbf{y}$ may become independent once the value of $\mathbf{z}$ is given.

Definition (Conditional independence)

Two r.v.s $\mathbf{x}$ and $\mathbf{y}$ are *conditionally independent given the value $\mathbf{z} = z$* ($\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} = z$) iff

$$P(\mathbf{x} = x, \mathbf{y} = y \mid \mathbf{z} = z) = P(\mathbf{x} = x \mid \mathbf{z} = z)P(\mathbf{y} = y \mid \mathbf{z} = z) \quad \forall x, y$$

Two r.v.s $\mathbf{x}$ and $\mathbf{y}$ are *conditionally independent given $\mathbf{z}$* ($\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$) iff they are conditionally independent for all values of $\mathbf{z}$.

Since from the chain rule

$$P(\mathbf{x} = x, \mathbf{y} = y \mid \mathbf{z} = z) = P(\mathbf{x} = x \mid \mathbf{z} = z)P(\mathbf{y} = y \mid \mathbf{x} = x, \mathbf{z} = z)$$

if $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} = z$ then

$$P(\mathbf{y} = y \mid \mathbf{x} = x, \mathbf{z} = z) = P(\mathbf{y} = y \mid \mathbf{z} = z)$$

- ▶ Conditional dependence makes formal the intuition that a variable may bring (or not) information about a second one, according to the context.
- ▶ $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z} = z$ means that $\mathbf{x}$ and $\mathbf{y}$ are independent if $\mathbf{z} = z$ occurs but does not say anything about the relation between $\mathbf{x}$ and $\mathbf{y}$ if $\mathbf{z} = z$ does not occur.
- ▶ Two variables may be independent but not conditional independent (or the other way round).



BALLEKES





BALLEKES



Pizza homework

$z_1(\text{owner})$	$z_2(\text{cook})$	$z_3(\text{pizza})$	$P(z_1 = z_1, z_2 = z_2, z_3 = z_3)$
ITALIAN	ITALIAN	GOOD	0.378
BELGIAN	ITALIAN	GOOD	0.168
ITALIAN	BELGIAN	GOOD	0.012
BELGIAN	BELGIAN	GOOD	0.032
ITALIAN	ITALIAN	BAD	0.162
BELGIAN	ITALIAN	BAD	0.072
ITALIAN	BELGIAN	BAD	0.048
BELGIAN	BELGIAN	BAD	0.128

Show that

1. $z_3 \not\perp z_1$
2. $z_3 \perp z_1 | z_2 = \text{ITALIAN}$ and
3. $z_3 \perp z_1 | z_2 = \text{BELGIAN}$.

1. $\mathbf{z}_3 \not\perp \mathbf{z}_1$:

$$\text{Prob}\{\mathbf{z}_3 = \text{GOOD}\} = 0.59 \neq \text{Prob}\{\mathbf{z}_3 = \text{GOOD} | \mathbf{z}_1 = \text{IT}\} = 0.65 \neq \text{Prob}\{\mathbf{z}_3 = \text{GOOD} | \mathbf{z}_1 = \text{BE}\} = 0.5$$

2. $\mathbf{z}_3 \perp \mathbf{z}_1 | \mathbf{z}_2 = \text{ITALIAN}$:

$$\begin{aligned} \text{Prob}\{\mathbf{z}_3 = \text{GOOD} | \mathbf{z}_2 = \text{IT}\} &= 0.7 = \\ \text{Prob}\{\mathbf{z}_3 = \text{GOOD} | \mathbf{z}_1 = \text{BE}, \mathbf{z}_2 = \text{IT}\} &= \\ \text{Prob}\{\mathbf{z}_3 = \text{GOOD} | \mathbf{z}_1 = \text{IT}, \mathbf{z}_2 = \text{IT}\} & \end{aligned}$$

3. $\mathbf{z}_3 \perp \mathbf{z}_1 | \mathbf{z}_2 = \text{BELGIAN}$:

$$\begin{aligned} \text{Prob}\{\mathbf{z}_3 = \text{GOOD} | \mathbf{z}_2 = \text{BE}\} &= 0.2 = \\ \text{Prob}\{\mathbf{z}_3 = \text{GOOD} | \mathbf{z}_1 = \text{BE}, \mathbf{z}_2 = \text{BE}\} &= \\ \text{Prob}\{\mathbf{z}_3 = \text{GOOD} | \mathbf{z}_1 = \text{IT}, \mathbf{z}_2 = \text{BE}\} & \end{aligned}$$

Conditional independence and dependence

From the pizza example we see that

Conditional independence does not imply independence

but also

Independence does not imply conditional independence.

Think about two independent causes of a common effect

Random numerical variables

- ▶ Machine learning deals with data, i.e. numeric measures.
- ▶ A measure is mapping from an outcome to a real number. For instance the temperature you read in the thermometer is the result of mapping from a thermal state to a number.
- ▶ Random numerical variables as numeric summaries of (sometimes extremely complex) random experiments.
- ▶ The probability distribution in the sample space and the mapping induce a probability distribution in the space of values of the random variable.
- ▶ A random variable has a **probability distribution** i.e. a **mathematical function** which returns the probability of occurrence of events related to $\mathbf{z}$.
- ▶ The probability distribution of a random variable (or set of variables) $\mathbf{z}$ is a complete description of the probabilistic behavior of $\mathbf{z}$.

- ▶ If not made explicit a random variable is a random numeric variables.
- ▶ To distinguish between random variables and their values, we use the **boldface notation for denoting a random variable (e.g. $\mathbf{z}$)** and the normal face notation for the observed value (e.g. $z = 11$).
- ▶ Example: $\mathbf{z}$ could be the age of a student before asking and $z = 22$ could be his value after the observation.
- ▶ Given a probability distribution $P_{\mathbf{z}}(z)$ the notation

$$P_{\mathbf{z}} \rightarrow \{z_1, z_2, \dots, z_N\}$$

means that the dataset $D_N = \{z_1, z_2, \dots, z_N\}$ is a i.i.d. random sample observed from the probability distribution $F_{\mathbf{z}}(\cdot)$.

Definition (Probability function of a discrete r.v.)

The **probability (mass) function** of a discrete r.v. $\mathbf{z}$ is the combination of

1. the countable² set $\mathcal{Z}$ of values that this r.v. can take (also called **range** or **sample space**)
2. the set of probabilities associated to each value of $\mathcal{Z}$

We attach to the random variable some specific mathematical function $P(\mathbf{z})$ that gives for each $z \in \mathcal{Z}$ the probability that $\mathbf{z}$ assumes the value z

$$P_{\mathbf{z}}(z) = \text{Prob} \{ \mathbf{z} = z \}$$

This function must satisfy

$$\begin{cases} P_{\mathbf{z}}(z) \geq 0 \\ \sum_{z \in \mathcal{Z}} P_{\mathbf{z}}(z) = 1 \end{cases}$$

²finite or in one-to-one correspondence with integers

Probability mass function of a discrete r.v.(II)

- ▶ For small range of $\mathbf{z}$, the probability function can be presented in the form of a table.
- ▶ If we plan to toss a *fair* coin twice, and the random variable $\mathbf{z}$ is the **number of heads** that eventually turn up, the probability function can be presented as follow

Values of the random variable $\mathbf{z}$	0	1	2
Associated probabilities	0.25	0.50	0.25

Definition (Expected value)

Expected value of a discrete random variable $\mathbf{z}$ is

$$E[\mathbf{z}] = \mu = \sum_{z \in \mathcal{Z}} z P_{\mathbf{z}}(z)$$

- ▶ $E[\mathbf{z}]$ is a weighted average of the possible values that $\mathbf{z}$ could assume, where each value is weighted with the probability that $\mathbf{z}$ would assume that value.
- ▶ $E[\mathbf{z}]$ must not be confused with the "most probable value"
- ▶ $E[\mathbf{z}]$ is not necessarily a value that belongs to $\mathcal{Z}$ (i.e. the expected value of a die roll is 3.5)
- ▶ In this class we will use "mean" is used as a synonymous of "expected value"...
- ▶ .. but the word "average" is NOT a synonymous of "expected value".

Definition (Variance)

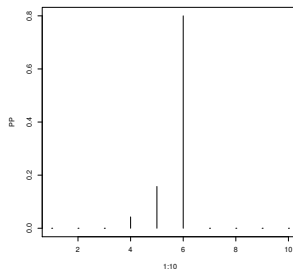
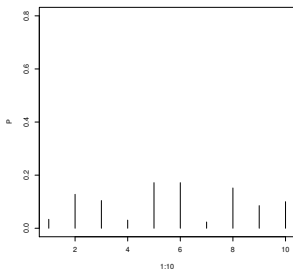
Variance of a discrete random variable $\mathbf{z}$ is

$$\text{Var}[\mathbf{z}] = \sigma^2 = E[(\mathbf{z} - E[\mathbf{z}])^2] = \sum_{z \in \mathcal{Z}} (z - E[\mathbf{z}])^2 P_{\mathbf{z}}(z)$$

- ▶ Measure of the dispersion of the probability function around its mean.
- ▶ Note that since $(\mathbf{z} - \mu)^2 = \mathbf{z}^2 - 2\mu\mathbf{z} + \mu^2$:

$$E[(\mathbf{z} - E[\mathbf{z}])^2] \equiv E[\mathbf{z}^2] - (E[\mathbf{z}])^2 = E[\mathbf{z}^2] - \mu^2$$

Examples of probability mass functions



Two discrete r.v. probability functions having the same mean but different variance.

Definition (Standard deviation)

Standard deviation of a discrete random variable $\mathbf{z}$ is defined as the positive square root of the variance.

$$\text{Std}[\mathbf{z}] = \sqrt{\text{Var}[\mathbf{z}]} = \sigma$$

Note that standard deviation is expressed in the same unit as $\mathbf{z}$

Definition (Moment)

For any positive integer r , the r th moment of the probability function is

$$\mu_r = E[\mathbf{z}^r] = \sum_z z^r P_z(z)$$

Definition (Skewness)

The skewness of a discrete random variable $\mathbf{z}$ is

$$\gamma = \frac{E[(\mathbf{z} - \mu)^3]}{\sigma^3}$$

Positive (negative) sk. means long right (left) tails.

Continuous random variable

Continuous random variables take their value in some continuous range of values. Consider a real random variable $\mathbf{z}$ whose range is the set of real numbers. The following quantities can be defined:

Definition

The (*cumulative*) *distribution function* of $\mathbf{z}$ is the function

$$F_z(z) = \text{Prob} \{ \mathbf{z} \leq z \}$$

Definition

The *density function* of a real random variable $\mathbf{z}$ is the derivative of the distribution function:

$$p_z(z) = \frac{dF_z(z)}{dz}$$

Continuous random variable

- ▶ Any individual value has probability zero for a continuous random variable
- ▶ Probabilities of continuous r.v. are not allocated to specific values but rather to interval of values. Specifically

$$\text{Prob}\{a \leq \mathbf{z} \leq b\} = \int_a^b p_{\mathbf{z}}(z) dz, \quad \int_{\mathcal{Z}} p_{\mathbf{z}}(z) dz = 1$$

Mean, variance,... of a continuous r.v.

Consider a continuous scalar r.v. having range (l, h) and density function $p(z)$. We can define

Expectation (mean):

$$\mu = \int_l^h zp(z)dz$$

Variance:

$$\sigma^2 = \int_l^h (z - \mu)^2 p(z)dz$$

Other quantities of interest are the *moments* :

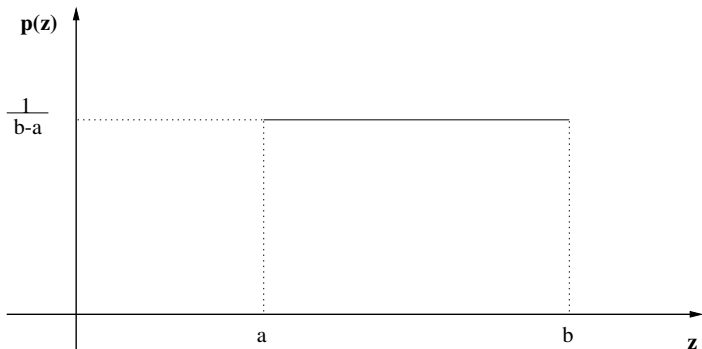
$$\mu_r = E[z^r] = \int_l^h z^r p(z)dz$$

The moment of order $r = 1$ is the *mean* of $\mathbf{z}$.

Uniform distribution

A random variable z is said to be **uniformly distributed** on the interval (a, b) (also $z \sim \mathcal{U}(a, b)$) if its probability density function is given by

$$p(z) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq z \leq b \\ 0, & \text{otherwise} \end{cases}$$



Question: compute the variance of $\mathcal{U}(a, b)$.

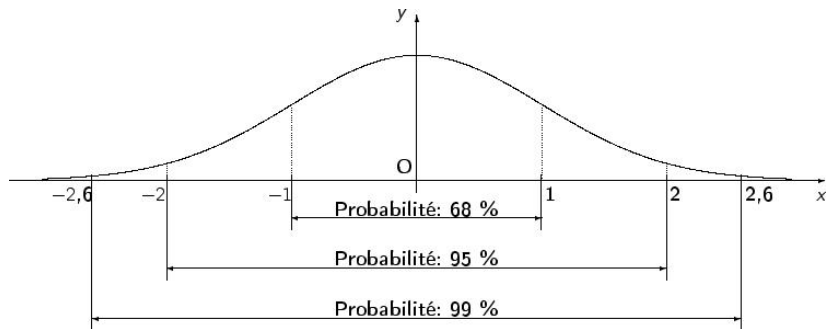
Normal distribution: the scalar case

A continuous scalar random variable $\mathbf{x}$ is said to be **normally distributed** with parameters μ and σ^2 (also $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$) if its probability density function is given by

$$p_{\mathbf{x}}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- ▶ The mean of $\mathbf{x}$ is μ ; the variance of $\mathbf{x}$ is σ^2 .
- ▶ The coefficient in front of the exponential ensures that $\int p(x)dx = 1$.
- ▶ The probability that an observation x is within 1 (2) standard deviations from the mean is 0.68 (0.95).
- ▶ If $\mu = 0$ and $\sigma^2 = 1$ the distribution $F_z(z) = \Phi(z)$ is defined **standard normal**.
- ▶ If $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$, the r.v. $\mathbf{z} = (\mathbf{x} - \mu)/\sigma$ has a standard normal distribution.
- ▶ $\mathbf{z} \sim \mathcal{N}(0, 1) \Rightarrow \mathbf{x} = \mu + \sigma\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$

Standard distribution $\mu = 0, \sigma^2 = 1$



Approximated rule

For a sample drawn from a normal distribution $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$, the interval with endpoints

- ▶ $\mu \pm \sigma$ contains approximately 68% of the measurements,
- ▶ $\mu \pm 2\sigma$ contains approximately 95% of the measurements,
- ▶ $\mu \pm 3\sigma$ contains approximately all the measurements,

Important relations

$\mathbf{x} \in \mathcal{N}(\mu, \sigma^2)$
$\text{Prob}\{\mu - \sigma \leq \mathbf{x} \leq \mu + \sigma\} \approx 0.683$
$\text{Prob}\{\mu - 1.282\sigma \leq \mathbf{x} \leq \mu + 1.282\sigma\} \approx 0.8$
$\text{Prob}\{\mu - 1.645\sigma \leq \mathbf{x} \leq \mu + 1.645\sigma\} \approx 0.9$
$\text{Prob}\{\mu - 1.96\sigma \leq \mathbf{x} \leq \mu + 1.96\sigma\} \approx 0.95$
$\text{Prob}\{\mu - 2\sigma \leq \mathbf{x} \leq \mu + 2\sigma\} \approx 0.954$
$\text{Prob}\{\mu - 2.57\sigma \leq \mathbf{x} \leq \mu + 2.57\sigma\} \approx 0.99$
$\text{Prob}\{\mu - 3\sigma \leq \mathbf{x} \leq \mu + 3\sigma\} \approx 0.997$

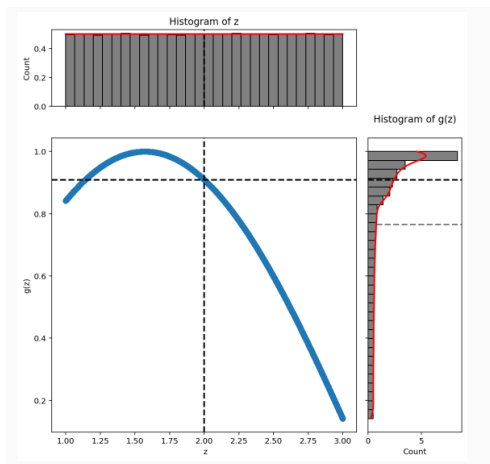
The sum $\mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2$ where $\mathbf{x}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathbf{x}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent is still Normal:

$$\mathbf{z} \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Test yourself these relations by random sampling and simulation using the script `Probability/norm.py`

Transformation of random variables

Let z a continuous r.v. and $y = g(z)$ a deterministic function of z . What about the pdf of y ?



See script `Probability/transform.py`

Transformation of random variables

Let $\mathbf{z}$ a continuous r.v. and $\mathbf{y} = g(\mathbf{z})$ a deterministic function of $\mathbf{z}$. What about $E[\mathbf{y}|\mathbf{z}]$ and $E[\mathbf{y}]$?

- ▶ $E[\mathbf{y}|\mathbf{z}] = g(\mathbf{z})$: since g is a deterministic function.
- ▶ $E[\mathbf{y}]$: *Law of the unconscious statistician* (LOTUS):

$$E[\mathbf{y}] = E[g(\mathbf{z})] = \int_{\mathcal{Z}} g(\mathbf{z}) dF_{\mathbf{z}}(\mathbf{z}) = \int g(\mathbf{z}) p_{\mathbf{z}}(\mathbf{z}) d\mathbf{z} \quad (1)$$

No need to know the distribution of $g(\mathbf{z})$ to compute its expectation.

Note also that for a generic $g(\cdot)$,

$$E[g(\mathbf{z})] \neq g(E[\mathbf{z}])$$

Monte Carlo simulation

For a generic g , the analytical computation may be too complex. A numerical alternative is represented by the Monte Carlo simulation which **requires a pseudo-random generator** of $\mathbf{z}$ samples.

The Monte Carlo approximation of $E[g(\mathbf{z})]$ is

1. generating a large number S of samples $\mathbf{z}_i \sim F_{\mathbf{z}}, i = 1, \dots, S$,
2. computing $g(\mathbf{z}_i)$,
3. returning the estimation

$$E[g(\mathbf{z})] \approx \frac{\sum_{i=1}^S g(\mathbf{z}_i)}{S}$$

If S is sufficiently large we may consider such approximation as reliable. We will have often recourse to Monte Carlo simulation to provide a numerical illustration of probabilistic formulas or concepts

Monte Carlo approximation

Let us consider a uniform $z \sim \mathcal{U}(a, b)$ and suppose we want to compute its variance.

► Analytical derivation

$$\text{Var}[z] = E[(z - \mu)^2] = \int_{a,b} \left(z - \frac{b+a}{2}\right)^2 p(z) dz = \frac{1}{b-a} \int_{a,b} \left(z - \frac{b+a}{2}\right)^2 dz = \frac{(b-a)^2}{12}$$

► Script Probability/mcarloU.py: MC approximation of $\text{Var}[z] = E[g(z)]$ where $g(z) = (z - \mu)^2$

```
import numpy as np

R = 50000 # number of MC trials

a = -1.0
b = 1.0
muz = (b + a) / 2

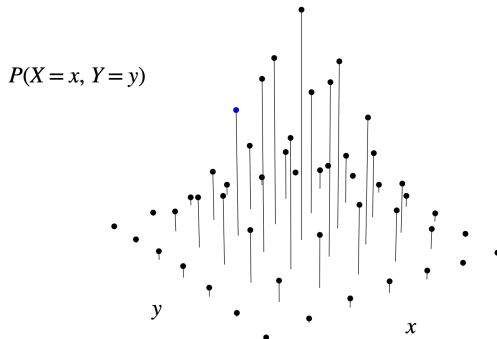
Z = np.zeros(R)
for r in range(R):
    z = np.random.uniform(a, b)
    Z[r] = (z - muz)**2

print(f"Var. th= {(b-a)**2/12:.6f}, MC approximation= {np.mean(Z):.6f}")
```

Compute by Monte Carlo simulation

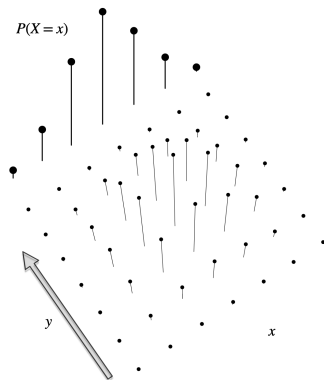
- ▶ $E[|z|]$ and $\text{Var}[|z|]$ where z is Gaussian. Script `Probability/mcarlo.py`
- ▶ the validity of Jensen's inequality $f(E[\mathbf{x}]) \leq E[f(\mathbf{x})]$ where $\mathbf{x} \sim \mathcal{U}(a, b)$ and f is convex. Script `Probability/mcarlo3.py`
- ▶ $E[\max(\mathbf{x}, \mathbf{y})]$ where $\mathbf{x}$ is Gaussian and $\mathbf{y}$ is Uniform. Script `Probability/mcarlo5.py`

Bivariate discrete distribution



From the book J.K. Blitzstein, J, Hwang, *Introduction to probability, 2nd edition*

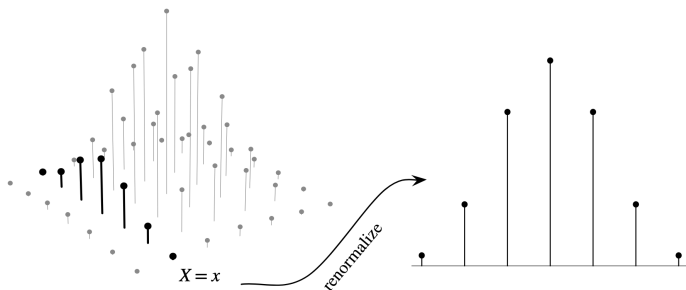
From bivariate discrete distribution to marginal



From law of total probability

$$\text{Prob}\{\mathbf{x} = x\} = \sum_{y \in \mathcal{Y}} \text{Prob}\{\mathbf{x} = x, \mathbf{y} = y\}$$

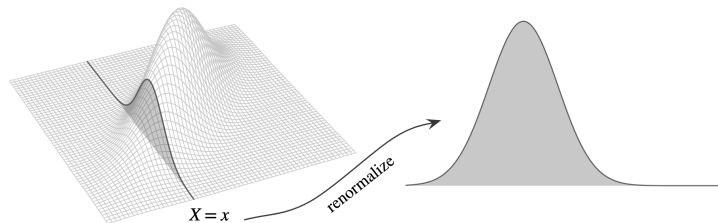
From bivariate discrete distribution to conditional



From Bayes' theorem

$$\text{Prob}\{\mathbf{y} = y | \mathbf{x} = x\} = \frac{\text{Prob}\{\mathbf{y} = y, \mathbf{x} = x\}}{\text{Prob}\{\mathbf{x} = x\}}$$

From bivariate continuous density to conditional density



From Bayes' theorem

$$p(\mathbf{y} = y | \mathbf{x} = x) = \frac{p_{\mathbf{x},\mathbf{y}}(x, y)}{p_{\mathbf{x}}(x)}$$

- ▶ the distribution of a single r.v. $\mathbf{z}$ provides **complete information** about the probability of $\mathbf{z}$ falling into any subset of the real line.
- ▶ But the distribution of two r.v.s $\mathbf{x}$ and $\mathbf{y}$ does **NOT** provide **complete information** about the probability of the pair $\mathbf{x}, \mathbf{y}$ falling into any subset of the plane.
- ▶ Complete information is provided only by the **joint information**.
- ▶ The marginal distribution of $\mathbf{x}$ is the individual distribution of $\mathbf{x}$, ignoring the value of $\mathbf{y}$
- ▶ Conditional distribution of $\mathbf{x}$ given $\mathbf{y} = y$ is the updated distribution for $\mathbf{x}$ after observing $\mathbf{y} = y$.

Linear combinations

- ▶ The expectation value of a linear combination of r.v.'s is simply the linear combination of their respective expectation values

$$E[ax + by] = aE[x] + bE[y]$$

i.e., expectation is a linear statistic.

- ▶ Since the variance is not a linear statistic, we have

$$\begin{aligned}\text{Var}[ax + by] &= a^2\text{Var}[x] + b^2\text{Var}[y] + 2ab(E[xy] - E[x]E[y]) = \\ &= a^2\text{Var}[x] + b^2\text{Var}[y] + 2ab\text{Cov}[x, y]\end{aligned}$$

where

$$\text{Cov}[x, y] = E[(x - E[x])(y - E[y])] = E[xy] - E[x]E[y]$$

is called **covariance**.

- ▶ Covariance measures the degree to which two variables vary simultaneously in the same way around their average. This is a measure of (linear) association.
- ▶ See the Shiny dashboard `mcarlo.R`.

Correlation

- ▶ The **correlation coefficient** is

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\text{Cov}[\mathbf{x}, \mathbf{y}]}{\sqrt{\text{Var}[\mathbf{x}] \text{Var}[\mathbf{y}]}}$$

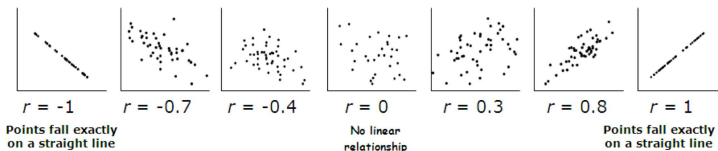
It is easily shown that $-1 \leq \rho(\mathbf{x}, \mathbf{y}) \leq 1$.

- ▶ Two r.v. are called *uncorrelated* if

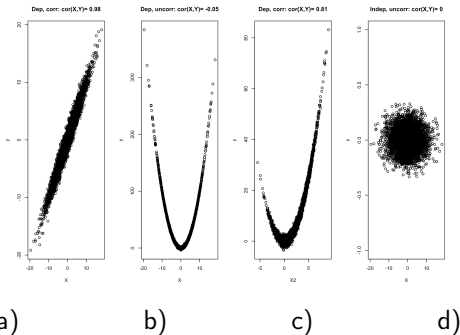
$$E[\mathbf{xy}] = E[\mathbf{x}]E[\mathbf{y}]$$

- ▶ If $\mathbf{x}$ and $\mathbf{y}$ are two independent random variables then $\text{Cov}[\mathbf{x}, \mathbf{y}] = 0$ or equivalently $E[\mathbf{xy}] = E[\mathbf{x}]E[\mathbf{y}]$.
- ▶ If $\mathbf{x}$ and $\mathbf{y}$ are two independent random variables then also $\text{Cov}[g(\mathbf{x}), h(\mathbf{y})] = 0$ or equivalently $E[g(\mathbf{x})h(\mathbf{y})] = E[g(\mathbf{x})]E[h(\mathbf{y})]$.
- ▶ Independence $\Rightarrow$ Uncorrelation but not viceversa for a generic distribution.
- ▶ Independence $\Leftrightarrow$ Uncorrelation if $\mathbf{x}$ and $\mathbf{y}$ are jointly gaussian.

Correlation



Dependence and correlation



- ▶ Independence $\Rightarrow$ Uncorrelation : d)
- ▶ Correlation $\Rightarrow$ Dependence : a), c)
- ▶ Dependence $\nRightarrow$ Correlation : b)
- ▶ Uncorrelation $\nRightarrow$ Independence: b)

Monte Carlo homework

Let us consider the r.v. $\mathbf{x}$ and $\mathbf{y} = K\mathbf{x}$.

1. Compute analytically the covariance of such two variables
2. Verify that the result is correct by using a Monte Carlo simulation for a Normal $\mathbf{x}$.
3. Verify that the result is correct by using a Monte Carlo simulation for a Uniform $\mathbf{x}$.

Script Probability/mcarlo2.py

Normal distribution: the multivariate case

Let $\mathbf{z}$ be a random vector ($n \times 1$).

The vector is said to be **normally distributed** with parameters $\mu_{(n \times 1)}$ and $\Sigma_{(n \times n)}$ (also $\mathbf{z} \sim \mathcal{N}(\mu, \Sigma)$) if its probability density function is given by

$$p_{\mathbf{z}}(\mathbf{z}) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(\Sigma)}} \exp \left\{ -\frac{1}{2} (\mathbf{z} - \mu)^T \Sigma^{-1} (\mathbf{z} - \mu) \right\}$$

It follows that

- ▶ the mean $E[\mathbf{z}] = \mu = [\mu_1, \dots, \mu_n]^T$ is a $[n, 1]$ -dimensional vector, where $\mu_i = E[\mathbf{z}_i]$, $i = 1, \dots, n$,
- ▶ the $[n, n]$ matrix

$$\Sigma_{(n \times n)} = E[(\mathbf{z} - \mu)(\mathbf{z} - \mu)^T] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \dots & \sigma_n^2 \end{bmatrix}$$

is the covariance matrix where $\sigma_i^2 = \text{Var}[\mathbf{z}_i]$ and $\sigma_{ij} = \text{Cov}[\mathbf{z}_i, \mathbf{z}_j]$. Σ is squared, symmetric and positive definite and has $n(n+1)/2$ parameters.

Normal multivariate distribution (II)

The quantity

$$\Delta = (z - \mu)^T \Sigma^{-1} (z - \mu)$$

is the *Mahalanobis distance* from z to μ .

It can be shown that the surfaces of constant probability density

- ▶ are hyperellipsoids on which Δ^2 is constant;
- ▶ their *principal axes* are given by the eigenvectors u_i , $i = 1, \dots, n$ of Σ which satisfy

$$\Sigma u_i = \lambda_i u_i \quad i = 1, \dots, n$$

where λ_i are the corresponding eigenvalues.

- ▶ the eigenvalues λ_i give the variances along the principal directions.

Normal multivariate distribution (III)

If the covariance matrix Σ is **diagonal** then

- ▶ the contours of constant density are hyperellipsoids with the principal directions aligned with the coordinate axes.
- ▶ the components of $\mathbf{z}$ are then *statistically independent* since the distribution of $\mathbf{z}$ can be written as the product of the distributions for each of the components separately in the form

$$p_{\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^n p(z_i)$$

- ▶ the total number of independent parameters in the distribution is $2n$.
- ▶ if $\sigma_i = \sigma$ for all i , the contours of constant density are hyperspheres.

Bivariate normal distribution

Consider a bivariate normal density whose mean is $\mu = [\mu_1, \mu_2]^T$ and the covariance matrix is

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

The correlation coefficient is the dimensionless number

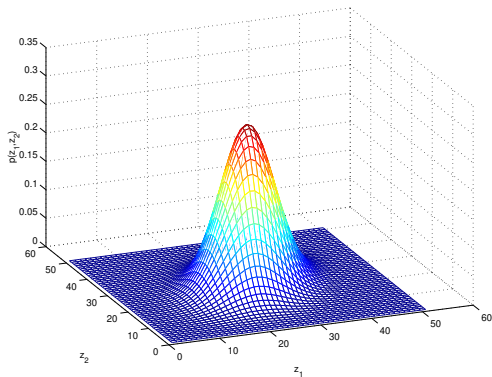
$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

It can be shown that the general bivariate normal density has the form

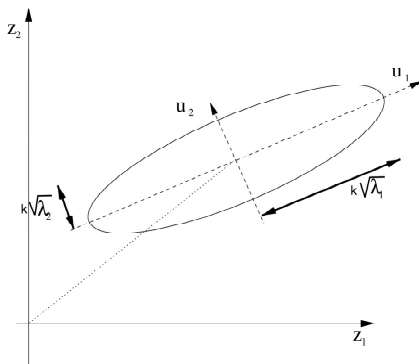
$$p(z_1, z_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} \left[\left(\frac{z_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{z_1 - \mu_1}{\sigma_1} \right) \left(\frac{z_2 - \mu_2}{\sigma_2} \right) + \left(\frac{z_2 - \mu_2}{\sigma_2} \right)^2 \right] \right]$$

Bivariate normal distribution

Let $\Sigma = [1.2919, 0.4546; 0.4546, 1.7081]$.



Bivariate normal distribution (prj)



Marginal and conditional distributions

In a multivariate normal density, all conditional and marginal probabilities are also normal.

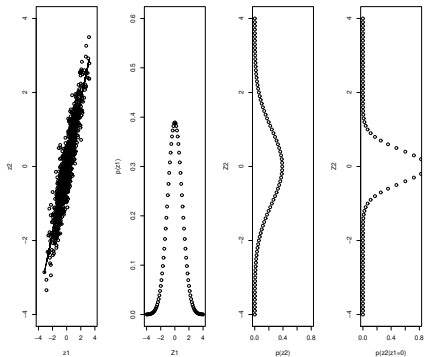
In the bivariate case it can be shown that $p(z_2|z_1)$ is a normal distribution $\mathcal{N}(\mu_{2|1}, \sigma_{2|1}^2)$, where

$$\begin{aligned}\mu_{2|1} &= \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (z_1 - \mu_1) \\ \sigma_{2|1}^2 &= \sigma_2^2 (1 - \rho^2)\end{aligned}$$

Note that

- ▶ $\mu_{2|1}$ is a linear function of z_1 : if the correlation coefficient ρ is positive, the larger z_1 , the larger $\mu_{2|1}$.
- ▶ if there is no correlation between z_1 and z_2 , we can ignore the value of z_1 to estimate μ_2 .

$\rho = 0.905$; $\text{Var}[z_2] = 1.05$; $\text{Var}[z_2|z_1] = 0.19$



See the Shiny dashboard [condpro](#).