Statistical foundations of machine learning

Foundations of probability with Examples

Karan Nathwani

Why probability and machine learning?

What are the advantages of adopting a probability formalism?

Probability is a convenient way of

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

If we want to generalise, we need to assume that there is something more than data.

Data are observations of a latent probabilistic process.

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 in the afternoon and takes value in {DRY,WET}.

Weather example: sample space

The associated sample space 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

Events

- 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 terminology related to sets (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 $\mathcal{E}_1$ or $\mathcal{E}_2$ or both occur.
- $\mathcal{E}_1 \cap \mathcal{E}_2 = \{\omega \in \Omega : \omega \in \mathcal{E}_1 \text{ AND } \omega \in \mathcal{E}_2\}$ 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 *disjoints* 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 $\mathcal{E}_j$, $j = 1, \ldots, n$ such that

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

Axiomatic definition of probability

- The axiomatic approach to probability consists in assigning to each event $\mathcal{E}$ a real number $\text{Prob}\{\mathcal{E}\}$ which is called the *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*.
- These conditions are the axioms of the theory of probability (Kolmogoroff,1933): they do not say anything about how probability should be interpreted.
- It follows that the probability of an event is the sum of the probability of the outcomes it comprises

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

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 (Barometer)	z_3 (Humidity)	$P(\mathbf{z}_1 = z_1, \mathbf{z}_2 = z_2, \mathbf{z}_3 = z_3)$
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 model and NOT a dataset...

Weather example: a dataset

z_1 (Sky)	z_2 (Barometer)	z_3 (Humidity)
CLOUDY	RISING	WET
CLEAR	RISING	DRY
CLEAR	RISING	DRY
CLOUDY	FALLING	WET
CLEAR	RISING	WET
CLEAR	FALLING	DRY
CLEAR	RISING	DRY
...	...	...
CLEAR	FALLING	WET
CLOUDY	RISING	DRY
CLOUDY	RISING	WET
CLEAR	RISING	DRY
CLOUDY	FALLING	DRY
CLEAR	RISING	DRY
...	...	...

Independent events

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

Exercises

1. Are two mutually exclusive events (with non null probability) independent?
2. Are two identical events (with probability smaller than one) independent?
3. Is two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent, what about their complements?

   Hint $\text{Prob}\left\{\mathcal{E}_1^c \cap \mathcal{E}_2^c\right\} = 1 - \text{Prob}\left\{\mathcal{E}_1 \cup \mathcal{E}_2\right\}$
4. Find a pair of independent events in the Weather example.
5. Let $\mathcal{E}_i$ the event of a a positive student and that the positivity of one student is independent of that of another. Compute the probability of "at least one positive in the class" if $\text{Prob}\left\{\mathcal{E}_i\right\} = 0.005$ and there are $N = 120$ students in the class.

Solutions

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. Is two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent, what about their complements?
   - Yes, 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 $\text{Prob}\{\mathcal{E}_1 \cap \mathcal{E}_2\} = \text{Prob}\{\omega_4\} = 0.1 = \text{Prob}\{\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_2\} = 0.5 * 0.2$
6. Let $\mathcal{E}_i$ the event of a a positive student and that the positivity of one student is independent of that of another. Compute the probability of "at least one positive in the class" if $\text{Prob}\{\mathcal{E}_i\} = 0.005$ and there are $N = 120$ students in the class.
   - It is the probability of the complement of the event "all negatives": $1 - \prod_{i=1}^{N}(1 - \text{Prob}\{\mathcal{E}_i\}) = 1 - (1 - \text{Prob}\{\mathcal{E}_i\})^N = 0.45$

Independent variables

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

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

- NOTA BENE: **for all** values of x and y !
- This means that we do not expect that the observation of a certain value of one variable will affect the probability of observing a certain value of the other.
- If two variables **x** and **y** are independent, then the transformed r.v. $g(\mathbf{x})$ and $h(\mathbf{y})$, where g and h are two given functions, are also independent.

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? Are the two variables independent?

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

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

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

Examples of non independent events:

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

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

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

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

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

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

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

Theorem (Law of total probability)

Let us consider a set of mutually exclusive and exhaustive events $\mathcal{E}_1$, $\mathcal{E}_2$,..., $\mathcal{E}_k$, *i.e. they form a partition of* Ω.
Let $Prob\{\mathcal{E}_i\}$, $i = 1, \ldots, k$ *denote the probabilities of the* i*th event* $\mathcal{E}_i$ *and* $Prob\{\mathcal{E}|\mathcal{E}_i\}$, $i = 1, \ldots, k$ *the conditional probability of a generic event* $\mathcal{E}$ *given that* $\mathcal{E}_i$ *has occurred.*
It can be shown that

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

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

$P(CLOUDY) = 0.35$ can be obtained in several manners:

$P(CLOUDY) = P(CLOUDY, RISING) + P(CLOUDY, FALLING) = 0.2 + 0.15$

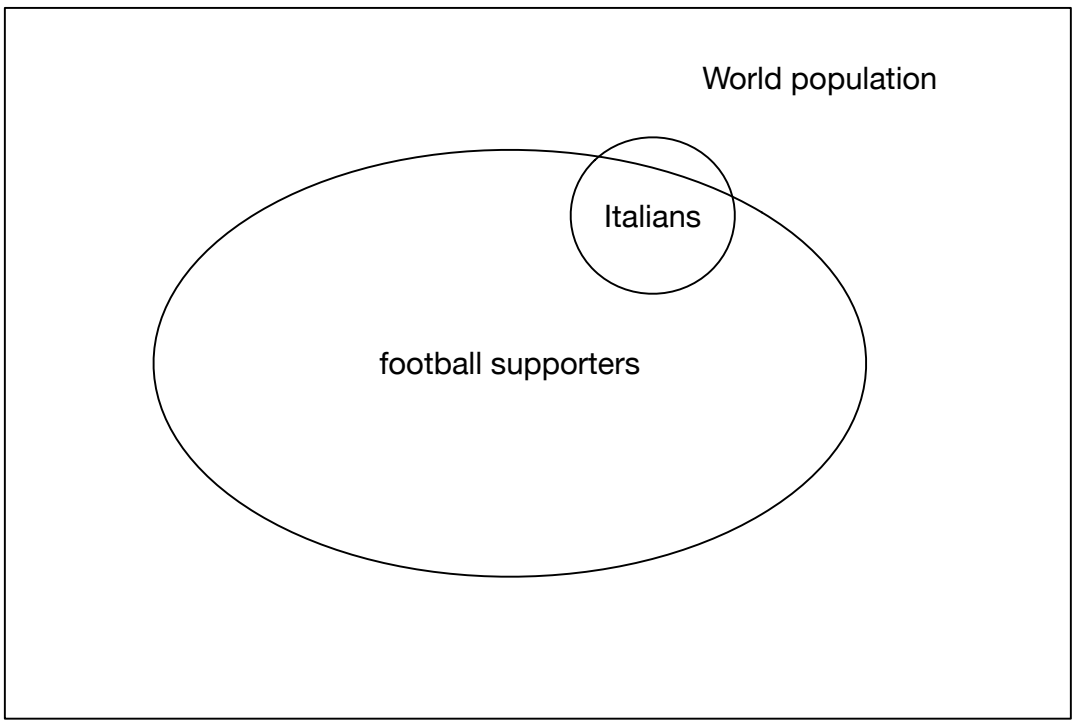
$P(CLOUDY) = P(CLOUDY, DRY) + P(CLOUDY, WET) = 0.12 + 0.23$

$P(CLOUDY) = P(CLOUDY|DRY)P(DRY) + P(CLOUDY|WET)P(WET) = 0.2 * 0.6 + 0.575 * 0.4$

Direct and inverse conditional probability

- NOTA BENE: Direct and inverse conditional probability are not the same !!
- it is generally NOT the case that Prob $\{\mathcal{E}_2|\mathcal{E}_1\} = \text{Prob}\,\{\mathcal{E}_1|\mathcal{E}_2\}$.
- Think about the probability that someone "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



P(Italian|Supporter)+ P(NOT Italian|Supporter)=1 but ...

P(Supporter| Italian)+ P(Supporter| 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

$P(CLEAR|RISING) = 0.47/0.67 = 0.701$ is different
$P(RISING|CLEAR) = 0.47/0.65 = 0.723$

$P(CLEAR|FALLING) = 0.18/0.33 = 0.545$ is different
$P(FALLING|CLEAR) = 0.18/0.65 = 0.277$

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

$P(RISING|CLEAR) + P(FALLING|CLEAR) = 1$ but
$P(CLEAR|RISING) + P(CLEAR|FALLING) \neq 1$

Conditional probabilities are probabilities...

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

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

- However 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 $\bar{\mathcal{E}}$ denotes the complement of $\mathcal{E}$ then

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

but

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

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 $\text{Prob}\{\mathbf{z} = \text{Italian}\}$ 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 $\text{Prob}\{\mathbf{z} = \text{Italian}\}$ should be better written as $\text{Prob}\{\mathbf{z} = \text{Italian}|\text{K}\}$ 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 evoution

$P(\mathcal{E})$

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

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

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

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

Medical study

Let us consider a medical study about the relationship between the outcome of a medical test and the presence of a disease. We model this study with two random attributes:

1. state of the patient: its sample space is $\Omega^s = \{H, S\}$ where H and S stand for Healthy and Sick patient, respectively.
2. outcome of the medical test: its sample space is $\Omega^o = \{+, -\}$ where $+$ and $-$ stand for positive and negative outcome of the test, respectively.

Suppose that out of $N = 1000$ patients,

	$\mathcal{E}^s = S$	$\mathcal{E}^s = H$
$\mathcal{E}^o = +$	9	99
$\mathcal{E}^o = -$	1	891

$\Rightarrow$

	$\mathcal{E}^s = S$	$\mathcal{E}^s = H$
$\mathcal{E}^o = +$	.009	.099
$\mathcal{E}^o = -$	.001	.891

What is the probability of having a positive (negative) test outcome when the patient is sick (healthy)? What is the probability of being in front of a sick (healthy) patient when a positive (negative) outcome is obtained?

Medical study (II)

From the definition of conditional probability we derive

$$\text{Prob}\{\mathcal{E}^o = +|\mathcal{E}^s = S\} = \frac{\text{Prob}\{\mathcal{E}^o = +, \mathcal{E}^s = S\}}{\text{Prob}\{\mathcal{E}^s = S\}} = \frac{.009}{.009 + .001} = .9$$

$$\text{Prob}\{\mathcal{E}^o = -|\mathcal{E}^s = H\} = \frac{\text{Prob}\{\mathcal{E}^o = -, \mathcal{E}^s = H\}}{\text{Prob}\{\mathcal{E}^s = H\}} = \frac{.891}{.891 + .099} = .9$$

According to these figures, the test appears to be accurate. Do we have to expect a similar high probability of being sick when the test is positive? The answer is NO as shown by

$$\text{Prob}\{\mathcal{E}^s = S|\mathcal{E}^o = +\} = \frac{\text{Prob}\{\mathcal{E}^o = +, \mathcal{E}^s = S\}}{\text{Prob}\{\mathcal{E}^o = +\}} = \frac{.009}{.009 + .099} \approx .08$$

Also such example shows that sometimes humans (e.g. doctors) tend to confound $\text{Prob}\{\mathcal{E}^s|\mathcal{E}^o\}$ with $\text{Prob}\{\mathcal{E}^o|\mathcal{E}^s\}$ and that the most intuitive response is not always the right one.

Bayes' theorem

Theorem (Bayes' theorem)

The conditional ("inverse") probability of any $\mathcal{E}_i$, $i = 1, \ldots, k$ given that $\mathcal{E}$ has occurred is given by

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

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

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

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

Definition (Entropy)

Given a categorical r.v. $\mathbf{z}$, the *entropy* of the probability function $P_{\mathbf{z}}(z)$ is defined by

$$H(\mathbf{z}) = -\sum_{z \in \mathcal{Z}} P_{\mathbf{z}}(z) \log P_{\mathbf{z}}(z)$$

$H(\mathbf{z})$ is a measure of the unpredictability of the r.v. $\mathbf{z}$.
Suppose that there are M possible values for the r.v. $\mathbf{z}$. The entropy is

- maximal (and takes the value $\log M$) if $P_{\mathbf{z}}(z) = 1/M$ for all z,
- is minimal iff $P(z) = 1$ for a value of $\mathbf{z}$ (i.e. all others probability values are null).

Example: conditioning reduce entropy

	BEER	WINE	
BELGIAN	0.7	0.1	0.8
ITALIAN	0.05	0.15	0.2
	0.75	0.25	

Prob $\{Beer|Italian\} = \frac{0.05}{0.2}$

$H(\text{alc}) = -\sum_{a=B,W} P(a) \log P(a) = -0.75 * \log(0.75) - 0.25 * \log(0.25) = 0.58$

$H(\text{alc}|\ \text{Italian}) = -\sum_{a=B,W} P(a|Italian) \log P(i|Italian) = 0.56$

$H(\text{alc}|\ \text{Belgian}) = -\sum_{a=B,W} P(a|Belgian) \log P(i|Belgian) = 0.37$

Average conditional entropy$= 0.2 * 0.56 + 0.8 * 0.37 = 0.408 < 0.58$

Chain rule

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 $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$,

$$\text{Prob}\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n\} =$$
$$\text{Prob}\{\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_2|\mathcal{E}_1\}\,\text{Prob}\{\mathcal{E}_3|\mathcal{E}_1, \mathcal{E}_2\}\ldots\text{Prob}\{\mathcal{E}_n|\mathcal{E}_1, \mathcal{E}_2\ldots\mathcal{E}_{n-1}\}$$

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

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

$P(CLEAR, FALLING, DRY) =$
$P(CLEAR|FALLING, DRY)P(FALLING|DRY)P(DRY) =$
$0.727 * 0.183 * 0.6$

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* $\mathbf{z} = z$ ($\mathbf{x} \perp\!\!\!\perp \mathbf{y} | \mathbf{z} = z$) iff

$$p(\mathbf{x} = x, \mathbf{y} = y | \mathbf{z} = z) = p(\mathbf{x} = x | \mathbf{z} = z) p(\mathbf{y} = y | \mathbf{z} = z) \qquad \forall x, y$$

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

Since from the chain rule

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

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

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

- In plain words, the notion of conditional dependence makes formal the intuition that a variable may bring (or not) information about a second one, according to the context.
- The statement $\mathbf{x} \perp\!\!\!\perp \mathbf{y}|\mathbf{z} = z$ means that $\mathbf{x}$ and $\mathbf{y}$ are independent if $\mathbf{z} = z$ occurs but does not say anything abut 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).


O SOLE MIO
RESTAURANT


BALLEKES


O SOLE MIO
RESTAURANT


BALLEKES

Pizza exercise

$z_1(owner)$	$z_2(cook)$	$z_3(pizza)$	$P(\mathbf{z_1} = z_1, \mathbf{z_2} = z_2, \mathbf{z_3} = z_3)$
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. $\mathbf{z}_3 \not\perp\!\!\!\perp \mathbf{z}_1$
2. $\mathbf{z}_3 \perp\!\!\!\perp \mathbf{z}_1 | \mathbf{z}_2 = \mathit{ITALIAN}$ and
3. $\mathbf{z}_3 \perp\!\!\!\perp \mathbf{z}_1 | \mathbf{z}_2 = \mathit{BELGIAN}$.

Solutions

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

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

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

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

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

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

Conditional independence and dependence

From the pizza example we see that

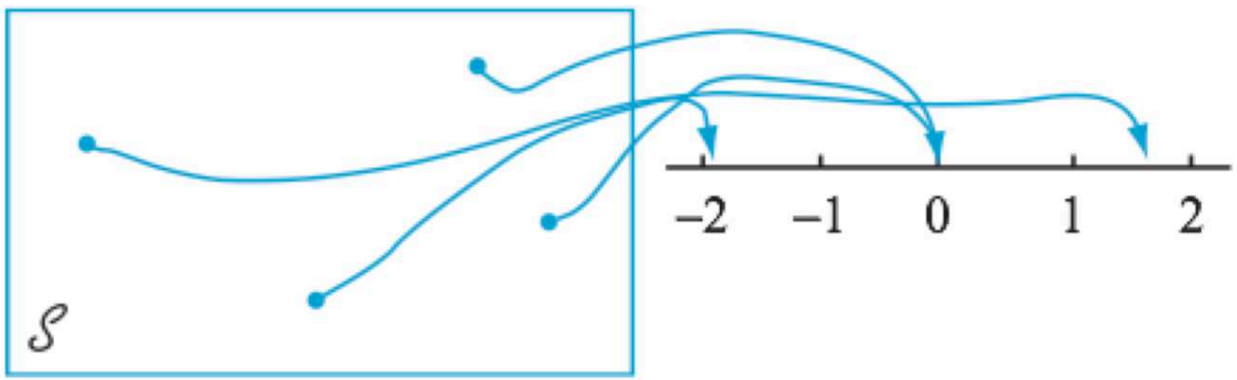
Conditional independence does not imply independence

but also

Independence does not imply conditional independence.

Think about two independent causes of a common effect

Random numerical variables



Random numerical variable is a deterministic mapping from sample space of a random experiment to a measurable space (typically real-valued).

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 $\mathbf{z}$.

Notation

- 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. 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

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

means that the dataset $D_N = \{z_1, z_2, \ldots, z_N\}$ 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 (or mass) function** of a discrete r.v. **z** is the combination of

1. the countable[1] 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 $P(z)$ that gives for each $z \in \mathcal{Z}$ the probability that **z** assumes the value z

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

This function must satify the two following conditions

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

[1] finite or in one-to-one correspondence with integers

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

- For a reduced number of possible values of **z**, the probability function can be presented in the form of a table.
- For example, if we plan to toss a *fair* coin twice, and the random variable **z** is the number of heads that eventually turn up, the probability function can be presented as follow

Values of the random variable **z**	0	1	2
Associated probabilities	0.25	0.50	0.25

Parametric probability function

Suppose that

1. $\mathbf{z}$ is a discrete r.v. that takes its value in $\mathcal{Z} = \{1, 2, 3\}$.
2. the probability function of $\mathbf{z}$ is

$$P_{\mathbf{z}}(z) = \frac{\theta^{2z}}{\theta^2 + \theta^4 + \theta^6}$$

where θ is some fixed non zero real number.

- Whatever the value of θ, $P_{\mathbf{z}}(z) \geq 0$ for $z = 1, 2, 3$ and $P_{\mathbf{z}}(1) + P_{\mathbf{z}}(2) + P_{\mathbf{z}}(3) = 1$. Therefore $\mathbf{z}$ is a well-defined random variable, even if the value of θ is unknown.
- We call θ a **parameter**, that is some constant, usually unknown involved in a probability function.
- The collection of all probability distributions for different values of the parameter is called a *family* of probability distributions.

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.
- $E[\mathbf{z}]$ is the best guess of $\mathbf{z}$ in a quadratic sense
  $\mu = \arg\min_m E[(\mathbf{z} - m)^2]$
- $E[\mathbf{z}]$ 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)

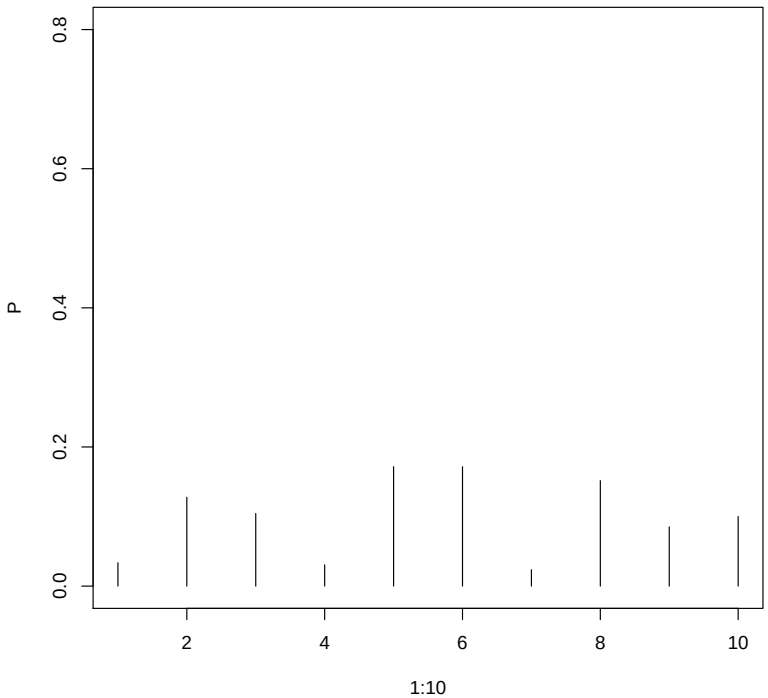
Variance of a discrete random variable $\mathbf{z}$ is

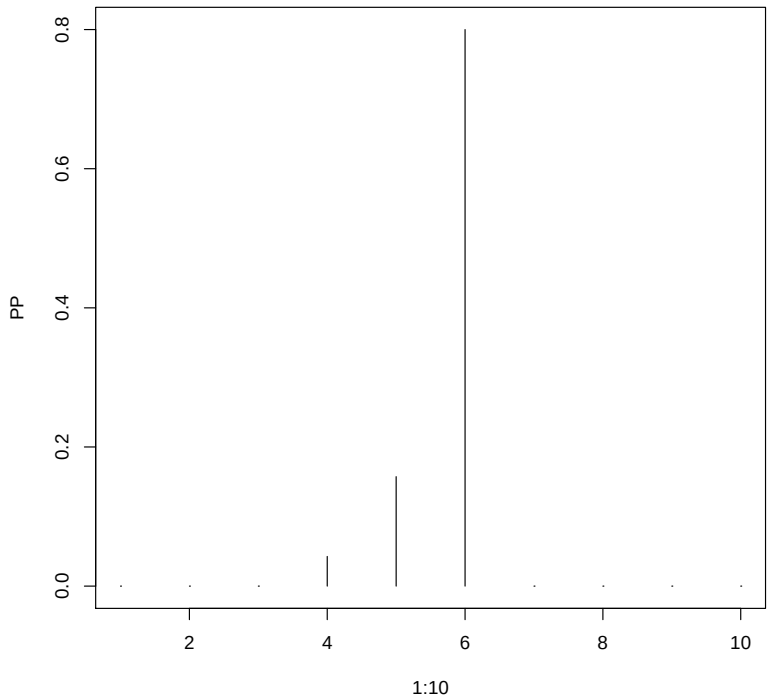
$$\text{Var}[\mathbf{z}] = \sigma^2 = E[(\mathbf{z} - E[\mathbf{z}])^2] = \sum_{z \in \mathcal{Z}} (z - E[\mathbf{z}])^2 P_{\mathbf{z}}(z)$$

- The variance is a measure of the dispersion of the probability function of the random variable around its mean.
- Note that since $(\mathbf{z} - \mu)^2 = \mathbf{z}^2 - 2\mu\mathbf{z} + \mu^2$ the following identity holds

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

Examples of probability 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 rth moment of the probability function is

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

Definition (Skewness)

Th 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

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

Definition

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

$$p_{\mathbf{z}}(z) = \frac{dF_{\mathbf{z}}(z)}{dz}$$

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

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

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):

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

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 $\mathbf{z}$.

Transformation of random variables

Let $\mathbf{z}$ a continuous r.v. and $\mathbf{y} = g(\mathbf{z})$ a function of $\mathbf{z}$, for example $g(\mathbf{z}) = \mathbf{z}^2$. What about $E[\mathbf{y}|z]$ and $E[\mathbf{y}]$?
Since g is a deterministic function it is evident that

$$E[\mathbf{y}|z] = g(z)$$

It can be shown that

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

This is also know as the law of the unconscious statistician (LOTUS).

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 $z_i \sim F_{\mathbf{z}}, i = 1, \ldots, S$,
2. computing $g(z_i)$,
3. returning the estimation

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

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

Linear combinations

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

$$E[a\mathbf{x} + b\mathbf{y}] = aE[\mathbf{x}] + bE[\mathbf{y}]$$

  i.e., expectation is a linear statistic.

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

$$\begin{aligned}\mathrm{Var}\left[a\mathbf{x} + b\mathbf{y}\right] = a^2\mathrm{Var}\left[\mathbf{x}\right] + b^2\mathrm{Var}\left[\mathbf{y}\right] + 2ab\left(E[\mathbf{xy}] - E[\mathbf{x}]E[\mathbf{y}]\right) = \\ = a^2\mathrm{Var}\left[\mathbf{x}\right] + b^2\mathrm{Var}\left[\mathbf{y}\right] + 2ab\mathrm{Cov}[\mathbf{x}, \mathbf{y}]\end{aligned}$$

  where

$$\mathrm{Cov}[\mathbf{x}, \mathbf{y}] = E\left[(\mathbf{x} - E[\mathbf{x}])(\mathbf{y} - E[\mathbf{y}])\right] = E[\mathbf{xy}] - E[\mathbf{x}]E[\mathbf{y}]$$

  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

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

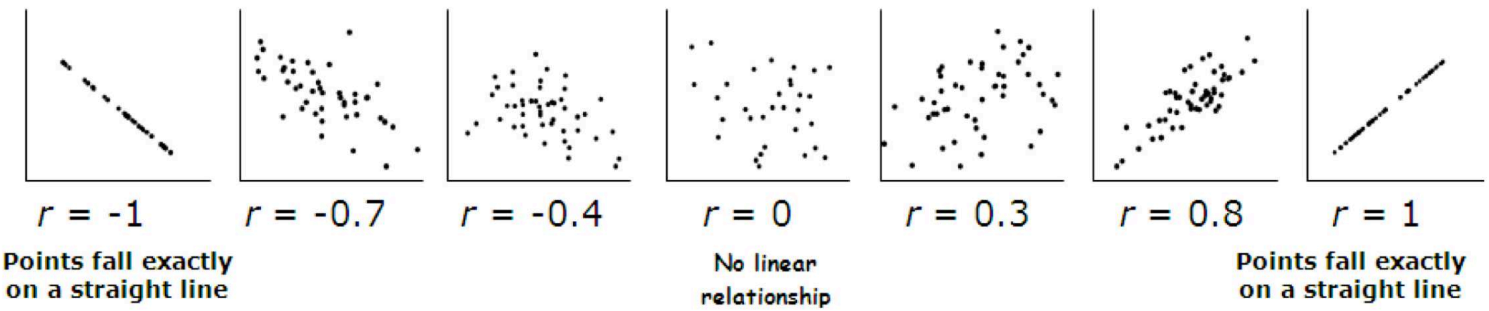
  It is easily shown that $-1 \leq \rho(\mathbf{x}, \mathbf{y}) \leq 1$.

- 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.

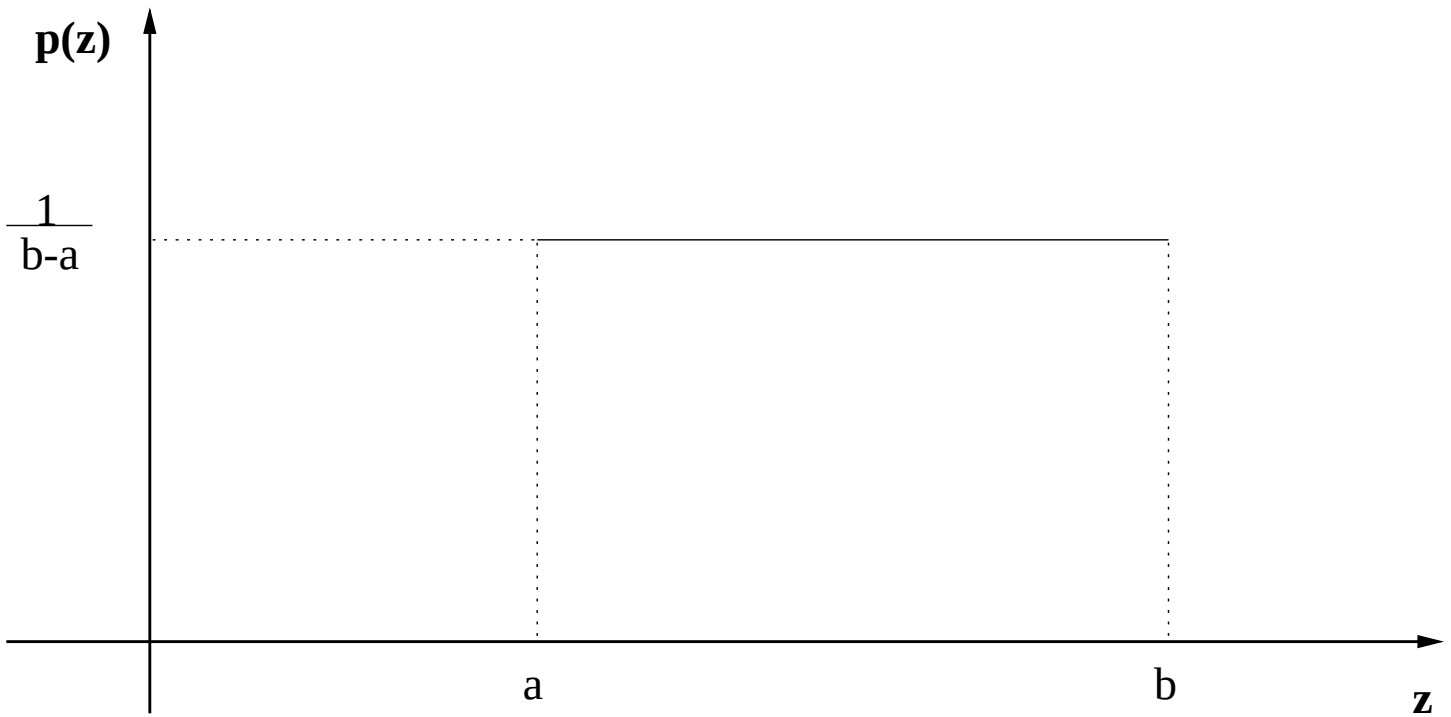
Correlation



Uniform distribution

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

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



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

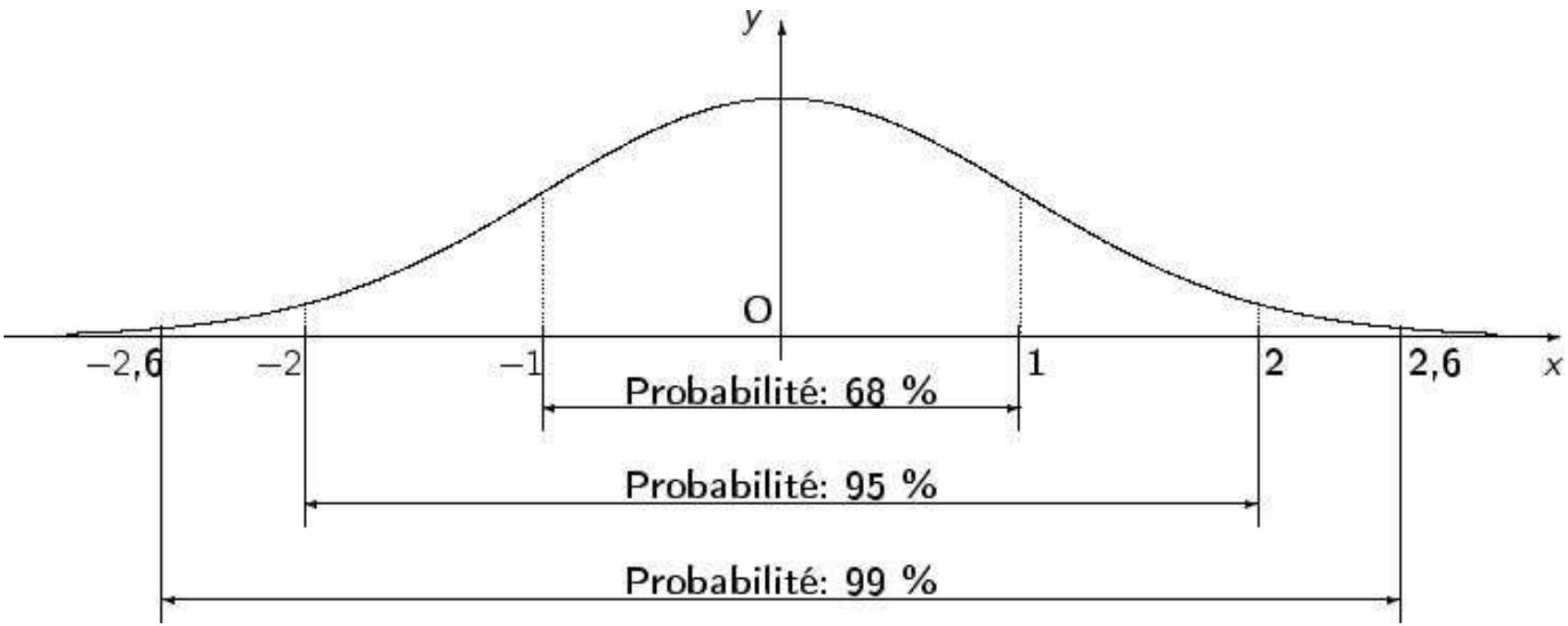
Normal distribution: the scalar case

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 from a normal r.v. is within 1 (2) standard deviations from the mean is 0.68 (0.95).
- If $\mu = 0$ and $\sigma^2 = 1$ the distribution is defined **standard normal**. We will denote its distribution function $F_{\mathbf{z}}(z) = \Phi(z)$.
- 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 $\mu = 0, \sigma^2 = 1$



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 `norm.R`

Monte Carlo exercise

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 $\mathbf{x}$.

The central limit theorem

Theorem

Assume that $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ *are i.i.d. random variables, discrete or continuous, each having a probability distribution with finite mean* μ *and finite variance* σ^2*. As* $N \to \infty$*, the standardized random variable*

$$\frac{(\bar{\mathbf{z}} - \mu)\sqrt{N}}{\sigma}$$

which is identical to

$$\frac{(\mathbf{S}_N - N\mu)}{\sqrt{N}\sigma}$$

converges in distribution to a r.v. having the standardized **normal distribution** $\mathcal{N}(0, 1)$*.*

- This result holds regardless of the common distribution of $\mathbf{z}_i$.
- This theorem justifies the importance of the normal distribution, since many r.v. of interest are either sums or averages.
- See R script `central.R` and the Shiny dashboard `random.R`.

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

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

It follows that

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

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

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

Normal multivariate distribution (II)

The quantity

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

which appears in the exponent of $p_{\mathbf{z}}$ is called 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, $i = 1, \ldots, n$ of Σ which satisfy

  $$\Sigma u_i = \lambda_i u_i \qquad i = 1, \ldots, n$$

  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

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

- 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

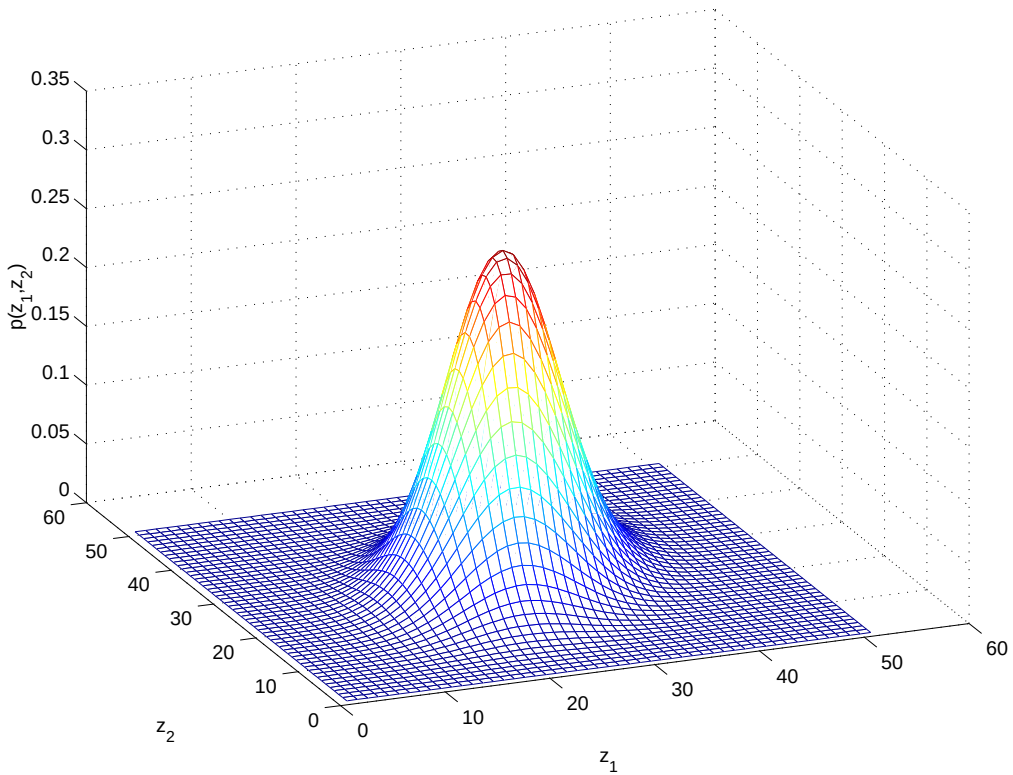
$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

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

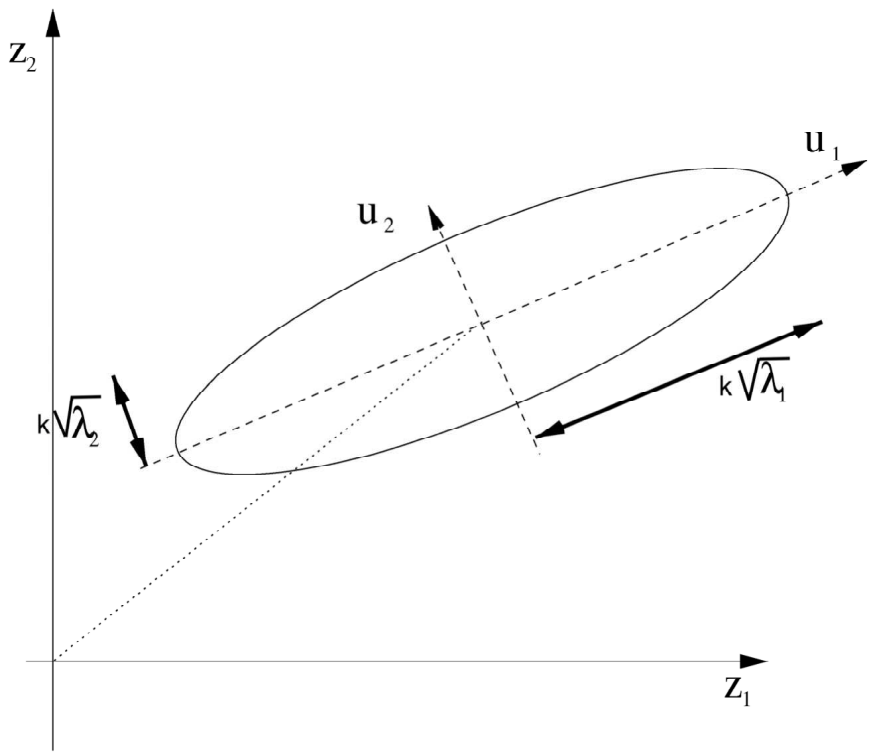
$$p(z_1, z_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$
$$\exp\left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{z_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{z_1-\mu_1}{\sigma_1}\right)\left(\frac{z_2-\mu_2}{\sigma_2}\right) + \left(\frac{z_2-\mu_2}{\sigma_2}\right)^2\right]\right]$$

Bivariate normal distribution

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



Bivariate normal distribution (prj)



Marginal and conditional distributions

One of the important properties of the multivariate normal density is that all conditional and marginal probabilities are also normal. Using the relation

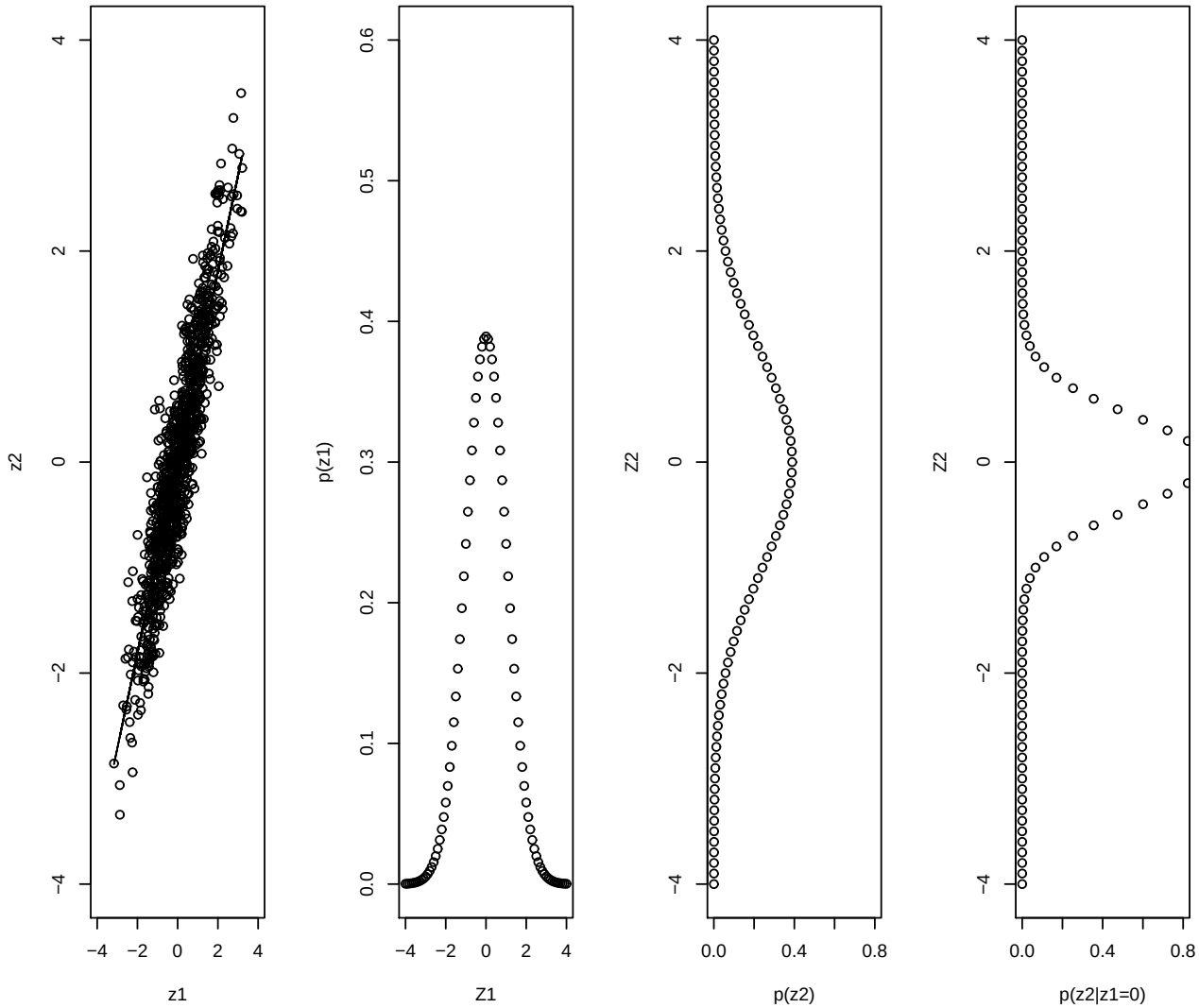
$$p(z_2|z_1) = \frac{p(z_1, z_2)}{p(z_1)}$$

we find that $p(z_2|z_1)$ is a normal distribution $\mathcal{N}(\mu_{2|1}, \sigma^2_{2|1})$, where

$$\mu_{2|1} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(z_1 - \mu_1)$$
$$\sigma^2_{2|1} = \sigma^2_2(1 - \rho^2)$$

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.



See the Shiny dashboard `condpro.R`.

Probability and miscommunication

Miscommunication of probability concepts is often the rule rather than the exception.
Consider the statement "there is a 30% probability that it will rain tomorrow".
What do people perceive from such statement about a single event?
What is its right interpretation?

1. Tomorrow it will rain 30% of the time
2. Tomorrow it will rain in 30% of this area
3. It will rain in 30% of the days that are like tomorrow

Example from the book *Reckoning with risk* by G. Gigerenzer.

Probability and miscommunication

Another common example is reduction of risk. Consider the statement "the drug D reduces the risk of death of 20 percent in patients". What does it mean?

1. Out of 100 persons, instead of 40 victims we will have 20 deaths (20% of 100 less) by administrating the drug,
2. Out of 100 persons, instead of 40 victims we will have 32 deaths (20% of 40 less)

The first is called *absolute risk reduction* while the second is the *relative risk reduction*.

Probability and miscommunication

Consider the following statement of a minister 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?

Beware of vague and inexact probabilistic (or in general mathematical) jargon !.

Gender bias In Berkeley

- In the 70's the university of Berkeley has been sued for bias against female applicants
- Though women generally outperform men at the undergraduate level in the US, on average 44% of men were accepted and 35% of women only.
- This difference is statistically significant
- However, if we breakdown the figures at the Department level the trend was inverted. The rate of acceptation in most departments was more favorable to women than men!
- Is it possible? No single department is gender biased while the entire university was!

Gender bias example: the Simpson paradox

	M=1	M=0	W=1	W=0
D1	400	100	90	10
D2	20	80	90	210
	420	180	180	220

	rate M	rate W	
D1	80%	90%	81%
D2	20%	30%	27.5%
	70%	45 %	

Overall the rate of success of men is superior to the one of women. However, for each distinct topic, women outperform men. This paradox illustrates the non monotonicity of the conditional probability

$$\text{Prob}\{M = 1\} > \text{Prob}\{W = 1\}$$

but

$$\text{Prob}\{M = 1|D1\} < \text{Prob}\{W = 1|D1\}$$

and

$$\text{Prob}\{M = 1|D2\} < \text{Prob}\{W = 1|D2\}$$

Typical pitfalls in probabilistic reasoning

- Illusion of certainty: for instance considering some tests (e.g. DNA or fingerprinting) as absolutely certain
- Wrong or unjustified estimation of probabilities (abusive manipulation of data principle of authority)
- Wrong hypothesis of independence (e.g. multiplying non-independent probabilities)
- Simpson paradox
- Prosecutor fallacy: inverting conditional probability
- Innumeracy
- Bad probabilistic model
- Vague statements.

Bad adoption of probabilistic reasoning led to major errors in some famous judicial cases.
See book *Math on trial* by L. Schneps and C. Colmez