From Statistics to Data Mining

Master 1
COlour in Science and Industry (COSI)
Cyber-Physical Social System (CPS2)
Saint-Étienne, France

Fabrice MUHLENBACH
https://perso.univ-st-etienne.fr/muhlfabr/
e-mail: fabrice.muhlenbach@univ-st-etienne.fr

Basics in Probabilities

- Definitions

- **Probability** is the branch of mathematics concerning numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true
- The probability of an event is a number between 0 and 1, where, roughly speaking, 0 indicates **impossibility** of the event and 1 indicates **certainty**

Basics in Probabilities

- Chance Experiment

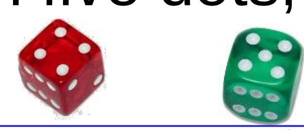
➢ A chance experiment is any activity or situation in which there is uncertainty about which of two or more possible outcomes will result

➢ Examples:

❑ Coin toss → Outcomes: heads or tails side up

❑ Card selection from a deck → Outcomes: ace of spades, five of diamonds, or one of the other 50 possibilities

❑ Red and green dice rolling → Outcomes: red die with four dots and green die with five dots, or one of the other 36 possibilities



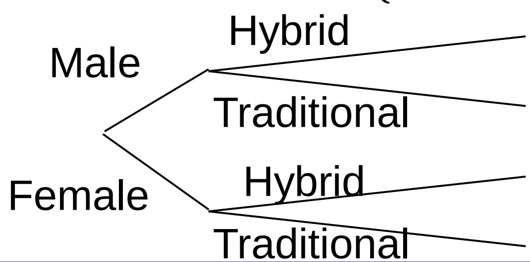
- Sample space

➢ The collection of all possible outcomes of a chance experiment is the **sample space** for the experiment

➢ Consider a chance experiment to investigate whether men or women are more likely to choose a hybrid car over a traditional internal combustion engine car

➢ The sample space can be described in two different ways:

$$sample\ space = \left\{ \begin{matrix} (male, hybrid), (female, hybrid) \\ (male, traditional), (female, traditional) \end{matrix} \right\}$$



→ Outcome: $(male, hybrid)$

Basics in Probabilities

- Event

➢ An **event** is any collection of outcomes from the sample space of a chance experiment

➢ A simple event is an event consisting of exactly one outcome

➢ Two events that have no common outcomes are said to be disjoint or mutually exclusive

Basics in Probabilities

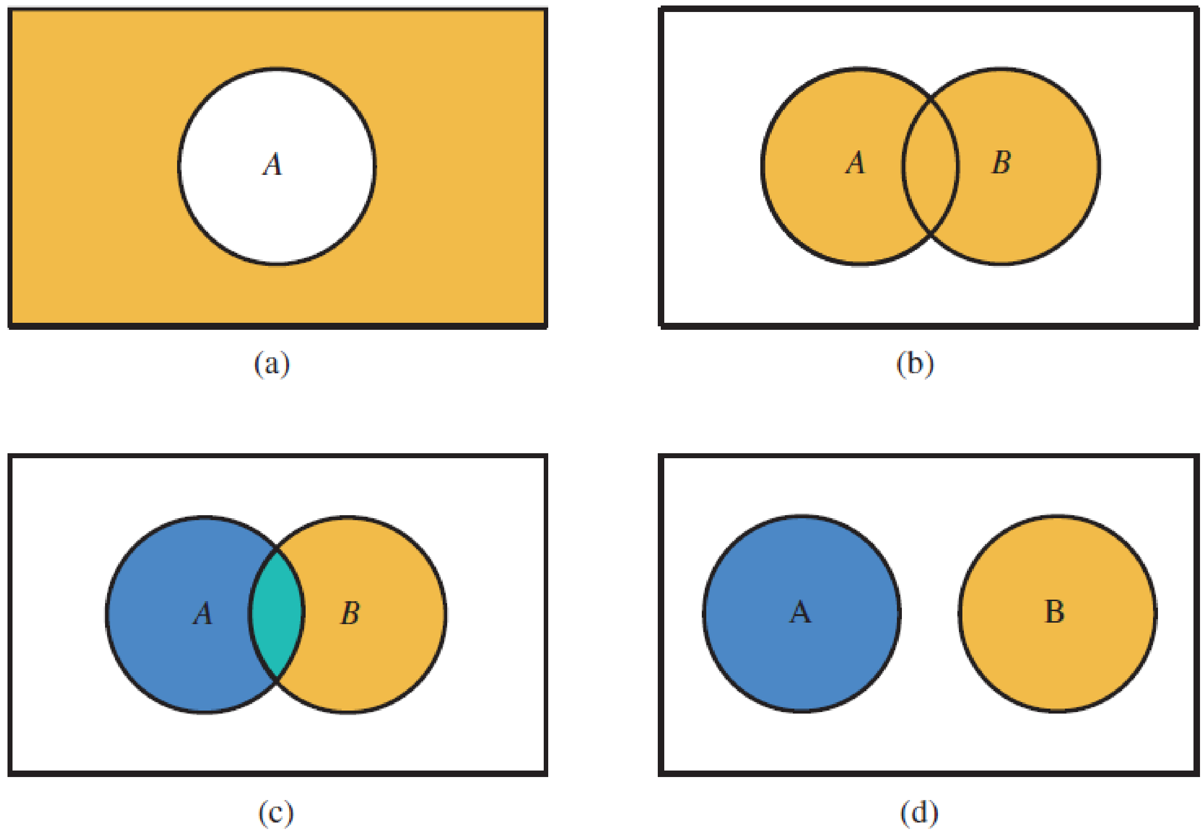
- Event

a) gold region = **not** *A*

b) gold region = *A* **or** *B*

c) green region = *A* **and** *B*

d) two **disjoint** events



Basics in Probabilities

• Probability

- When the outcomes in the sample space of a chance experiment are **equally likely**, the probability of an event *E*, denoted by *P*(*E*), is the ratio of the number of outcomes favorable to *E* to the total number of outcomes in the sample space:

$$P(E) = \frac{\text{number of outcomes favorable to } E}{\text{number of outcomes in the sample space}}$$

- Example: Chance experiments that involve tossing fair coins, rolling fair die (but not the sum of 2 dice!), or selecting cards from a well-mixed deck have equally likely outcomes

Basics in Probabilities

- Law of Large Numbers

- We are aware that chance experiments and observations do not always give the same results when repeated and that even in the most carefully replicated chance experiment, there is **variation**
- Examples: it is easy to imagine a fair coin landing heads up on only 3 or 4 of 10 tosses
- As the number of repetitions of a chance experiment increases, the chance that the relative frequency of occurrence for an event will differ from the true probability of the event by more than any small number approaches 0

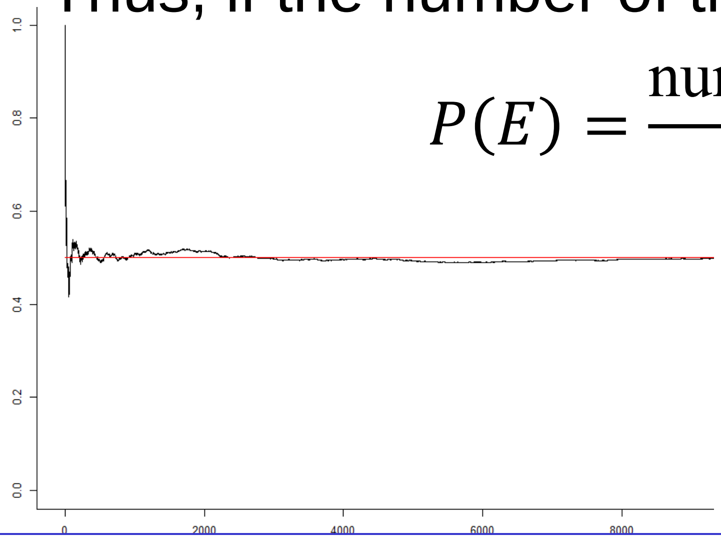
Basics in Probabilities

- Relative Frequency Approach to Probability

➢ The probability of an event *E* (possibly unknown), denoted by *P*(*E*), is defined to be the value approached by the relative frequency of occurrence of *E* in a very long series of trials of a chance experiment

➢ Thus, if the number of trials is quite large,

$$P(E) = \frac{\text{number of times } E \text{ occurs}}{\text{number of trials}}$$



→ Stabilization of the relative frequency of heads in coin tossing to 0.5

Basics in Probabilities

- Probabilistic model / Statistical model

- A statistical model is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population)
- A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variables
- A probabilistic model is a statistical model which incorporate probability distribution(s)

Basics in Probabilities

• Conditional Probability

- ➢ difficulty in synthesizing uncertain reasoning
  e.g., diagnosis of toothache
- ▪ Toothache ⇒ Dental caries
- ○ problem 1: wrong rule because there are other possible origins
- ▪ Toothache ⇒ Caries ∨ Gingivitis ∨ Abscess ...
- ○ problem 2: almost unlimited list of possible causes
- ○ transformation of the causal rule:
- ▪ Caries ⇒ Toothache
- ○ but still incorrect rule: some cavities are not painful

Basics in Probabilities

• Conditional Probability

- ➢ probabilistic propositions concern possible worlds and the set of possible worlds is the universe
- ➢ a complete probabilistic model associates a numerical value $P(\omega)$ to each possible world ω its probability
- ➢ the sum of the probabilities of all possible worlds is 1
- ➢ a proposition is associated with all the possible worlds in which it is true
- ➢ we distinguish the unconditional probabilities, or *a priori*, from the conditional probabilities, or *a posteriori*, in the case of probabilities depending on a known information: $P(a \mid b)$ is the (conditional) probability of a knowing b

Basics in Probabilities

• Conditional Probability

➢ e.g., the probability of having a dental caries when going to the dentist for a check-up is 0.2: $P(caries) = 0.2$

→ this is an unconditional probability

➢ on the other hand, going to the dentist for a toothache has a different value: $P(caries \mid toothache) = 0.6 > 0.2$

→ it is a conditional probability

➢ the unconditional probability is not invalidated by the conditional probability, it simply becomes less useful

Basics in Probabilities

• Conditional Probability

- conditional probabilities can be defined in terms of unconditional probabilities by the following equation:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

- product rule: $P\ (a \wedge b) = P\ (a \mid b)\ P\ (b)$

Basics in Probabilities

• Conditional Probability and Bayes' Rule

- the product rule: $P(a \wedge b) = P(a \mid b)P(b)$ and $P(a|b) = P(a \wedge b) / P(b)$ lead to the following formula by dividing by $P(a)$:
- $P(b \mid a)$ = (P ($a \mid b$) P (b)) / (P (a)):

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

- this equation is known as **Bayes' rule**:

$$P(cause|effect) = \frac{P(effect|cause)P(cause)}{P(effect)}$$

- Bayes' theorem underlies most modern AI systems for probabilistic inference

Basics in Probabilities

• Conditional Probability and Bayes' Rule

- interest of Bayes' theorem:

$$P(cause|effect) = \frac{P(effect|cause)P(cause)}{P(effect)}$$

- *P*(*effect* | *cause*) measures the relationship between a cause and one of its possible consequences → causal direction
- *P*(*cause* | *effect*) measures the relationship between a consequence and one of its possible causes → diagnostic
- e.g., in medical practice one often knows the conditional causal probabilities (i.e., the doctor knows *P*(*symptoms* | *disease*)) and he wishes to derive a diagnosis *P*(*disease* | *symptoms*)

Basics in Probabilities

- Total Probability Theorem

Let A_i, $i = 1,2,\ldots, M$, be the events so that $\sum_{i=1}^{M} p(A_i) = 1$.

Then the probability of an arbitrary event B is given by:

$$p(B) = \sum_{i=1}^{M} p(B \mid A_i).p(A_i)$$

where $p(B \mid A)$ denotes the conditional probability of B assuming A (= “the conditional probability of B, given A” or “the probability of B under the condition A”) which is defined as:

$$p(B \mid A) = \frac{p(B, A)}{p(A)}$$

and $p(B, A)$ is the joint probability of the two events.

$$\Rightarrow p(B \mid A).\, p(A) = p(A \mid B).\, p(B)$$

Basics in Probabilities

• The Bayesian Method and the Bayes Rule

- ➢ The Bayesian learning consists of detecting the optimal class $y^* \in Y$ of an example ω, given its feature vector $X(\omega)$.
- ➢ **Theorem**

$$\forall y_j \in Y,\ p(y_j \mid X(\omega)) = \frac{p(X(\omega)|y_j).p(y_j)}{p(X(\omega))}$$

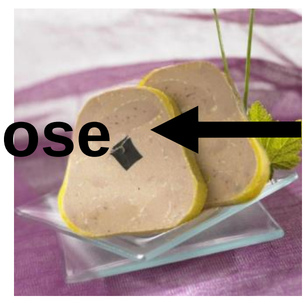
- ➢ This theorem is very powerful because it uses *a priori* information to take an *a posteriori* decision.
- ➢ We deduce that $y^*(\omega) = \arg\max_c p(y_c \mid X(\omega))$

  which means $y^*(\omega) = \arg\max_c p(X(\omega) \mid y_c).p(y_c)$

Basics in Probabilities

- Exercise

A student in data mining speaks with his uncle, expert in ornithology. The student asks: “How many types of birds there are on this lake?” His uncle answers that only gooses and swans come down on this lake, but they are very similar.



Goose

?

Swan

The expert gives the following information to the student: “Swans are 3 times more numerous than gooses. Moreover, given a dark bird, 9 times out of 10 it is a goose, while this occurs only one time out of 20 for swans.”

Basics in Probabilities

- Exercise

The expert gives the following information to the student:
"Swans are 3 times more numerous than gooses. Moreover, given a dark bird, 9 times out of 10 it is a goose, while this occurs only one time out of 20 for swans."

Seeing a new bird landing on the lake, the student claims:
"I bet you, at 6-to-1, it's a goose".

Can you justify this claim?

Basics in Probabilities

- Conditions to solve this exercise

If this calculation is possible, it is optimal from a probabilistic point of view.
However, some hypotheses are assumed:

1) The *a priori* probabilities of the different classes are known:
p(swan) and p(goose)

2) The probabilities of the observations given the classes are also known:
p(dark bird | swan) and p(dark bird | goose)

Without any knowledge *a priori*, this requires to estimate these two quantities.

Basics in Probabilities

- Estimation of the *a priori* probability of classes $p(y_j)$

Without any information on the domain, we assume that they are equivalent such that:

$$p(y_j) = 1 / C$$

where C is the number of classes.

We assume that the learning set has been drawn from the target probability distribution and so we use the frequencies of each class such that

$$p(y_j) = |LS_j| / |LS|$$

where $|LS_j|$ is the number of examples in the class y_j

Basics in Probabilities

- Estimation of the conditional probabilities $p(X(\omega)|y_j)$

We can distinguish two types of estimates:

1) The parametric methods which assume that $p(X(\omega)|y_j)$ follow a given statistical distribution.
In this case, the problem to solve boils down to the estimation of the parameters of the considered distribution (normal distribution, for instance, with μ and σ)

2) The non-parametric methods which do not impose any constraint on the underlying distributions, and for which the densities $p(X(\omega)|y_j)$ are locally estimated around $X(\omega)$.

→ Parzen windows and the nearest-neighbor algorithm

Basics in Probabilities

- Solution of the exercise – Reminder

A student in data mining speaks with his uncle, expert in ornithology. The student asks: “How many types of birds there are on this lake?” His uncle answers that only gooses and swans come down on this lake, but they are very similar.

Goose ←  ?  → Swan

The expert gives the following information to the student: “Swans are 3 times more numerous than gooses. Moreover, given a dark bird, 9 times out of 10 it is a goose, while this occurs only one time out of 20 for swans.”

Seeing a new bird landing on the lake, the student claims: “I bet you, at 6-to-1, it’s a goose”. Can you justify this claim?

Basics in Probabilities

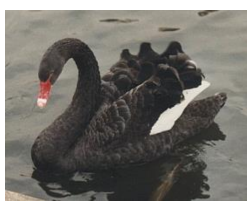
- Solution of the exercise

How many classes? Two classes

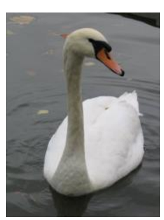
y_1 = Goose

y_2 = Swan

How many features? One feature with two values

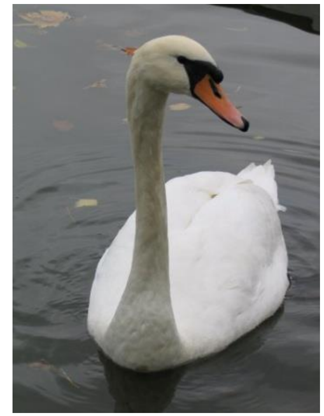



A_1 = Dark bird

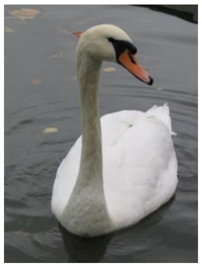
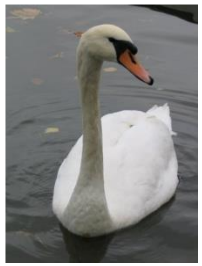
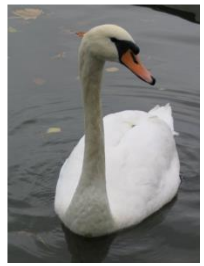
A_2 = Light bird

Basics in Probabilities

	Goose	Swan	
Light			p(light)
Dark			p(dark)
	p(goose)	p(swan)	

Basics in Probabilities

"Swans are 3 times more numerous than gooses".

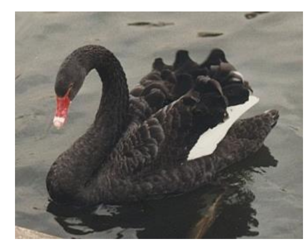
p(swan) = 3/4 p(goose) = 1/4

"Given a dark bird…"

"9 times out of 10 it is a goose…"

p(dark | goose) = 9/10

"while this occurs only one time out of 20 for swans."



p(dark | swan) = 1/20

Basics in Probabilities

	Goose	Swan	
Light	$p(L \mid G) = 1/10$	$p(L \mid S) = 19/20$	$p(L)$
Dark	$p(D \mid G) = 9/10$	$p(D \mid S) = 1/20$	$p(D)$
	$p(G) = 1 / 4$	$p(S) = 3 / 4$	

Basics in Probabilities

- Solution of the exercise

$p(D)$?

$p(L)$?

Total probability theorem:

$$p(B) = \sum_{i=1}^{M} p(B \mid A_i).p(A_i)$$

$$\rightarrow \quad p(D) = p(D \mid G).p(G) + p(D \mid S).p(S)$$

$$p(D) = 9/10 * 1/4 + 1/20 * 3/4 = 9/40 + 3/80 = 21/80$$

$$\rightarrow \quad p(L) = p(L \mid G).p(G) + p(L \mid S).p(S)$$

$$p(L) = 1/10 * 1/4 + 19/20 * 3/4 = 1/40 + 57/80 = 59/80$$

Basics in Probabilities

	Goose	Swan	
Light	$p(L \mid G) = 1/10$	$p(L \mid S) = 19/20$	$p(L)=59/80$
Dark	$p(D \mid G) = 9/10$	$p(D \mid S) = 1/20$	$p(D)=21/80$
	$p(G) = 1 / 4$	$p(S) = 3 / 4$	

Basics in Probabilities

We want to know the probability to be a swan or a goose by knowing the color (light or dark) of the bird.

What are the probabilities $p(S|L)$, $p(G|L)$, $p(S|D)$ and $p(G|D)$?

Bayes Rule:

$$\forall y_j \in Y,\ p(y_j \mid X(\omega)) = \frac{p(X(\omega)|y_j).p(y_j)}{p(X(\omega))}$$

➡ $p(S \mid L) = p(L \mid S).p(S) \,/\, p(L)$

$p(S \mid L) = ((19/20) * (3/4)) \,/\, (59/80) = 57 \,/\, 59$

➡ $p(G \mid L) = p(L \mid G).p(G) \,/\, p(L)$

$p(G \mid L) = ((1/10) * (1/4)) \,/\, (59/80) = 2 \,/\, 59$

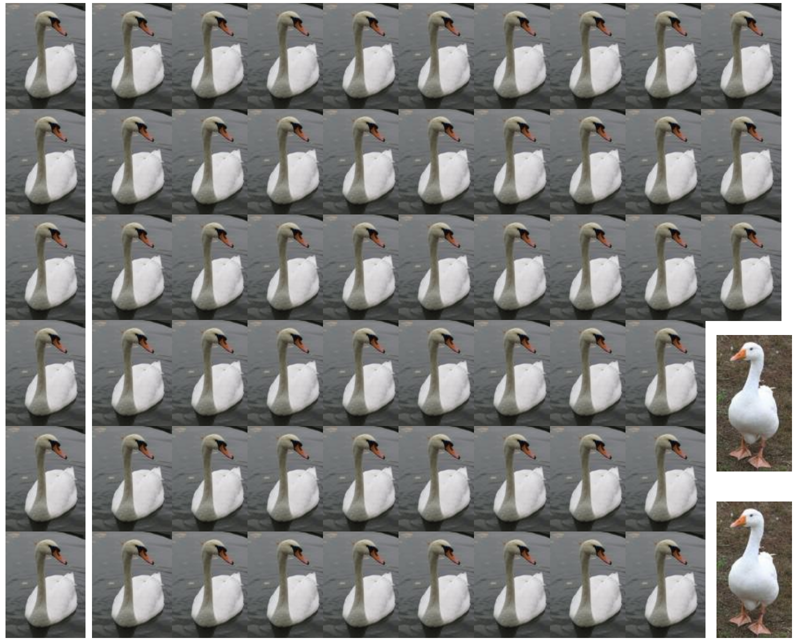
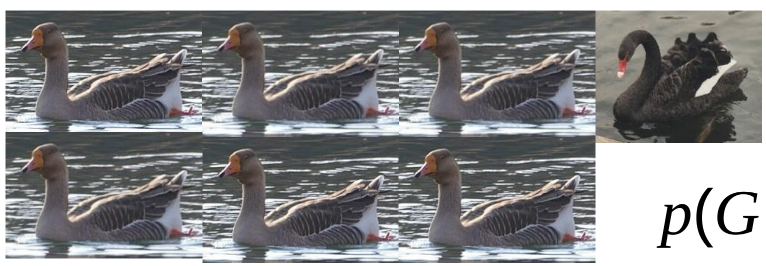
➡ $p(S \mid D) = p(D \mid S).p(S) \,/\, p(D)$

$p(S \mid D) = ((1/20) * (3/4)) \,/\, (21/80) = 3/21 = 1 \,/\, 7$

➡ $p(G \mid D) = p(D \mid G).p(G) \,/\, p(D)$

$p(G \mid D) = ((9/10) * (1/4)) \,/\, (21/80) = 18/21 = 6 \,/\, 7$

Basics in Probabilities

	Total Population		arg_max
Light bird		$p(S \mid L)$ = 57 / 59 $p(G \mid L)$ = 2 / 59	light bird → swan risk=2/59 of error (3.39%)
Dark bird		$p(S \mid D)$ = 1 / 7 $p(G \mid D)$ = 6 / 7	dark bird → goose risk=1/7 of error (14.3%)

- Exercise: Conclusion

Seeing a new bird landing on the lake, the student claims:
"I bet you, at 6-to-1, it's a goose".
Can you justify this claim?

The new bird is landing on the lake

What is its color?

It's a dark bird?
It's a goose!
(6-to-1 bet)

It's a light bird?
It's a swan!
(57-to-2 bet)

The Bayesian Classifier:

- if it's a dark bird
  → we can predict that it's a goose with a risk of 1/7
- if it's a light bird
  → we can predict that it's a swan with a risk of 2/59

Basics in Probabilities

- The Bayesian Method with Numerical attributes

➢ **Example:**

The height of the people in the classroom (and relatives)

→ Goal: design a model of the gender by knowing the height.

- Collect the data
- Draw graphs
- Calculate: mean (μ), variance (σ^2) and std deviation (σ)
- Assume that it is a normal distribution

  Gaussian Probability Density Function:

$$p(x) = \frac{1}{\sqrt{2\,\pi\,\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\,\sigma^2}\right)$$

- Extend to a Bayes classification rule

Basics in Probabilities

- The Bayesian Method with Numerical attributes

The Bayes classification rule (for two classes ω_1 and ω_2, $M = 2$)

- Given $\underline{x}$ classify it according to the rule

$$\boxed{\begin{array}{l} If\ P(\omega_1|\underline{x}) > P(\omega_2|\underline{x})\ \ \underline{x} \to \omega_1 \\ If\ P(\omega_2|\underline{x}) > P(\omega_1|\underline{x})\ \ \underline{x} \to \omega_2 \end{array}}$$

- Equivalently: classify $\underline{x}$ according to the rule

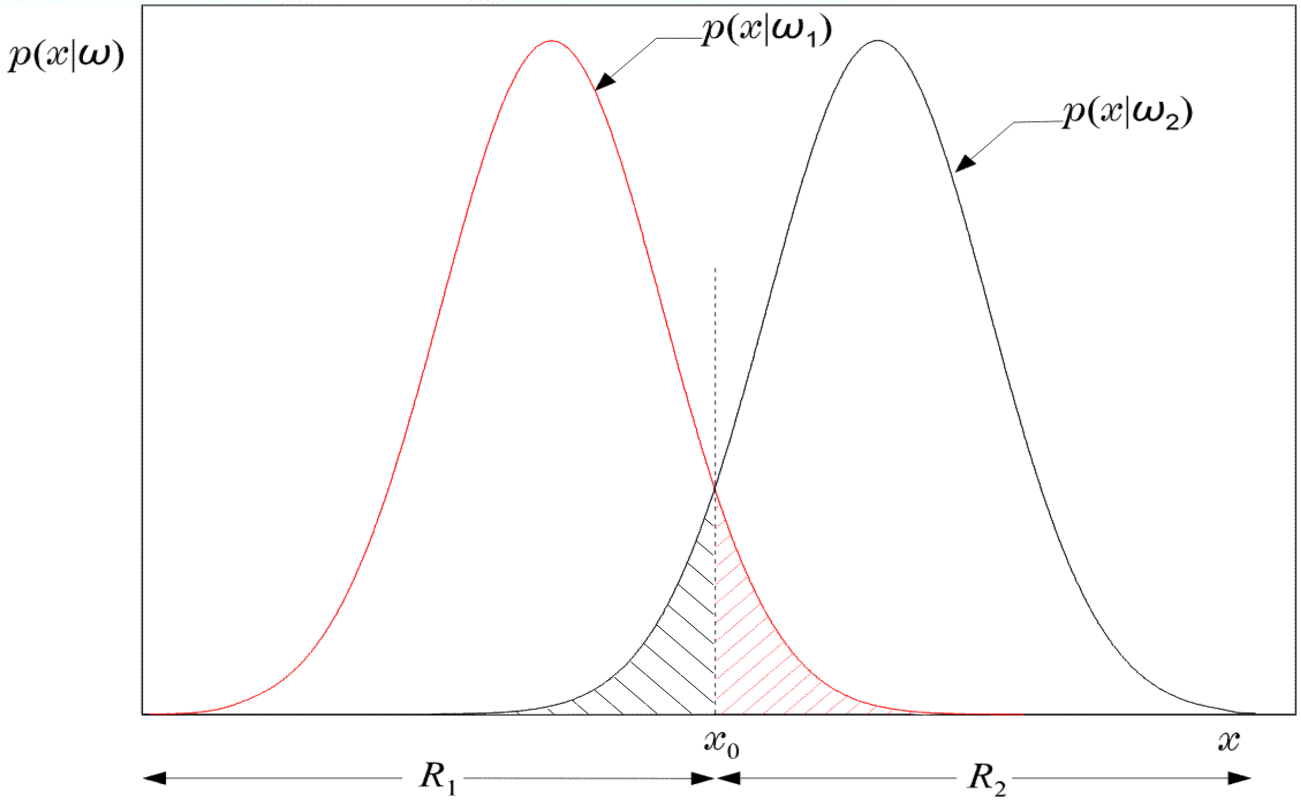
$$\boxed{p(\underline{x}|\omega_1)P(\omega_1)(><)p(\underline{x}|\omega_2)P(\omega_2)}$$

- For equally likely classes the test becomes

$$\boxed{p(\underline{x}|\omega_1)(><)P(\underline{x}|\omega_2)}$$

Basics in Probabilities

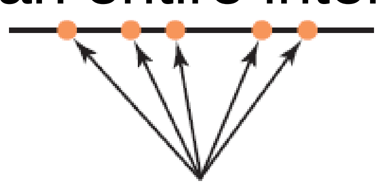
- The Bayesian Method with Numerical attributes



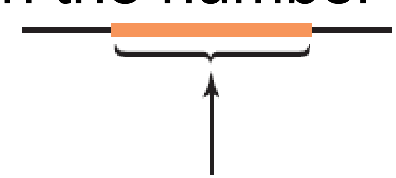
$$R_1(\rightarrow \omega_1) \ and \ R_2(\rightarrow \omega_2)$$

• Random Variables

- ➢ A random variable is a a numerical variable X whose value depends on the outcome of a chance experiment
- ➢ A random variable associates a numerical value with each outcome of a chance experiment
- ➢ A random variable X is discrete if its set of possible values x is a collection of isolated points along the number line
- ➢ A random variable X is continuous if its set of possible values x includes an entire interval on the number line



Possible values of a discrete random variable



Possible values of a continuous random variable

Basics in Probabilities

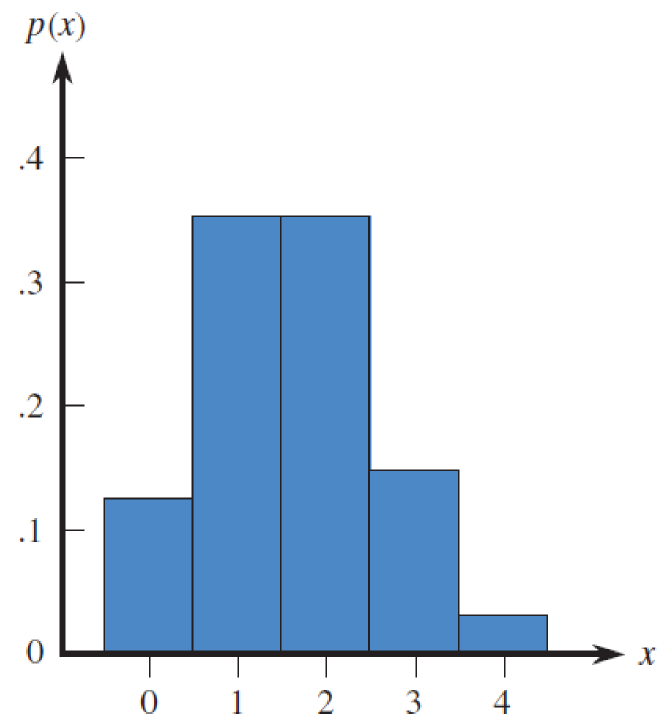
• Probability Distribution for Discrete Random Variables

- ➢ The probability distribution of a discrete random variable X gives the probability $P(X = x)$ (or $p(x)$ for sake of simplicity) associated with each possible x value
- ➢ Each probability is the long-run relative frequency of occurrence of the corresponding x value when the chance experiment is performed a very large number of times
- ➢ Properties of Discrete Probability Distributions:
- ❑ for every possible x value, $0 \leq p(x) \leq 1$
- ❑ the sum for all x possible values $p(x) = 1$

Basics in Probabilities

- Probability Distribution for Discrete Random Variables

➢ A pictorial representation of a discrete probability distribution is called a probability histogram



Basics in Probabilities

- Random Variables and Probability Law

- ➢ Random variable → dimension varying according to the result of a random experiment
- ➢ e.g., the toss of a coin with 1 for "heads"
- ➢ another example, the roll of two balanced dice: possible pairs = {{1,1}, {1,2},…, {6,6}} where each event has the same probability of occurrence $p(\omega)$ = 1 / 36
- ➢ sum of the points marked by the dice: possible results = 2, 3, 4,…, 12 with different probabilities of appearance: the result “2” appears only once out of 36: {1, 1} whereas the result “7” can appear 6 times out of 36: {1, 6}, {2, 5}, {3, 4}, {4, 3}, {5, 2} or {6, 1}

Basics in Probabilities

- Random Variables and Probability Law

- **probability law** of X = image of P by X and denoted by P_X
- for a discrete variable (i.e., only able to take a finite number of values), the P_X law is made up of point masses and can be represented by a bar chart (e.g., the throw of two dice)
- **distribution function** of a random variable X:
  application of F from $\mathbb{R}$ to [0; 1] defined by: $F(x) = P(X < x)$
- practical importance of the distribution function:
  allows to calculate the probability of any interval of $\mathbb{R}$:
  $$P(a \leq X < b) = F(b) - F(a)$$
- continuous variable → variable with a probability density

Basics in Probabilities

• Random Variables, Probability Law and Moments

- ➢ a probability law can be characterized by certain typical values associated with the notions of central value, dispersion and shape of the distribution, known as “moments”
- ➢ **expected value** (or mean value):
- ❑ for a discrete variable, the expected value $E(x)$ is defined by: $E(x) = x_i . P(X = x_i)$
- ❑ for a continuous variable admitting a density, $E(x)$ is the value, if the integral converges, of $\int_{\mathbb{R}} x . f(x) \mathrm{d}x$
- ➢ additivity of expected values: $E(X_1 + X_2) = E(X_1) + E(X_2)$

Basics in Probabilities

- Random Variables, Probability Law and Moments

➢ **variance**:

❑ the variance of X, denoted by $V(X)$ or σ^2, is the quantity

$$\sigma^2 = E\left((x-m)^2\right) = \int_{\mathbb{R}} (x-m)^2 dP_X(x) \qquad \text{where } m = E(x)$$

❑ the variance is the moment of order 2 of the distribution

❑ the variance is a measure of the dispersion of X around m

❑ σ is the standard deviation (= the square root of the variance)

❑ we call **covariance** of X and Y the quantity:

$$\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y) = E\left(\left(X - E(X)\right)\left(Y - E(Y)\right)\right)$$

Basics in Probabilities

- Random Variables, Probability Law and Moments

➢ study of the **correlation** between two or more random variables or numerical statistics = study of the strength of the link that may exist between these variables (measurement of the linear dependence between two variables *X* and *Y*)

➢ **correlation coefficient**: coefficient equal to the ratio of the covariance of two variables and the non-zero product of their standard deviations → $\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$

➢ the correlation coefficient is between -1 and 1

➢ warning: the fact that two variables are "strongly correlated" does not demonstrate that there is a causal relationship between one and the other (“*cum hoc ergo propter hoc*”)

Basics in Probabilities

- Probability Distributions

- ➢ examples: uniform discrete distribution, Benoulli distribution with parameter *p*, binomial distribution, Poisson distribution...
- ➢ Laplace-Gauss distribution (also called “Normal distribution”): continuous probability distribution which plays a fundamental role in probabilities and mathematical statistics → appears as the limiting law of characteristics linked to a large sample
- ➢ X follows a Normal distribution $\mathscr{LG}(m\,;\sigma)$ or $\mathscr{N}(m\,;\sigma)$ if its density is

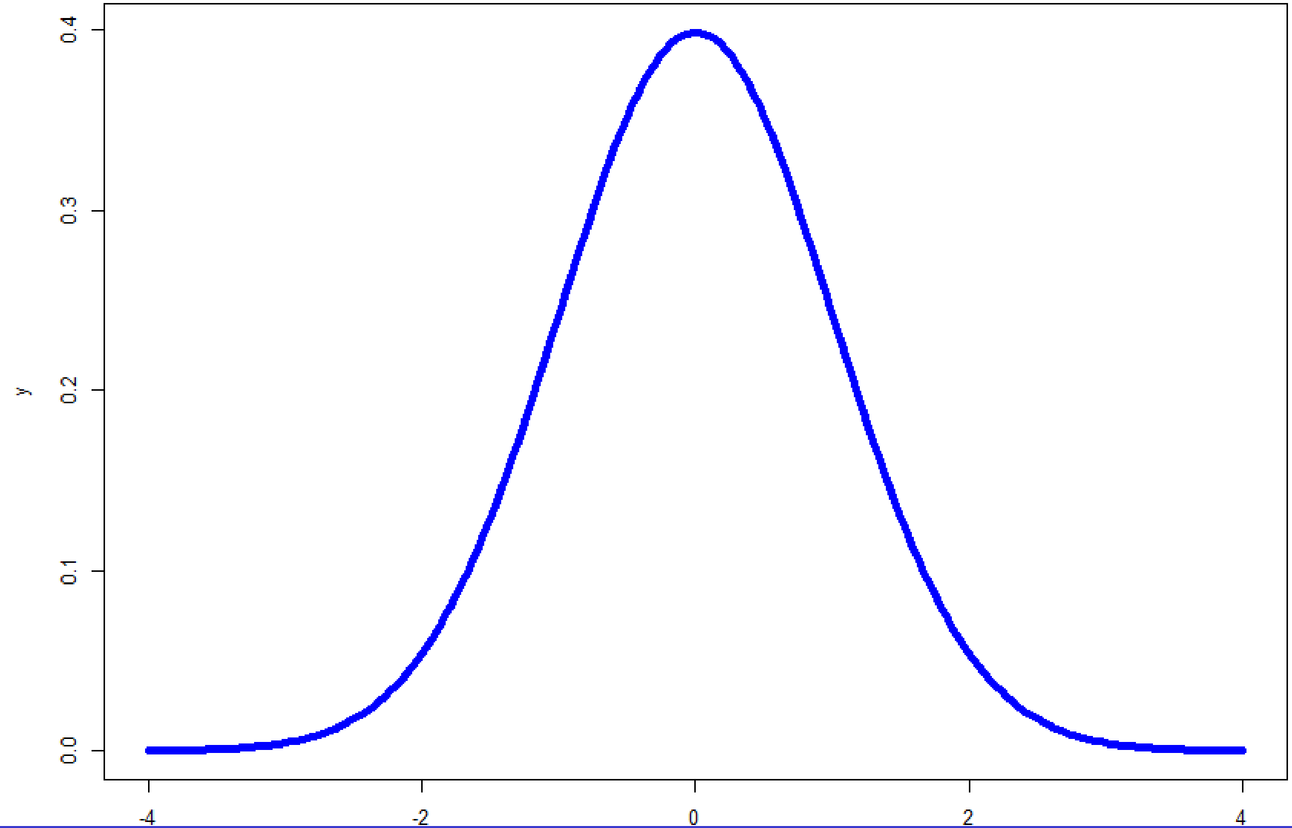
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right)$$

- ➢ as a result of the symmetry of f and since the integral of X converges: $E(X) = m$
- ➢ change of random variable: $U = (X - m) / \sigma$

Basics in Probabilities

- Probability Distributions

➢ Normal Distribution: density of X

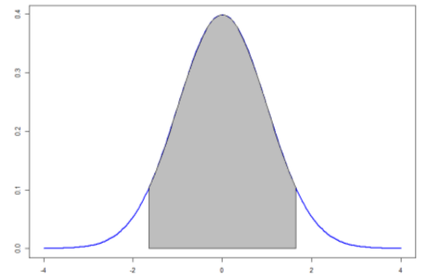


Basics in Probabilities

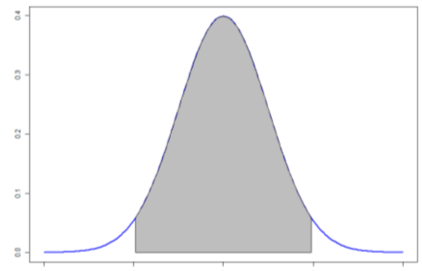
- Probability Distributions

➢ Normal Distribution: some interesting values

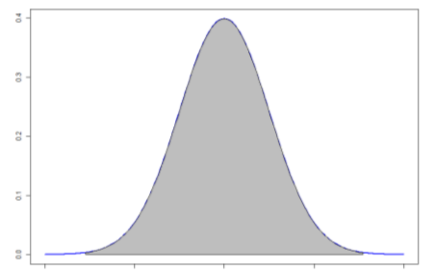
$$P(m - 1.64\sigma < X < m + 1.64\sigma) = 0.90$$



$$P(m - 1.96\sigma < X < m + 1.96\sigma) = 0.95$$



$$P(m - 3.09\sigma < X < m + 3.09\sigma) = 0.998$$



Basics in Probabilities

- Exercise

➢ A normal distribution with mean μ = 3500 grams and standard deviation σ = 600 grams is a reasonable model for the probability distribution of the continuous variable X : birth weight of a randomly selected full-term baby

➢ **Question 1:** What proportion of birth weights are between 2900 and 4700 grams?

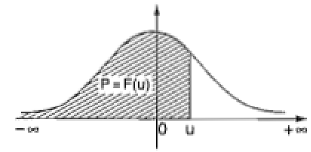
➢ the direct calculation of such probabilities (with the areas under a normal curve) is not easy

➢ to overcome this difficulty, we rely on the table of the distribution function of the reduced normal distribution

Basics in Probabilities

- Exercise

➢ use of a table of the distribution function of the reduced normal distribution (= probability of finding a value less than *u*)



u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de *u*

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
F(u)	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

Basics in Probabilities

- Exercise

➢ the birth weight of a newborn (any sex combined) follows a normal law with an average μ = 3500 grams and a standard deviation σ = 600 grams

➢ what is the proportion of birth weight between 2900 and 4700 grams?

➢ $$P(2900 < X < 4700) = P\left(\frac{2900-3500}{600} < \frac{X-\mu}{\sigma} < \frac{4700-3500}{600}\right)$$
$$= P\left(-1 < \frac{X-\mu}{\sigma} < 2\right)$$
$$= P(Z < 2) - P(Z < -1)$$

➢ we look in the table: $P(Z < 2)$ gives 0.9772 and $P(Z < -1)$ gives 1 – 0.8413 = 0.1587

Basics in Probabilities

- Exercise

- the birth weight of a newborn (any sex combined) follows a normal law with an average μ = 3500 grams and a standard deviation σ = 600 grams
- what is the proportion of birth weight between 2900 and 4700 grams?
- we look in the table: $P(Z < 2)$ gives 0.9772 and $P(Z < -1)$ gives 1 – 0.8413 = 0.1587
- so $(2900<X<4700)=P(Z<2)-P(Z<-1)=0.9772-0.1587=0.8185$
- therefore, the proportion of birth weight between 2900 and 4700 grams is 81.85%

Basics in Probabilities

- Exercise

➢ A normal distribution with mean μ = 3500 grams and standard deviation σ = 600 grams is a reasonable model for the probability distribution of the continuous variable X : birth weight of a randomly selected full-term baby

➢ **Question 2:** What birth weight w is exceeded only 2.5% of the time?

Basics in Probabilities

- Exercise

- $P(X > w) = 0.025$
- $\Leftrightarrow P\left(Z > \frac{w-x}{\sigma}\right) = 0.025$
- $\Leftrightarrow P\left(Z > \frac{w-3500}{600}\right) = 0.025$
- $\Leftrightarrow 1 - P\left(Z < \frac{w-3500}{600}\right) = 0.025$
- $\Leftrightarrow P\left(Z < \frac{w-3500}{600}\right) = 0.975$
- we look in the table at the value corresponding to a probability of 0.9750: it is 1.96
- $\Leftrightarrow w = 1.96 \times 600 + 3500 = 4676$ → a weight of 4676 grams

Bibliography

- ❑ Peck R., C. Olsen, and J. L. Devore (2016). *Introduction to Statistics and Data Analysis*, 5th edition, Boston: Cengage Learning
- ❑ On YouTube: *StatQuest* with Josh Starmer:
- ❖ Bayes' Theorem
  https://www.youtube.com/watch?v=ONCOkccpk3w
- ❖ Conditional Probability
  https://www.youtube.com/watch?v=iiN_J9S0KLM