Statistics for Business and Machine learning

Tien-Nam Le

tien-nam.le@ens-lyon.fr

Sciences U University
18 Octobre, 2018

Outline

Part I. Descriptive Statistics and Basic Probability

Descriptive Statistics

1. Population, sample and data
2. Key parameters, eg. mean, variance, covariance
3. Data interpretation and presentation

Basic Probability

1. Random variable (discrete and continuous)
2. Distributions: Binomial, Poisson, Normal, Exponential
3. Basic rules in probability and Central Limit Theorem

Outline

Part II. Inferential Statistics

Point estimators

1. Unbiased estimators
2. Consistence and efficiency

Confidence intervals

1. Confidence level and confidence intervals
2. Interpretation of the result

Hypothesis testing

1. Null/alternative hypothesis
2. Type I, type II errors, significance level
3. Student's t-distribution

Outline

Part III. Statistical design of experiments

Linear Regression

1. prediction and assessing model fit
2. Transforming to linear model
3. $\hat{\beta}_1$ and $\hat{\beta}_2$ and hypothesis testing

Analysis of Variance (ANOVA)

1. 1-, 2-, 3-way ANOVA
2. F-statistics and F-distribution

Advanced techniques

1. Regression trees and classification trees
2. Polynomial and Logistic Regression
3. Spatial Statistics and Time Series Analysis
4. Prior Information and Bayesian Inference

Part I. Descriptive Statistics and Basic Probability

Example

- *Population* (*population*): All customers who buy your product.
- *Échantillon* (*sample*): a set of customers who do the survey, each customer is a sample point (or data point)
- *Donnée* (*data*): for each sample point:
  - age
  - city
  - salary
  - satisfaction about the product(1-10)

  Note: Different samples give different data

Types of data:

- *Donnée quantitative* (*quantitative data*): numbers, either discrete (eg. age, satisfaction) or continuous (eg. salary).
- *Donnée qualitative* (*Qualitative data*): categorical (eg. city)

Statistics

Statistique (*Statistics*): a method to get information from data

1. Collection of data
2. Analysis of data
3. Interpretation of data to reach conclusions
4. Presentation of data

- *Statistique descriptive* (*descriptive statistics*) (Part I): draw conclusions about the data
- *Statistique inférentielle* (*inferential statistics*) (Part II): infer conclusions about entire population

Example:

- Average customer satisfaction of data: Descriptive statistics
- Average customer satisfaction of whole population: Inferential statistics.

Key parameters

Example: customer satisfaction data: $X = \{4, 8, 8, 9, 10, 9, 1, 6, 9, 7\}$
Formally, we write

- sample data: $X = \{X_1, ..., X_n\}$ with n sample points
- (unknown) population data: $\{x_1, ..., x_N\}$ with N elements

Key *paramètres* (*parameters*):

- *Moyenne* (*mean*):
  - population: $\mu = \frac{\sum x_i}{N}$
  - échantillon: $\overline{X} = \frac{\sum X_i}{n} = 7.1$

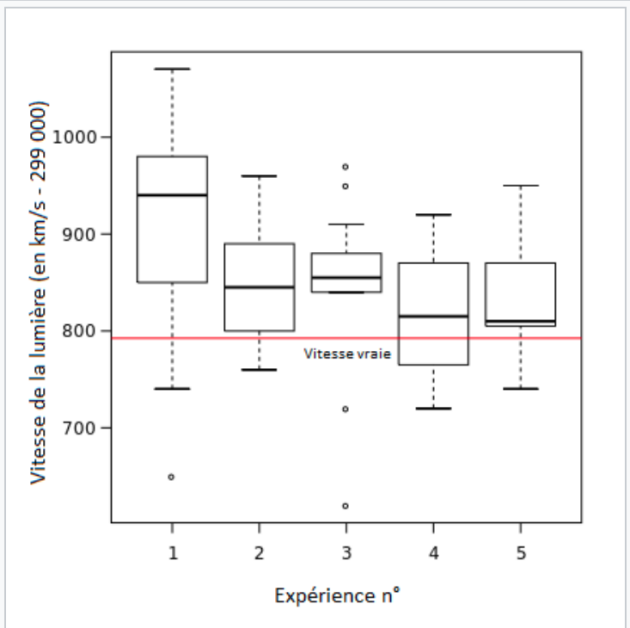
- *Médiane* (*median*): element in the middle (at the position 50%) of the sorted data.
  - If n is odd: median = the middle
  - If n is even: (left middle + right middle)/2

  Eg. Sorted data $= \{1, 4, 6, 7, 8, 8, 9, 9, 9, 10\}$ donc $median(X) = (8 + 8)/2 = 8$.

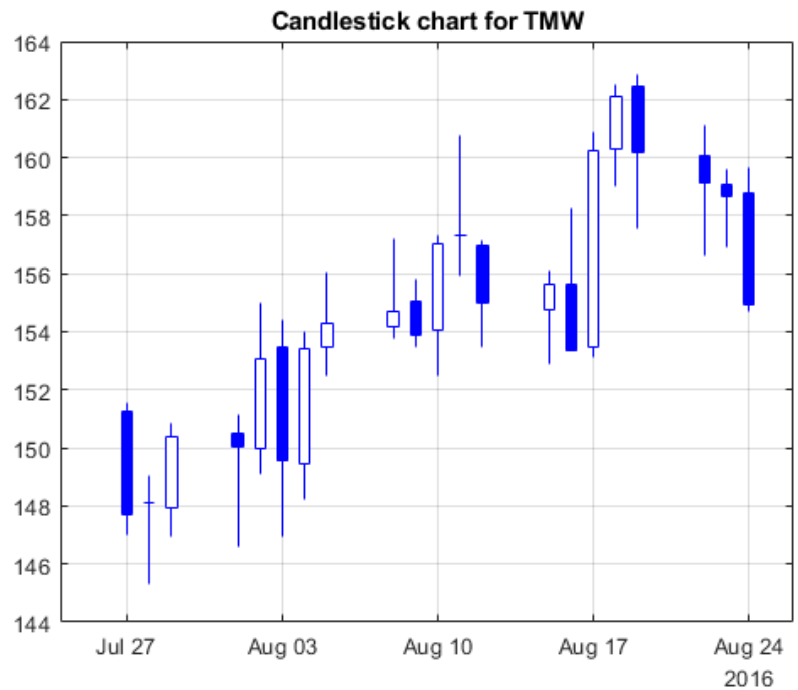
- *Mode* (*mode*): most frequent elements: Eg. 9

Key parameters

- *Étendue* (*range*): $\max - \min = 10 - 1 = 9$
- *Donnée aberrante* (*outlier*) qui est distante des autres. Eg. 1.
- 1^e *quartile* (*lower quartile*): median of lower half of data (i.e. at 25% of the sorted data. Eg. 6 in sorted data $\{1, 4, 6, 7, 8, 8, 9, 9, 9, 10\}$
- 3^e *quartile* (*upper quartile*): median of lower half of data (at 75% of the sorted data). Eg. 9.
- *Boîte à moustache* (*box-plot*). Don't mess up with candlestick chart.



Boîte à moustache issue des données obtenues grâce à l'expérience de Michelson-Morley. Il y a 4 données aberrantes dans la colonne du milieu et 1 dans la 1^{re} colonne.



Key parameters

Recall: sample data: $X = \{X_1, ..., X_n\}$ and population data: $\{x_1, ..., x_N\}$.

- *Variance* (*variance*):
  - population: $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$
  - échantillon: $Var(X) = s^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$
- *Écart type* (*standard deviation*):
  - population: $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$
  - échantillon: $s = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n-1}}$
- *Cote Z* (*Standard score*): $z_i = \frac{x_i - \sigma}{\mu}$

Covariance

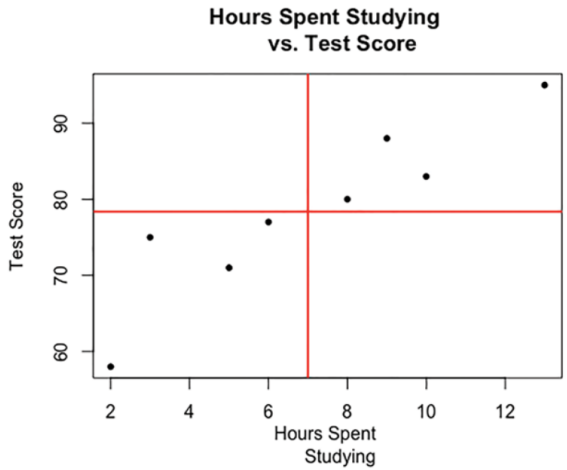
Example:

- Hours Studied $X = \{2, 3, 5, 6, 8, 9, 10, 13\}$
- Test results $Y = \{58, 75, 71, 77, 80, 88, 83, 95\}$
- *Covariance* (*covariance*):

$$Cov(X, Y) = \sum_i \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{n - 1}$$

- $Cov(X, X) = Var(X)$
- $Cov(X, Y) = Cov(Y, X)$

In this case, $Cov(X, Y) = 38$. What does it mean ?
Is it large or small ?



Corrélation

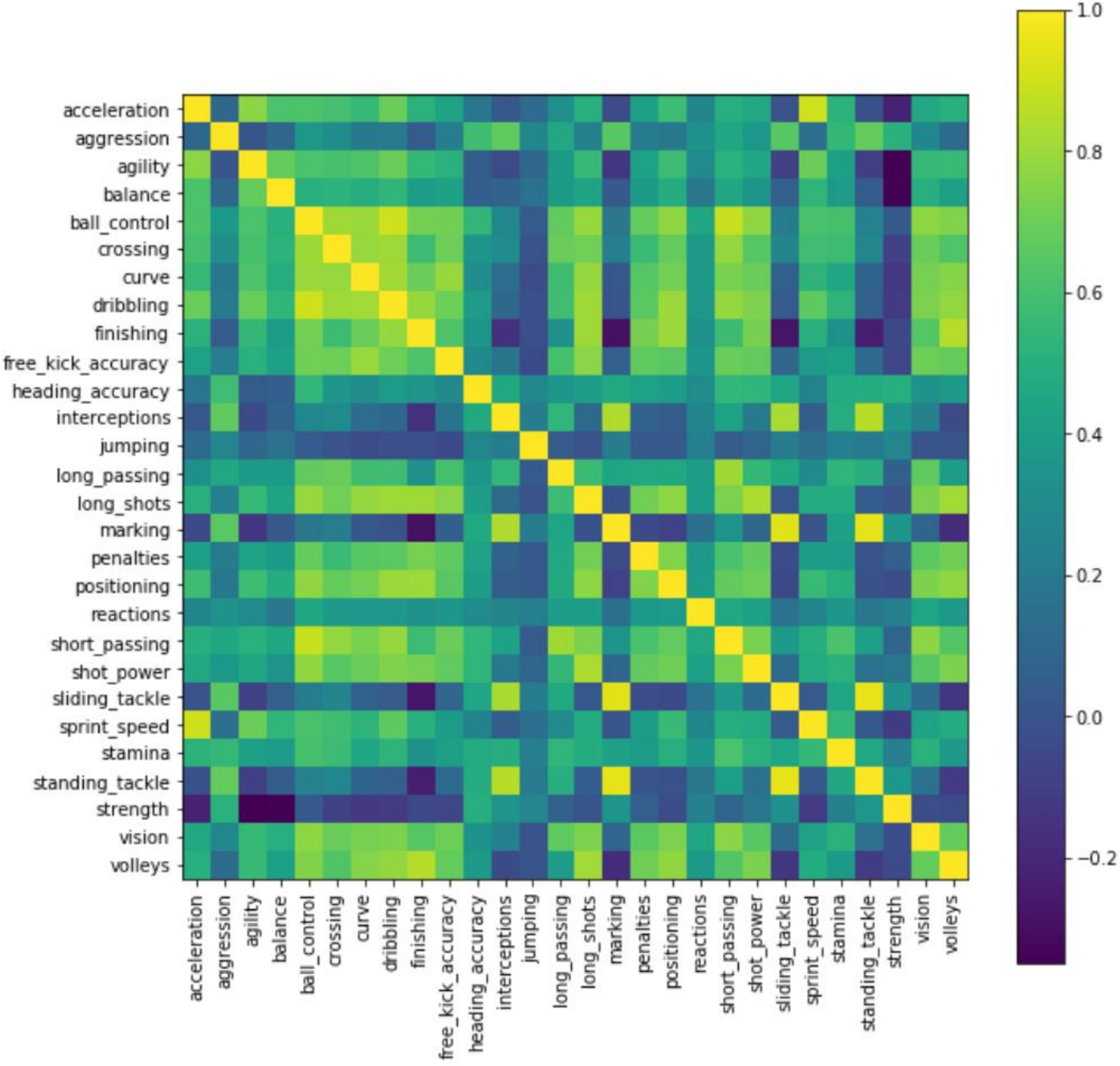
Corrélation (*correlation*):

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X, Y)}{s_X s_Y}$$

In this case: $Cor(X, Y) = \frac{38}{3.70 \times 11.19} = 0.92$, very strong positive relationship.

- $-1 \leq Cor(X, Y) \leq 1$
- $Cor(X, X) = 1$
- 1 or -1: linear relationship, does not implies causalation
- 0 does not implies there is no relation (i.e does not implies X and Y are independent)

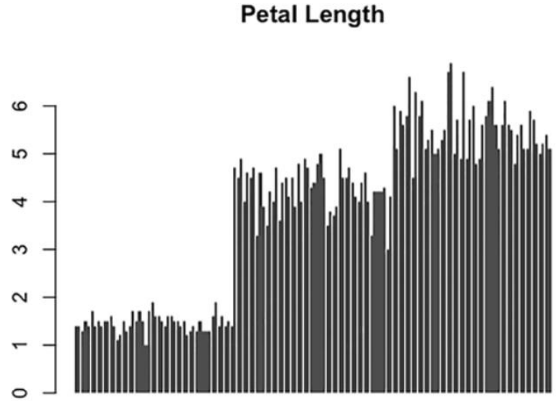
Example of Corrélation

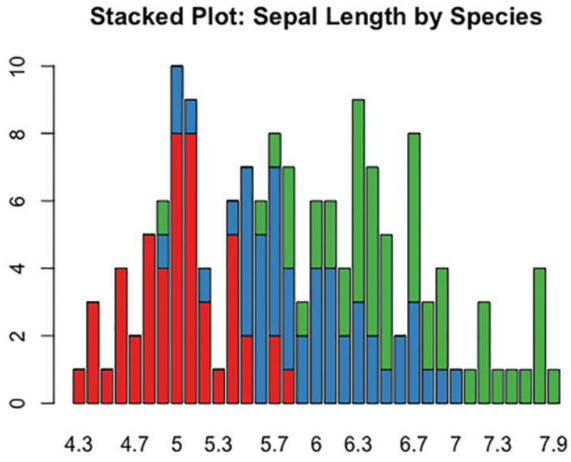


Data presentation

Iris Data: 150 data points, 3 species
numerical data: Sepal length, sepal width, petal length, petal width

Bar plot and **Stacked plot**

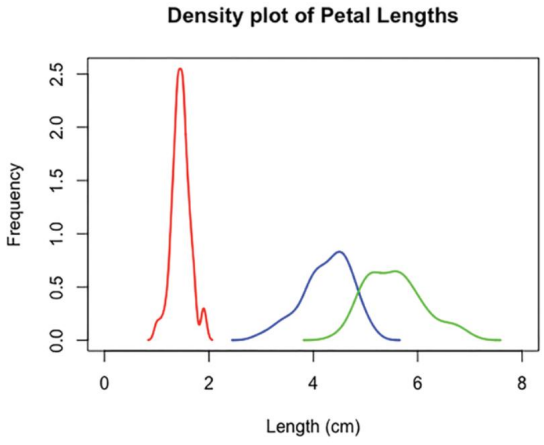


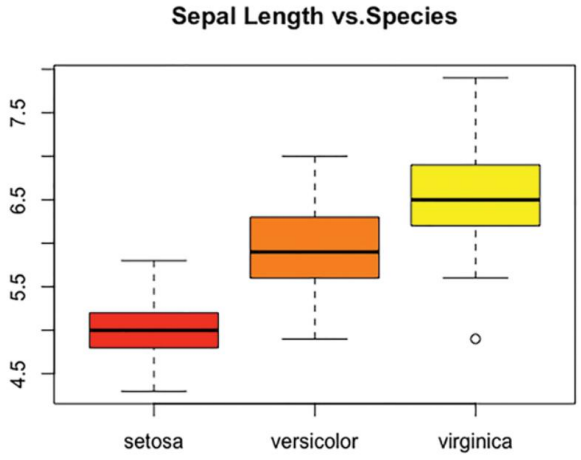


Histogram

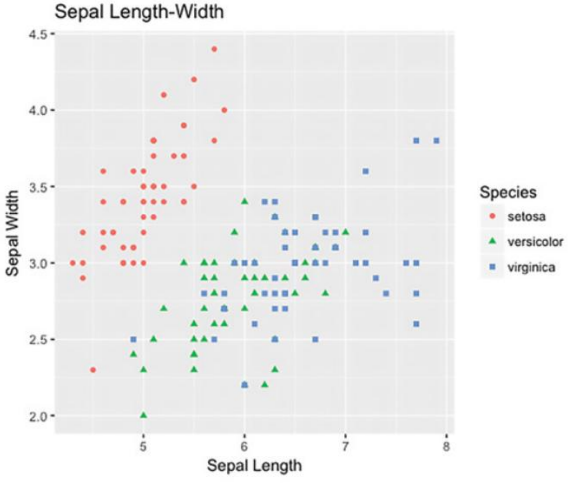
Data presentation

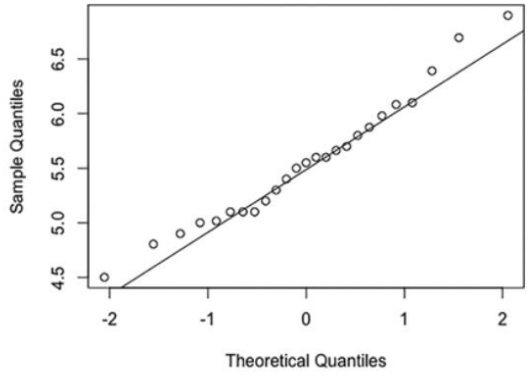
Density plot and **Box plot**





Scatter plot and **Quantile-Quantile plot**

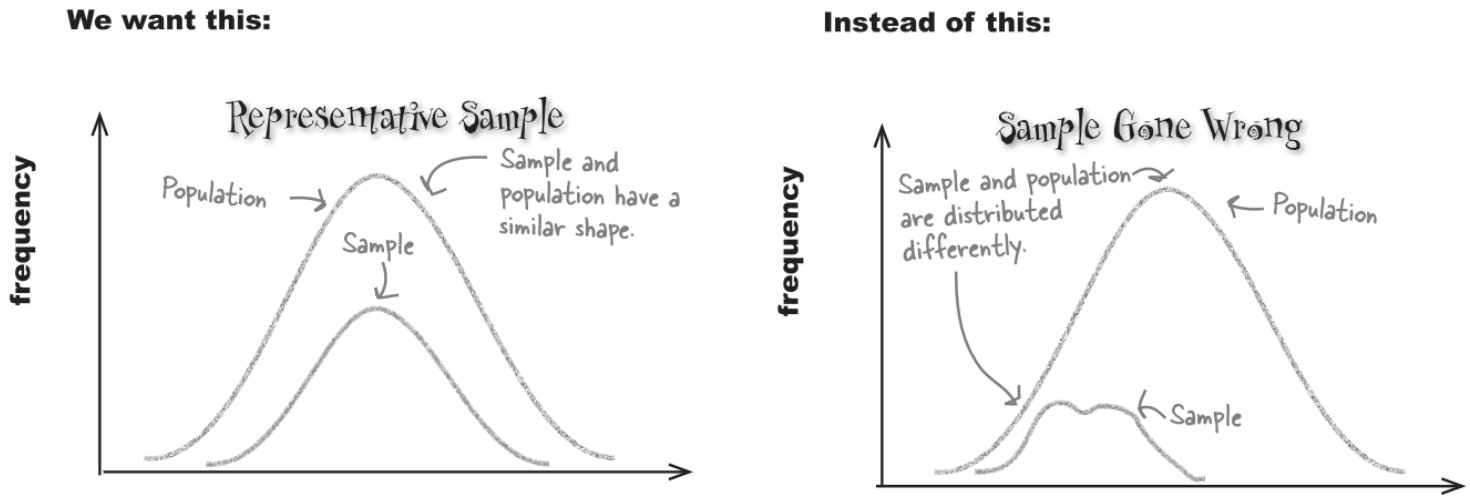





```
n(quantile.versicolor, main="Versicolor")
```

Donnée

- Each sample give a different data.
- Data can be good or bad.



- Good data: samples represents well the entire population.
- How to get good data? *Answer: understand Probability.*

Basic Probability

Probabilité

- *Résultat* (*outcome*): Eg. in customer stisfaction: any number from 1-10.
- *Univers* (*sample space*) (Ω): the set of all possible outcomes. Eg. in customer stisfaction: $\Omega = \{1, 2, ..., 10\}$.
  Note: Population is not sample space.
- *Probabilité* (*probability*): Probability of an outcome x: $0 \leq \mathbb{P}(x) \leq 1$ and $\sum_{x \in \Omega} \mathbb{P}(x) = 1$.
- *Événement* (*event*): a set of outcomes $A \subseteq \Omega$. Proability of an event: $\mathbb{P}(A) = \sum_{x \in A} \mathbb{P}(x)$.
- *Événements indépendents* (*independent event*): A and B are independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Example

Example: tossing a coin 3 times.

- Sample space $\Omega = \{hhh, hht, hth, ..., ttt\}$ (8 outcomes).
- Probability of an outcome: $\mathbb{P}(hhh) = 1/8$
- Event A: exactly two heads, so $A = \{hht, hth, thh\}$ and $\mathbb{P}(A) = 3/8$.
- Event B: tail in the middle, so $B = \{ttt, tth, hth, htt\}$ and $\mathbb{P}(B) = 1/2$.
- $A \cap B$: both A and B happen. Eg: $\mathbb{P}(A \cap B) = 1/8$
- $A \cup B$: either A or B happens. Eg: $\mathbb{P}(A \cup B) = 2/3$
- $B|A$: B happens knowing that A happens. Eg: $\mathbb{P}(B|A) = 1/3$
- A^c: A does not happens. Eg: $\mathbb{P}(A^c) = 5/8$
- Are A and B independent? No
- Find an event C such that B and C are independent.

Important formulas

Use Venn diagram when in doubt.

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$ (law of total probability)
- $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- $\Longrightarrow \quad \mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$
- $\Longrightarrow \quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$ (simple Bayes' Rule)
- $\Longrightarrow \quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)P(A) + P(B|A^c)P(A^c)}$ (Bayes' Rule)

proof: $\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c) = \mathbb{P}(B|A)P(A) + P(B|A^c)P(A^c)$.

Exercises

Exercise: There are 30 psychiatrists and 24 psychologists attending a certain conference. Two of these 54 people are randomly chosen to take part in a panel discussion. What is the probability that at least one psychologist is chosen?

Exercise: There are three cards in a hat. One is colored red on both sides, one is black on both sides, and one is red on one side and black on the other. The cards are thoroughly mixed in the hat, and one card is drawn and placed on a table. If the side facing up is red, what is the conditional probability that the other side is black?

Exercise. Suppose you're worried that you might have a rare disease. You visit your doctor to get tested, and the doctor tells you that the test is accurate 98% of the time. So, if you have the rare disease, it will correctly tell you that 98% of the time. Likewise, if you don't have the disease, it will correctly tell you that you don't 98% of the time.
The disease is rare and deadly and occurs in 1 out of every 10,000 people. Unfortunately, your test result is positive. What's the chance that you actually have the disease?

Random Variable

Variable aléatoire (*random variable*): X is a random variable if a variable that can takes on a set of values, where each value is associated with a specific probability.

Requirement: sum of all probabilities $= 1$, i.e. $\sum_i \mathbb{P}(X = x_i) = 1$

Examples:

- $X =$ number of heads in 3 tosses is a RV. $X \in \{0, 1, 2, 3\}$ and
  - $\mathbb{P}(X = 0) = 1/8$
  - $\mathbb{P}(X = 1) = 3/8$
  - $\mathbb{P}(X = 2) = 3/8$
  - $\mathbb{P}(X = 0) = 1/8$

- $Y =$ rating of a random customer is a RV. $Y \in \{1, 2, ..., 10\}$ and $\mathbb{P}(Y = i)$ is unknown

Note: If X and Y are random variables, then $aX + b$, $X + Y$, $X - Y$ are random variables.

Espérance

- *Espérance* (*expected value*):

$$\mathbb{E}[X] = \sum_i x_i \mathbb{P}(X = x_i)$$

- Espérance of any function $h(X)$:

$$\mathbb{E}\big[f(X)\big] = \sum_i h(x_i) \mathbb{P}(X = x_i)$$

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$

Exercise: You toss the coins 3 times.

- Suppose you win i€ if i heads appear. How much you win on average?
- Suppose you win i^2€ if i heads appear. How much you win on average?

More exercises

A distributor makes a profit of \$30 on each item that is received in perfect condition and suffers a loss of \$6 on each item that is received in less-than-perfect condition. If each item received is in perfect condition with probability 0.4, what is the distributor's expected profit per item?

An engineering firm must decide whether to prepare a bid for a construction project. It will cost \$800 to prepare a bid. If it does prepare a bid, then the firm will make a gross profit (excluding the preparation cost) of \$0 if it does not get the contract, \$3000 if it gets the contract and the weather is bad, or \$6000 if it gets the contract and the weather is not bad. If the probability of getting the contract is 0.4 and the probability that the weather will be bad is 0.6, what is the company's expected net profit if it prepares a bid?

Variance

Recall variance population in statistics: $\sigma^2 = \frac{(x_i - \mu)^2}{N}$

- *Variance* (*variance*):

$$\begin{aligned} Var[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \sum_i (x_i - \mathbb{E}[X])^2 \mathbb{P}(X = x_i) \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2. \end{aligned}$$

- $Var[aX + b] = a^2 Var[X]$
- $Var[aX + bY] = a^2 Var[X] + b^2 Var[Y]$ if X and Y are independent
- The variance of the number of heads appearing after 5 coin flips is?
- *Écart type* (*standard deviation*): $\sigma(X) = \sqrt{(Var[X])}$.

Exercises

Exercise: A lawyer must decide whether to charge a fixed fee of $2000 or to take a contingency fee of $8000 if she wins the case (and $0 if she loses). She estimates that her probability of winning is 0.3. Determine the standard deviation of her fee if
(a) She takes the fixed fee.
(b) She takes the contingency fee.

Exercise: The amount of money that Robert earns has expected value $30,000 and standard deviation $3000. The amount of money that his wife Sandra earns has expected value $32,000 and standard deviation $5000. Determine the
(a) Expected value
(b) Standard deviation of the total earnings of this family, assume that Robert's earnings and Sandra's earnings are independent.

Lois de probabilité

Loi de Bernoulli

- *Fonction de masse* (*probability mass function*): fonction qui donne la probabilité d'un résultat élémentaire.

Question: Consider $f(x) = \frac{1}{14}x^2$ for $x \in \{1, 2, 3\}$. Est-ce que f est un fonction de masse ?

Loi de Bernoulli $Bern(p)$: Only two outcomes: success or failure (eg. tossing a coin)

- Paramètre: $0 \leq p \leq 1$
- Fonction de masse: $\begin{cases} \mathbb{P}(X = 1) = p \\ \mathbb{P}(X = 0) = 1 - p \end{cases}$
- Espérance $= p$
- Variance $= p(1 - p)$

Loi binomiale

$Binom(n, p)$ Eg: A biased coin return head with probability p. The number of head after tossing coins n times.

- Paramètre: $0 \leq p \leq 1$ and $n \geq 0$
- Possible values: $\{0, 1, 2, ..., n\}$
- Fonction de masse:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- $Binom(n, p) \sim n \times Bern(p)$
- Espérance $= np$
- Variance $= np(1-p)$
- Compute $Binom(n, p)$ distribution at https://stattrek.com/online-calculator/binomial.aspx

Exercises

1. (a) Determine $\mathbb{P}(X \geq 12)$ when X is a binomial random variable withparameters 20 and 0.4.
(b) Determine $\mathbb{P}(Y \leq 12)$ when Y is a binomial random variable with parameters 16 and 0.5.

2. A satellite system consists of 4 components and can function if at least 2 of them are working. If each component independently works with probability 0.8, what is the probability the system will function?

3. The National Basketball Association championship series is a best-of-seven series, meaning that the first team to win four games is declared the champion. In its history, no team has ever come back to win the championship after being behind three games to one. Assuming that each of the games played in this year's series is equally likely to be won by either team, independent of the results of earlier games, what is the probability that the upcoming championship series will be the first time that a team comes back from a three-game-to-one deficit to win the series?

Probabilité continue

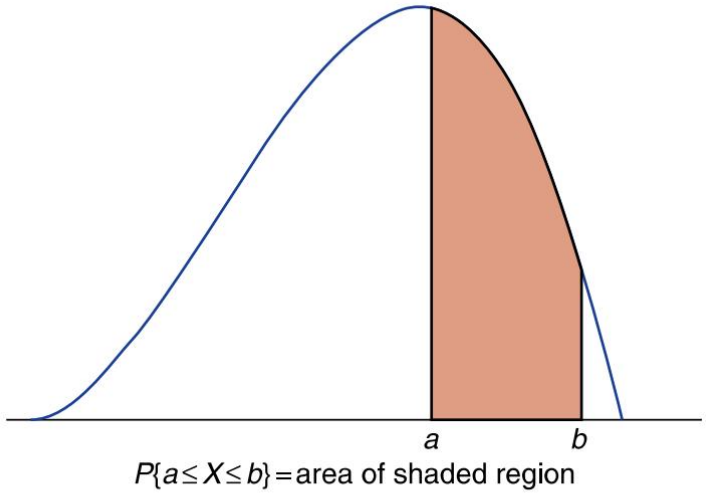


FIGURE 6.1

Probability density function of X.

- *Densité de probabilité* (*probability density function*) f: the curve
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx =$ area of shaded region.

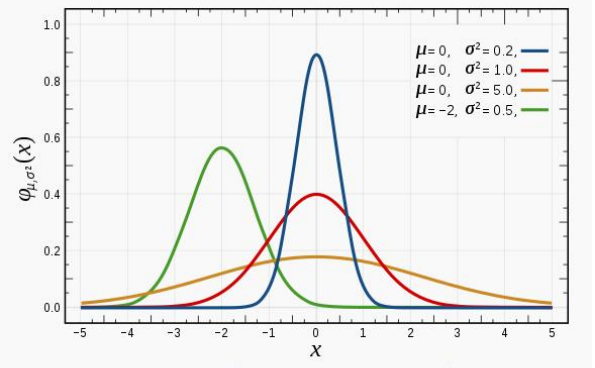
Loi normale

$\mathcal{N}(\mu, \sigma^2)$

- Paramètre: $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$
- Possible values: $\mathbb{R}$
- Densité de probabilité:

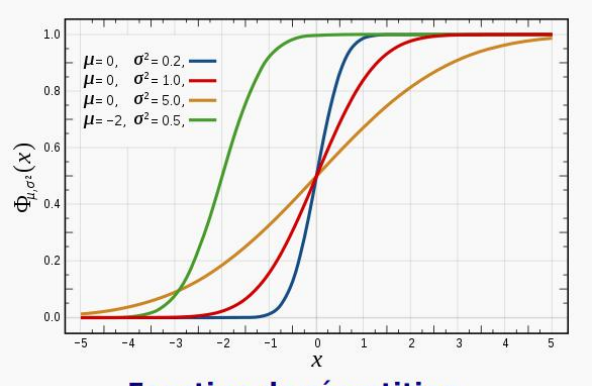
$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Espérance $= \mu$
- Variance $= \sigma^2$



Densité de probabilité

La courbe rouge représente la fonction φ, densité de probabilité de la loi normale centrée réduite.



Fonction de répartition

Loi normale centrée réduite

- $Z \sim \mathcal{N}(0,1)$: *loi normale centrée réduite* (*standard normal law*)
- Calculate probability of Z at https://www.mathsisfun.com/data/standard-normal-distribution-table.html

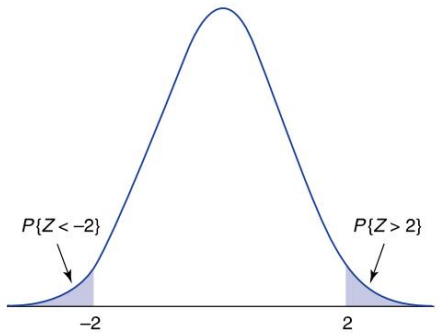


FIGURE 6.7
$P\{Z < -2\} = P\{Z > 2\}$.

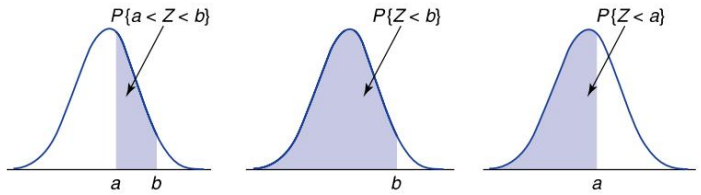


FIGURE 6.8
$P\{a < Z < b\} = P\{Z < b\} - P\{Z < a\}$.

Exercise: Find (a) $\mathbb{P}(1 < Z < 2)$ (b) $\mathbb{P}(-1.5 < Z < 2.5)$

Conver loi normale à centrée réduite

- If $X \sim N(\mu, \sigma^2)$ then $Z = \dfrac{X - \mu}{\sigma}$
- $\Longrightarrow \mathbb{P}(X < a) = \mathbb{P}\left(\dfrac{X - \mu}{\sigma} < \dfrac{a - \mu}{\sigma}\right) = \mathbb{P}\left(Z < \dfrac{a - \mu}{\sigma}\right)$

Exercises

1. *IQ examination scores for sixth-graders are normally distributed with mean value 100 and standard deviation 14.2.*
(a) What is the probability a randomly chosen sixth-grader has a score greater than 130?
(b) What is the probability a randomly chosen sixth-grader has a score between 90 and 115?

2. *Let X be normal with mean μ and standard deviation σ. Find*
(a) $\mathbb{P}(|X - \mu| > \sigma)$
(b) $\mathbb{P}(|X - \mu| > 2\sigma)$
(c) $\mathbb{P}(|X - \mu| > 3\sigma)$