Statistique pour Business et Machine learning

Tien-Nam Le

tien-nam.le@ens-lyon.fr
slides: http://perso.ens-lyon.fr/tien-nam.le/su

Sciences U University
18 Octobre, 2018

Plan

Partie I. Statistiques descriptives et probabilité

Statistiques descriptives

1. Population, échantillon et données
2. Paramètres importantes
3. Interprétation et présentation des données

Probabilité

1. Règles de probabilité
2. Variables aléatoires (discrètes et continues)
3. Loi binomiale et loi normale

Plan

Partie II. Statistiques déductives

Échantillonnage

1. Théorème central limite
2. Proportion d'échantillonnage

Estimateurs ponctuels

1. Estimateurs sans biais
2. Estimation de la moyenne et de la variance

Intervalles de confiance

1. Estimation de la moyenne lorsque la variance est connue
2. Estimation de la moyenne lorsque la variance est inconnue

Tests d'hypothèses

1. niveau de signification, p-valeur
2. Proportion de la population

Plan

Partie III. Plan d'expériences

Régression linéaire

Analyse de la variance (ANOVA)

Techniques avancées

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

Partie I. Statistiques descriptives et probabilité

Exemple

- *Population* (*population*): Tous les clients qui achètent votre produit.
- *Échantillon* (*sample*): un ensemble de clients qui font l'enquête de satisfaction
- *Donnée* (*data*): Pour chaque client qui fait l'enquête:
  - age
  - ville
  - salaire
  - satisfaction (1-10)

  Remarque: différents échantillons donnent des données différentes

Types de données:

- *Donnée quantitative* (*quantitative data*): nombres discrets (ex. âge, satisfaction) ou continus (ex. salaire).
- *Donnée qualitative* (*Qualitative data*): catégorique (ex. ville)

Statistics

Statistique (*Statistics*): une méthode pour obtenir des informations à partir de données

1. Collecte de données
2. Analyse des données
3. Interprétation des données pour obtenir des conclusions
4. Présentation des données

- *Statistique descriptive* (*descriptive statistics*) (Partie I): obtenir des conclusions sur les échantillon
- *Statistique déductive* (*inferential statistics*) (Partie II): déduire des conclusions sur l'ensemble de la population

Exemple:

- Satisfaction moyenne des clients de l'échantillon: statistique descriptive
- Satisfaction moyenne des clients de la population: statistique déductive

Paramètres importants

Exemple: Satisfaction sur l'échantillon: $X = \{4, 8, 8, 9, 10, 9, 1, 6, 9, 7\}$
Formellement, nous écrivons

- données sur l'échantillon: $X = \{X_1, ..., X_n\}$ avec n éléments
- données (inconnu) sur population: $\{x_1, ..., x_N\}$ avec N éléments

paramètres (*parameters*):

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

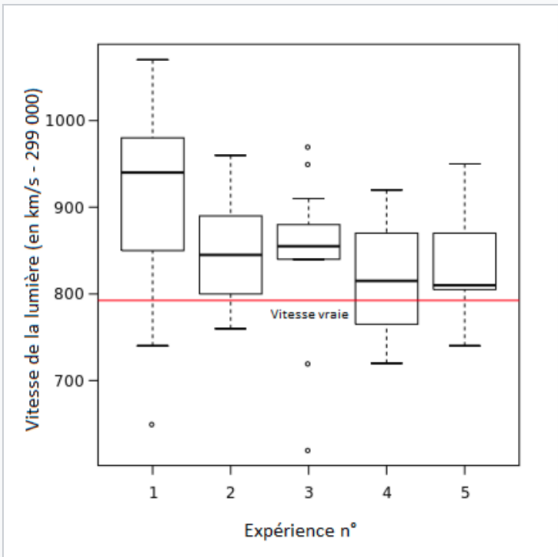
- *Médiane* (*median*): élément au milieu (à la position 50 %) des données triées.
  - Si n est impair: médiane = l'élément du milieu
  - Si n est pair: (milieu gauche + milieu droit) /2

  Ex. données triées $= \{1, 4, 6, 7, 8, 8, 9, 9, 9, 10\}$ donc $median(X) = (8 + 8)/2 = 8$.

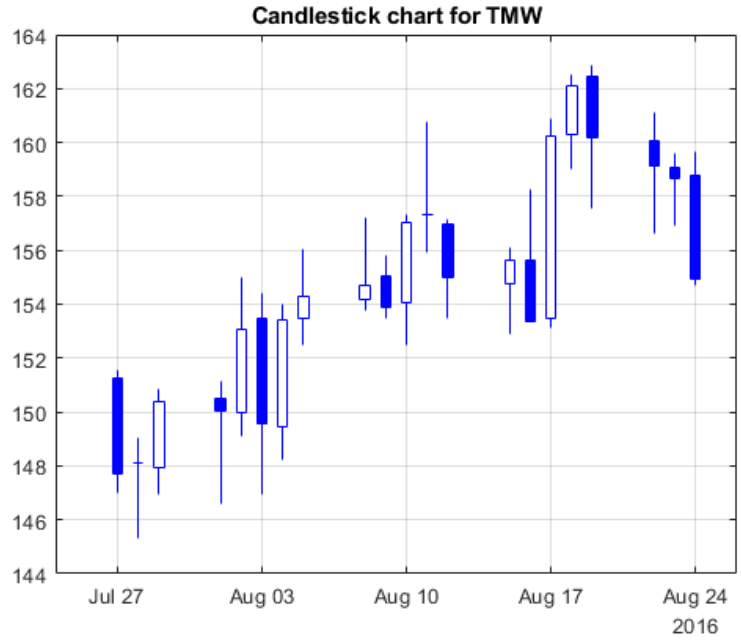
- *Mode* (*mode*): éléments les plus fréquents: ex. 9

Paramètres importants

- *Étendue* (*range*): $\max - \min = 10 - 1 = 9$
- *Donnée aberrante* (*outlier*) qui est distante des autres. ex. 1.
- 1^e *quartile* (*lower quartile*): médiane de la moitié inférieure des données triées (i.e. à la position 25% de données entières).
  ex. 6 dans $\{1, 4, 6, 7, 8, 8, 9, 9, 9, 10\}$
- 3^e *quartile* (*upper quartile*): médiane de la moitié supérieure des données triées (à la position 75% de données entières). ex. 9.
- *Boîte à moustache* (*box-plot*). Ne soyez pas confondu avec les chandeliers japonais



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.



Paramètres importants

Rappel: données sur l'échantillon $X = \{X_1, ..., X_n\}$ et données sur la population: $\{x_1, ..., x_N\}$.

- *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*): $z_i = \frac{x_i - \sigma}{\mu}$

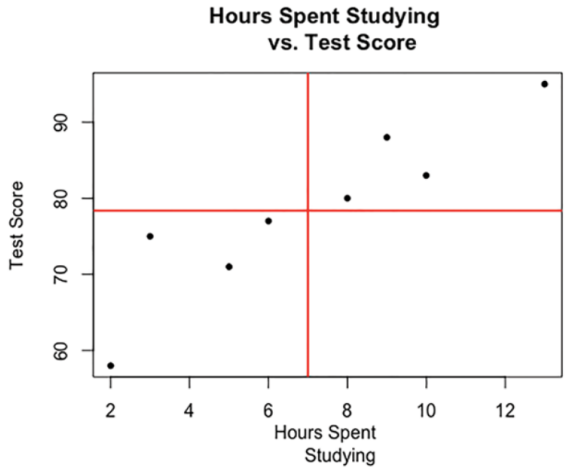
Covariance

Example:

- Heures étudiées $X = \{2, 3, 5, 6, 8, 9, 10, 13\}$
- Résultats de test $Y = \{58, 75, 71, 77, 80, 88, 83, 95\}$
- *Covariance* (*covariance*):

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

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



Dans cet exemple, $Cov(X, Y) = 38$.

Qu'est-ce que ça veut dire ? Est-ce grand ou petit?

Corrélation

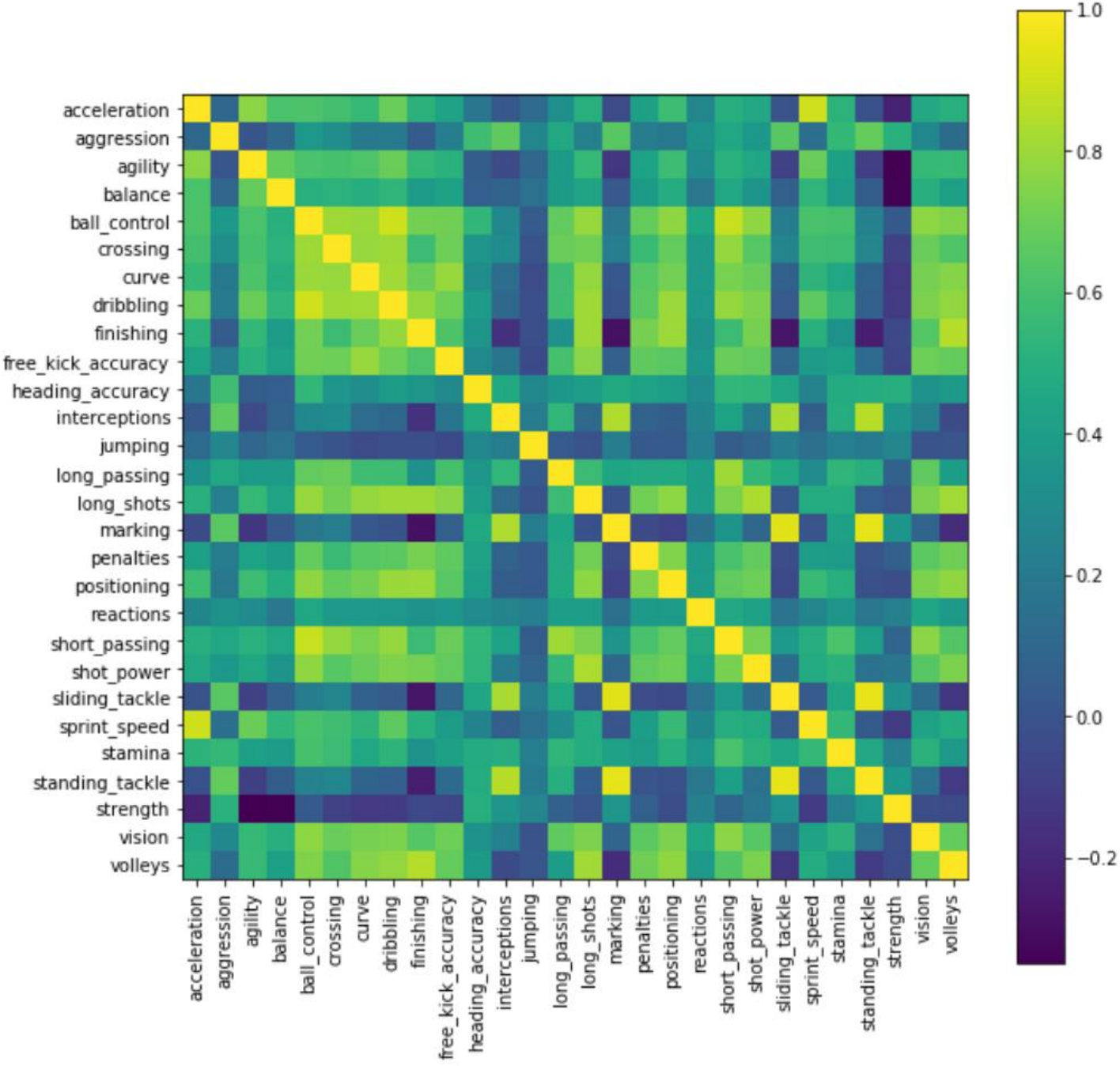
Corrélation (*correlation*):

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

Dans cet exemple, $Cor(X, Y) = \frac{38}{3.70 \times 11.19} = 0.92$, une relation positive très forte.

- $-1 \leq Cor(X, Y) \leq 1$
- $Cor(X, X) = 1$
- 1 or -1: relation linéaire, mais n'implique pas de causalité
- 0: n'implique pas qu'il n'y a pas de relation entre les deux (c'est-à-dire que X et Y ne sont pas indépendants)

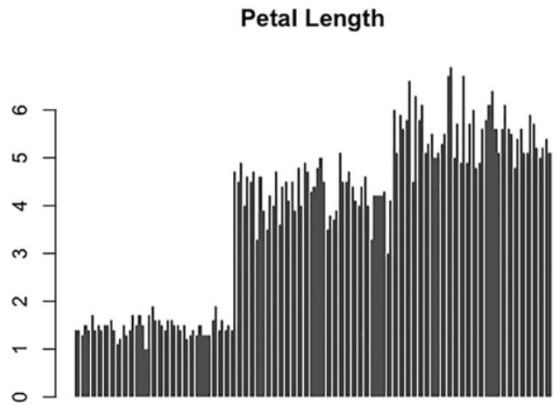
Exemple de la Corrélation

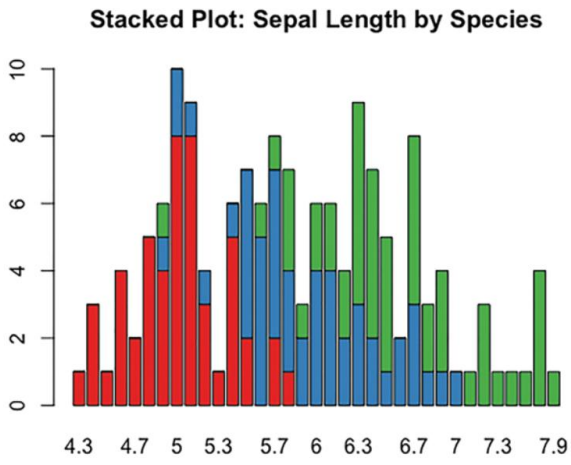


Présentation des données

Données Iris: 150 éléments, 3 espèces
données numériques: longueur des sépales, largeur des sépales, longueur des pétales, largeur des pétales

Bar plot et **Stacked plot**

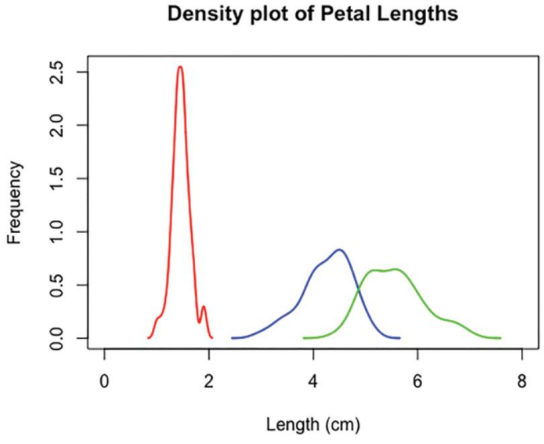


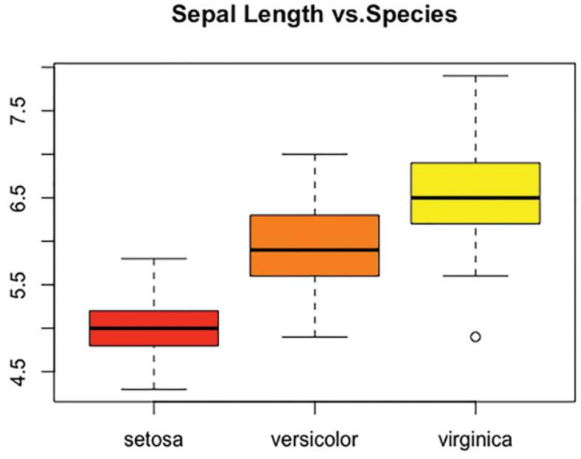


Histogram

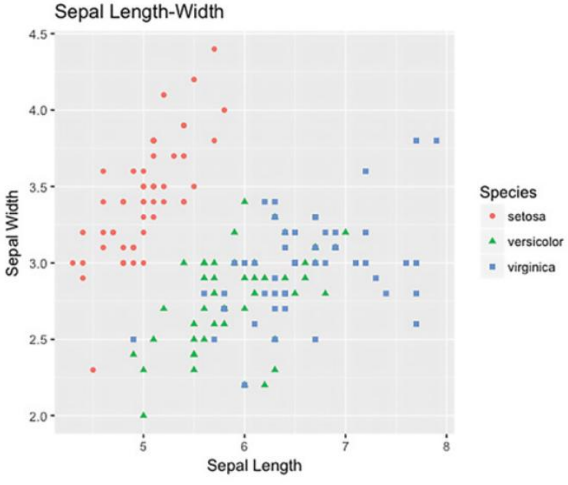
Présentation des données

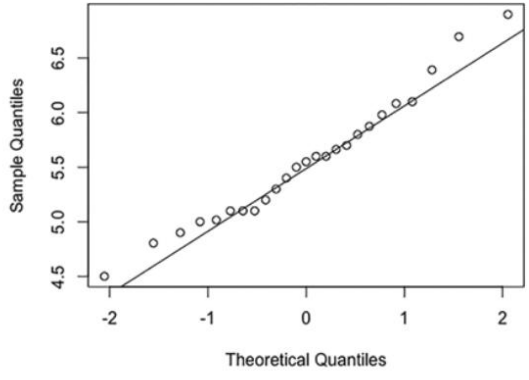
Density plot et **Box plot**





Scatter plot et **Quantile-Quantile plot**

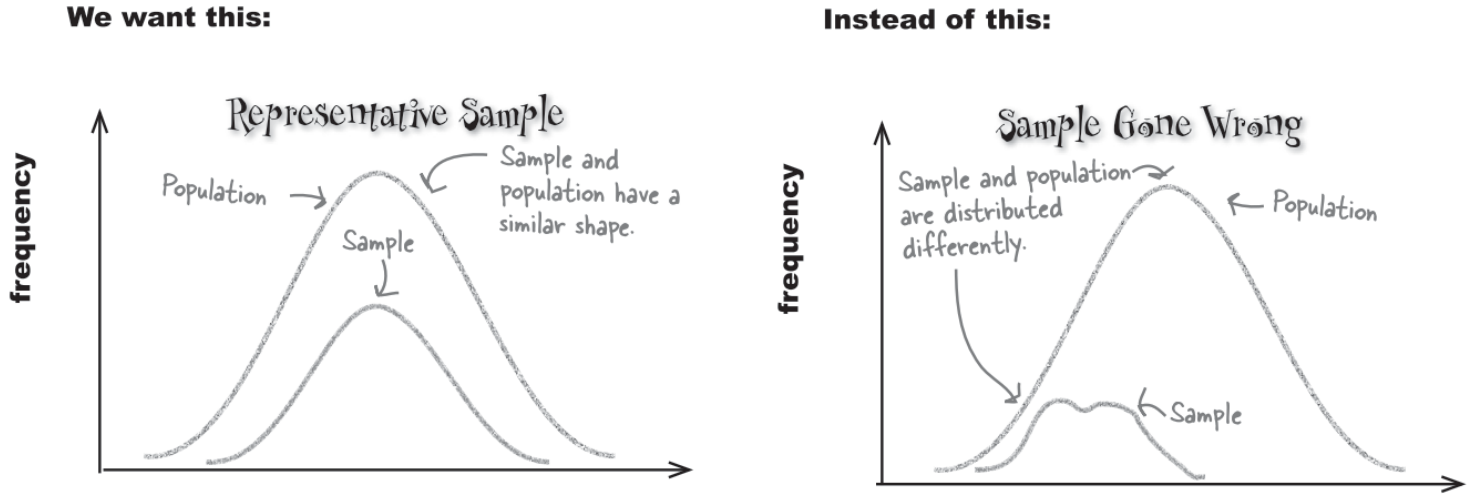





```
n(quantile.versicolor, main="Versicolor")
```

Donnée

- Chaque échantillon donne une donnée différente.
- Les données peuvent être bonnes ou mauvaises.



- Bonnes données: l'échantillon représentent bien la population entière.
- Comment obtenir de bonnes données?
- Réponse: comprendre la probabilité.

Probabilité

Probabilité

- *Résultat* (*outcome*): ex. satisfaction des clients: n'importe quel nombre de 1-10.
- *Univers* (*sample space*) (Ω): l'ensemble de tous les résultats possibles. ex. satisfaction des clients: $\Omega = \{1, 2, ..., 10\}$.
  Note: L'univers n'est pas la population.
- *Probabilité* (*probability*): Probabilité d'un résultat x: $0 \leq \mathbb{P}(x) \leq 1$ and $\sum_{x \in \Omega} \mathbb{P}(x) = 1$.
- *Événement* (*event*): un ensemble de résultats $A \subseteq \Omega$. Probabilité d'un événement: $\mathbb{P}(A) = \sum_{x \in A} \mathbb{P}(x)$.
- *Événements indépendents* (*independent event*): A et B sont indépendants si et seulement si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Exemple

Exemple: lancer une pièce 3 fois.

- L'univers $\Omega = \{hhh, hht, hth, ..., ttt\}$ (8 outcomes).
- Probabilité d'un résultat: $\mathbb{P}(hhh) = 1/8$
- Évenement A: exactement deux têtes, donc $A = \{hht, hth, thh\}$ et $\mathbb{P}(A) = 3/8$.
- Évenement B: deuxième lancer est la queue, donc $B = \{ttt, tth, hth, htt\}$ et $\mathbb{P}(B) = 1/2$.
- $A \cap B$: tous les deux A et B arrivent. Ex: $\mathbb{P}(A \cap B) = 1/8$
- $A \cup B$: soit A ou B arrive. Ex. $\mathbb{P}(A \cup B) = 2/3$
- $B|A$: B happens knowing that A happens. Ex. $\mathbb{P}(B|A) = 1/3$
- A^c: A n'arrive pas. Ex. $\mathbb{P}(A^c) = 5/8$
- A et B sont indépendent?
- Trouvez un événement C tel que B et C sont indépendants.

Formules importantes

Utilisez le diagramme de Venn en cas de doute.

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$ (loi de probabilité totale)
- $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- $\Longrightarrow \quad \mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$
- $\Longrightarrow \quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$ (simple règle de Bayes)
- $\Longrightarrow \quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)P(A) + P(B|A^c)P(A^c)}$ (règle de Bayes)

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

Exercises

1. 30 psychiatres et 24 psychologues assistent à une conférence donnée. Deux de ces 54 personnes sont choisies au hasard pour participer à une discussion en groupe. Quelle est la probabilité qu'au moins un psychologue soit choisi?

2. Il y a trois cartes dans un chapeau. L'un est de couleur rouge des deux côtés, l'un est noir des deux côtés et l'un est rouge d'un côté et noir de l'autre. Les cartes sont soigneusement mélangées dans le chapeau, et une carte est tirée et placée sur une table. Si le côté tourné vers le haut est rouge, quelle est la probabilité conditionnelle que l'autre côté soit noir?

3. Supposons que vous craignez d'avoir une maladie rare. Vous rendez visite à votre médecin pour vous faire tester, et le médecin vous dit que le test est précis 98% du temps. Donc, si vous avez la maladie rare, elle vous dira correctement que 98% du temps. De même, si vous n'avez pas la maladie, elle vous dira correctement que vous n'avez pas 98% du temps.
La maladie est rare et mortelle et survient chez 1 personne sur 10 000. Malheureusement, le résultat de votre test est positif. Quelle est la chance que vous ayez réellement la maladie?

Random Variable

Variable aléatoire (*random variable*): X est une variable aléatoire si X peut recevoir une valeur d'un ensemble de valeurs, chaque valeur étant associée à une probabilité spécifique.

Condition: somme de toutes les probabilités $= 1$, i.e. $\sum_i \mathbb{P}(X = x_i) = 1$

Exemples:

- $X =$ Le nombre de têtes dans 3 lancers est un variable aléatoire. $X \in \{0, 1, 2, 3\}$ et
  - $\mathbb{P}(X = 0) = 1/8$
  - $\mathbb{P}(X = 1) = 3/8$
  - $\mathbb{P}(X = 2) = 3/8$
  - $\mathbb{P}(X = 0) = 1/8$

- $Y =$ l'évaluation d'un client aléatoire est un variable aléatoire. $Y \in \{1, 2, ..., 10\}$ and $\mathbb{P}(Y = i)$ est inconnu.

Remarque: Si X et Y sont des variables aléatoires, alors $aX + b$, $X + Y$, $X - Y$ sont des variables aléatoires aussi.

Espérance

- *Espérance* (*expected value*):

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

- Espérance de n'importe quel fonction $h(X)$:

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

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

Exercise: *Vous lancez une pièce 3 fois.*
(a) Supposons que vous gagniez i € si i têtes apparaissent. Combien vous gagnez en moyenne?
(b) Supposons que vous gagniez i^2 € si i têtes apparaissent. Combien vous gagnez en moyenne?

Exercises

1. Un distributeur réalise un bénéfice de 30 USD sur chaque article reçu en parfait état et subit une perte de 6 USD sur chaque article reçu dans un état imparfait. Si chaque article reçu est en parfait état avec une probabilité de 0,4, quel est le bénéfice escompté du distributeur par article?

2. Une firme d'ingénierie doit décider de préparer une offre pour un projet de construction. Il en coûtera 800 $ pour préparer une offre. Si elle prépare une offre, l'entreprise réalisera un bénéfice brut (à l'exclusion du coût de préparation) de 0 $ si elle n'obtient pas le contrat, de 3 000 $ si elle obtient le contrat et que les conditions météorologiques sont mauvaises, ou de 6 000 $ si obtient le contrat et le temps n'est pas mauvais. Si la probabilité d'obtenir le contrat est de 0,4 et la probabilité que les conditions météorologiques soient défavorables est de 0,6, quel sera le bénéfice net attendu de la société si elle prépare une offre?

Variance

Rappel: variance de la population (dans statistique): $\sigma^2 = \frac{(x_i - \mu)^2}{N}$

- *Variance* (*variance*):

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

- $Var[aX + b] = a^2 Var[X]$
- $Var[aX + bY] = a^2 Var[X] + b^2 Var[Y]$ si X et Y sont indépendants
- *Écart type* (*standard deviation*): $\sigma(X) = \sqrt{(Var[X])}$.

Exercise: *La variance du nombre de têtes apparaissant après 5 lancers de pièce est?*

Exercises

1. Un avocat doit décider s'il convient de facturer des frais fixes de 2 000 dollars ou de percevoir une indemnité forfaitaire de 8 000 dollars s'il gagne le procès (et de 0 dollar s'il perd). Elle estime que sa probabilité de gagner est de 0,3. Déterminer l'écart type de ses frais si
(a) Elle prend les frais fixes.
(b) Elle prend les honoraires conditionnels.

2. Le montant que gagne Robert a une valeur attendue de 30 000 dollars et un écart type de 3 000 dollars. Le montant que gagne son épouse Sandra a une valeur attendue de 32 000 dollars et un écart type de 5 000 dollars. Détermine le
a) Valeur attendue
b) Écart type des gains totaux de cette famille, en supposant que les revenus de Robert et de Sandra soient indépendants.

Lois de probabilité

Loi de Bernoulli

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

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

Loi de Bernoulli $Bern(p)$: Seulement deux résultats: succès ou échec (par exemple, lancer une pièce de monnaie)

- 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)

Ex. Étant donné une pièce biaisée qui retourne la tête avec la probabilité p, Le nombre de têtes après avoir lancé le pièce n fois suit *Binom*(n, p)

- Paramètre: $0 \leq p \leq 1$ and $n \geq 0$
- Univers: $\{0, 1, 2, ..., n\}$
- 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 $= np(1-p)$
- Calculer la probabilité de *Binom*(n, p) à https://stattrek.com/online-calculator/binomial.aspx

Exercises

1. (a) Déterminez $\mathbb{P}(X \geq 12)$ lorsque X est une variable aléatoire binomiale avec les paramètres 20 et 0.4.
(b) Déterminez $\mathbb{P}(Y \leq 12)$ lorsque Y est une variable aléatoire binomiale avec les paramètres 16 et 0.5.

2. Un système satellite comprend 4 composants et peut fonctionner si au moins 2 d'entre eux fonctionnent. Si chaque composant fonctionne indépendamment avec une probabilité de 0,8, quelle est la probabilité que le système fonctionne?

3. La série de championnats de la National Basketball Association est une série au meilleur des sept, ce qui signifie que la première équipe à remporter quatre matchs est déclarée championne. Dans son histoire, aucune équipe n'est revenue pour remporter le championnat après avoir été derrière trois matchs contre un. En supposant que chacune des parties jouées dans la série de cette année a toutes les chances d'être gagnée par l'une ou l'autre des équipes, indépendamment des résultats des précédentes éditions, quelle est la probabilité que la prochaine série de championnat soit la première fois qu'une équipe revient d'une déficit de trois matchs contre un pour gagner la série?

Probabilité continue

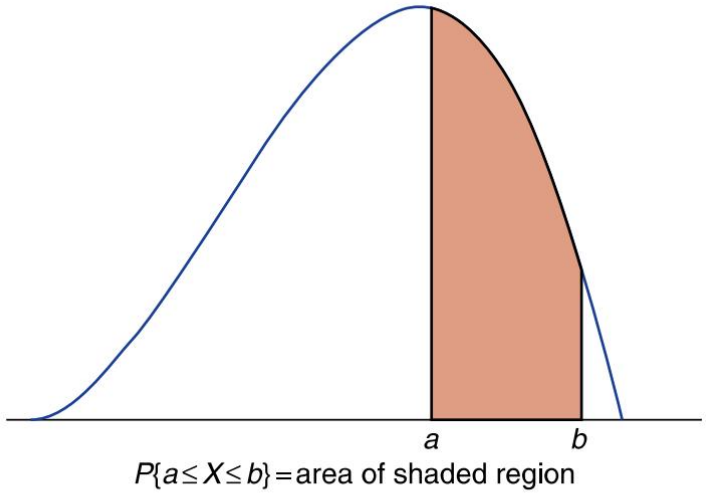


FIGURE 6.1
Probability density function of X.

- *Densité de probabilité* (*probability density function*) f : la courbe
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx =$ zone de la région ombragée.

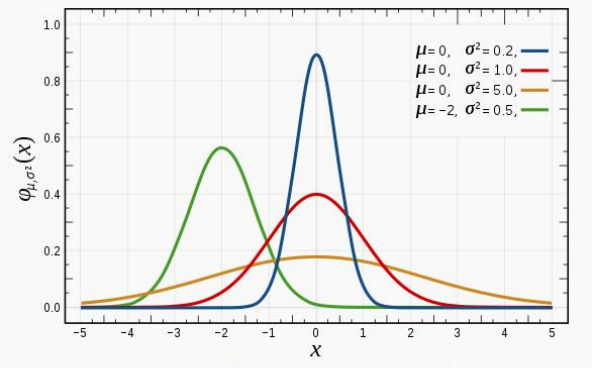
Loi normale

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

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

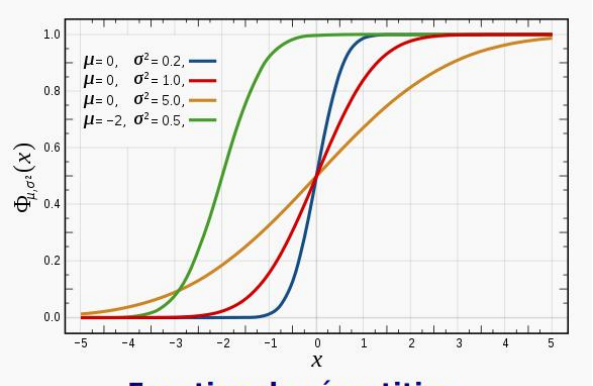
$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- 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*)
- Calculer la probabilité de Z à https://www.mathsisfun.com/data/standard-normal-distribution-table.html

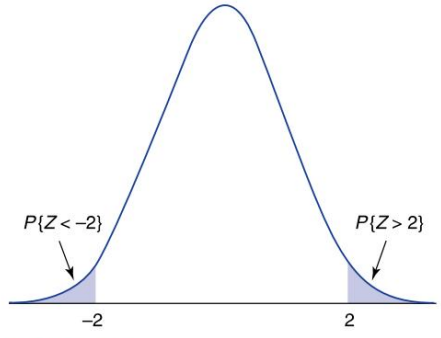


FIGURE 6.7
$P\{Z < -2\} = P\{Z > 2\}$.

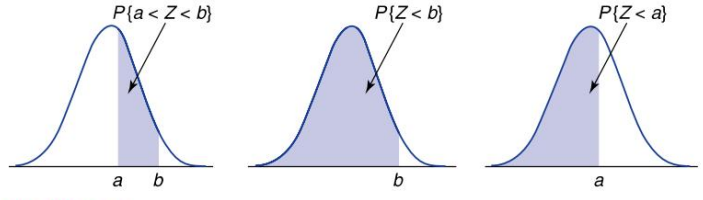


FIGURE 6.8
$P\{a < Z < b\} = P\{Z < b\} - P\{Z < a\}$.

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

Convert loi normale à centrée réduite

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

Exercises

1. *Les scores aux examens de QI des élèves de sixième année sont normalement répartis entre la valeur moyenne 100 et l'écart type 14.2.*
(a) Quelle est la probabilité pour qu'un élève de sixième année choisi au hasard ait un score supérieur à 130?
(b) Quelle est la probabilité pour qu'un élève de sixième année choisi au hasard ait un score compris entre 90 et 115?

2. *Soit X normal avec la moyenne μ et l'écart type σ. Trouver*
(a) $\mathbb{P}(|X - \mu| > \sigma)$
(b) $\mathbb{P}(|X - \mu| > 2\sigma)$
(c) $\mathbb{P}(|X - \mu| > 3\sigma)$