

Description et Inférence Statistiques

Pré-Orientation ICBE

3ème année

2021-2022

Avant-propos

Ce polycop est présenté, à l'exception des deux annexes qui le clôturent, sous forme de slides, proches de (mais non identiques à) celles qui seront projetées en cours. Il aborde un spectre de notions très large et sa seule lecture ne saurait suffire à leur assimilation. Des supports de cours détaillés sont disponibles sur Moodle.

N'hésitez pas à me contacter pour toute remarque ou suggestion.

Table des matières

1	Chapitre 1 : Introduction	1
2	Chapitre 2 : Statistique descriptive	11
3	Chapitre 3 : Probabilités	68
4	Chapitre 4 : Estimation statistique	125
5	Chapitre 5 : Tests statistiques	148
6	Chapitre 6 : Régression linéaire simple	184
7	Chapitre 7 : Régression linéaire multiple	219
A	Annexe 1 : Tables statistiques	237
B	Annexe 2 : Examen 2018/2019	242

Chapitre 1 : Une introduction

Vous avez dit "statistique(s)" ?

Qu'est ce que ce mot vous inspire ?

D'après Wikipédia, *"la statistique est l'étude d'un phénomène par la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous."*

En dehors d'une salle de maths, le terme *statistique* évoque :

- les sondages, enquêtes de satisfaction, d'intentions de vote, ...
- une anecdote avec des chiffres au milieu : "95% des statistiques que l'on raconte en cours de maths sont de pures inventions".
- les "stats" d'un joueur de basket ou de foot.

Dans le cadre mathématique, dans le **domaine des Statistiques** :

- **Une statistique** : une quantité définie par un rapport à un modèle qui permet d'obtenir une indication sur le comportement de la population. On parle notamment de **statistique de test**.
- **Les statistiques** : tableaux de chiffres, d'observations.

En parlant de vocabulaire...

NB : rigueur et précision sont indispensables dans l'emploi des mots et des concepts en statistique !

- **Population** (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. Souvent notée Ω .
↪ exemple : la population de la région Occitanie.
- **Individu** (ou unité statistique) : élément de la population, souvent noté $\omega \in \Omega$. ↪ exemple : un habitant de Lavelanet.
- **Echantillon** : sous-ensemble (effectivement observé) de la population. Sa taille est généralement notée n .
↪ exemple : les 3ICBE de l'INSA de Toulouse.

+ représentativité ?
+ déontologie !

En parlant de vocabulaire...

NB : rigueur et précision sont indispensables dans l'emploi des mots et des concepts en statistique !

- **Enquête** : opération consistant à observer l'ensemble des éléments/individus d'un échantillon.
- **Recensement** (ou enquête exhaustive) : enquête dans laquelle l'échantillon observé est la population dans son intégralité.
- **Sondage** (ou enquête non exhaustive) : enquête dans laquelle l'échantillon observé est un sous-ensemble strict de la population.

En parlant de vocabulaire...

- **Variable** (statistique) : caractère, propriété, "chose" que l'on étudie ; application définie pour tout individu de la population Ω ; elle peut être quantitative ou qualitative ; souvent notée X .
↪ exemple : "chocolatine ou pain au chocolat ?" : la variable associée X_C est définie de la population Ω dans l'ensemble $\mathcal{E} = \{\text{chocolatine, pain au chocolat}\}$.
 - ▶ **Variable quantitative** : variable à valeurs dans \mathbb{R} (ou dans une partie de \mathbb{R}), i.e. $X : \Omega \rightarrow \mathbb{R}$. ↪ exemple : taille, âge, note de philo au bac, etc.
 - ▶ **Variable qualitative** : variable à valeurs dans un ensemble \mathcal{E} qui n'est pas (une partie de) \mathbb{R} , i.e. $X : \Omega \rightarrow \mathcal{E}$. ↪ exemple : couleur des chaussettes, lieu de naissance, sexe, prénom.
- **Modalités** : valeurs que peut prendre une variable statistique.
↪ exemple : la variable X_C a deux modalités, "chocolatine" et "pain au chocolat".
- **Données** (statistiques) : ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus.

A propos du titre.

- **Description statistique**, ou **statistique descriptive** :
 - ▶ Objectif : **décrire les données** disponibles pour une population. "Résumer l'information contenue dans les données de manière synthétique et efficace."
 - ▶ Comment ? Représentations graphiques, résumés numériques.
 - ▶ **Sans** outils probabilistes.

On parle aussi de **statistique exploratoire** ou d'**analyse des données**.
- **Inférence statistique**, ou **statistique inférentielle** :
 - ▶ Objectif : à **partir** de l'observation d'**un échantillon**, **induire** les caractéristiques inconnues d'une **population**. "Inférence du particulier au général à visée explicative."
 - ▶ Comment ? Estimation, tests, modélisation.
 - ▶ **Avec** des outils probabilistes.

● Expérimentation

- ▶ Quelle est la question ? Quel est l'objectif ?
Est-il descriptif, explicatif, prédictif ?
- ▶ Quelle est la population étudiée ? De quels échantillons dispose-t-on ?
- ▶ Que peut-on dire des conditions expérimentales, de la précision des mesures ?

● Objectif descriptif

- ▶ Etape indispensable : description des données
- ▶ Valeurs manquantes, erronées ou atypiques ?
- ▶ Distributions anormales ?
- ▶ Incohérences ?

● Objectif explicatif

- ▶ Traduction mathématique/statistique de l'hypothèse
- ▶ Choix du modèle
- ▶ Estimation des paramètres et/ou calcul de la statistique de test
- ▶ Prise de décision : acceptation ou rejet de l'hypothèse

Exemples d'applications

Parmi les (très) nombreux champs d'application...

● Biostatistique :

- ▶ Traitements cliniques
- ▶ Génomique (ex : GenBank)
- ▶ Dynamique de population
- ▶ Modèles épidémiologiques

● Fouille de données (ou *data mining*) :

- ▶ Traitement des *Big Data* (mégadonnées)
- ▶ Marketing
- ▶ Gestion de la relation client

● Industrie :

- ▶ Contrôle de qualité
- ▶ Optimisation de la durée de vie
- ▶ Planification expérimentale

A propos du cours.

I. Statistique descriptive

1. Unidimensionnelle
2. Bidimensionnelle
3. Multidimensionnelle : introduction à l'ACP.

II. Quelques notions de probabilités

III. Statistique inférentielle

1. Estimation statistique
2. Tests statistiques
3. Régression linéaire

En conclusion.

Quelques questions à se poser :

- Que veut-on savoir ?
- Quelle va être la méthode appropriée ?
- Quelles en sont les limites ?
- Comment modéliser l'expérience ?
- Comment interpréter le résultat ?

Chapitre 2 : Statistique descriptive

Programme des réjouissances

- 1 **Statistique descriptive unidimensionnelle**
 - Une variable quantitative
 - Une variable quantitative discrète
 - Une variable quantitative continue
 - Une variable qualitative
 - Détection de problèmes et transformation des données

- 2 **Statistique descriptive bidimensionnelle**
 - Deux variables quantitatives
 - Une variable quantitative et une variable qualitative
 - Deux variables qualitatives

- 3 **Statistique descriptive multidimensionnelle**
 - Généralisation du cadre bidimensionnel
 - Introduction à l'Analyse en Composantes Principales

On enchaîne avec...

- 1 Statistique descriptive unidimensionnelle
 - Une variable quantitative
 - Une variable quantitative discrète
 - Une variable quantitative continue
 - Une variable qualitative
 - Détection de problèmes et transformation des données
- 2 Statistique descriptive bidimensionnelle
 - Deux variables quantitatives
 - Une variable quantitative et une variable qualitative
 - Deux variables qualitatives
- 3 Statistique descriptive multidimensionnelle
 - Généralisation du cadre bidimensionnel
 - Introduction à l'Analyse en Composantes Principales

Complément sur les variables statistiques

- **Variable quantitative** : variable à valeurs dans \mathbb{R} (ou dans une partie de \mathbb{R}), i.e. $X : \Omega \rightarrow \mathbb{R}$. \leftrightarrow exemple : taille, âge, note de philo au bac, etc.
 - ▶ **Variable quantitative discrète** : variable quantitative à valeurs entières (ou, rarement, décimales) ; variable n'ayant qu'un nombre fini, ou dénombrable, de modalités. \leftrightarrow exemple : nombre de cousins.
 - ▶ **Variable quantitative continue** : variable ayant un nombre infini non dénombrable de modalités \leftrightarrow exemple : taille d'un individu.
- **Variable qualitative** : variable à valeurs dans un ensemble \mathcal{E} qui n'est pas (une partie de) \mathbb{R} , i.e. $X : \Omega \rightarrow \mathcal{E}$. \leftrightarrow exemple : couleur des chaussettes, lieu de naissance, sexe, prénom.
 - ▶ **Variable qualitative ordinale** : variable qualitative aux modalités naturellement ordonnées. \leftrightarrow exemple : taille de vêtements, mention au bac.
 - ▶ **Variable qualitative nominale** : variable dont les modalités sont traduites par un "nom". \leftrightarrow exemple : couleur des chaussettes, groupe sanguin.

Une variable quantitative discrète : la population des cantons de Haute-Garonne.

N°	Cantons et métropoles	Population municipale
1	Auterive	51 182
2	Bagnères-de-Luchon	33 424
3	Blagnac	50 522
4	Castanet-Tolosan	43 240
5	Castelginest	51 928
6	Cazères	43 219
7	Escalquens	41 463
8	Léguévin	52 665
9	Muret	52 902
10	Pechbonnieu	40 715
11	Plaisance-du-Touch	46 768
12	Portet-sur-Garonne	47 588
13	Revel	39 500
14	Saint-Gaudens	35 097

N°	Cantons et métropoles	Population municipale
15	Toulouse 1	54 307
16	Toulouse 10	56 446
17	Toulouse 11	48 078
18	Toulouse 2	55 397
19	Toulouse 3	52 536
20	Toulouse 4	52 753
21	Toulouse 5	45 521
22	Toulouse 6	55 866
23	Toulouse 7	57 622
24	Toulouse 8	60 233
25	Toulouse 9	54 205
26	Tournefeuille	52 721
27	Villemur-sur-Tarn	41 770

Figure – Population des cantons de Haute-Garonne au 1er janvier 2017. Source : INSEE.

Série brute arrondie au millier : 51, 33, 51, 43, 52, 43, 41, 53, 53, 41, 47, 48, 40, 35, 54, 56, 48, 55, 53, 53, 46, 56, 58, 60, 54, 53, 42.

Une variable quantitative discrète : présentation des données

Vocabulaire et notations

- $(m_i)_{i \in \{1, \dots, r\}}$: modalités rangées par ordre croissant ;
- $(n_i)_{i \in \{1, \dots, r\}}$: effectifs associés à chaque modalité ;
- $(f_i)_{i \in \{1, \dots, r\}}$: fréquences associées à chaque modalité, définies par $f_i = n_i / n$;
- $(N_i)_{i \in \{1, \dots, r\}}$: effectifs cumulés, définis par $N_i = \sum_{j=1}^i n_j$;
- $(F_i)_{i \in \{1, \dots, r\}}$: fréquences cumulées, définies par $F_i = \sum_{j=1}^i f_j$.

Tableau statistique

m_i	n_i	N_i	f_i en %	F_i en %
33	1	1	3,70 %	3,70 %
35	1	2	3,70 %	7,41 %
40	1	3	3,70 %	11,11 %
41	2	5	7,41 %	18,52 %
42	1	6	3,70 %	22,22 %
43	2	8	7,41 %	29,63 %
46	1	9	3,70 %	33,33 %
47	1	10	3,70 %	37,04 %
48	2	12	7,41 %	44,44 %
51	2	14	7,41 %	51,85 %
52	1	15	3,70 %	55,56 %
53	5	20	18,52 %	74,07 %
54	2	22	7,41 %	81,48 %
55	1	23	3,70 %	85,19 %
56	2	25	7,41 %	92,59 %
58	1	26	3,70 %	96,30 %
60	1	27	3,70 %	100,00 %

Figure – Observations distinctes, effectifs, effectifs cumulés, fréquences et fréquences cumulées.

Une variable quantitative discrète : représentation graphique

Diagramme en bâtons.

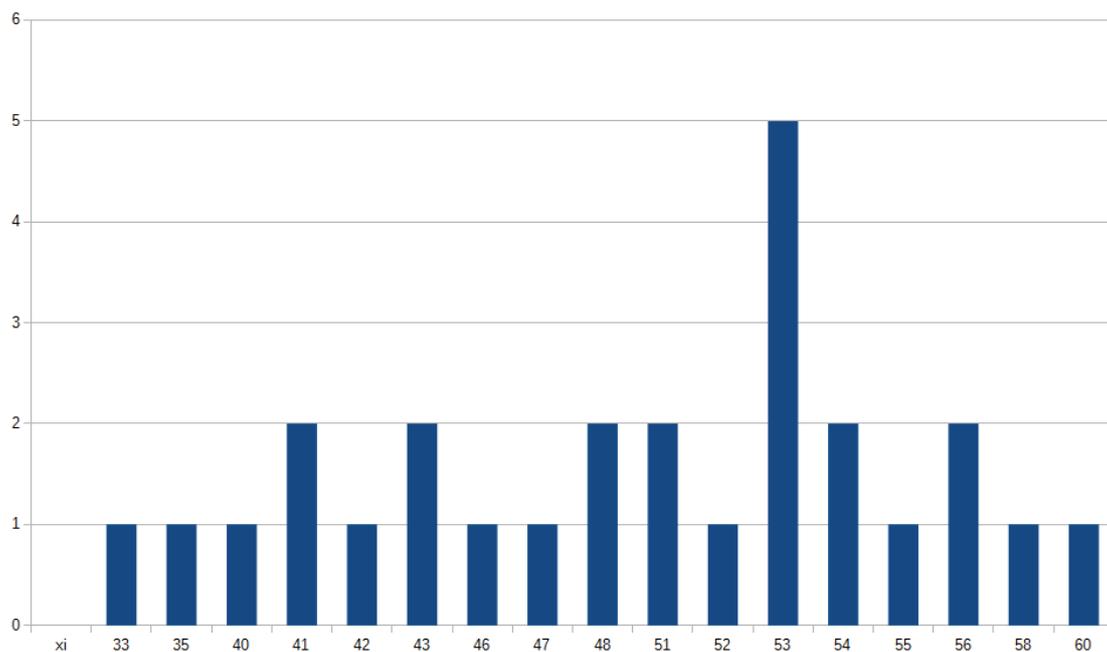


Figure – Abscisse : population en milliers d'habitants. Ordonnée : effectif.

Une variable quantitative discrète : indicateurs numériques

- **Moyenne** : définie par $\bar{x} = \frac{1}{n} \sum_{j=1}^r n_j m_j$. Ici, $\bar{x} \approx 48,9 \approx 49$.
- **Variance** : définie par $\text{var}(X) = \sigma_X^2 = \frac{1}{n} \sum_{l=1}^r n_l (m_l - \bar{x})^2 = \left(\frac{1}{n} \sum_{l=1}^r n_l m_l^2 \right) - (\bar{x})^2$
Ici, $\text{var } X \approx 47,4$.
- **Ecart-type** : racine carrée de la variance, noté σ_X . Ici, $\sigma_X \approx 6,9$.
- **Médiane** : valeur, notée $x_{1/2}$, telle que la moitié des observations est plus grande que $x_{1/2}$ (et donc telle que la moitié des observations est plus petite que $x_{1/2}$). Ici, $x_{1/2} = 51$.
- **Quantile** : le quantile d'ordre $\alpha \in [0, 1]$, noté x_α est une valeur telle qu'une proportion α des observations est plus petite que x_α (et donc qu'une proportion $1 - \alpha$ est plus grande que x_α). Ici, $x_{0,1} = 40$, $x_{0,25} = 43$, $x_{0,75} = 53,5$. NB : la médiane est le quantile d'ordre 1/2 (et second quartile).

Une variable quantitative discrète : le diagramme-boîte

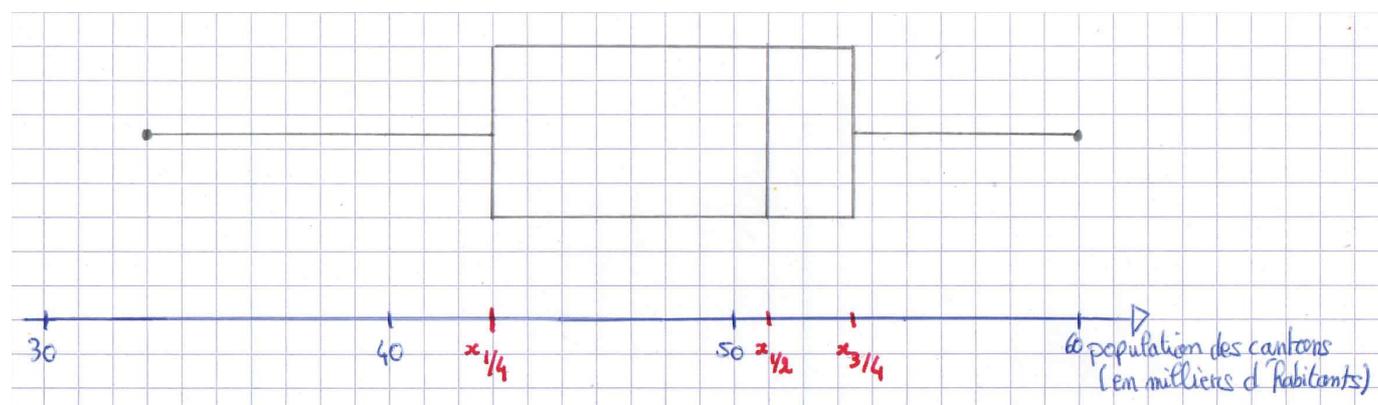


Figure – Diagramme-boîte pour la population des cantons de Haute-Garonne.

Une variable quantitative continue : la superficie des exploitations agricoles de Midi-Pyrénées

Surface agricole	Nombre d'exploitations (en milliers)	Fréquence (en %)
Moins de 20ha	202,3	42,84 %
De 20ha à 50ha	79	16,73 %
De 50ha à 100ha	93,3	19,76 %
De 100ha à 200ha	74	15,67 %
Plus de 200ha	23,6	5,00 %

Figure – Surface agricole utile des exploitations agricoles de Midi-Pyrénées en 2013.
Source : SSP, Agreste, enquête structure 2013.

Fonction de répartition et densité empiriques

On note b_0, \dots, b_r les bornes des classes, et f_j (resp. F_j) la fréquence (resp. fréquence cumulée) de la classe $[b_{j-1}, b_j]$.

↪ dans notre exemple : $b_0 = 0$, $b_1 = 20$, $b_2 = 50$, $b_3 = 100$, $b_4 = 200$ et $b_5 = 1000$.

- **Fonction de répartition empirique :**

$$F_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ F_{j-1} + \frac{f_l}{b_l - b_{l-1}}(x - b_{j-1}) & \text{si } b_{l-1} < x < b_l, \quad l = 1, \dots, r, \\ 1 & \text{si } x \geq b_r. \end{cases}$$

Son graphe est appelé **courbe cumulative** de la variable X .

Fonction de répartition et densité empiriques

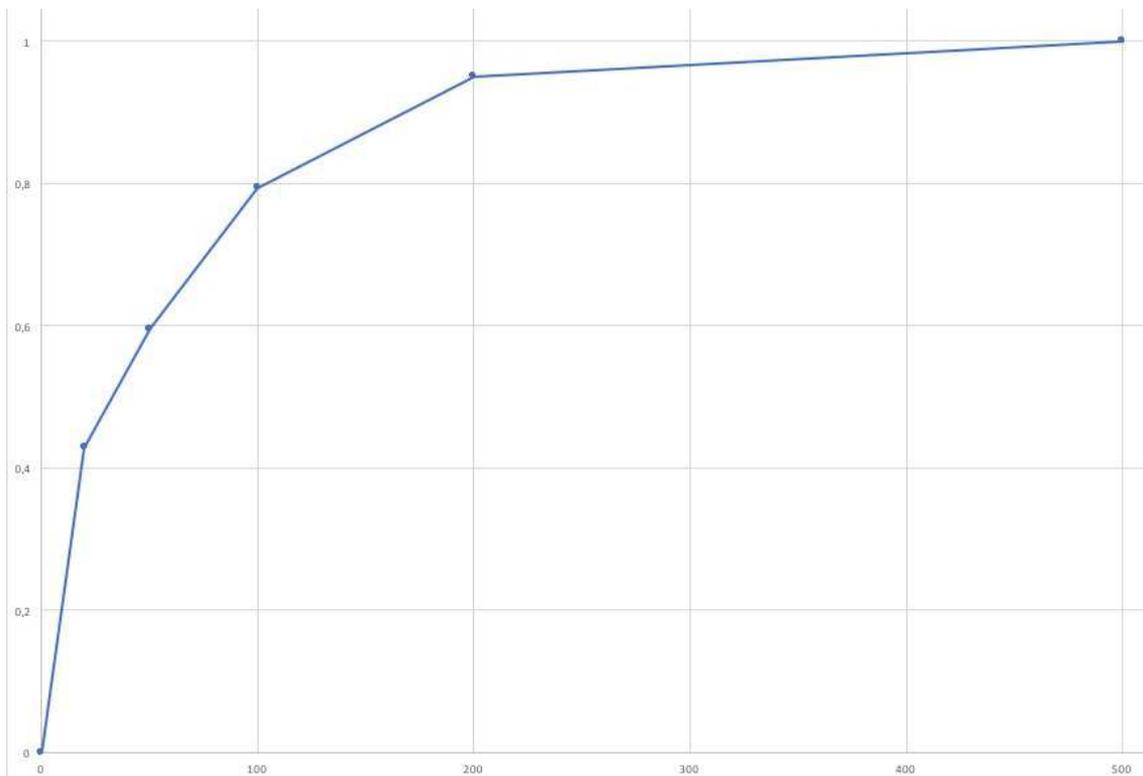


Figure – Courbe cumulative associée à la surface des exploitations agricoles de Midi-Pyrénées en 2013.

Fonction de répartition et densité empiriques

On note b_0, \dots, b_r les bornes des classes, et f_j (resp. F_j) la fréquence (resp. fréquence cumulée) de la classe $[b_{j-1}, b_j]$.

↪ dans notre exemple : $b_0 = 0$, $b_1 = 20$, $b_2 = 50$, $b_3 = 100$, $b_4 = 200$ et $b_5 = 1000$.

- **Densité empirique** : notée, f_X , il s'agit de la fonction dérivée de F_X .

$$f_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ \frac{f_j}{b_j - b_{j-1}} & \text{si } b_{j-1} < x < b_j, \quad j = 1, \dots, r, \\ 0 & \text{si } x \geq b_r. \end{cases}$$

Son graphe est appelé **histogramme** de la variable X .

- **Construction de l'histogramme** : en pratique, la hauteur associée à la classe $[b_{j-1}, b_j]$ est donnée par

$$h_j = \frac{f_j}{b_j - b_{j-1}}.$$

Fonction de répartition et densité empiriques

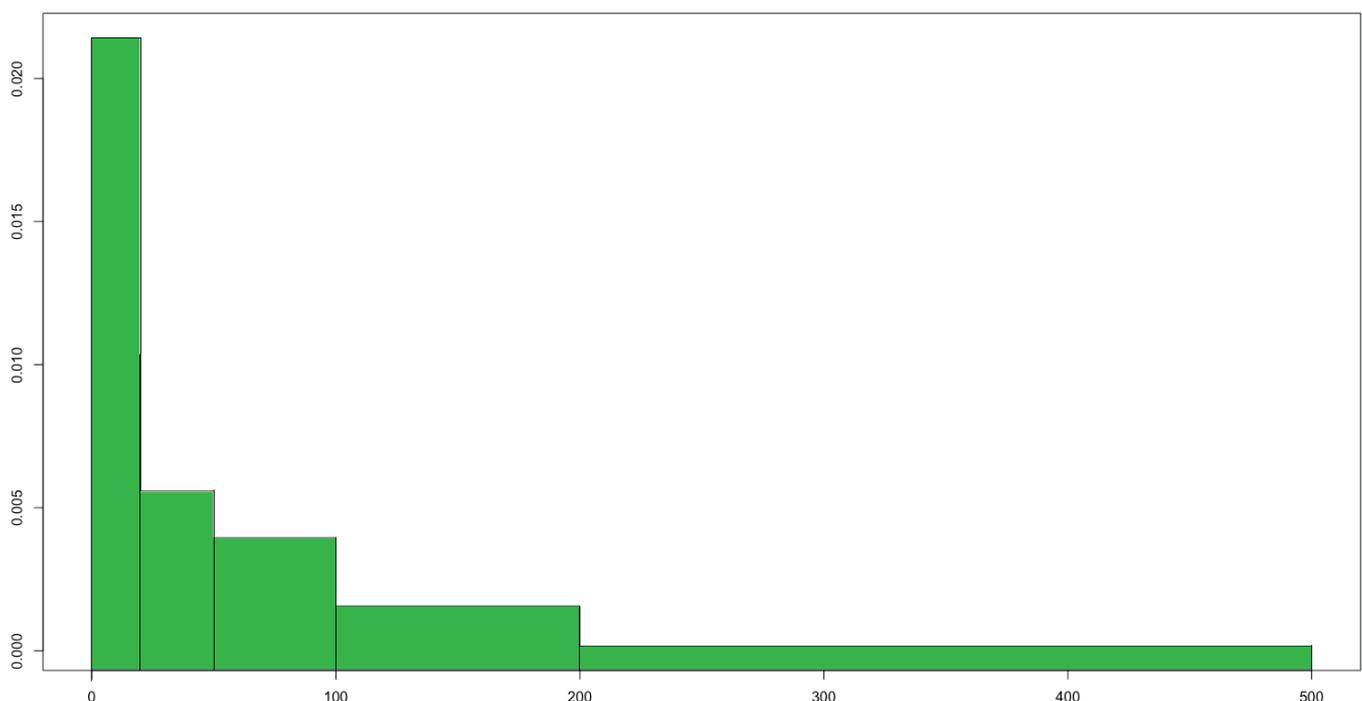


Figure – Histogramme associée à la surface des exploitations agricoles de Midi-Pyrénées en 2013.

Une variable quantitative continue : indicateurs numériques

La moyenne, la variance et l'écart-type d'une variable continue se calculent de manière analogue au cas discret, en prenant pour x_j le centre des classes, ie $\frac{b_j + b_{j-1}}{2}$ lorsque que l'on ne connaît que la classe de chaque observation.

Le cas d'une variable qualitative

⇒ Uniquement des tableaux statistiques et des représentations graphiques.
↔ Diagrammes en colonnes, en barres et en secteurs.

Exemple :

	Licenciés	Fréquence
Football	2 135 193	24,49 %
Tennis	1 052 127	12,07 %
Equitation	673 026	7,72 %
Judo et associés	552 815	6,34 %
Basket	513 717	5,89 %
Handball	513 194	5,89 %
Golf	407 569	4,67 %
Rugby	323 571	3,71 %
Gymnastique	287 358	3,30 %
Natation	300 926	3,45 %
Autres	1 959 982	22,48 %
Total	8 719 478	

Figure – Nombre de licenciés des fédérations olympiques françaises en 2016. *Source : INJEP, mars 2017.*

Diagramme en secteurs

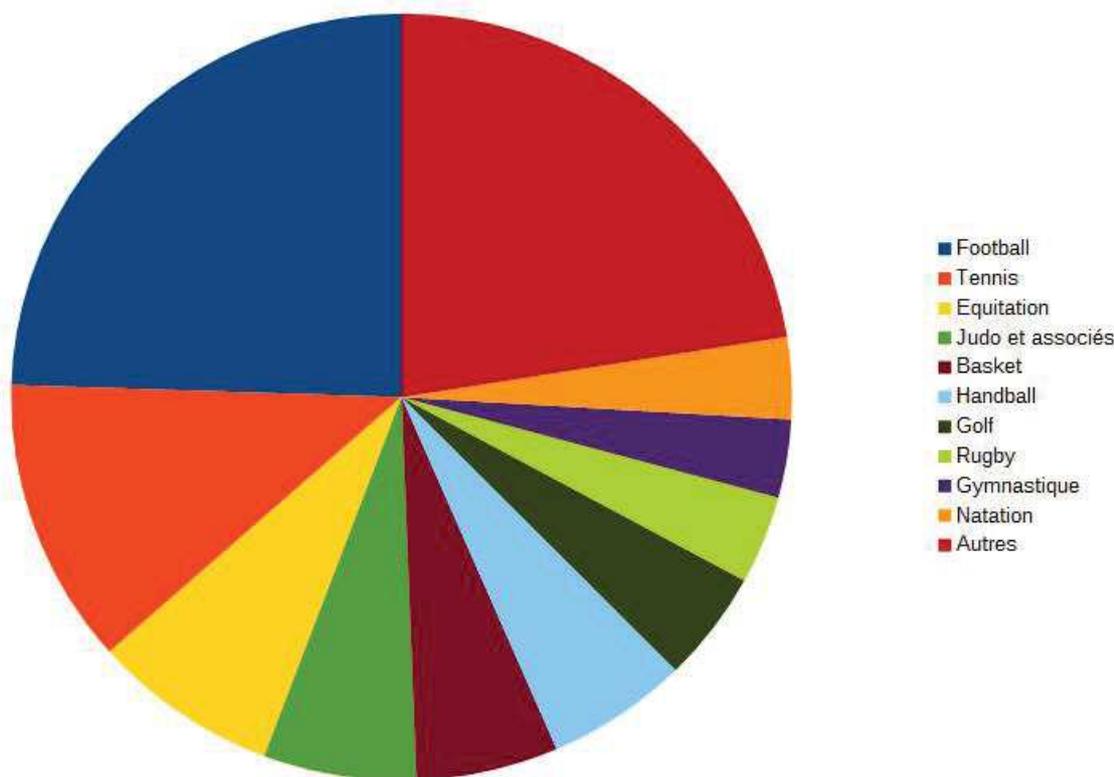


Figure – Diagramme en secteurs des licenciés en 2016.

Détection de problèmes et transformation de données

- **Détection de problèmes** : un des buts de l'analyse de données est de repérer les **données aberrantes** (résultant par exemple d'une erreur de mesure) ; on parle parfois "d'outlier".
- **Transformation des données** : afin de rendre les données plus faciles à lire, mais surtout pour **normaliser la distribution** (afin de satisfaire certaines hypothèses, voir les chapitres sur la statistique inférentielle), on applique parfois une transformation aux données.
Le plus souvent, une **transformation logarithmique**, mais parfois il est parfois plus judicieux de prendre la racine carrée, le carré ou l'inverse des observations.

Transformation logarithmique des données

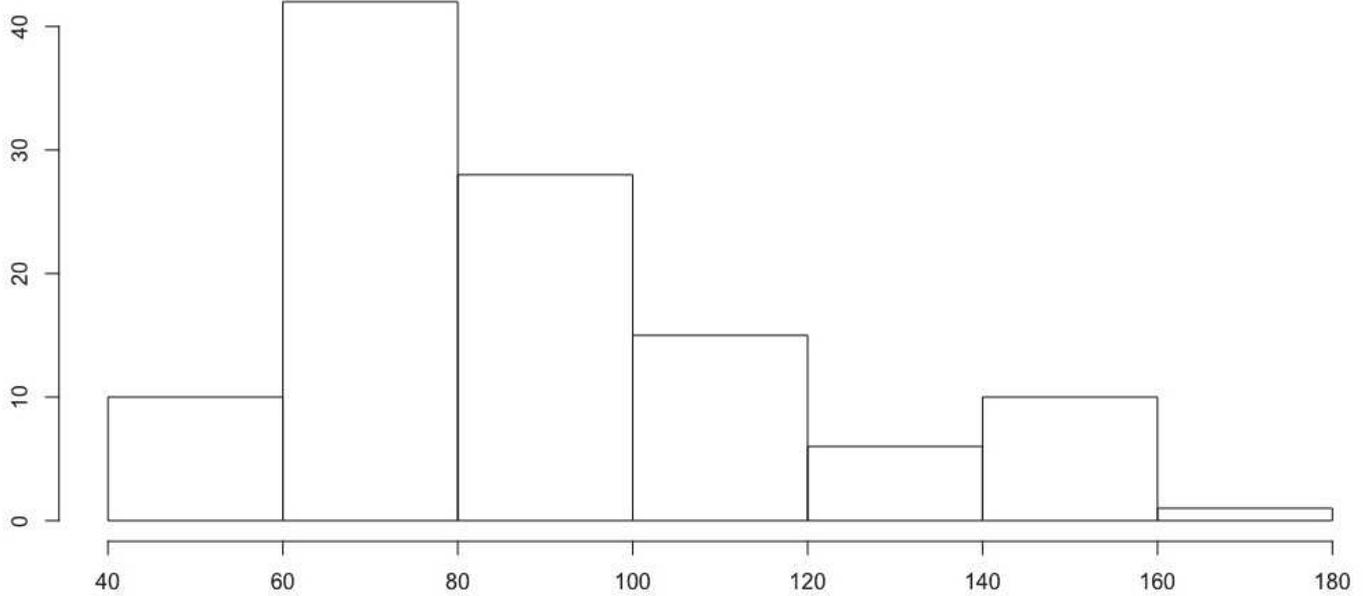


Figure – Histogramme des données originales.

Transformation logarithmique des données

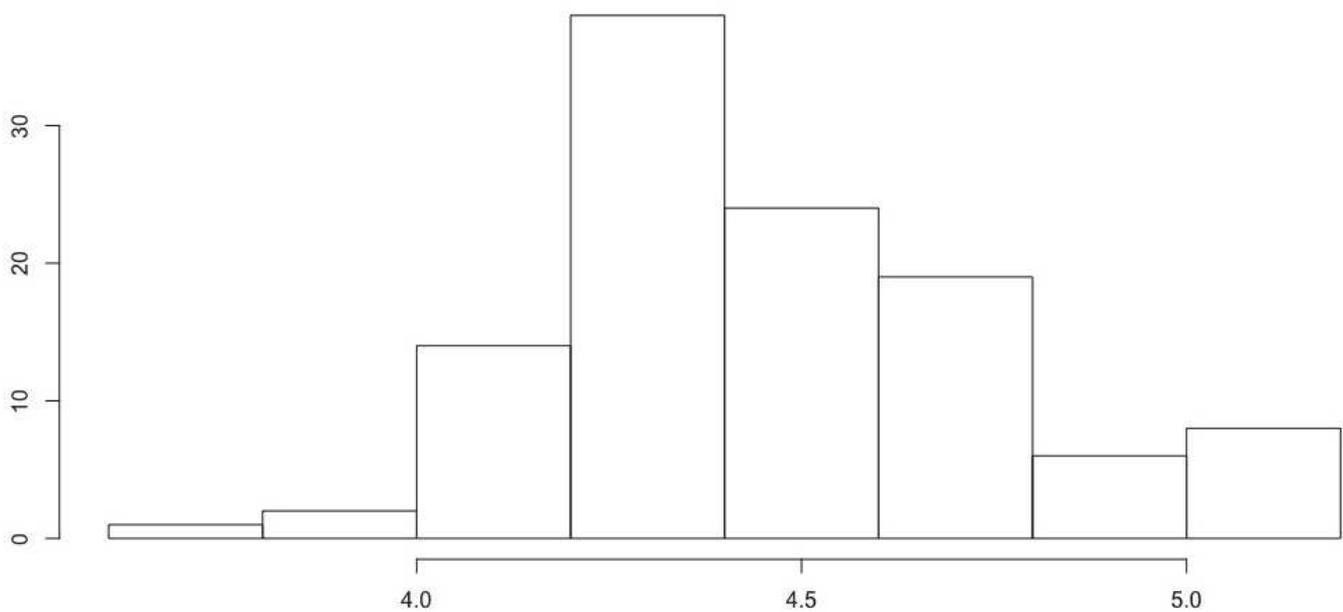


Figure – Histogramme des données transformées.

On enchaîne avec...

1 Statistique descriptive unidimensionnelle

- Une variable quantitative
 - Une variable quantitative discrète
 - Une variable quantitative continue
- Une variable qualitative
- Détection de problèmes et transformation des données

2 Statistique descriptive bidimensionnelle

- Deux variables quantitatives
- Une variable quantitative et une variable qualitative
- Deux variables qualitatives

3 Statistique descriptive multidimensionnelle

- Généralisation du cadre bidimensionnel
- Introduction à l'Analyse en Composantes Principales

Deux variables quantitatives : un exemple

Joueur	Minutes jouées	Evaluation	Passes décisives	Contré
Pearson	853	454	49	6
Brown	968	427	52	5
Sulaimon	876	304	85	9
Bigote	739	336	46	8
Alingue	834	516	68	6
Julien	819	312	162	3
Holston	388	219	96	0
Loum	464	203	12	0
Frazier	335	116	41	4
Taylor	314	99	10	5
Kennedy	209	117	11	3
Dorez	13	-3	0	2
Burrell	13	0	0	0

Figure – Quelques statistiques des joueurs de l'équipe de basket de Pau-Lacq-Orthez lors de la saison 2017-2018. Source : LNB.

Nuage de points : temps de jeu et évaluation

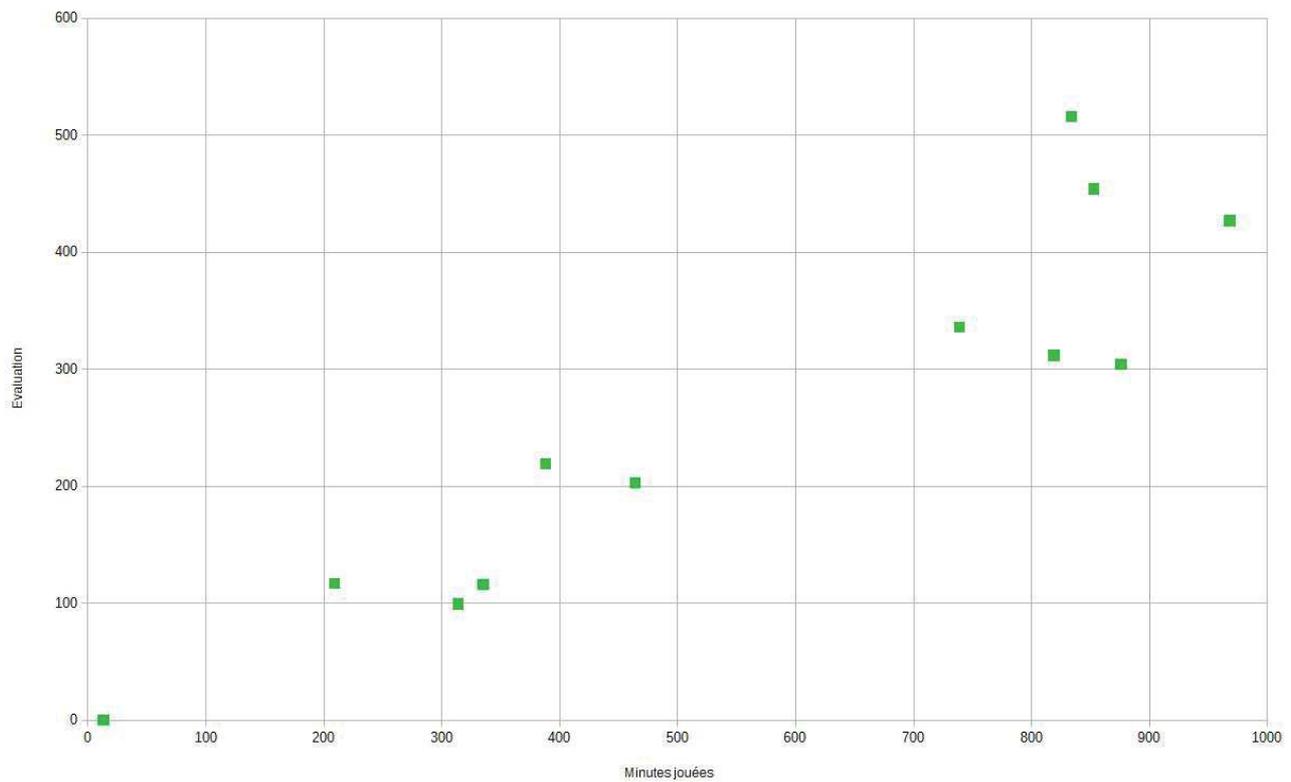


Figure – Nuage de points illustrant la corrélation entre le temps de jeu et l'évaluation.

Nuage de points : passes décisives et contres subis

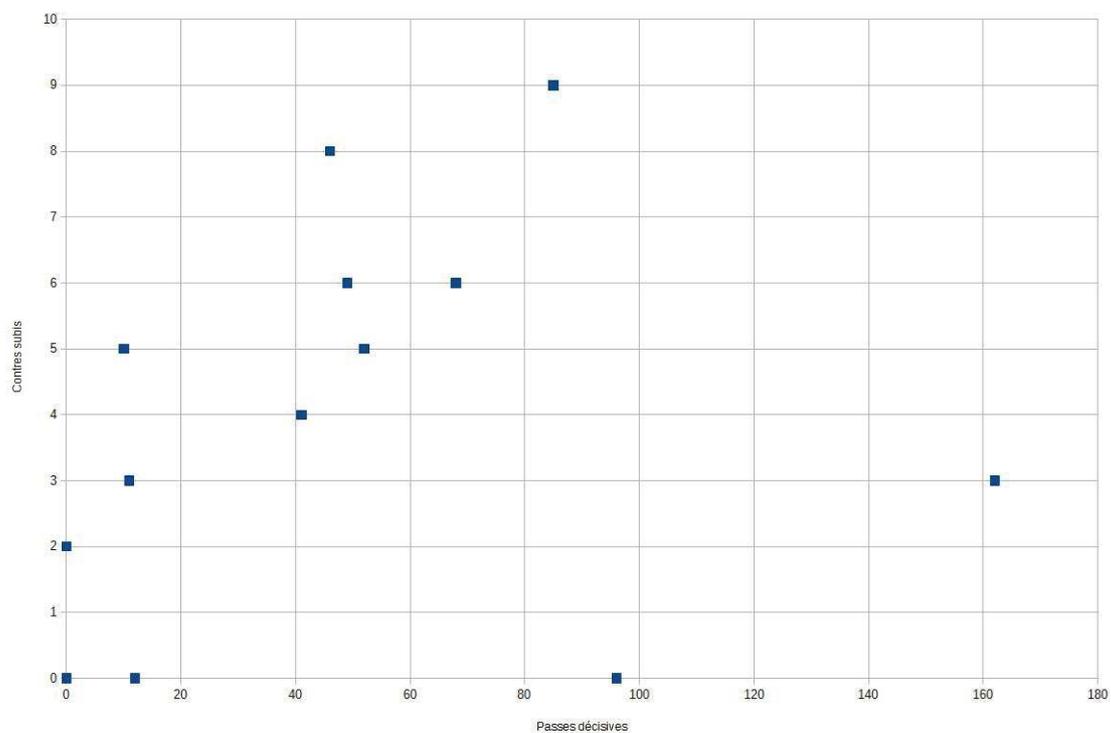


Figure – Nuage de points illustrant la faible corrélation entre les nombres de passes décisives et de contres subis.

Corrélation entre deux variables quantitatives

On considère deux variables X et Y définies sur un même ensemble Ω à valeurs dans \mathbb{R} . \rightarrow dans notre exemple, Ω est l'ensemble des 13 joueurs ; cas 1 : X_1 correspond aux minutes jouées, et Y_1 à l'évaluation ; cas 2 : X_2 aux passes décisives et Y_2 aux contres subis.

- **Covariance** de X et Y : généralisation de la variance, définie par

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{x} \bar{y}.$$

\rightarrow exemples : $\text{cov}(X_1, Y_1) \approx 49777$ et $\text{cov}(X_2, Y_2) \approx 26,82$.

Remarque : "moyenne des produits des écarts aux moyennes", et dépend des unités. Pas de signification concrète.

\rightarrow Indicateur interprétable : **coefficient de corrélation**.

Corrélation entre deux variables quantitatives

On considère deux variables X et Y définies sur un même ensemble Ω à valeurs dans \mathbb{R} . \rightarrow dans notre exemple, Ω est l'ensemble des 13 joueurs ; cas 1 : X_1 correspond aux minutes jouées, et Y_1 à l'évaluation ; cas 2 : X_2 aux passes décisives et Y_2 aux contres subis.

- **Coefficient de corrélation linéaire** de X et Y : défini par

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

\rightarrow exemples : $\text{corr}(X_1, Y_1) \approx 0,94$ et $\text{corr}(X_2, Y_2) \approx 0,21$.

Remarque : coefficient varie entre -1 et 1 ; plus sa valeur absolue est grande plus la liaison est forte ; le signe indique le sens.

Une variable qualitative et une variable quantitative

Cadre : une **variable qualitative** X définie sur Ω à r modalités m_1, \dots, m_r et une **variable quantitative** Y également définie sur Ω , de moyenne \bar{y} et de variance σ_Y^2 .

↪ exemple : X réponse à "chocolatine ou pain au chocolat ?" avec 2 modalités, et Y la note au premier contrôle de thermodynamique.

- **Partition de Ω avec les modalités de X** : $\Omega = \cup_{j=1}^r \Omega_j$ avec Ω_j de cardinal n_j ensemble des individus pour lesquels on a observé m_j .
↪ exemple : Ω_1 les 6 individus qui ont répondu "chocolatine", Ω_2 les 4 qui ont répondu "pain au chocolat".
- Sur chaque Ω_j , \bar{y}_j **moyenne partielle** et σ_j^2 **variance partielle**.
↪ exemple : $\bar{y}_1 = 14$, $\bar{y}_2 = 9,5$, $\sigma_1^2 = 22,4$ et $\sigma_2^2 = 17$.

Une variable qualitative et une variable quantitative

- **Formules de décomposition** :

$$\text{Moyenne : } \bar{y} = \frac{1}{n} \sum_{j=1}^r n_j \bar{y}_j$$

$$\text{Variance : } \sigma_Y^2 = \frac{1}{n} \sum_{j=1}^r n_j (\bar{y}_j - \bar{y})^2 + \frac{1}{n} \sum_{j=1}^r n_j \sigma_j^2 = \sigma_E^2 + \sigma_R^2$$

↪ exemple : $\bar{y}_1 = 14$, $\sigma_Y^2 = 18,4$ et $\sigma_E^2 = 4,86$.

- **Rapport de corrélation** : défini par $c_{Y/X} = \frac{\sigma_E}{\sigma_Y}$, varie entre 0 et 1.

↪ exemple : $c_{Y/X} = 0,51$.

Plus le coefficient de corrélation est proche de 1, plus la liaison entre les deux variables est forte.

Un exemple : buts marqués et poste de jeu

N°	Poste	Buts
10	Milieu	12
11	Attaquant	44
19	Attaquant	20
9	Attaquant	27
5	Milieu	2
21	Milieu	5
14	Milieu	1
6	Défenseur	2
23	Milieu	6
32	Défenseur	1
26	Défenseur	1
12	Défenseur	1
7	Milieu	1
1	Gardien	0
22	Gardien	0
66	Défenseur	3
4	Défenseur	1
17	Défenseur	1
29	Attaquant	1
15	Attaquant	3

- X : poste de jeu ; 4 modalités : $m_1 =$ gardien, $m_2 =$ défenseur, $m_3 =$ milieu, $m_4 =$ attaquant.
- Y : buts marqués ; $\bar{y} = 6,6$ et $\sigma_Y \approx 11,03$.
- Moyennes partielles : $\bar{y}_1 = 0$, $\bar{y}_2 = 1,4$, $\bar{y}_3 = 4,5$ et $\bar{y}_4 = 19$.
- Variances partielles : $\sigma_1^2 = 0$, $\sigma_2^2 = 0,5$, $\sigma_3^2 = 14,0$ et $\sigma_4^2 = 254$.
- Rapport de corrélation : $\sigma_{Y/X} \approx 0,89$.

Figure – Buteurs du Liverpool FC lors de la saison 2017/2018.

Représentation graphique : boîtes parallèles

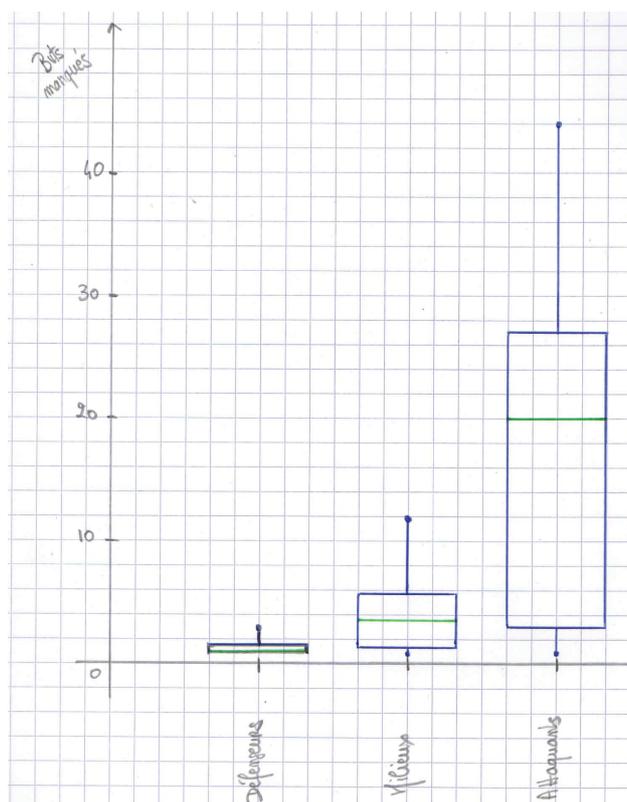


Figure – Diagrammes-boîtes parallèles selon les postes de jeu.

Deux variables qualitatives : table de contingence

On considère deux variables qualitatives observés simultanément sur n individus : X a m modalités x_1, \dots, x_m et Y a r modalités y_1, \dots, y_r .

- **Effectif conjoint** $n_{i,j}$: nombre d'observations des modalités x_i et y_j .
- **Effectifs marginaux** n_{i+} et n_{+j} : définis par $n_{i+} = \sum_{j=1}^r n_{i,j}$ et $n_{+j} = \sum_{i=1}^m n_{i,j}$.
- **Table de contingence** :

	y_1	...	y_j	...	y_r	total
x_1	$n_{1,1}$...	$n_{1,j}$...	$n_{1,r}$	n_{1+}
...
x_i	$n_{i,1}$...	$n_{i,j}$...	$n_{i,r}$	n_{i+}
...
x_m	$n_{m,1}$...	$n_{m,j}$...	$n_{m,r}$	n_{m+}
total	n_{+1}	...	n_{+j}	...	n_{+r}	n

Un exemple : lecture et genres

	Roman	Histoire	BD	Arts de vivre	Sciences	Dictionnaires	Total
Hommes	58	50	55	47	44	42	296
Femmes	81	42	46	62	57	43	331
Total	139	92	101	109	101	85	627

Figure – Préférence de lecture selon les genres. Chiffres inexacts.

Profils et représentations graphiques

- **i ème profil-ligne** : $\left\{ \frac{n_{i,1}}{n_{i+}}, \dots, \frac{n_{i,r}}{n_{i+}} \right\}$.

↪ dans notre exemple : en exprimant les valeurs en pourcentages, pour la ligne "Hommes" : {20,17,19,16,15,14}, pour la ligne "Femmes" : {24,13,14,19,17,13}.

- **j ème profil-colonne** : $\left\{ \frac{n_{1,j}}{n_{+j}}, \dots, \frac{n_{m,j}}{n_{+j}} \right\}$.

↪ dans notre exemple : en exprimant les valeurs en pourcentages, pour la colonne "Roman" : {42,58}, pour la colonne "Histoire" : {54,46}, pour la colonne "BD" : {54,46}, pour la colonne "Arts de vivre" : {43,57}, pour la colonne "Sciences" : {44,56} et pour la colonne "Dictionnaires" : {49,51}.

Diagramme en barres des profils-lignes

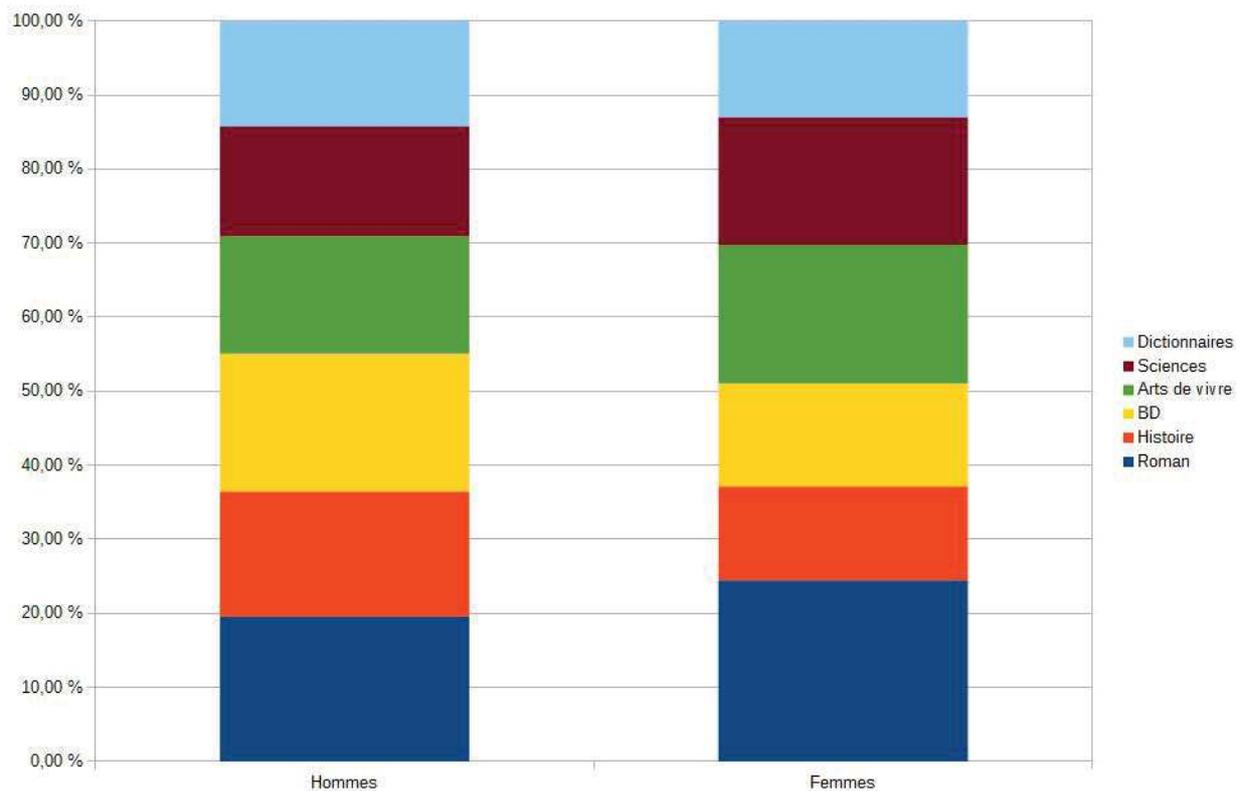


Figure – Diagramme en barres des profils-lignes

Diagramme en barres des profils-colonnes

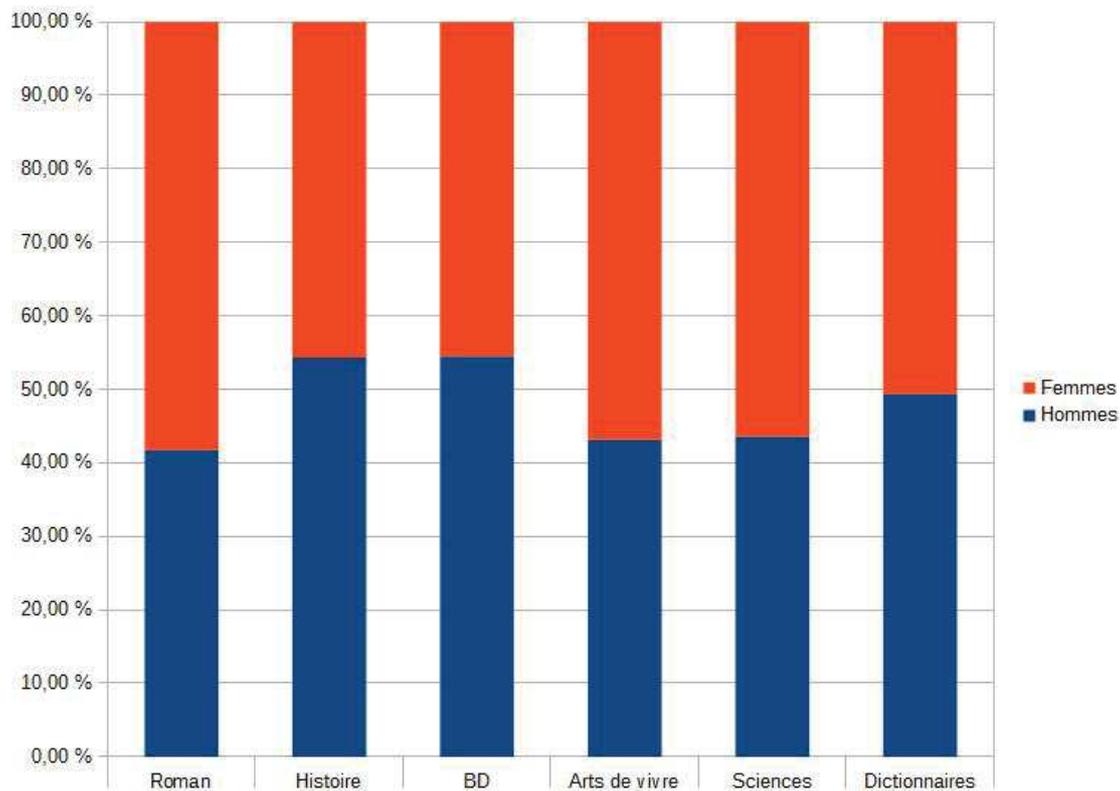


Figure – Diagramme en barres des profils-colonnes

Indices de liaison

Comment jauger la force de la liaison entre X et Y ?

- **Aucune liaison** entre X et Y \Leftrightarrow tous les profils-lignes sont égaux
 \Leftrightarrow tous les profils-colonnes sont égaux \Leftrightarrow pour tous i et j , $n_{i,j} = \frac{n_{i+} n_{+j}}{n}$.
- **Indicateur khi-deux** : défini par

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^r \frac{\left(n_{i,j} - \frac{n_{i+} n_{+j}}{n} \right)^2}{\frac{n_{i+} n_{+j}}{n}};$$

la liaison est d'autant plus grande qu'il est important.

↪ dans notre cas, on trouve 7,12.

- Il existe d'autres indicateurs : le **phi-deux**, le **coefficient T de Tschuprow**, le **coefficient C de Cramer**, ...

On enchaîne avec...

1 Statistique descriptive unidimensionnelle

- Une variable quantitative
 - Une variable quantitative discrète
 - Une variable quantitative continue
- Une variable qualitative
- Détection de problèmes et transformation des données

2 Statistique descriptive bidimensionnelle

- Deux variables quantitatives
- Une variable quantitative et une variable qualitative
- Deux variables qualitatives

3 Statistique descriptive multidimensionnelle

- Généralisation du cadre bidimensionnel
- Introduction à l'Analyse en Composantes Principales

Statistique descriptive multidimensionnelle : problématique

- Comment traiter simultanément plus de 2 variables ?
- Peut-on "généraliser" le cas bidimensionnel ?
- Comment représenter graphiquement la structure des corrélations entre ces facteurs ?

↔ Les p variables considérées auront même nature : ou toutes quantitatives, ou toutes qualitatives.

Pour p variables qualitatives : le tableau de Burt.

→ Généralisation de la **table de contingence**.

	Chocolatine	Pain au cho.	BD	Roman	Foot	Rugby
Chocolatine	154	0	86	68	80	74
Pain au cho.	0	132	60	72	58	74
BD	86	60	146	0	70	76
Roman	68	72	0	140	68	72
Foot	80	58	70	68	138	0
Rugby	74	74	76	72	0	148

Figure – Tableau de Burt. *Chiffres totalement fictifs.*

Pour p variables quantitatives : matrices de corrélation.

On reprend un exemple déjà vu :

Joueur	Minutes jouées	Evaluation	Passes décisives	Contré
Pearson	853	454	49	6
Brown	968	427	52	5
Sulaimon	876	304	85	9
Bigote	739	336	46	8
Alingue	834	516	68	6
Julien	819	312	162	3
Holston	388	219	96	0
Loum	464	203	12	0
Frazier	335	116	41	4
Taylor	314	99	10	5
Kennedy	209	117	11	3
Dorez	13	-3	0	2
Burrell	13	0	0	0

Figure – Quelques statistiques des joueurs de l'équipe de basket de Pau-Lacq-Orthez lors de la saison 2017-2018. *Source : LNB.*

Matrices des covariances et de corrélations

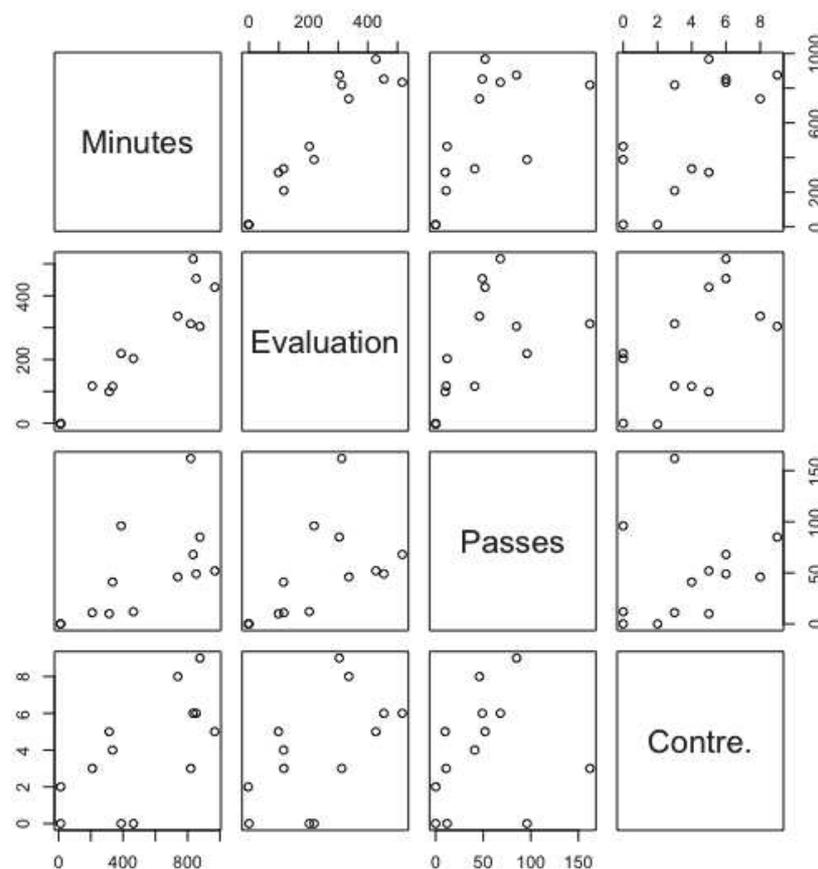
	Minutes jouées	Evaluation	Passes décisives	Contres subis
Minutes jouées	106308	49777	9244	616
Evaluation	49777	26481	3936	266
Passes décisives	9244	3936	1989	27
Contres subis	616	266	27	8

Figure – Matrice des variances/covariances.

	Minutes jouées	Evaluation	Passes décisives	Contres subis
Minutes jouées	1,00	0,94	0,64	0,66
Evaluation	0,94	1,00	0,54	0,58
Passes décisives	0,64	0,54	1,00	0,21
Contres subis	0,66	0,58	0,21	1,00

Figure – Matrice des corrélations.

Représentation graphique : matrice des nuages de points



Chameau ou dromadaire ?

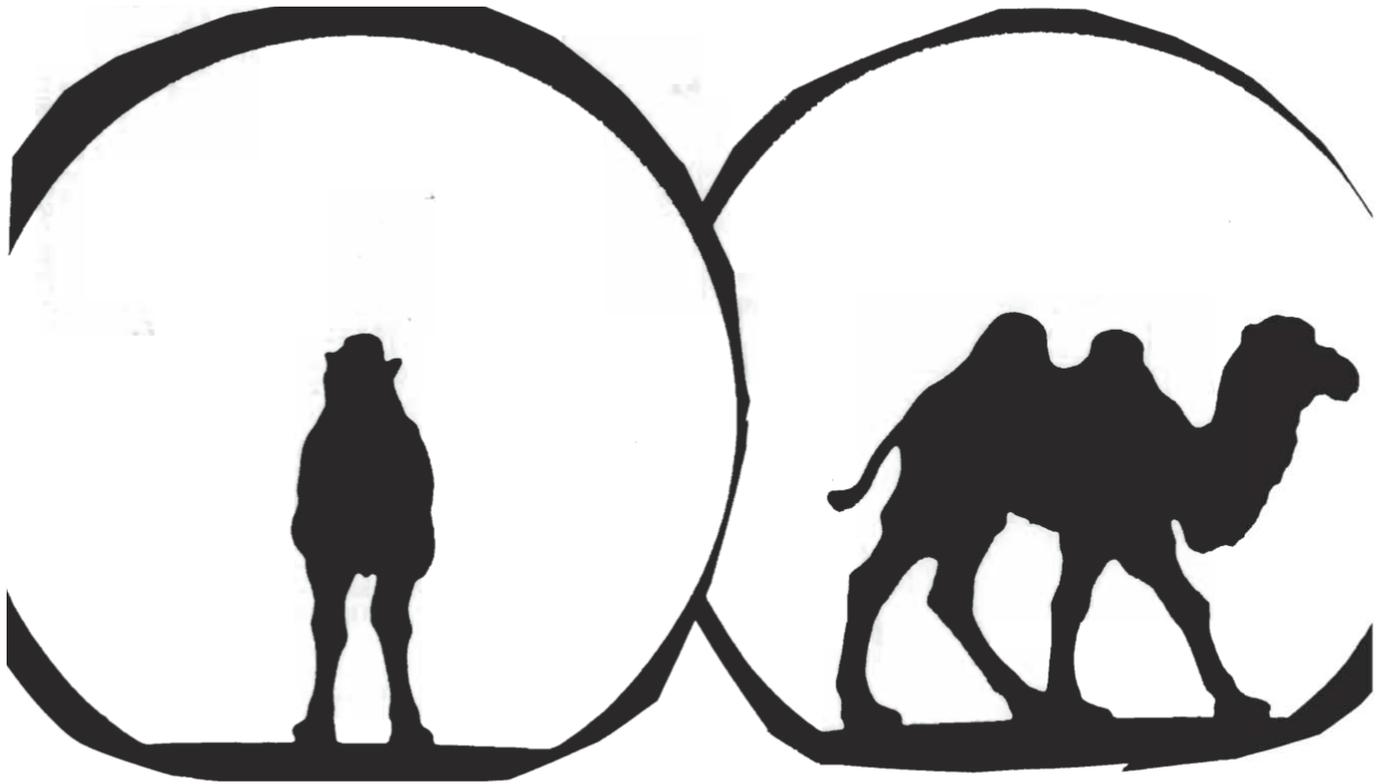


Figure – Source : J.P. Fenelon

Analyse en composantes principales.

- Idée : se ramener à un espace de dimension 2 en déformant le moins possible la réalité ; "**trouver l'angle de vue optimal**".
- Mathématiquement : **diagonalisation de la matrice des covariances (ou des corrélations)**, ie changement de base de la base canonique vers la base des vecteurs propres.

Un exemple : littéraires ou scientifiques ?

X	Français	Maths	Physique	Anglais
1 Nathan	5.0	6.0	6.0	5.5
2 Floris	8.0	8.0	8.0	8.0
3 Elsa	11.0	6.0	7.0	9.5
4 Hicham	15.5	14.5	14.5	15.0
5 Faustine	12.0	14.0	14.0	12.5
6 Bastien	5.5	11.0	10.0	7.0
7 Carmen	14.0	5.5	7.0	11.5
8 Maguelo...	8.5	13.0	12.5	9.5
9 Timéo	12.5	9.0	9.5	12.0

Figure – Notes d'étudiants, données fictives.

Un exemple : littéraires ou scientifiques ?

Français	Maths	Physique	Anglais
Min. : 5.00	Min. : 5.500	Min. : 6.000	Min. : 5.50
1st Qu.: 8.00	1st Qu.: 6.000	1st Qu.: 7.000	1st Qu.: 8.00
Median :11.00	Median : 9.000	Median : 9.500	Median : 9.50
Mean :10.22	Mean : 9.667	Mean : 9.833	Mean :10.06
3rd Qu.:12.50	3rd Qu.:13.000	3rd Qu.:12.500	3rd Qu.:12.00
Max. :15.50	Max. :14.500	Max. :14.500	Max. :15.00

Figure – Statistiques élémentaires. Logiciel : R Studio

```
Français Maths Physique Anglais
3.683673 3.579455 3.172144 2.983752
```

Figure – Ecart types. Logiciel : R Studio

NB : homogénéité des 4 variables considérées.

Un exemple : littéraires ou scientifiques ?

	Français	Maths	Physique	Anglais
Français	13.569444	2.989583	4.635417	10.454861
Maths	2.989583	12.812500	11.156250	5.427083
Physique	4.635417	11.156250	10.062500	6.166667
Anglais	10.454861	5.427083	6.166667	8.902778

Figure – Matrices des variances/covariances. *Logiciel : R Studio*

	Français	Maths	Physique	Anglais
Français	1.0000000	0.2267319	0.3966932	0.9512058
Maths	0.2267319	1.0000000	0.9825357	0.5081440
Physique	0.3966932	0.9825357	1.0000000	0.6515305
Anglais	0.9512058	0.5081440	0.6515305	1.0000000

Figure – Matrices des corrélations. *Logiciel : R Studio*

NB : tous les coefficients de corrélation sont positifs.

Un exemple : littéraires ou scientifiques ?

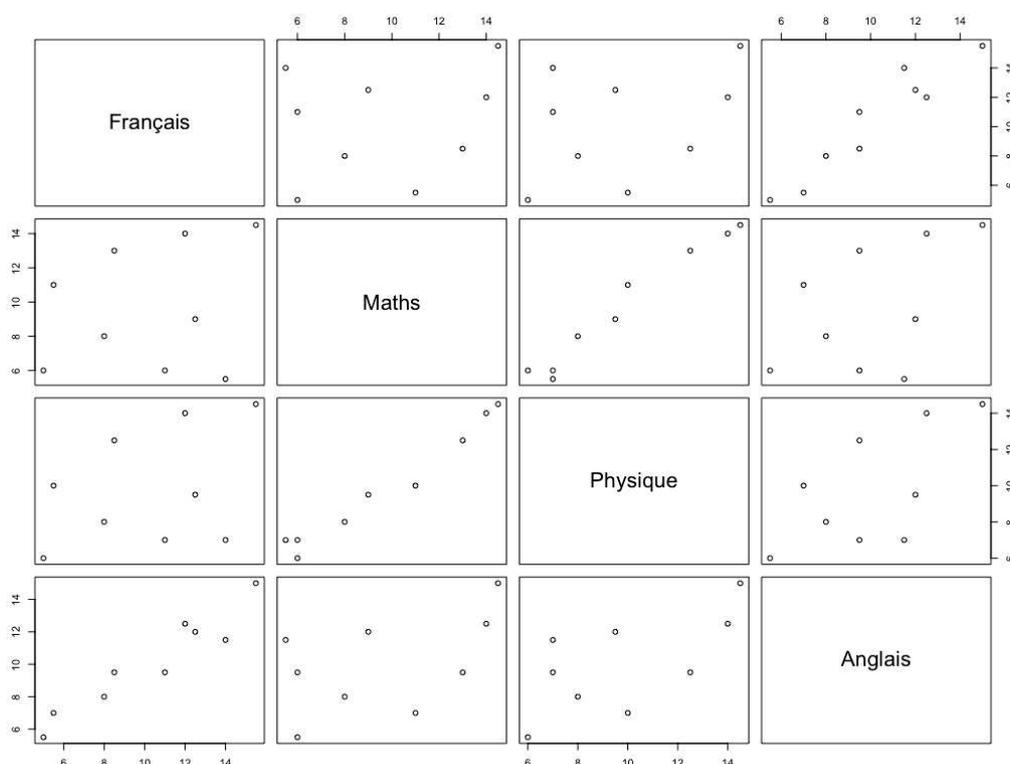


Figure – Matrices des nuages de points. *Logiciel : R Studio*

Un exemple : littéraires ou scientifiques ?

Idée : diagonaliser la matrice des covariances.

```
Svalues
[1] 31.76423012 13.53436431 0.03671101 0.01191678

      PC1      PC2      PC3      PC4
Français 0.4922789 0.6581534 0.4603842 -0.3354728
Maths    0.5151694 -0.5686517 -0.1852853 -0.6139259
Physique 0.5076129 -0.3712665 0.4499844 0.6340381
Anglais  0.4843461 0.3250085 -0.7424485 0.3294671
```

Figure – Valeurs propres et vecteurs propres de la matrices des covariances. *Logiciel : R Studio*

- Les vecteurs propres (ici PC1, PC2, PC3 et PC4) sont appelés **axes principaux**.
- Les **facteurs principaux** sont les **variables** (virtuelles) obtenues par projection sur les axes principaux. ↔ Ici, correspondent à des "matières virtuelles". Par exemple, pour le premier facteur, $0,49 \times \text{Français} + 0,51 \times \text{Maths} + 0,51 \times \text{Physique} + 0,48 \times \text{Anglais}$.

Un exemple : littéraires ou scientifiques ?

Y	PC1	PC2	PC3	PC4
1 Nathan	-8.612059	-1.4093727	-0.06752404	0.07158969
2 Floris	-3.878793	-0.5022279	-0.01309446	-0.07093634
3 Elsa	-3.213388	3.4683149	0.17497150	0.01065973
4 Hicham	9.851807	0.5995132	-0.03680819	-0.14998275
5 Faustine	6.406574	-2.0465857	0.07561885	0.19044801
6 Bastien	-3.033102	-4.9211080	-0.07749344	-0.13542301
7 Carmen	-1.025444	6.3771179	0.16386970	-0.02986136
8 Maguelone	1.953971	-4.1995965	0.20192835	0.03907002
9 Timéo	1.550436	2.6339447	-0.42146828	0.07443601

Figure – Composantes principales. *Logiciel : R Studio*.

- Les **composantes principales** correspondent aux réalisations des facteurs principaux pour chaque individu. ↔ Ici, chaque coordonnée d'une composante principale correspond à la "note virtuelle" obtenue par l'étudiant dans la "matière virtuelle" correspondante, après recentrage.

Un exemple : littéraires ou scientifiques ?



Figure – Valeur propre (variance) associée à chaque composante. *Logiciel : R Studio.*

- Somme des variances des variables = somme des valeurs propres.
Ici, répartie essentiellement sur les deux premières valeurs propres.
↪ la quasi-totalité de l'information est contenue **dans les deux premiers facteurs**.
- On peut donc **négliger l'impact des deux dernières composantes**.

Un exemple : littéraires ou scientifiques ?

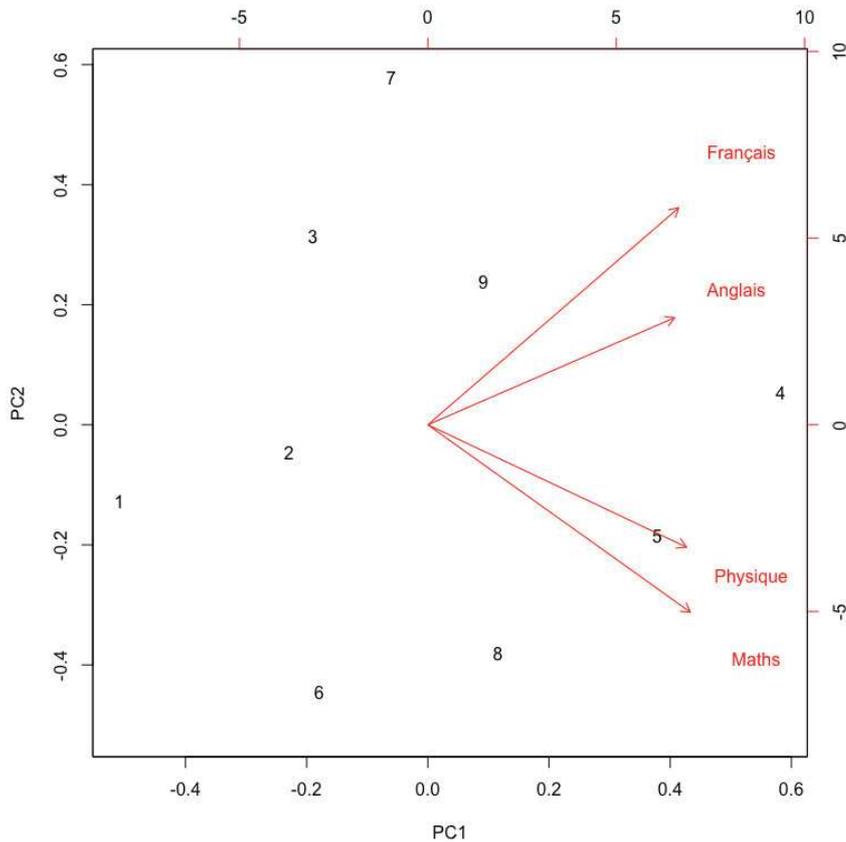
Objectif : interprétation des composantes principales.

- Calcul des coefficients de corrélation entre les "anciennes" variables et les "nouvelles" variables (facteurs) :

Variable	Facteur1	Facteur2	Facteur3	Facteur4
Français	0,753	0,657	0,032	-0,012
Maths	0,811	-0,584	-0,012	-0,021
Physique	0,902	-0,430	0,026	0,019
Anglais	0,915	0,401	-0,041	0,009

Figure – Tableau de corrélation variables-facteurs.

- Première composante : corrélée positivement avec toutes les matières.
↪ **d'autant plus grande que les notes sont bonnes.**
- Deuxième composante : corrélée positivement avec les maths et la physique, négativement avec le français et l'anglais.
↪ **valeur importante si l'étudiant a de bons résultats dans les matières littéraires et de mauvaises notes dans les matières scientifiques**



- Carmen (7) a de bonnes notes dans les matières littéraires, pas dans les matières scientifiques.
- A contrario, Maguelone (8) a de bonnes notes en maths et en physique, et de mauvaises en français et en anglais.
- Hicham (4) a lui de bons résultats dans toutes les matières, à l'inverse de Nathan (1).

Figure – Représentation graphique. Logiciel : R Studio.

Application : analyse des statistiques au basket

Ici, **données centrées et réduites** car variables non homogènes.

Joueur	Minutes jouées	Evaluation	Passes décisives	Contré
Pearson	853	454	49	6
Brown	968	427	52	5
Sulaimon	876	304	85	9
Bigote	739	336	46	8
Alingue	834	516	68	6
Julien	819	312	162	3
Holston	388	219	96	0
Loum	464	203	12	0
Frazier	335	116	41	4
Taylor	314	99	10	5
Kennedy	209	117	11	3
Dorez	13	-3	0	2
Burrell	13	0	0	0

Figure – Quelques statistiques des joueurs de l'équipe de basket de Pau-Lacq-Orthez lors de la saison 2017-2018. Source : LNB.

Application : analyse des statistiques au basket

	PC1	PC2	PC3	PC4
Minutes.jouées	-0.5809227	0.004618784	-0.2261835	-0.7818878
Evaluation	-0.5531577	0.015536543	-0.5947605	0.5831253
Passes.décisives	-0.4136083	-0.734105067	0.5161478	0.1536534
Contres.subis	-0.4306664	0.678842422	0.5733171	0.1581358

Figure – Facteurs principaux. *Logiciel : R Studio.*

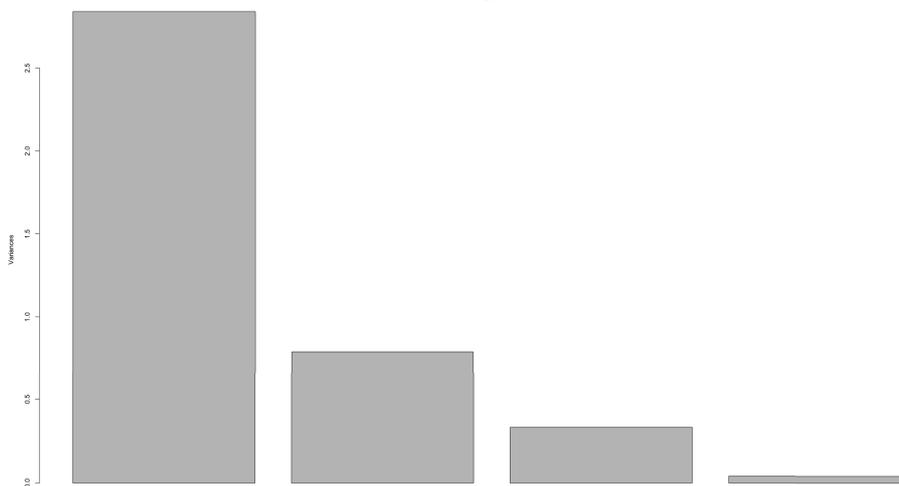


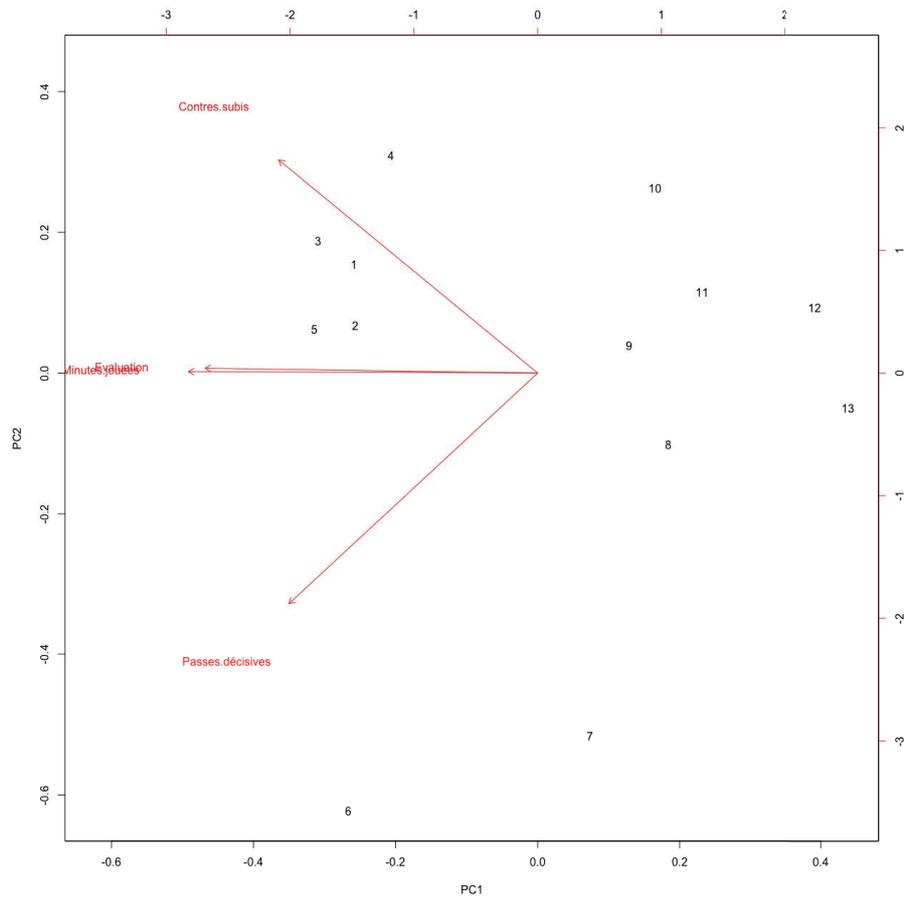
Figure – Valeurs propres / variance des facteurs. *Logiciel : R Studio.*

Application : analyse des statistiques au basket

	Y	PC1	PC2	PC3	PC4
1	Pearson	-1.5713191	0.4949611	-0.56851108	0.098697583
2	Brown	-1.5610864	0.2170254	-0.71087304	-0.302765363
3	Sulaimon	-1.8785488	0.6008461	0.92488904	-0.191102037
4	Bigote	-1.2553544	0.9891844	0.27624215	0.052126357
5	Alingue	-1.9105946	0.1998768	-0.56227180	0.418827944
6	Julien	-1.6194012	-1.9945123	0.62774231	-0.098199228
7	Holston	0.4471960	-1.6537464	-0.07404453	0.195707478
8	Loum	1.1179013	-0.3255963	-1.00263821	-0.312562835
9	Frazier	0.7818488	0.1242889	0.48688742	-0.004952189
10	Taylor	1.0039217	0.8423289	0.40973085	-0.064242562
11	Kennedy	1.4072544	0.3675850	0.03985091	0.135999253
12	Dorez	2.3783448	0.2983164	0.27565284	0.084547831
13	Burrell	2.6598375	-0.1605579	-0.12265686	-0.012082232

Figure – Composantes principales. *Logiciel : R Studio.*

Application : analyse des statistiques au basket



Chapitre 3 : Notions de probabilités

Programme des réjouissances

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

On enchaîne avec...

- 1 Événements et probabilités
 - Événements
 - Probabilités
 - Indépendance
- 2 Variables aléatoires réelles
 - Définitions
 - Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
 - Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles
- 3 Indépendance, moments, théorèmes limites
 - Indépendance de variables aléatoires
 - Propriétés de l'espérance et de la variance
 - Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Expériences et événements aléatoires

- **Expérience aléatoire** : expérience dont on ne peut pas prévoir le résultat, qui peut avoir plusieurs issues possibles.
↔ exemples : lancer d'un dé, attente d'un bus, durée de vie.
Mais aussi : fiabilité d'un système, fatigue des matériaux, bruitage des télécommunications.
- **Univers** : ensemble des résultats possibles d'une expérience aléatoire, noté Ω .
↔ dans le cas d'un lancer de dé, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- **Événement aléatoire** : partie de l'ensemble des résultats possibles, ie sous-ensemble de l'univers Ω . L'événement A est réalisé si le résultat ω de l'expérience appartient à A . ↔ ex : dans le cas d'un lancer de dé, $A = \{1, 5, 6\}$.

Notation ensemblistes et événements

- **Événement impossible** : $A = \emptyset$.
- **Événement certain** : $A = \Omega$.
- **Événement contraire** : $A^c = \{\omega \in \Omega \mid \omega \notin A\}$.
- **Union d'événements** : $A \cup B = \{\omega \in \Omega \mid \omega \in A \text{ ou } \omega \in B\}$.
- **Intersection d'événements** : $A \cap B = \{\omega \in \Omega \mid \omega \in A \text{ et } \omega \in B\}$.
- **Inclusion d'événements** : $A \subset B$ si $\omega \in A \Rightarrow \omega \in B$.
- **Incompatibilité d'événements** : A et B sont incompatibles (ou disjoints) si $A \cap B = \emptyset$.

On enchaîne avec...

- 1 Événements et probabilités
 - Événements
 - Probabilités
 - Indépendance
- 2 Variables aléatoires réelles
 - Définitions
 - Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
 - Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles
- 3 Indépendance, moments, théorèmes limites
 - Indépendance de variables aléatoires
 - Propriétés de l'espérance et de la variance
 - Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Probabilité : définition et premières propriétés

Considérons Ω l'univers associé à une expérience aléatoire et \mathcal{A} l'ensemble des parties de Ω (i.e. l'ensemble des événements).

Une **probabilité** \mathbb{P} sur l'espace (Ω, \mathcal{A}) est une application de \mathcal{A} dans $[0, 1]$ telle que :

- $\mathbb{P}(\Omega) = 1$.
- Pour une famille d'événements $(A_n)_n$ 2 à 2 incompatibles,

$$\mathbb{P}(\cup_n A_n) = \sum_n \mathbb{P}(A_n).$$

Propriétés fondamentales

- $\mathbb{P}(\emptyset) = 0$
- $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

Probabilités conditionnelles

Considérons deux événements aléatoires A et B avec $\mathbb{P}(B) \neq 0$.

La **probabilité conditionnelle de A sachant B** est définie par

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Propriétés fondamentales

- $\mathbb{P}(B|B) = 1$
- $\mathbb{P}(A|B) + \mathbb{P}(A^c|B) = 1$
- $\mathbb{P}(B|A) = \frac{\mathbb{P}(B)}{\mathbb{P}(A)} \mathbb{P}(A|B)$

Formule des probabilités totales

Si les $(A_i)_i$ forment une partition de Ω et $\mathbb{P}(A_i) \neq 0$ pour tout i , alors

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A|A_i)\mathbb{P}(A_i).$$

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- **Indépendance**

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Événements indépendants

Deux événements A et B sont dits **indépendants** si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Propriétés fondamentales

- Si A et B sont indépendants, $\mathbb{P}(A|B) = \mathbb{P}(A)$.
- Si A et B sont indépendants, alors A^c et B^c sont indépendants, tout comme A et B^c , et A^c et B .

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Variable aléatoire et loi de probabilité

On se place sur l'univers Ω , muni d'une probabilité \mathbb{P} .

- **Variable aléatoire réelle** : une application X de l'univers Ω , associé à une expérience aléatoire, à valeurs dans \mathbb{R} :

$$X : \omega \in \Omega \mapsto X(\omega) \in \mathbb{R}.$$

↪ exemple : expérience : lancer d'un dé ; on considère

$$X : \omega \in \{1, 2, 3, 4, 5, 6\} \mapsto \begin{cases} 0 & \text{si } \omega \in \{1, 2\}, \\ 1 & \text{si } \omega \in \{3, 4\}, \\ 0 & \text{si } \omega \in \{5, 6\}. \end{cases}$$

- **Loi de probabilité** de X : application, notée \mathbb{P}_X qui à une partie A de \mathbb{R} associe $\mathbb{P}_X(A) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\})$ (aussi noté $\mathbb{P}(X \in A)$).

↪ exemple : $\mathbb{P}_X(\{0, 2\}) = 2/3$.

NB : on note également $\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\})$.

NB2 : \mathbb{P}_X définit une probabilité sur \mathbb{R} .

Fonction de répartition

On considère une variable aléatoire réelle X définie sur Ω .

- **Fonction de répartition** de X : application de \mathbb{R} dans $[0, 1]$, notée F_X , définie par

$$F_X(x) = \mathbb{P}(X \leq x).$$

↪ exemple : $F_X(-2) = 0$, $F_X(0) = 1/3$, $F_X(1/2) = 1/3$, $F_X(1) = 2/3$, $F_X(2) = 1$, $F_X(8\pi) = 1$, $F_X(543) = 1$.

Propriétés fondamentales

- F est croissante et continue à droite.
- F est comprise entre 0 et 1.
- $\lim_{x \rightarrow -\infty} F_X(x) = 0$ et $\lim_{x \rightarrow +\infty} F_X(x) = 1$
- Pour $b > a$, $\mathbb{P}(a \leq X < b) = F_X(b) - F_X(a)$.
- $\mathbb{P}(X > a) = 1 - F_X(a)$.

Quantiles et médiane

- Si la fonction de répartition F_X d'une v.a.r. X est strictement croissante d'un intervalle I dans $[0, 1]$, le **quantile d'ordre α** , noté x_α , est défini – pour $\alpha \in]0, 1[$ – par

$$F_X(x_\alpha) = \mathbb{P}(X \leq x_\alpha) = \alpha.$$

- Si F_X n'est pas strictement croissante, on définit le **quantile d'ordre α** par

$$x_\alpha = \inf\{x \in \mathbb{R} \mid F_X(x) \geq \alpha\}.$$

- La **médiane de X** est $x_{1/2}$ et vérifie $\mathbb{P}(X \leq x_{1/2}) = \mathbb{P}(X > x_{1/2}) = 1/2$.

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Variables aléatoires discrètes

- **Variable aléatoire discrète** : variable aléatoire réelle à valeurs dans un ensemble \mathcal{X} fini ou dénombrable (en particulier \mathbb{N}).
- Loi de X : déterminée par la famille des $\mathbb{P}_X(x) = \mathbb{P}(X = x)$ pour tout $x \in \mathcal{X}$.
 \Rightarrow pour toute partie A de \mathcal{X} , $\mathbb{P}_X(A) = \sum_{x \in A} \mathbb{P}(X = x)$.

- **Espérance** (ou moyenne) de X : quantité notée $\mathbb{E}[X]$ et définie par

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \mathbb{P}(X = x).$$

- **Variance** de X : quantité (positive) notée $\text{var}(X)$ (ou σ_X^2) et définie par

$$\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

- **Ecart-type** de X : racine carrée de la variance, $\sigma_X = \sqrt{\text{var}(X)}$.

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Loi de Bernoulli $B(p)$ avec $p \in [0, 1]$

- Schéma de Bernoulli : succès ou échec ; "lancer de pièce biaisée".
- $\Omega = \{0, 1\}$; p : probabilité de succès.
- $\mathbb{P}(X = 1) = p$ et $\mathbb{P}(X = 0) = 1 - p$.
- $\mathbb{E}[X] = p$ et $\text{var}(X) = p(1 - p)$.
- Exemples : jeux de hasard, état de fonctionnement d'un système.

Loi binomiale $B(n, p)$ avec $n \in \mathbb{N}$ et $p \in [0, 1]$

- Nombre de succès lors de n répétitions indépendantes d'un même schéma de Bernoulli $B(p)$.
- $\Omega = \{0, 1, \dots, n\}$.
- Pour tout $k \in \{0, 1, \dots, n\}$, $\mathbb{P}(X = k) = C_n^k p^k (1 - p)^{n-k}$ avec $C_n^k = \frac{n!}{k!(n-k)!}$.
- $\mathbb{E}[X] = np$ et $\text{var}(X) = np(1-p)$.
- Exemples : nombre d'individus porteurs d'un certain gène dans une population, nombre de pièces défectueuses dans un lot de n pièces.

Loi binomiale $B(n, p)$ avec $n \in \mathbb{N}$ et $p \in [0, 1]$

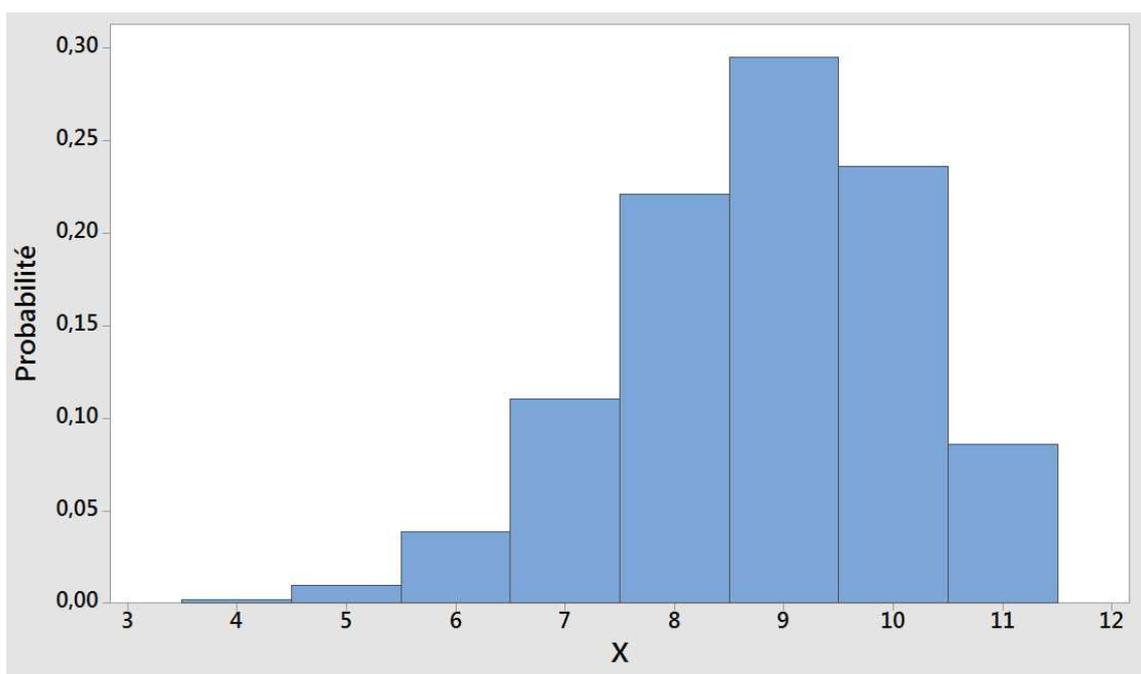


Figure – Loi de probabilité de $B(11; 0,8)$

Loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda > 0$

- "Loi des événements rares" ; "limite de la loi binomiale".
- $\Omega = \mathbb{N}$; λ fréquence moyenne des occurrences.
- Pour tout $k \in \mathbb{N}$, $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$.
- $\mathbb{E}[X] = \lambda$ et $\text{var}(X) = \lambda$.
- Exemples historiques : arrivées de bateaux dans les ports, accidents dus aux coups de sabots de chevaux dans les armées.
- Stabilité par somme : si $X \sim \mathcal{P}(\lambda)$ et $Y \sim \mathcal{P}(\mu)$, avec X et Y indépendantes, alors $X + Y \sim \mathcal{P}(\lambda + \mu)$

Loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda > 0$

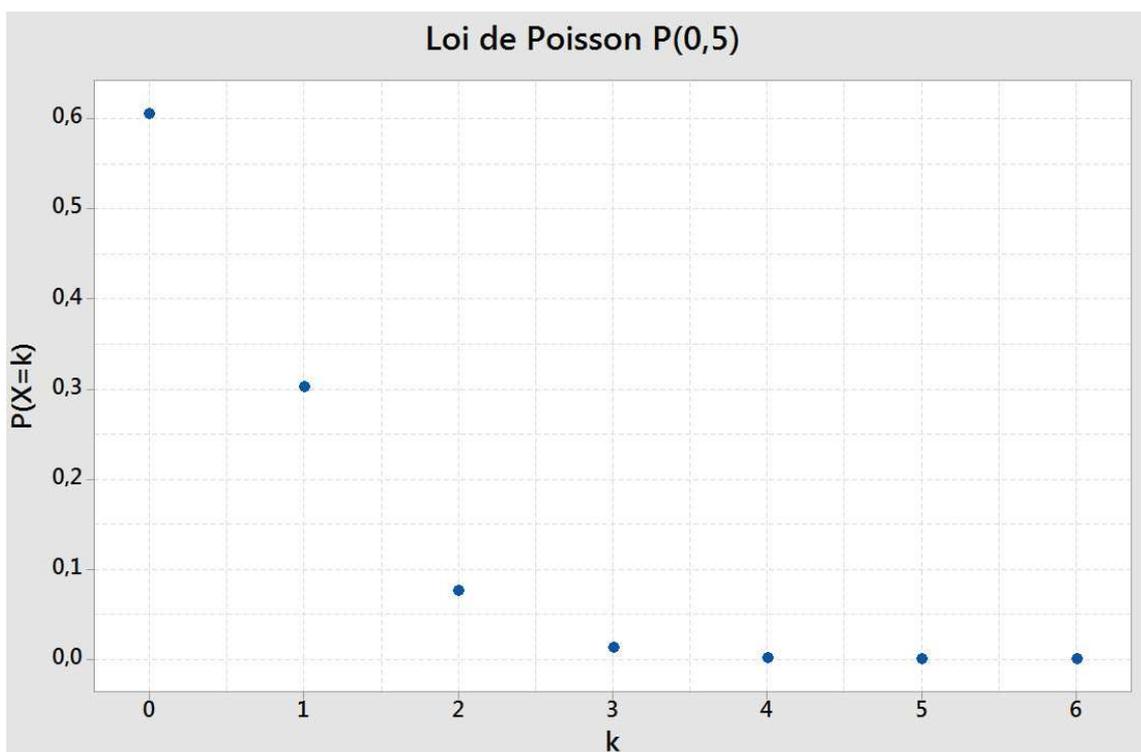


Figure – Loi de probabilité de $\mathcal{P}(0,5)$

Loi de Poisson $\mathcal{P}(\lambda)$ avec $\lambda > 0$

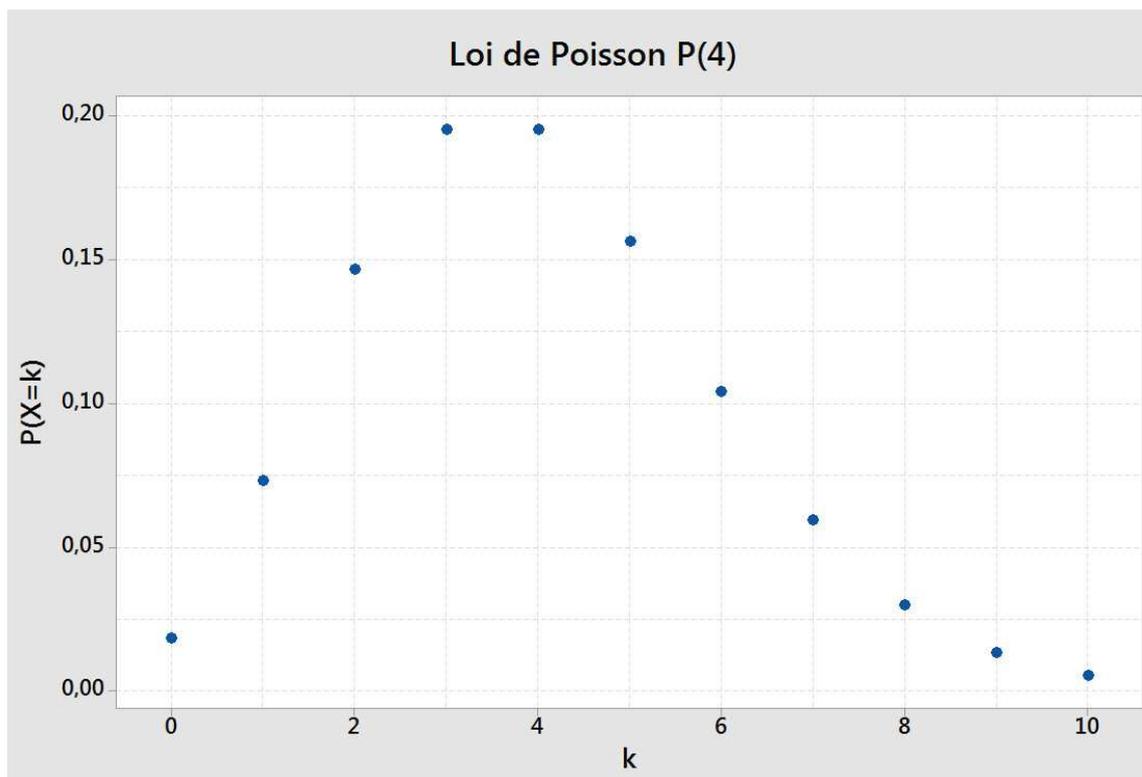


Figure – Loi de probabilité de $\mathcal{P}(4)$

Loi géométrique $\mathcal{G}(p)$ avec $p \in [0, 1]$

- Nombre de réalisations indépendantes d'une loi de Bernoulli $B(p)$ jusqu'à l'obtention du premier succès.
- $\Omega = \mathbb{N}^*$.
- Pour tout $k \in \mathbb{N}^*$, $\mathbb{P}(X = k) = p(1 - p)^{k-1}$.
- $\mathbb{E}[X] = \frac{1}{p}$ et $\text{var}(X) = \frac{1 - p}{p^2}$.
- Exemples : désintégration d'une particule radioactive, collectionneur de vignettes.

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Variables aléatoires continues

- **Variable aléatoire continue** : variable aléatoire réelle à valeurs dans un intervalle de \mathbb{R} .
- Loi de X : déterminée par la famille des $\mathbb{P}(a \leq X \leq b)$ pour tous $a < b$.
- Egalement caractérisée, si elle existe, par sa **densité** f_X , définie par

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(u) du.$$

Propriétés de la densité

- f_X est la dérivée de F_X .
- $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(u) du.$
- f_X est positive et $\int_{-\infty}^{+\infty} f_X(u) du = 1.$

Espérance et variance

- **Espérance** de X : définie dans le cas continue par

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} u f_X(u) du.$$

- **Variance** de X : encore définie par

$$\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

- **Ecart-type** de X : racine carrée de la variance, $\sigma_X = \sqrt{\text{var}(X)}$.

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Loi uniforme $\mathcal{U}([a, b])$ avec $a < b$

- "Tirage au hasard sur l'intervalle $[a, b]$ ".

- $\Omega = [a, b]$.

- Densité : $f_X(x) = \frac{1}{b-a} \mathbf{1}_{[a,b]}(x)$.

- $\mathbb{E}[X] = \frac{a+b}{2}$ et $\text{var}(X) = \frac{(a-b)^2}{12}$.

Loi uniforme $\mathcal{U}([a, b])$ avec $a < b$

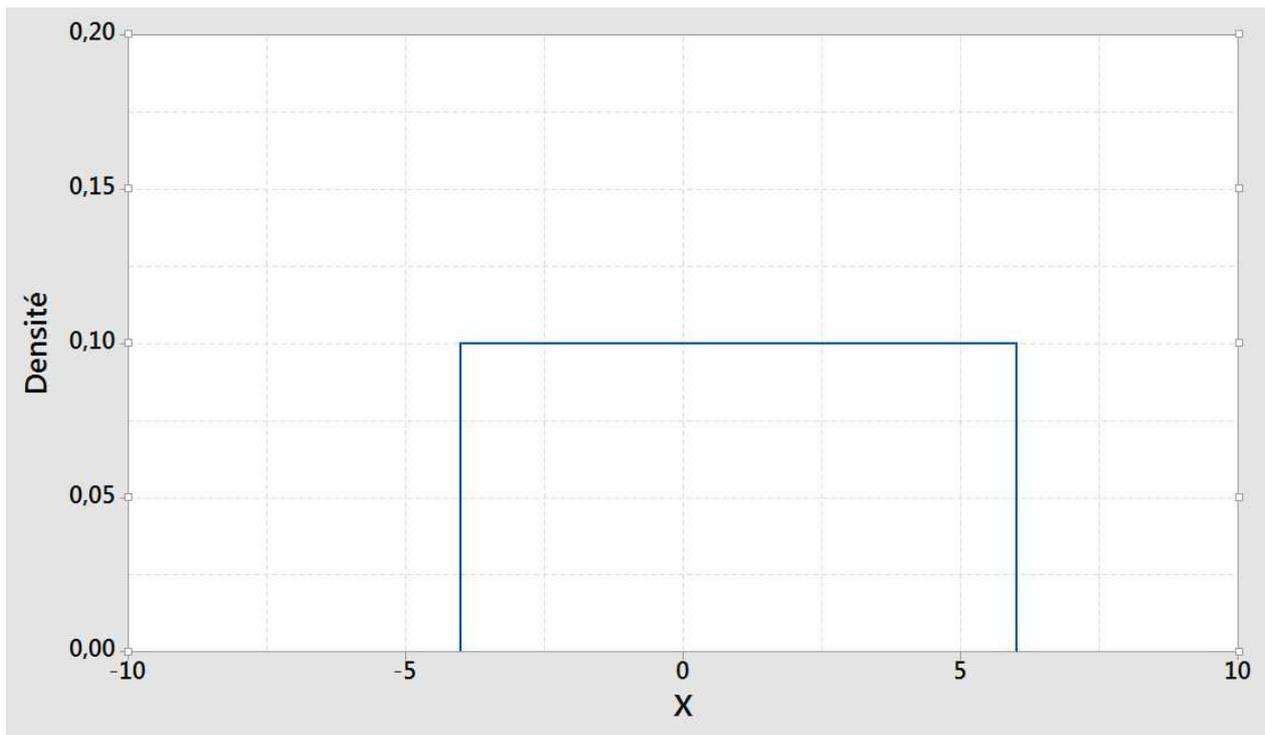


Figure – Densité de $\mathcal{U}([-4, 6])$

Loi exponentielle $\varepsilon(\lambda)$ avec $\lambda > 0$

- $\Omega = \mathbb{R}_+$; λ "taux moyen de défaillance".
- Densité : $f_X(x) = \lambda e^{-\lambda x} \mathbf{1}_{\mathbb{R}_+}(x)$.
- Fonction de répartition : $F_X(x) = (1 - e^{-\lambda x}) \mathbf{1}_{\mathbb{R}_+}(x)$.
- $\mathbb{E}[X] = \frac{1}{\lambda}$ et $\text{var}(X) = \frac{1}{\lambda^2}$.
- Exemples : fiabilité des matériaux électroniques, modélisation de populations sans vieillissement, désintégration radioactive, attente d'un bus, etc.
- "Loi sans mémoire" : $\mathbb{P}(X > t + u \mid X > t) = \mathbb{P}(X > u)$.

Loi exponentielle $\varepsilon(\lambda)$ avec $\lambda > 0$

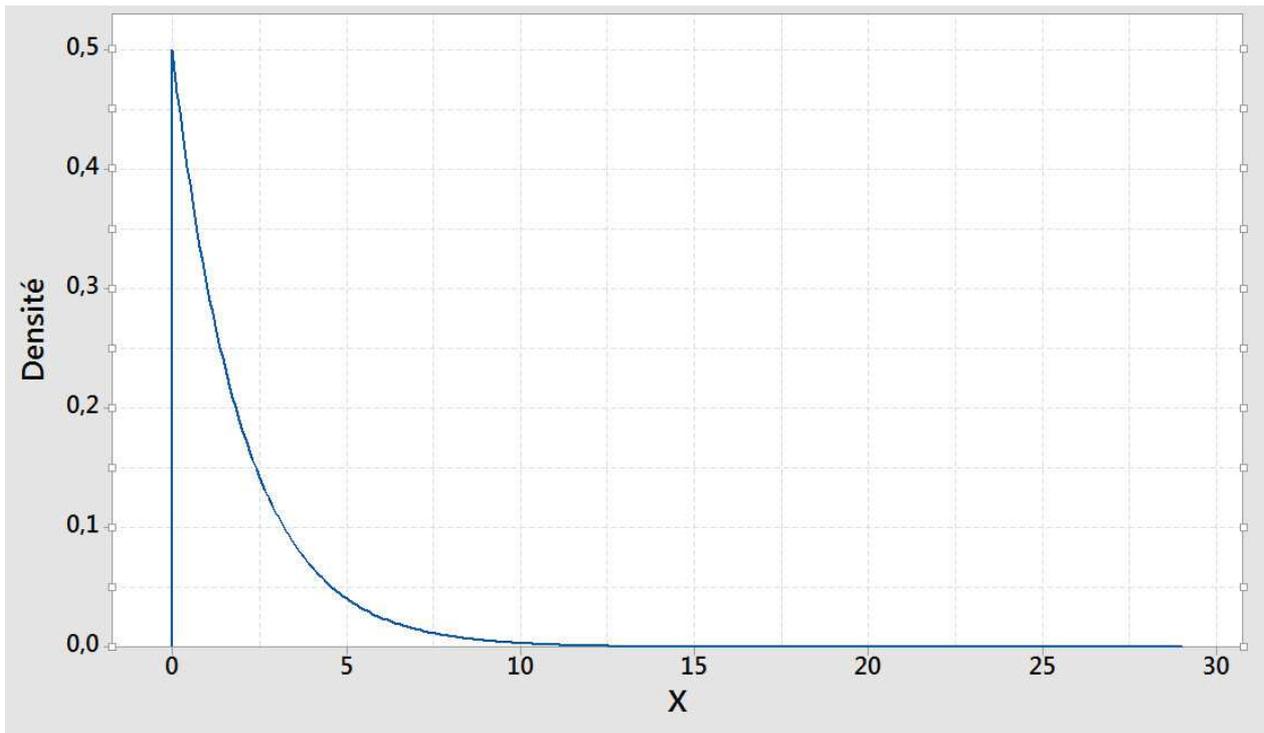


Figure – Densité de $\varepsilon(2)$.

Loi normale $\mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma > 0$

- Aussi appelée loi gaussienne ou de Laplace-Gauss.
- $\Omega = \mathbb{R}$.
- Densité : $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- $\mathbb{E}[X] = \mu$ et $\text{var}(X) = \sigma^2$.
- Exemples : modélisation de très nombreux phénomènes naturels, physiques, économiques, etc.
- **Loi normale centrée réduite** : $\mathcal{N}(0, 1)$

Si $X \sim \mathcal{N}(\mu, \sigma^2)$ alors $\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$.

Loi normale $\mathcal{N}(\mu, \sigma^2)$ avec $\mu \in \mathbb{R}$ et $\sigma > 0$

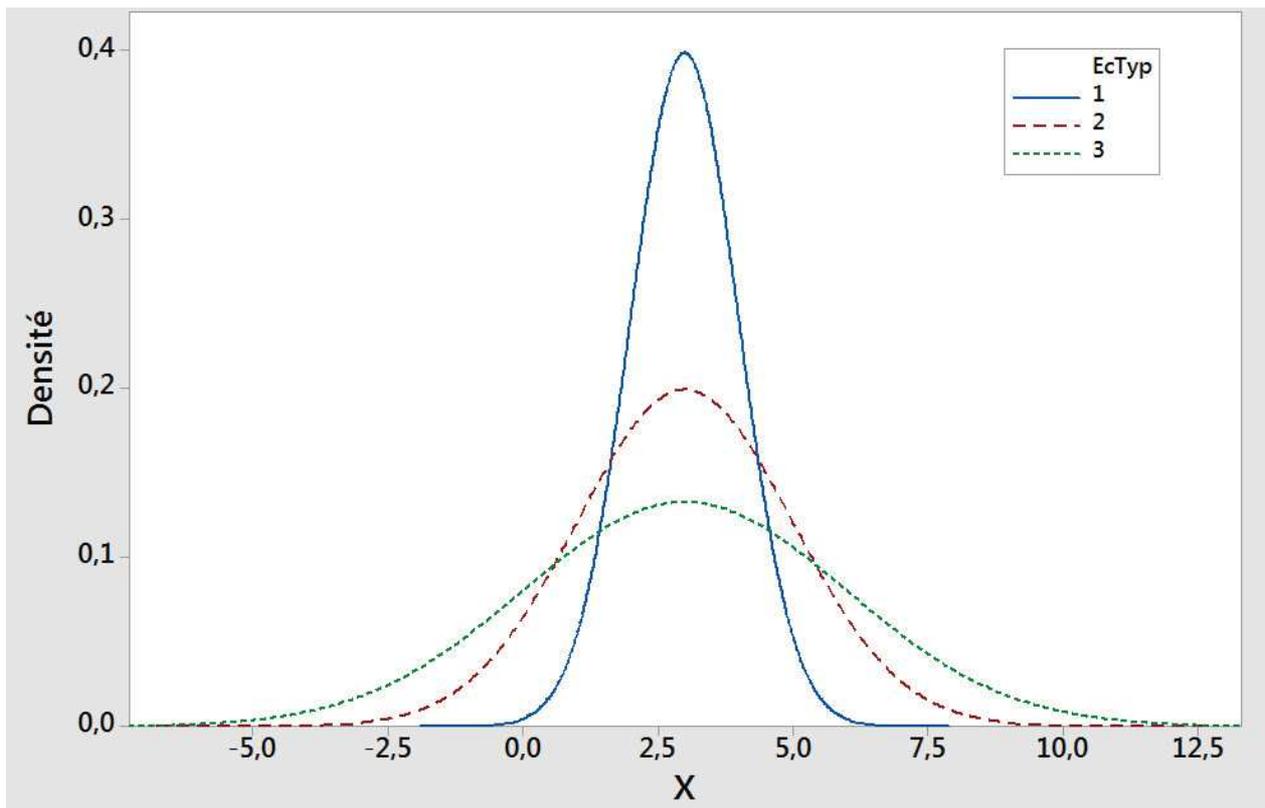


Figure – Densités de $\mathcal{N}(3,1)$, $\mathcal{N}(3,4)$ et $\mathcal{N}(3,9)$.

Loi du Khi-deux $\chi^2(d)$ avec $d \in \mathbb{N}^*$

- $\Omega = \mathbb{R}_+$; d : nombre de "degrés de liberté".
- $Y \sim \chi^2(d) \iff Y = X_1^2 + \dots + X_d^2$ avec $X_1, \dots, X_d \sim \mathcal{N}(0,1)$ indépendantes.
- $\mathbb{E}[Y] = d$ et $\text{var}(Y) = 2d$.

Loi du Khi-deux $\chi^2(d)$ avec $d \in \mathbb{N}^*$

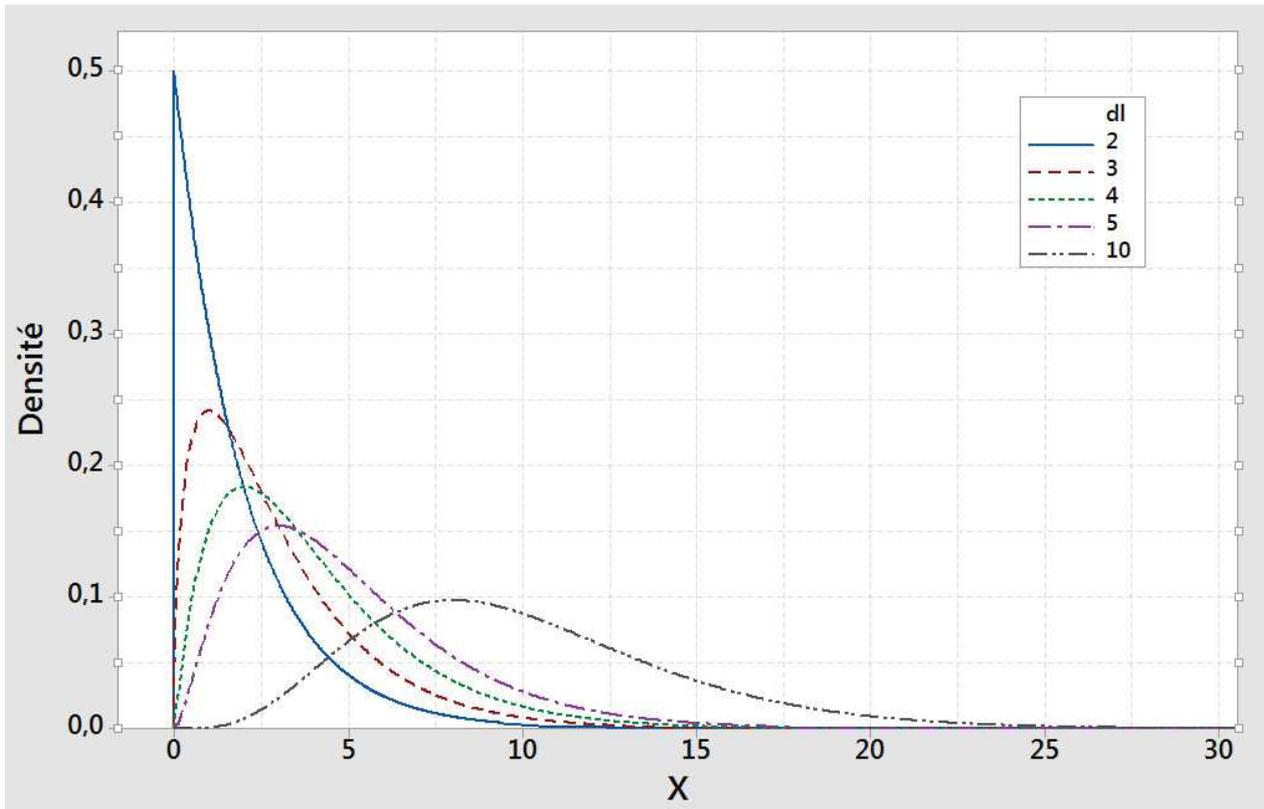


Figure – Densités de $\chi^2(2)$, $\chi^2(3)$, $\chi^2(4)$, $\chi^2(5)$ et $\chi^2(10)$.

Loi de Student $\mathcal{T}(d)$ avec $d \in \mathbb{N}^*$

- $\Omega = \mathbb{R}$; d : nombre de "degrés de liberté".
- $T \sim \mathcal{T}(d) \iff T = \frac{X}{\sqrt{Y/d}}$ avec $X \sim \mathcal{N}(0, 1)$ et $Y \sim \chi^2(d)$ indépendantes.
- $\mathbb{E}[T] = 0$ si $d > 1$ et $\text{var}(T) = \frac{d}{d-2}$ si $d > 2$.

Loi de Student $\mathcal{T}(d)$ avec $d \in \mathbb{N}^*$

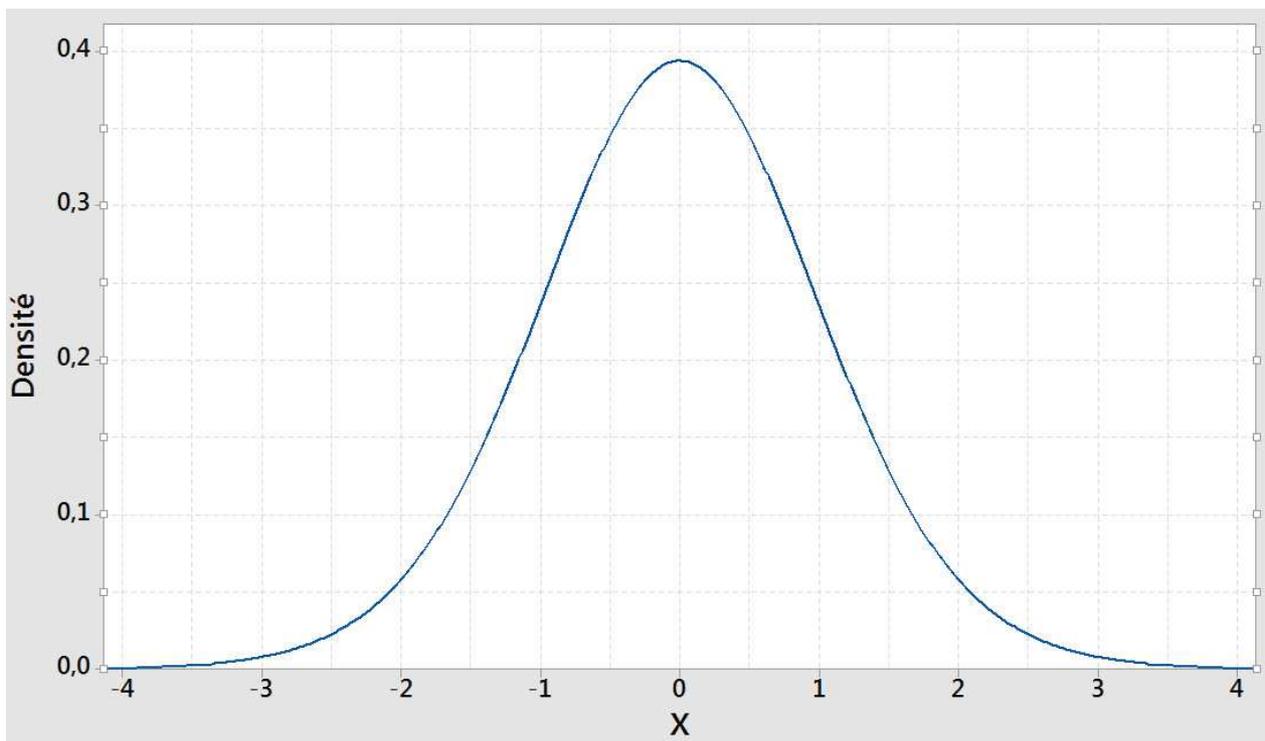


Figure – Densités de $\mathcal{T}(20)$.

Loi de Fisher $\mathcal{F}(d_1, d_2)$ avec $d_1, d_2 \in \mathbb{N}^*$

- $\Omega = \mathbb{R}_+$; d_1, d_2 : nombres de "degrés de liberté".

- $F \sim \mathcal{F}(d_1, d_2) \iff F = \frac{Y_1/d_1}{\sqrt{Y_2/d_2}}$ avec $Y_1 \sim \chi^2(d_1)$ et $Y_2 \sim \chi^2(d_2)$
indépendantes.

Loi de Fisher $\mathcal{F}(d_1, d_2)$ avec $d_1, d_2 \in \mathbb{N}^*$

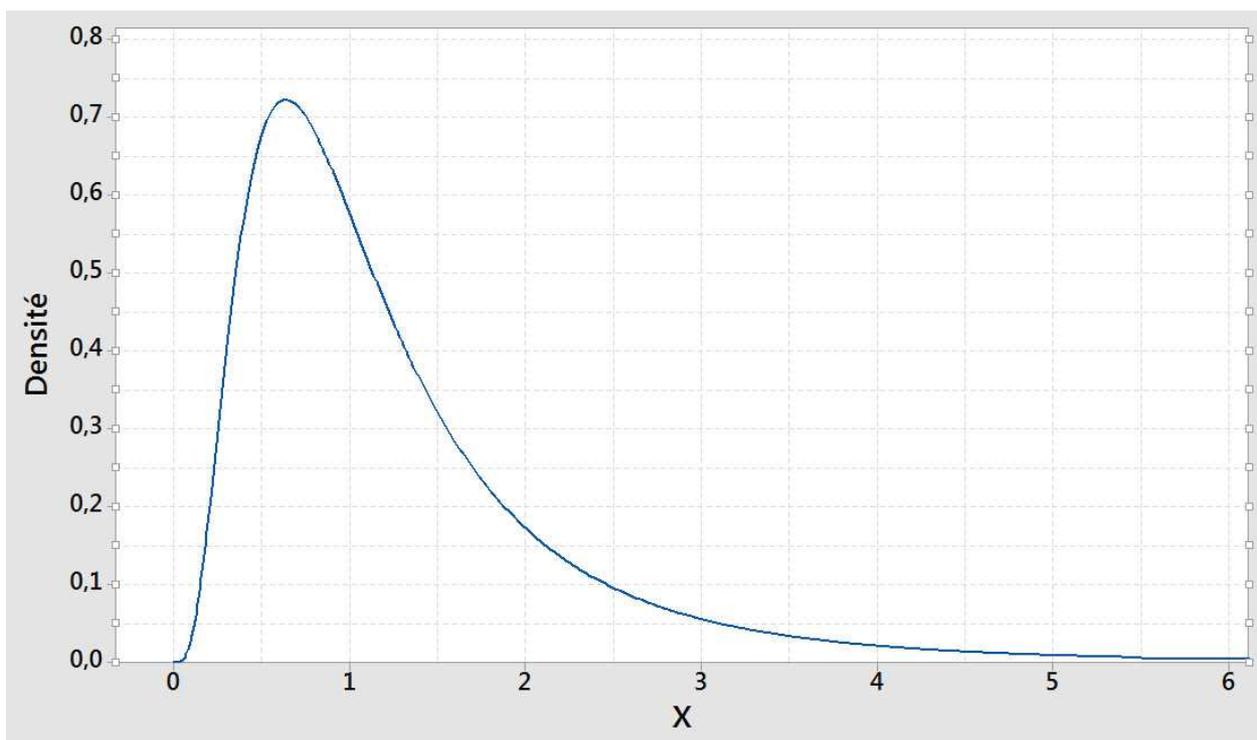


Figure – Densités de $\mathcal{F}(10, 8)$.

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Indépendance de variables aléatoires

Deux variables aléatoires X et Y sont **indépendantes** si, pour toutes parties A et B de \mathbb{R} ,

$$\mathbb{P}(X \in A \text{ et } Y \in B) = \mathbb{P}(X \in A) \mathbb{P}(Y \in B)$$

Conditions équivalentes

- Pour tous réels a et b , $\mathbb{P}(X \leq a \text{ et } Y \leq b) = \mathbb{P}(X \leq a) \mathbb{P}(Y \leq b)$.
- Si X et Y discrètes : pour tous x et y , $\mathbb{P}_{X,Y}(x, y) = \mathbb{P}_X(x) \mathbb{P}_Y(y)$.
- Si X et Y continues : pour tous x et y , $f_{X,Y}(x, y) = f_X(x) f_Y(y)$.

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Propriétés de l'espérance et de la variance

Espérance de l'indicatrice

Pour A partie de \mathbb{R} , $\mathbb{E}[\mathbf{1}_{X \in A}] = \mathbb{P}(X \in A)$.

Positivité

Si $X \leq Y$ alors $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Combinaison linéaire

Pour X et Y deux variables aléatoires, α et β deux réels,

- $\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$.
- Si X et Y indépendantes, $\text{var}(\alpha X + \beta Y) = \alpha^2 \text{var}(X) + \beta^2 \text{var}(Y)$.

Inégalités de Markov et de Chebychev

On considère une variable aléatoire X .

Inégalité de Markov

Si X est positive, pour tout $\epsilon > 0$,

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}.$$

Inégalité de Chebychev

Si X admet une variance, pour tout $\epsilon > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{var}(X)}{\epsilon^2}.$$

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Loi faible des grands nombres

- "Phénomène de stabilisation autour de la valeur moyenne quand la taille de l'échantillon tend vers l'infini" ; "convergence de la moyenne empirique vers l'espérance".
- On considère une famille $(X_n)_{n \in \mathbb{N}}$ de variables aléatoires de même loi, d'espérance μ et admettant une variance σ^2 .

Théorème : loi faible des grands nombres

Pour tout $\epsilon > 0$,

$$\mathbb{P} \left(\left| \frac{X_1 + \dots + X_n}{n} - \mu \right| > \epsilon \right) \xrightarrow{n \rightarrow +\infty} 0$$

- On dit aussi que $\frac{X_1 + \dots + X_n}{n}$ converge vers μ en probabilité.
- Ce résultat est la base de très nombreux résultats que nous verrons dans les chapitres suivants.

On enchaîne avec...

1 Événements et probabilités

- Événements
- Probabilités
- Indépendance

2 Variables aléatoires réelles

- Définitions
- Variables aléatoires discrètes
 - Espérance et variance
 - Variables discrètes usuelles
- Variables aléatoires continues
 - Espérance et variance
 - Variables continues usuelles

3 Indépendance, moments, théorèmes limites

- Indépendance de variables aléatoires
- Propriétés de l'espérance et de la variance
- Théorèmes limites
 - Loi des grands nombres
 - Théorème central limite

Théorème Central Limite

- Idée : "à **quelle vitesse** s'effectue la convergence précédente ?"
- On considère une famille $(X_n)_{n \in \mathbb{N}}$ de variables aléatoires de même loi, d'espérance μ et admettant une variance σ^2 .

Théorème : théorème central limite

$\frac{\sqrt{n}}{\sigma} \left(\frac{X_1 + \dots + X_n}{n} - \mu \right)$ converge "en loi" vers $\mathcal{N}(0, 1)$.

- La convergence en loi est la convergence des fonctions de répartition.
- La convergence de la moyenne empirique vers l'espérance s'effectue donc à une vitesse de l'ordre de $\frac{\sigma}{\sqrt{n}}$.

Approximation d'une loi binomiale par une loi normale

Conséquence du TCL : sous certaines conditions, on peut approximer une loi binomiale $B(n, p)$ par une loi normale $\mathcal{N}(np, np(1 - p))$.

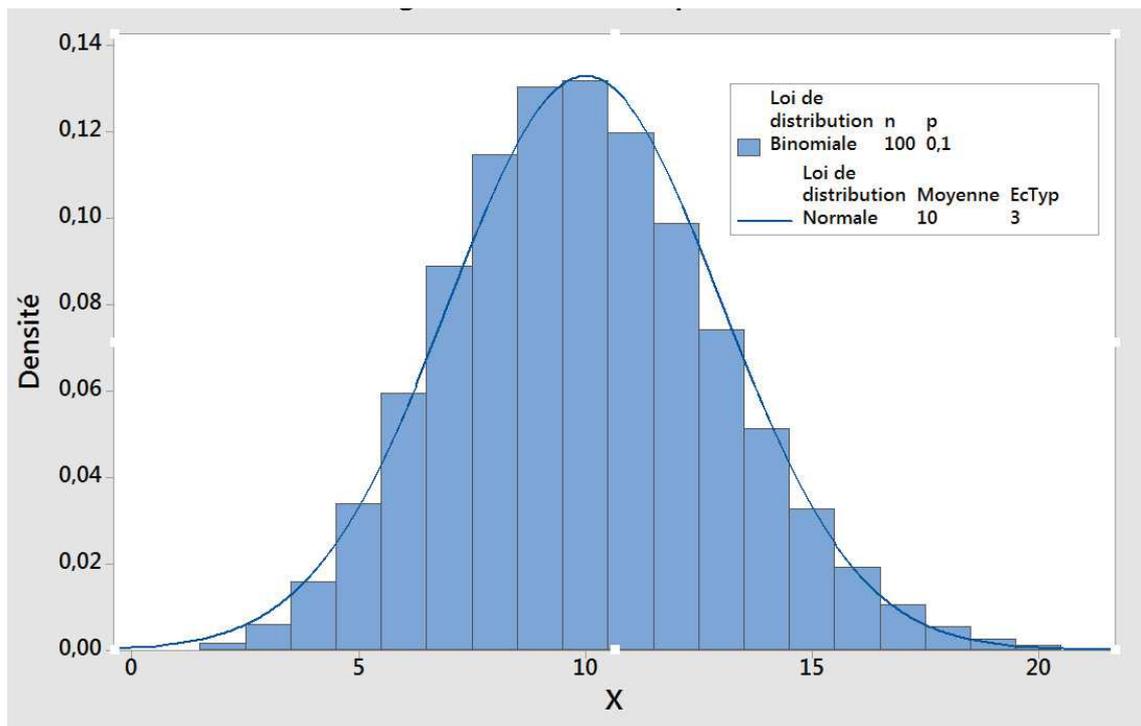


Figure – Approximation d'une loi $B(100; 0, 1)$ par la loi $\mathcal{N}(10, 9)$.

LGN et TCL

A retenir

Sous certaines conditions, si les X_i sont des variables aléatoires de moyenne μ et de variance σ^2 alors la loi de la moyenne empirique $\frac{X_1 + \dots + X_n}{n}$ peut être approximée, pour n grand (≥ 30 par exemple), par une loi normale $\mathcal{N}(\mu, \sigma^2)$.

Chapitre 4 : Estimation statistique

Vers la statistique inférentielle

- **Statistique descriptive** : donner une description aussi fidèle que possible des données.
- **Statistique inférentielle** : à partir d'un échantillon, induire les caractéristiques inconnues d'une population.

↪ Par exemple, les paramètres de cette population.

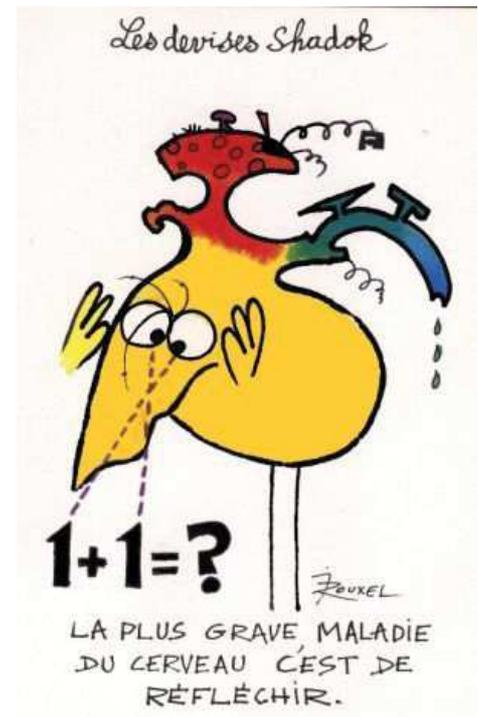
Un exemple avec les mains : temps moyen d'attente au bar.

Objectif : temps d'attente moyen des toulousains dans les bars de la ville.

Contrainte : seuls les 3ICBE sont disponibles pour répondre.

Question : comment se faire, malgré tout, une idée de ce temps d'attente ?

Question subsidiaire : peut-on donner une fourchette dans laquelle sera "très probablement" le temps d'attente moyen ?



La problématique de l'estimation

- Grande **population** ; seulement un **échantillon** de taille réduite.
- **Paramètre** associé à la population, que l'on souhaite déterminer (en pratique : espérance μ , variance σ^2 ou proportion p).
- **Modélisation** à l'aide de variables aléatoires indépendantes et de même loi.
- Obtention d'une **"estimation"** (une valeur approchée) du paramètre pour la population à l'aide des valeurs observables de l'échantillon, via une variable appelée **"estimateur"**.
- Construction d'un **"intervalle de confiance"** : intervalle auquel le paramètre à estimer a une grande probabilité d'appartenir.

Programme des réjouissances

- 1 Estimation ponctuelle et intervalle de confiance
 - Estimateurs : vocabulaire et propriétés
 - Intervalle de confiance

- 2 Estimation d'une moyenne et d'une variance
 - Estimateurs de la moyenne et de la variance
 - Intervalles de confiance pour la moyenne et la variance

- 3 Estimation d'une proportion

On enchaîne avec...

- 1 Estimation ponctuelle et intervalle de confiance
 - Estimateurs : vocabulaire et propriétés
 - Intervalle de confiance

- 2 Estimation d'une moyenne et d'une variance
 - Estimateurs de la moyenne et de la variance
 - Intervalles de confiance pour la moyenne et la variance

- 3 Estimation d'une proportion

Echantillon, estimateur et estimation

On considère une variable aléatoire X et θ un paramètre associé à la loi de X .

- **n -échantillon** issu d'une variable aléatoire X : n variables aléatoires X_1, \dots, X_n indépendantes et de même loi que X .
- **estimateur** de θ : **variable aléatoire**, souvent notée $\hat{\theta}_n$, qui dépend uniquement du n -échantillon X_1, \dots, X_n .
- **estimation** de θ : réalisation de $\hat{\theta}_n$; il s'agit d'un **nombre réel**.

Attente dans les bars

X : temps d'attente d'un toulousain dans les bars de la ville.

θ : espérance de X , temps d'attente moyen ; n : nombre d'étudiants en 3ICBE.

X_1, \dots, X_n : temps d'attente de chacun des 3ICBE.

x_1, \dots, x_n : réalisations des variables aléatoires X_1, \dots, X_n .

Estimateur de θ : $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Estimation de θ : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Estimateur convergent, estimateur sans biais

- **Estimateur sans biais** de θ : estimateur tel que $\mathbb{E}[\hat{\theta}_n] = \theta$.
- **Estimateur consistant** de θ : "proche de θ " au sens suivant : pour tout $\epsilon > 0$:

$$\mathbb{P}(|\hat{\theta}_n - \theta| \geq \epsilon) \xrightarrow{n \rightarrow +\infty} 0$$

Si $\lim_{n \rightarrow +\infty} \mathbb{E}[\hat{\theta}_n] = \theta$ et $\lim_{n \rightarrow +\infty} \text{var}(\hat{\theta}_n) = 0$, alors $\hat{\theta}_n$ est un estimateur consistant.

↔ La moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur consistant et sans biais de l'espérance. ⇒ **exercice**.

Intervalle de confiance

- **Idée** : **contrôler l'erreur** commise ; disposant d'une estimation \bar{x} , trouver un **intervalle** de la forme $[\bar{x} - \epsilon, \bar{x} + \epsilon]$ dans lequel "on est **presque certain**" de **trouver** le paramètre que l'on souhaite estimer.
- On considère X_1, \dots, X_n un échantillon, θ un paramètre à estimer et $\alpha \in]0, 1[$.

Définition formelle

Si θ_{min} et θ_{max} sont deux variables aléatoires, dépendant uniquement de X_1, \dots, X_n , telles que $\mathbb{P}(\theta \in [\theta_{min}, \theta_{max}]) = 1 - \alpha$, alors on dit que $[\theta_{min}, \theta_{max}]$ est un **intervalle de confiance pour θ avec coefficient de sécurité $1 - \alpha$** , que l'on note $IC_{1-\alpha}(X_1, \dots, X_n)$.

- **Exemple** : pour $\alpha = 0,05$, on parle d'**intervalle de confiance à 95%**.

Construction d'un intervalle de confiance $IC_{1-\alpha}(X_1, \dots, X_n)$ pour un paramètre θ inconnu

Mode d'emploi

- Prendre un "bon" estimateur $\hat{\theta}_n$ de θ .
- Déterminer la loi $\hat{\theta}_n$ en fonction de θ .
- Transformer $\hat{\theta}_n$ pour se ramener à une variable aléatoire dont la loi est connue (loi normale, loi de Student).
- Aller chercher (dans les tables) les quantiles ad hoc de cette loi.
- Construire l'intervalle de confiance en manipulant les probabilités.

↪ Peut-on construire un intervalle de confiance à 95 % pour le temps d'attente dans les bars toulousains ?

On enchaîne avec...

- 1 Estimation ponctuelle et intervalle de confiance
 - Estimateurs : vocabulaire et propriétés
 - Intervalle de confiance
- 2 Estimation d'une moyenne et d'une variance
 - Estimateurs de la moyenne et de la variance
 - Intervalles de confiance pour la moyenne et la variance
- 3 Estimation d'une proportion

Estimateur de la moyenne

On considère un échantillon (X_1, \dots, X_n) issu d'une loi de moyenne μ et de variance σ^2 , supposées inconnues.

↪ Exemple : hauteur moyenne des selles des vélos VelôToulouse en fin de journée ; échantillon : vélos de la station Faculté de Pharmacie.

- **Estimateur de la moyenne** : moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Propriétés de l'estimateur

- \bar{X}_n est un estimateur sans biais et convergent de μ .
- Variance de \bar{X}_n : $\text{var}(\bar{X}_n) = \frac{\sigma^2}{n}$.
- Loi de \bar{X}_n :
 - ▶ Si $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ alors $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
 - ▶ Sinon, pour n grand, la loi de \bar{X}_n est proche de $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ (avec le TCL).

Estimateur de la variance

On considère un échantillon (X_1, \dots, X_n) issu d'une loi de moyenne μ et de variance σ^2 , supposées inconnues.

- **Estimateur de la variance** : variance empirique

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n \bar{X}_n^2 \right).$$

Propriétés de l'estimateur

- S_n^2 est un estimateur sans biais et convergent de σ^2 .
- Loi de S_n^2 : si $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ alors $\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$.

Intervalle de confiance pour la moyenne avec une variance connue.

On considère un échantillon X_1, \dots, X_n de variables aléatoires de loi $\mathcal{N}(\mu, \sigma^2)$.
On suppose la variance σ^2 connue.

- De $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, on tire $\sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \sim \mathcal{N}(0, 1)$, puis

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

- On en déduit un intervalle de confiance pour μ :

$$IC_{1-\alpha}(X_1, \dots, X_n) = \left[\bar{x}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

avec $z_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de $\mathcal{N}(0, 1)$ et \bar{x}_n l'estimation ponctuelle de μ .

Intervalle de confiance pour la moyenne avec une variance inconnue.

On considère un échantillon X_1, \dots, X_n de variables aléatoires de loi $\mathcal{N}(\mu, \sigma^2)$. μ et σ^2 sont inconnues.

- On admet le résultat suivant :

$$\sqrt{n} \left(\frac{\bar{X}_n - \mu}{S_n} \right) \sim \mathcal{T}(n-1)$$

- Sur le même principe, on aboutit à un intervalle de confiance pour μ à 95% :

$$IC_{1-\alpha}(X_1, \dots, X_n) = \left[\bar{x}_n - t_{1-\alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{1-\alpha/2} \frac{s_n}{\sqrt{n}} \right]$$

avec $t_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de $\mathcal{T}(n-1)$ et \bar{x}_n et s_n les estimations ponctuelles de μ et S_n .

Intervalle de confiance asymptotique pour la moyenne.

On considère un échantillon X_1, \dots, X_n d'une variable aléatoire X **de loi quelconque** de moyenne μ et de variance σ^2 .

- Variance σ^2 connue :
 - ▶ On ne peut rien dire sur la loi exacte de \bar{X}_n .
 - ▶ D'après le **théorème central limite** (voir **chapitre 3**),

$$\sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

→ Pour n grand, on construit un **intervalle de confiance asymptotique** pour μ de niveau asymptotique $1 - \alpha$, ie $\mathbb{P}(\mu \in IC_{1-\alpha}(X_1, \dots, X_n)) \xrightarrow[n \rightarrow +\infty]{} 1 - \alpha$.

$$IC_{1-\alpha}(X_1, \dots, X_n) = \left[\bar{x}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

avec $z_{1-\alpha/2}$ quantile d'ordre $1 - \alpha/2$ de $\mathcal{N}(0, 1)$ et \bar{x}_n estimation de μ .

- ▶ Même intervalle de confiance que dans le cas gaussien, mais il est asymptotique.

Intervalle de confiance asymptotique pour la moyenne.

On considère un échantillon X_1, \dots, X_n d'une variable aléatoire X **de loi quelconque** de moyenne μ et de variance σ^2 .

- Variance σ^2 inconnue :

- ▶ En remplaçant σ^2 par S_n^2 , la convergence est loi est conservée :

$$\sqrt{n} \left(\frac{\bar{X}_n - \mu}{S_n} \right) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1)$$

- ▶ **Pour n grand**, on construit un intervalle de confiance asymptotique pour μ de niveau asymptotique $1 - \alpha$,

$$IC_{1-\alpha}(X_1, \dots, X_n) = \left[\bar{x}_n - z_{1-\alpha/2} \frac{S_n}{\sqrt{n}}, \bar{x}_n + z_{1-\alpha/2} \frac{S_n}{\sqrt{n}} \right]$$

avec $z_{1-\alpha/2}$ quantile d'ordre $1 - \alpha/2$ de $\mathcal{N}(0, 1)$ et \bar{x}_n estimation de μ .

Intervalle de confiance pour la variance

- Le résultat crucial ici est

$$\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$$

- L'intervalle de confiance obtenu pour σ^2 est alors

$$IC_{1-\alpha}(X_1, \dots, X_n) = \left[\frac{(n-1) s_n^2}{\nu_{1-\alpha/2}}, \frac{(n-1) s_n^2}{\nu_{\alpha/2}} \right]$$

avec $\nu_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de $\chi^2(n-1)$ et s_n l'estimation ponctuelle de S_n .

On enchaîne avec...

- 1 Estimation ponctuelle et intervalle de confiance
 - Estimateurs : vocabulaire et propriétés
 - Intervalle de confiance
- 2 Estimation d'une moyenne et d'une variance
 - Estimateurs de la moyenne et de la variance
 - Intervalles de confiance pour la moyenne et la variance
- 3 Estimation d'une proportion

Un exemple : la résistance des *smartphones*

Contexte

50 millions de smartphones d'un même modèle sont produits chaque année. On souhaite connaître le taux de résistance à la chute à un mètre sur de l'asphalte.

Cadre expérimental

On dispose d'un échantillon de $n = 100$ smartphones. On veut estimer la proportion p de résistance à la chute.

Modélisation

Au i ème smartphone, on associe X_i une v.a.r. de Bernoulli $B(p)$, avec p inconnu :

$$X_i = \begin{cases} 1 & \text{si résistant,} \\ 0 & \text{sinon .} \end{cases}$$

et on suppose les $(X_i)_{i \in \{1, \dots, n\}}$ indépendantes. On obtient n réalisations $(x_i)_{i \in \{1, \dots, n\}}$.

Un exemple : la résistance des *smartphones*

Idée naturelle

Estimer p par $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Justification mathématique : loi des grands nombres.

Application numérique

Sur l'échantillon, 68 smartphones sont résistants. $\Rightarrow \bar{x} = 0,68$.

Estimation ponctuelle de la proportion p .

Intervalle de confiance

Intervalle de confiance à 95% : $[0,59;0,77]$ \leftrightarrow voir les slides suivantes !

Estimateur d'une proportion

On considère un échantillon X_1, \dots, X_n issu d'une loi $B(p)$, avec p inconnue.

\Rightarrow Même principe que pour la moyenne, sauf que $\mu = p$ et $\sigma^2 = p(1-p)$.

- **Estimateur de la proportion** : $P_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Propriétés de l'estimateur

- P_n est un estimateur sans biais et convergent de p .
- Variance de P_n : $\text{var}(P_n) = \frac{1}{n} p(1-p)$.
- Loi de P_n : pour n grand, la loi de P_n est proche de $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

Intervalle de confiance pour une proportion

Comme les X_i suivent des $B(p)$, ils ne sont pas gaussiens : on utilise une approximation.

Pour n suffisamment grand, p_n suit approximativement une loi $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$.

- L'intervalle de confiance correspondant est

$$IC_{1-\alpha}(X_1, \dots, X_n) = \left[p_n - z_{1-\alpha/2} \sqrt{\frac{p_n(1-p_n)}{n}}, p_n + z_{1-\alpha/2} \sqrt{\frac{p_n(1-p_n)}{n}} \right]$$

avec $z_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de $\mathcal{N}(0, 1)$ et p_n l'estimation ponctuelle de P .

Intervalle de confiance pour la résistance moyenne des smartphones

Intervalle de confiance à 95% : $IC_{0,95}(x_1, \dots, x_n) = [0,59; 0,77]$.

Chapitre 5 : Tests statistiques

A propos des vertus insoupçonnées du jus de betterave...

A l'automne 2019, une rumeur parcourt avec insistance le campus de l'INSA Toulouse : le jus de betterave serait bénéfique à l'assimilation des connaissances scientifiques. Décidant de mettre toutes les chances de leur côté, un groupe d'étudiants en 3ICBE en boivent chacun un verre la veille d'un examen de mécanique des fluides...

Les 30 étudiants ayant bu du jus de betterave obtiennent une note moyenne de $m_1 = 11,8$ avec un écart type de $s_1 = 1,8$, alors que les 50 étudiants n'ayant pas testé ce breuvage ont en moyenne une note de $m_2 = 11,1$ pour un écart-type $s_2 = 1,2$.

Problématique : l'écart de notes est-il significatif, ou dû au hasard ?
Le jus de betterave a-t-il un réel impact sur la note obtenue par les étudiants ?

Le principe des tests statistiques

Principe d'un test : une **hypothèse** que l'on **accepte ou rejette**, avec une certaine **marge d'erreur**, selon le résultat d'une expérience.

- Question avec réponse OUI/NON.
↪ Le jus de betterave a-t-il un impact sur la note des étudiants ?
- Données relatives à cette question à disposition suite à une expérimentation.
↪ Les résultats au contrôle de mécanique des fluides des 3ICBE.
- Modélisation statistique de l'expérience : les données sont vues comme la réalisation de variables aléatoires.
↪ Deux populations. Population 1 : les buveurs de jus de betterave ; la note de chaque étudiant est modélisée par une variable aléatoire de moyenne μ_1 et d'écart type σ_1 . Population 2 : les non-buveurs de jus de betterave ; idem avec une moyenne μ_2 et un écart type σ_2 .
- Réponse à la question = acceptation ou rejet d'une hypothèse – appelée **hypothèse nulle** et notée H_0 – caractéristique du modèle.
↪ Hypothèse nulle H_0 : " $\mu_1 = \mu_2$ ".

Le principe des tests statistiques

Principe d'un test : une **hypothèse** que l'on **accepte ou rejette**, avec une certaine **marge d'erreur**, selon le résultat d'une expérience.

- Question avec réponse **OUI/NON**.
↪ Le jus de betterave a-t-il un impact sur la note des étudiants ?
- Réponse à la question = acceptation ou rejet d'une hypothèse – appelée **hypothèse nulle** et notée H_0 – caractéristique du modèle.
↪ Hypothèse nulle H_0 : " $\mu_1 = \mu_2$ ".
 - ▶ Si on **accepte l'hypothèse** H_0 , on répond **NON** à la question.
↪ "*Les différences observées ne sont dues qu'au hasard*".
↪ Les deux moyennes sont égales (**acceptation**), le jus de betterave n'a donc pas d'impact (**NON**).
 - ▶ Si on **rejette l'hypothèse** H_0 , on répond **OUI** à la question.
↪ "*Les différences observées sont trop improbables pour n'être l'oeuvre que du seul hasard, et sont donc significatives*".
↪ Les deux moyennes sont différentes (**rejet**), le jus de betterave a donc un impact significatif (**OUI**).

La problématique des tests

Deux problèmes à gérer : la **prise de décision** et le **contrôle des erreurs**.

- Comment sait-on quelle décision prendre, i.e. accepter ou rejeter l'hypothèse ?
↪ **Rejet de l'hypothèse** \Leftrightarrow "Statistique de test" dans la "**zone de rejet**".
↪ Rejeter H_0 : " $\mu_1 = \mu_2$ " pour un écart $m_1 - m_2$ suffisamment grand.
 - Deux façons de se tromper :
 - ▶ "**Erreur de première espèce**" (faux-positif) : rejeter H_0 alors qu'elle est vraie.
↪ Conclure que le jus de betterave a un impact alors que ce n'est pas le cas.
 - ▶ "**Erreur de seconde espèce**" (faux-négatif) : accepter H_0 alors qu'elle est fautive. ↪ Conclure que le jus de betterave n'a pas d'impact sur la note, alors qu'il en a un.
- ↪ Compromis à trouver entre ces deux erreurs.

Une introduction à la théorie des tests statistiques

- 1 Généralités sur les tests
- 2 Un premier exemple : un test de conformité bilatéral sur la moyenne
- 3 Un deuxième exemple : un test d'homogénéité bilatéral sur la moyenne
- 4 Un panorama non-exhaustif des tests statistiques classiques
 - Tests paramétriques de conformité
 - Tests paramétriques d'homogénéité
 - Comparaison de plusieurs échantillons : l'ANOVA
 - Tests d'ajustement
 - Tests non paramétriques : comparaison de médianes.

On enchaîne avec...

- 1 Généralités sur les tests
- 2 Un premier exemple : un test de conformité bilatéral sur la moyenne
- 3 Un deuxième exemple : un test d'homogénéité bilatéral sur la moyenne
- 4 Un panorama non-exhaustif des tests statistiques classiques
 - Tests paramétriques de conformité
 - Tests paramétriques d'homogénéité
 - Comparaison de plusieurs échantillons : l'ANOVA
 - Tests d'ajustement
 - Tests non paramétriques : comparaison de médianes.

Définition théorique d'un test

On considère un n -échantillon X_1, \dots, X_n et $\alpha \in]0; 1[$.

Test de niveau α

Tester l'**hypothèse nulle** H_0 contre l'**hypothèse alternative** H_1 au niveau α c'est se donner une **zone de rejet** \mathcal{R}_α et une variable aléatoire T fonction de X_1, \dots, X_n , appelée **statistique de test**, telles que

$$\mathbb{P}_{H_0}(T \in \mathcal{R}_\alpha) \leq \alpha$$

On rejette alors l'hypothèse H_0 ssi $T \in \mathcal{R}_\alpha$.

Remarque 1 : α correspond à l'erreur de première espèce.

Remarque 2 : Il n'y a pas unicité du test.

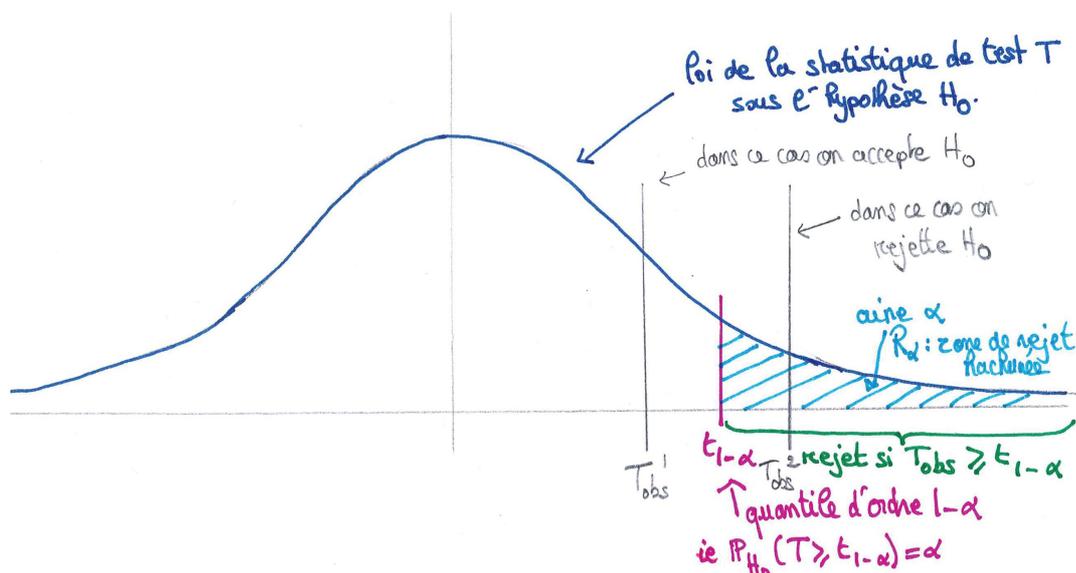
Heuristique d'un test

- **Test = règle de décision.**

- Illustration de la prise de décision sur un exemple (test unilatéral à droite) :

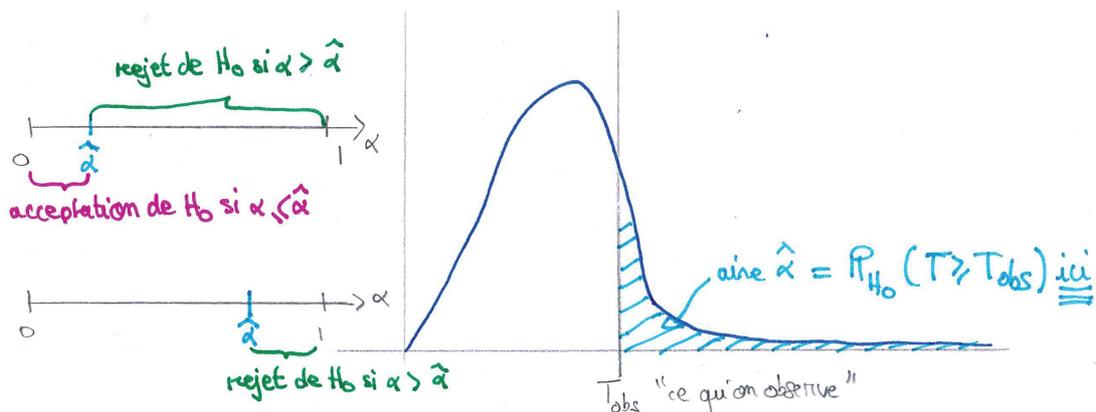
$$\text{rejet de } H_0 \Leftrightarrow T_{\text{obs}} > t_{1-\alpha},$$

avec $t_{1-\alpha}$ quantile d'ordre $1 - \alpha$ de T .



Notion de p-valeur

- **Probabilité critique / p-valeur / p-value** = plus grand niveau α autorisant l'acceptation de H_0 au vu des données ; ou encore la "probabilité sous l'hypothèse H_0 d'être plus extrême que ce qu'on observe", notée $\hat{\alpha}$ ou \hat{p} .
- p-valeur et prise de décision : pour un test de niveau α ,
rejet de $H_0 \Leftrightarrow$ p-valeur $< \alpha$
- Illustration sur un **cas particulier** (test unilatéral à droite) :



Erreurs, risques et puissance

● Différents types d'erreurs

Pour un test de niveau α , on a les probabilités suivantes :

	H_0 vraie	H_1 vraie
H_0 acceptée	$1 - \alpha =$ confiance du test	$\beta =$ risque de 2nde espèce
H_1 acceptée	$\alpha =$ risque de 1ère espèce	$1 - \beta =$ puissance du test

- Idéalement, α et β petits mais si α diminue, β augmente, et vice versa.
 \hookrightarrow Trouver un consensus.

● Toutes les erreurs n'ont pas les mêmes conséquences !

\hookrightarrow Exemple de l'eau radioactive.

μ niveau moyen de radioactivité ; μ_0 valeur critique.

On teste $H_0 : \mu \geq 5$ contre $H_1 : \mu < 5$.

Erreur de première espèce : **boire de l'eau toxique**.

Erreur de deuxième espèce : jeter de l'eau potable.

L'élaboration d'un test : un mode d'emploi

1. Modélisation statistique de l'expérience.
2. Choix de H_0 (**hypothèse nulle**) et H_1 (**hypothèse alternative**); choix du **niveau α du test** (souvent $\alpha = 0,05$).
3. Détermination de la **statistique de test T** (fonction de l'échantillon, loi standard connue).
4. **Rejet ou acceptation** de H_0 au risque α ; deux méthodes :
 - ▶ Calcul de la **région de rejet R_α** (quantile à lire dans les tables) et de la **valeur observée T_{obs}** de la statistique de test, **en supposant H_0 vérifiée**.
 - ▶ Calcul de la **p-valeur $\hat{\alpha}$** (avec les tables ou R Studio [voir TP]).

On enchaîne avec...

- 1 Généralités sur les tests
- 2 Un premier exemple : un test de conformité bilatéral sur la moyenne
- 3 Un deuxième exemple : un test d'homogénéité bilatéral sur la moyenne
- 4 Un panorama non-exhaustif des tests statistiques classiques
 - Tests paramétriques de conformité
 - Tests paramétriques d'homogénéité
 - Comparaison de plusieurs échantillons : l'ANOVA
 - Tests d'ajustement
 - Tests non paramétriques : comparaison de médianes.

Une histoire de panneaux solaires ariégeois



Figure – Eglise de Crampagna, dans le canton du Val d'Ariège. Photo : ladepeche.fr.

Sur la Région Occitanie, la production photovoltaïque de panneaux solaires est en moyenne $\mu_0 = 1200$ kWh/kWc. Dans le canton du Val d'Ariège, où l'ensoleillement est plus faible, on a relevé une production photovoltaïque moyenne relevée est de $m = 1050$ kWh/kWc et un écart-type empirique $s = 200$ pour les $n = 100$ panneaux du canton (données fictives).

Les conditions météorologiques ont-elles un impact significatif sur la production photovoltaïque ?

Construction d'un test de conformité bilatéral

● Etape n°1 : modélisation

- ▶ On note X la variable aléatoire qui représente la production photovoltaïque moyenne d'un panneau solaire du Val d'Ariège, de moyenne μ et de variance σ^2 (inconnues).
- ▶ Soient \bar{X}_n et S_n^2 les estimateurs de μ et σ^2 pour un n -échantillon (ici $n = 100$).

→ ici **test de conformité** : comparaison entre un échantillon et une population de référence.

● Etape n°2 : choix des hypothèses

- ▶ **Hypothèse nulle** H_0 : " $\mu = \mu_0$ "; autrement dit, pas de différence sur la production moyenne des panneaux solaires en Val d'Ariège par rapport à l'Occitanie.
- ▶ **Hypothèse alternative** H_1 : " $\mu \neq \mu_0$ "; ce sera la conclusion du test si on rejette H_0 .
- ▶ On prend comme **niveau du test** $\alpha = 0,05$.

→ pour des hypothèses de cette forme, on parle de **test bilatéral**.

Construction d'un test de conformité bilatéral

• Etape n°3 : construction de la statistique de test

- ▶ On définit la variable aléatoire T :

$$T = \sqrt{n} \left(\frac{\bar{X}_n - \mu}{S_n} \right)$$

- ▶ Pour n grand ($n = 100$ convient), T suit approximativement une loi normale centrée réduite $\mathcal{N}(0, 1)$ [voir le **chapitre 4** et la construction des intervalles de confiance].
⇒ On prend T comme **statistique de test**.

Construction d'un test de conformité bilatéral

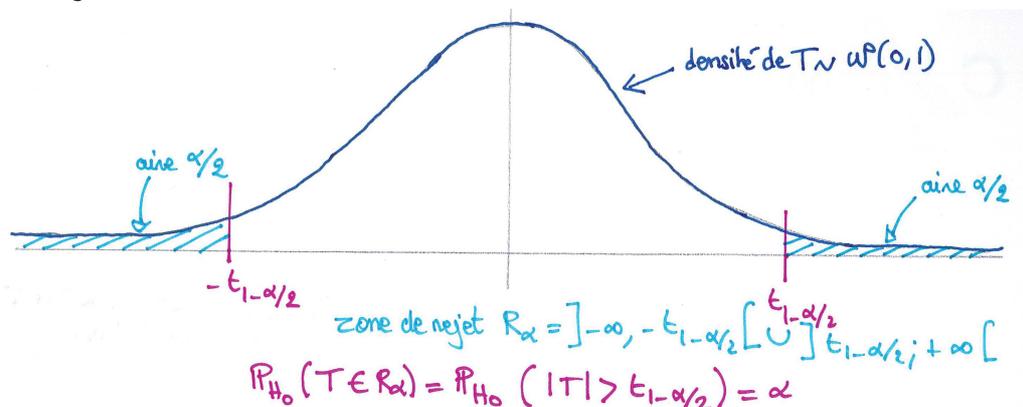
• Etape n°4 : acceptation ou rejet de l'hypothèse :

Première méthode, avec la zone de rejet et le quantile

- ▶ Pour $t_{1-\alpha/2}$ quantile d'ordre $1 - \alpha/2$ de T , $\mathbb{P}(|T| > t_{1-\alpha/2}) = \alpha$.
Pour $\alpha = 0,05$,

$$\mathbb{P}(|T| > 1,96) = 0,05.$$

Zone de rejet : $R_\alpha =]-\infty; -1,96[\cup]1,96; +\infty[$.



- ▶ En supposant H_0 , et donc $\mu = \mu_0 = 1200$, on calcule la **valeur observée de la statistique de test** : $T_{obs} = -7,5$.
- ▶ $T_{obs} \in R_\alpha$ donc **rejet de H_0** au niveau $\alpha = 0,05$..
L'écart de production photovoltaïque observé est donc significatif, et pas le fait du seul hasard.

Construction d'un test de conformité bilatéral

- **Etape n°4 : acceptation ou rejet de l'hypothèse :**
Deuxième méthode, avec la p-valeur.

- ▶ Dans le cadre bilatéral, la p-valeur $\hat{\alpha}$ est définie par

$$\hat{\alpha} = \mathbb{P}_{H_0}(|T| \geq |T_{obs}|).$$

- ▶ Dans la table de $\mathcal{N}(0,1)$, on se rend compte que $\hat{\alpha} = \mathbb{P}_{H_0}(|T| \geq 7,5) < 0,001$ donc

$$\hat{\alpha} < \alpha.$$

- ▶ **Rejet de H_0** au niveau $\alpha = 0,05$.

L'écart de production photovoltaïque observé est donc significatif, et pas le fait du seul hasard.

On enchaîne avec...

1 Généralités sur les tests

2 Un premier exemple : un test de conformité bilatéral sur la moyenne

3 Un deuxième exemple : un test d'homogénéité bilatéral sur la moyenne

4 Un panorama non-exhaustif des tests statistiques classiques

- Tests paramétriques de conformité
- Tests paramétriques d'homogénéité
- Comparaison de plusieurs échantillons : l'ANOVA
- Tests d'ajustement
- Tests non paramétriques : comparaison de médianes.

Le retour des betteraves



A l'automne 2019, une rumeur parcourt avec insistance le campus de l'INSA Toulouse : le jus de betterave serait bénéfique à l'assimilation des connaissances scientifiques. Décidant de mettre toutes les chances de leur côté, un groupe d'étudiants en 3ICBE en boivent chacun un verre la veille d'un examen de mécanique des fluides...

Les 30 étudiants ayant bu du jus de betterave obtiennent une note moyenne de $m_1 = 11,8$ avec un écart type de $s_1 = 1,8$, alors que les 50 étudiants n'ayant pas testé ce breuvage ont en moyenne une note de $m_2 = 11,1$ pour un écart-type $s_2 = 1,2$.

Problématique : l'écart de notes est-il significatif, ou dû au hasard ?

Le jus de betterave a-t-il un réel impact sur la note obtenue par les étudiants ?

Construction d'un test d'homogénéité bilatéral

● Etape n°1 : modélisation

- ▶ **Population 1** : les buveurs de jus de betterave ; la note de chaque étudiant est modélisée par une variable aléatoire X_1 de moyenne μ_1 et d'écart type σ_1 . Soient \bar{X}_{1,n_1} et S_{1,n_1}^2 les estimateurs de μ_1 et σ_1^2 pour un n_1 -échantillon (ici $n_1 = 30$).
- ▶ **Population 2** : les non-buveurs de jus de betterave : on définit de manière analogue X_2 , μ_2 et σ_2 ainsi que les estimateurs \bar{X}_{2,n_2} et S_{2,n_2}^2 avec $n_2 = 50$.

↔ il s'agit d'un **test d'homogénéité** : comparaison de deux échantillons qui représentent deux populations.

● Etape n°2 : choix des hypothèses

- ▶ **Hypothèse nulle** H_0 : " $\mu_1 = \mu_2$ "; le jus de betteraves n'a pas d'impact significatif sur la note en mécanique des fluides.
- ▶ **Hypothèse alternative** H_1 : " $\mu_1 \neq \mu_2$ ".
- ▶ On prend comme **niveau du test** $\alpha = 0,05$.

Construction d'un test d'homogénéité bilatéral

● Etape n°3 : construction de la statistique de test

- ▶ On définit la variable aléatoire T :

$$T = \frac{(\bar{X}_{1,n_1} - \bar{X}_{2,n_2}) - (\mu_1 - \mu_2)}{\sqrt{S_{1,n_1}^2/n_1 + S_{2,n_2}^2/n_2}}$$

- ▶ Résultat admis : pour n_1 et n_2 grands (en pratique, ≥ 30), T suit approximativement une loi normale centrée réduite $\mathcal{N}(0,1)$.
⇒ On prend T comme **statistique de test**.

● Etape n°4 : acceptation ou rejet de l'hypothèse :

Première méthode, avec la zone de rejet et le quantile

- ▶ Comme dans l'exemple précédent, pour $\alpha = 0,05$,

$$\mathbb{P}(|T| > 1,96) = 0,05.$$

Zone de rejet : $R_\alpha =]-\infty; -1,96[\cup]1,96; +\infty[$.

- ▶ En supposant H_0 : " $\mu_1 = \mu_2$ " vérifiée, on calcule $T_{obs} = \frac{m_1 - m_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = 1,89$.
- ▶ $T_{obs} \notin R_\alpha$ donc on ne peut pas rejeter H_0 au niveau $\alpha = 0,05$... donc **acceptation de H_0** .

Construction d'un test d'homogénéité bilatéral

● Etape n°4 : acceptation ou rejet de l'hypothèse :

Deuxième méthode, avec la p-valeur.

- ▶ Dans le cadre bilatéral, la p-valeur $\hat{\alpha}$ est définie par

$$\hat{\alpha} = \mathbb{P}_{H_0}(|T| \geq |T_{obs}|).$$

- ▶ Dans la table de $\mathcal{N}(0,1)$, on se rend compte que

$$\hat{\alpha} = \mathbb{P}_{H_0}(|T| \geq 1,89) = 2 \mathbb{P}_{H_0}(T \geq 1,89) = 0,0588$$

donc $\hat{\alpha} > \alpha$.

- ▶ **Acceptation de H_0** au niveau $\alpha = 0,05$.

On ne peut donc pas conclure à un impact significatif du jus de betterave...

On enchaîne avec...

1 Généralités sur les tests

2 Un premier exemple : un test de conformité bilatéral sur la moyenne

3 Un deuxième exemple : un test d'homogénéité bilatéral sur la moyenne

4 Un panorama non-exhaustif des tests statistiques classiques

- Tests paramétriques de conformité
- Tests paramétriques d'homogénéité
- Comparaison de plusieurs échantillons : l'ANOVA
- Tests d'ajustement
- Tests non paramétriques : comparaison de médianes.

Les différents types de tests

- **Test de paramètres / test d'ajustement :**
 - ▶ **Test de paramètres :** l'hypothèse sur la valeur d'un **paramètre réel** (moyenne, médiane, variance) associé à l'échantillon.
 - ▶ **Test d'ajustement :** l'hypothèse porte sur la **loi de l'échantillon** (souvent via sa fonction de répartition).
- **Test unilatéral / bilatéral :** pour un test d'un paramètre réel θ :
 - ▶ **Test bilatéral :** hypothèses de la forme $H_0 : \theta = \theta_0$ et $H_1 : \theta \neq \theta_0$.
 - ▶ **Test unilatéral à droite :** hypothèses de la forme $H_0 : \theta \leq \theta_0$ et $H_1 : \theta > \theta_0$.
 - ▶ **Test unilatéral à gauche :** hypothèses de la forme $H_0 : \theta \geq \theta_0$, $H_1 : \theta < \theta_0$.
- **Test de conformité / test d'homogénéité :**
 - ▶ **Test de conformité :** un **échantillon** que l'on souhaite comparer à une population connue.
 - ▶ **Test d'homogénéité :** **deux échantillons** issus de deux populations, que l'on souhaite comparer.
- **Test paramétrique / non-paramétrique :** pour un n -échantillon X_1, \dots, X_n ,
 - ▶ **Tests paramétriques :** valides lorsque la loi commune des X_i est une loi connue (**loi gaussienne**, ou, plus rarement, binomiale) OU si **n est assez grand** (on se ramène au cas gaussien grâce au théorème central limite).
 - ▶ **Tests non paramétriques :** valides sans hypothèse sur la distribution des X_i ; moins puissants que les tests paramétriques. Utilisés uniquement pour les **petits échantillons non gaussiens**.

Tests paramétriques sur un échantillon

Hypothèses données dans le cas bilatéral, mais possible aussi en unilatéral.

Population de référence de moyenne μ_0 et de variance σ_0^2 .

n -échantillon X_1, \dots, X_n de loi $\mathcal{N}(\mu, \sigma^2)$ (resp. $\mathcal{B}(n, p)$); estimateurs \bar{X}_n et S_n^2 .

- **Moyenne d'une loi gaussienne** $H_0 : \mu = \mu_0$.

- ▶ **Variance σ^2 connue.** Statistique de test $T = \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \sim \mathcal{N}(0, 1)$.

- ▶ **Variance σ^2 inconnue.** Statistique de test $T = \sqrt{n} \left(\frac{\bar{X}_n - \mu}{S_n} \right) \sim \mathcal{T}(n-1)$

⇒ Test de Student **[commande t.test avec R Studio]**.

Pour n grand, approximation par une $\mathcal{N}(0, 1)$ (voir exemple photovoltaïque).

- **Variance d'une loi gaussienne** $H_0 : \sigma^2 = \sigma_0^2$.

Loi de la statistique de test : $\chi^2(n-1)$.

- **Proportion d'une loi binomiale** $H_0 : p = \mu_0$.

Loi de la statistique de test : approximation par $\mathcal{N}(0, 1)$ pour n grand.

NB : pour un test bilatéral, rejet de $H_0 : \mu = \mu_0 \Leftrightarrow \mu_0 \notin IC_{1-\alpha}(X_1, \dots, X_n)$.

Tests paramétriques sur deux échantillons

Population 1 : loi $\mathcal{N}(\mu_1, \sigma_1^2)$ (resp. $\mathcal{B}(n_1, p_1)$). Échantillon de taille n_1 .

Population 2 : loi $\mathcal{N}(\mu_2, \sigma_2^2)$ (resp. $\mathcal{B}(n_2, p_2)$). Échantillon de taille n_2 .

Idée : se ramener aux cas précédents en considérant $\mu = \mu_1 - \mu_2$ et $\mu_0 = 0$.

- **Comparaison des moyennes** $H_0 : \mu_1 = \mu_2$.

- ▶ **Variances σ_1^2 et σ_2^2 connues.** Statistique de test suit une $\mathcal{N}(0, 1)$.

- ▶ **Variances σ_1^2 et σ_2^2 inconnues.** Statistique de test suit une loi de Student (ou approximation par une $\mathcal{N}(0, 1)$ si n_1 et n_2 grands).

- **Comparaison des proportions** $H_0 : p_1 = p_2$.

- ▶ Statistique de test : approximation par une $\mathcal{N}(0, 1)$.

L'ANOVA, pour au moins 3 échantillons

- **ANOVA** : **AN**alyse **Of** **VA**riance; test de **comparaison des moyennes** ! (mais basé sur la décomposition de la variance)
- Étude de l'impact d'une variable qualitative X , à k modalités sur une variable quantitative Y . ↪ **exemple** : un fabricant d'éoliennes souhaite tester l'efficacité de quatre modèles de pales différentes et relève la production d'énergies sur une partie de son champ éolien, obtenant 4 échantillons en regroupant les observations selon le type de pale.
- Hypothèses à vérifier pour que le test soit valide :
 - ▶ l'indépendance des échantillons ;
 - ▶ la variable Y suit une loi gaussienne (ou chaque échantillon est suffisamment grand) ;
 - ▶ les variances à l'intérieur de chaque sous-groupe (à modalité fixée) sont identiques : on parle d'**homoscédasticité**. Pour le vérifier : test de Bartlett [**commande bartlett.test avec R Studio**].
- Test : H_0 : "les moyennes de chaque sous-groupe (ou classe) sont égales" contre H_1 : "au moins deux moyennes sont différentes".

L'ANOVA, pour au moins 3 échantillons

- **Principe de l'ANOVA** :
 - ▶ **Décomposition de la variance** de Y : variance inter-classe + variance intra-classe.
 - ▶ **Statistique de test F** : d'autant plus grande que la variance inter-classe l'emporte que la variance intra-classe, car signe de la dispersion des moyennes des classes.
 - ▶ Sous H_0 , F suit une loi de Fisher.
 - ▶ Rejet de H_0
 - ⇒ F "grand"
 - ⇒ grande dispersion des moyennes des échantillons
 - ⇒ au moins deux moyennes sont différentes
 - ⇒ la variable qualitative X a un impact sur la variable quantitative Y .
- [**commande aov avec R Studio**]
- Si les hypothèses de l'ANOVA ne sont pas satisfaites, test non paramétrique de Kruskal-Wallis (voir plus loin).

Tests d'ajustement. Cas discret : les test du χ^2 .

● Le test d'adéquation du χ^2 [commande `chisq.test` avec R Studio]

- ▶ **Cadre** : soient (X_1, \dots, X_n) un n -échantillon d'une variable aléatoire X à valeurs dans $\{x_1, \dots, x_k\}$ et $p = (p_1, \dots, p_k) \in \mathbb{R}^k$ tel que $p_1 + \dots + p_k = 1$.
- ▶ **Hypothèse nulle** : $H_0 : \forall i \in \{1, \dots, k\}, \mathbb{P}(X = x_i) = p_i$ \rightsquigarrow **quid de H_1 ?**
- ▶ **Statistique de test** :

$$D^2 = \sum \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}} = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \sim \chi^2(k-1)$$

avec N_i nombre d'occurrences de x_i dans l'échantillon observé.

- ▶ **Règle de décision** : cadre unilatéral
 - ★ Rejet de $H_0 \Leftrightarrow D_{obs}^2 > d_{1-\alpha}$ avec $d_{k-1, 1-\alpha}$ quantile d'ordre $1-\alpha$ de $\chi^2(k-1)$;
 - ★ p -valeur $\hat{\alpha} = \mathbb{P}_{H_0}(D^2 \geq D_{obs}^2)$.

Exemple : le dé pipé (ou pas)

Un enseignant de maths souhaitant faire une introduction aux probabilités à ses élèves veut vérifier a priori que son dé n'est pas pipé... pour ce faire, il lance son dé à 100 reprises.

- Modélisation : soit (X_1, \dots, X_{100}) un échantillon d'une variable aléatoire X à valeurs dans $\{1, \dots, 6\}$.
- Hypothèse nulle : $H_0 : "X$ suit la loi uniforme discrète sur $\{1, \dots, 6\}"$ ie $H_0 : \forall i \in \{1, \dots, 6\}, \mathbb{P}(X = i) = 1/6"$.

Face	1	2	3	4	5	6
$N_{i,obs}$: Effectifs	7	18	26	15	18	16
np_i	16.67	16.67	16.67	16.67	16.67	16.67
$\frac{(N_{i,obs} - np_i)^2}{np_i}$	5.61	0.11	5.23	0.17	0.11	0.03

- Statistique de test : $D_{obs}^2 = 11,24$; quantiles : $d_{5,0,95} = 11,07$, $d_{5,0,975} = 12,83$.
- p -valeur : $\hat{\alpha} = 0,0468 \rightsquigarrow$ **conclusion ?**

Tests d'ajustement. Cas discret, cas continu.

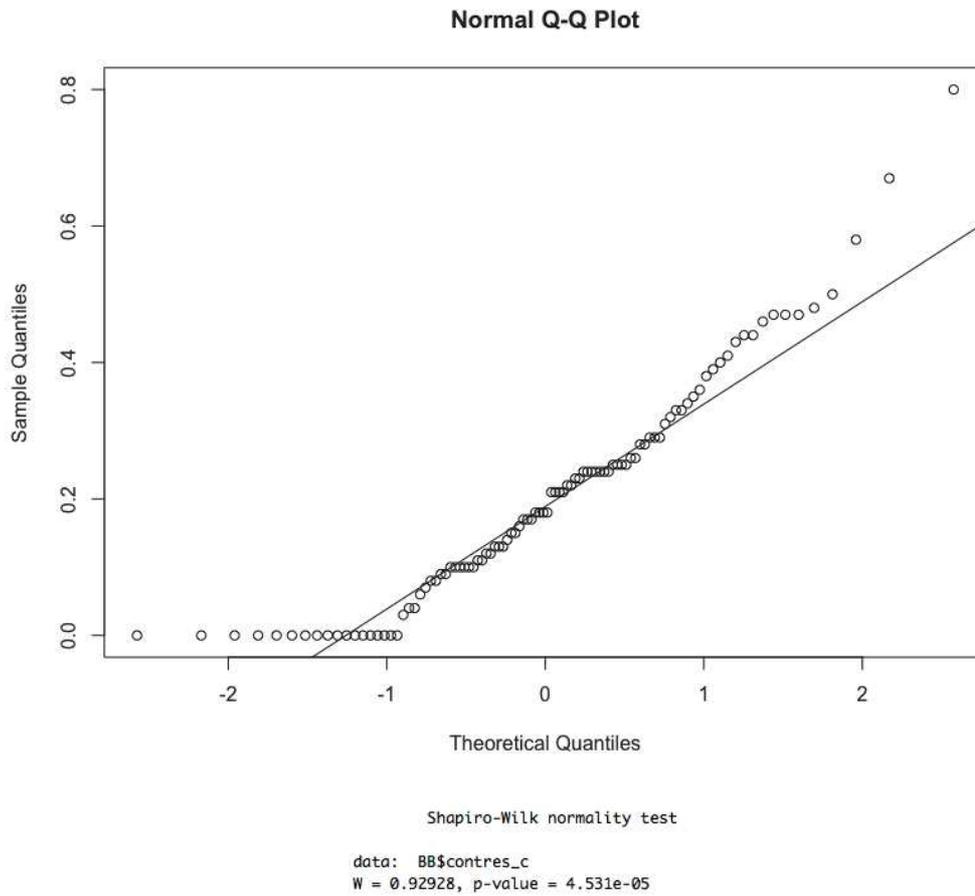
- Cas **discret**, suite et fin : le **test de contingence du χ^2**
 - ▶ Cas particulier du test d'adéquation.
 - ▶ Teste l'**indépendance de deux variables qualitatives**.
 - ▶ Basé sur la notion de **table de contingence**.
 - ▶ Statistique de test : indicateur du khi-deux défini dans le **chapitre de Statistique descriptive**.
 - ▶ Test asymptotique, avec une loi du χ^2 .
- Cas **continu** : le **test de Kolmogorov**
 - ▶ Soient (X_1, \dots, X_n) un n -échantillon d'une variable aléatoire continue X de fonction de répartition F_X et $F_{théo}$ une fonction de répartition continue connue (loi normale, exponentielle).
 - ▶ Hypothèse nulle : $H_0 : "F_X = F_{théo}"$.
 - ▶ Statistique de test = distance entre la fonction de répartition empirique et la fonction $F_{théo}$.
 - ▶ Peut en particulier tester la normalité d'une distribution.
 - ▶ **[commande ks.test avec R Studio]**

Tests d'ajustement. Cas continu, indicateurs de normalité.

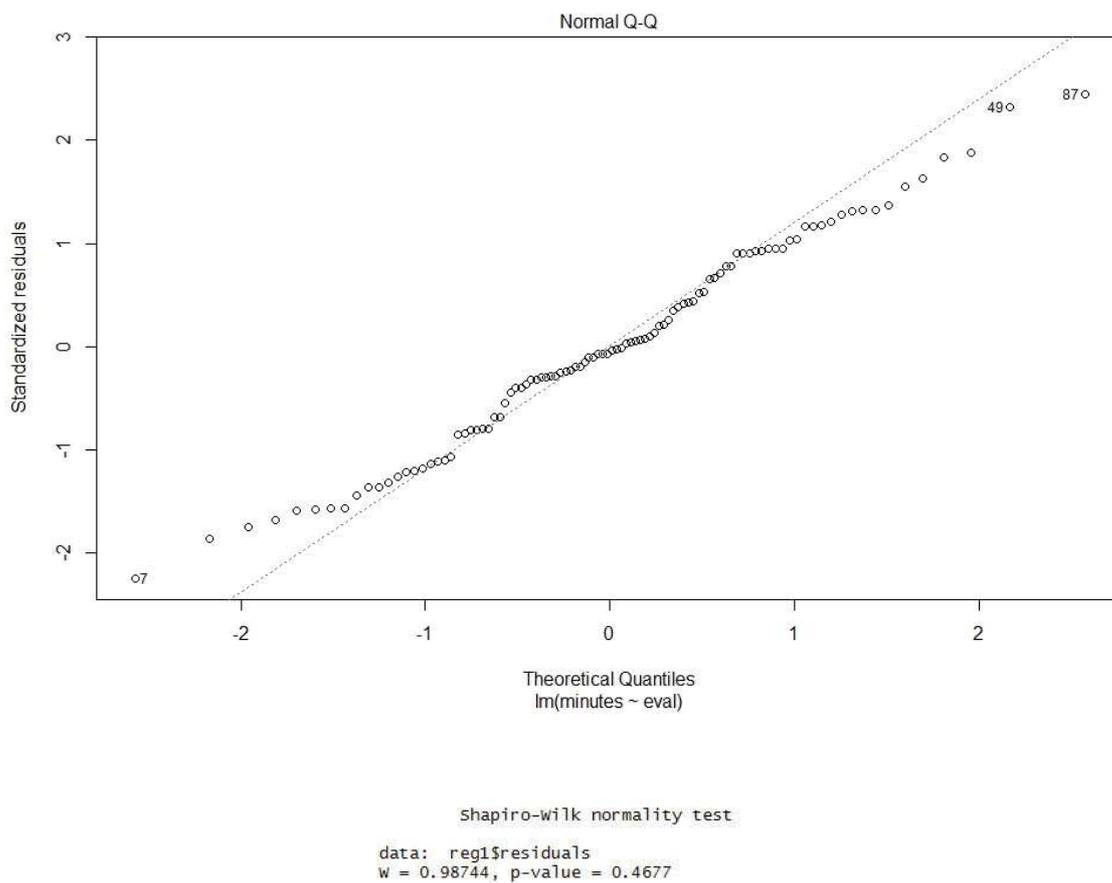
- **Droite de Henri** (ou QQ-plot)
 - ▶ **Visualisation graphique** de la normalité d'une distribution.
 - ▶ Basée sur une transformation de l'échelle des ordonnées et une inverse la fonction de répartition gaussienne.
 - ▶ Principe : $X \sim \mathcal{N}(m, \sigma^2) \Leftrightarrow$ **points alignés sur la droite** d'équation $y = \frac{x - m}{\sigma}$; autrement dit X suit une loi normale si on obtient un nuage de points le long d'une droite ; on peut ensuite déterminer les coefficients de cette distribution :
 - ▶ m : valeur pour laquelle on coupe l'axe des abscisses ; σ : inverse du coefficient directeur de la droite.
 - ▶ **[commandes qqnorm et qqline avec R Studio]**
- **Test de Shapiro-Wilk**
 - ▶ Hypothèse nulle H_0 : "la population suit une distribution gaussienne".
 - ▶ Très utilisé pour vérifier les hypothèses des tests, modèles linéaires, etc.
 - ▶ **[commande shapiro.test avec R Studio]**

↪ En pratique : droite de Henri, puis test de Shapiro-Wilk pour confirmer.

Droite de Henri et test de Shapiro-Wilk : exemple 1



Droite de Henri et test de Shapiro-Wilk : exemple 2



Tests non paramétriques d'homogénéité

↪ Pour les **petits échantillons non gaussiens**.

- **Test de Wilcoxon-Mann-Whitney**

- ▶ Test non paramétrique d'homogénéité pour deux échantillons.
- ▶ Hypothèse nulle : H_0 : "les médianes sont égales".
- ▶ Si l'hypothèse est rejetée, on peut en déduire que les populations dont sont issues les échantillons sont probablement différentes.
- ▶ [\[commande wilcox.test avec R Studio\]](#)

- **Test de Kruskal-Wallis**

- ▶ Généralisation du test de Wilcoxon-Mann-Whitney pour k échantillons, $k \geq 2$.
- ▶ Hypothèse nulle : H_0 : "les médianes sont égales".
- ▶ [\[commande kruskal.test avec R Studio\]](#)

Chapitre 6 : Régression linéaire simple

Location d'un T1 et prix au m^2 ... un indicateur cohérent ?

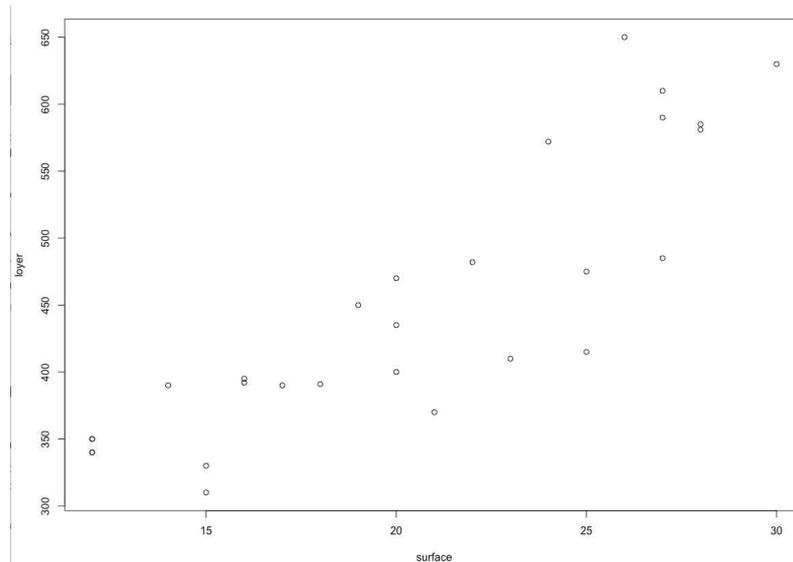


Figure – Loyer en fonction de la surface pour des T1 dans l'ultracentre de Toulouse. Annonces issues du site leboncoin.fr.

- Que peut-on dire sur la **liaison** entre la surface d'un appartement et son loyer ? Est-elle **linéaire**, **affine** ? Quid de la notion de prix au m^2 ?
- Peut-on **prévoir** le loyer d'un studio de $23 m^2$?

La régression linéaire simple

1 Le modèle de régression linéaire simple

- Modélisation et hypothèses
- Estimation des paramètres
- Prévision et résidus

2 Validation, et qualité, de la modélisation

- Diagnostic des résidus
- Significativité du modèle
- Qualité d'ajustement du modèle

3 Étude de trois cas

- La taille d'un nourrisson
- La distance de freinage
- Extrait d'annale : l'évaluation au basket

On enchaîne avec...

- 1 Le modèle de régression linéaire simple
 - Modélisation et hypothèses
 - Estimation des paramètres
 - Prévission et résidus
- 2 Validation, et qualité, de la modélisation
 - Diagnostic des résidus
 - Significativité du modèle
 - Qualité d'ajustement du modèle
- 3 Étude de trois cas
 - La taille d'un nourrisson
 - La distance de freinage
 - Extrait d'annale : l'évaluation au basket

Régression linéaire : principe et modélisation

Idée : **modéliser** une **relation entre deux variables quantitatives** X et Y en **expliquant** Y comme une **fonction affine** de X .

↔ par exemple : expliquer Y , le loyer d'un studio toulousain, comme une fonction affine de X , sa surface.

Modèle de régression linéaire simple

$$Y = \beta_0 + \beta_1 X + \epsilon$$

avec :

- Y variable aléatoire à expliquer ;
- X variable non aléatoire explicative ;
- ϵ terme d'erreur aléatoire ;
- β_0 et β_1 paramètres à estimer.

Pour n observations,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Commande lm sur R Studio pour un modèle de régression linéaire.

Régression linéaire : hypothèses à vérifier

Modèle de régression linéaire simple

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Hypothèses de validité du modèle

- X est déterministe (i.e. non aléatoire)
- **Homoscédasticité** des erreurs (i.e. égalité des variances) :

$$\forall i \in \{1, \dots, n\}, \text{var}(\epsilon_i) = \sigma^2.$$

- Indépendance des erreurs : les $(\epsilon_i)_i$ sont indépendantes.
- Normalité des erreurs : $\forall i \in \{1, \dots, n\}, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Hypothèses à vérifier a posteriori pour voir si le modèle est valide.

Un peu de terminologie :

- régression : ensemble des méthodes pour expliquer une variable par une autre ;
- linéaire : linéarité en β_0 et β_1 ;
- simple : une seule variable explicative.

Estimation des paramètres

Idée : estimer les paramètres β_0 , β_1 et σ^2 à partir des observations $(x_i, y_i)_{i \in \{1, \dots, n\}}$.

- **Estimations de β_0 et β_1**
→ Estimateur des moindres carrés : (voir **TP d'optimisation en 2ICBE**)

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

→ Le critère des moindres carrés est minimisé par

$$b_1 = \frac{s_{xy}}{s_x^2} \text{ et } b_0 = \bar{y} - b_1 \bar{x}$$

avec \bar{x} et s_x^2 (resp. \bar{y} et s_y^2) estimations de la moyenne et de la variance de X (resp. Y) et s_{xy} estimation de la covariance entre X et Y .

Les estimations b_0 et b_1 sont des réalisations des **estimateurs** $\hat{\beta}_0$ et $\hat{\beta}_1$.

- **Estimation de σ^2**

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

↪ dans notre exemple, $b_1 = 14,04$, $b_0 = 160,88$ et $s^2 = 3853$.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 160.876    27.714   5.805 1.11e-05 ***
surface     14.044     1.003  14.001 8.52e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62.08 on 20 degrees of freedom
Multiple R-squared:  0.9074,    Adjusted R-squared:  0.9028
F-statistic: 196 on 1 and 20 DF,  p-value: 8.521e-12
```

- En **vert**, le terme constant, aussi appelé *Intercept*, et son estimation b_0 .
- En **orange**, le coefficient devant la variable X (ici la surface de l'appartement), et son estimation b_1 .
- En **rouge**, l'estimation de la variance commune des erreurs/résidus.
- **Si le modèle est valide**, la régression obtenue est

$$\text{Loyer} = 160,88 + 14,04 \times \text{Surface}.$$

Valeur estimée et résidus

- **Valeur estimée** ou ajustée de Y : pour tout $i \in \{1, \dots, n\}$,

$$\hat{y}_i = b_0 + b_1 x_i$$

- **Valeur prédite** : connaissant une valeur x_0 ,

$$\hat{y}_0 = b_0 + b_1 x_0$$

↪ dans notre exemple, pour $x_0 = 23m^2$, $\hat{y}_0 = 484$ euros/mois.

Sous réserve de validité du modèle.

- **Résidus** : différence entre la valeur observée et la valeur estimée

$$e_i = y_i - \hat{y}_i$$

↪ correspond à l'erreur entre ce qui est observé, et ce que le modèle prédit.

On enchaîne avec...

- 1 Le modèle de régression linéaire simple
 - Modélisation et hypothèses
 - Estimation des paramètres
 - Préviation et résidus

- 2 Validation, et qualité, de la modélisation
 - Diagnostic des résidus
 - Significativité du modèle
 - Qualité d'ajustement du modèle

- 3 Étude de trois cas
 - La taille d'un nourrisson
 - La distance de freinage
 - Extrait d'annale : l'évaluation au basket

De la pertinence d'un modèle de régression linéaire

↪ Comment juger du bien-fondé de la modélisation par régression linéaire ?

- **Étape 1 : diagnostic des résidus.**
Vérifier, à l'aide des résidus, que les hypothèses de **validité du modèle** sont satisfaites, notamment la **linéarité** entre Y et X , et l'**homoscédasticité** et la **normalité des résidus**.

- **Étape 2 : test de Fisher, ou de Student.**
Tester la **significativité du modèle** revient à **tester** l'hypothèse H_0 : " $\beta_1 = 0$ " : si on accepte l'hypothèse, la modélisation n'a pas de sens.

- **Étape 3 : calcul du coefficient de détermination.**
Donner une indication de la **qualité d'ajustement du modèle**.

Etape 1 : diagnostic des résidus

- **Normalité des résidus** : utilisation de deux outils vus au **chapitre précédent**.
 - ▶ Droite de Henri
 - ▶ Test de Shapiro-Wilk

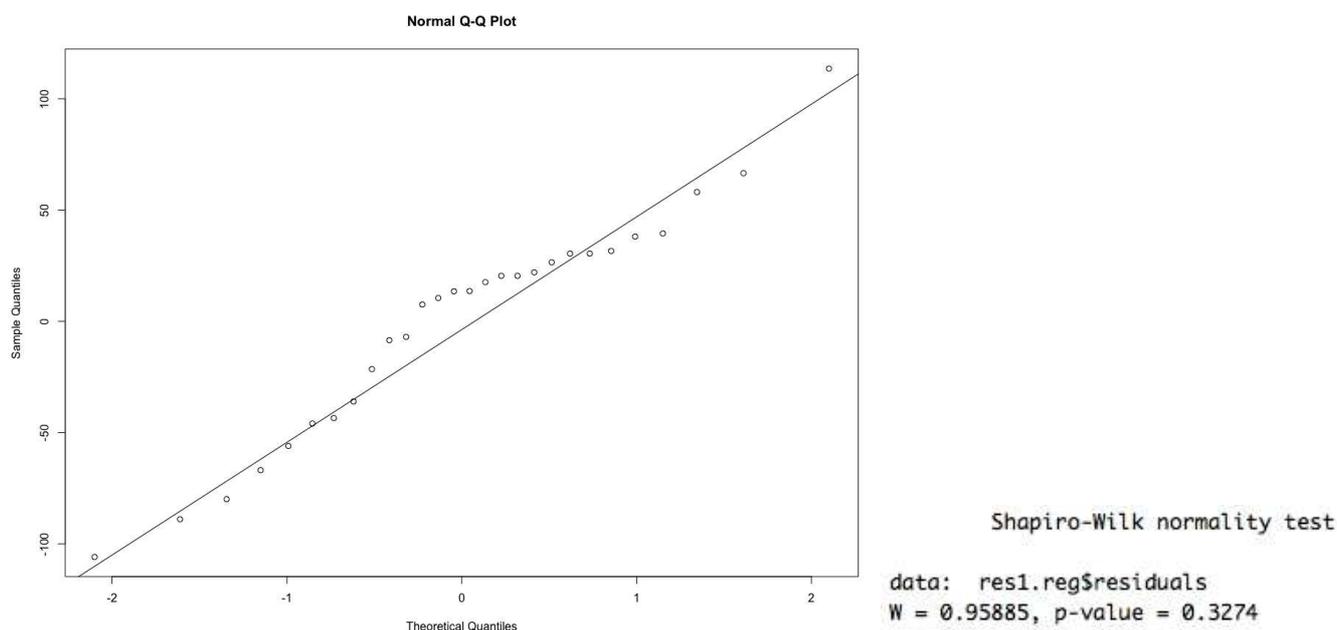


Figure – Droite de Henri, à gauche, test de Shapiro-Wilk, à droite, pour l'exemple des studios toulousains. Ici, l'hypothèse de normalité n'est pas rejetée. *Logiciel : R Studio.*

Etape 1 : diagnostic des résidus

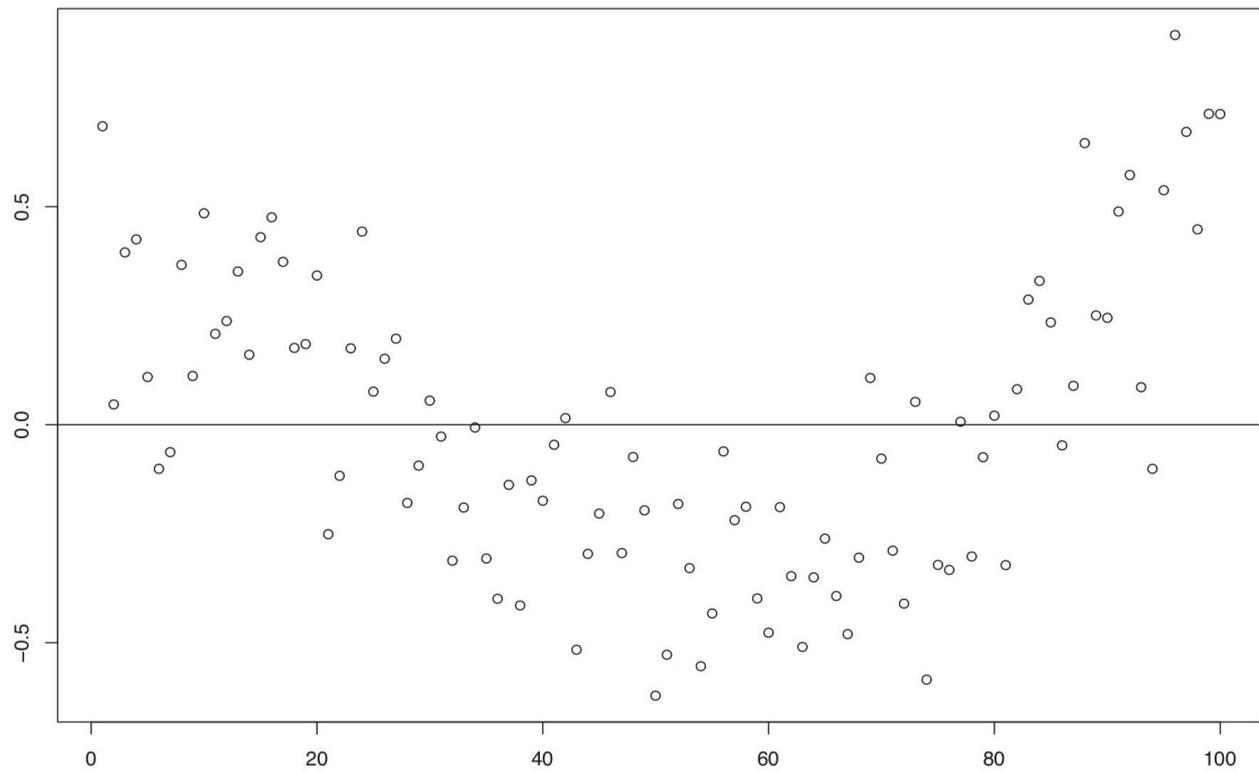
- **Grphe des résidus** : on y lit les résidus (renormalisés) ; selon la forme du nuage de points, pour l'**homoscédasticité des résidus** et la **linéarité de la liaison entre Y et X** :
 - ▶ **Hétéroscédasticité** : si le nuage est en "forme d'entonnoir ou de diabol".
 - ▶ **Non-linéarité** : si le nuage est en "forme de banane".
 - ▶ **Homoscédasticité et linéarité** : si la dispersion est "normale" de part et d'autre de l'axe, c'est-à-dire symétrique et sans forme particulière.

Deux remarques sur la vérification des hypothèses :

- ★ si l'échantillon est grand, on peut passer outre l'hypothèse de **normalité** ;
- ★ en cas de **non-linéarité**, parfois judicieux de transformer les données (voir **chapitre 2**).

- **Valeurs aberrantes, influentes et distance de Cook** :
 - ▶ **Observation atypique** : valeur qui dépasse 2 en valeur absolue sur le graphe des résidus. Potentiellement influente.
 - ▶ **Distance de Cook** : mesure de l'influence d'une observation sur l'ensemble des prévisions.
 - ▶ **Valeur influente** : observation atypique dont la distance de Cook est plus grande que 1.
 - ▶ La présence de valeurs influentes peut mettre en doute la validité du modèle.

Un exemple de non-linéarité : le cas de la banane



L'exemple des loyers toulousains

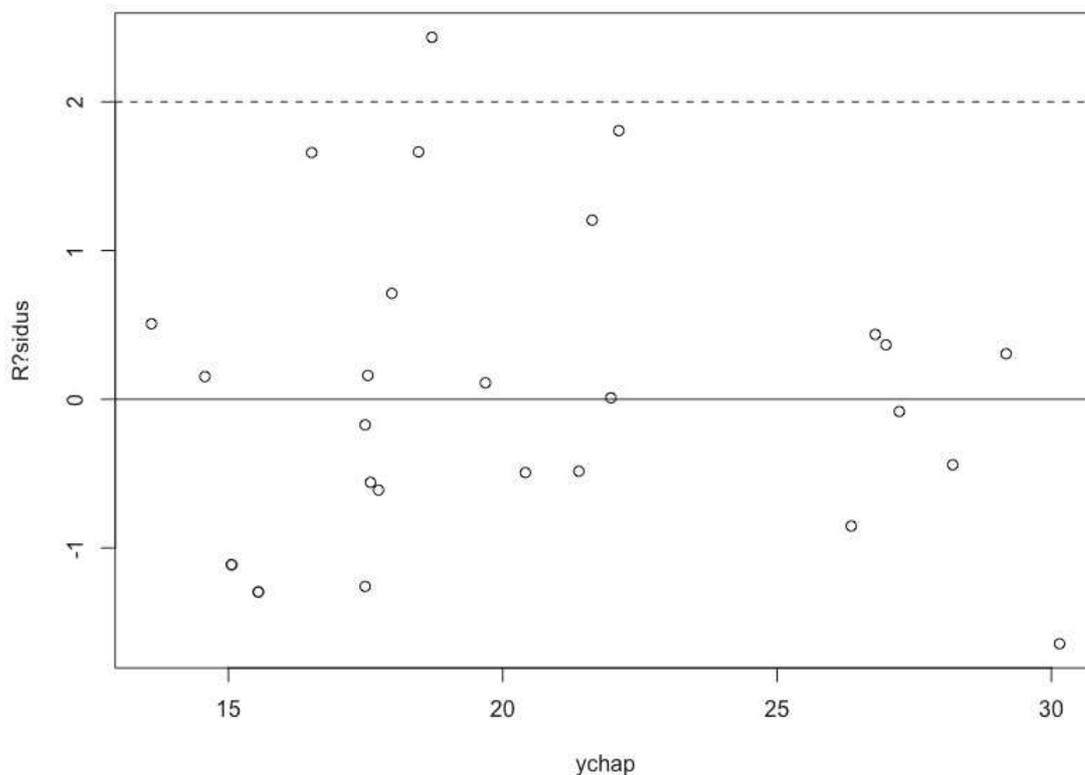


Figure – Résidus renormalisés associés à la régression linéaire du loyer par rapport à la surface pour un échantillon d'appartements toulousains de moins de 30 m^2 .

L'exemple des loyers toulousains

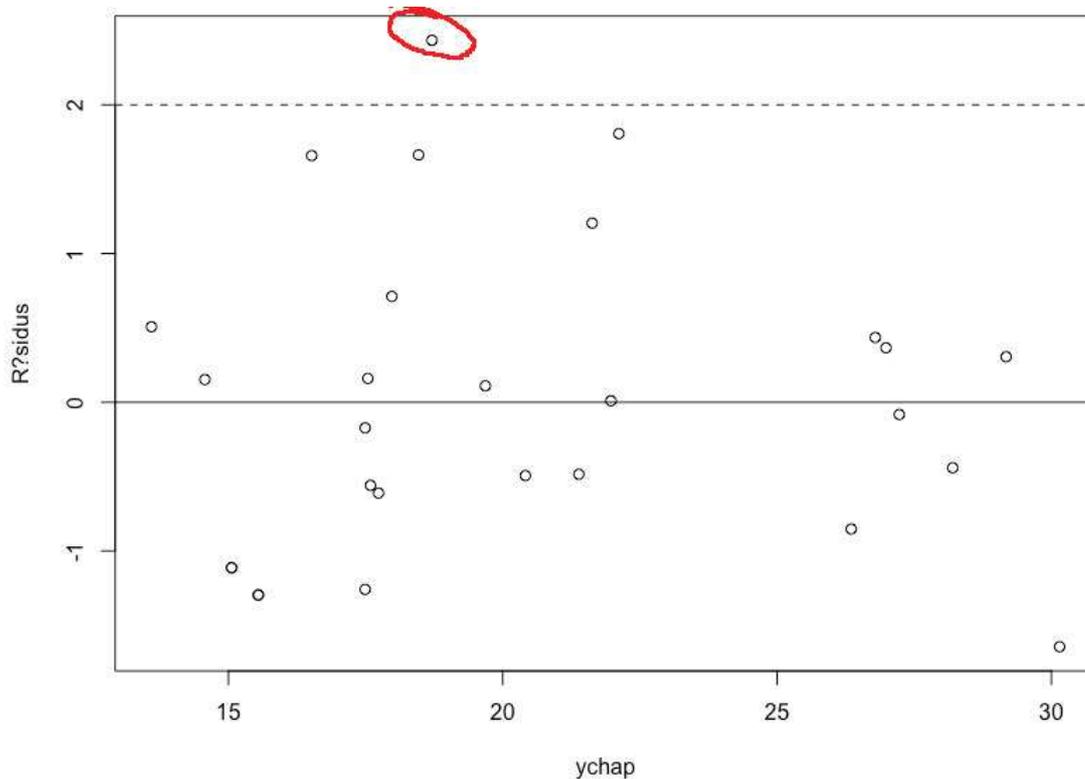


Figure – Résidus renormalisés associés à la régression linéaire du loyer par rapport à la surface pour un échantillon d'appartements toulousains de moins de $30 m^2$.

L'exemple des loyers toulousains

Une observation atypique \Rightarrow Étude des distances de Cook.

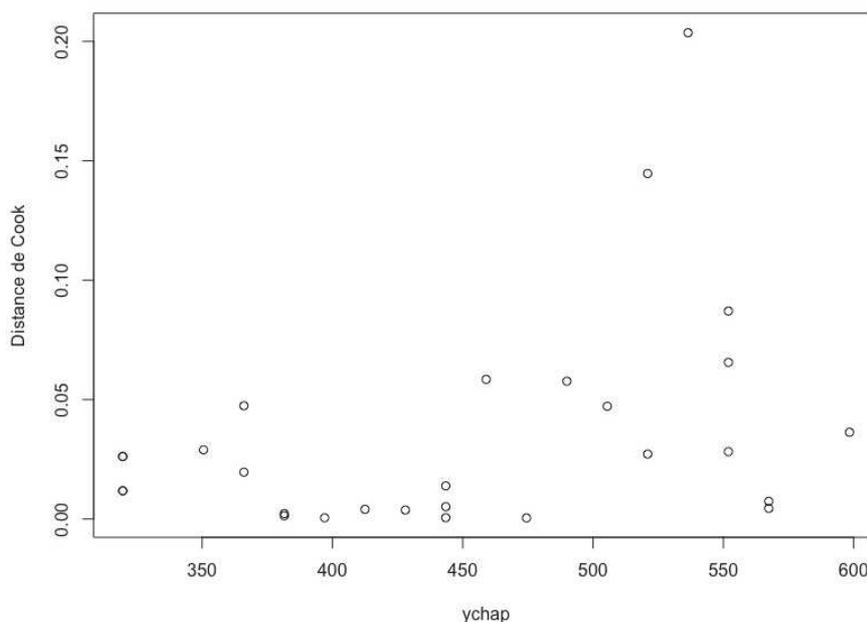


Figure – Distance de Cook associée aux résidus.

Toutes les valeurs sont inférieures à 1 \Rightarrow pas de valeurs influentes.

Etape 2 : significativité du modèle, test de Fisher/Student

Test de Fisher ou Student d'hypothèse $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$.

- Tester la **significativité du modèle** revient à **tester** l'hypothèse $H_0 : \beta_1 = 0$:
 - ▶ si on rejette H_0 , c'est que le coefficient directeur de la droite de régression est non nul, et donc que le modèle est significatif ;
 - ▶ a contrario, si on accepte H_0 , c'est que la liaison entre les deux variables est très faible et/ou non-linéaire.
- Lois des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ connues \Rightarrow construction d'une statistique de test. Deux statistiques de test équivalentes, une qui suit une **loi de Fisher**, l'autre une **loi de Student**.
- **Commande summary sur R Studio** pour lire la p-valeur (identique) de ces deux tests.

Significativité du modèle : les studios toulousains.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  160.876    27.714    5.805 1.11e-05 ***
surface      14.044     1.003   14.001 8.52e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 62.08 on 20 degrees of freedom
Multiple R-squared:  0.9074,    Adjusted R-squared:  0.9028
F-statistic: 196 on 1 and 20 DF,  p-value: 8.521e-12
```

- En **orange**, la statistique de test et la p -valeur du test de Student.
 - En **vert**, la statistique de test et la p -valeur du test de Fisher.
- \rightarrow La p -valeur est très faible ($\ll 0,01$) donc on rejette l'hypothèse $\beta_1 = 0$.
 \rightarrow Le modèle est donc significatif.

Un autre test (en **rouge**) est effectué : l'hypothèse testée est $\beta_0 = 0$; autrement dit, on rejette l'hypothèse si la droite ne passe pas par l'origine. Rarement utile...
... mais nous permet ici de conclure que la liaison entre loyer et surface n'est pas linéaire (rejet de l'hypothèse).

Étape 3 : qualité d'ajustement du modèle

● Coefficient de détermination :

$$R^2 = \frac{\text{variance expliquée par le modèle}}{\text{variance totale}} = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

- ▶ compris entre 0 et 1 ;
- ▶ mesure la qualité d'ajustement du modèle ;
- ▶ le modèle de régression linéaire "explique $100 R^2$ % de la variance des données" ;
- ▶ $R^2 = 1 \Rightarrow$ modélisation parfaite, qui explique tout ($\sigma^2 = 0$, donc erreur nulle) ;
- ▶ $R^2 = 0 \Rightarrow$ modélisation complètement inadaptée.
- ▶ R analogue du coefficient de corrélation ;

↪ dans notre exemple, $R^2 = 0,91 \Rightarrow$ modèle très bien ajusté.

En **bleu** sur la diapo précédente.

Attention ! Pas le premier critère à regarder pour juger un modèle !

Et pas le seul indicateur pour comparer des modèles... par exemple **le PRESS**, qui est d'autant plus faible que la qualité de la prévision est importante (voir le **chapitre 7** sur la Régression linéaire multiple).

On enchaîne avec...

- 1 Le modèle de régression linéaire simple
 - Modélisation et hypothèses
 - Estimation des paramètres
 - Prévision et résidus
- 2 Validation, et qualité, de la modélisation
 - Diagnostic des résidus
 - Significativité du modèle
 - Qualité d'ajustement du modèle
- 3 Étude de trois cas
 - La taille d'un nourrisson
 - La distance de freinage
 - Extrait d'annale : l'évaluation au basket

Premier exemple : la croissance des nourrissons français

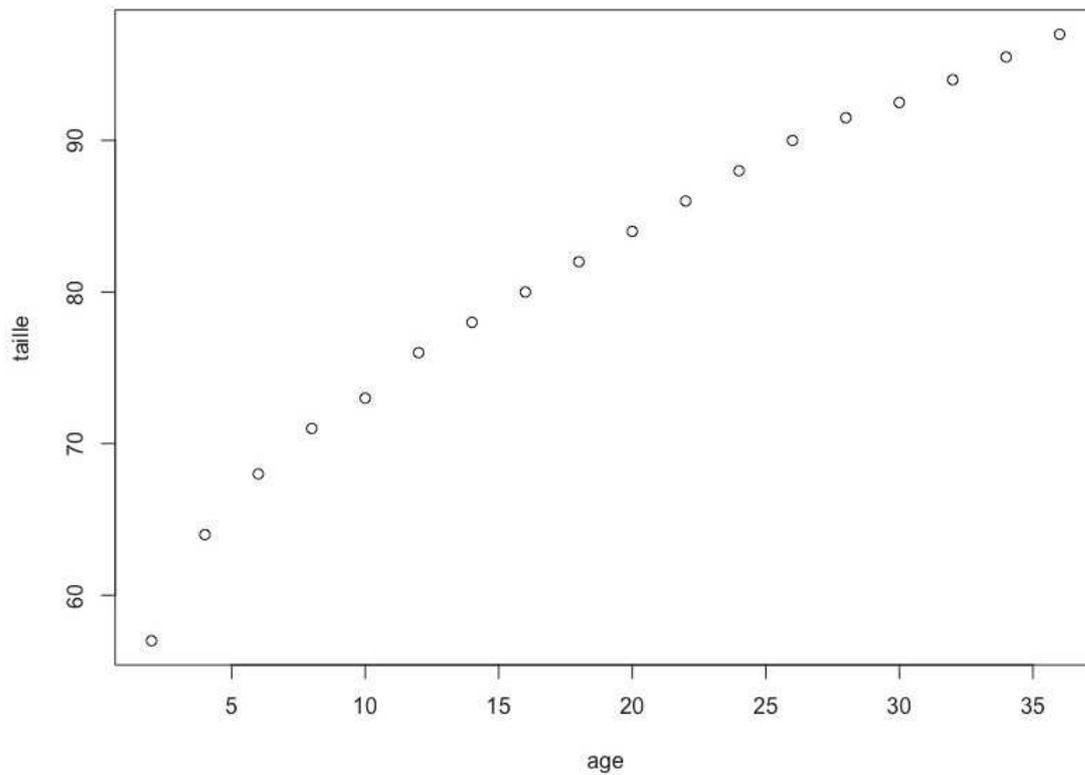


Figure – Observations basées sur la courbe de croissance des garçons français de 1 mois à 3 ans. Données : Association Française de Pédiatrie Ambulatoire.

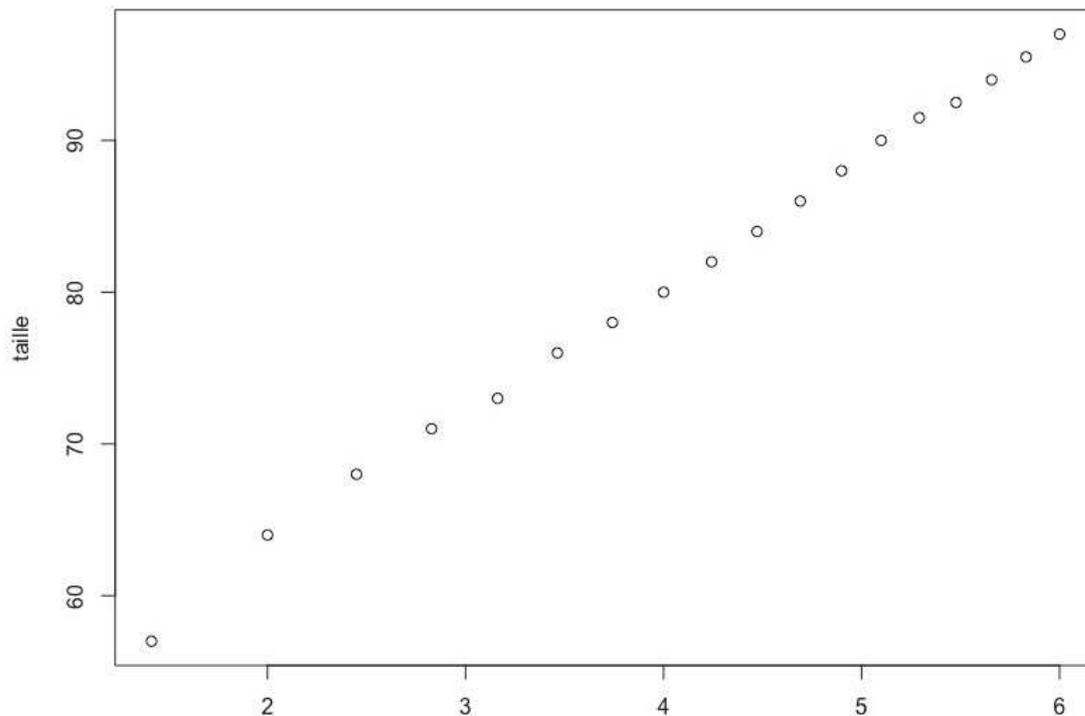


Figure –

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	46.50029	0.40491	114.84	<2e-16 ***
age_2	8.43811	0.09289	90.84	<2e-16 ***

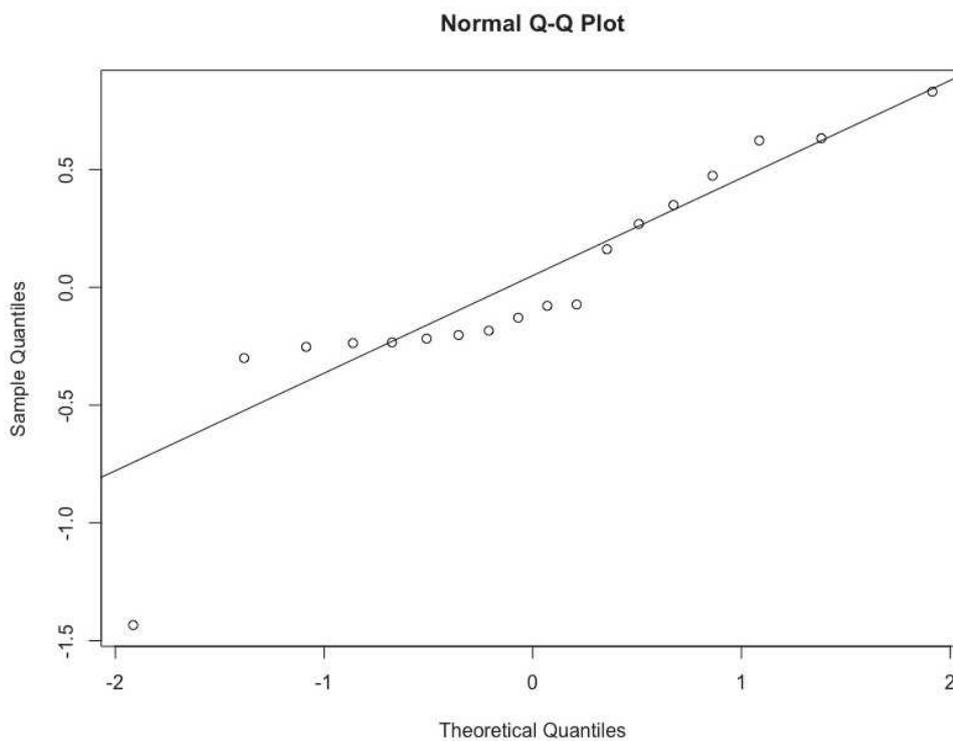
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5241 on 16 degrees of freedom

Multiple R-squared: 0.9981, Adjusted R-squared: 0.9979

F-statistic: 8251 on 1 and 16 DF, p-value: < 2.2e-16

Figure –



Shapiro-Wilk normality test

data: res1.reg\$residuals

W = 0.87859, p-value = 0.02472

Figure –

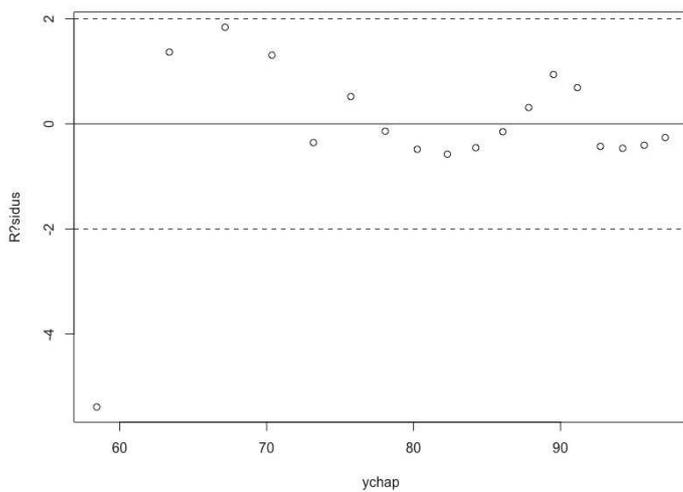


Figure –

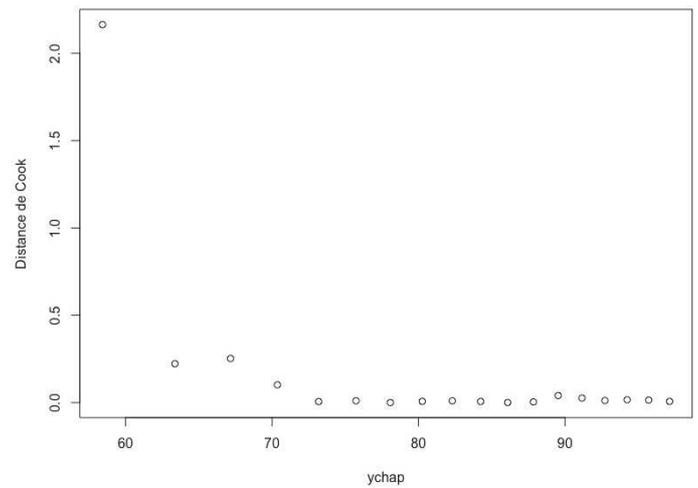


Figure –

Deuxième exemple : distance de freinage d'une Clio

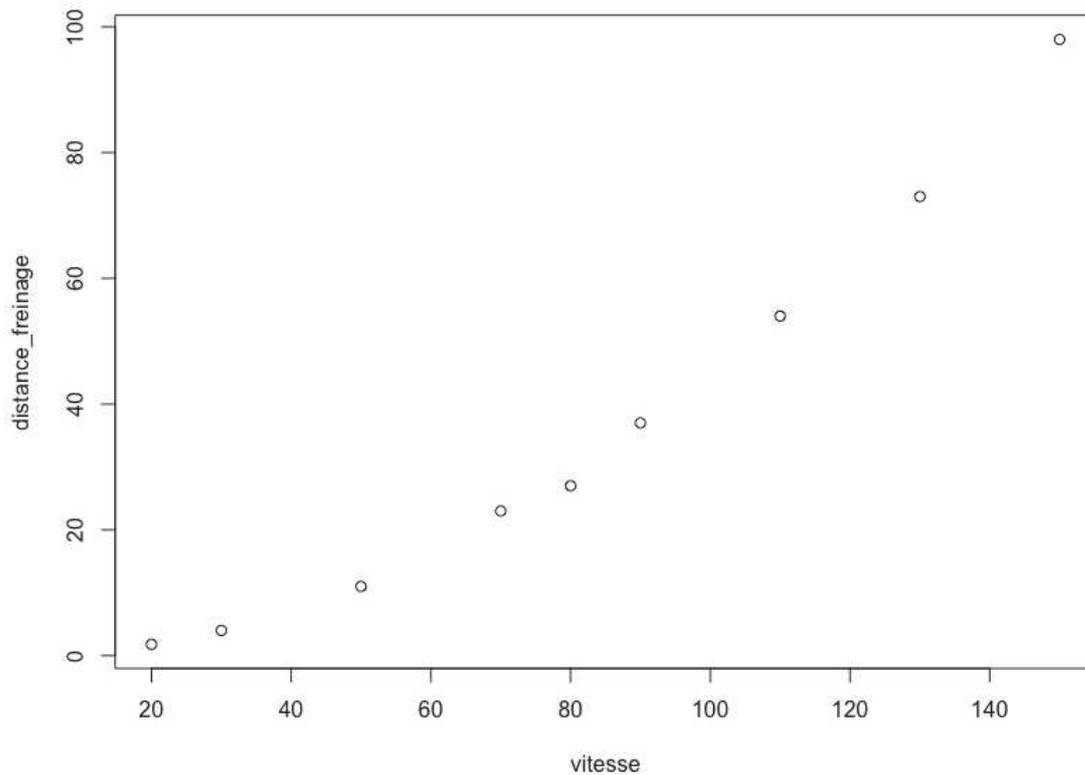


Figure – Données largement fictives. Distance de freinage en fonction de la vitesse du véhicule.

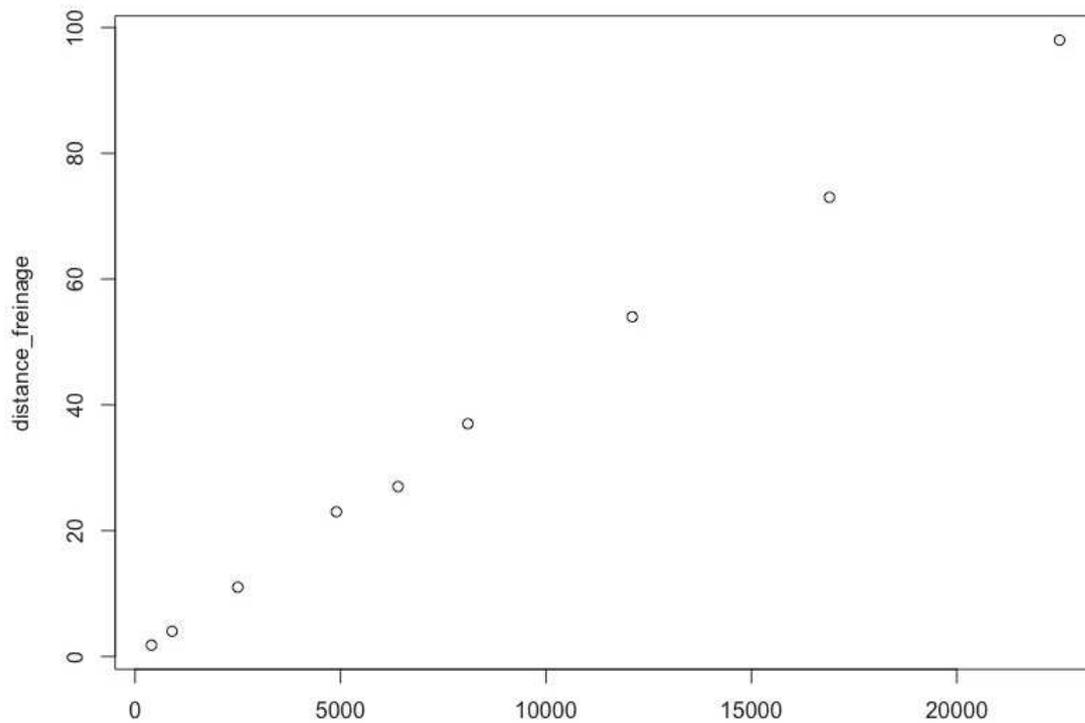


Figure –

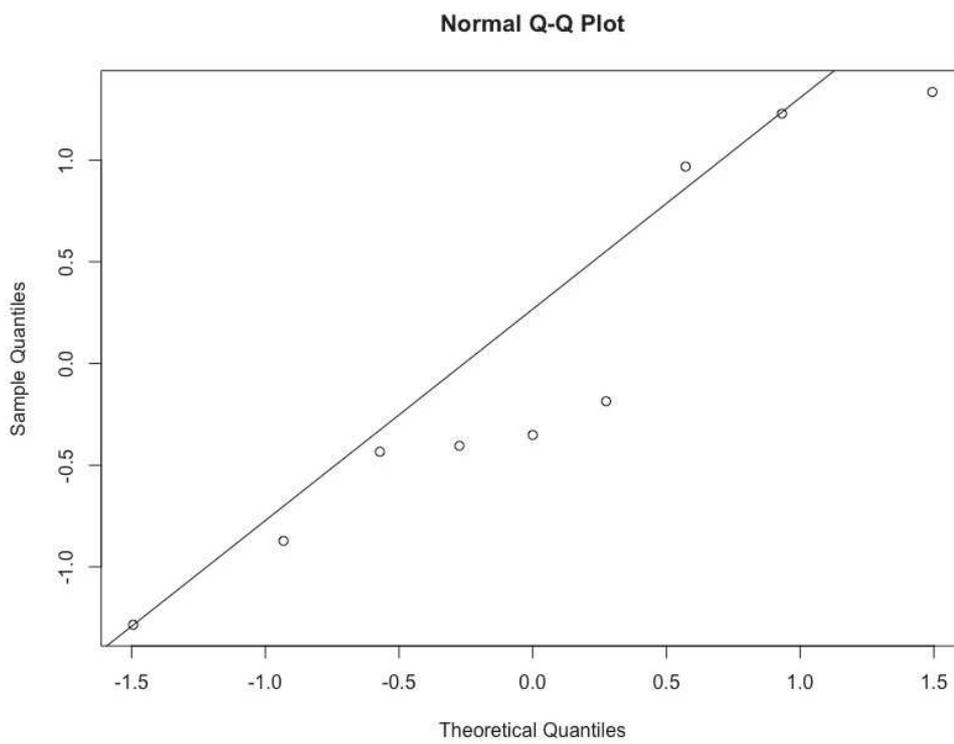
```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.971e-01  5.174e-01  0.961  0.369
vitesse_2    4.342e-03  4.732e-05  91.755 4.81e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.011 on 7 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9991
F-statistic: 8419 on 1 and 7 DF, p-value: 4.811e-12

```

Figure –



Shapiro-Wilk normality test

```

data: res1.reg$residuals
W = 0.88954, p-value = 0.1973

```

Figure –

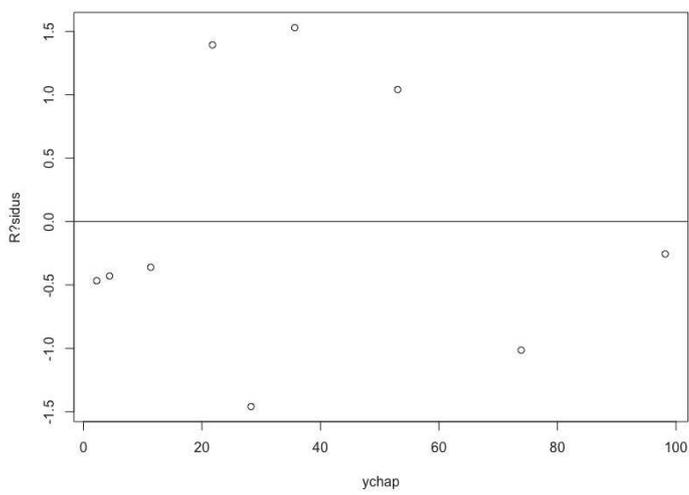


Figure –

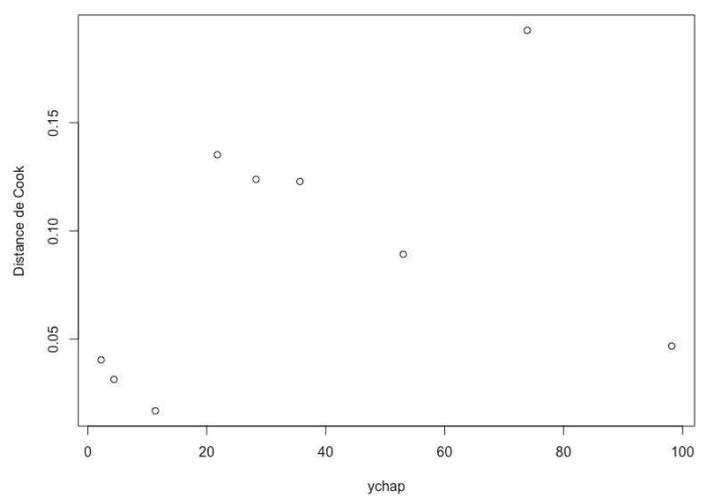


Figure –

Troisième exemple : l'évaluation au basket

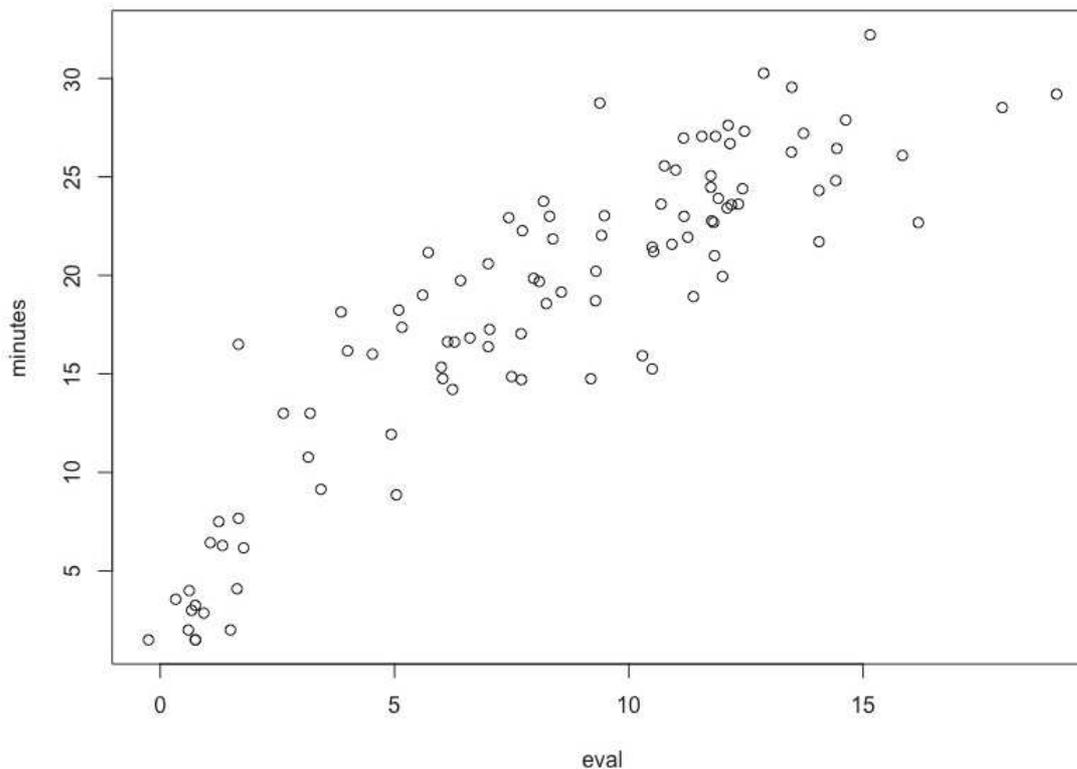
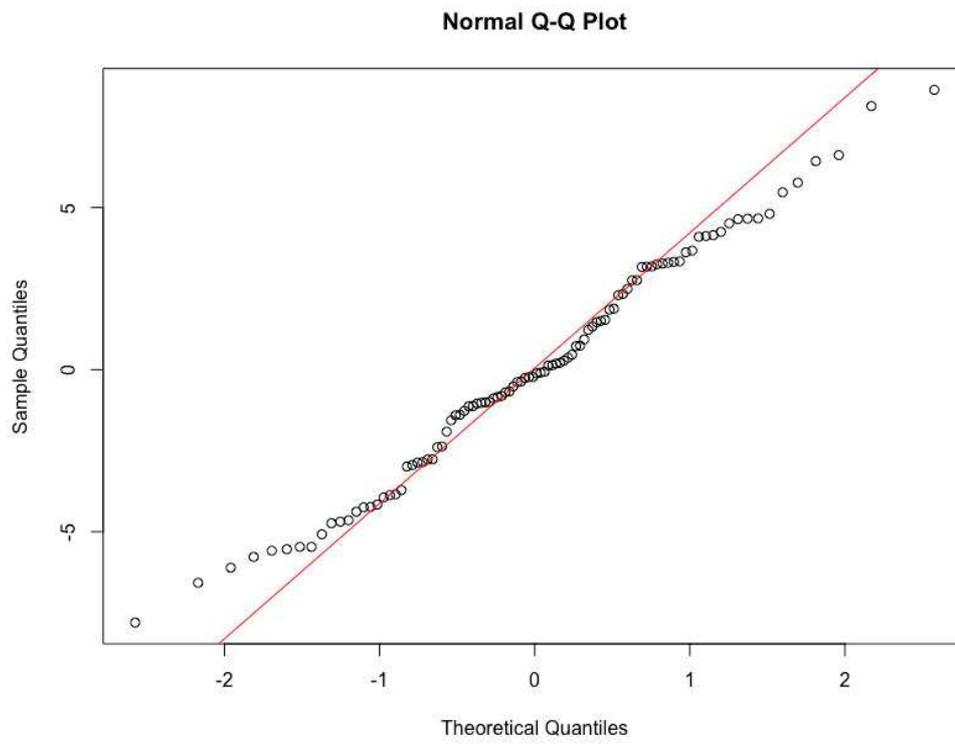


Figure – Evaluation en fonction du temps de jeu de 100 joueurs de basket de Jeep Elite.
Source : Ligue Nationale de Basket.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.82491    0.72375   8.048   2e-12 ***
eval         1.52390    0.07679  19.846  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.546 on 98 degrees of freedom
Multiple R-squared:  0.8008,    Adjusted R-squared:  0.7987
F-statistic: 393.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

Figure –



Shapiro-Wilk normality test

data: reg1\$residuals
 $W = 0.98744$, $p\text{-value} = 0.4677$

Figure –

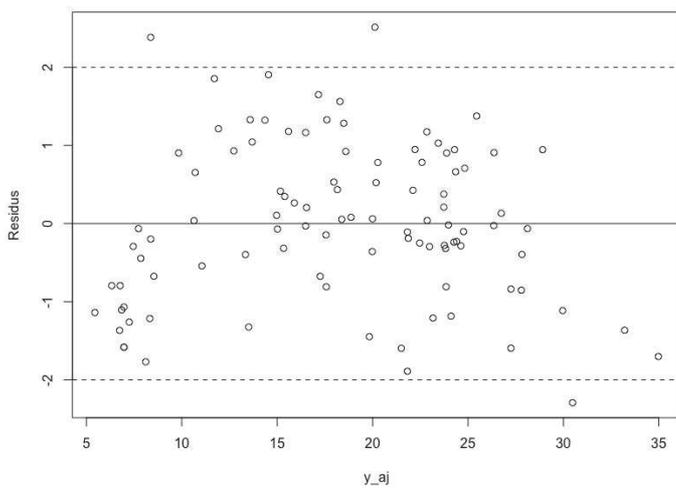


Figure –

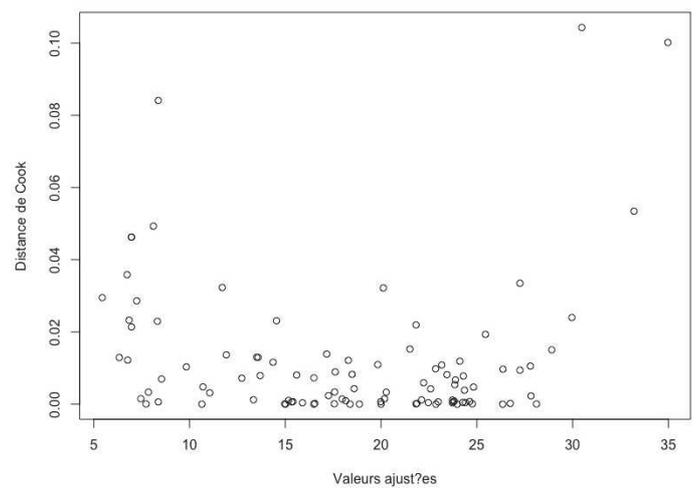


Figure –

Chapitre 7: Introduction à la régression linéaire multiple

1 Modèle de régression linéaire multiple

- Généralisation du cas à un facteur
- Inférence statistique
- Qualité de la modélisation

2 Un exemple météorologique

On enchaîne avec...

1 Modèle de régression linéaire multiple

- Généralisation du cas à un facteur
- Inférence statistique
- Qualité de la modélisation

2 Un exemple météorologique

Modèle de régression linéaire multiple

- **Généralisation** multidimensionnelle de la **régression linéaire simple**.
- Cas où une variable X ne suffit pas à expliquer Y .
- Explication d'une variable quantitative Y comme une **combinaison affine de p variables explicatives** $X^{(1)}, \dots, X^{(p)}$.

Modèle de régression linéaire multiple

$$Y = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \dots + \beta_p X^{(p)} + \epsilon$$

avec :

- Y variable aléatoire à expliquer ;
- $X^{(1)}, \dots, X^{(p)}$ variables non aléatoires explicatives ;
- ϵ terme d'erreur aléatoire ;
- $\beta_0, \beta_1, \dots, \beta_p$ paramètres à estimer.

Pour n observations $(x_i^{(1)}, \dots, x_i^{(p)}, y_i)_{i \in \{1, \dots, n\}}$,

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \dots + \beta_p x_i^{(p)} + \epsilon_i$$

Régression linéaire : hypothèses à vérifier

Modèle de régression linéaire multiple

$$Y = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \dots + \beta_p X^{(p)} + \epsilon$$

Hypothèses de validité du modèle

- Les variables $X^{(1)}, \dots, X^{(p)}$ sont déterministes.
- **Homoscédasticité** des erreurs (i.e. égalité des variances) :

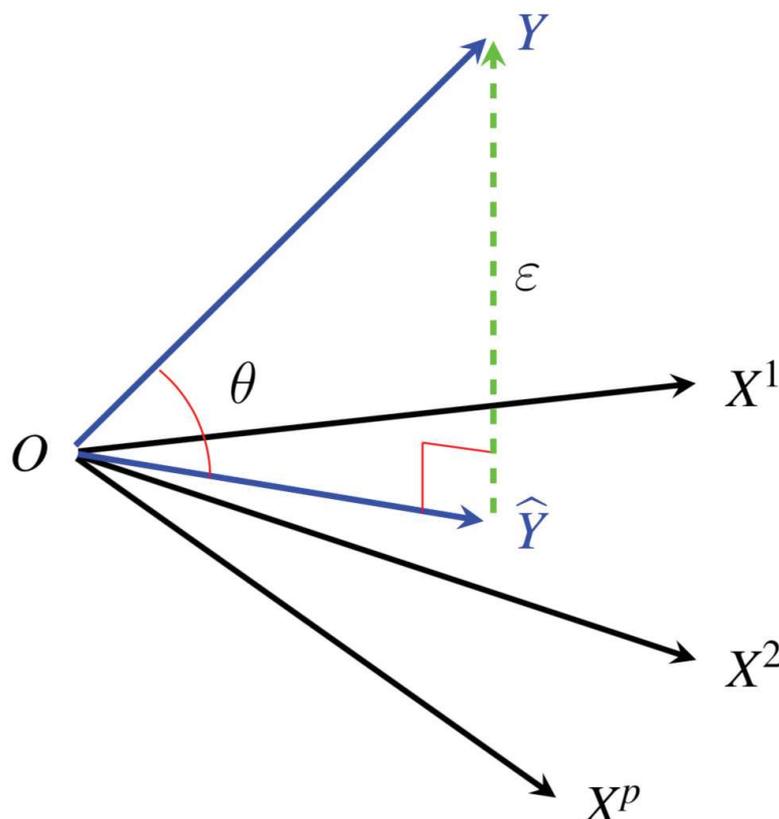
$$\forall i \in \{1, \dots, n\}, \text{var}(\epsilon_i) = \sigma^2.$$

- Indépendance des erreurs : les $(\epsilon_i)_i$ sont indépendantes.
- Normalité des erreurs : $\forall i \in \{1, \dots, n\}, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Hypothèses à vérifier a posteriori pour s'assurer de la validité du modèle : mêmes démarches que pour la régression linéaire simple.

Interprétation géométrique de la régression linéaire multiple

Géométriquement, la régression de Y par $X^{(1)}, \dots, X^{(p)}$ est la projection \hat{Y} de Y sur l'espace vectoriel $\text{Vect}\{\mathbf{1}, X^{(1)}, \dots, X^{(p)}\}$.



Estimation, vocabulaire et diagnostics

- Comme dans le cas simple, **estimations** b_0, \dots, b_p de β_0, \dots, β_p avec la **méthode des moindres carrés** (multidimensionnelle, avec du calcul matriciel).

- Mêmes définitions et notations :

- ▶ **Valeur estimée** ou ajustée de Y : pour tout $i \in \{1, \dots, n\}$,

$$\hat{y}_i = b_0 + b_1 x_i^{(1)} + \dots + b_p x_i^{(p)}$$

- ▶ **Valeur prédite** : connaissant une valeur $(x_0^{(1)}, \dots, x_0^{(p)})$,

$$\hat{y}_0 = b_0 + b_1 x_0^{(1)} + \dots + b_p x_0^{(p)}$$

- ▶ **Résidus** : différence entre la valeur observée et la valeur estimée

$$e_i = y_i - \hat{y}_i$$

- **Diagnostic sur les résidus** : à faire de manière analogue à la régression linéaire simple.

Tests d'hypothèse

Là encore, deux tests différents, avec des rôles cette fois-ci distincts :

- **Test de significativité globale du modèle** : $H_0 : \beta_0 = \dots = \beta_p = 0$.

- ▶ Statistique de test suit une **loi de Fisher**.
- ▶ Acceptation : modélisation inadaptée.
- ▶ Rejet : au moins un des paramètres β_i est non nul.
- ▶ Intérêt limité pour un modèle avec beaucoup de variables.

- **Test sur les coefficients** : pour un $j \in \{1, \dots, p\}$ fixé, $H_0 : \beta_j = 0$.

- ▶ Statistique de test suit une **loi de Student**.
- ▶ Acceptation : le facteur explicatif $X^{(j)}$ est peu significatif.
- ▶ Rejet : le facteur $X^{(j)}$ a un poids non négligeable dans l'explication de Y .
- ▶ Pas très fiable dû aux corrélations entre les différents facteurs et donc coefficients. Dépend des variables présentes/absentes dans le modèle.

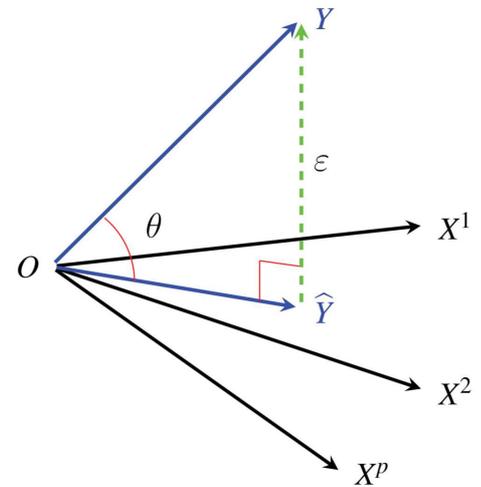
Qualité d'ajustement du modèle

• Coefficient de détermination R^2 :

- ▶ Défini de manière analogue au cas simple :

$$R^2 = \frac{\text{variance expliquée par le modèle}}{\text{variance totale}}$$

- ▶ **Interprétation géométrique** : $R^2 = \cos^2 \theta$. Plus l'angle entre \hat{Y} et Y faible, plus R^2 grand.
- ▶ Varie entre 0 et 1, d'autant plus grand que l'ajustement est bon.
- ▶ Augmente avec le nombre de facteurs, qu'ils soient pertinents ou non.
 - > Introduction d'un " R^2 ajusté" qui prend en compte le nombre de variables.
- ▶ R **coefficient de corrélation multiple**, correspond au coefficient de corrélation entre \hat{Y} et Y .



Qualité de prévision du modèle

• Critère PRESS de validation croisée :

$$\text{PRESS} = \sum_{i=1}^n (\text{prédiction de } y_i \text{ obtenue en oubliant l'observation } i - y_i)^2$$

- ▶ Validation croisée : estimation à partir d'une partie des données, puis mesure de l'erreur de prédiction sur les données laissées de côté.
- ▶ Somme des carrés de l'erreur de prédiction en "oubliant" une observation. "Estimation de l'erreur quadratique de prévision".
- ▶ **Plus le PRESS est petit, plus la qualité de prévision est importante.**

• Comment choisir entre plusieurs modèles ?

- ▶ R^2 proche de 1 \Rightarrow bonne prévision
 - ★ Beaucoup de variables \Rightarrow **bon ajustement** et **modèle précis** mais souvent **mauvais conditionnement**.
 - ★ Mauvais conditionnement : grandes variances pour les estimations des paramètres et pour les prévisions \Rightarrow **mauvaises prévisions**.
- ▶ Pour une **bonne prévision**, ne garder que les variables les plus pertinentes. Garder le modèle avec le **PRESS le plus faible**.

On enchaîne avec...

1 Modèle de régression linéaire multiple

- Généralisation du cas à un facteur
- Inférence statistique
- Qualité de la modélisation

2 Un exemple météorologique

Peut-on expliquer l'humidité de l'air ?

On dispose de **données météorologiques** (source : meteofrance.fr) recueillies le 16 juillet 2019 à midi dans des stations météorologiques françaises : pour chacune des 34 stations, on a relevé l'**humidité** de l'air, la **vitesse du vent**, la **température** de l'air, le **point de rosée** (i.e. la température de condensation), la **visibilité** et la **pression** atmosphérique.

On souhaite **expliquer l'humidité à l'aide des 5 autres variables.**

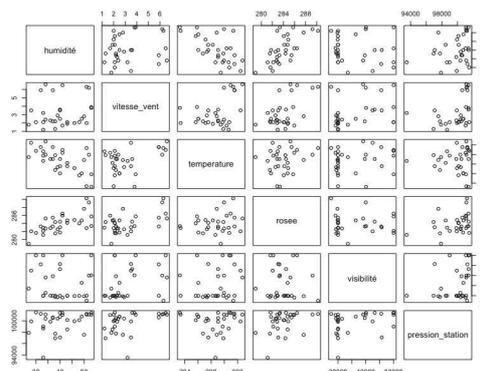
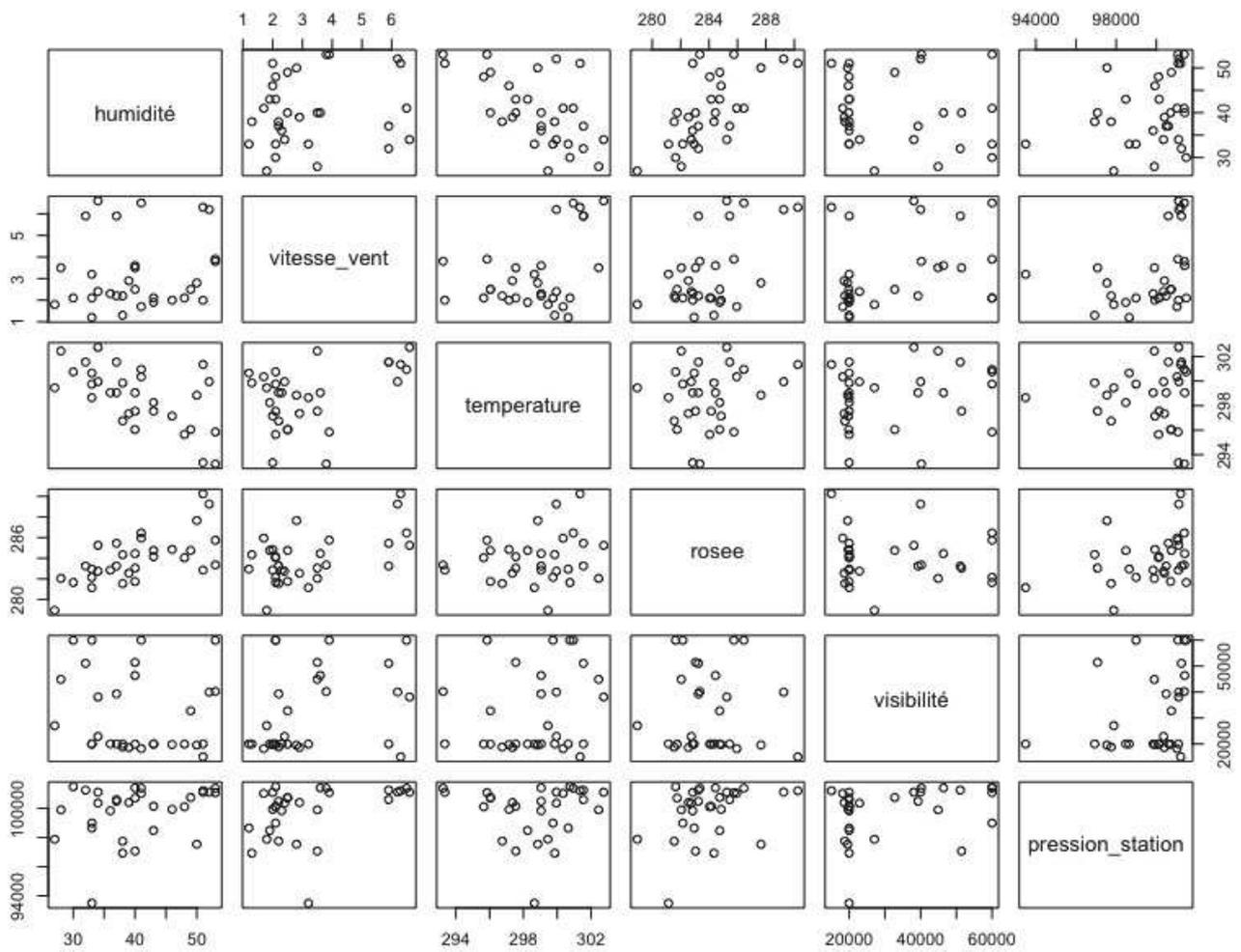


Figure – Matrice des nuages de points.



Régression linéaire multiple avec les 5 facteurs explicatifs

On vérifie la validité du modèle :

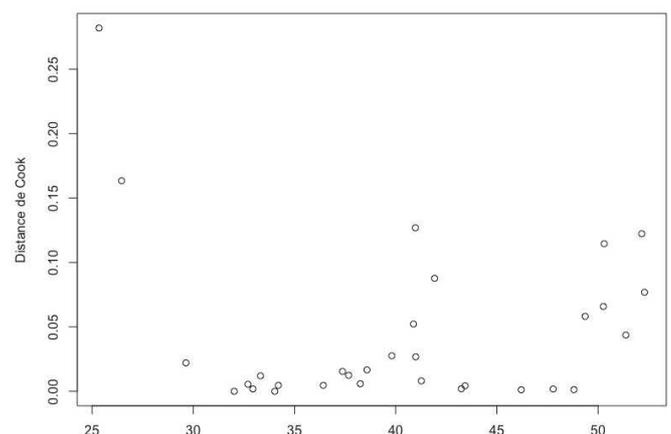
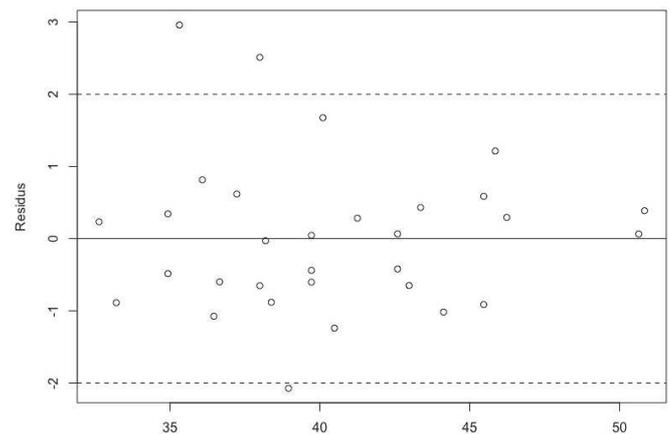
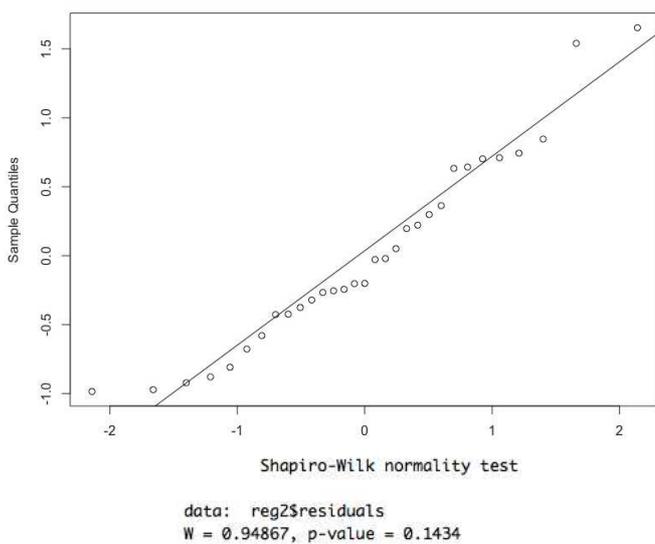


Figure – Droite de Henri (à gauche, en haut), test de Shapiro-Wilk (à gauche, en bas), graphe des résidus (à droite, en haut) et graphe des distances de Cook (à droite en bas) pour le modèle de régression linéaire multiple à 5 facteurs.

Analyse des résultats

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.645e+01  2.895e+01  0.914  0.369
vitesse_vent  7.272e-02  1.186e-01  0.613  0.545
rosee        2.594e+00  7.625e-02  34.021 <2e-16 ***
temperature  -2.427e+00  6.344e-02 -38.257 <2e-16 ***
visibilité   1.099e-05  1.040e-05  1.057  0.301
pression_station 1.599e-05  8.771e-05  0.182  0.857
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7583 on 25 degrees of freedom
Multiple R-squared:  0.9918,    Adjusted R-squared:  0.9902
F-statistic: 607.4 on 5 and 25 DF,  p-value: < 2.2e-16
```

- En **rouge**, les estimations pour chacun des paramètres β_0, \dots, β_5 .
- En **orange**, la statistique de test et la p -valeur du test de Fisher : on rejette l'hypothèse, le modèle est donc globalement significatif.
- En **vert**, les statistiques de test et p -valeurs des test de Student : pour les 5 variables explicatives considérées, seuls les paramètres associés à au point de rosée et à la température semblent significatifs (i.e. rejet de l'hypothèse " $\beta_i = 0$ ").
- En **bleu**, la valeur du coefficient de détermination R^2 : le modèle semble ici très bien ajusté (mais beaucoup de variables).

Un sous-modèle à deux facteurs

- *Idée : pour obtenir un modèle à **meilleure prévision**, garder uniquement les deux facteurs qui semblent significatifs : le **point de rosée** et la **température**.*
- **Validité du modèle** : on admet ici que les hypothèses sont satisfaites.
- **Analyse des résultats** :

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.35827  21.82354  0.612  0.545
rosee        2.61805   0.05980  43.784 <2e-16 ***
temperature  -2.39851   0.05796 -41.384 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7563 on 28 degrees of freedom
Multiple R-squared:  0.9909,    Adjusted R-squared:  0.9903
F-statistic: 1525 on 2 and 28 DF,  p-value: < 2.2e-16
```

- **Critère PRESS** : ici il vaut 0,69, contre 0,73 pour le modèle précédent.
- **Que peut-on conclure ?**

Comparaisons avec la régression linéaire simple

On considère les hypothèses de validité des deux modèles ci-dessous satisfaites. Analyser les résultats.

Était-il judicieux de choisir un modèle de régression linéaire multiple ?

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  612.620    139.207   4.401 0.000134 ***
temperature  -1.916     0.466  -4.111 0.000296 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.193 on 29 degrees of freedom
Multiple R-squared:  0.3682, Adjusted R-squared:  0.3464
F-statistic: 16.9 on 1 and 29 DF, p-value: 0.0002955

> press(reg1)
[1] 40.08924
```

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -569.4638    129.1586  -4.409 0.000131 ***
rosee        2.1473     0.4548   4.721 5.5e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.859 on 29 degrees of freedom
Multiple R-squared:  0.4346, Adjusted R-squared:  0.4151
F-statistic: 22.29 on 1 and 29 DF, p-value: 5.5e-05

> press(reg1)
[1] 35.34686
```

Figure – Modèles de régression linéaire simple expliquant l'humidité par la température (à gauche), et l'humidité par la rosée (à droite).

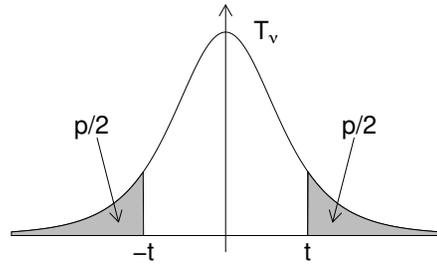
Loi de Student

Table de t en fonction du degré de liberté ν et de la probabilité p tels que :

$$\mathbb{P}(|T_\nu| > t) = p,$$

avec

$$T_\nu = \frac{U}{\sqrt{Y/\nu}} \quad \text{où } U \sim \mathcal{N}(0,1) \perp\!\!\!\perp Y \sim \chi^2(\nu).$$



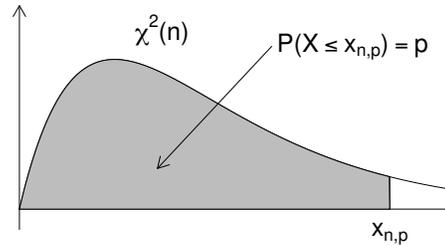
$\nu \backslash p$	0.9	0.7	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01
1	0.158	0.510	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.142	0.445	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.137	0.424	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.134	0.414	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.132	0.408	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.131	0.404	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.130	0.402	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.130	0.399	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.129	0.398	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.129	0.397	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.129	0.396	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.128	0.395	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.128	0.394	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.128	0.393	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.128	0.393	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.128	0.392	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.128	0.392	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.127	0.392	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.127	0.391	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.127	0.391	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.127	0.391	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.127	0.390	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.127	0.390	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.127	0.390	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.127	0.390	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.127	0.390	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.127	0.389	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.127	0.389	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.127	0.389	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.127	0.389	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
$+\infty$	0.12566	0.38532	0.67449	0.84162	1.03643	1.28155	1.64485	1.95996	2.32635	2.57583

La loi limite, lorsque ν tend vers l'infini, est une loi normale centrée réduite.

Loi du khi-deux

Table des quantiles de $X \sim \chi^2(n)$:

$$\mathbb{P}(X \leq x_{n,p}) = p.$$



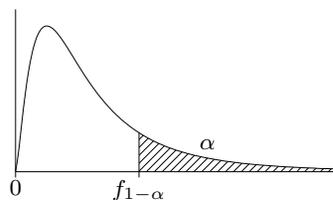
$n \backslash p$	0.005	0.01	0.025	0.5	0.1	0.25	0.5	0.75	0.9	0.95	0.975	0.99	0.995
1	0.00	0.00	0.00	0.45	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	1.39	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	2.37	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	3.36	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	4.35	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	5.35	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	6.35	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	7.34	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	8.34	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	9.34	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	10.34	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	11.34	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	12.34	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	13.34	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	14.34	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	15.34	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	16.34	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	17.34	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	18.34	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	19.34	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	20.34	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	21.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	22.34	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	23.34	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	24.34	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	25.34	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	26.34	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	27.34	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	28.34	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	29.34	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	39.34	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	49.33	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	59.33	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	69.33	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	79.33	64.28	71.14	79.33	88.13	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	89.33	73.29	80.62	89.33	98.65	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	99.33	82.36	90.13	99.33	109.1	118.5	124.3	129.6	135.8	140.1

Loi de Fisher

Si F est une variable aléatoire suivant la loi de Fisher–Snedecor à (ν_1, ν_2) degrés de liberté, la table donne la valeur $f_{1-\alpha}$ telle que

$$\mathbb{P}\{F \geq f_{1-\alpha}\} = \alpha = 0,05.$$

Ainsi, $f_{1-\alpha}$ est le quantile d'ordre $1 - \alpha = 0,95$ de la loi de Fisher–Snedecor à (ν_1, ν_2) degrés de liberté.



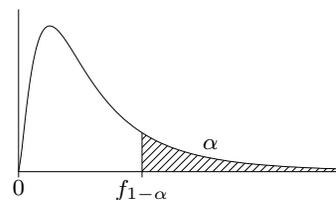
$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	8	10	15	20	30	∞
1	161	200	216	225	230	234	239	242	246	248	250	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,70	8,66	8,62	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,86	5,80	5,75	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,62	4,56	4,50	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	3,94	3,87	3,81	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64	3,51	3,44	3,38	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35	3,22	3,15	3,08	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,14	3,01	2,94	2,86	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,98	2,85	2,77	2,70	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,85	2,72	2,65	2,57	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,85	2,75	2,62	2,54	2,47	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,77	2,67	2,53	2,46	2,38	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,60	2,46	2,39	2,31	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,54	2,40	2,33	2,25	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,49	2,35	2,28	2,19	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,45	2,31	2,23	2,15	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,41	2,27	2,19	2,11	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,38	2,23	2,16	2,07	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,35	2,20	2,12	2,04	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,30	2,15	2,07	1,98	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,25	2,11	2,03	1,94	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,32	2,22	2,07	1,99	1,90	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,29	2,19	2,04	1,96	1,87	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,16	2,01	1,93	1,84	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,08	1,92	1,84	1,74	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	2,03	1,87	1,78	1,69	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,10	1,99	1,84	1,75	1,65	1,39
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,95	1,79	1,70	1,60	1,32
100	3,94	3,09	2,70	2,46	2,31	2,19	2,03	1,93	1,77	1,68	1,57	1,28
∞	3,84	3,00	2,60	2,37	2,21	2,10	1,94	1,83	1,67	1,57	1,46	1,00

Loi de Fisher (suite)

Si F est une variable aléatoire suivant la loi de Fisher–Snedecor à (ν_1, ν_2) degrés de liberté, la table donne la valeur $f_{1-\alpha}$ telle que

$$\mathbb{P}\{F \geq f_{1-\alpha}\} = \alpha = 0,025.$$

Ainsi, $f_{1-\alpha}$ est le quantile d'ordre $1 - \alpha = 0,975$ de la loi de Fisher–Snedecor à (ν_1, ν_2) degrés de liberté.



$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	8	10	15	20	30	∞
1	648	800	864	900	922	937	957	969	985	993	1001	1018
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,5	14,4	14,3	14,2	14,1	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	8,98	8,84	8,66	8,56	8,46	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,76	6,62	6,43	6,33	6,23	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,60	5,46	5,27	5,17	5,07	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,90	4,76	4,57	4,47	4,36	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,43	4,30	4,10	4,00	3,89	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,10	3,96	3,77	3,67	3,56	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,85	3,72	3,52	3,42	3,31	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,66	3,53	3,33	3,23	3,12	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,51	3,37	3,18	3,07	2,96	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,39	3,25	3,05	2,95	2,84	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,29	3,15	2,95	2,84	2,73	2,49
15	6,20	4,76	4,15	3,80	3,58	3,41	3,20	3,06	2,86	2,76	2,64	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,12	2,99	2,79	2,68	2,57	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,06	2,92	2,72	2,62	2,50	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,01	2,87	2,67	2,56	2,44	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	2,96	2,82	2,62	2,51	2,39	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	2,91	2,77	2,57	2,46	2,35	2,09
22	5,79	4,38	3,78	3,44	3,22	3,05	2,84	2,70	2,50	2,39	2,27	2,00
24	5,72	4,32	3,72	3,38	3,15	2,99	2,78	2,64	2,44	2,33	2,21	1,94
26	5,66	4,27	3,67	3,33	3,10	2,94	2,73	2,59	2,39	2,28	2,16	1,88
28	5,61	4,22	3,63	3,29	3,06	2,90	2,69	2,55	2,34	2,23	2,11	1,83
30	5,57	4,18	3,59	3,25	3,03	2,87	2,65	2,51	2,31	2,20	2,07	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,53	2,39	2,18	2,07	1,94	1,64
50	5,34	3,98	3,39	3,06	2,83	2,67	2,46	2,32	2,11	1,99	1,87	1,55
60	5,29	3,93	3,34	3,01	2,79	2,63	2,41	2,27	2,06	1,94	1,82	1,48
80	5,22	3,86	3,28	2,95	2,73	2,57	2,36	2,21	2,00	1,88	1,75	1,40
100	5,18	3,83	3,25	2,92	2,70	2,54	2,32	2,18	1,97	1,85	1,71	1,35
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,19	2,05	1,83	1,71	1,57	1,00

DESCRIPTION ET INFÉRENCE STATISTIQUES 3ICBE – I3BEBC11_02

Examen du vendredi 26 octobre 2018. Durée : 1h15.

Documents, calculatrices et téléphones portables sont interdits.

Ce sujet comporte 8 pages.

1. PREMIÈRE PARTIE : LA NOTATION DES HÔTELS DE LAS VEGAS SUR LE SITE TRIPADVISOR

Le site web TripAdvisor est un site de voyage permettant, entre autres, aux internautes de comparer, et réserver, les hôtels des destinations touristiques les plus prisées. Il est basé sur un système d'avis : après un séjour dans un des établissements référencés sur le site, l'utilisateur laisse une note entre 1 et 5.

Dans le cadre d'une étude statistique¹, 492 avis (ou *reviews*) laissés sur le site à propos de 21 hôtels de Las Vegas ont été analysés : dans chaque instance, outre la note laissée à l'établissement, ont également été recueillies des informations telles que le pays d'origine du client, le mois de la visite, le jour de la semaine où l'avis a été laissé, la présence d'une piscine, d'un spa (centre d'hydrothérapie), d'un cours de tennis, d'une salle de gym dans l'enceinte de l'établissement ou encore la possibilité d'un accès gratuit à internet.

Nous nous concentrerons ici sur les cinq variables suivantes :

- **Score** : la note, un entier entre 1 et 5, laissée par l'utilisateur ;
- **Pool** : la présence (YES) ou l'absence (NO) d'une piscine au sein de l'hôtel ;
- **Spa** : la présence (YES) ou l'absence (NO) d'un spa ;
- **Free.internet** : la présence (YES) ou l'absence (NO) d'un accès gratuit à internet ;
- **Review.weekday** : le jour de la semaine (de lundi à dimanche) où a été laissé l'avis sur le site.

Le tableau contenant les données est noté **V**.

A l'aide des documents de l'Annexe 1, répondez aux questions suivantes, en justifiant ce que vous affirmez.

1. Quelle est la taille de l'échantillon considéré ? Quelle est la nature de chacune des variables ?
2. Quelle est la valeur moyenne des avis laissés ? Leur variance ?
3. Comment appelle-t-on les graphiques représentés sur la figure 5 ? Comment peut-on les interpréter ?
4. Que peut-on dire de la figure 4 ?

¹que le lecteur intéressé pourra trouver dans l'article *Moro, S., Rita, P., and Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. Tourism Management Perspectives, 23, 41-52.*

5. Le choix du test effectué dans la figure 7 vous semble-t-il cohérent ? Déterminez votre réponse en présentant ce test et en vous aidant des différents éléments de l'Annexe 1.
6. Les observations graphiques sont-elles confortées par les résultats numériques du test ?
7. Que pouvez-vous en inférer quant à l'influence des paramètres sur la note attribuée aux hôtels sur TripAdvisor ?

2. DEUXIÈME PARTIE : L'ÉVALUATION... AU BASKETBALL

Au basketball, l'évaluation est une donnée statistique permettant d'obtenir une indication sur la performance d'un joueur lors d'un match. Il s'agit d'un nombre entier, positif ou négatif, obtenu selon une formule mathématique en prenant en compte les contributions numériques d'un joueur pendant la rencontre (le nombre de paniers marqués par exemple)².

Pour 100 joueurs du championnat de France Jeep Elite (anciennement Pro A), on a collecté lors la saison 2017-2018 les données de 16 paramètres statistiques à chaque rencontre. On a ensuite, pour chaque joueur, calculé la moyenne sur la saison de chacun de ces 16 paramètres³ : on dispose donc d'un tableau à 100 lignes et 16 colonnes, noté **BB**.

On s'intéresse aux corrélations pouvant exister entre l'évaluation et le temps de jeu moyen d'un joueur d'une part, et entre l'évaluation et les différents paramètres observés pendant une rencontre d'autre part.

On considère les variables suivantes :

- **minutes** : nombre moyen de minutes jouées par match ;
- **points** : nombre moyen de points marqués par match ;
- **tirs_mar** : nombre moyen de tirs marqués par match ;
- **tirs_tentes** : nombre moyen de tirs tentés par match ;
- **reussite** : pourcentage de réussite au tir, obtenu en faisant le ratio $\frac{tirs_mar}{tirs_tentes}$;
- **lf_reussis** : nombre de lancers francs réussis en moyenne par match ;
- **lf_tentes** : nombre de lancers francs tentés en moyenne par match ;
- **rebonds** : nombre moyen de rebonds gagnés par match ;
- **contres_p** : nombre moyen de contres réalisés par match ;
- **contres_c** : nombre moyen de contres subis par match ;
- **passes_dec** : nombre moyen de passes décisives réalisées par match ;
- **intercep** : nombre moyen d'interceptions par match ;
- **balles_per** : nombre moyen de balles perdues par match ;
- **fautes_com** : nombre moyen de fautes commises par match ;

²Il n'est évidemment pas nécessaire de connaître le basket pour pouvoir répondre correctement aux questions de l'énoncé !

³Ces données peuvent être trouvées sur le site de Ligue Nationale de Basket, www.lnb.fr

- **fautes_pro** : nombre moyen de fautes provoquées par match ;
- **eval** : évaluation moyenne par match.

Elles sont fournies dans cet énoncé davantage par souci d'exhaustivité que pour leur réelle importance dans ce qui suit.

A l'aide des documents de l'Annexe 2, répondez, en justifiant, aux questions suivantes.

8. Dans quelle fourchette se situe l'évaluation de la moitié des joueurs ?
9. Comment s'appelle le graphique représenté en figure 9 ? Comment peut-on l'interpréter ? Quelles semblent être les variables, parmi celles qui y figurent, les plus corrélées ?
10. On entre dans RStudio la ligne de code :
- $$\text{reg1}=\text{lm}(\text{minutes} \sim \text{eval}, \text{data} = \text{BB})$$
- A quoi est-ce que cela correspond ?
- Compte tenu des observations précédentes, cette modélisation paraît-elle raisonnable ?
11. Comment appelle-t-on la figure 11 ? Que permet-elle d'évaluer ? Un test permet de confirmer cette observation : donner son nom, l'hypothèse nulle H_0 et l'hypothèse alternative H_1 associées.
12. Quelles sont les trois hypothèses principales à vérifier avant de pouvoir traiter les résultats de reg1? Sont-elles vérifiées ici ? Justifier à l'aide de l'Annexe 2.
13. Que nous permettent de dire les figures 12 et 13 sur les valeurs atypiques et/ou influentes ?
14. Ecrire l'équation du modèle obtenu avec la ligne de code de la question 10. Interpréter les résultats obtenus. Que peut-on conclure quant à la significativité et l'ajustement du modèle ?
15. L'évaluation est en fait obtenue à l'aide d'une formule mathématique combinaison linéaire de certaines des 15 autres variables étudiées ici. On donne dans le tableau présenté ci-dessous en figure 1 les valeurs du coefficient R^2 et du *PRESS* pour trois modèles de régression linéaire multiple, pour lesquels les hypothèses classiques sont supposées satisfaites.

Modèle	R^2	PRESS
M1	0,999	0,0001
M2	0,97	0,53
M3	0,92	1,88

FIGURE 1

Quel est le modèle qui ajuste le mieux ? Et celui qui prévoit le mieux ? Pourquoi ?
 A votre avis, quel est le modèle qui correspond à la formule de l'évaluation ?

ANNEXE 1 : LA NOTATION DES HÔTELS DE LAS VEGAS

```

      Score      Pool      Spa      Free.internet      Review.weekday
Min.   :1.000    NO : 24    NO :108    NO : 24    Friday   :65
1st Qu.:4.000    YES:468   YES:384   YES:468   Monday   :73
Median :4.000
Mean   :4.112
3rd Qu.:5.000
Max.   :5.000

> sd(V$score)
[1] 1.014007

```

FIGURE 2

```

shapiro-wilk normality test

data:  V$score
w = 0.79867, p-value < 2.2e-16

```

FIGURE 3

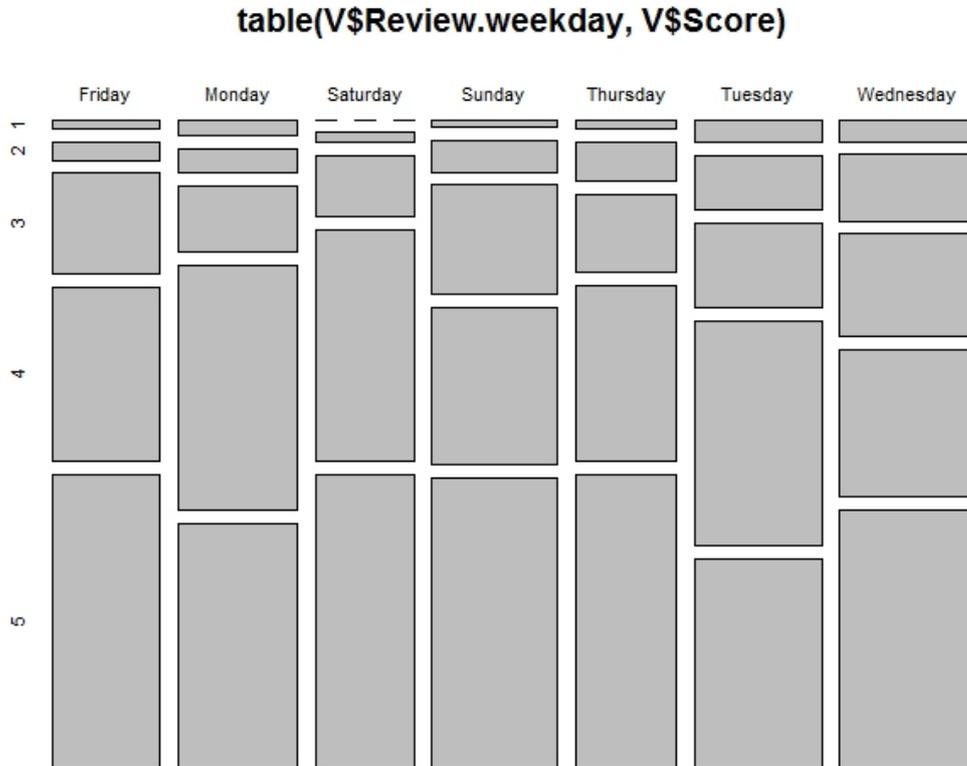


FIGURE 4

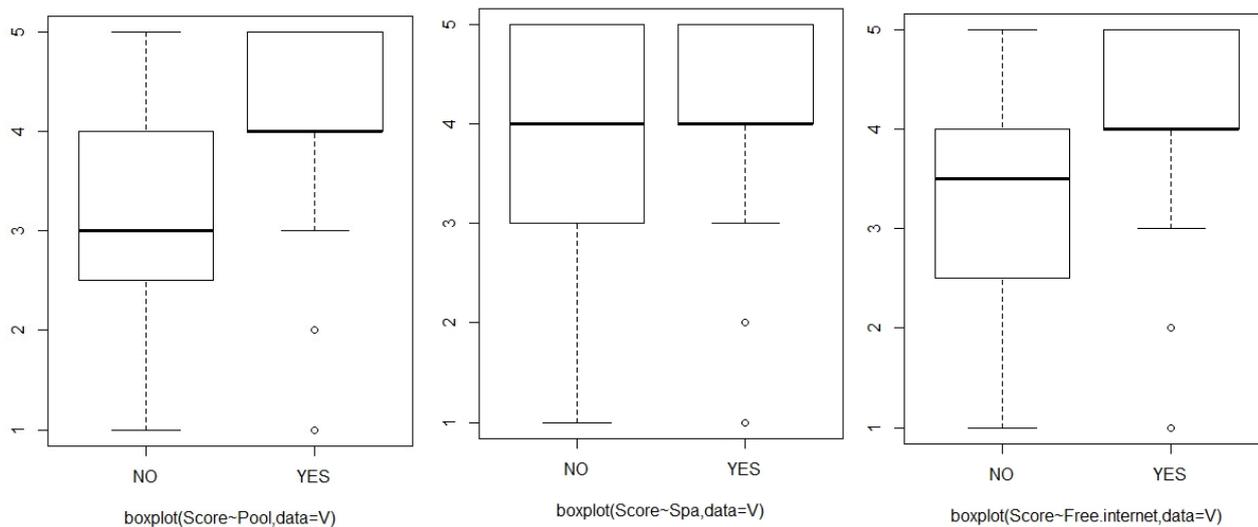
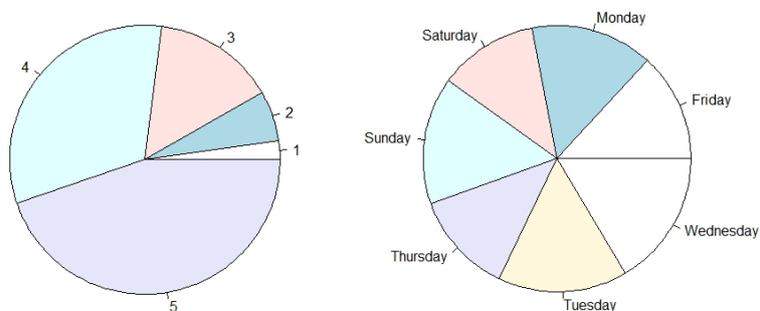


FIGURE 5

FIGURE 6. A gauche, la variable **Score**, à droite la variable **Review.weekday**

```
> kruskal.test(score~Spa,data=v)
      kruskal-wallis rank sum test

data:  score by Spa
Kruskal-wallis chi-squared = 2.7472, df = 1, p-value = 0.09742

> kruskal.test(score~Pool,data=v)
      kruskal-wallis rank sum test

data:  score by Pool
Kruskal-wallis chi-squared = 18.81, df = 1, p-value = 1.444e-05

> kruskal.test(score~Free.internet,data=v)
      kruskal-wallis rank sum test

data:  score by Free.internet
Kruskal-wallis chi-squared = 16.998, df = 1, p-value = 3.742e-05

> kruskal.test(score~Review.weekday,data=v)
      kruskal-wallis rank sum test

data:  score by Review.weekday
Kruskal-wallis chi-squared = 7.4771, df = 6, p-value = 0.279
```

FIGURE 7

ANNEXE 2 : L'ÉVALUATION AU BASKET

```

> summary(BB)
  minutes      points      tirs_mar      tirs_tentes      reussite      lf_reussis      lf_tentes      rebonds
Min.   : 1.50   Min.   : 0.000   Min.   :0.000   Min.   : 0.250   Min.   : 0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
1st Qu.:14.76   1st Qu.: 4.897   1st Qu.:1.673   1st Qu.: 3.397   1st Qu.: 42.67   1st Qu.:0.5975   1st Qu.:0.8425   1st Qu.:1.478
Median :19.90   Median : 7.040   Median :2.745   Median : 5.570   Median : 46.85   Median :1.1900   Median :1.6050   Median :2.705
Mean   :18.35   Mean   : 7.378   Mean   :2.718   Mean   : 5.689   Mean   : 47.27   Mean   :1.2336   Mean   :1.6740   Mean   :2.846
3rd Qu.:23.80   3rd Qu.:10.902   3rd Qu.:4.000   3rd Qu.: 8.023   3rd Qu.: 53.23   3rd Qu.:1.7725   3rd Qu.:2.4175   3rd Qu.:4.100
Max.   :32.21   Max.   :16.270   Max.   :5.940   Max.   :11.790   Max.   :100.00   Max.   :3.9700   Max.   :4.7000   Max.   :7.480

  contres_p      contres_c      passes_dec      intercep      balles_per      fautes_com      fautes_pro      eval
Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   : -0.250
1st Qu.:0.0000   1st Qu.:0.0875   1st Qu.:0.685   1st Qu.:0.3050   1st Qu.:0.665   1st Qu.:1.430   1st Qu.:1.090   1st Qu.: 5.013
Median :0.1200   Median :0.1800   Median :1.215   Median :0.5900   Median :1.130   Median :1.855   Median :1.585   Median : 8.345
Mean   :0.2321   Mean   :0.2036   Mean   :1.687   Mean   :0.5894   Mean   :1.104   Mean   :1.743   Mean   :1.755   Mean   : 8.217
3rd Qu.:0.4000   3rd Qu.:0.2900   3rd Qu.:2.400   3rd Qu.:0.7825   3rd Qu.:1.567   3rd Qu.:2.188   3rd Qu.:2.538   3rd Qu.:11.815
Max.   :1.7500   Max.   :0.8000   Max.   :7.150   Max.   :1.8700   Max.   :2.680   Max.   :3.030   Max.   :4.560   Max.   :19.130

> sapply(BB,sd)
  minutes      points      tirs_mar      tirs_tentes      reussite      lf_reussis      lf_tentes      rebonds      contres_p      contres_c      passes_dec
7.9032969   4.2123648   1.5660210   3.0375996   12.7373632   0.8647923   1.1133878   1.7402307   0.2969178   0.1652578   1.5125075
  intercep      balles_per      fautes_com      fautes_pro      eval
0.3811408   0.6357429   0.6876061   1.0631423   4.6408962
  
```

FIGURE 8

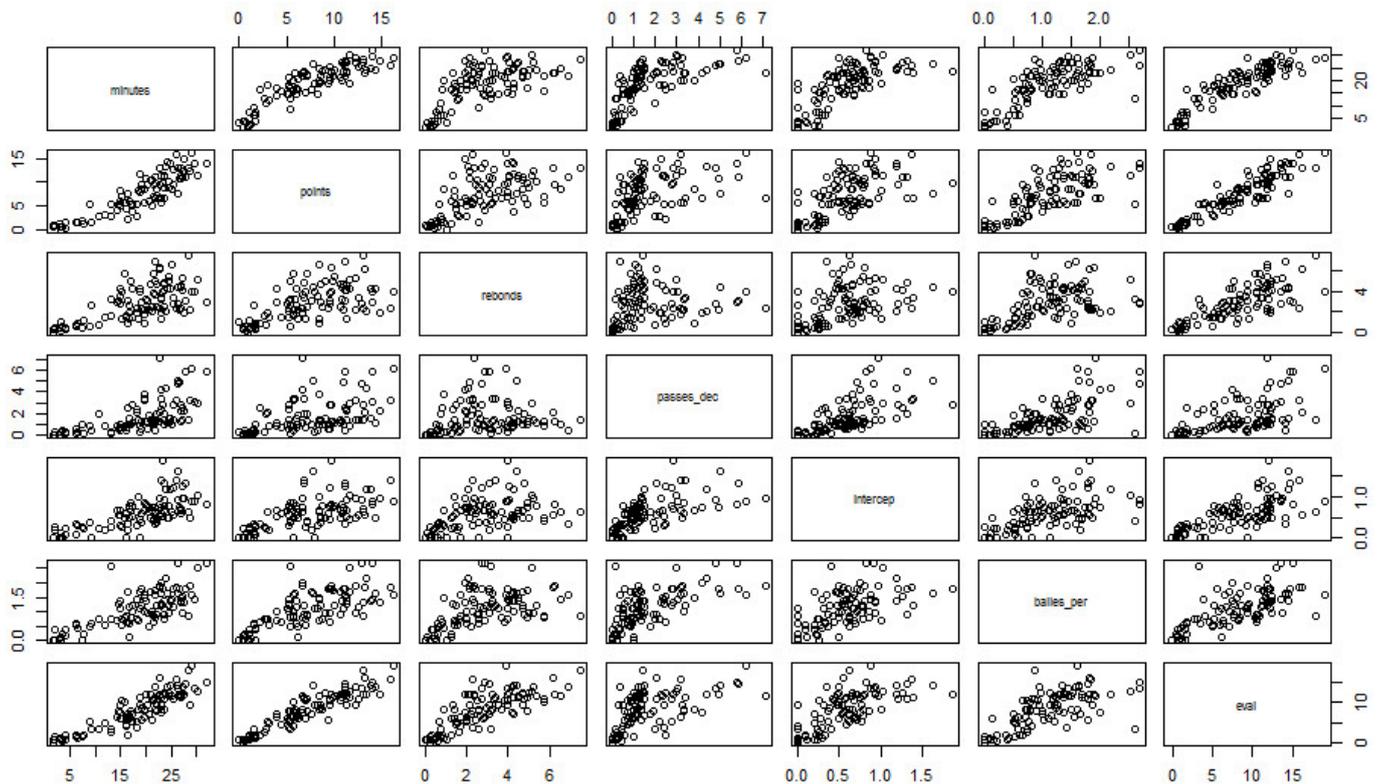


FIGURE 9. Variables (de g. à d.) : minutes, points, rebonds, passes_dec, intercep, balles_per, eval.

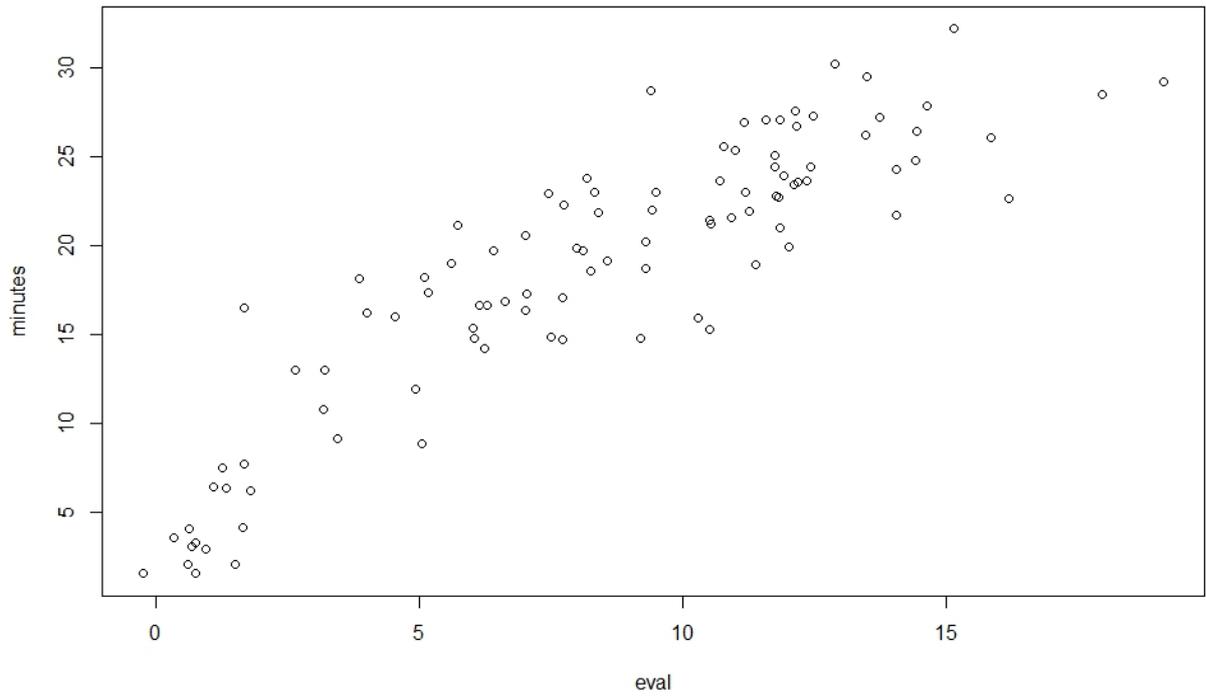


FIGURE 10

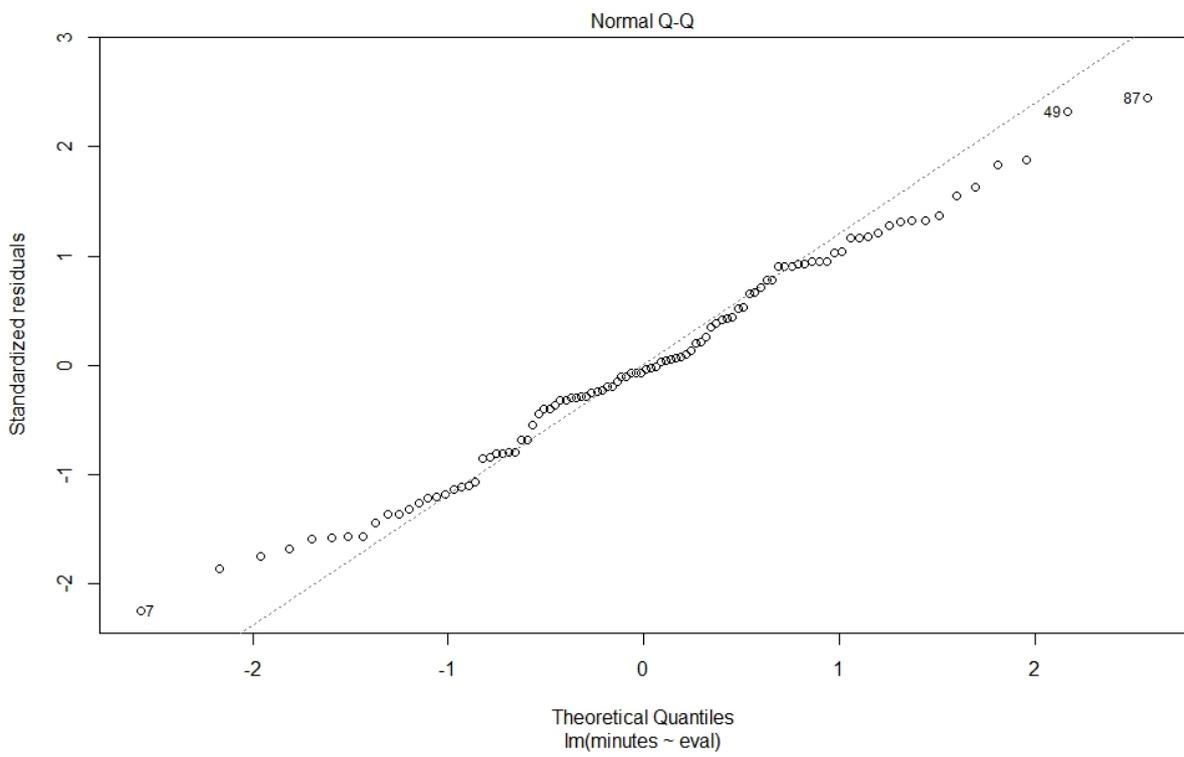


FIGURE 11

```

> shapiro.test(reg1$residuals)
      shapiro-wilk normality test

data:  reg1$residuals
w = 0.98744, p-value = 0.4677

> #Répartition des résidus
> res.student=rstudent(reg1)
> y_aj=reg1$fitted.values
> plot(res.student~y_aj,ylab="Residus")
> abline(h=c(-2,0,2),lty=c(2,1,2))

```

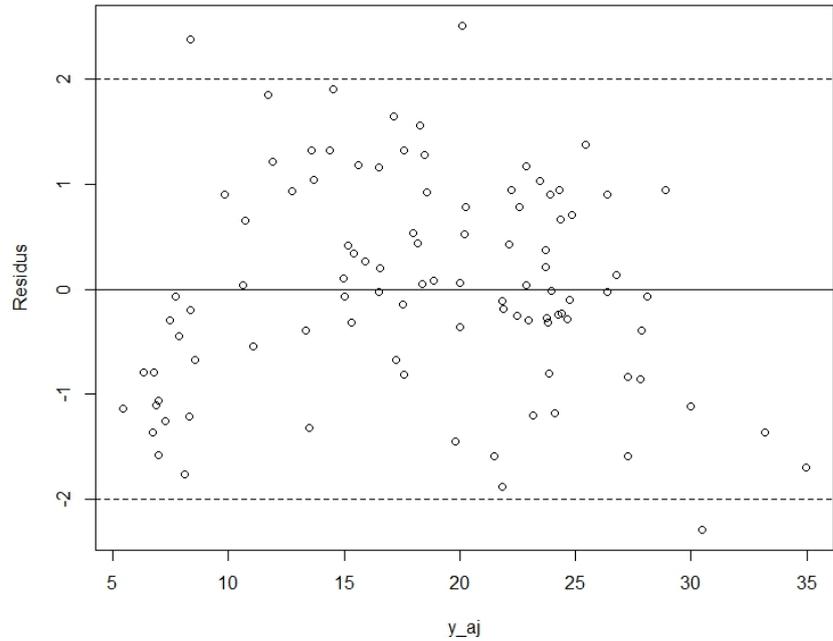


FIGURE 12

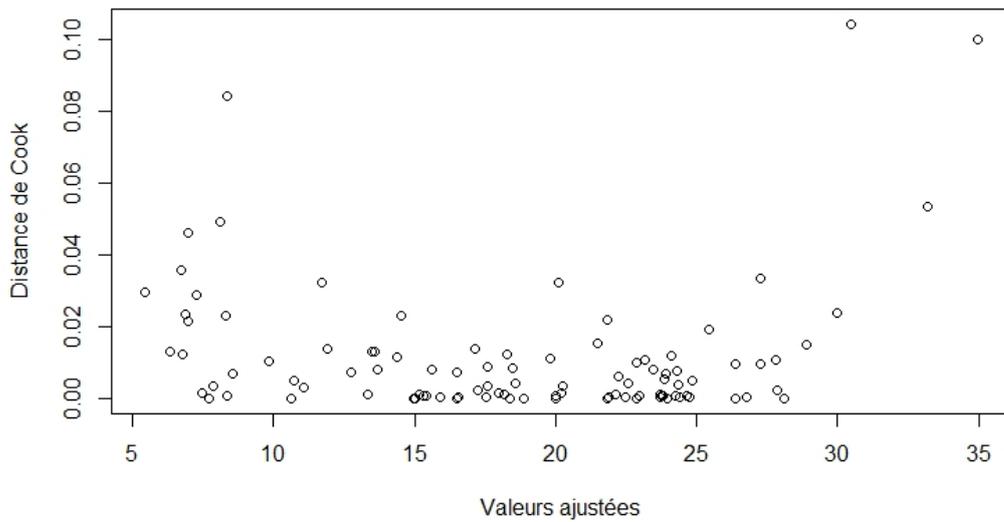


FIGURE 13

```

> summary(reg1)

Call:
lm(formula = minutes ~ eval, data = BB)

Residuals:
    Min       1Q   Median       3Q      Max
-7.8015 -2.7683 -0.1708  2.8635  8.6310

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.82491    0.72375   8.048  2e-12 ***
eval         1.52390    0.07679  19.846 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.546 on 98 degrees of freedom
Multiple R-squared:  0.8008,    Adjusted R-squared:  0.7987
F-statistic: 393.8 on 1 and 98 DF,  p-value: < 2.2e-16

```

FIGURE 14