Chapitre 2 : Statistique descriptive

- Terminologie
- Représentations graphiques
- Indicateurs statistiques

Statistique descriptive

La **statistique descriptive** a pour but de **résumer l'information** contenue dans les données de façon à en dégager les caractéristiques essentielles sous une forme simple et intelligible.

Les deux principaux outils de la statistique descriptive sont :

- les **représentations graphiques**
- les **indicateurs statistiques**

Terminologie

- **Données** : mesures faites sur des **individus** (ou unités statistiques) issus d'une **population**.
- **Variables** : particularités des individus.
- **Echantillon** : ensemble des individus.

Terminologie

- **Données** : mesures faites sur des **individus** (ou unités statistiques) issus d'une **population**.
- **Variables** : particularités des individus.
- **Echantillon** : ensemble des individus.

Exemple : si l'échantillon est un groupe de TD à l'Ensimag,

- un individu est un étudiant,
- la population peut être l'ensemble des étudiants de l'Ensimag, des élèves ingénieur de France, des habitants de Grenoble, etc...
- les variables étudiées peuvent être la taille, la filière choisie, la moyenne d'année, la couleur des yeux, la catégorie socio-professionnelle des parents,...

Recensement et sondages

Si l'échantillon est constitué de tous les individus de la population, on dit que l'on fait un **recensement**.

Quand l'échantillon n'est qu'une partie de la population, on parle de **sondage**.

Le principe des sondages est d'étendre à l'ensemble de la population les enseignements tirés de l'étude de l'échantillon.

Pour que cela ait un sens, il faut que l'échantillon soit représentatif de la population.

Statistique unidimensionnelle / multidimensionnelle

- **Statistique unidimensionnelle** : on ne mesure qu'**une seule variable** sur les individus, comme dans l'exemple des ampoules.

  Les données sont sous la forme de la série des valeurs prises par la variable pour les n individus, notées $x_1, ..., x_n$.

  On supposera que ces données sont les réalisations de n variables aléatoires $X_1, ..., X_n$ indépendantes et de même loi.

  On notera X une variable aléatoire de cette loi.

Statistique unidimensionnelle / multidimensionnelle

- **Statistique unidimensionnelle** : on ne mesure qu'**une seule variable** sur les individus, comme dans l'exemple des ampoules.

  Les données sont sous la forme de la série des valeurs prises par la variable pour les n individus, notées $x_1, ..., x_n$.

  On supposera que ces données sont les réalisations de n variables aléatoires $X_1, ..., X_n$ indépendantes et de même loi.

  On notera X une variable aléatoire de cette loi.

- **Statistique multidimensionnelle** : on mesure **plusieurs variables** sur les mêmes individus.

  $\Longrightarrow$ régression linéaire et cours de Statistical Analysis and Document Mining en 2A.

Variables discrètes / continues

Une variable statistique peut être **discrète** ou **continue**, **qualitative** ou **quantitative**. Les méthodes de représentation des données diffèrent suivant la nature des variables étudiées.

Une **variable discrète** est une variable à valeurs dans un ensemble fini ou dénombrable.

Les variables qui s'expriment par des nombres réels sont appelées **variables quantitatives** ou **numériques** (ex : longueur, durée, coût,...).

Les variables qui s'expriment par l'appartenance à une catégorie sont appelées **variables qualitatives** ou **catégorielles** (ex : couleur, catégorie socio-professionnelle, ...).

Variables discrètes qualitatives

Modalités : $E = \{e_1, \ldots, e_k\}$.

Fréquence absolue de la modalité e_j : nombre total n_j d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j :

$$n_j = \sum_{i=1}^{n} \mathbf{1}_{\{e_j\}}(x_i)$$

Fréquence relative de la modalité e_j : pourcentage n_j/n d'individus de l'échantillon pour lesquels la variable a pris la modalité e_j

couleur des yeux	bleu	vert	brun	pers	noir
fréquences absolues	66	34	80	15	5
fréquences relatives	33%	17%	40%	7.5%	2.5%

Table – couleur des yeux d'un échantillon de $n = 200$ personnes

Représentations graphiques pour des variables qualitatives

- **diagrammes en colonnes** ou **en bâtons** : à chaque modalité correspond un rectangle vertical dont la hauteur est proportionnelle à la fréquence relative de cette modalité.

- **diagrammes sectoriels** ou **camemberts** : à chaque modalité correspond un secteur de disque dont l'aire (ou l'angle au centre) est proportionnelle à la fréquence relative de cette modalité.

Listes	PC	LFI	Gen	EEES	EE	LREM	UDI	UDC	DLF	RN	PA	Autres
% Voix	2.5	6.3	3.3	6.2	13.5	22.4	2.5	8.5	3.5	23.3	2.2	5.8

Table – résultats des élections européennes de 2019

Listes	PC	LFI	Gen	EEES	EE	LREM	UDI	UDC	DLF	RN	PA	Autres
% Voix	2.5	6.3	3.3	6.2	13.5	22.4	2.5	8.5	3.5	23.3	2.2	5.8

Table – résultats des élections européennes de 2019

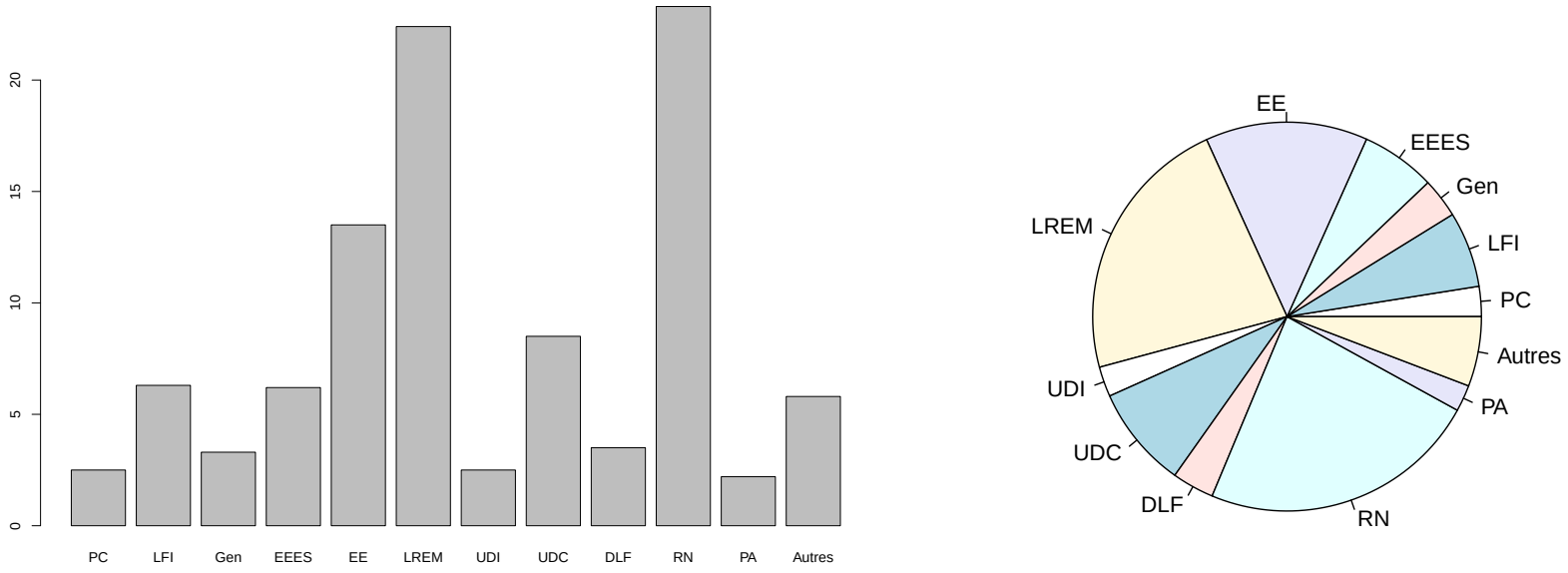

Figure – élections européennes, diagramme en colonnes et diagramme sectoriel

Variables discrètes quantitatives

Seule différence : ordre sur les modalités.

Nombre d'enfants	0	1	2	3	4	5	6	> 6
fréquence absolue	235	183	285	139	88	67	3	0
fréquence relative	23.5%	18.3%	28.5%	13.9%	8.8%	6.7%	0.3%	0

Table – nombre d'enfants de 1000 couples

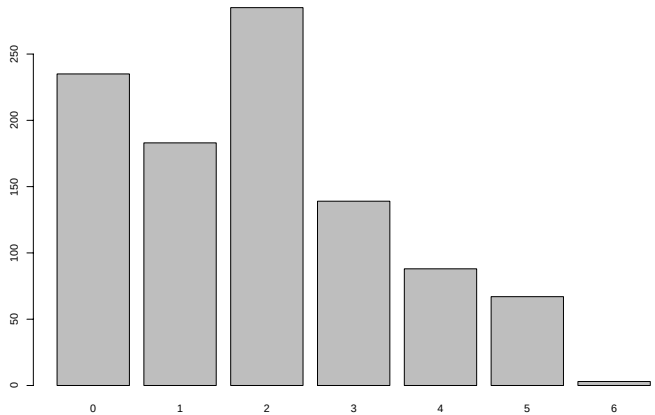

Figure – nombre d'enfants de 1000 couples, diagramme en bâtons

Variables continues

Une **variable continue** est à valeurs dans un ensemble non dénombrable comme $\mathbf{R}$ ou $[a, b]$.

Les représentations du type diagramme en bâtons sont sans intérêt, car les données sont en général toutes distinctes, donc les fréquences absolues sont toutes égales à 1. On va avoir besoin d'**ordonner les données**.

Variables continues

Une **variable continue** est à valeurs dans un ensemble non dénombrable comme $\mathbf{R}$ ou $[a, b]$.

Les représentations du type diagramme en bâtons sont sans intérêt, car les données sont en général toutes distinctes, donc les fréquences absolues sont toutes égales à 1. On va avoir besoin d'**ordonner les données**.

Exemple des durées de vie d'ampoules

- échantillon initial :

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
91.6	35.7	251.3	24.3	5.4	67.3	170.9	9.5	118.4	57.1

- échantillon ordonné :

x_1^*	x_2^*	x_3^*	x_4^*	x_5^*	x_6^*	x_7^*	x_8^*	x_9^*	x_{10}^*
5.4	9.5	24.3	35.7	57.1	67.3	91.6	118.4	170.9	251.3

Histogrammes

k classes $]a_{j-1}, a_j]$.

Largeur de la classe : $h_j = a_j - a_{j-1}$.

Effectif de la classe j : $n_j = \sum_{i=1}^{n} \mathbb{1}_{]a_{j-1},a_j]}(x_i)$.

Fréquence de la classe j : n_j/n.

L'**histogramme** est la figure constituée des rectangles dont les **bases** sont les classes et dont les **aires** sont égales aux fréquences de ces classes. Autrement dit, la hauteur du $j^{\text{ème}}$ rectangle est n_j/nh_j.

Histogrammes

k classes $]a_{j-1}, a_j]$.

Largeur de la classe : $h_j = a_j - a_{j-1}$.

Effectif de la classe j : $n_j = \sum_{i=1}^{n} \mathbb{1}_{]a_{j-1},a_j]}(x_i)$.

Fréquence de la classe j : n_j/n.

L'**histogramme** est la figure constituée des rectangles dont les **bases** sont les classes et dont les **aires** sont égales aux fréquences de ces classes.
Autrement dit, la hauteur du $j^{\text{ème}}$ rectangle est n_j/nh_j.

Règles conseillées :

- Entre 5 et 20 classes.
- *Règle de Sturges* : $k \approx 1 + \log_2 n = 1 + \ln n / \ln 2$
  $\Longrightarrow$ $k = 5$ pour $n \leq 22$, $k = 6$ pour $23 \leq n \leq 45$, etc...
- $a_0 = x_1^* - 0.025(x_n^* - x_1^*)$ et $a_k = x_n^* + 0.025(x_n^* - x_1^*)$.

Histogramme à classes de même largeur

classes $]a_{j-1}, a_j]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_j	4	3	1	1	1
fréquences n_j/n	40%	30%	10%	10%	10%
hauteurs n_j/nh	0.0077	0.0058	0.0019	0.0019	0.0019

Table – Ampoules, répartition en classes de même largeur

Histogramme à classes de même largeur

classes $]a_{j-1}, a_j]$	$]0, 52]$	$]52, 104]$	$]104, 156]$	$]156, 208]$	$]208, 260]$
effectifs n_j	4	3	1	1	1
fréquences n_j/n	40%	30%	10%	10%	10%
hauteurs n_j/nh	0.0077	0.0058	0.0019	0.0019	0.0019

Table – Ampoules, répartition en classes de même largeur

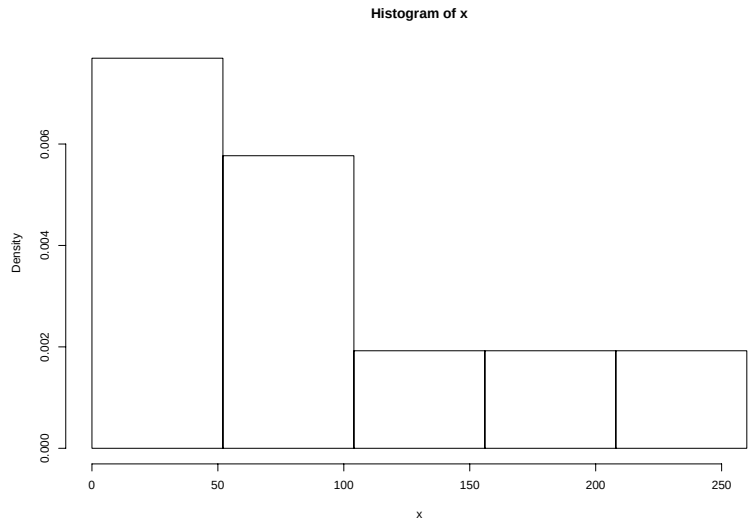

Figure – Ampoules, histogramme à classes de même largeur

Histogramme à classes de même effectif

classes $]a_{j-1}, a_j]$	$]0, 17]$	$]17, 46]$	$]46, 79]$	$]79, 145]$	$]145, 260]$
largeurs h_j	17	29	33	66	115
effectifs n_j	2	2	2	2	2
fréquences n_j/n	20%	20%	20%	20%	20%
hauteurs n_j/nh_j	0.0118	0.0069	0.0061	0.0030	0.0017

Table – Ampoules, répartition en classes de même effectif

Histogramme à classes de même effectif

classes $]a_{j-1}, a_j]$	$]0, 17]$	$]17, 46]$	$]46, 79]$	$]79, 145]$	$]145, 260]$
largeurs h_j	17	29	33	66	115
effectifs n_j	2	2	2	2	2
fréquences n_j/n	20%	20%	20%	20%	20%
hauteurs n_j/nh_j	0.0118	0.0069	0.0061	0.0030	0.0017

Table – Ampoules, répartition en classes de même effectif

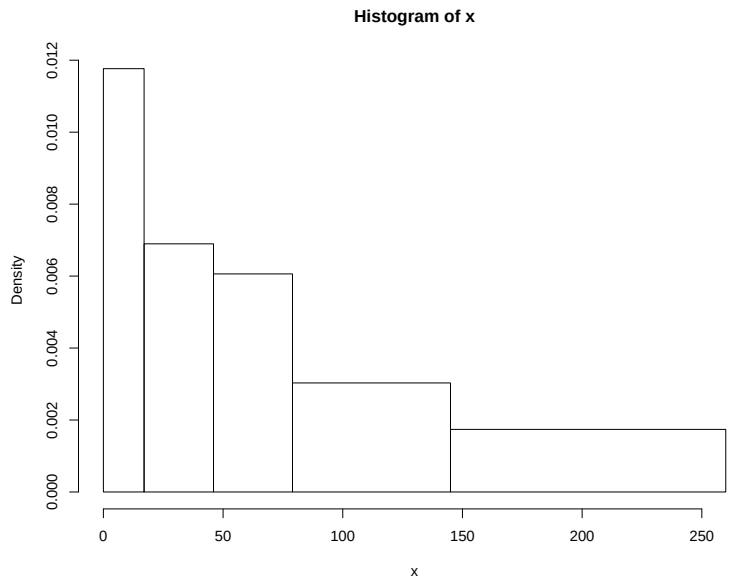

Figure – Ampoules, histogramme à classes de même effectif

Remarques générales sur les histogrammes

- Un histogramme fournit une estimation de la densité des observations.
- L'allure d'un histogramme permet de proposer des modèles probabilistes vraisemblables pour la loi des observations en comparant la forme de l'histogramme à celle des densités de lois de probabilité usuelles.
- Plusieurs histogrammes peuvent être dessinés à partir des mêmes données et avoir des allures assez différentes. On se contentera donc de dire qu'un histogramme donne l'allure générale de la densité des observations.
- Les histogrammes à classes de même effectif décrivent plus finement la distribution que les histogrammes à classes de même largeur. Mais leur usage est moins répandu car ils sont moins faciles à tracer.

Fonction de répartition empirique - 1/3

La **fonction de répartition empirique** (FdRE) F_n associée à un échantillon $x_1, \ldots, x_n$ est la fonction définie par :

$\forall x \in \mathbf{R}$, $F_n(x) =$ pourcentage d'observations inférieures à x

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i \leq x\}} = \begin{cases} 0 & \text{si } x < x_1^* \\ \dfrac{i}{n} & \text{si } x_i^* \leq x < x_{i+1}^* \\ 1 & \text{si } x \geq x_n^* \end{cases}$$

Fonction de répartition empirique - 2/3

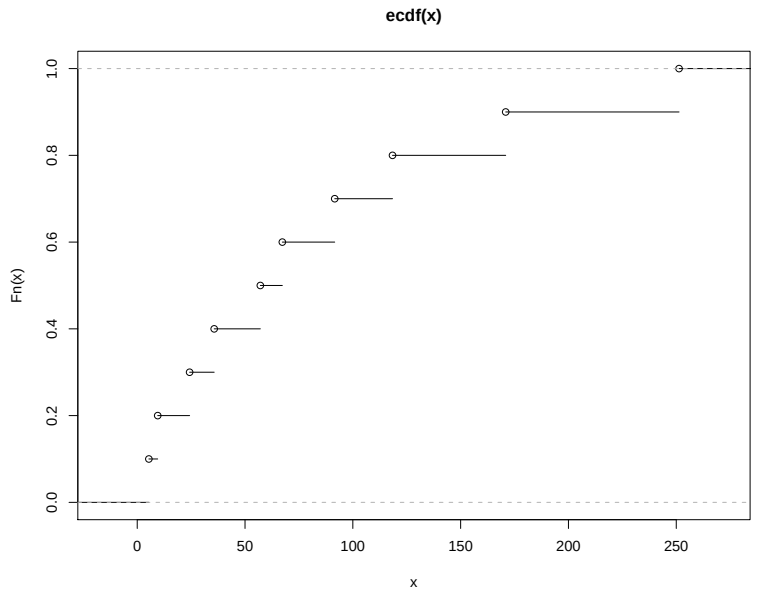

Figure – Ampoules, fonction de répartition empirique

Fonction de répartition empirique - 3/3

- La fonction de répartition de X, $F(x) = P(X \leq x)$, donne la probabilité qu'une observation soit inférieure à x.
- $F_n(x)$ est le pourcentage d'observations inférieures à x.

$\Longrightarrow$ $F_n(x)$ est une estimation de $F(x)$.

On peut montrer que cette estimation est d'excellente qualité, en un sens que l'on verra plus tard.

Choix de modèle à partir de la FdRE

- **Première idée** : tracer le graphe de la fonction de répartition empirique et déterminer si ce graphe ressemble à celui de la fonction de répartition d'une loi connue.

  Problème : les fonctions de répartition de toutes les lois de probabilité se ressemblent.

Choix de modèle à partir de la FdRE

- **Première idée** : tracer le graphe de la fonction de répartition empirique et déterminer si ce graphe ressemble à celui de la fonction de répartition d'une loi connue.

  Problème : les fonctions de répartition de toutes les lois de probabilité se ressemblent.

- **Seconde idée** : appliquer une transformation à la fonction de répartition empirique qui permette de reconnaître visuellement une caractéristique d'une loi de probabilité.

  Un **graphe de probabilités** est un nuage de points tracé à partir de la fonction de répartition empirique, tel que les points doivent être approximativement alignés si les observations proviennent d'une loi de probabilité bien précise.

Graphe de probabilités 1/2

On souhaite savoir si les observations sont issues de la loi de probabilité, dépendant d'un paramètre θ inconnu, dont la fonction de répartition est F.

Graphe de probabilités 1/2

On souhaite savoir si les observations sont issues de la loi de probabilité, dépendant d'un paramètre θ inconnu, dont la fonction de répartition est F.

Principe : chercher une relation affine du type $h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$, où h et g sont des fonctions qui ne dépendent pas de θ.

Graphe de probabilités 1/2

On souhaite savoir si les observations sont issues de la loi de probabilité, dépendant d'un paramètre θ inconnu, dont la fonction de répartition est F.

Principe : chercher une relation affine du type $h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$, où h et g sont des fonctions qui ne dépendent pas de θ.

Si la vraie fonction de répartition des observations est F, $h[F_n(x)]$ devrait être "proche" de $\alpha(\theta)g(x) + \beta(\theta)$, pour tout x.

Pour $x = x_i^*$, $h[F_n(x_i^*)] = h(i/n)$.

Graphe de probabilités 1/2

On souhaite savoir si les observations sont issues de la loi de probabilité, dépendant d'un paramètre θ inconnu, dont la fonction de répartition est F.

Principe : chercher une relation affine du type $h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$, où h et g sont des fonctions qui ne dépendent pas de θ.

Si la vraie fonction de répartition des observations est F, $h[F_n(x)]$ devrait être "proche" de $\alpha(\theta)g(x) + \beta(\theta)$, pour tout x.

Pour $x = x_i^*$, $h[F_n(x_i^*)] = h(i/n)$.

Donc, si la vraie fonction de répartition est F, les points $\big(g(x_i^*), h(i/n)\big)$ seront approximativement alignés.

Graphe de probabilités 1/2

On souhaite savoir si les observations sont issues de la loi de probabilité, dépendant d'un paramètre θ inconnu, dont la fonction de répartition est F.

Principe : chercher une relation affine du type $h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$, où h et g sont des fonctions qui ne dépendent pas de θ.

Si la vraie fonction de répartition des observations est F, $h[F_n(x)]$ devrait être "proche" de $\alpha(\theta)g(x) + \beta(\theta)$, pour tout x.

Pour $x = x_i^*$, $h[F_n(x_i^*)] = h(i/n)$.
Donc, si la vraie fonction de répartition est F, les points $\big(g(x_i^*), h(i/n)\big)$ seront approximativement alignés.

La pente et l'ordonnée à l'origine de cette droite fourniront des estimations de $\alpha(\theta)$ et $\beta(\theta)$, donc la plupart du temps de θ.

Graphe de probabilités 2/2

Soit F la fonction de répartition d'une loi de probabilité, dépendant d'un paramètre inconnu θ.

S'il existe des fonctions h, g, α et β telles que,

$$\forall x \in \mathbf{R},\ h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$$

alors le nuage des points

$$\big(g(x_i^*), h(i/n)\big),\ i \in \{1, \ldots, n\}$$

est appelé **graphe de probabilités** pour la loi de fonction de répartition F.

Si les points du nuage sont approximativement alignés, on admettra que F est une fonction de répartition plausible pour les observations.

Graphe de probabilités pour la loi exponentielle

Nuage des points $\left(x_i^*, \ln(1 - i/n)\right),\ i \in \{1, \ldots, n-1\}$

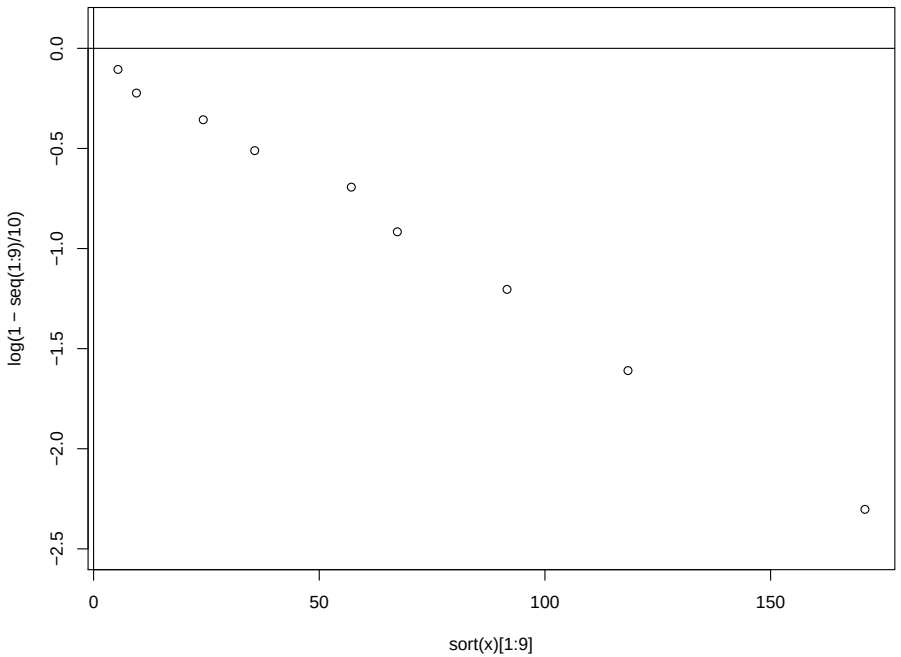

Figure – Ampoules, graphe de probabilités pour la loi exponentielle

Estimation de λ par la pente de la droite : 0.013.

Graphe de probabilités pour la loi normale

Nuage des points $\left(x_i^*, \phi^{-1}(i/n)\right),\ i \in \{1, \ldots, n-1\}$

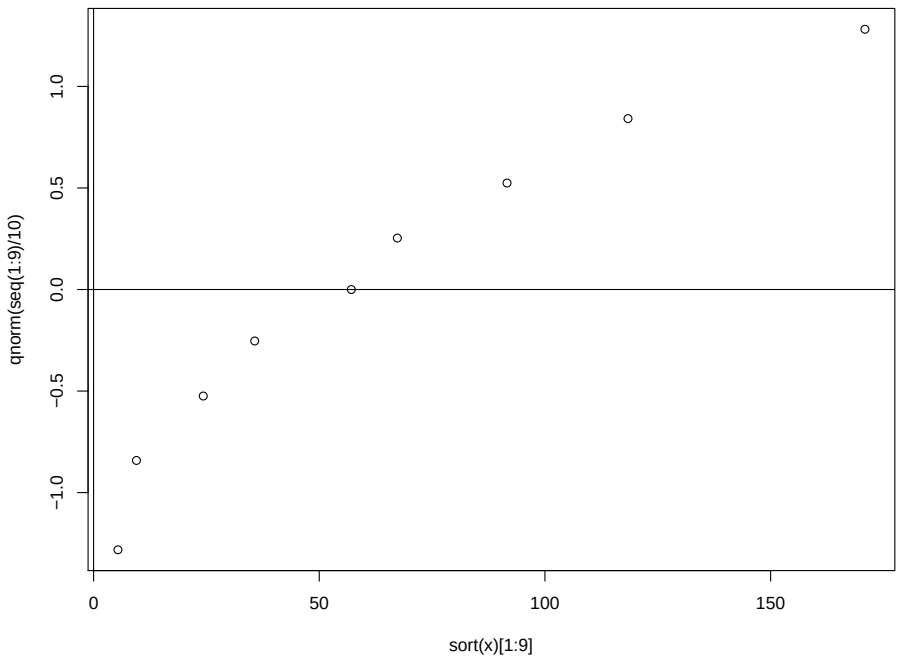

Figure – Ampoules, graphe de probabilités pour la loi normale

Graphe de probabilités, conclusion

Sur l'exemple des ampoules, les graphes de probabilités amènent à retenir le modèle de loi exponentielle et à rejeter le modèle de loi normale.

Principal défaut de la méthode : comment juger visuellement si des points sont "suffisamment alignés" ?
La réponse est soumise à la subjectivité de l'utilisateur.
Il est donc nécessaire de compléter cette approche graphique par des techniques objectives : les **tests d'adéquation**.

Néanmoins, les graphes de probabilités sont une première étape indispensable dans une étude statistique :

- Ils sont faciles à mettre en oeuvre.
- Ils permettent de détecter facilement des modèles clairement pas adaptés aux données.

Indicateurs statistiques

Représentations graphiques : ne permettent qu'une analyse visuelle de la répartition des données.

Indicateurs statistiques : indicateurs numériques permettant de caractériser au mieux des données quantitatives.

Deux familles d'indicateurs :

- Indicateurs de localisation
- Indicateurs de dispersion.

Indicateurs de localisation - 1/3

But : donner un **ordre de grandeur général** des observations, un nombre unique qui résume au mieux les données.

Indicateurs de localisation - 1/3

But : donner un **ordre de grandeur général** des observations, un nombre unique qui résume au mieux les données.

- La **moyenne empirique** : $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$. Commande `R` : `mean(x)`.

Indicateurs de localisation - 1/3

But : donner un **ordre de grandeur général** des observations, un nombre unique qui résume au mieux les données.

- La **moyenne empirique** : $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Commande `R` : `mean(x)`.

  Exemple des ampoules : $\bar{x}_{10}$= 83.15.

  Si on admet comme modèle la loi exponentielle, on sait que l'espérance de la loi exp(λ) est $1/\lambda$.

  Loi des grands nombres : la moyenne empirique converge vers l'espérance de la loi.

  Il est donc logique de considérer qu'une estimation de λ est $1/\bar{x}_{10}$= 0.012.

  Valeur cohérente avec la valeur trouvée à l'aide du graphe de probabilités, 0.013.

Indicateurs de localisation - 2/3

- Les **valeurs extrêmes** : $x_1^* = \min x_i$ et $x_n^* = \max x_i$.

  $(x_1^* + x_n^*)/2$ est un indicateur de localisation.

  Inconvénient de la moyenne et des valeurs extrêmes : sensibilité aux valeurs aberrantes.

Indicateurs de localisation - 2/3

- Les **valeurs extrêmes** : $x_1^* = \min x_i$ et $x_n^* = \max x_i$.
  $(x_1^* + x_n^*)/2$ est un indicateur de localisation.

  Inconvénient de la moyenne et des valeurs extrêmes : sensibilité aux valeurs aberrantes.

- La **médiane empirique** : $\tilde{x}_n$ ou $\tilde{q}_{n,1/2}$, est un réel qui partage l'échantillon ordonné en deux parties de même effectif. La moitié des observations sont inférieures à $\tilde{x}_n$ et l'autre moitié lui sont supérieures.

$$\tilde{x}_n = \tilde{q}_{n,1/2} = \begin{cases} \frac{1}{2}\left(x^*_{\frac{n}{2}} + x^*_{\frac{n}{2}+1}\right) & \text{si } n \text{ est pair} \\ x^*_{\frac{n+1}{2}} & \text{sinon} \end{cases}$$

Indicateurs de localisation - 3/3

La médiane empirique est insensible aux valeurs aberrantes.

- 1 3 5 8 10 1 3 5 8 10000

  $\tilde{x}_5 = x^*_3 = 5$ pour les deux échantillons, alors que $\bar{x}_5$ vaut 5.4 pour le premier échantillon et 2003.4 pour le deuxième.

- *Ampoules* : $\tilde{x}_{10} = (57.1 + 67.3)/2 = 62.2$ alors que $\bar{x}_{10} = 83.1$.

- *Salaires nets mensuels des français en 2020* :
  - salaire médian : 2005 € salaire moyen : 2520 €

Conclusion : la moyenne et la médiane empiriques sont deux résumés de l'échantillon dont la connaissance simultanée peut être riche d'enseignements.

Quand la distribution est symétrique, moyenne et médiane empiriques sont proches.

Indicateurs de dispersion - 1/3

But : mesurer la **variabilité** des observations.

	J	F	M	A	M	J	J	A	S	O	N	D
New-York	0	1	5	12	17	22	25	24	20	14	8	2
San Francisco	9	11	12	13	14	16	17	17	18	16	13	9

Table – températures mensuelles moyennes à New-York et à San Francisco

	$\bar{x}_n$	$\tilde{x}_n$
t° New-York	12.5	13.0
t° San Francisco	13.7	13.5

Indicateurs de dispersion - 2/3

- **Variance empirique** : $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$. Elle mesure l'écart quadratique moyen de l'échantillon à sa moyenne.

  En `R`, `var(x)` donne $s'^2_n = \frac{n}{n-1}s_n^2$ au lieu de s_n^2. Explications plus tard.

Indicateurs de dispersion - 2/3

- **Variance empirique** : $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$. Elle mesure l'écart quadratique moyen de l'échantillon à sa moyenne.

  En `R`, `var(x)` donne $s'^2_n = \frac{n}{n-1}s_n^2$ au lieu de s_n^2. Explications plus tard.

- **Ecart-type empirique** : $s_n = \sqrt{s_n^2}$. *Intérêt* : il s'exprime dans la même unité que les données.

Indicateurs de dispersion - 2/3

- **Variance empirique** : $s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$. Elle mesure l'écart quadratique moyen de l'échantillon à sa moyenne.

  En `R`, `var(x)` donne $s'^2_n = \frac{n}{n-1}s_n^2$ au lieu de s_n^2. Explications plus tard.

- **Ecart-type empirique** : $s_n = \sqrt{s_n^2}$. *Intérêt* : il s'exprime dans la même unité que les données.

- **Coefficient de variation empirique** : $cv_n = \frac{s_n}{\bar{x}_n}$.

  Intérêt : indicateur sans dimension. On considère que l'échantillon possède une variabilité significative si $cv_n > 0.15$. Si $cv_n \leq 0.15$, les données présentent peu de variabilité et on considère que la moyenne empirique à elle seule est un bon résumé de tout l'échantillon.

Indicateurs de dispersion - 3/3

	J	F	M	A	M	J	J	A	S	O	N	D
New-York	0	1	5	12	17	22	25	24	20	14	8	2
San Francisco	9	11	12	13	14	16	17	17	18	16	13	9

Table – températures mensuelles moyennes à New-York et à San Francisco

	$\bar{x}_n$	$\tilde{x}_n$	s_n^2	s_n	cv_n
ampoules	83.15	62.2	5540.2	74.4	0.89
t° New-York	12.5	13.0	77.7	8.8	0.70
t° San Francisco	13.7	13.5	8.9	3.0	0.22

Quantiles empiriques

Les **quantiles empiriques** sont des valeurs qui partagent l'échantillon ordonné en un certain nombre de parties de même effectif.

- 2 parties : médiane empirique $\tilde{x}_n$.
- 4 parties : **quartiles**, $\tilde{q}_{n,1/4}$, $\tilde{q}_{n,1/2}$ $(= \tilde{x}_n)$ et $\tilde{q}_{n,3/4}$.
- 10 parties : **déciles**, $\tilde{q}_{n,1/10}, \ldots, \tilde{q}_{n,9/10}$.
- 100 parties : **centiles**, $\tilde{q}_{n,1/100}, \ldots, \tilde{q}_{n,99/100}$.

Quantiles empiriques

Les **quantiles empiriques** sont des valeurs qui partagent l'échantillon ordonné en un certain nombre de parties de même effectif.

- 2 parties : médiane empirique $\tilde{x}_n$.
- 4 parties : **quartiles**, $\tilde{q}_{n,1/4}$, $\tilde{q}_{n,1/2}$ $(= \tilde{x}_n)$ et $\tilde{q}_{n,3/4}$.
- 10 parties : **déciles**, $\tilde{q}_{n,1/10}, \ldots, \tilde{q}_{n,9/10}$.
- 100 parties : **centiles**, $\tilde{q}_{n,1/100}, \ldots, \tilde{q}_{n,99/100}$.

Quantiles empiriques de l'échantillon $x_1, \ldots, x_n$:

$$\forall p \in]0,1[\ ,\ \tilde{q}_{n,p} = \begin{cases} \frac{1}{2}\left(x^*_{np} + x^*_{np+1}\right) & \text{si } np \text{ est entier} \\ x^*_{\lfloor np \rfloor + 1} & \text{sinon} \end{cases}$$

Ampoules : $\tilde{q}_{n,1/4} = x^*_3 = 24.3, \tilde{q}_{n,1/2} = \tilde{x}_n = 62.2, \tilde{q}_{n,3/4} = x^*_8 = 118.4$

Remarque finale

Puisqu'on considère les observations $x_1, \ldots, x_n$ comme des réalisations de variables aléatoires $X_1, \ldots, X_n$, toutes les quantités définies dans ce chapitre sont elles-mêmes des réalisations de variables aléatoires :

$$\mathbf{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}} \qquad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$\widetilde{Q}_{n,p} = \begin{cases} \frac{1}{2}(X^*_{np} + X^*_{np+1}) & \text{si } np \text{ est entier} \\ X^*_{\lfloor np \rfloor + 1} & \text{sinon} \end{cases}$$