Statistiques descriptives

LMAFY1101

Anouar El Ghouch

LSBA, UCLouvain

Plan

Vocabulaire de la méthodologie statistique

La statistique est une discipline scientifique qui concerne les aspects suivants:

- **Recueillir** et organiser des données.
- Les **décrire** et les résumer à l'aide, par exemple, de graphiques.
- **Les analyser avec le but d'en extraire le plus possible d'informations pertinentes**.
- **Interpréter** et communiquer les résultats de cette analyse et développer des outils d'aider à la décision.

Les données étudiées peuvent être de **toute nature**, ce qui rend la statistique utile dans **tous les champs disciplinaires** et explique pourquoi elle est enseignée dans toutes les filières universitaires, de l'économie à la biologie en passant par la psychologie et les sciences mathématiques.

La statistique dans divers domaines

Donnons quelques exemples d'utilisation de la statistique dans divers domaines.

- Biologie, médecine : essais thérapeutiques, épidémiologie, dynamique des populations, analyse du génome, ...
- Sciences de l'ingénieur : contrôle de qualité, sureté de fonctionnement (fiabilité, disponibilité, sécurité,...), maîtrise des risques industriels, ...
- Physique, sciences de la terre : prévisions météorologiques, exploration pétrolière, mécanique statistique, théorie cinétique des gaz, ...
- Sciences humaines : enquêtes d'opinion, sondages, études de populations, ...

Si on omet la collecte et l'organisation des données, les méthodes statistiques se répartissent, principalement, en deux classes :

Statistique descriptive

La statistique descriptive a pour but de **résumer l'information contenue dans les données** de façon synthétique et efficace. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux de synthèse et d'indicateurs numériques. Elle permet de dégager les caractéristiques essentielles et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée.

Statistique inférentielle

La statistique inférentielle va au-delà de la simple description des données. Elle a pour but de **tirer des conclusions** et/ou faire des prévisions **concernant** le phénomène (**la population**) étudié **à partir des observations** (l'échantillon) récoltées. Il faut pour cela proposer des modèles probabilistes et savoir évaluer et gérer les risques d'erreurs.

Exemple: Ampoule

Voici les durées de vie (en jours) de 10 ampoules toutes fabriquées et exploitées à l'identique.

```
91.6  89.7  201.3  94.3  95.4  87.3  170.9  159.5  118.4  107.1
```

• On peut constater que ces mesures varient "considérablement". S'agit'il de variations qui sont dues purement au hasard ? Ou, au contraire, ce sont là des variations "pas normales". Des variations dues, par exemple, à des erreurs de mesures ou dues au fait qu'on a oublié de contrôler un (ou des) facteur(s) qui a une influence significative sur la durée de vie ?

• Comment peut-on quantifier la variation observée ? Aussi, y a-t-il des valeurs atypiques parmi ces mesures ? Comment les repérer facilement ? Pour répondre à ces questions, il suffit de faire une analyse descriptive des données.

• Au vu de ces observations, est-il raisonnable de supposer que la durée de vie moyenne d'une ampoule quelconque est inférieure à une valeur fixe, disons 95 ? Quelle est la probabilité pour qu'une ampoule tombe en panne avant une date t. Il s'agit ici d'un problème d'inférence.

Les données représentent la matière première sans laquelle aucune étude statistique n'est possible.

On obtient classiquement des données à partir de deux sources distinctes : les études observationnelles et les études expérimentales:

Étude observationnelle

Dans une étude observationnelle, **on observe** et on mesure des caractéristiques spécifiques **sans intervention** qui a pour but d'influencer ou modifier les sujets de l'étude.

Exemple: Sondage d'opinion ou enquête. Données récoltées dans la nature ou sur internet.

Étude expérimentale

Dans une étude expérimentale (ou interventionnelle), les individus sont exposés à un ou plusieurs **facteurs fixés et contrôlés par l'expérimentateur** et on passe ensuite à l'observation de **leur effets** sur les sujets de l'étude.

Exemple: Essai dans un laboratoire de recherche (clinique, physique ou chimique).

La première étape d'une étude statistique consiste à bien définir la **population** visée par l'étude, c.-à-d. l'ensemble d'objets ou d'individus sur lesquels l'information est recherchée.

Puis à planifier la récolte des données dans le but d'obtenir un **échantillon**, c.-à-d. un sous ensemble, représentatif de la population.

Exemples de population

- La collection de tous les individus souffrant d'une certaine maladie (population tangible).
- La collection de toutes les mesures qui pourraient être faites dans certaines conditions expérimentales (population conceptuelle ou hypothétique).

L'échantillon doit être collecté, de façon appropriée, par **tirage aléatoire**.
L'échantillonnage aléatoire donne à tous les individus une **chance égale d'être choisi**, éliminant (ou réduisant) ainsi le risque de biais de sélection (sous ou sur-représentativité de certaines catégories ou variétés).

Type de données

Lors de la création d'un plan de récolte de données, il est aussi nécessaire de déterminer le ou les caractères qu'on souhaite observer ou mesurer.

On distingue deux types de caractère: quantitatif et qualitatif.

Caractère quantitatif (appelé aussi variable)

Les valeurs possibles d'un caractère quantitatif sont des nombres. Il y a deux sous-types:

- **Discret**. C'est une variable à valeurs dans un ensemble fini ou dénombrable.
  Exemple: combien de frères avez-vous? Les valeurs possibles sont 0,1,2,...
- **Continu**. Les valeurs possibles peuvent être tout nombre réel dans un intervalle donné.
  Exemple: hauteur, durée de vie, vitesse.

Caractère qualitatif ou catégoriel.

Les valeurs possibles d'un caractère qualitatif ne sont pas des nombres. Ces variables expriment l'appartenance à une catégorie.
Exemple: couleur, sexe, catégorie socio-professionnelle, groupe sanguin.

Certains caractères catégoriels viennent dans un ordre naturel et sont donc appelés caractère ou variable **ordinales**.
Exemple: qualité de l'eau (mauvaise, moyenne, bonne, excellente).

Remarques

- Il faut distinguer le type d'un caractère de l'encodage utilisé pour l'enregistrer. Par exemple le sexe est un caractère catégoriel, mais il peut être encodé sous forme d'une variable quantitative (1 pour femme et 2 pour homme, par exemple).
- L'encodage d'un caractère ne change en rien sa nature. Faire des calculs arithmétiques sur un caractère catégoriel (encodé sous forme numérique) n'a évidemment aucun sens.

Données catégorielles

À partir d'ici, pour mieux comprendre ce qui suit, vous êtes invité à consulter la documentation sur R intitulée Exploration de données avec R.

Commençons par un exemple simple: voici les notes d'un test (en anglais) dans une petite classe de 10 élèves: A, D, C, D, C, C, C, C, F et B.

• Commençons par introduire ces données en R

```
> Grades <- c("A", "D", "C", "D", "C", "C", "C", "C", "F", "B") |> factor()
```

• Puis calculons le tableau des **effectifs** (ou fréquences absolues). Pour cela, nous pouvons utiliser les fonctions *table* ou *xtabs*

```
> table(Grades)

Grades
A B C D F
1 1 5 2 1
```

```
> xtabs(~Grades)

Grades
A B C D F
1 1 5 2 1
```

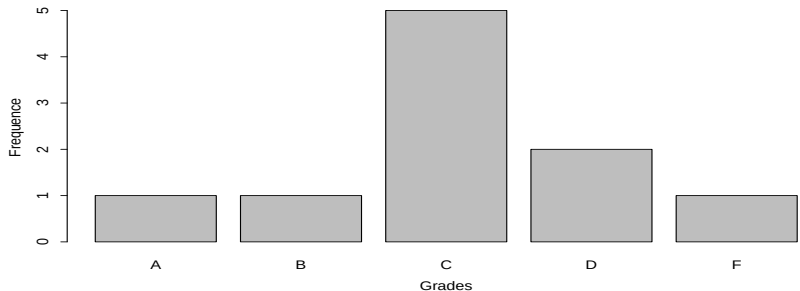
Notez que nous affichons l'invite ">" pour chaque commande telle qu'elle apparaîtrait dans R, mais cette invite ne fait pas partie du code.

- Nous pouvons aussi calculer les **proportions** (ou fréquence relatives)

```
> table(Grades) |> proportions()

Grades
  A   B   C   D   F
0.1 0.1 0.5 0.2 0.1
```

- Et présenter ces chiffres sous forme d'un **diagramme en barres** (Bar plot, en anglais)

```
> plot(Grades, xlab = "Grades", ylab = "Frequence")
```



Exemple : Tabagisme

En pratique dans la plupart des cas, les données à analyser se trouvent enregistrées dans R sous forme d'un data-frame.

Pour cet exemple, les données se trouvent dans le data.frame *Whickham* du package *mosaicData*. Pour commencer, vous devez (i) installer (si ce n'est pas déjà fait) et (ii) charger ce package.

Installation

```
> install.packages("mosaicData")
```

Chargement

```
> require("mosaicData")
```

Consultez ce document pour plus de détails.

Whickham donne une petite partie d'une enquête menée à Whickham au Royaume-Uni au début des années 1970. On a demandé aux participantes (uniquement des femmes) leurs âges et si elles fumaient. Un suivi vingt ans plus tard a révélé si la participante était encore en vie.

Whickham contiennent 1314 observations et les variables suivantes :

- outcome : statut de survie après 20 ans: un facteur avec deux niveaux (Alive, Dead)
- smoker : statut de fumeur au départ: un facteur avec deux niveaux (No, Yes)
- age: âge (en années) au départ

Pour avoir plus d'information sur ces données, taper

```
> head(Whickham)
```

outcome	smoker	age
Alive	Yes	23
Alive	Yes	18
Dead	Yes	71
Alive	No	67
Alive	No	64
Alive	Yes	38

```
> str(Whickham)

'data.frame': 1314 obs. of  3 variables:
 $ outcome: Factor w/ 2 levels "Alive","Dead": 1 1 2 1 1 1 1 2 1 1 ...
 $ smoker : Factor w/ 2 levels "No","Yes": 2 2 2 1 1 2 2 1 1 1 ...
 $ age    : int  23 18 71 67 64 38 45 76 28 27 ...
```

```
> help(Whickham)
```

La distribution de la variable smoker

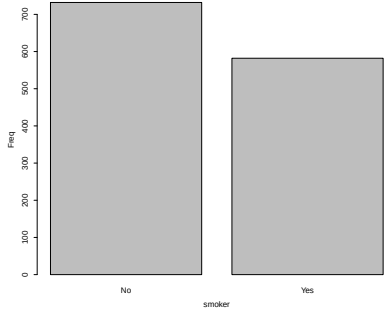
```
> Smo <- xtabs(~smoker, data = Whickham)
> Smo
```

No	Yes
732	582

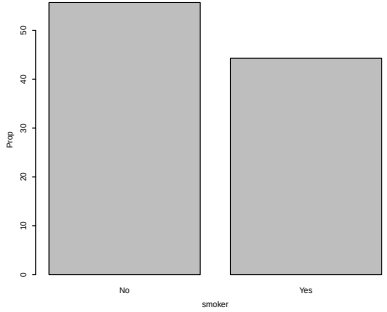
```
> ProSmo <- proportions(Smo) * 100
> ProSmo
```

No	Yes
55.7	44.3

```
> plot(Whickham$smoker)
> title(xlab = "smoker", ylab = "Freq")
```




```
> barplot(ProSmo)
> title(xlab = "smoker", ylab = "Prop")
```



Relation entre deux variables catégorielles

Table de contingence

Le tableau suivant est un exemple d'un tableau de contingence (ou tableau croisé).

```
> SmoOut <- xtabs(~outcome + smoker, data = Whickham)
> SmoOut
```

	smoker	
outcome	No	Yes
Alive	502	443
Dead	230	139

Il s'agit d'une façon de présenter des données d'énumération (comptage) d'individus classés en catégories selon les modalités définies par deux variables catégorielles (ici $X =$ smoker et $Y =$ outcome). À l'intersection de chaque ligne et de chaque colonne, on indique le nombre d'individus possédant les mêmes d'attributs.

```
> addmargins(SmoOut)
```

	smoker		
outcome	No	Yes	Sum
Alive	502	443	945
Dead	230	139	369
Sum	732	582	1314

Fréquences marginales

À l'aide de la fonction *addmargins*, on obtient les fréquences marginales et le nombre total d'obs. $n = 1314$

```
> proportions(SmoOut) * 100
```

	smoker	
outcome	No	Yes
Alive	38.20	33.71
Dead	17.50	10.58

Proportions conjointes

Exemple d'interprétation: 38.2% de l'échantillon est constitué de femmes non fumeuses encore en vie 20 ans après le début de l'étude.

La variable *smoker* partage la population en deux groupes: fumeurs et non-fumeurs. Dans chacun de ces groupes, on s'intéresse à la distribution de la variable *outcome* (statut de survie). On appelle cela une distribution conditionnelle.

```
> prSmoOut <- proportions(SmoOut, "smoker") * 100
> prSmoOut
```

	smoker	
outcome	No	Yes
Alive	68.58	76.12
Dead	31.42	23.88

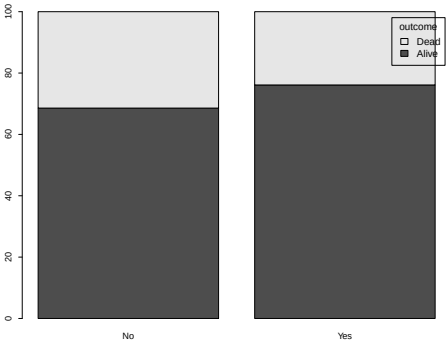
Proportions conditionnelles

Exemple d'interprétation: 68.6% des femmes non-fumeuses ont survécu 20 ans après.

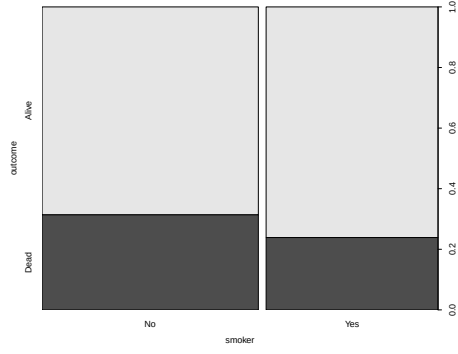
Représentation graphique

Pour visualiser l'association entre deux variables qualitatives, on utilise un Bar plot ou un Spine plot. Pour ce dernier la largeur des barres est proportionnelle aux effectives des catégories horizontales.

```
> barplot(prSmoOut, legend= TRUE,
          args.legend=list(title="outcome"))
```



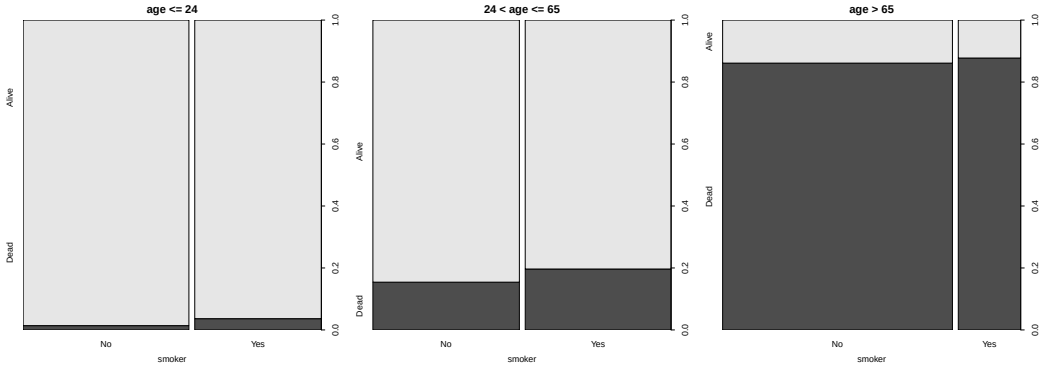

```
> plot(outcome ~ smoker, data = Whickham)
```



Cette analyse semble indiquer un effet “protecteur” du tabagisme: seulement 69% des non-fumeuses ont survécu au bout de 20 ans contre 76% des fumeuses!

Est-ce que ceci peut être l’interprétation correcte?
Examinons les données en intégrant l’âge des femmes.

```
> plot(outcome ~ smoker, data = subset(Whickham, age <= 24), main = "age <= 24")
> plot(outcome ~ smoker, data = subset(Whickham, age > 24 & age <= 65), main = "24 < age <= 65")
> plot(outcome ~ smoker, data = subset(Whickham, age > 65), main = "age > 65")
```



On peut voir que dans chaque tranche d'âge, la survie chez les fumeuses est inférieure à celle des non-fumeuses.

Comment expliquer alors que la tendance s'inverse lorsqu'on mélange toutes les tranches d'âges?

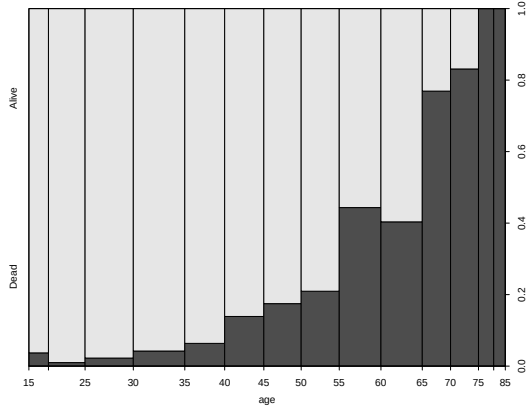
Cela peut être expliqué en examinant la relation entre l'âge et les autres variables (*outcome* et *smoker*).

En effet, sur le graphe précédent, on peut constater que peu de femmes âgées (plus de 65 ans lors de l'enquête initiale) étaient des fumeuses, mais beaucoup d'entre elles décèdent naturellement au bout de 20 ans. Ce qui a induit à l'effet "protecteur" du tabagisme observé lorsque l'âge a été ignoré.

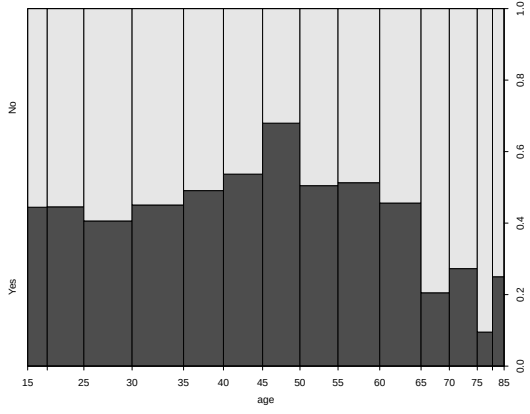
Le fait d'ignorer une variable importante, ici l'âge, peut cacher ou inverser un effet ou une tendance. C'est une illustration de ce qui est connu sous le nom du **paradoxe de Simpson**.

Notez qu'il est possible d'utliser un Spine plot pour représenter une variable qualitative en fonction d'une variable continue. Dans ce cas, cette dernière est découpée en classes selon la même méthode qu'un histogramme.

```
> plot(outcome ~ age, data = Whickham)
```




```
> plot(smoker ~ age, data = Whickham)
```



Description de données numériques

Il y a beaucoup à dire sur les variables numériques, et de nombreux outils ont été développés pour les étudier.

Par la suite, nous allons nous focaliser sur les éléments suivants:

- la répartition/distribution graphique des données
- les indicateurs dits de localisation
- les indicateurs dits de dispersion

Nous apprendrons davantage sur ces termes, et d'autres encore, au fur et à mesure de notre avancement dans la matière.

Exemple: Qualité de l'air

Les données *airquality* utilisées pour cet exemple donnent des informations sur la qualité de l'air à New York entre mai et septembre 1973. Une observation correspond à des mesures effectuées sur un jour de l'année. Voici les variables dans le jeu de données :

- Ozone: Taux d'ozone (ppb)
- Solar.R : Radiation solaire (lang)
- Wind : Vitesse moyenne du vent (mph)
- Temp : Température quotidienne maximale (degré Fahrenheit)
- Month : Mois (1-12)
- Day : Jours (1-31)

Notez qu'il s'agit ici d'une étude observationnelle.

Pour avoir plus d'information sur ces données, taper

```
> head(airquality)
```

Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7	67	5	1
36	118	8	72	5	2
12	149	13	74	5	3
18	313	12	62	5	4
NA	NA	14	56	5	5
28	NA	15	66	5	6

```
> str(airquality)

'data.frame': 153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...


> help(airquality)
```

Avec de telles données, on peut être intéressé à étudier la distribution (les variations) dans les taux d'ozone, par exemple. Cette question porte sur **une seul variable numérique**. On parle alors d'une **analyse univariée**.

Une autre question d'intérêt est de décrire le lien entre, par exemple, le tau d'ozone et la température ou entre le tau d'ozone et et le mois. Ces questions portent sur **deux variables** (numérique-numérique ou numérique-factoriel). On parle alors d'une **analyse bivariée**.

Ces questions peuvent être explorés à la fois **graphiquement** et **numériquement**.

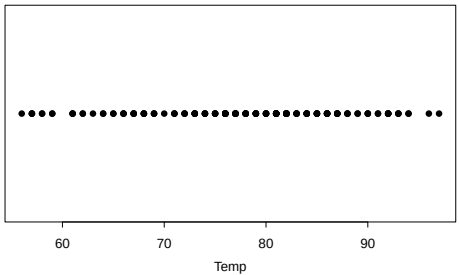
Description de données numériques

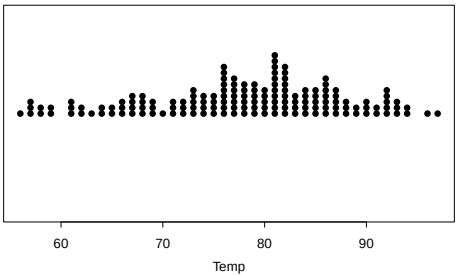
Visualiser la répartition/distribution des données

Diagramme à points

Un diagramme à points (Dotplot ou Stripchart, en anglais) permet de représenter des données sur la droite graduée.

```
> stripchart(airquality$Temp, xlab = "Temp", pch = 16)
> stripchart(airquality$Temp, xlab = "Temp", pch = 16, method = "stack")
```

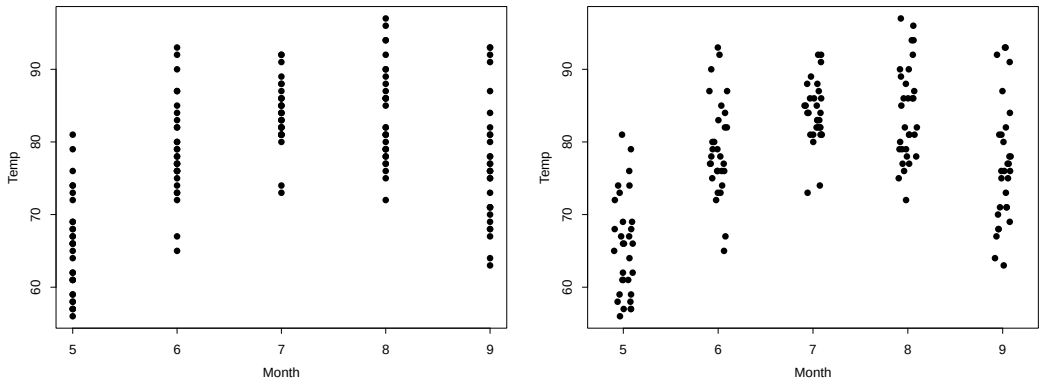




L'option *method* = *"stack"* empile les valeurs identiques, créant un graphique qui représente mieux les données.

Voici un dotplot de la température par mois.

```
> stripchart(Temp ~ Month, data = airquality, vertical = TRUE, xlab = "Month")
> stripchart(Temp ~ Month, data = airquality, vertical = TRUE, xlab = "Month", method = "jitter")
```

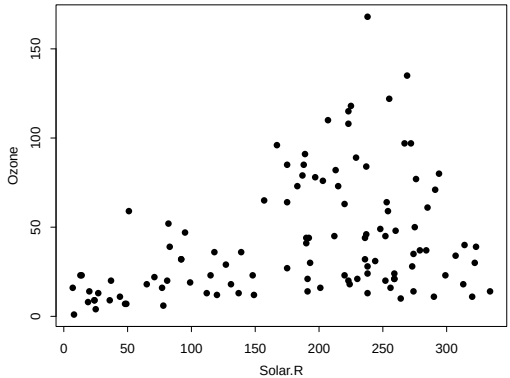


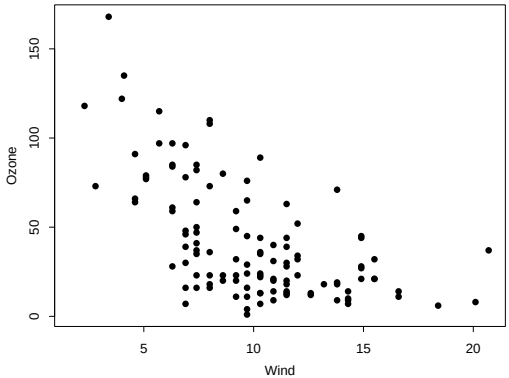
Ici, *"jitter"* est utilisée pour ajouter un bruit aléatoire et réduire ainsi le problème des points qui se chevauchent.

Diagramme de dispersion

Un diagramme de dispersion, ou nuage de points (scatter plot, en anglais), est une représentation graphique qu'on utilise typiquement pour visualiser/analyser la possible relation/association entre deux variables quantitatives continues.

```
> plot(Ozone ~ Solar.R, data = airquality)
> plot(Ozone ~ Wind, data = airquality)
```





Histogramme

C'est un graphique utile pour examiner la forme et la dispersion des données.

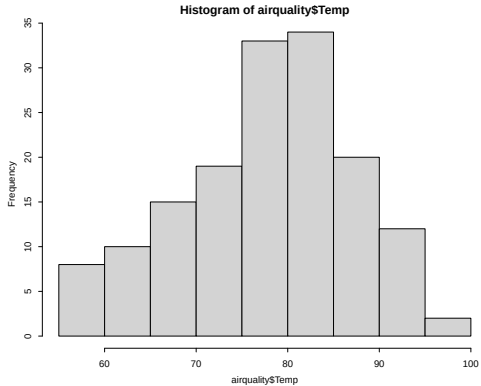
Pour construire un histogramme il nous faut (1) une série $x_1, \ldots, x_n$ de n observations, (2) se fixer une origine $a_0 \leqslant \min_i x_i$, (3) choisir K le nombre de classes (bins, en anglais), ou, de façon équivalente, choisir h la largeur des classes (bin-width, en anglais), (4) découper les données en classes $[a_0, a_0 + h)$, $[a_0 + h, a_0 + 2h)$,..., et compter le nombre n_k d'observations se trouvant dans la k—ième classe.

Il y a deux types d'histogrammes:

- Histogramme des fréquences (classique): Les hauteur des barres correspondent aux effectives n_k.
- Histogramme des densités (normalisé): Les hauteur des barres correspondent aux $\frac{n_k}{nh}$, ce qui résulte en une surface totale égale à 1, nous permettant ainsi de plus facilement comparer plusieurs distributions et/ou de confronter nos données à une fonction de densité théorique.

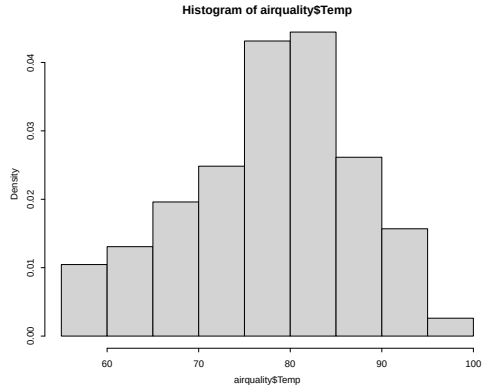
CLASSIQUE

```
> hist(airquality$Temp)
```



NORMALISÉ

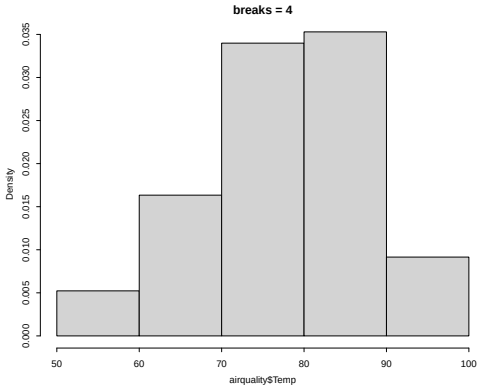
```
> hist(airquality$Temp, freq = FALSE)
```

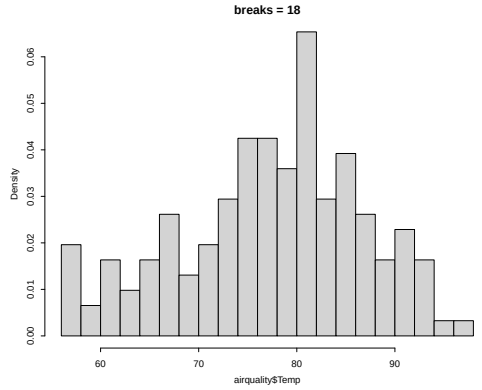


NOMBRE DE CLASSES: de nombreuses règles existent pour choisir une valeur optimale pour ce paramètre qui peut impacter fortement le graphique. Par défaut, la fonction *hist* utilise la formule de Sturge.

L'argument *breaks* permet de contrôler le nombre ou les bornes des classes.

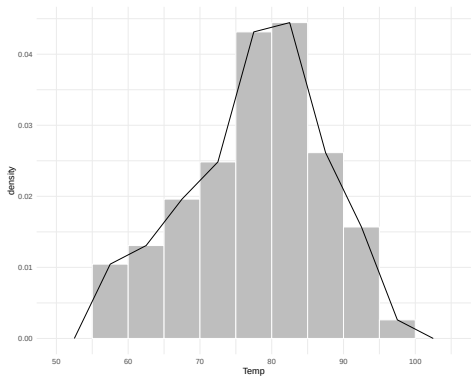
```
> hist(airquality$Temp, freq = FALSE, breaks = 4)
> hist(airquality$Temp, freq = FALSE, breaks = 18)
```

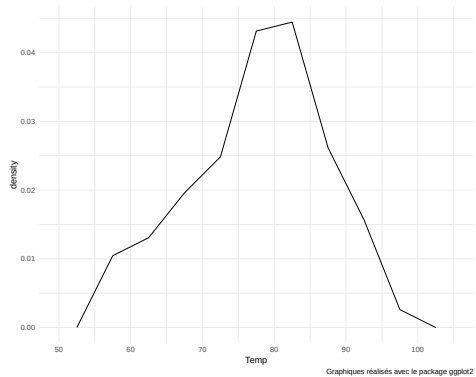




Notez que R traduit le nombre indiqué dans *breaks* comme une suggestion qu'il adapte pour que la lecture du graphique soit facile.

Une autre manière de résumer la distribution des données est d'utiliser un polygone que l'on obtient en joignant les milieux des bases supérieures de chaque rectangle de l'histogramme par des segments de droite. On adjoint généralement une classe d'effectif nul, de part et d'autre de l'histogramme.





Courbe de densité estimée

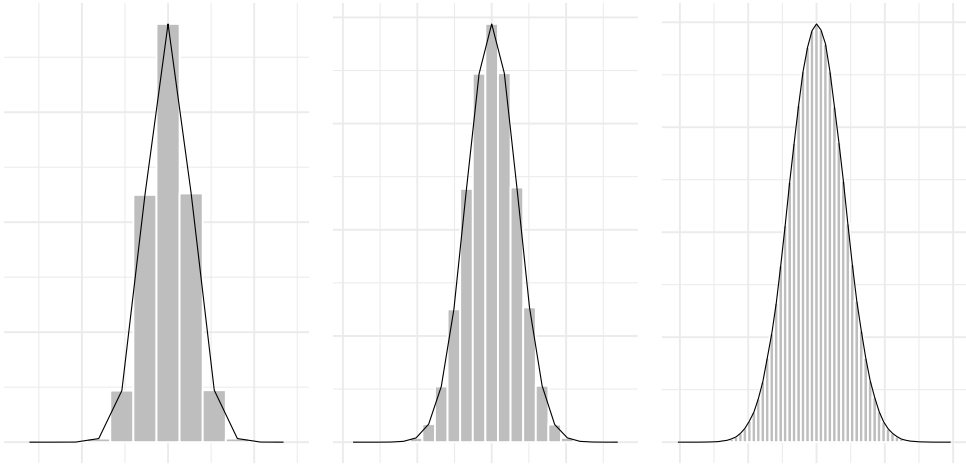
Une courbe de densité est la version continue d'un histogramme normalisé. Il s'agit ici de remplacer ce dernier par une courbe lisse (smooth, en anglais) enfermant une surface d'aire égale à 1 en supprimant les discontinuités (les sauts entre les segments horizontaux).

La courbe de densité permet aussi de corriger l'un des défauts majeurs de l'histogramme, à savoir sa rigidité : la hauteur d'un rectangle est seulement fonction du nombre de données contenues dans sa base sans tenir compte de leur répartition sur cet intervalle.

En réalité, une courbe de densité est la représentation graphique d'un estimateur de la fonction de densité théorique, un sujet que nous aborderons plus loin dans ce cours. Cependant, nous fournissons ici une explication heuristique pour comprendre les bases de cet outil très utile.

Le concept principal est le suivant: nous supposons que nos observations proviennent d'une immense population. Dans notre cas, imaginez que, au lieu des 153 observations, nous disposant de 10^6 de prélèvements. Nous pouvons alors créer un histogramme/polygone avec h (largeur des classe) très très petite. Si nous faisons cela, les bars consécutifs seront très similaires et l'histogramme/polygone qu'on obtiendrait sera très lisse.

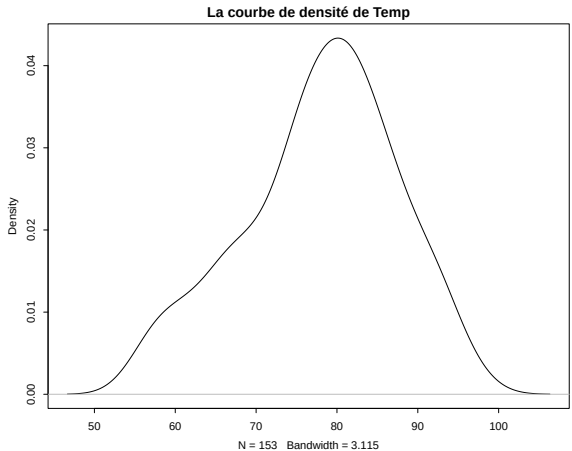
Cette idée est illustrée dans le slide suivant avec un échantillon de 10^6 observations.



REMARQUE. De gauche à droite $\texttt{breaks} = 10, 20$ et 60.

En pratique, avec un nombre relativement petit d'observations, nous ne pouvons pas appliquer le raisonnement décrit ci-dessus, mais nous pouvons espérer approximer la densité théorique (celle de la population) à l'aide d'un estimateur. En R, nous pouvons calculer ce dernier à l'aide de la fonction *density*.

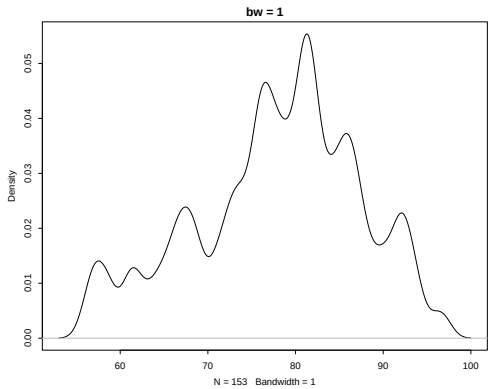
```
> density(airquality$Temp) |> plot(main = "La courbe de densité de Temp")
```

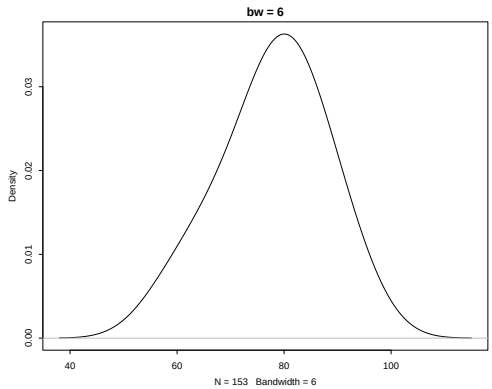


Un des arguments les plus importants de *density* est *bw* (bandwidth, càd. paramètre de lissage) qui permet de contrôler le niveau du lissage à appliquer. Par défaut, *density* utilise la règle de Silverman.

La figure suivante illustre l'effet du bandwidth.

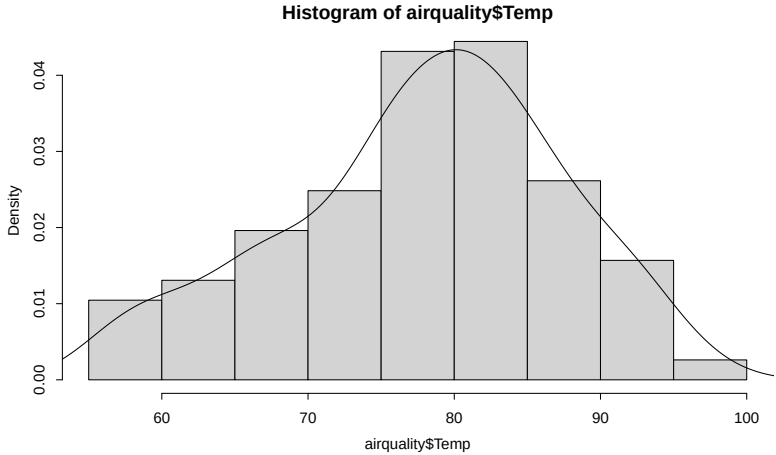
```
> density(airquality$Temp, bw = 1) |> plot(main = "bw = 1")
> density(airquality$Temp, bw = 6) |> plot(main = "bw = 6")
```





Notez qu'il est possible de combiner un histogramme et une courbe de densité sur le même graphique. Voici un exemple.

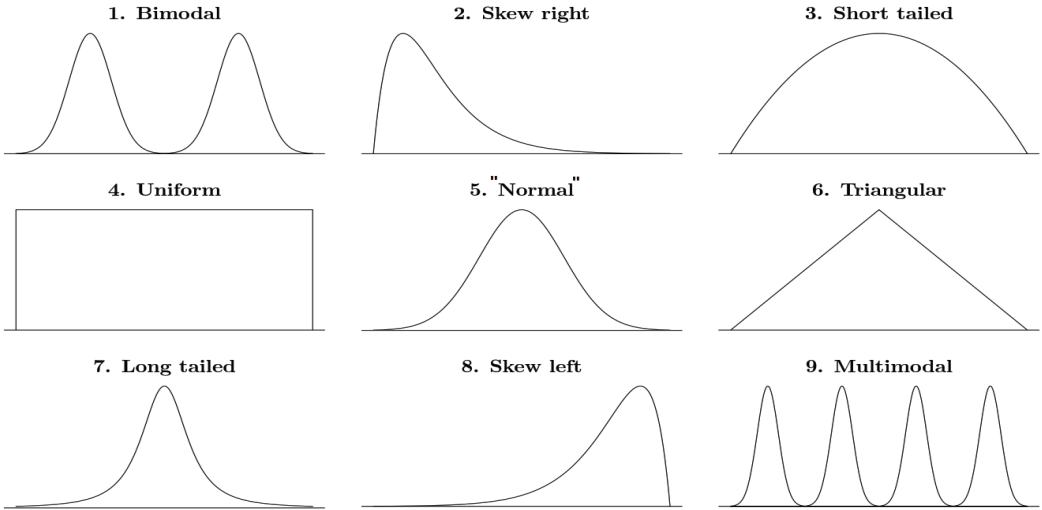
```
> hist(airquality$Temp, freq = FALSE)
> density(airquality$Temp) |> lines()
```



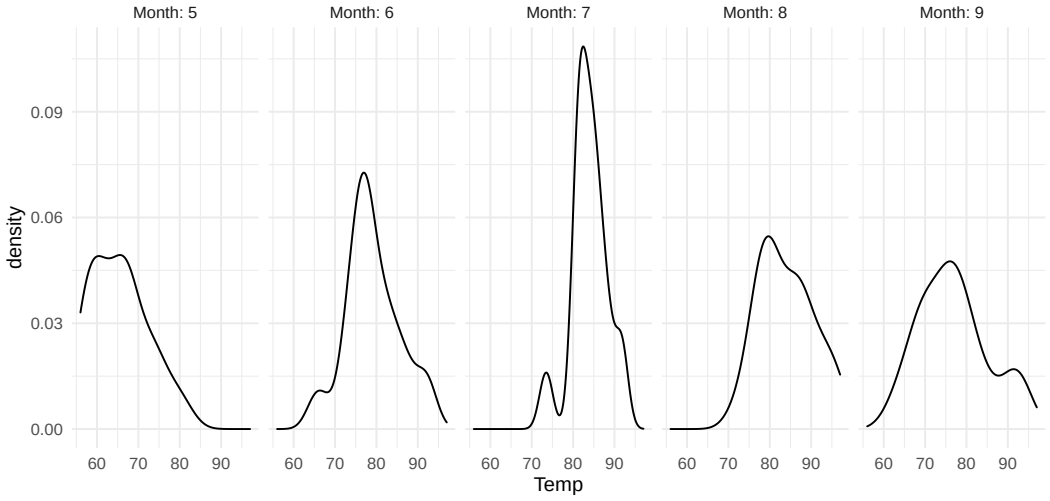
Une courbe de densité nous fournit beaucoup d'information, notamment : les valeurs les plus représentées (pics), concentration/étalement des données par zone, présence de valeurs extrêmes, symétrie/asymétrie, aplatissement, centre, etc.

Elle nous permet aussi de comparer la distribution des données à des densités théoriques (e.g., Normal, Student, Gamma, ...).

Le graphe suivant montre quelques exemples courbes de densité (théorique) typiques.


1. Bimodal
2. Skew right
3. Short tailed
4. Uniform
5. "Normal"
6. Triangular
7. Long tailed
8. Skew left
9. Multimodal

Le graphe suivant montre la densité estimée de la variable *Temp* pour chaque mois d'étude (avec les mêmes échelles pour les axes x et y).



Description de données numériques

Indicateurs de localisation

Les indicateurs de localisation ainsi que ceux de dispersion (qu'on va voir par la suite) aident à caractériser les données et à décrire leurs spécificités.

Ces indicateurs **peuvent** s'appliquer à la population ou à l'échantillon.

> Les résumés numériques sur une population sont appelés paramètres, le plus souvent ces paramétrés sont inconnus. Les résumés numériques sur un échantillon sont appelés statistiques ou mesures empiriques ou encore échantillonnales. Ces derniers sont typiquement utilisés pour estimer/approximer les paramètres.

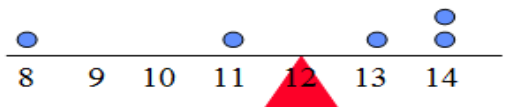
Des définitions plus formelles des paramètres qui caractérisent une population seront données ultérieurement.

Dans ce chapitre, on considère uniquement les mesures empiriques.

La MOYENNE EMPIRIQUE (mean, en anglais) est la moyenne arithmétique des observations, notée par $\bar{x}_n$, est donnée par

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

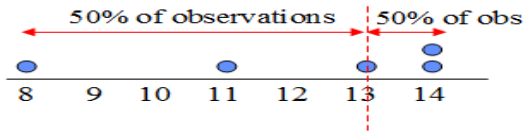
La moyenne empirique peut être interprété comme étant le centre de masse des observations ou leur point d'équilibre.




```
> mean(c(8, 11, 13, 14, 14))

[1] 12
```

La MÉDIANE EMPIRIQUE (median) est le nombre qui est juste au milieu (au sens proportionnelle) des données. Ce nombre sera notée par $q_{0.5}$

$q_{0.5}$ est telle que la moitié des observations lui sont inférieures et l'autre moitié lui sont supérieures.



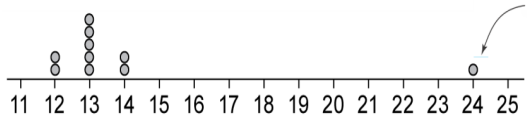

```
> median(c(8, 11, 13, 14, 14))

[1] 13
```

S'il y a un nombre pair d'observations, alors la médiane est la moyenne des deux points les plus proches du milieu.

```
> median(c(8, 11, 13, 14))

[1] 12
```

Moyenne ou Médiane ?

La moyenne se laisse influencer par les observations atypiques (outliers), qui peuvent être **parfois** erronées.



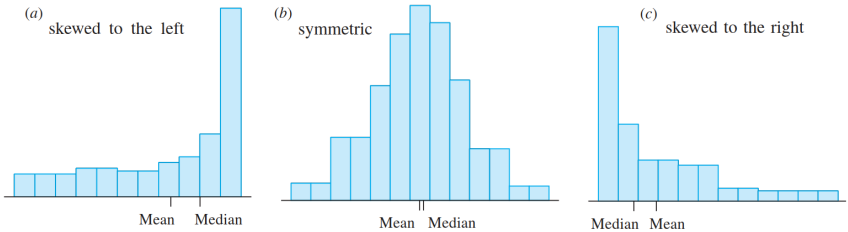
La médiane est plus résistante ou robuste que la moyenne.

```
> x <- c(12, 12, 13, 13, 13, 13, 13, 14, 14)
> y <- c(12, 12, 13, 13, 13, 13, 13, 14, 14, 24)
```

```
> mean(x)

[1] 13

> mean(y)

[1] 14.1
```

```
> median(x)

[1] 13

> median(y)

[1] 13
```

Moyenne, médiane et asymétrie



- Une densité/histogramme est dite **symétrique** si sa moitié droite est une image en miroir de sa moitié gauche, voire (b).
- Une densité/histogramme **asymétrique** possède une queue plus longue que l'autre. Une densité/histogramme possède avec une longue queue droite, voir (c), est dite **asymétrique à droite**, dans le cas contraire elle est dite **asymétrique à gauche**.
- Lorsque une distribution est parfaitement symétrique, la moyenne et la médiane seront identiques.
- Lorsque une distribution est asymétrique, la moyenne sera plus proche de la queue des données que la médiane.

Quantile empirique

Pour un $p \in (0,1)$ donné, le quantile empirique d'ordre p, noté q_p est la valeur qui partage l'échantillon en deux : une proportion p des observations sont inférieures à q_p et une proportion $1-p$ lui sont supérieures.

Pour calculer q_p, il faut d'abord **ordonner les données** du plus petit au plus grand. Si l'échantillon initial est noté $x_1, x_2, \ldots, x_n$ l'échantillon ordonné sera noté $x_{(1)}, x_{(2)} \ldots, x_{(n)}$, avec

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots, x_{(n)}$$

q_p est l'observation du rang $(1+(n-1)p)$, ce qu'on peut écrire comme

$$q_p = x_{(1+(n-1)p)},$$

Autrement dit, c'est l'observation qui occupe le rang $(1+(n-1)p)$ dans l'échantillon ordonné.

Exemple

```
> x <- c(22.3, 17.9, 20.4, 24.6, 19.5, 26.2, 18.7)
> quantile(x, p = 0.3)

 30%
19.3
```

$q_{0.3}$ est la valeur qui correspond au rang $1 + (7-1) \times 0.3 = 2.8$. On peut la calculer, manuellement, par interpolation linéaire. Autrement dit, c'est la valeur situé à 80% du chemin entre $x_{(2)}$ et $x_{(3)}$.

$$\begin{aligned} q_{0.3} &= x_{(2)} + 0.8 \times (x_{(3)} - x_{(2)}) \\ &= 18.7 + 0.8 \times (19.5 - 18.7) = 19.34 \end{aligned}$$

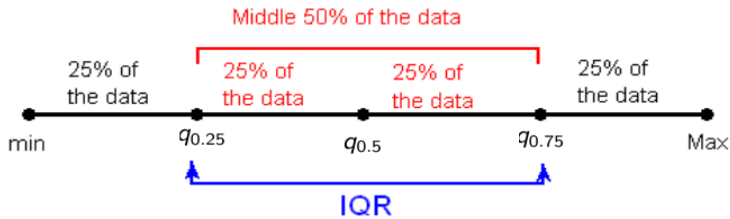
Vous pouvez obtenir autant de quantiles que vous le souhaitez!

```
> quantile(x, p = c(0.25, 0.5, 0.75))

 25%  50%  75%
19.1 20.4 23.5
```

Certains quantiles ont des noms particuliers et des notations particulières. Les quantiles d'ordre 0.25, 0.50 et 0.75 sont également appelésle premier quartile, le deuxième quartile (ou médiane) et le troisième quartile et sont souvent dénotés par Q_1, Q_2 et Q_3.

Le résumé à cinq nombres

Le vecteur $(x_{(1)}, Q_1, Q_2, Q_3, x_{(n)})$ est souvent appelé le résumé à cinq nombres. $x_{(1)}$ est simplement la plus petite observation et est souvent dénotée min (pour minimum). De même $x_{(n)}$, est simplement la plus grande observation et est souvent dénotée max (pour maximum).



Pour obtenir ces statistiques, vous pouvez utiliser la fonction *quantile* ou *summary*

```
> quantile(x)

  0%  25%  50%  75% 100%
17.9 19.1 20.4 23.5 26.2
```

```
> summary(x)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17.9	19.1	20.4	21.4	23.4	26.2

Calculant ce résumé pour la variable *Ozone* de l'exemple de la qualité de l'air:

```
> summary(airquality$Ozone)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
1	18	31.5	42.1	63.2	168	37

Notez que la fonction *summary* retourne, en plus du résumé à cinq nombres, la moyenne des données et le nombre de valeurs manquantes (si de telles valeurs existent).

L'avantage de la fonction *summary* c'est qu'on peut l'appliquer aussi sur un data-frame pour obtenir les statistiques de toutes les variables. Voici un exemple

```
> summary(airquality)
```

Ozone	Solar.R	Wind	Temp	Month	Day
Min. : 1.0	Min. : 7	Min. : 1.70	Min. :56.0	Min. :5.00	Min. : 1.0
1st Qu.: 18.0	1st Qu.:116	1st Qu.: 7.40	1st Qu.:72.0	1st Qu.:6.00	1st Qu.: 8.0
Median : 31.5	Median :205	Median : 9.70	Median :79.0	Median :7.00	Median :16.0
Mean : 42.1	Mean :186	Mean : 9.96	Mean :77.9	Mean :6.99	Mean :15.8
3rd Qu.: 63.2	3rd Qu.:259	3rd Qu.:11.50	3rd Qu.:85.0	3rd Qu.:8.00	3rd Qu.:23.0
Max. :168.0	Max. :334	Max. :20.70	Max. :97.0	Max. :9.00	Max. :31.0
NA's :37	NA's :7				

La fonction *summary* s'adapte à la nature des données fournies. Ainsi, lorsqu'on transforme la variable *Month* en facteur, ce qui est logique puisqu'il s'agit d'une variable catégorielle, nous obtenus la sortie suivante.

```
> airquality <- transform(airquality, Month = factor(Month))
> summary(airquality)
```

Ozone	Solar.R	Wind	Temp	Month	Day
Min. : 1.0	Min. : 7	Min. : 1.70	Min. :56.0	5:31	Min. : 1.0
1st Qu.: 18.0	1st Qu.:116	1st Qu.: 7.40	1st Qu.:72.0	6:30	1st Qu.: 8.0
Median : 31.5	Median :205	Median : 9.70	Median :79.0	7:31	Median :16.0
Mean : 42.1	Mean :186	Mean : 9.96	Mean :77.9	8:31	Mean :15.8
3rd Qu.: 63.2	3rd Qu.:259	3rd Qu.:11.50	3rd Qu.:85.0	9:30	3rd Qu.:23.0
Max. :168.0	Max. :334	Max. :20.70	Max. :97.0		Max. :31.0
NA's :37	NA's :7				

Avant d'analyser un data-frame en R, il faut vérifier la nature de chaque colonne/variable et la transformer, si nécessaire. Cela est important, car, certaines fonctions, comme *summary*, traitent chaque variable différemment en fonction de sa nature (numérique, caractère, facteur, etc.).

Statistiques par groupe

Pour calculer des statistiques descriptives tels que le minimum, la moyenne, etc., d'une variable pour chaque niveau d'un facteur nous utilisons la fonction *aggregate* dont voici quelques exemples typiques.

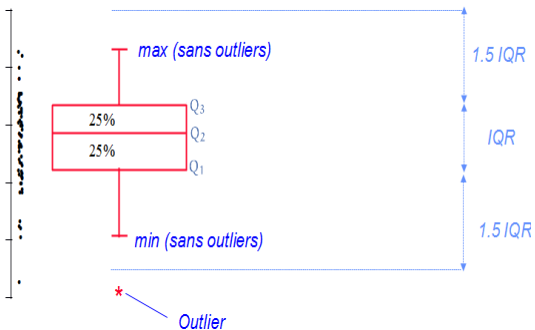
```
> # moyenne de la température par mois
> aggregate(Temp ~ Month, data = airquality, FUN = mean)

  Month Temp
1     5 65.5
2     6 79.1
3     7 83.9
4     8 84.0
5     9 76.9

> # moyenne de la température et de ozone par mois
> aggregate(cbind(Temp, Ozone) ~ Month, data = airquality, FUN = mean)

  Month Temp Ozone
1     5 66.7  23.6
2     6 78.2  29.4
3     7 83.9  59.1
4     8 84.0  60.0
5     9 76.9  31.4
```

```
> # médiane de la température par mois
> aggregate(Ozone ~ Month, data = airquality, FUN = quantile, p = 0.5)

  Month Ozone
1     5    18
2     6    23
3     7    60
4     8    52
5     9    23

> # summary de la température par mois
> aggregate(Ozone ~ Month, data = airquality, FUN = summary)

  Month Ozone.Min. Ozone.1st Qu. Ozone.Median Ozone.Mean Ozone.3rd Qu.
1     5        1.0          11.0         18.0       23.6          31.5
2     6       12.0          20.0         23.0       29.4          37.0
3     7        7.0          36.2         60.0       59.1          79.8
4     8        9.0          28.8         52.0       60.0          82.5
5     9        7.0          16.0         23.0       31.4          36.0
  Ozone.Max.
1      115.0
2       71.0
3      135.0
4      168.0
5       96.0
```

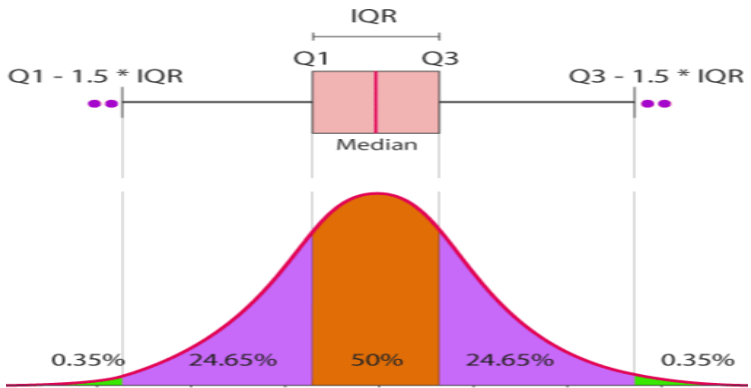
Le diagramme en boîte

Le diagramme en boîte (en anglais boxplot) ou boîte à moustache est une simple représentation graphique du résumé à cinq nombres dont le principe est décrit par le schéma suivant :

- une boîte du premier quartile au troisième quartile
- une ligne à la hauteur de la médiane
- les lignes (moustaches) qui vont des extrémités de la boîte aux points extrêmes se trouvant dans l'intervalle $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$.
- les observations classées comme atypiques (**outliers**). C'est tous les points en dehors des moustaches.

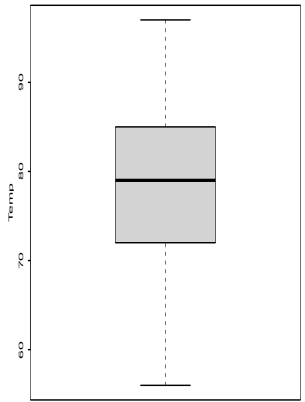


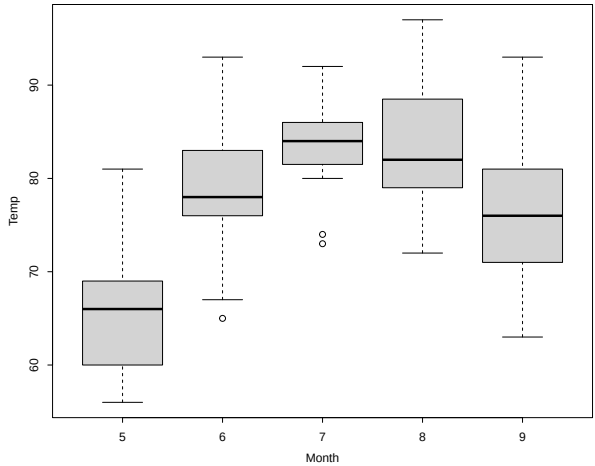
La figure suivante illustre le lien entre un boxplot et une courbe de densité "normal".



Boxplot avec R

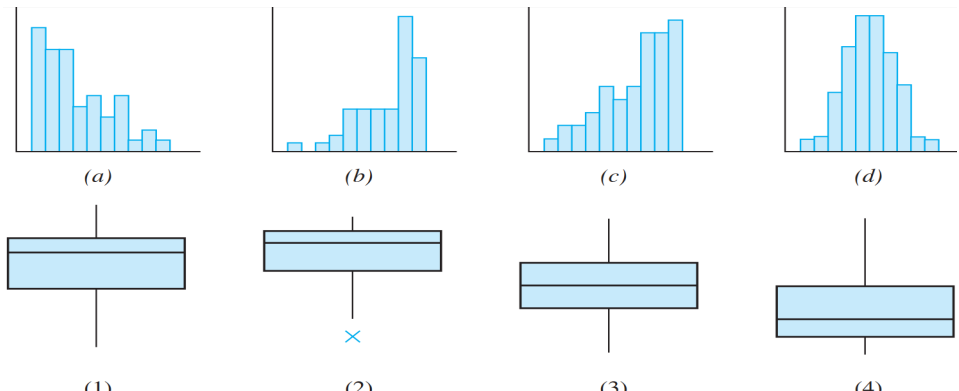
```
> boxplot(airquality$Temp, ylab = "Temp")
> boxplot(Temp ~ Month, data = airquality)
```



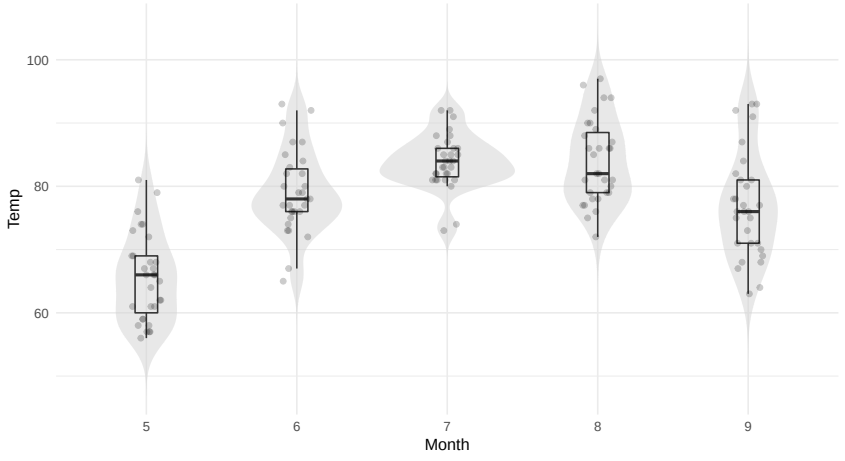


Exercice

Faites correspondre chaque histogramme à la boîte à moustaches qui représente le même ensemble de données.



Un boxplot permet de comparer facilement et rapidement plusieurs distribuions mais, en comparaison à une densité ou même un histogramme, il cache souvant beaucoup trop de détails. Il est possible de combiner à la fois, un boxplot, une courbe de densité et un diagramme à points sur le même graphique.



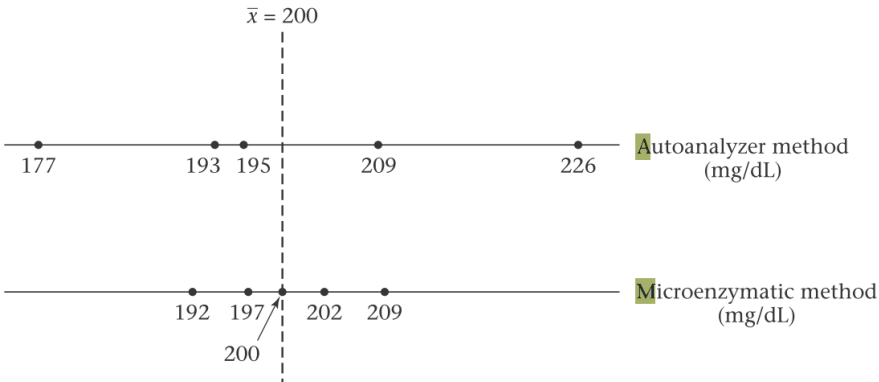
Description de données numériques

Indicateurs de dispersion

Pour mieux cerner les caractéristiques d'un échantillon, il est nécessaire de compléter les indicateurs de localisation par des indicateurs de dispersion, qui mesureront la variabilité des données.

La figure suivante montre deux jeux de données qui partagent la même moyenne mais pas la même dispersion.

Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



L'ÉTENDUE d'un échantillon est l'écart entre ces valeurs extrêmes, autrement dit c'est $\max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n)$

Pour la méthode A:

```
> A <- c(177, 193, 195, 209, 226)
> max(A) - min(A)

[1] 49
```

Pour la méthode M:

```
> M <- c(192, 197, 202, 209)
> max(M) - min(M)

[1] 17
```

L'ÉCART INTERQUANTILE (IQR) d'un échantillon est l'écart entre sont troisième et son premier quartile, autrement dit $\mathrm{IQR} = Q_3 - Q_1$.

Pour la méthode A:

```
> IQR(A)

[1] 16
```

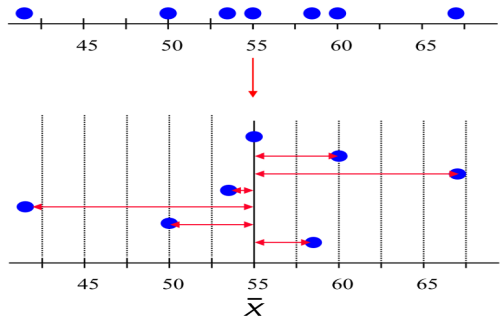
Pour la méthode M:

```
> IQR(M)

[1] 8
```

Variance et écart-type empiriques

La variance, notée par s_n^2, de n observations est la somme des carrés des écarts à la moyenne divisée par $n-1$.

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$



Pour la méthode A:

```
> var(A)

[1] 340
```

Pour la méthode M:

```
> var(M)

[1] 52.7
```

Il est facile de vérifier que

$$s_n^2 = \frac{1}{n-1}\left(\sum_i x_i^2 - n\bar{x}^2\right)$$

L'écart-type (standard deviation ou sd) est la racine carrée de la variance

$$s_n = \sqrt{s_n^2}$$

Il s'exprime dans la même unité que les données, ce qui rend son interprétation plus facile que celle de la variance.

L'écart-type quantifie la distance entre les observations et leur moyenne. Autrement dit, l'écart-type mesure la dispersion des données autour de leur moyenne.

Pour la méthode A:

```
> sd(A)

[1] 18.4
```

Pour la méthode M:

```
> sd(M)

[1] 7.26
```

Ce qui exprime bien la différence de variabilité des mesures entre les deux méthodes. La méthode A a un écart-type presque trois fois plus grand que celui de la méthode M. On peut dire que cette dernière est trois fois plus précise.

Notez que, comme la moyenne, l'écart-type se laisse facilement influencer par les valeurs atypiques :

```
> x <- c(12, 12, 13, 13, 13, 13, 13, 14, 14)
> y <- c(12, 12, 13, 13, 13, 13, 13, 14, 14, 24)
```

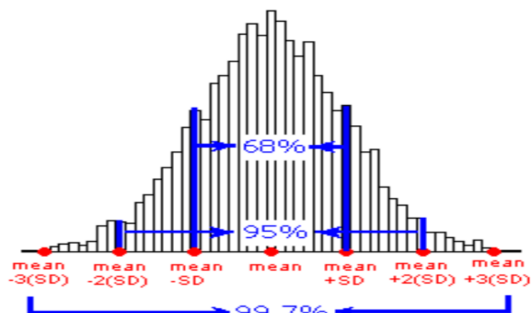
```
> sd(x)

[1] 0.707
```

```
> sd(y)

[1] 3.54
```

Règle empirique (dite des $68 - 95 - 99.7$)

Une autre façon utile pour interpréter les valeurs d'un écart type est la règle empirique qui dit que pour beaucoup de jeux de données, la majorité (95%) des valeurs d'échantillon sont éloignées de moins de deux écarts types de la moyenne. Plus précisément, pour des données avec une distribution raisonnablement **symétrique** et **sans valeurs** aberrantes:

- Environ 68% des valeurs sont situées à moins d'un écart type de la moyenne (càd entre $\bar{x} - s_n$ et $\bar{x} + s_n$);
- Environ 95% des valeurs sont situées à moins de deux écarts types de la moyenne ($\bar{x} - 2s_n$ et $\bar{x} + 2s_n$);
- Environ 99.7% des valeurs sont situées à moins de trois écarts types de la moyenne ($\bar{x} - 3s_n$ et $\bar{x} + 3s_n$).



Quelques propriétés

Supposons qu'on observe $x_1 \ldots, x_n$. Soit $y_i = a + bx_i$, $i = 1, \ldots, n$, alors

$$\bar{y} = a + b\bar{x}, \quad s_y^2 = b^2 s_x^2, \quad s_y = |b| s_x$$

Exemple. La moyenne et l'écart-type de la variable *Temp* du data *airquality* sont 77.882 (°F, degree Fahrenheit) et 9.465 (°F), respectivement. Sachant que

$$T_C = (T_F - 32) \times \frac{5}{9}$$

On déduit que, en Celsius, la moyenne et l'écart-type sont

```
> c((77.882 - 32) * (5/9), 9.465 * (5/9))

[1] 25.49  5.26
```

Un inconvénient de l'écart-type en tant que mesure de variation est qu'il **dépend de l'unité de mesure**.

Aussi, l'interprétation de l'écart-type dépend de **la magnitude des données**. En effet, un écart-type de 10^o n'a pas le même sens si la température moyenne de référence est 20^o ou 20000^o. Les données présentent une forte variabilité si l'**écart-type est grand par rapport à la moyenne**.

On ne peut comparer la variation de deux ensembles de données que si (1) ils sont mesurés en même unité et (2) ils ont (approximativement) la même moyenne.

Coefficient de variation empirique

Le coefficient de variation, donné par $cv_x = \frac{s_x}{\bar{x}}$, est une mesure de dispersion relative. Ce nombre, sans unité, peut être utilisé pour comparer deux (ou plusieurs) séries de données d'unités et/ou de moyennes différentes.

Excercice. Supposons qu'on observe $x_1 \ldots, x_n$. Soit $y_i = bx_i$, $i = 1, \ldots, n$, avec $b > 0$. Montrez que $cv_x = cv_y$.

Exemple. Voici quelques statistiques obtenues sur deux échantillons correspondants à deux méthodes de mesures différentes.

	moyenne	écart-type	cv
Méthode A	38.3	3.13	0.082
Méthode B	95.9	5.28	0.055

La méthode B est, vraisemblablement, plus précise que la méthode A, mais la comparaison de l'écart-type (uniquement) peut entraîner une fausse conclusion.

Commandes R les plus utiles vues dans ce chapitre

- c, factor
- table, xtabs, proportions
- head, str, help, require
- summary, aggregate, subset, transform
- plot, stripchart, barplot, boxplot, hist, desnity, lines, title
- mean, median, quantile, var, sd, min, max, IQR
- install.packages, require
- Packages: mosaicData