Modélisation

Marie-Luce Taupin
marie-luce.taupin@univ-evry.fr

Laboratoire LaMME, Université d'Evry val d'Essonne
http://www.math-evry.cnrs.fr/members/mtaupin/welcome

2018

Organisation du cours

Séances

1. 05 mars
2. 12 mars
3. 19 mars
4. 26 mars - contrôle
5. 09 avril
6. 16 avril - contrôle
7. 30 avril

Evaluation basée sur

- Rapport rédigé sur l'étude du jeu de données
- Notes des deux contrôles
- Examen

Statistiques descriptives uni-dimensionnelles - Plan

1. Généralités
2. Tableaux statistiques
3. Représentations graphiques
4. Indicateurs statistiques

- **Quelques références bibliographiques**

  - Statistique Descriptive. M. Lethielleux. Dunod. Collection "Express". 2005.
  - Éléments de Statistique. J.J. Droesbeke. Ellipses Marketing. 2002.
  - Probabilités, Analyse des données et Statistiques. G. Saporta. Editions Technip. 2006.
  - Statistique descriptive : Cours et exercices corrigés A. Hamon, N. Jégou. Presses Univ. de Rennes, Collection Didact. Statistique. 2008.
  - Statistiques avec R. P. A. Cornillon et co-auteurs. Presses Univ. de Rennes, Collection Didact. Statistique. 2010.

1. Vocabulaire

- **Population**
  - Ensemble d'individus concernés par une étude.
  - **Exemples**
    - ★ étudiants de l'Université d'Evry en 2016.
    - ★ étudiants inscrits en mathématiques à l'université d'Evry en 2016.
    - ★ salariés aux états-unis en 2012.
    - ★ électeurs français en 2016.
- **Individu ou unité statistique**
  - Un élément de la population étudiée.
  - **Exemples**
    - ★ un étudiant de l'Université d'Evry en 2016.
    - ★ un salarié aux états unis en 2012.
    - ★ un électeur français en 2016.

- **Taille** de la population : nombre d'individus de la population.
- **Échantillon**
  - Sous-ensemble de la population dont les individus feront l'objet de mesure.
  - Doit être représentatif de la population (selon certains critères).
  - **Exemples** : un groupe d'étudiants, un lot de pièces, ...
- **Enquête** : opération consistant à questionner ou observer les individus d'un échantillon.
- **Recensement** : enquête exhaustive sur toute la population.
- **Sondage** : enquête sur un échantillon représentatif de la population.

- **Caractère ou variable statistique** X
  - caractéristique de l'individu à laquelle l'étude s'intéresse.
  - **Exemples**
    - La taille, le poids, l'âge, le salaire.
    - le sexe, la catégorie socio-professionnelle, la nationalité.
    - le niveau d'étude
    - le taux de cholestérol
    - le statut fumeur ou non fumeur
- **Ensemble des modalités** d'une variable : ensemble des valeurs possibles que peut prendre la variable.
  - Tout individu doit présenter une et une seule modalité de chaque variable étudiée.
  - Deux **types** de variables : **qualitative** et **quantitative**.

- **Observation** : valeur x_k prise par la variable X pour un individu donné k de la population (ou de l'échantillon).

- **Série statistique** univariée (tableau à 1 entrée) :
  ensemble des observations $x_1, x_2, \ldots, x_k, \ldots, x_n$ (recueillies sur n individus).

- **Tableau des données brutes** :

  Ensemble des données recueillies sur n individus :
  tableau individus $\times$ variables.

Exemple de tableau
Données : Current Population Survey

Extrait des résultats d'une enquête réalisée aux États-Unis en juillet 2012 par le bureau du recensement et le bureau des statistiques du travail. Les données sont extraite de l'enquête de juillet 2012.

id.	Age	Sexe	Region	Stat-Mari	Sal-Hor	Syndicat	Categorie	⋯
1	58	F	NE	C	13.25	non	5	⋯
2	40	M	W	M	12.50	non	7	⋯
3	29	M	S	C	14.00	non	5	⋯
4	59	M	NE	D	10.60	oui	3	⋯
5	51	M	W	M	13.00	non	3	⋯
6	19	M	NW	C	7.00	non	3	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

1.2 Les différents types de variables

Variable qualitative

- ensemble des modalités : ensemble de valeurs non numériques.
- ne peut pas faire l'objet de mesure.
- codage possible des modalités par un nombre, mais pas d'opérations algébriques possibles sur ces nombres.
- deux catégories :
  - variable qualitative **nominale** : modalités non ordonnées.
    Exemple : sexe, région d'habitation, ...
  - variable qualitative **ordinale** : modalités ordonnées.
    Exemple : niveau d'études.

Variable quantitative

- ensemble des modalités : ensemble de valeurs numériques.
- opérations algébriques possibles sur les observations (moyenne, variance,....).
- deux catégories :
  - variables quantitatives **discrètes** : les modalités appartiennent à un ensemble fini ou dénombrable (le plus souvent $\mathbb{N}$ ou une partie de $\mathbb{N}$).
    Exemple : nombre d'enfants, ...
  - variables quantitatives **continues** : l'ensemble des modalités est $\mathbb{R}$ ou un intervalle de $\mathbb{R}$.
    Exemple : le salaire horaire, l'âge, ...
- Attention, la nature des variables dépend des valeurs possibles de la variable, et non des valeurs mesurées.

  Exemple : l'âge est une variable continue, même s'il est mesuré en mois ou en année.

Exemple : Current Population Survey

Variables : Âge, Sexe, Région d'habitation, Statut Marital, Salaire horaire, Appartenance à un syndicat, Catégorie professionnelle, Niveau d'études, Nombre de personnes/foyer, Nombre d'enfants, Revenu du foyer.

- **variable qualitative nominale** : ....

  ....
- **variable qualitative ordinale** : ....

  ....
- **variable quantitative discrète** : ....

  ....
- **variable quantitative continue** : ....

  ....

2. Tableaux et distributions statistiques

Objectifs

- résumer les données brutes.
- établir la distribution statistique en effectifs/fréquences de X : pour chaque modalité x de la variable X, compter le nombre d'individus de la population pour lesquels X prend la valeur x.

2.1 Variables qualitatives et quantitatives discrètes

- $x_1, x_2, \cdots x_p$ les p valeurs distinctes ou **modalités** de X.
- pour toute modalité x_i, $i = 1, \cdots, p$
  - **effectif** n_i :

$$n_i = \text{nombre d'individus } k = 1, \ldots, n \text{ dans la population tq } x_k^* = x_i.$$

  - **fréquence** f_i :

$$f_i = \frac{n_i}{n} = \text{fréquence d'apparition de la modalité } x_i \text{ dans la population.}$$

- $\displaystyle\sum_{i=1}^{p} n_i = n \quad ; \quad \forall i,\ 0 \leq f_i \leq 1 \quad ; \quad \sum_{i=1}^{p} f_i = 1.$

Tableaux des effectifs et des fréquences

modalités	effectifs	fréquences
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$
x_p	n_p	f_p
Total	n	1

Distribution statistique en effectifs et en fréquences de X : donnée par les colonnes (ou lignes) effectifs et fréquences du tableau statistique

fonctions R : **table** et **prop.table**

Exemple : Current Population Survey

- Population : .... de taille ....
- unité statistique : ....
- variables **sexe** et **appartenance à un syndicat** : type .... $p = \dots$ modalités.

Sexe	effectifs	fréquences (en%)
féminin	297	...
masculin	302	...
Total	599	100

Répartition Femmes/Hommes

Syndiqué	effectifs	fréquences (en%)
oui	103	17.2
non	496	82.8
Total	599	100

Répartition syndiqué/non syndiqué

- le jeu de donnée est constitué de ....% de femmes et de ....% d'hommes.
- 17.2% sont membres d'un syndicat et 82.8 % ne sont pas membres d'un syndicat .

2.2 Variables quantitatives continues

- données brutes $x_1^*, \ldots, x_n^*$ (presque toutes) distinctes.
- calcul d'effectifs et fréquences par **classe**.
- découpage de l'ensemble des valeurs possibles $[e_0, e_p]$ de X, en p classes
$$[e_0, e_1[\,,\ [e_1, e_2[\,, \ldots,\ [e_{p-1}, e_p].$$
- classes contigües et non chevauchantes.
- chaque observation x_k^* rangée dans une unique classe.
- $e_0, e_1, \ldots, e_p$: **extrémités** (ou bornes) des classes.
- a_i : **amplitude** de la i-ème classe $a_i = e_i - e_{i-1}$, $i = 1, \ldots, p$.

Remarques

- classes d'amplitudes égales ou inégales, première classe du type "moins de ", dernière du type "plus de ".
- choix du nombre p de classes délicat, dépend de n et de la dispersion des données.
- grande perte d'information sur la série si p trop petit : négligence des aspects importants de la distribution.
- nombreuses classes vides si p trop grand : rôle exagéré pour les variations accidentelles.
- par défaut, classes d'amplitudes égales dans les logiciels spécialisés.
- nombre suffisant d'individus par classe.

Effectif et fréquence de la classe $[e_{i-1}, e_i[$

$$
\begin{aligned}
n_i &= \text{nombre d'individus pour lesquels } x_k^* \in [e_{i-1}, e_i[\\
f_i &= \frac{n_i}{n} : \text{fréquence de la classe } [e_{i-1}, e_i[\,.
\end{aligned}
$$

$$
\text{avec } \sum_{i=1}^{p} n_i = n \quad ; \quad 0 \leq f_i \leq 1,\ \forall i \in \{1, \ldots, p\} \ ; \quad \sum_{i=1}^{p} f_i = 1\,.
$$

Tableau des effectifs et des fréquences

classes	effectifs	fréquences
$[e_0, e_1[$	n_1	f_1
$[e_1, e_2[$	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
$[e_{i-1}, e_i[$	n_i	f_i
$\vdots$	$\vdots$	$\vdots$
$[e_{p-1}, e_p[$	n_p	f_p
Total	n	1

Exemple : Current Population Survey
Salaire horaire

Découpage de $[0, 100]$ en $p = 9$ classes d'amplitudes inégales.

Classe $[e_{i-1}, e_i[$	amplitude $a_i = e_i - e_{i-1}$	effectifs n_i	fréquences (en %) f_i
$[0, 5[$	5	9	1.5
$[5, 10[$	5	100	16.7
$[10, 15[$	5	186	31.1
$[15, 20[$	5	112	18.7
$[20, 35[$	5	73	12.2
$[25, 30[$	5	47	7.8
$[30, 35[$	5	31	5.2
$[35, 40[$	5	14	2.3
$[40, 100]$	60	27	4.5
Total		$n = 599$	100

fonctions R : **hist** et **prop.table**

Exemple : Current Population Survey
Salaire horaire

Découpage de $[2, 99]$ en $p = 10$ classes ayant des effectifs équilibrés (à partir des déciles).

Classe $[e_{i-1}, e_i[$	amplitude $a_i = e_i - e_{i-1}$	effectifs n_i	fréquences (en %) f_i
$[2, 8[$	6	41	6.8
$[8, 10[$	2	68	11.4
$[10, 11[$	1	52	8.7
$[11, 13[$	2	66	11
$[13, 15[$	2	68	11.4
$[15, 17[$	2	61	10.2
$[17, 20[$	3	51	8.5
$[20, 24[$	4	66	11
$[24, 30[$	6	54	9
$[30, 99]$	69	72	12
Total		$n = 599$	100

Présentation d'un tableau statistique

- Un bon tableau
  - doit contenir un titre explicite.
  - doit comporter les sources le cas échéant.
  - doit être compréhensible dès la lecture.
  - doit contenir des informations pertinentes.
  - doit être cohérent (somme des fréquences égales à 1, ...).
  - doit comporter des chiffres arrondis de manière raisonnable (0.35 plutôt que 0.34726373838).
  - doit être référencé dans le texte le cas échéant.
- Ne jamais hésiter à retoucher un tableau produit par un logiciel.

2.3. Effectifs et fréquences cumulés

Uniquement pour les variables quantitatives.

modalité ou classe	effectifs n_i	effectifs cumulés croissants N_i
x_1 ou $[e_0, e_1[$	n_1	n_1
x_2 ou $[e_1, e_2[$	n_2	$n_1 + n_2$
$\vdots$	$\vdots$	$\vdots$
x_i ou $[e_{i-1}, e_i[$	n_i	$n_1 + \cdots + n_i$
$\vdots$	$\vdots$	$\vdots$
x_{p-1} ou $[e_{p-2}, e_{p-1}[$	n_{p-1}	$n_1 + \cdots + n_{p-1}$
x_p ou $[e_{p-1}, e_p[$	n_p	$\sum_{i=1}^{p} n_i = n$

On peut produire un tableau du même genre pour les fréquences (les fréquences cumulées sont notées F_i).

Exemple : Current Population Survey
Salaire horaire

Classe $[e_{i-1}, e_i[$	fréq. (en %) f_i	fréq. cumulées (en %) F_i
$[0, 5[$	1.5	1.5
$[5, 10[$	16.7	18.2
$[10, 15[$	31.1	49.3
$[15, 20[$	18.7	68
$[20, 35[$	12.2	80.2
$[25, 30[$	7.8	88
$[30, 35[$	5.2	93.2
$[35, 40[$	2.3	95.5
$[40, 100]$	4.5	100
Total	100	

Interprétation : 49.3 % des salariés touchent moins de 15$ de l'heure. 88% touchent moins de 30$ de l'heure.

fonction R : **cumsum**

3 Représentations graphiques
3.1. Variables qualitatives

- diagramme circulaire ou “camembert”.
- diagramme en barres verticales/horizontales.
- pas d’échelle car valeurs non numériques.

Diagramme circulaire ou camembert

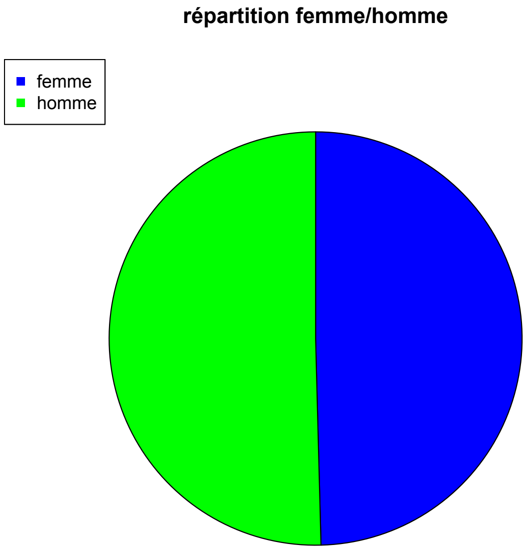
- disque d'aire décomposée en secteurs circulaires représentant respectivement la part de chaque modalité x_i (angle au centre α_i)

$$\alpha_i = 360 \times f_i$$
$$\alpha_1 + \alpha_2 + \cdots + \alpha_p = 360\,.$$

- l'information pertinente est **l'aire** de chaque secteur. Ne jamais tracer de diagramme circulaire en 3D.
- ne pas utiliser cette représentation lorsque la variable possède beaucoup de modalités.
- pas de sens pour une variable ordinale (modalités ordonnées).

Exemple : Current Population Survey

fonction R : **pie**



Exemple : Current Population Survey

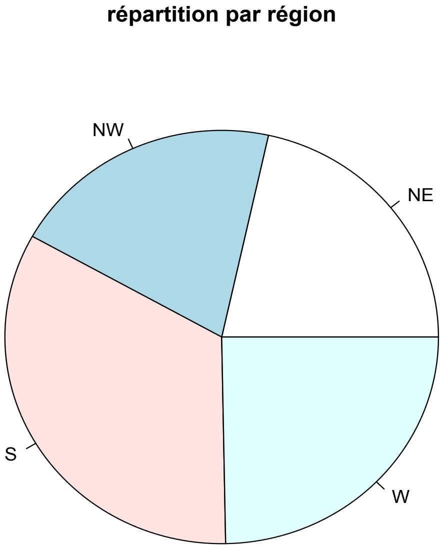
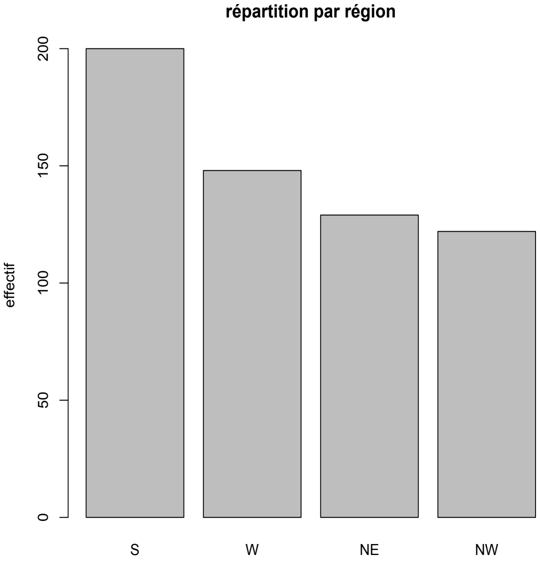


Diagramme en barres

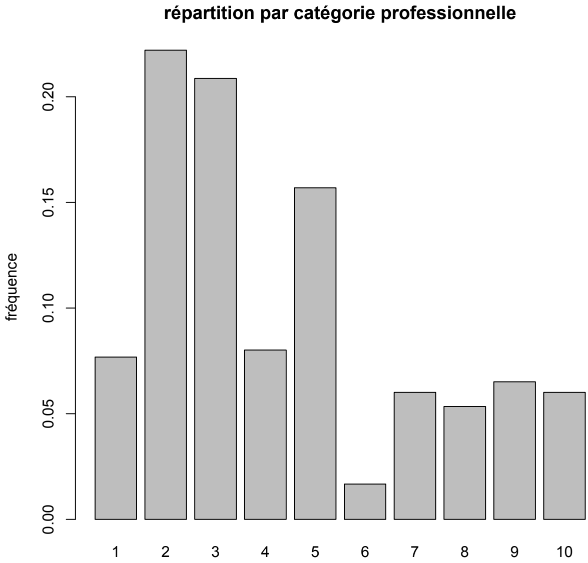
- diagramme en barres verticales ou horizontales.
- surfaces des barres égales ou proportionnelles aux effectifs ou aux fréquences des modalités.
- conserver l'ordre naturel des modalités pour les variables ordinales.
- tri possible des modalités selon leurs fréquences pour des variables nominales.

Exemple : Current Population Survey

fonctions R : **barplot** et **sort**



Exemple : Current Population Survey

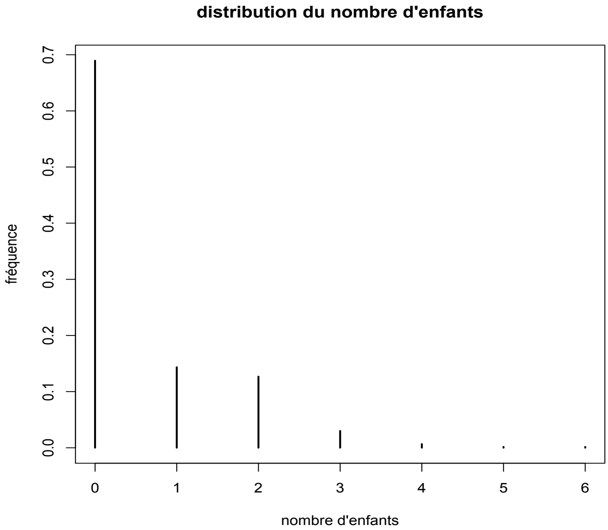


3.2 Variables quantitatives discrètes
Diagramme en bâtons

- fréquences (effectifs) des x_i représentées par des bâtons (ou barres selon le logiciel) verticales de hauteur proportionnelle à f_i (ou n_i).
- les hauteurs des bâtons, mais aussi l'écart et l'ordre des modalités x_i ont un sens.

Exemple : Current Population Survey

fonctions R : **plot** et **prop.table**



3.3 Variable quantitative continue

- histogramme (densité de fréquence).
- boîte à moustaches ou box plot (cf. section 4).
- fonction de répartition empirique/courbe des fréquences cumulées.

Histogramme

- représentation graphique du tableau des fréquences des classes $[e_{i-1}, e_i[$, avec $i = 1, \ldots, p$.

- représentation de chaque classe par un rectangle d'**aire**, et non de hauteur, proportionnelle à l'effectif ou la fréquence.

- comparaison impossible des fréquences si les amplitudes des classes sont inégales.

- notion de **densité de fréquence**.

Pour la classe $[e_{i-1}, e_i[$, ou $i = 1, \ldots, p$,

- $a_i = e_i - e_{i-1} =$ base du rectangle représentant la classe i.
- $d_i = \dfrac{f_i}{a_i}$ **densité** de fréquence de la classe i.
- la hauteur du rectangle représentant la classe i est égale à d_i.
- **l' aire du rectangle est égale à f_i**

$$aire = a_i \times d_i = \; f_i \, .$$

- représentation simplifiée de la distribution de X en supposant que la distribution est uniforme dans chaque classe.
- **interprétation** : un intervalle de longueur 1 inclus dans la classe $[e_{i-1}, e_i[$ a pour fréquence approchée d_i.

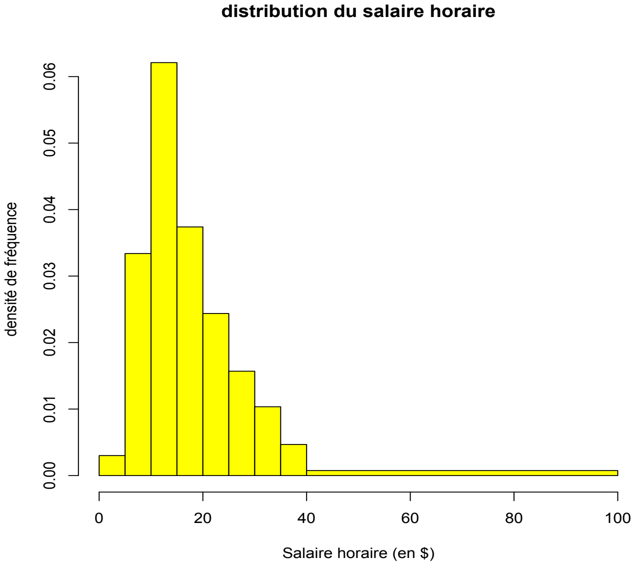
Exemple : Current Population Survey
Salaire horaire

Découpage de $[0, 100]$ en $p = 9$ classes d'amplitudes inégales.

Classe $[e_{i-1}, e_i[$	amplitude $a_i = e_i - e_{i-1}$	effectifs n_i	fréq. $f_i \times 100$	densité de fréq. $d_i \times 100$
$[0, 5[$	5	9	1.5	0.17
$[5, 10[$	5	100	16.7	3.34
$[10, 15[$	5	186	31.1	6.22
$[15, 20[$	5	112	18.7	3.74
$[20, 35[$	5	73	12.2	2.44
$[25, 30[$	5	47	7.8	1.56
$[30, 35[$	5	31	5.2	1.04
$[35, 40[$	5	14	2.3	0.46
$[40, 100]$	60	27	4.5	0.07
Total		$n = 599$	100	

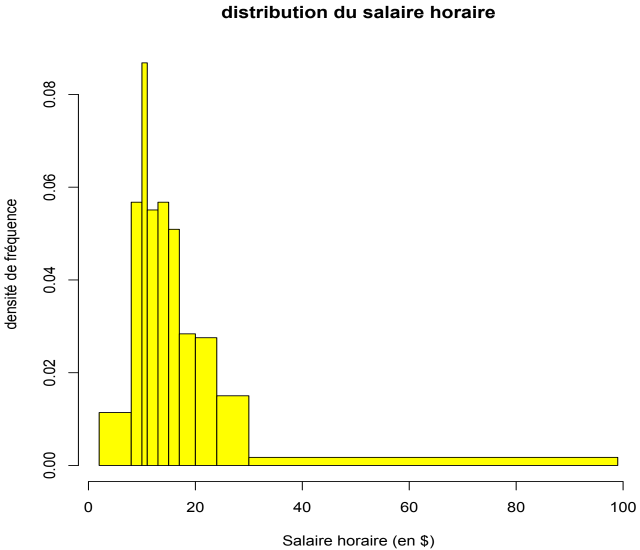
Exemple : Current Population Survey, suite

fonction R : **hist**



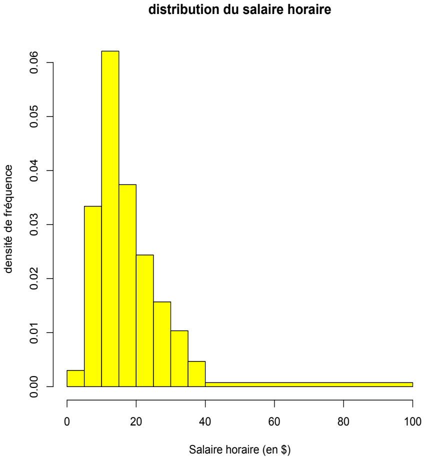
Exemple : Current Population Survey, suite

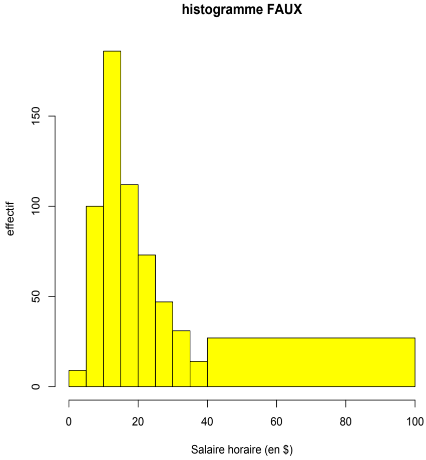
Découpage de $[2, 99]$ en $p = 10$ classes ayant des effectifs équilibrés.



Attention

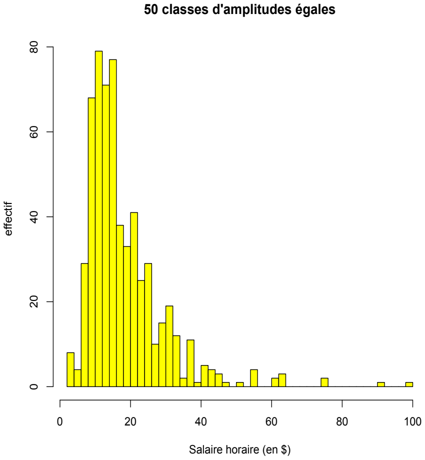
Ne **jamais** tracer d'histogramme en effectifs/fréquences si les amplitudes sont inégales.

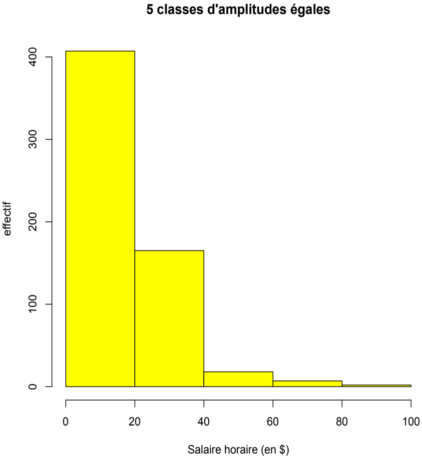




Attention au choix des classes

Trop ou **trop peu** de classes $\rightarrow$ mauvaise visualisation de la distribution.
La règle par défaut sous R n'est pas forcément pertinente.





Fonction de répartition empirique

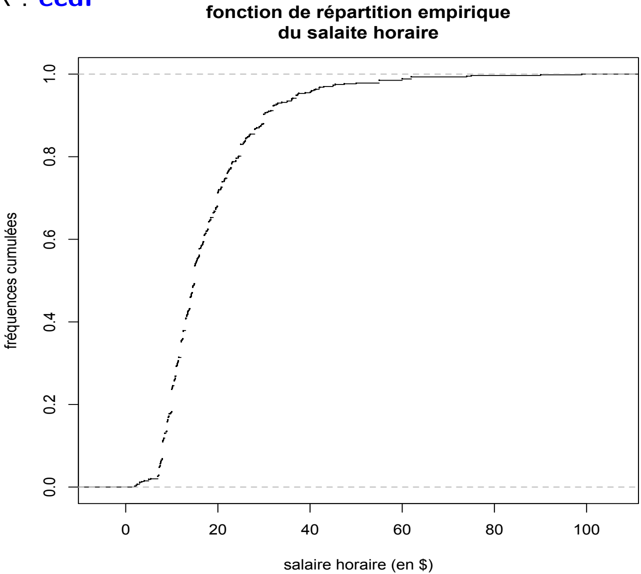
- représentation des fréquences cumulées.
- tracé de F_n, **fonction de répartition empirique** de X.
- pour tout réel x, $F_n(x)$ est la proportion d'individus pour lesquels X est inférieur ou égal à x.

$$\begin{array}{llll} F_n : & \mathbb{R} & \to [0,1] \\ & x & \mapsto F_n(x) = \dfrac{1}{n} Card\{k = 1, \ldots, n \quad x_k \leq x\}. \end{array}$$

- F_n : fonction en "escalier" (constante par morceaux), continue à droite, croissante de 0 à 1.
- sauts à chaque passage par les valeurs x_k prises par X.

Exemple : Current Population Survey

fonction R : **ecdf**



Courbe des fréquences cumulées

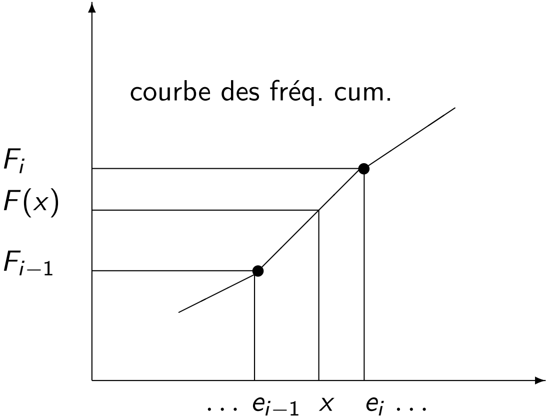
- représentation graphique du tableau des fréquences cumulées des classes $[e_{i-1}, e_i[$, avec $i = 1, \ldots, p$.

- on approche F_n par la fonction F définie aux bornes des classes par

$$F(bleue_0) = 0 \quad ; F(e_1) = f_1 \quad ; F(e_i) = F_i \quad ; \quad F(e_p) = 1.$$

- hypothèse de répartition uniforme dans les classes.

- **interpolation linéaire** de F_n.



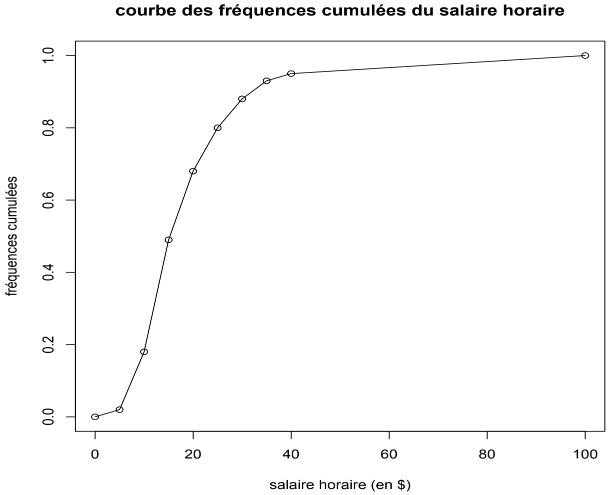
Interpolation linéaire :

$$\text{pour } x \in [e_{i-1}, e_i[, \quad F(x) = F_{i-1} + \frac{f_i}{a_i}(x - e_{i-1}).$$

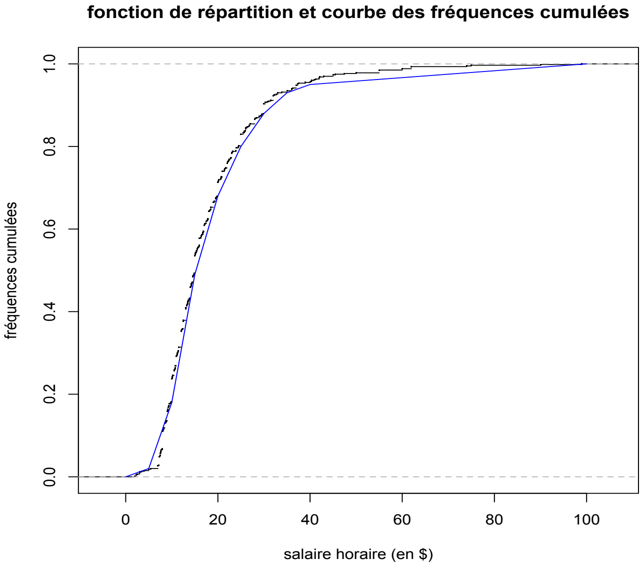
Exemple : Current Population Survey

fonctions R : **cumsum**, **prop.table** et **plot**.

Découpage de $[0, 100]$ en $p = 9$ classes d'amplitudes inégales.

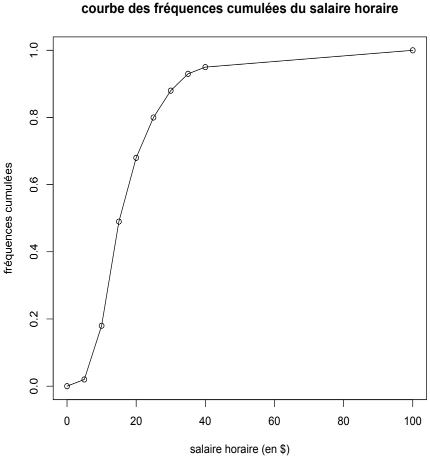


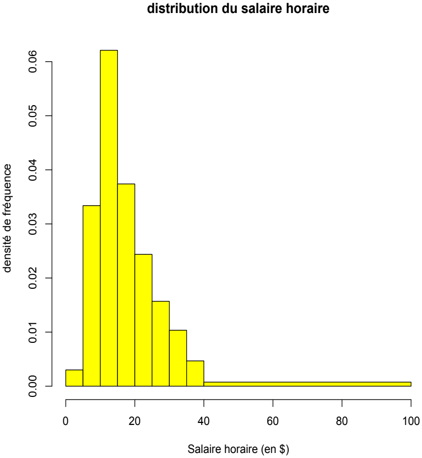
Visualisation de l'approximation



Lien avec l'histogramme

L'histogramme est la fonction dérivée de la courbe des fréquences cumulées.





Conclusion sur les objectifs de l'analyse graphique

- visualisation graphique des distributions statistiques en effectifs/fréquences de la variable.

- obtenir des indications de forme (symétrie,...), de concentration, de dispersion,... de la distribution observée de la variable.

- choix d'une représentation graphique en fonction
  - du type de la variable étudiée.
  - du problème posé.

- indications pour une étape ultérieure de modélisation des données.

Quelques remarques pour une bonne présentation de graphiques

- lisibilité du graphique
  - un titre clair, concis et complet.
  - une légende pour chaque axe en indiquant le cas échéant le nom de la variable, son unité de mesure, si la distribution est en effectifs, en fréquences, ...
  - choix des couleurs ... Impression en noir et blanc ...
  - graphique "auto-suffisant" : contient les informations nécessaires à sa compréhension par le lecteur ou l'auditeur.
- représentations graphiques en 3D à proscrire.
- représentation simple ... mais adaptée et commentée.

- choix de l'échelle primordial

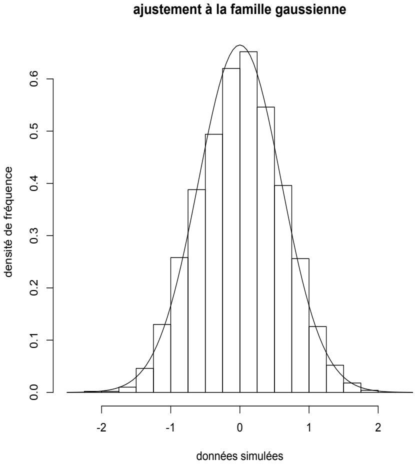
  - mise en évidence de caractéristiques de la distribution purement artificielles pour une échelle non adaptée.

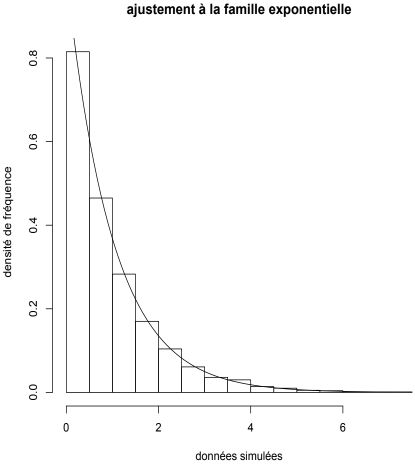
  - **Règle** : le graphique doit tenir dans un carré (ou un rectangle un peu allongé).

- utiliser la même échelle pour comparer graphiquement deux ou plusieurs distributions.

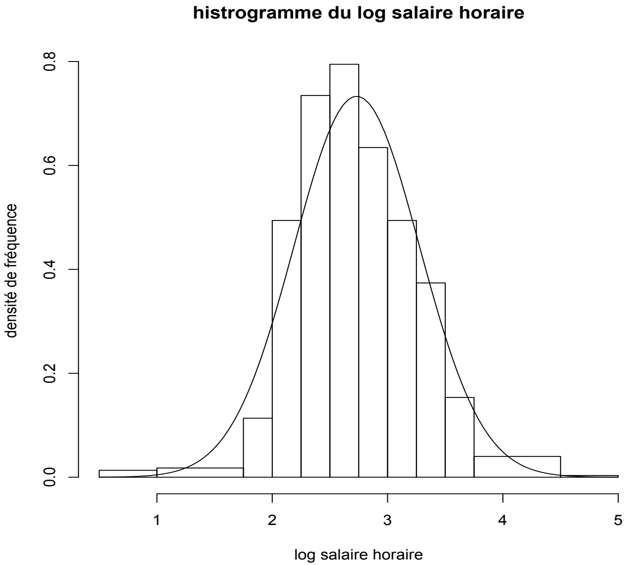
Vers la modélisation des données

Comparaison de la **forme de la distribution observée** à des formes connues de **distributions théoriques**.





Exemple : Current Population Survey



4. Indicateurs statistiques

Objectifs

- résumer des données en quelques valeurs numériques.
- comparer plusieurs séries statistiques.
- UNIQUEMENT pour des variables **quantitatives** (excepté le mode).
- Deux types de caractéristiques
  - indicateurs de **tendance centrale et de position** : mode, médiane, moyenne et quantiles.
  - indicateurs de **dispersion** : variance, écart-type, écart inter-quartile.

4.1. Indicateurs de tendance centrale et de position
Le mode ou la classe modale

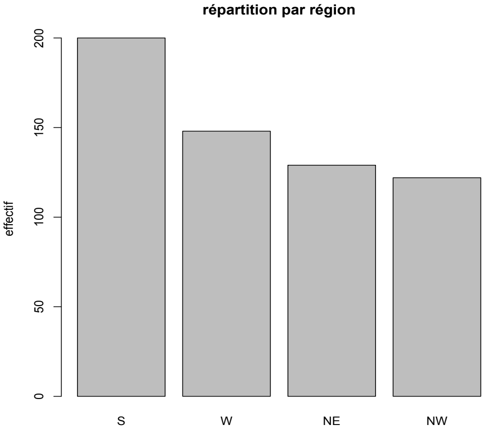
Définition

- le **mode** de la distribution d'une variable **quantitative discrète (ou qualitative)**, est la valeur la plus fréquemment observée, c'est à dire celle d'effectif (ou fréquence) le plus élevé.
- il est repérable sur le diagramme en bâtons (ou en barres) ou sur tableau des effectifs (ou fréquences).

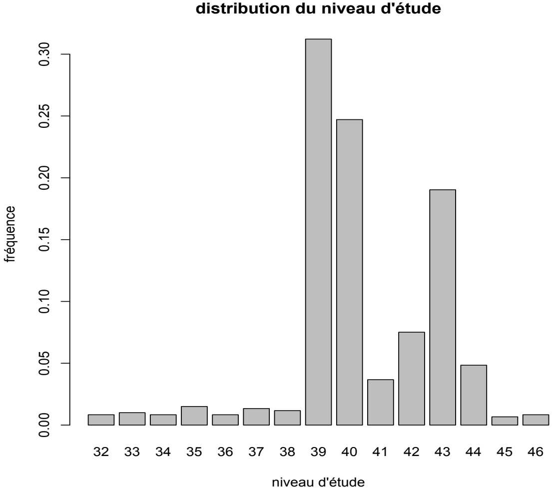
- le mode (ou la classe modale) peut ne pas être unique.
- distribution **unimodale** : un seul maximum marqué dans les diagrammes en bâtons (ou les histogrammes).
- **multimodale** : plusieurs maxima relatifs.

Exemple : Current Population Survey

Région	fréq. (en %)
NE	21.5
NW	20.4
S	**33.4**
W	24.7



Exemple : Current Population Survey
Une distribution multimodale



La classe modale

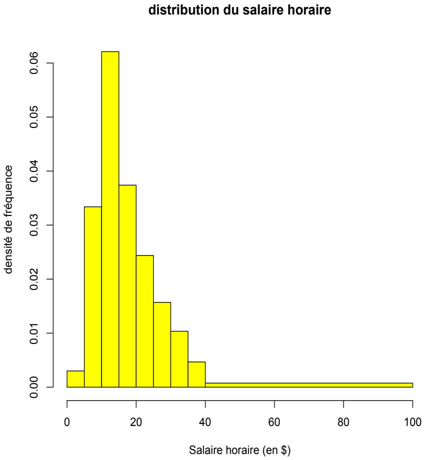
Définition

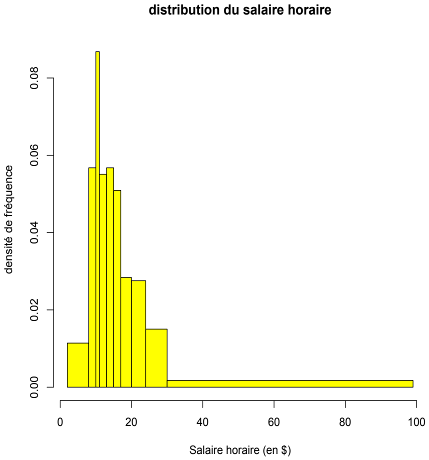
- La **classe modale** de la distribution d'une variable **quantitative continue**, est la classe de densité de fréquence (ou d'effectif) la plus élevée.
- C'est la classe de hauteur maximale dans l'histogramme.
- C'est la classe de fréquence la plus élevée **si les amplitudes des classes sont égales**.

Remarque : la classe modale change selon le découpage en classes.

Interprétation : Une sous-classe de longueur ℓ est plus fréquente si elle est incluse dans la classe modale que si elle est incluse dans n'importe quelle autre classe.

Exemple : Current Population Survey





Classe modale : $[10, 15[$

Classe modale : $[10, 11[$

La médiane

- Intuitivement, la médiane est la valeur de la variable quantitative pour laquelle il y a autant d'observations supérieures à cette valeur que d'observations inférieures.
- Le plus souvent une telle quantité n'existe pas. On donne alors la définition plus précise suviante :

Définition

Une médiane, M_e, de la distribution d'une variable quantitative est une valeur de la variable pour laquelle il y a plus de ($\geq$) 50% d'observations supérieures ou égales à M_e et plus de ($\geq$) 50% d'observations inférieures ou égales à M_e.

Détermination de la médiane

- ranger les observations par ordre croissant (arbitrairement si répétition).
- si n impair, de la forme $n = 2k + 1$, $M_e = x_{k+1}$.
  Exemple :
  - 15, 3, 16, 10, 15, 18, 5
  - tri des données : 3, 5, 10, 15, 15, 16, 18 $\Rightarrow M_e = \ldots\ldots\ldots$
- si n pair, de la forme $n = 2k$,
  - pour une variable **discrète**, $M_e = x_k$ **ou** x_{k+1}.
    Exemple
    - ★ 3, 5, 10, 10, 15, 16 $\Rightarrow M_e = \ldots\ldots$
  - pour une variable **continue**, $M_e \in [x_k, x_{k+1}]$.
    Exemples
    - ★ 3.5, 5.1, 10, 15.1, 15.3, 16 $\Rightarrow M_e \in \ldots\ldots\ldots$
    - ★ 3.5, 5.1, 10, 10, 15.3, 16 $\Rightarrow M_e = \ldots\ldots\ldots$
- tous les logiciels ne font pas le même choix !

Propriétés de la médiane

- **robustesse de la médiane** : peu sensible aux valeurs extrêmes.
- M_e comprise entre l'observation la plus petite et l'observation la plus grande.
- linéarité de la médiane : si $y_k = ax_k + b$, alors $aM_e(X) + b$ est une médiane de Y.
- Plus généralement : si $y_k = f(x_k)$ et f est monotone, alors $f(M_e(X))$ est une médiane de Y.

Les quantiles

Définition

Soit $\alpha \in]0,1[$. Un quantile d'ordre α est une valeur Q_α telle qu'il y ait une proportion supérieure à ($\geq$) α d'observations inférieures à Q_α et une proportion supérieure à ($\geq$) $1-\alpha$ d'observations supérieures à Q_α.

À partir de la fonction de répartition :

- Le plus petit x_i tel que $F_n(x_i) \geq \alpha$ est un quantile Q_α.
- si $F_n(x_i) < \alpha$ et $F_n(x_{i+1}) \geq \alpha$, alors x_{i+1} est un quantile Q_α.

fonctions R : **median** et **quantile**

Remarque

- **Les quartiles** : quantiles $Q_{1/4}$, $Q_{1/2}$ et $Q_{3/4}$.
- un quartile $Q_{1/2}$ est une médiane M_e.
- les quantiles sont des **indicateurs de position** :
  - Les 3 **quartiles** subdivisent la série en 4 intervalles contenant (à peu près) le même nombre d'observations chacun ($\sim$ 25% chacun).
  - Les 9 **déciles** subdivisent la série en 10 intervalles contenant (à peu près) le même nombre d'observations chacun ($\sim$ 10% chacun).
  - Les 99 **centiles** subdivisent la série en 100 intervalles contenant (à peu près) le même nombre d'observations chacun ($\sim$ 1% chacun).

Exemple : pour le salaire horaire (en $), R fournit :

$$Q_{1/4} = 10.5,\ M_e = Q_{1/2} = 15,\ Q_{3/4} = 22\,.$$

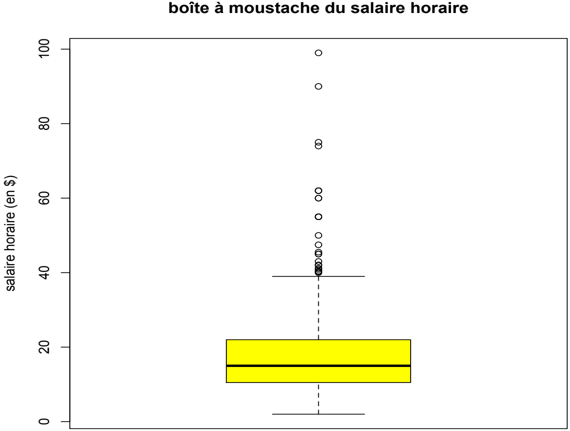
Interprétation : 50% des individus touchent moins de 15$ de l'heure. 50% des individus touchent entre 10.5$ et 22$ de l'heure. 25% des individus touchent moins de 10.5$ de l'heure.

La boîte à moustaches ou box-plot

- sur un axe gradué (horizontal ou vertical), le minimum, le maximum et les quartiles.
- tracé d'un rectangle parallèlement à l'axe entre $Q_{1/4}$ et $Q_{3/4}$.
- On peut prolonger la boîte par des moustaches de longueur $1,5 \times (Q_{3/4} - Q_{1/4})$.
- les observations extérieures aux moustaches sont dites statistiquement "aberrantes" et repérées par des $\circ$
- $1,5$: critère arbitraire, ordre de grandeur raisonnable.
- selon les logiciels, moustaches raccourcies aux observations adjacentes aux moustaches.
- moustaches raccourcies au min et max si aucune observation n'arrive au delà des moustaches.
- examen attentif des individus "hors norme" d'un point de vue statistique, pas d'attitude systématique.

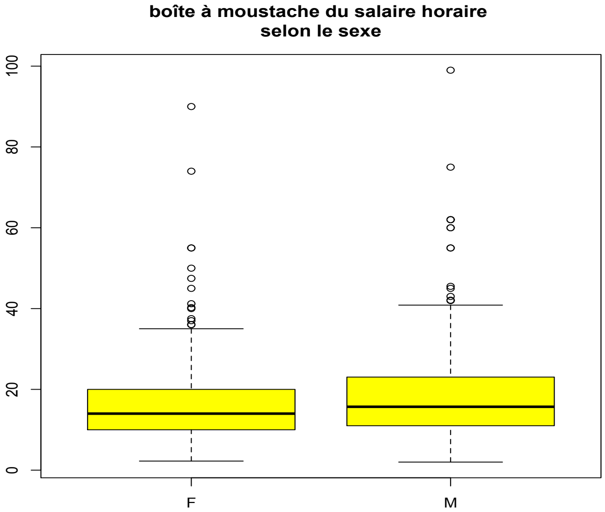
Exemple : Current Population Survey

fonction R : **boxplot**



- la moustache inférieure est raccourcie au minimum.
- présence de nombreuses valeurs statistiquement aberrantes ("très gros salaires").

Comparaison de la distribution de X sur deux sous-populations



La moyenne arithmétique

Définition

La **moyenne arithmétique** est la quantité définie

- à partir des données brutes, par

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \, .$$

- à partir des fréquences des modalités (cas des variables discrètes), par

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i \, .$$

La moyenne est linéaire :
soient $a, b \in \mathbb{R}$, si $y_k = a x_k + b$, alors $\bar{y} = a\bar{x} + b$.

fonction R : **mean**

Moyenne des moyennes

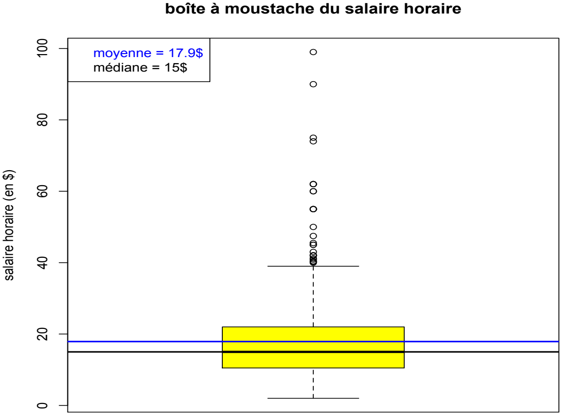
La moyenne arithmétique de deux séries d'effectifs n_1 et n_2 et de moyennes $\bar{x}_1$ et $\bar{x}_2$ est

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}\,.$$

Exemple : le salaire horaire moyen des 297 femmes est de $\bar{x}_F = 16.6\$$. le salaire horaire moyen des 302 hommes est de $\bar{x}_H = 19.17\$$. Le salaire horaire moyen sur l'ensemble de la population est de

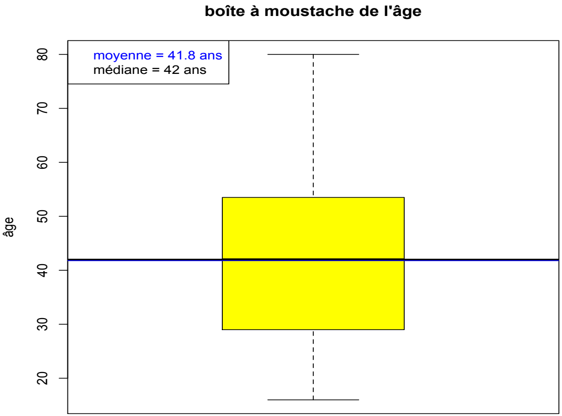
$$\bar{x} = \frac{297\bar{x}_F + 302\bar{x}_H}{599} = 17.9\$$$

Exemple : Current Population Survey
Moyenne et médiane du salaire horaire



Distribution asymétrique : la moyenne est très différente de la médiane.
Quelques gros salaires "tirent la moyenne vers le haut".

Exemple : Current Population Survey
Moyenne et médiane de l'âge

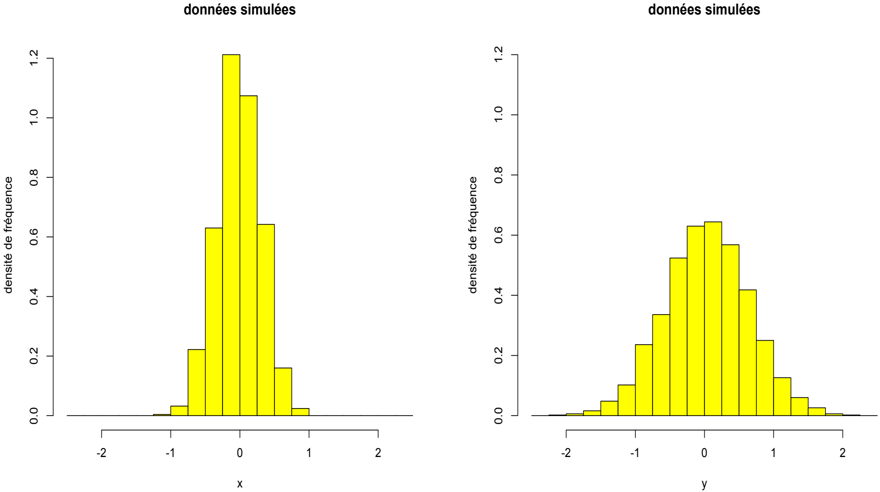


Distribution quasi symétrique : la moyenne et la médiane sont presque identiques.

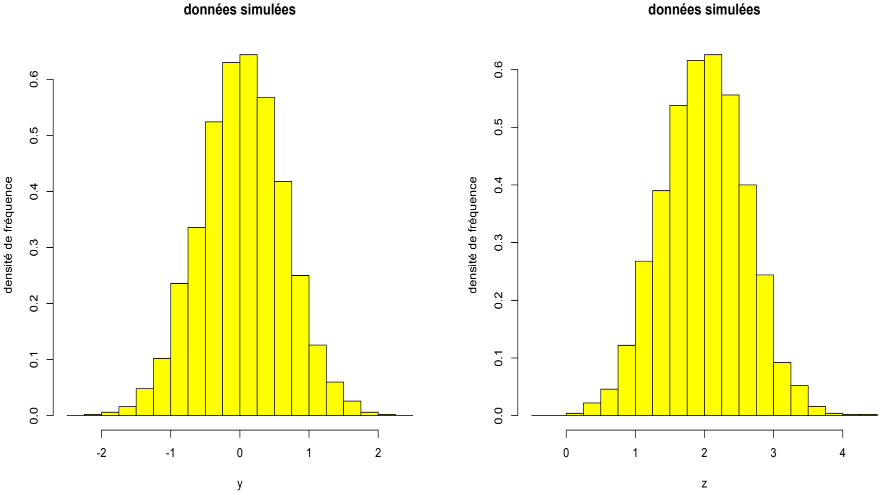
4.2. Indicateurs de dispersion

- information sur la variabilité des observations autour d'une valeur centrale.
- **UNIQUEMENT** pour des variables **quantitatives**.
- indicateur d'autant plus grand que la variable est dispersée (grande variabilité autour d'une caractéristique de tendance centrale).
- indicateur toujours positif.

Exemple : moyennes égales, écart-types différents



Exemple : moyennes différentes, écart-types égaux



La variance et l'écart-type

Définition

La **variance** est définie

- à partir des données brutes par

$$\mathbb{V}(x) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 .$$

- à partir des fréquences des modalités (cas des variables discrètes), par

$$\mathbb{V}(x) = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2 .$$

L'**écart-type** est la racine carrée de la variance :

$$\sigma_x = \sqrt{\mathbb{V}(x)}$$

- $x_k - \bar{x}$ est l'**écart à la moyenne** de x_k.
- $\mathbb{V}(x)$ est la moyenne des carrés des écarts à la moyenne.
- $\mathbb{V}(x)$ est d'autant plus faible que les données sont groupées autour de la moyenne.
- σ_x exprimé dans la même unité que les données x_i.
- Exemple : X taille en cm, $\bar{x}$ exprimée en cm, σ_x^2 en cm^2 et σ_x en cm.
- les variances et les écart-types des logiciels (par exemple sour R) sont le plus souvent calculés avec la formule

$$\mathbb{V}(x) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2 \,.$$

- fonctions R : **var** et **sd**

Propriétés de la variance

On a l'égalité :

- pour les données brutes

$$\mathbb{V}(x) = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - (\bar{x})^2 .$$

- à partir des fréquences des modalités (cas des variables discrètes),

$$\mathbb{V}(x) = \underbrace{\frac{1}{n}\sum_{i=1}^{p} n_i x_i^2}_{\text{moyenne des } x_i^2} - \underbrace{\bar{x}^2}_{\left(\text{moyenne des } x_i\right)^2} = \sum_{i=1}^{p} f_i x_i^2 - (\bar{x})^2 .$$

- **La variance n'est pas linéaire** : soient $a, b \in \mathbb{R}$, si $y_k = ax_k + b$, alors
$$\mathbb{V}(y) = a^2\mathbb{V}(x) \qquad \text{et} \qquad \sigma_y = |a|\sigma_x.$$
- **variable centrée réduite** : si $(x_k, k = 1, \ldots, n)$ a pour moyenne $\bar{x}$ et pour variance σ_x^2, alors la série définie par
$$\left(y_k = \frac{x_k - \bar{x}}{\sigma_x}, k = 1, \ldots, n\right)$$
a pour moyenne $\bar{y} = 0$ et variance $\sigma_y^2 = 1$. On dit que Y est une **variable centrée réduite**.
- une valeur centrée réduite s'appelle un **z-score** (score normalisé).

Intérêt des variables centrées réduites

- **situation d'un individu dans la population**
  - exemple : X note math au bac, $\bar{x} = 13.21$, $\sigma = 3.19$.
  - si $x_k = 12$, $x_k - \bar{x} = -1.21$, donc élève k en-dessous de la moyenne de l'ensemble des élèves (et même à plus d'1 point).
  - variable centrée réduite : l'échelle de référence (unité de mesure) est l'écart-type.
  - $x_k = 12$, $\frac{x_k - \bar{x}}{\sigma} = -0.38$. La note est inférieure à la moyenne générale (valeur centrée négative), mais assez proche de la moyenne compte tenu de la dispersion des notes (écart inférieur à moins d'un écart-type).
- **détection de valeur "anormalement" grandes ou petites**
  - en math, $\bar{x}_{math} = 13.21$, $\sigma_{math} = 3.19$.
  - en philo, $\bar{x}_{philo} = 7.84$, $\sigma_{philo} = 3.29$.
  - valeur centrée réduite associée à une note de 18 :
    - en math, $\frac{18-13.21}{3.19} = 1.5$ et en philo, $\frac{18-7.84}{3.29} = 3.09$.
    - 18 en philo est plus "remarquable" que 18 en math.
    - 18 en math au bac est "aussi exceptionnel" que 13 en philo : $7.84 + 1.5 * 3.29 = 12.77$.

L'écart inter-quartile

Définition

- L'**intervalle inter-quartile** est l'intervalle $[Q_{1/4}, Q_{3/4}]$.
- Il contient au moins 50% des observations les plus médianes :

  au moins 50% des observations appartiennent à $[Q_{1/4}, Q_{3/4}]$, et au moins 50 % n'appartiennent pas à $]Q_{1/4}, Q_{3/4}[$;
- L'**écart inter-quartile** est la longueur de l'intervalle inter-quartile : $Q_{3/4} - Q_{1/4}$.
- Si l'écart est petit, au moins 50% des observations se trouvent dans ce "petit" intervalle.
- 50% au moins des observations sont en dehors de $]Q_{1/4}, Q_{3/4}[$; plus l'écart inter-quartile est grand, plus il existe des valeurs éloignées de la médiane.

Exemple : médianes égales, dispersions différentes