Préparation des données

Année universitaire 2020/2021

I. Vue Globale/Compréhension des données

I. Transformations des données

I. Données Aberrantes/Manquantes/Spéciales

Vue Globale/Compréhension des données

Les structures des données:

Les données à analyser sont organisées sous deux formes différentes:

Données structurées: toutes informations (mots, chiffres…) présentées dans des cases (les champs d'un tableau) qui permettent leur interprétation et leur traitement. Plus simplement un tableau individus × variable.

Données non structurées: out ce qui n'est pas organisé sous forme d'un tableau de données la messagerie, les images, les vidéos, etc…

Concepts de base

Population : ensemble des unités statistiques observées.
Individu : unité statistique de base ou élément de la population étudiée.
Caractère ou variable : aspect particulier de l'individu auquel on s'intéresse et qui pourra prendre différentes valeurs selon l'individu concerné.

Variables Statistiques

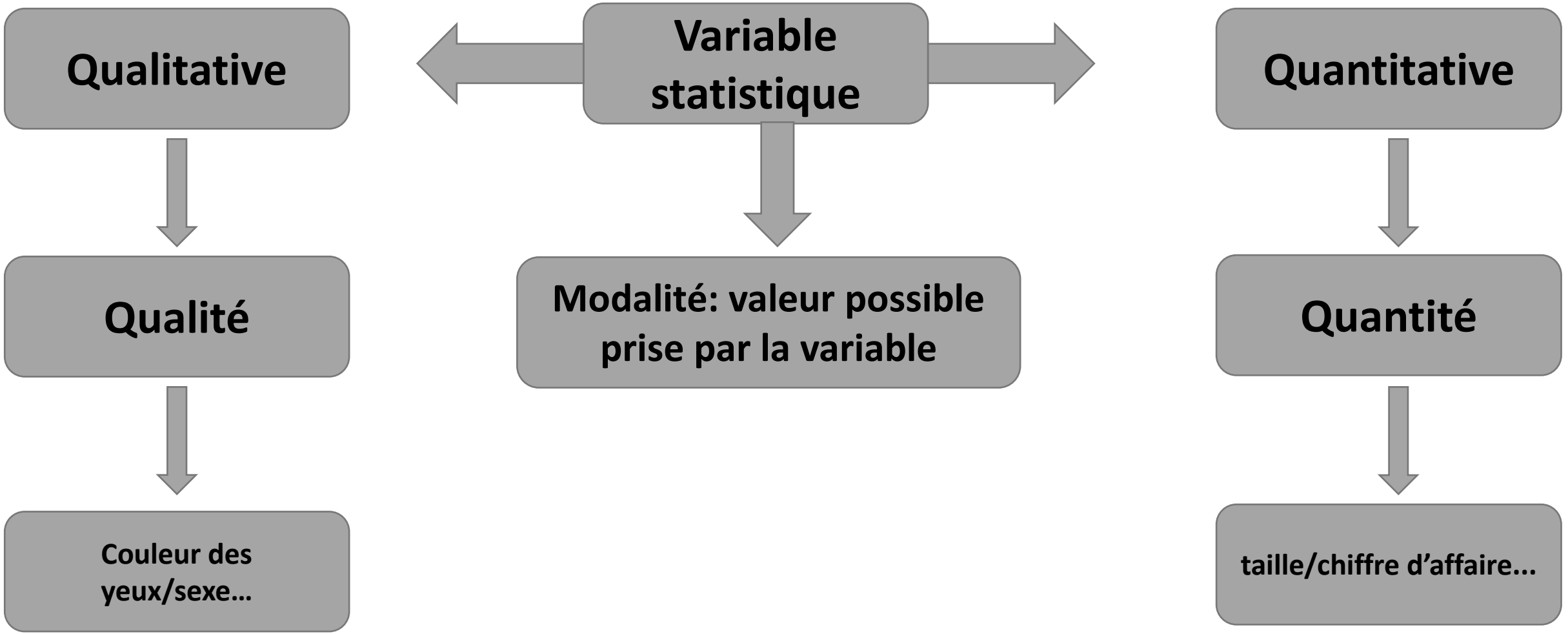
individus Statistiques

N°	Poids	Nombre de frères et sœurs	Situation familiale	Force	Nationalité	Sexe	Date	Heure
1	41,5	1	Célibataire	Faible	Indien	Homme	31/10/1982	10:12
2	33,4	1	Célibataire	Normal	Suisse	Femme	31/10/1983	15:14
3	37,5	2	Célibataire	Fort	Sénégalais	Femme	31/12/1983	14:48
4	33,5	1	Divorcé	Faible	Australien	Femme	01/01/1984	20:15
5	39,7	2	Divorcé	Normal	Birman	Homme	02/01/1984	06:02
6	30,8	1	Marié	Fort	Belge	Homme	31/10/2010	05:45
7	37,4	3	Marié	Très fort	Portugais	Femme	21/10/1983	14:47
8	38,2	1	Marié	Invincible	Brésilien	Homme	02/10/1956	21:14
9	43	3	Veuf	Fort	Russe	Femme	14/09/1929	14:54
10	38,5	2	Veuf	Normal	Serbe	Homme	27/08/1902	02:12

Observation Statistique

Sexe	Effectifs n
Homme	5
Femme	5
Total	10

Types de variables statistiques



Variables qualitative: une variable dont les modalités sont des mots

Nominale

Les modalités ne peuvent pas être ordonnées selon leur sens

état civil: célibataire, marié, divorcé, veuf

ordinale

Les modalités sont ordonnées selon leur sens

Degré de satisfaction: très satisfait, satisfait, insatisfait

Variables quantitative: une variable dont les modalités sont des nombres.

Discrète

Les modalités sont des nombres précis, isolés et dans la plupart du temps des entiers

Nombre d'années de scolarité achevées: 9,10,11,12,13…

Continue

Les modalités sont des nombres issues d'un intervalle de nombres réels

Tranche d'âge: [10,20[,[20,30[,[30,40[..

Choix des techniques statistiques

Les différents types de variables vont conditionner le choix des techniques, des paramètres et des tests utilisés.
Variable quantitative: tous les indicateurs numériques (moyenne, variance, corrélation…)la régression linéaire-ACP…

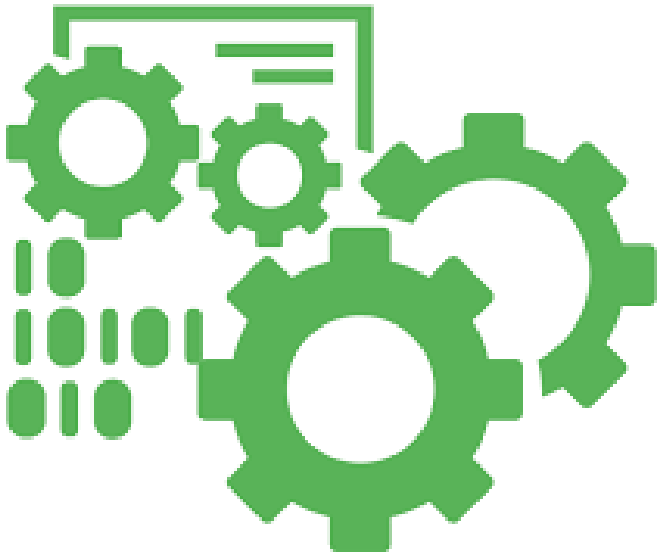
Variable qualitative: test de khi-deux-ACM…

A vous maintenant

N°	Poids	Nombre de frères et sœurs	Situation familiale	Force	Nationalité	Sexe
1	41,5	1	Célibataire	Faible	Indien	Homme
2	33,4	1	Célibataire	Normal	Suisse	Femme
3	37,5	2	Célibataire	Fort	Sénégalais	Femme
4	33,5	1	Divorcé	Faible	Australien	Femme
5	39,7	2	Divorcé	Normal	Birman	Homme
6	30,8	1	Marié	Fort	Belge	Homme
7	37,4	3	Marié	Très fort	Portugais	Femme
8	38,2	1	Marié	Invincible	Brésilien	Homme
9	43	3	Veuf	Fort	Russe	Femme
10	38,5	2	Veuf	Normal	Serbe	Homme

1. Rentrer les valeur du tableau dans une structure R la plus appropriée.
2. Retrouver le type de variable de chaque colonne du tableau suivant.

Transformations des données



1. Codage des variables

Le codage des données consiste à classer en catégories numériques les diverses modalités d'une variable statistique.

Etapes:

- Etablir un code spécifique pour désigner chaque variable ou modalité
- Réfléchir aux types d'échelles auxquelles on a à faire
- Créer le tableau de données ou matrice générale.

Codage des variables qualitatives

Certains traitements et analyses sur des variables qualitatives nécessitent que ces dernières présentent une forme « **pseudo quantitative** » en lieu et place de leur forme « **nominale** ».

C'est notamment le cas lorsqu'il s'agit d'utiliser des variables qualitatives dans un traitement multivarié (faire une régression linéaire etc…)ou simplement lorsque l'on désire les rendre manipulables et compatibles avec des logiciels statistiques.

Les types de codage:

Les échelles nominales

Elles correspondent à des valeurs qualitatives nominales (masculin/féminin). Les modalités de réponses sont une série de catégories, de classes non hiérarchisées.
La mesure utilisée est une valeur arbitraire qui correspond au fait de posséder ou non la propriété

Exemples

Sexe: Masculin Féminin
Codage: 1 pour masculin et 2 pour féminin
Etes vous pour l'euthanasie?
Codage: 1 pour « oui » et 0 pour « non ».

Les types de codage:

Les échelles ordinales

Les modalités de réponses sont des catégories ordonnées logiquement

Ces catégories sont représentées par des valeurs numériques ou des lettres, non arbitraires et ordonnées.

Exemple:

Comment envisagez vous votre avenir professionnel?

Très précis	Précis	Assez précis	Peu précis	Imprécis
1	2	3	4	5

Les échelles d'opinion, d'attitudes (Likert) sont les plus connues des échelles ordinales.

Codage des variables quantitatives

Les échelles d'intervalles

Les modalités de réponses sont des catégories, des classes ordonnées logiquement et qui présentent en plus des distances, des intervalles clairement définis entre les catégories.

Il s'agit de variables quantitatives (ou numériques).

2.Normalisation des données

- Il existe plusieurs transformations utilisées. Par exemple l'analyse centrée qui consiste à modifier les données d'un tableau **X** en remplaçant les valeurs des x_{ik} par des valeurs dans l'intervalle 0 et 1 pour ne pas privilégier les attributs ayant les plus grands domaines de variation (par exemple salaire/age)

- L'analyse **centrée réduite** ou encore **normée** est liée à la transformation des données du tableau **X** en remplaçant les valeurs des x_{ik} par **(x_{ik} – moyenne)/écart type).**

- Réduire les données permet d'uniformiser les unités de mesures.

2.Normalisation des données (Sous R)

scale(): permet de normaliser par colonne les valeurs numériques dans une matrice ou un data frame.

Exemple: Standardisons les valeurs du jeu de donnes Data avec la fonction **scale().**

```
Data<-matrix(c(1,0,3,0,5,0,3,2,5,3,7,2,4,9,0),3,5)
Data_standard <- scale(Data)
```

Les valeurs dans chacune des colonnes de Data_standard ont une moyenne de 0 et un écart-type de 1. Elles sont donc sur la même échelle.

2.Normalisation des données (Sous R)

Les arguments center et scale de la fonction scale() permettent de contrôler, respectivement, les valeurs soustraites par colonnes et les valeurs par lesquelles les colonnes sont divisées. Il est possible, par exemple, de centrer sans réduire avec:

center = TRUE et scale = FALSE.

2.Normalisation des données

Un autre exemple de normalisation possible avec scale est de ramener les mesures de toutes les variables entre 0 et 1. En notation statistique, la formule pour obtenir cette normalisation est la suivante**:**

(xi-min(x))/(max(x)-min(x))

Exemple: effectuons cette normalisation sur le jeu de données **Data**.

```
Data<-matrix(c(1,0,3,0,5,0,3,2,5,3,7,2,4,9,0),3,5)
mmums <- apply(Data, MARGIN = 2, FUN = min)  # obtention des minimums par variable
mmums <- apply(Data, MARGIN = 2, FUN = max)  # obtention des maximums par variable
Datanorm <- scale(Data, center = mmums, scale = mmums - mmums)
```

Données Aberrantes
Données Manquantes
Données Spéciales

Données Aberrantes/Manquantes/Spéciales

Avant de traiter les données, vérifier la qualité des données

✓Identifier les données aberrantes.
✓Traiter ou gérer les données manquantes.
✓Enlever les caractères spéciaux dans les données numériques.

Les données peuvent être:

- Manquantes
- Aberrantes
- Spéciales

1. Données Aberrantes

Définition :
Un *outlier*, ou donnée *aberrante* est une valeur ou une *observation* qui est « *distante » des autres observations* effectuées sur le même phénomène, c'est-à-dire qu'elle contraste grandement avec les valeurs « normalement » mesurées.

Une donnée aberrante peut être due à la variabilité inhérente au phénomène observé ou bien elle peut aussi indiquer une erreur expérimentale.

1. Données Aberrantes

Formes :

Les données aberrantes peuvent prendre plusieurs formes:

- Données catégorielles.
- Valeur extrême (positive ou négative) de probabilité faible

Une valeur **aberrante** est toujours une **valeur extrême de l'échantillon**.

1. Données Aberrantes

Exemples :

<u>Contrôle sur le domaine des valeurs</u> :
Pour la variable « Total des heures effectuées », une borne maximale (208 heures) est fixée à partir de la convention collective. Les valeurs supérieures à 208 heures sont aberrantes.

<u>Comptes fournisseurs</u>:
Dans un fichier de compte fournisseurs, l'ensemble complet des factures varie entre 40 $ et 5 000 $. Toutefois, trois factures présentent un montant supérieur à 20 000 $.

1. Données Aberrantes

Méthodes de détection:

*Calcul du Z-Score

*Test de Grubbs

*Détection graphique : Pour détecter la présence de valeurs aberrantes On peut utiliser :

- Boxplot
- Histogrammes
- Nuages de points
- diagramme de dispersion des observations classées en fonction de leur rang

Détection graphique: Boîte à moustaches

Les boîtes à moustaches sont des représentations graphiques utilisable pour des variables numériques.
La « boîte » est délimitée par le premier et le troisième quartiles, elle contient donc 50% de la population.

Les moustaches encadrent les individus “proches” du centre. Au-delà des moustaches, on trouve soit **les valeurs aberrantes** (erreur de saisie), soit **les valeurs éloignées du centre** (valeurs extrêmes).

Détection graphique: Boîte à moustaches

Packages R : ggplot2

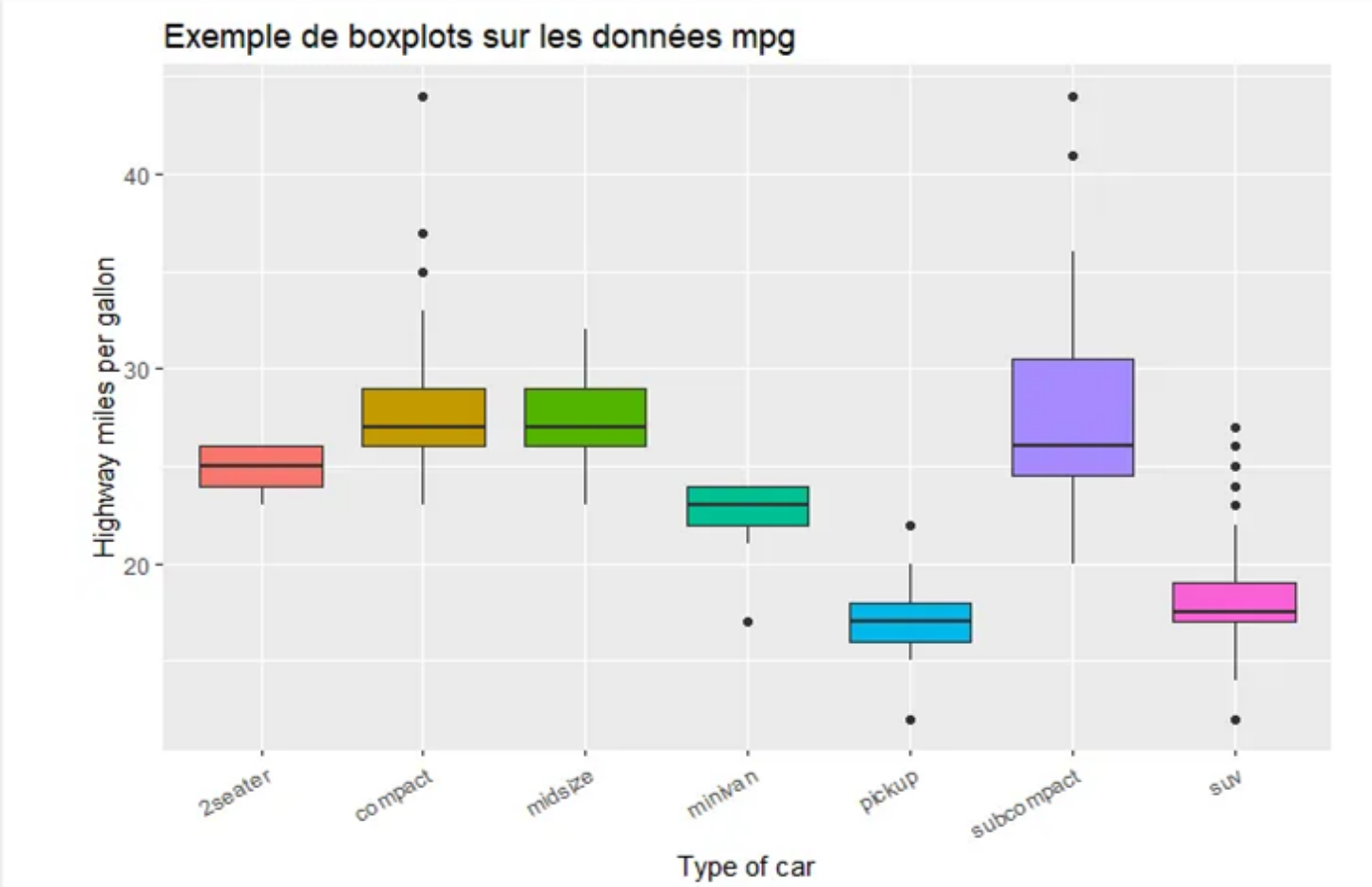
Boîte à moustaches:Exemple

Le jeu de données **mpg** contenues dans le package **ggplot2**.

On s'intéresse à la variable **hwy** (Highway miles per gallon) en fonction du type de véhicule (variable class).

```
>library(ggplot2)
>ggplot(mpg, aes( x=class,y=hwy, fill=class)) + geom_boxplot()+ xlab(label = "Type of car") + ylab(label = "Highway miles per gallon") + theme(axis.text.x = element_text(angle=30, hjust=1, vjust=1))+theme(legend.position="none")+ggtitle("Exemple de boxplots sur les données mpg")
```

Boîte à moustaches:Exemple



Sur cette visualisation des données, l**es valeurs aberrantes sont représentés sous forme de points**. Ils correspondent à des observations dont les valeurs sont :
supérieures à la valeur du 3ème quartile plus 1.5 fois l'intervalle inter-quartile, ou inférieures à la valeur du 1er quartile moins 1.5 fois l'intervalle inter-quartile.

Détection analytique: Z-Score

- Le Z-score ; calculé statistiquement en se basant sur la valeur assigné (référence). Soit la variable aléatoire $X \sim N(\mu, \sigma)$.

Le Z score est calculé comme suit:

$$Z_{score} = \frac{x - \mu}{\sigma}$$

Avec;

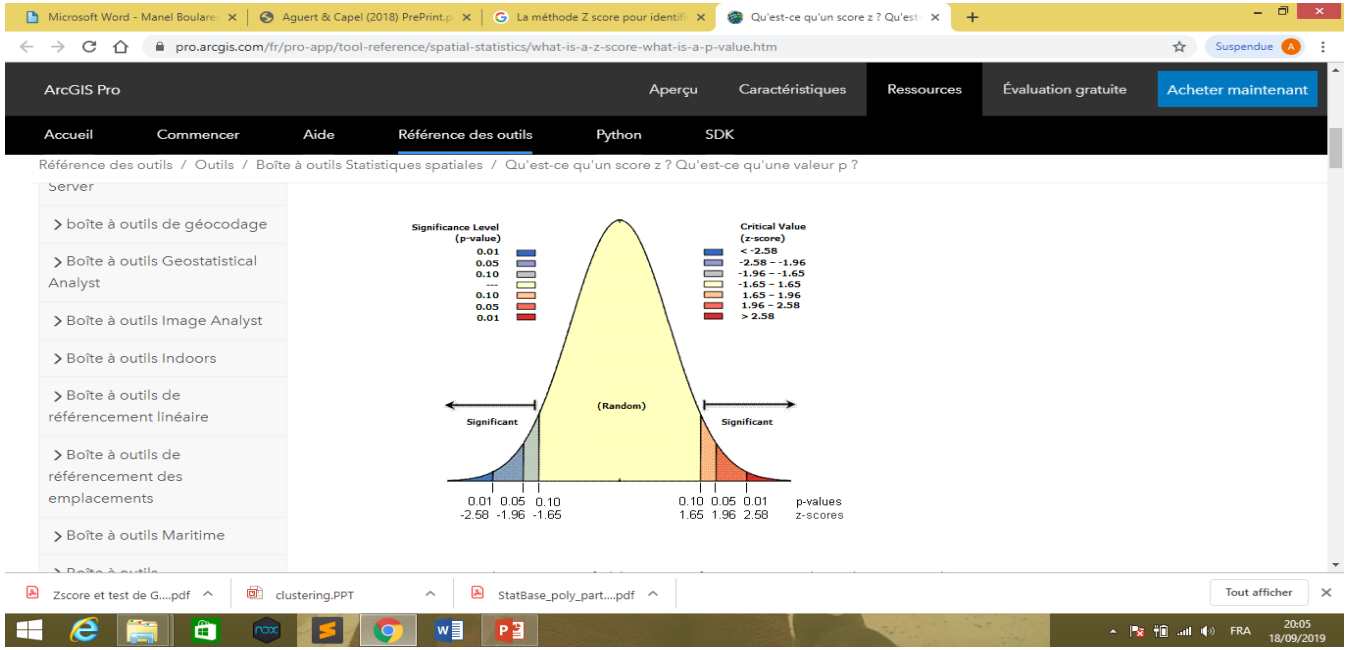
x: La valeur de mesure

μ: La moyenne de mesure

σ: l'écart type

Détection analytique: Z-Score

Supposons qu'un laboratoire ait fourni le résultat X= 4,10 L. Dans ces conditions, Z = (4,10 – 3,22) / 0,267 = 3,30. Le résultat X peut donc être considéré comme "hors limites", puisqu'il se situe à plus de 3 écart type) de la valeur assignée.



2. Données Manquantes

- Les tableaux de données couramment utilisés dans l'analyse statistique présentent souvent des données manquantes.
- Il est fréquent que les personnes qui ont volontairement répondu à un sondage d'opinion refusent de répondre à certaines questions
- Les techniques statistiques partent des données pour apporter de l'information, et ne sont généralement pas propres à traiter des ensembles de données dans lesquels certaines d'entre elles sont inconnues.

2. Données Manquantes

Types:
On distingue deux catégories de **non-réponse** :

- la **non-réponse totale**, lorsque aucune information n'est recueillie sur une unité échantillonnée.

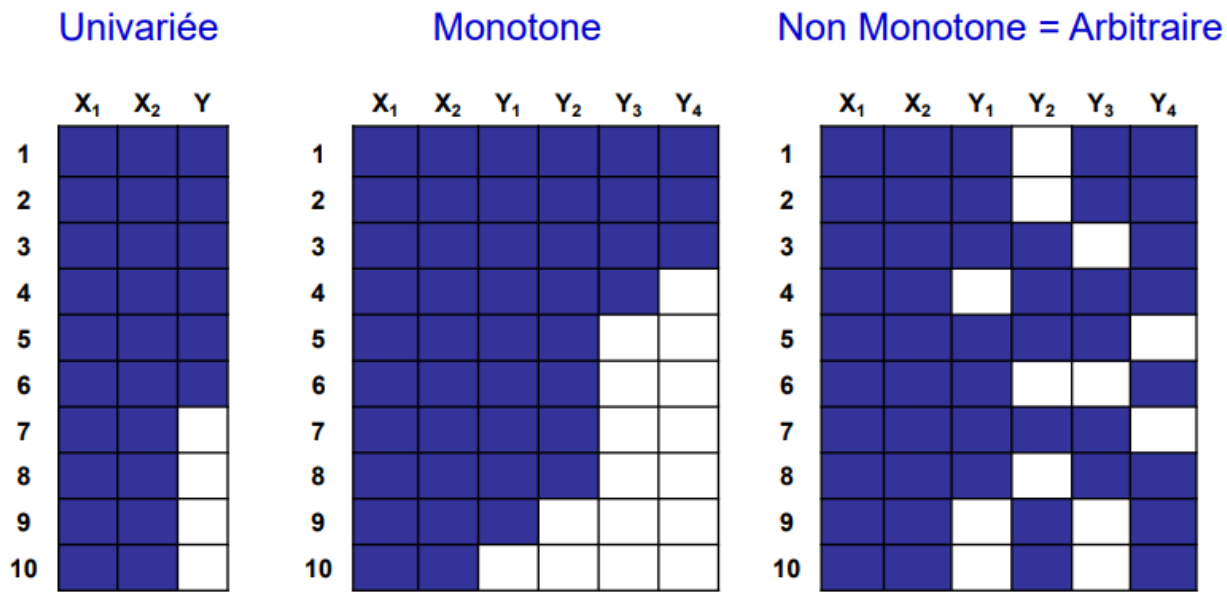
- la **non-réponse partielle**, lorsque le manque d'information est limité à certaines variables.

2. Données Manquantes

Soient:

X un vecteur de variables complètes

Y vecteur de variables incomplètes



2. Données Manquantes

Solutions:

Les deux formes, radicalement différentes, pour contourner ou résoudre le problème de données manquantes sont les suivantes :

- **Elimination**: réduire l'échantillon aux individus «complets», parmi lesquels on ne trouve aucune donnée manquante.

 Retirer les individus qui présentent des données manquantes, s'ils représentent un faible pourcentage de l'échantillon (**5% ou moins**), peut paraître raisonnable.

- **Imputation** : attribuer des valeurs aux données manquantes selon un critère rationnel.

2. Données Manquantes

Imputation Simple:

Principe: Remplacer chaque DM par une valeur plausible. Cela peut être: **la moyenne** ou **la médiane** pour les variables quantitatives ou **le mode** pour les variables qualitatives.

Cette méthode peut comprendre deux types:

- **Generalized imputation**

  On calcule la moyenne ou bien la médiane de toutes les valeurs non manquantes que prend la variable, puis on remplace les DM par ces valeurs (Mode dans le cas des variables qualitatives)

- **Similar case imputation**

  On remplace les DM par des valeurs provenant d'individus similaires pour lesquels, toute l'information a été observée

2. Données Manquantes

Imputation par l'algorithme KNN (1/2)

Dans cette méthode:

*Les DM d'une variable sont imputées en utilisant les variables les plus similaires de celui en question

*La similarité entre deux attributs est déterminée en utilisant une fonction de distance

*Le package **VIM** dans **R**,utilise une fonction appelé kNN qui implémente **une distance de Gowers** pour déterminer les K plus proches voisins

2. Données Manquantes

Imputation par l'algorithme KNN (2/2)

Pour effectuer une prédiction, l'algorithme K-NN va se baser sur le jeu de données en entier. En effet, pour une observation, qui ne fait pas parti du jeu de données, qu'on souhaite prédire, l'algorithme va chercher les K instances du jeu de données les plus proches de notre observation. Ensuite pour ces voisins, l'algorithme se basera sur leurs variables de sortie (output variable) pour calculer la valeur de la variable de l'observation qu'on souhaite prédire.

2. Données Manquantes

Imputation Hot deck

- Imputer la valeur manquante avec une valeur observée de la même base de donnée aléatoirement (Sous R, la fonction **impute** du package **Hmisc** implémente cette méthode en utilisant la fonction)
- **Exemple:** Soit « height » les tailles extraites du jeu de données « Woman »

```
> height <- women$height
> height[c(6,9)]<-NA #Ajouter des DM
> height<-Hmisc::impute(height, "random")
> height
```

```
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
 58  59  60  61  62 65*  64  65 61*  67  68  69  70  71  72
```

2. Données Manquantes

Sous R:

- Une valeur manquante est indiquée par NA :

```
u = c(31, 43, NA, 36, NA)
is.na(u)
```

On ignore les valeurs manquantes :

```
mean(u, na.rm = TRUE)
```

- Si le fichier texte données contient des données manquantes, on ignore les individus correspondants en utilisant la commande na.strings :

```
donnees = read.table("donnees.txt", header = T, na.strings = "NA")
```

- On peut également créer un nouveau jeu de données sans valeur manquante en faisant :

```
donnees2 = na.omit(donnees)
```

2. Données Manquantes

Gestion des valeurs manquantes:

➢Compter le nombre de valeurs manquantes dans chacune des colonnes

```
sapply(données, sum(is.na(x)))
```

➢Compter le nombre de NA dans une colonne en particulier

```
sum(is.na(données$colonne))
```

➢Afficher les lignes dont la valeur dans la colonne "colonne" est NA

```
dataset[is.na(données$colonne),]
```