1 Préparation de données

Repères visuels

1. La **position** sur une **même échelle** (variable numérique), souvent à l'aide de points ou de boîte à moustache ;
2. La **position** sur une **échelle identique**, mais **non alignée** (variable numérique) ;
   - Pour exemple, lorsqu'on utilise des facettes.
3. La **longueur** sur une **même échelle** (variable numérique) ;
   - Pour exemple, dans un diagramme à bande ou un histogramme.
4. L'angle ou **la pente** (variable numérique) ;
   - Pour exemple, en représentant les données sous forme de lignes (pente) ;
   - Pour exemple, la redoutable pointe à tarte (angle).
5. La **forme des points** (variable catégorielle) peut représenter le groupe ;
6. L'aire ou **le volume** (variable numérique) ;
   - Pour exemple, dans un diagramme à bande ou un histogramme ;
7. La **saturation de la couleur** (numérique ou catégorielle) peut représenter "à quel point" sur une échelle de claire à foncé ;
   - Pour exemple, gris vs noir.
8. La **teinte de couleur** (numérique ou catégorielle).
   - Pour exemple, bleu vs rouge.

Étapes du nettoyage de données

Cette liste n'est pas séquentielle, il est surtout important de *tout* le faire.

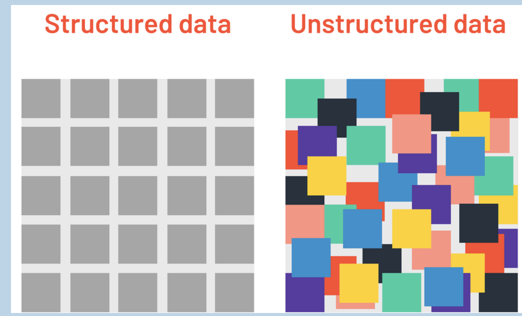
- ☐ **Comprendre** la structure des données ;
  - dimensions, types de variables, `str` et `summary`.
- ☐ **Visualiser** les données ;
  - `head`, `summary` et graphiques exploratoires.
- ☐ **Mettre en forme *(format)*** les données ;
  - Chaque ligne est une observation ;
  - Chaque colonne est une variable ;
  - Supprimer les doublons ;
- ☐ Vérifier et corriger les **types** de variables ;
  - booléens, entiers, numériques, facteurs, chaînes de caractères, dates, etc.
- ☐ **Manipuler** les **chaînes** de caractères ;
  - Corriger les typos ;
  - Changer la casse avec `tolower` ;
  - Extraire des informations avec les expressions régulières.
- ☐ Identifier les **données aberrantes** ;
  - Mettre `NA` et gérer plus tard.
- ☐ **Détecter** les **erreurs** flagrantes ou les changements structurels dans les données ;
  - Prendre compte des réformes ;
  - Constater les structures importantes.
- ☐ **Augmenter** les données à l'aide d'autres sources ;
  - optionnel.

Types de variables

Les jeux de données peuvent être :

Structuré : Données ayant une structure prédéfinie.
Pour exemple :
- Tableaux ;
- « spreadsheet » ;
- « Relational databases ».

Non structuré : Données sans structure prédéfinie venant en toute forme et ne pouvant pas être facilement résumées à un tableau.
Pour exemple :
- Texte ;
- Images ;
- Audio ;



Types de variables :

Quantitatives : Données numériques pouvant être :
- Discrète, pour exemple des données de comptage ;
- Continue, pour exemple des montants de sinistre.

Qualitatives : Données catégorielles pouvant être :
- Nominales, pour exemple le sexe ;
- Ordinales, pour exemple des groupes d'âge.

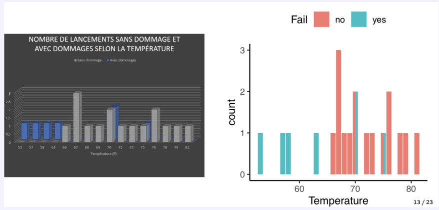
Temporelles Données chronologiques représentant un état dans le temps ;

- Pour exemple, la quantité totale de pluie au Canada le 1er juillet 1990.

Spatiales Données géographiques avec des coordonnées (pour exemple, la latitude et longitude).

Conseils

- Conserver un script séparé pour le traitement des données ;
- Documenter nos choix ;
  - Si l'on omet quelques observations en raison de données aberrantes, il faut être conscients que ça peut biaiser les résultats.
- Effacer l'encre superflue



- Considérer l'étique :
  - Avons-nous l'autorisation d'utiliser les données ?

    Pour exemple, les données ouvertes du gouvernement avec peu de restrictions ;
  - Est-ce qu'il y a des considérations éthiques à l'utilisation des données ?

    Pour exemple, Facebook qui utilise les données de façon immorale ;
  - Est-ce qu'il y a des biais dans les données pouvant causer préjudice ?

    Pour exemple, des données avec un biais sexiste ou raciste, des données financées par une compagnie.

Biais

d'échantillonnage Lorsque les données d'entraînement ne représentent pas de façon représentative la vraie population.
Pour exemple :

- Entraîner une voiture autonome seulement le jour alors qu'elles peuvent être conduites jour et soir ;
- Sonder les lecteurs du Devoir sur le parti pour lequel ils vont voter à l'élection et donc ignorer ceux ne lisant pas le journal.

de stéréotypes Données influencées par des stéréotypes (consciemment ou pas).
Pour exemple :

- Entraîner un algorithme pour comprendre comment les personnes travaillent avec des images d'hommes sur des ordinateurs et de femmes à la maison ;
- L'algorithme aura tendance à penser que les hommes sont des programmeurs et les femmes des cuisinières.

de mesure Données influencées par un problème avec l'instrument de mesure.
Pour exemple :

- Entraîner un algorithme de reconnaissance d'image avec des images d'une caméra avec un filtre de couleur ;
- L'algorithme serait fondé sur des images ayant systématique mal-représentée le vrai environnement.

de modèle Le compromis de biais-variance $B(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta$.
Pour exemple :

- Certaines variables ne sont pas considérées (ou mesurées) ;
- Le modèle n'est pas assez flexible (compromis linéaire vs lisse)

2 Données manquantes

Le chapitre utilise la **mise en contexte** suivante :

- Il y a une réclamation pour un accident d'auto en Ontario ;
- Le contrat d'assurance couvre les frais médicaux ;
- On désire calculer la probabilité de paiement (variable réponse) en fonction de :
  1. La gravité de l'accident (variable explicative) ;

     3 niveaux : mineur-majeur-catastrophique ;
  2. La souffrance du réclamant ;

     Échelle de 1 (peu) à 5 (beaucoup) ;

Problèmes de modélisation :

- Comment analyser les données malgré les valeurs manquantes ?
- Quels enjeux ou problèmes devrait-on considérer dans la modélisation ?

Terminologie

Notation

Y_{ij} : Valeur de la variable explicative j pour l'observation i où $j \in \{1,\ldots,p\}$ et $i \in \{1,\ldots,n\}$;

$\mathbf{Y}_{n\times p}$: Matrice contenant les données **complètes** ;

$\mathbf{Y}$ est partitionné en deux, $\mathbf{Y} = \{\mathbf{Y}_{obs}, \mathbf{Y}_{mis}\}$

$\mathbf{Y}_{obs}$: matrice avec les données ayant toutes les valeurs observées ;

$\mathbf{Y}_{mis}$: matrice avec les données comportant des valeurs manquantes ;

$\mathbf{R}_{n\times p}$: **Matrice de réponse** des variables indicatrices $R_{ij} = \mathbf{1}_{\{Y_{ij} \text{ observé}\}}$;

θ : **Paramètre de nuisance**

Exemple de notation

$$\boldsymbol{Y} = \begin{bmatrix} 2 & 3 \\ 8 & 6 \\ 3 & 12 \end{bmatrix}$$

$$\boldsymbol{Y}_{obs} = \begin{bmatrix} 2 & . \\ 8 & 6 \\ . & . \end{bmatrix} \qquad \boldsymbol{Y}_{mis} = \begin{bmatrix} . & 3 \\ . & . \\ 3 & 12 \end{bmatrix}$$

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$$

Mécanisme de non-réponse

La distribution de **R** est le *mécanisme de non-réponse* ;

Types de données manquantes :

1. **MCAR** : Missing Completely at Random ;
   - Le patron de non-réponse (pattern of missing values) est indépendant des données **Y** ;
   - Il s'ensuit que la probabilité de réponse $f(R|\mathbf{Y},\theta)$ ne dépend pas des données complètes **Y** :

     $f(R|\mathbf{Y},\theta) = f(R|\theta)$

     Exemple avec un θ de 10%

     On perd 10% des valeurs mesurées alors, $\forall i \in \{1,\ldots,n\}, j \in \{1,\ldots,p\}$, la distribution du mécanisme de non-réponse :

     $R_{ij} \sim \text{Bernoulli}(p = 1-\theta = 90\%)$

   - **Tester** la différence de moyennes :

     $\mathcal{H}_0 : \{p_{\text{Cat, mis}} - p_{\text{Cat, obs}} = 0\}$ et $\left\{p_{\text{Maj, mis}} - p_{\text{Maj, obs}} = 0\right\}$

     Est équivalent à tester :

     $\mathcal{H}_0$: les données sont MCAR

     avec un test du khi-carré de Pearson ;

     C'est-à-dire que le test est équivalent, mais **pas** le hypothèses.
2. **MAR** : Missing at Random ;

   Can think of it as Missing *Conditionally* at Random ;
   - La probabilité de réponse $f(R|\boldsymbol{Y}1616,\theta)$ dépend seulement des variables qui ont été observées dans le jeu de données $\mathbf{Y}_{\text{obs}}$:

     $f(R|\mathbf{Y},\theta) = f(R|\mathbf{Y}_{\text{obs}},\theta)$
   - Exemple de patients d'un hôpital : les données sont MAR lorsque la probabilité de non-réponse ne dépend pas de la qualité de vie sachant l'âge ;
   - Le négatif est qu'il est impossible de tester que, sachant l'âge, la probabilité de non-réponse ne dépend pas de la qualité de vie ;
   - Il est **inconcevable** d'avoir un test pour MAR.
3. **NMAR** : Not Missing at Random ;
   - Le patron de non-réponse pour **Y** est relié à sa valeur et les variables observées ;

     Ce même si on conditionne sur les valeurs observées ;
   - La probabilité de réponse $f(R|\mathbf{Y},\theta)$ dépend également de $\mathbf{Y}_{\text{mis}}$ et ne peut pas être simplifiée ;
   - Pour exemple, les patients malades ne répondent pas aux sondages en plus des patients plus jeunes et donc la probabilité de réponse dépend de la qualité de vie ;
   - Pour exemple, la probabilité de réponse dépend d'une autre variable non observée ;

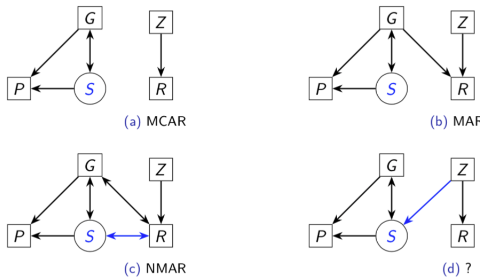
Visuellement on peut comparer les 3 patrons de non-réponse :

- On observe (MCAR) que la variable *indépendante* S ne dépend pas du patron de non-réponse R ;
- On observe (MAR) que G influe la variable réponse S et le patron de non-réponse R ;
- On observe (NMAR) qu'il y a un lien direct entre S et R ;

- En dernier, puisque Z n'est pas mesuré et que Y y dépend c'est NMAR.

Principe d'inclusion : Si une variable est exclue, cela peut créer une corrélation entre la variable indépendante S et le patron de non-réponse R. De plus, ça peut changer le patron lui-même !

P = paiement, G = gravité, S = souffrance, Z = variables non-mesurées



Pour exemple, si on cherchait à supprimer des valeurs d'une BD :

MCAR supprime des valeurs aléatoirement ;

MAR supprime 60% des valeurs pour les femmes et 40 % des valeurs pour les hommes ;

NMAR plus il y a de sinistres observés, plus il y a de chances que les valeurs soient manquantes.

Ordre de restriction des différents patrons

1. MCAR : plus « restrictif », car les données doivent être manquantes complètement aléatoirement ;

   Même idée que l'hypothèse de normalité est restrictive puisque les données *doivent* l'être, **mais** cela nous permet d'utiliser plein de tests ;

   Comme une Bernoulli, il n'y a pas de paramètres ;

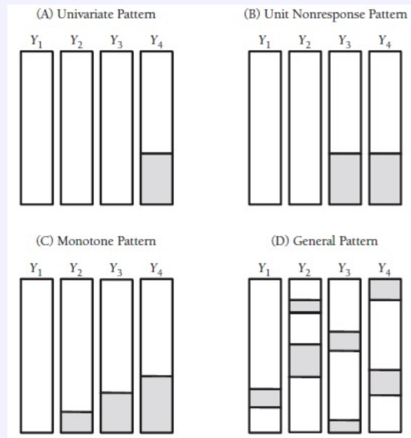
   plus restrictif alias moins flexible.
2. MAR : inclue les variables observées et donc il y a plus de paramètres ;
3. NMAR : inclue toutes les variables et est donc le patron le plus flexible (alias, le moins restrictif).

Visualisation et détection

Traiter les données manquantes

1. Détectez, visualisez et documentez les données manquantes ;
2. Identifier le patron de non-réponse ;

   Pour exemple, voici quelques patrons de non-réponse pour quelques variables :

   Exemple de patron univarié : concessionnaire ne pose jamais à ses assurés s'ils ont une piscine et donc la variable est manquante dans toutes ses données ;
3. Comparer les distributions des autres variables selon la valeur des variables indicatrices $R_{1j}, \ldots, R_{nj}$;

Identification des types de non-réponses

- Pour les variables continues, on fait un **test** t **sur les différences de moyenne** au lieu du khi-carré de Pearson comme pour MCAR ;
- Problème de comparaisons multiples ;
- Le test MCAR de Little est peu utile, mais peut adresser le problème de comparaisons multiple avec une hypothèse testant toutes les variables ;

Traitement des données manquantes

En continuant la mise en contexte, on suppose qu'on veut estimer le vecteur **ﬁ** des coefficients de la régression logistique pour prédire la probabilité de paiement ; Une question valide est si les options pour le traitement des données manquantes dépendent du **type** de non-réponse

Options de traitement :

1. Utiliser seulement les **cas complets** *(complete-case analysis)* ;
   - L'option par défaut pour les fonctions :

     `lm, glm, na.rm, na.omit`
   - **Impact** :

     ↓ taille de l'échantillon

     ↓ taille de la corrélation des variables

     ↑ variance des estimateurs

     ↓ puissance des tests
   - Uniquement valide sous MCAR ;
2. Utiliser seulement les **cas disponibles** *(available-case analysis)* ;
   - Utilise uniquement les données observées pour l'analyse ;
   - Rarement applicable ;
   - ↓ la taille de l'échantillon **moins** qu'en utilisant d'uniquement les cas complets ;
   - Sans biais **uniquement** sous MCAR ;
3. **Imputation** simple par la **moyenne ou** la **médiane**
   - Substitue les `NA` par la moyenne ou médiane de la variable ;
   - **Impact** :

     ↓ variabilité de la variable

     ↓ corrélation de la variable avec les autres

- Même sous MCAR, les données sont **sévèrement** « distorted » ;

4. Imputation simple par une régression ;
    - Substitue les `NA` par la prévision d'une régression de la variable sur les autres avec les cas complets ;
    - Si plusieurs variables ont des données manquantes, leurs patrons doivent être traités séparément ;
    - L'inter **corrélation** des variables est **conservée**, mais est **surestimée** (même si MCAR) ;
    - La variance est **sous-estimée**, mais **moins** qu'avec l'imputation par la moyenne ;
5. Imputation stochastique par une régression ;
    - Ajoute un terme d'erreur ε (normalement distribué) à la prévision de la régression ;
    - *Si plusieurs variables sont manquantes dans un patron, les erreurs sont corrélées*
    - Corrige les biais pour la méthode d'imputation par la régression (sous-estimation de la variance et surestimation de l'inter corrélation des variables) ;
    - La variance des paramètres est **sous-estimée**, *sauf si on en tient compte dans les calculs* ;
    - Fonctions `R` utiles du paquetage `mice` :

      `mice.impute.norm.nob()`, `mice.impute.norm()`
6. Imputation simple *hot-deck* ;
    - Substitue les valeurs `NA` d'une observation par les valeurs observées d'une autre observation choisie aléatoirement ; ;
    - Habituellement, cette observation fait parmi d'un sous-ensemble d'observations *proches* (pensez au *K*-NN, clustering, etc.) ;
    - Souvent utilisée pour les sondages ;
    - N'altère par les distributions univariées
    - ↓ l'inter corrélation des variables ;
    - Biais des estimations des coefficients fi de régression ;
7. Imputation **multiple** ;
    - Répète l'imputation stochastique et agrège les résultats ;
    - Ce faisant, la variabilité additionnelle dût à l'imputation des valeurs manquante est adressée et la variance des estimateurs est *non biaisée* ;

Autres méthodes :

- MLE avec données manquantes ;
- Algorithme EM (expectation-maximisation)
- Inférence bayésienne ;

Conseils

- Conserver un script pour le traitement de données manquantes et ne **pas hard-coder** ;
- *Utiliser une méthode d'imputation qui respecte le format de la variable ;*
- Plus la proportion de non-réponses est élevée, plus l'impact sur l'analyse sera important ;
- S'il y a plusieurs patrons de non-réponse différents, l'ordre dans lequel les données sont imputées est important ;

3 Analyse en composantes principales

4 Classification non supervisée

Concepts de distance et similarité

Classification

supervisée Lorsqu'on connaît les **étiquettes** des groupes et on veut prédire l'appartenance à un groupe pour de nouvelles observations ;
- « *Classification* ».

non supervisée Lorsqu'on ne sait pas combien de groupes il y a ni comment qu'ils sont formés ;
- Alias partitionnement ou regroupement ;
- « *Clustering* ».

Conditions des mesures de distance

Une mesure de distance dans l'espace $\mathbb{E}$ est une application $d : E \times E \to \mathbb{R}^+$ respectant ces 3 conditions :

1. Symétrique
   $d(i,j) = d(j,i), \forall i,j \in E$;
2. Nulle pour un même objet
   $d(i,j) = 0 \Leftrightarrow i = j, \forall i,j \in E$;
3. Satisfait l'inégalité du triangle
   $d(i,k) \leq d(i,j) + d(j,k), \forall i,j,k \in E$.

Distance de Minkowski ℓ_q

Soit 2 points $(x_{11}, \ldots, x_{1d})$ et $(x_{21}, \ldots, x_{2d})$ dans l'espace $\mathbb{R}^d$.
Alors, pour $q \geq 1$:

$$\ell_q = \|\mathbf{x_1} - \mathbf{x_2}\|_q = \left(\sum_{i=1}^{d} |x_{1i} - x_{2i}|^q\right)^{1/q}$$

Deux cas particuliers sont bien connus :

- $q = 1$: Distance de Manhattan ℓ_1 ;
- $q = 2$: Distance euclidienne ℓ_2.

La version standardisée de la distance euclidienne se simplifie à :

$$d(\mathbf{x_1}, \mathbf{x_2}) = \sqrt{\sum_{i=1}^{d} \left(\frac{x_{1i} - x_{2i}}{s_i} \right)^2}$$

Conditions des indices de similarité

Une mesure de similarité entre 2 objets dans l'espace $\mathbb{E}$ est une application $s : E \times E \to [0, 1]$ respectant ces 2 conditions :

1. Symétrique
   $s(i, j) = s(j, i)$, $\forall i, j \in E$;
2. Un même objet est identique et maximise la similarité
   $s(i, j) = 1 \geq s(i, j)$, $\forall i, j \in E$.

De plus, l'indice de similarité peut être obtenu d'une mesure de distance $s(i, j) = \frac{1}{1+d(i,j)}$.

Également, on peut définir un indice de **dissimilarité** avec $\tilde{d}(i, j) = 1 - s(i, j)$.

Types de variables binaires

symétrique Si une variable binaire est **symétrique**, c'est que ses deux niveaux sont aussi fréquents ;
- Par exemple, qu'une personne soit un homme ou une femme.

asymétrique Si une variable binaire est **asymétrique**, c'est qu'un niveau est plus rare que l'autre.
- On dénote le niveau plus rare par 1 ;
- Par exemple, deux personnes ayant brisé leur orteil sont plus semblables que deux personnes ayant un mal de tête.

Mesures de similarité (variables binaires ou ordinales)

Variable binaire symétrique

La **proportion d'accord** (« *simple matching coefficient* ») :

$$s(\mathbf{x_1} - \mathbf{x_2}) = \frac{1}{d} \sum_{i=1}^{d} \mathbf{1}_{\{x_{1i} = x_{2i}\}}$$

Variable binaire asymétrique

L'indice de Jaccard :

$$s(\mathbf{x_1} - \mathbf{x_2}) = \frac{\sum_{i=1}^{d} x_{1i} x_{2i}}{\sum_{i=1}^{d} \{1 - (1 - x_{1i})(1 - x_{2i})\}}$$

Variable binaire asymétrique

On assigne un score numérique (positif) à chaque niveau et traite la variable comme une variable numérique.

Similarité de Gower

Dans le cas où les variables sont de plusieurs types :

$$G(\mathbf{x_1} - \mathbf{x_2}) = \frac{\sum_{i=1}^{d} w_i \gamma_i(x_{1i}, x_{2i}) s_i(x_{1i}, x_{2i})}{\sum_{i=1}^{d} w_i \gamma_i(x_{1i}, x_{2i})}$$

où le poids des variables $w_i > 0$.

Également, avec l'étendu dénoté $r_i = \max(x_{1i}, \ldots, x_{ni}) - \min(x_{1i}, \ldots, x_{ni})$, si $\mathbf{x_i}$ est :

- numérique ou ordinale, $\gamma_i(x_{1i}, x_{2i}) = 1$ et $s_i(x_{1i}, x_{2i}) = 1 - \frac{|x_{1i} - x_{2i}|}{r_i}$;
- binaire symétrique, $\gamma_i(x_{1i}, x_{2i}) = 1$ et $s_i(x_{1i}, x_{2i}) = 1_{\{x_{1i} = x_{2i}\}}$;
- binaire asymétrique, $\gamma_i(x_{1i}, x_{2i}) = 1 - (1 - x_{1i})(1 - x_{2i})$ et $s_i(x_{1i}, x_{2i}) = 1_{\{x_{1i} = x_{2i}\}}$.

Similarité du Cosinus

Un exemple d'application de la similarité du Cosinus est l'analyse de texte. On assigne une variable indicatrice pour chaque mot selon s'il est présent ou non (il y a donc *beaucoup* de colonnes). Par la suite, on assigne un poids w_i à chaque variable selon son importance. Donc, par exemple, la mesure peut calculer la similarité entre deux textes.

$$s_i(x_{1i}, x_{2i}) = \frac{\sum_{i=1}^{d} w_i x_{1i} x_{2i}}{\sqrt{\sum_{i=1}^{d} w_i x_{1i}^2 \sum_{i=1}^{d} w_i x_{2i}^2}}$$

K-means clustering

Hierarchical clustering