Etude de la base de données :

**Table des matières**

[Présentation de la base données 2](#_Toc111644466)

[Extraction de la base de données 2](#_Toc111644467)

[Quelle est la structure de la base de données d’étude ? 2](#_Toc111644468)

[Traitement de la donnée 4](#_Toc111644469)

[Ciblage sur le risque arrêt de travail 4](#_Toc111644470)

[Quels choix a-t-on effectué pour le nettoyage de la base de données ? 4](#_Toc111644471)

[Variables d’étude à calculer 4](#_Toc111644472)

[Durée d’un arrêt de travail 4](#_Toc111644473)

[Analyse descriptive 5](#_Toc111644474)

[Construction d’une base de données agrégée en vue du calcul du taux d’absentéisme 5](#_Toc111644475)

# Présentation de la base données

## Extraction de la base de données

Les données que nous utilisons pour cette étude sont des données de prestation en prévoyance collective du portefeuille de Generali VIE fournies par la direction de la Techniques Assurance.

Cette base de données est alimentée par les équipes de gestion, actualisée et mise à jour chaque mois puis importé dans le logiciel de gestion AVT, utilisé par l’ensemble des équipes de souscription et surveillance de portefeuille pour leurs opérations au quotidien.

Cette base de données nous donne accès à l’ensemble des sinistres ITIP survenu sur le périmètre collectif depuis 2012.

Nous allons à présent nous focaliser dans le détail de la structure de cette base de données et des informations qu’elle pourra nous apporter dans notre démarche.

## Quelle est la structure de la base de données d’étude ?

La base de données d’étude est structurée par ligne, chacune d’elle correspond à un sinistre prévoyance, voici une liste non-exhaustive des garanties couvertes :



La base de données initiale est composée de 63 variables pour chacun des sinistres, nous avons sélectionné parmi celles-ci, les variables suivantes :



Après avoir sélectionné les informations intéressantes pour notre étude, un travail de data-cleaning était nécessaire pour différentes raisons, d’abord pour homogénéiser les données car elles ont été saisies par différentes équipes de gestions qui n’ont pas les mêmes process et les mêmes normes depuis 2012 donc il s’agira d’effectuer un travail d’homogénéisation des données, de plus il faut pouvoir être capable de capter les erreurs de saisie et de faire des choix quant au nettoyage ou au remplacement de celles-ci afin d’avoir la base de données la plus « propre » possible.

# Traitement de la donnée

## Ciblage sur le risque arrêt de travail

Pour étudier le risque arrêt de travail, il s’agit de sélectionner dans la base de données uniquement les sinistres de type ITT : Incapacité temporaire de travail. Pour ce faire, nous avons sélectionné dans la suite uniquement les sinistres dont la garantie élémentaire couverte contient les mentions “ITT” ou “INV” ou tout autres garanties incapacité/invalidité. De plus, parmi les données sélectionnées en incapacité, nous souhaitions ne considérer que les sinistres d’une durée inférieure à 3 ans, durée maximum légale de l’incapacité, c’est pourquoi nous avons calculé des variables à partir des données initiales pour en tirer les informations utiles à notre étude, ces variables vont être présentées et expliquées dans la suite.

## Quels choix a-t-on effectué pour le nettoyage de la base de données ?

Dans le processus de nettoyage d’une base de données différents cas de figures peuvent exister lorsque l’on parle d’une donnée que l’on juge impropre, celle-ci peut être erroné, décalée à cause d’une erreur de saisie manuelle par exemple, ou tout simplement manquante. Dans ces différents cas de figure plusieurs façons de résoudre le problème existent, on peut soit essayer de corriger les erreurs de saisie ce qui en pratique peut s’avérer très compliqué lorsque l’on travaille avec une base de données conséquente (notre base de données initiale contient 697000 lignes), soit trouver la source de l’erreur lorsqu’elle est automatique et la réparer, soit simplement supprimer la donnée. Dans le cas de notre étude, l’objectif était d’avoir la base de données la plus précise et donc pour ce faire nous avons fait les choix suivants :

* Supprimer les lignes qui possédaient des informations importantes manquantes (ex : le code option)
* Homogénéiser et simplifier les informations de certaines variables (ex : les garanties élémentaires, les codes options), par exemple en regroupant, dans le cas des garanties élémentaires, l’ensemble des garanties faisant référence à de l’incapacité sous la mention « ITT »
* Corriger lorsque c’était possible les erreurs de saisie : par exemple lors de la saisie du nom et du prénom de l’assuré, dans de nombreux cas le nom et le prénom étaient saisis dans le même champ, celui du nom, ce qui fait qu’un grand nombre de prénoms d’assuré étaient manquants dans la variable prénom car saisi dans le champ nom. Cette anomalie par exemple était corrigeable pour la plupart des dossiers

## Variables d’étude à calculer

### Durée de l’arrêt de travail

L’une des informations importantes que nous devons capter grâce à cette base de données pour notre étude est la durée de l’arrêt de travail, pour cela nous devons procéder avec les informations que nous possédons et en particulier grâce à la période de couverture, ainsi on calcule la durée d’un sinistre incapacité de la façon suivante :

D’ailleurs il faut aussi tenir compte que cette façon de calculer la durée d’un arrêt de travail n’est pas totalement exacte car elle ignore la période de franchise.

### Taux d’absentéisme

Cette variable à calculer est très importante et représente un enjeu important de cette étude, notamment pour la modélisation GLM dans la suite, elle est cependant assez complexe à calculer dans le contexte de l’étude avec les données en notre possession. La difficulté de ce calcul provient du fait que nous avons besoin de l’information du notre de jour arrêté total sur une année pour chaque assuré et que cette base référence uniquement les sinistres un à un, la solution que nous avons choisie pour calculer ce taux dans la suite est la création d’une nouvelle base de données qui cette fois-ci sera une base de données agrégée par assurés et par année de survenance.

# Analyse descriptive

# Construction d’une base de données agrégée en vue du calcul du taux d’absentéisme