

Les sciences autour de nous : laboratoires de probabilités et statistiques

Yasmine Tawfik et Jean-Sébastien Turcotte



Les sciences autour de nous

Laboratoires de probabilités et statistiques

Les sciences autour de nous

Laboratoires de probabilités et statistiques

Yasmine Tawfik
Cégep Gérald-Godin

Jean-Sébastien Turcotte
Cégep Gérald-Godin

2025/03/31

Édition: Première édition

Site Web: <https://jeansebastienturcotte.github.io/LabosStats/LSAN.html>¹

©2024–2025 Jean-Sébastien Turcotte

Sauf indications contraires, le contenu de ce manuel électronique est disponible en vertu des conditions de la licence "Creative Commons Attribution - Partage dans les Mêmes Conditions 4.0 International (CC BY-SA 4.0)", qu'il est possible de consulter à l'adresse [suivante](https://creativecommons.org/licenses/by-sa/4.0/deed.fr)², symbolisée par le logo ci-dessous.



Vous êtes autorisé(e) à :

Partager — copier, distribuer et communiquer le matériel par tous moyens et sous tous formats.

Adapter — remixer, transformer et créer à partir du matériel pour toute utilisation, y compris commerciale.

Selon les conditions suivantes :

Attribution — Vous devez créditer l'Oeuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'Oeuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son Oeuvre.

Partage dans les Mêmes Conditions — Dans le cas où vous effectuez un remix, que vous transformez, ou créez à partir du matériel composant l'Oeuvre originale, vous devez diffuser l'Oeuvre modifiée dans les mêmes conditions, c'est-à-dire avec la même licence avec laquelle l'Oeuvre originale a été diffusée.

Pas de restrictions complémentaires — Vous n'êtes pas autorisé(e) à appliquer des conditions légales ou des mesures techniques qui restreindraient légalement autrui à utiliser l'Oeuvre dans les conditions décrites par la licence.

¹

²<https://creativecommons.org/licenses/by-sa/4.0/deed.fr>

Remerciements

Un grand merci à l'organisme la [fabrique REL](#)³ qui a financé une partie de ce projet. En particulier, merci à Claude Potvin pour support tout au long du processus.

Merci au cégep Gérald-Godin, qui a complété la libération allouée par la Fabrique afin de nous permettre de produire ce matériel. Merci à notre spécialiste en moyens et techniques d'enseignement, Sandra Lenneville, qui a renouvelé l'expérience de l'univers des ressources libres. Un immense merci à Sylvain Pelletier, conseiller pédagogique et enseignant en français, langue et littérature, qui a assuré de main de maître la révision linguistique de ce projet! Merci pour tes précieux conseils.

Merci aux élèves de la session d'hiver 2025, qui ont eu à tester un produit imparfait, incomplet, et qui ont joué le rôle de cobayes dans cette aventure. Vos commentaires ont été les bienvenus.

Merci à Rob Beezer, pour PreTeXt, la machinerie qui permet de produire les différentes formes que peut prendre ce manuel. Merci pour ta patience et tes réponses aux nombreuses questions. Merci également aux autres développeurs et contributeurs du groupe d'aide [Google](#)⁴.

Finalement, un merci à l'avance à tous ceux et celles, élèves autant qu'enseignantes et enseignants, qui utiliseront ce manuel et qui, nous l'espérons, contribueront à l'améliorer et à le bonifier.



fabrique REL

RESSOURCES ÉDUCATIVES LIBRES



³<https://fabriquerel.org/>

⁴<https://groups.google.com/g/pretext-support>

Versions et source

Ce livre existe en deux formats principaux. La version la plus complète et la plus à jour sera toujours la version web. Elle est disponible à l'adresse <https://jeansebastienturcotte.github.io/LabosStats/LSAN.html>. Au besoin, des mises à jour seront effectuées chaque début de session, dans les premières semaines du mois d'août ou du mois de janvier. Si une mise à jour ne change pas la structure de la numérotation, elle pourrait être effectuée à un autre moment. Une liste des changements de l'édition en cours est disponible sur GitHub.

Une version PDF existe et est disponible en cliquant sur ce [lien](#). Cette version sera mise à jour moins régulièrement que la version en ligne. Elle le sera uniquement en cas de changement majeur, dans le but de la garder arrimée avec la version en ligne.

Le code source du manuel se trouve sur GitHub, à l'adresse [suivante](#).

Table des matières

Remerciements	iv
Versions et source	v
1 Introduction	1
1.1 Prélab	1
1.2 Laboratoire	3
1.3 Réflexions	28
2 Variables qualitatives	32
2.1 Prélab	32
2.2 Laboratoire	36
2.3 Réflexions	41
3 Variables quantitatives	43
3.1 Prélab	43
3.2 Laboratoire	48
3.3 Réflexions	111
4 Distribution d'échantillonnage	113
4.1 Prélab	113
4.2 Laboratoire	118
4.3 Réflexions	126
5 Estimation par intervalle de confiance	129
5.1 Prélab	129
5.2 Laboratoire	132
5.3 Réflexions	134
6 Tests d'hypothèses	136
6.1 Prélab	136
6.2 Laboratoire	140
6.3 Réflexions	165

7 Tests du khi-deux	167
7.1 Prélab167
7.2 Laboratoire169
7.3 Réflexions187
8 Corrélation et régression	189
8.1 Prélab189
8.2 Laboratoire193
8.3 Réflexions197

Appendices

A Bases de données	201
A.1 Armée américaine201
A.2202
A.3202
A.4 Base canadienne de données sur les collisions202
A.5 Diabète204
A.6 Polluants205
A.7 Précipitations206
A.8206
B Importer des données	207
C Manipuler la feuille de calcul	210
D Divers	212
D.1 Mise en forme212
E Fonctions utiles	213
E.1 Adresse213
E.2 Indirect214
E.3 Substitue215
F Les macros	219
F.1 Sauvegarder un fichier avec l'extension <code>xlsm</code>219
F.2 Afficher l'onglet Développeur220
G Raccourcis pratiques	224
G.1 Utilité224
G.2 Navigation224
G.3 Sélection variée225
G.4 Mise en forme225

Annexes

Chapitre 1

Introduction

Ce chapitre sert d'introduction au logiciel Excel en contexte scientifique. On y présente la première base de données qui sera utilisée lors des prochains laboratoires afin d'étudier la présence de diabète chez une population de femmes d'origine pima. Il couvre des sujets tels que l'interface d'Excel, la manipulation des lignes et des colonnes, l'attribution de noms aux cellules, le filtrage et le tri des données, ainsi que la création de tableaux croisés dynamiques. Ces compétences sont essentielles pour effectuer des analyses statistiques efficaces et organiser les données de manière optimale.

1.1 Prélab

Les trois premiers laboratoires sont consacrés à l'étude du diabète, en mettant l'accent sur les femmes d'origine pima. Les Pimas sont un peuple autochtone de la région du Sonora, au Mexique, et de l'État de l'Arizona, aux États-Unis. Les données utilisées proviennent du National Institute of Diabetes and Digestive and Kidney Diseases (l'Institut national du diabète et des maladies digestives et rénales des États-Unis). Les femmes ont été suivies sur une longue période, soit de 1965 à 1995. Ce type de recherche, appelée étude longitudinale, vise à observer l'évolution d'une variable dans le temps. L'année exacte de la collecte de données pour ce laboratoire n'est toutefois pas connue.

Selon Santé Canada, le diabète est défini comme « une maladie chronique qui se développe lorsque le corps ne produit pas l'insuline dont il a besoin pour transformer le sucre en énergie ou qu'il ne l'utilise pas efficacement ». Cette maladie se divise en trois catégories : le diabète de type 1, le diabète de type 2 et le diabète gestationnel. Le diabète de type 1 se caractérise par une production insuffisante d'insuline par le corps. Le diabète de type 2 est lié à une résistance du corps à l'insuline. Le diabète gestationnel, quant à lui, se manifeste par une élévation récente du taux de glucose dans le sang d'une femme enceinte.

Dans l'étude présentée, on se concentre sur le diabète de type 2, une forme de diabète fréquemment associée à l'inactivité physique, à l'obésité, à l'âge avancé d'un individu, ainsi qu'à des antécédents familiaux de diabète de type 2. Plusieurs facteurs expliquent cette prévalence chez les communautés autochtones. Notamment, ces dernières rencontrent des obstacles pour accéder à des soins de santé adéquats en raison de ressources limitées et de conditions économiques précaires.

Le fichier Excel à télécharger, ouvrir et enregistrer est [Données_Diabète.xlsx](#)

disponible en cliquant [ici](#)¹.

La base de données de l'enquête comprend les mesures diagnostiques d'un échantillon de 768 femmes pimas de l'Arizona. Les variables à l'étude sont :

- l'**identifiant** des participantes. L'éthique en matière de recherche exige l'anonymat des personnes participantes. Ainsi, un numéro est attribué à chaque individu afin d'éviter de divulguer leur identité;
- l'**âge** des participantes en année;
- le **nombre de grossesses**;
- la concentration de **glucose** plasmatique après deux heures lors d'un test de tolérance au glucose par voie orale en milligramme de glucose par décilitre de sang (mg/dL). Un taux élevé de glucose est un signe précoce du diabète de type 2. Après ce test, une valeur considérée saine est inférieure à 140 mg/dL. Une valeur comprise entre 140 et 199 mg/dL est considérée comme un prédiabète. Une valeur de 200 mg/dL ou plus indique un diabète;
- la **pression artérielle diastolique** en millimètre de mercure (mmHg). La pression diastolique indique la pression dans les artères lorsque le cœur se repose entre deux battements. Une valeur comprise entre 60 et 80 est considérée comme normale. Une valeur entre 80 et 90 est qualifiée de préhypertension. Une valeur supérieure à 90 est classifiée comme hypertension;
- l'**épaisseur du pli cutané du triceps** en mm;
- l'**insuline** dans le sang en micro unité internationale par millilitre ($\mu\text{U}/\text{mL}$). L'insuline est une hormone produite par le pancréas. Elle joue un rôle crucial dans le maintien de l'équilibre énergétique du corps et la régulation du taux de glucose sanguin. Après un repas riche en glucides, le taux de glucose peut augmenter rapidement; l'insuline intervient alors pour l'abaisser. En cas de production insuffisante d'insuline ou si le corps devient résistant à son action, le glucose reste en excès dans le sang, ce qui peut entraîner des maladies comme le diabète de type 1 ou le diabète de type 2. Dans l'étude des femmes d'origine pima, l'accent est mis sur le risque de développement du diabète de type 2;
- l'**indice de masse corporelle** (IMC) en kilogramme par mètre carré (kg/m^2). Comme l'indiquent les unités, l'IMC est la valeur obtenue en divisant la masse d'un individu par sa taille au carré. Il s'agit d'un indicateur permettant d'estimer le surpoids d'une personne. Selon Statistique Canada, l'IMC fournit « un moyen de classer le poids corporel en fonction de risque pour la santé »². Cependant, l'IMC n'est pas sans faille. Il ne fournit aucune information concernant la distribution de la matière grasse dans le corps;
- les stades d'**obésité** selon Santé Canada (catégorie de l'IMC) :
 1. Poids insuffisant (< 18,5)
 2. Poids normal (18,5 – 24,9)

¹github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Diab%C3%A8te.xlsx?raw=true

²<https://www150.statcan.gc.ca/n1/pub/82-229-x/2009001/status/abm-fra.htm>, page consultée le 20 août 2024

3. Excès de poids (25,0 – 29,9)
 4. Obésité classe I (30,0 – 34,9)
 5. Obésité classe II (35 – 39,9)
 6. Obésité classe III ($\geq 40,0$)
- la **fondation pedigree du diabète**. C'est un score mesurant le risque familial du diabète. Cette valeur mesure entre 0,08 et 2,42;
 - l'**atteinte** au diabète (avoir ou non le diabète).

1. , start=0 Non
2. , start=0 Oui

1.2 Laboratoire

Ce laboratoire a pour objectif d'introduire le logiciel Excel.

1.2.1 Interface d'Excel

Avant de commencer à utiliser un nouveau logiciel, il est essentiel de se familiariser avec son interface et ses outils. Les fichiers d'Excel sont appelés des **classeurs** et chaque classeur contient des **feuilles de calcul**. La Figure 1.2.1 présente les noms des différents éléments d'une feuille d'Excel.

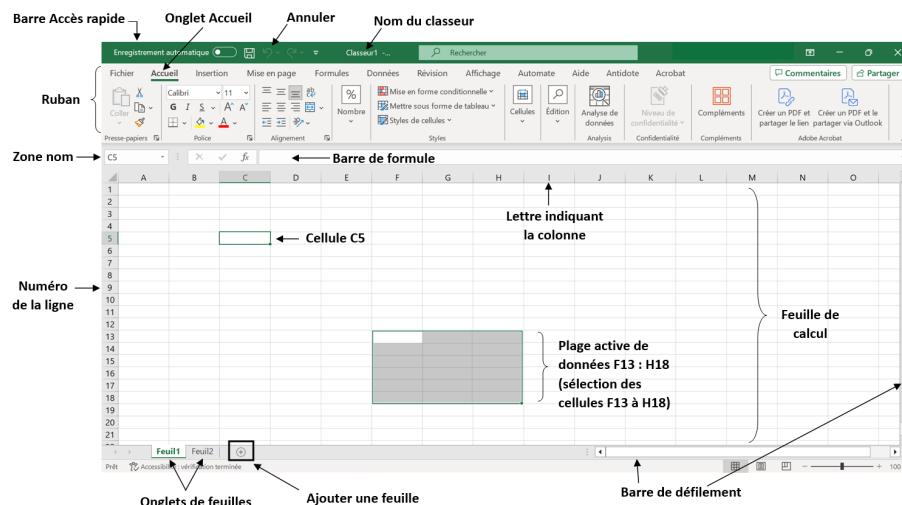


Figure 1.2.1 Interface d'une fenêtre Excel

- La **barre Accès rapide** est une barre personnalisable qui affiche les icônes des commandes les plus courantes. Par défaut, elle comprend les commandes suivantes : l'enregistrement de fichier, l'annulation d'une action effectuée et le rétablissement d'une action annulée.
- Le **nom du classeur** est écrit dans la barre d'Accès rapide. Par défaut, Excel nomme le fichier Classeur 1. Les versions plus récentes d'Excel permettent de sauvegarder automatiquement le travail fait.

- Le **ruban** est composé d'une barre de menus (Fichier, Accueil, Insertion, Mise en page, Formules, Données, Affichage, entre autres) et d'un groupe de commandes courantes dans chaque menu (outils courants regroupés par nature dans chaque onglet).
- Une **feuille de calcul** est un quadrillage comportant des cases, appelées cellules, dans lesquelles on peut écrire du texte ou effectuer des calculs à l'aide de formules et de fonctions.
- La **cellule de référence C5** est la cellule située à l'intersection de la colonne C et de la ligne 5. C5 est **l'adresse de la cellule**. Une cellule est dite **active** lorsqu'elle est sélectionnée. Dans ce cas, sa bordure est verte et épaisse. Les entêtes de sa colonne et de sa ligne sont alors grisés.
- La **plage de données F13:H18** correspond à l'ensemble des cellules contiguës comprises entre les colonnes F et H et entre les lignes 13 et 18.
- La **zone nom** permet d'afficher la référence d'une cellule (la lettre de sa colonne et le numéro de sa ligne) ou le nom qui lui a été attribué. Cette zone permet également d'afficher le nom d'une plage de données.
- La **barre de formule** permet d'afficher, d'entrer ou de modifier le contenu d'une cellule active. Le contenu peut être composé de texte, de nombres, de formules ou de fonctions.
- Les **onglets de feuilles** représentent l'ensemble des feuilles de calcul qui composent le **classeur**. En cliquant droit sur l'onglet d'une feuille, un menu contextuel apparaît. Il est possible, entre autres, de renommer la feuille, de la déplacer, de la copier, de la masquer ou de la supprimer.
- En cliquant l'icône + située à droite des onglets des feuilles, on peut **ajouter une nouvelle feuille** au classeur.
- Les **barres de défilement** permettent de déplacer horizontalement et verticalement la feuille de calcul.

1.2.2 Exploration de l'élève

L'élève est invité à explorer les onglets du ruban et les groupes de commandes. Pour ce faire, il faut ouvrir le fichier **Données_Diabète.xlsx**.

1.2.2.1 Sélection d'une ligne ou d'une colonne

Dans une feuille de calcul, il est possible de sélectionner une ligne entière ou une colonne entière.

1. Placer le curseur sur le numéro de la ligne à sélectionner jusqu'à ce qu'une flèche noire apparaisse au-dessus du numéro de la ligne (voir la [Figure 1.2.2](#)).
2. Une fois que la flèche apparaît, cliquer. La ligne entière sera sélectionnée. Elle sera grise et entourée d'un cadre vert.

Identifiante	Âge	Nombre de grossesse	Glucose	Pression artérielle diastolique	Épaisseur peau	Insuline	IMC	Obésité	Fonction pedigree du diabète
5	1	50	6	148	72	35	0	33,6	4
6	2	31	1	85	66	29	0	26,6	3
7	3	32	8	139	64	0	0	23,3	2
8	4	21	1	89	66	23	84	20,1	3
9	5	33	0	137	40	35	168	42,1	5
10	6	30	5	116	74	0	0	25,6	3
11	7	26	3	78	50	32	88	31	4
12	8	29	10	115	0	0	35,3	5	0,134
13	9	53	2	197	70	45	543	30,5	4
14	10	54	8	125	96	0	0	0	0,158
15	11	30	4	110	92	0	0	37,6	5
16	12	34	10	168	74	0	0	38	5
17	13	57	10	139	80	0	0	27,1	3
18	14	59	1	189	60	23	846	30,1	4
19	15	51	5	166	72	19	175	25,8	3
20									0,587

Figure 1.2.2 Sélection d'une ligne

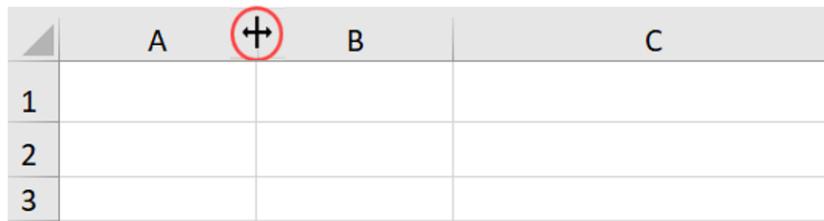
Pour la sélection d'une colonne entière, la procédure est similaire. Il faut placer le curseur sur la lettre de la colonne à sélectionner, attendre l'apparition de la flèche noire et cliquer.

Il ne faut pas oublier de sauvegarder régulièrement son travail en appuyant sur les touches **CTRL-S** ou en activant la sauvegarde automatique.

1.2.2.2 Modifier la largeur des colonnes ou des lignes

Dans une feuille de calcul, il est possible d'élargir ou de rétrécir la largeur d'une colonne ou d'une ligne.

- Placer le curseur entre deux colonnes jusqu'à ce qu'une flèche bidirectionnelle apparaisse (voir la [Figure 1.2.3](#)).

**Figure 1.2.3** Modifier la largeur d'une colonne

- Cliquer et déplacer la flèche vers la gauche pour rétrécir la largeur de la colonne de gauche et vers la droite pour élargir la largeur de la colonne de gauche. Pour modifier la largeur d'une ligne, il faut placer le curseur entre deux lignes et déplacer la flèche vers le haut ou le bas selon le résultat souhaité.
- Il est possible d'ajuster la largeur d'une colonne ou d'une ligne au contenu de celle-ci. Il suffit de doublecliquer lorsque la flèche bidirectionnelle apparaît.

1.2.3 Attribuer des noms

Dans cette section, les étapes pour nommer des cellules sont présentées. Attribuer un nom à une cellule signifie lui donner un identifiant autre que son adresse, afin de pouvoir y faire référence dans toutes les feuilles de calcul d'un classeur Excel.

1.2.3.1 Attribuer un nom à une cellule ou à une plage de cellules

La sélection de données est une opération courante dans Excel. En effet, que ce soit pour l'utilisation de formules ou la création de graphiques, il est souvent nécessaire de sélectionner des cellules. Cette tâche peut être laborieuse lorsque

la base de données est volumineuse : on risque de glisser trop loin avec la souris, de devoir revenir en arrière, de ne pas sélectionner suffisamment de cellule et d'omettre des données, etc. Pour éviter ces problèmes, il est pratique d'attribuer un nom à une cellule ou à une plage de cellules. Cela permet de faire référence à ce nom dans une formule, peu importe la feuille de calcul dans laquelle on travaille.

Dans la feuille de calcul **Données** du fichier Excel, on attribue un nom au tableau principal, soit le nom « Échantillon ». Il existe deux façons d'effectuer cette opération.

Première méthode : *zone nom.*

1. Sélectionner l'entièreté du tableau, soit la plage A5:K773. Il est possible de faire ceci en sélectionnant la cellule A5 et en tapant la combinaison **Ctrl**+**A**.
 2. Dans la **zone nom** (voir la figure Figure 1.2.1), taper le mot *Échantillon* (voir la Figure 1.2.4) et appuyer sur la touche **Enter**. Les noms ne doivent pas contenir d'espace ni de caractères spéciaux.

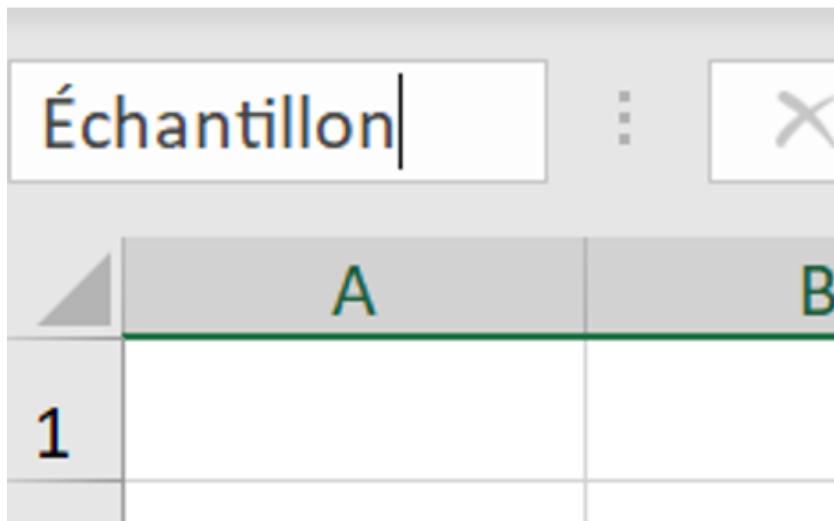


Figure 1.2.4 Attribuer un nom à une plage de cellules à partir de la zone *nom*.

La plage de données est désormais nommée *Échantillon*.

Remarque 1.2.5 Nommer une plage de cellules. Lorsque l'on attribue un nom à une plage de cellules avec la première méthode, il ne faut pas oublier d'appuyer sur la touche **Enter** pour s'assurer de l'enregistrement du nom.

Deuxième méthode: ruban.

1. Sélectionner l'entièreté du tableau, soit la plage A5:K773. Il est possible de faire ceci en sélectionnant la cellule A5 et en tapant la combinaison **Ctrl**+**A**.
 2. Cliquer sur l'onglet **Formules** du ruban (voir la Figure 1.2.1).
 3. Dans le groupe **Noms définis**, cliquer sur l'icône **Définir un nom** (voir la Figure 1.2.6). Une boîte de dialogue apparaît à l'écran.

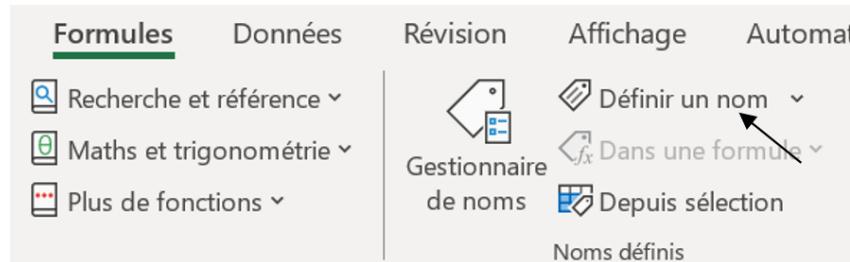


Figure 1.2.6 L'onglet Formules et le groupe Noms définis

4. Dans l'encadré vide de l'option **Nom**, taper le mot *Échantillon*, le nom attribué au tableau (voir la [Figure 1.2.7](#)). Les noms ne doivent pas contenir d'espace ni de caractères spéciaux.

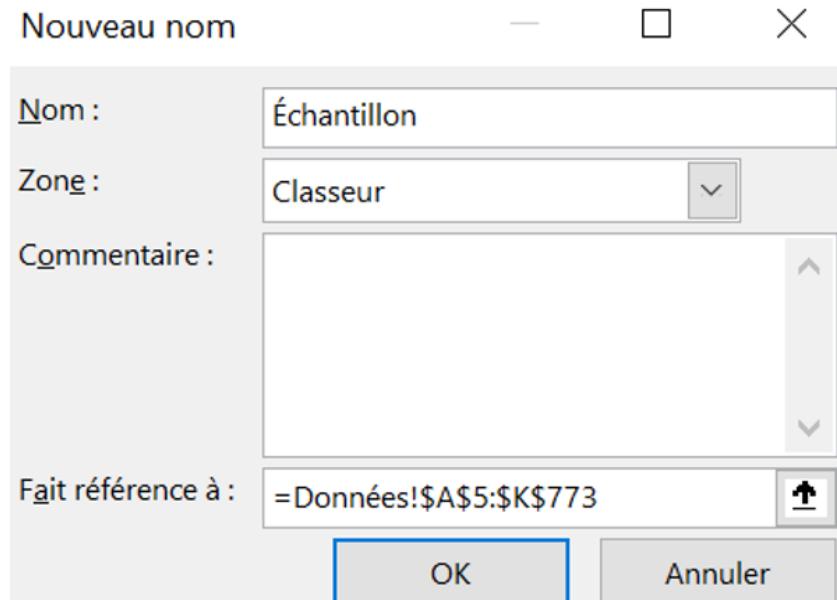


Figure 1.2.7 Attribuer un nom à une plage de cellules à partir de l'onglet **Formules**

5. S'assurer que, dans l'encadré de l'option **Zone**, il soit écrit *Classeur*. Ceci garantit que le nom attribué à une plage de cellules est défini dans toutes les feuilles du classeur. Cela permet aussi à l'utilisateur ou à l'utilisatrice d'y faire référence quelle que soit la feuille de travail.
6. Cliquer sur **OK**.

1.2.3.2 Vérification de l'attribution d'un nom

Il est important de vérifier que l'on a bien attribué un nom à une plage de cellules.

1. Dans la **zone nom**, cliquer sur la flèche du menu déroulant (voir la [Figure 1.2.8](#)).

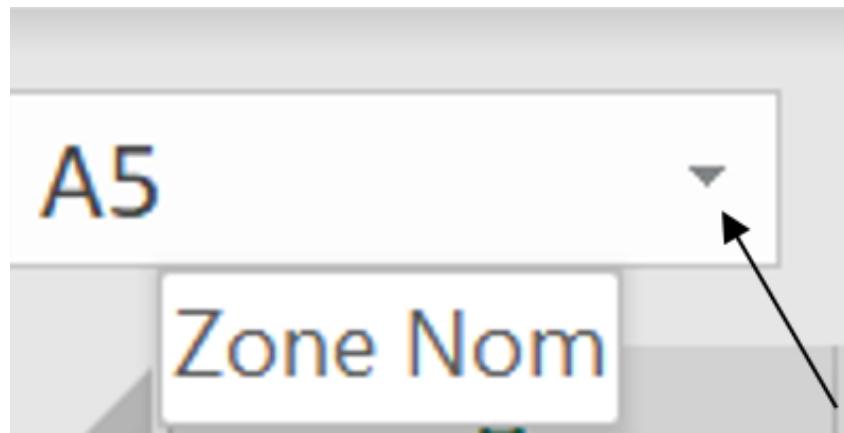


Figure 1.2.8 Vérification de l'attribution d'un nom à une plage de celles Formules

2. Le nom *Échantillon* devrait apparaître (voir la [Figure 1.2.9](#)). S'il n'apparaît pas, l'attribution de nom n'a pas été bien saisie.



Figure 1.2.9 Vérification de l'attribution d'un nom *Échantillon* au tableau principal

1.2.3.3 Attribuer un nom aux colonnes d'un tableau

Il est pratique d'attribuer un nom aux colonnes d'un tableau puisqu'il est possible de faire référence à celles-ci dans des formules Excel, et ce, peu importe la feuille de calcul dans laquelle on travaille.

1. Dans la **zone nom**, sélectionner *Échantillon* (voir la [Figure 1.2.9](#)).



Figure 1.2.10 Sélection du tableau *Échantillon*

2. Cliquer sur l'onglet **Formules** du ruban.
3. Dans le groupe **Noms définis**, cliquer sur **Depuis sélection** (voir la Figure 1.2.11). Une boîte de dialogue s'ouvre.

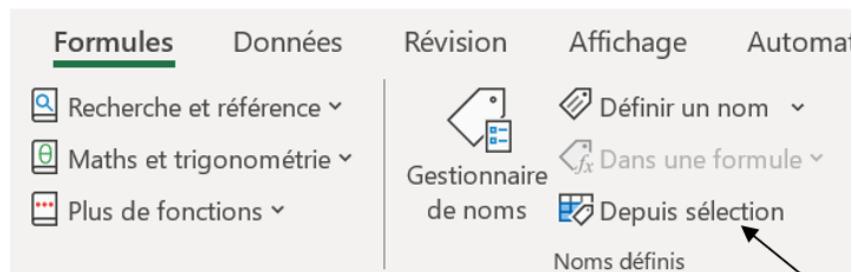


Figure 1.2.11 L'onglet **Formules** et le groupe **Noms définis**

4. Cliquer sur l'option **Ligne du haut** (voir la Figure 1.2.12). Ceci permettra à Excel d'attribuer l'entête de la première ligne comme nom à la colonne.

Créer des noms à partir de la s... ? ×

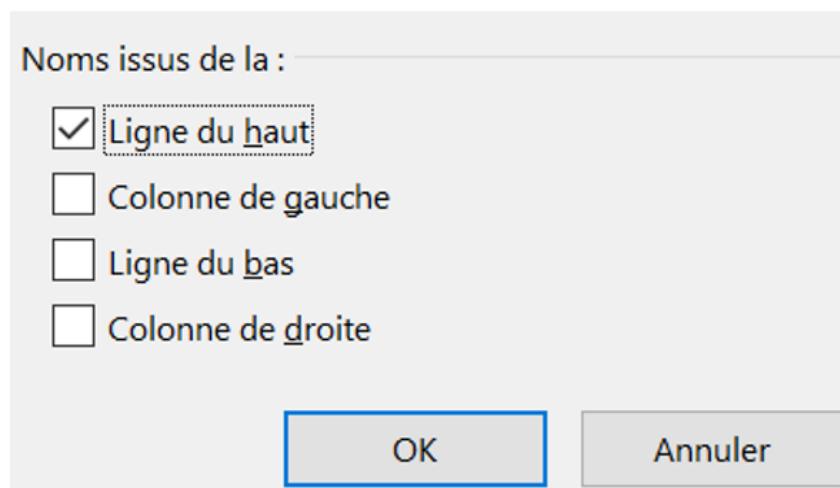


Figure 1.2.12 Sélection de l'option **Lignes du haut** dans le groupe **Noms définis**

5. Dans la **zone nom**, cliquer sur la flèche du menu déroulant pour vérifier qu'un nom a été attribué à chaque colonne (voir la figure [Figure 1.2.13](#)).

Échantillon	
Âge	B
Atteint	
Échantillon	
Épaisseur_pea	
Fonction_pedi.	
Glucose	
Identifiant	
IMC	
Insuline	
Nombre_de_g.	
Obésité	
Pression_arté..	

The screenshot shows a software interface with a sidebar on the left containing a list of column names: Échantillon, Âge, Atteint, Échantillon, Épaisseur_pea, Fonction_pedi., Glucose, Identifiant, IMC, Insuline, Nombre_de_g., Obésité, and Pression_arté.. The column 'Âge' is highlighted with a blue border. To the right of the sidebar is a table with two columns. The first column contains the column names from the sidebar. The second column contains numerical values: 50, 31, 32, 21, 33, 30, and 26. The row for 'Âge' has a green header bar above it, and the value '50' is also highlighted with a green background.

Figure 1.2.13 Vérification de l'attribution du nom de chaque colonne

6. Sauvegarder le travail.

1.2.4 Figer et libérer les volets

Lorsqu'on souhaite parcourir un tableau de grande taille (beaucoup de lignes ou de colonnes), il est pratique de maintenir les titres des colonnes ou des lignes

visibles pendant le défilement de la feuille de calcul.

1.2.4.1 Figer la ligne de titres d'un tableau

1. Sélectionner la deuxième ligne du tableau, soit la ligne 6 (voir la [Figure 1.2.14](#)). Il faut placer son curseur à la ligne 6 et cliquer.

Identifiant	Âge	Nombre de grossesses	Glucose	Pression artérielle diastolique	Épaisseur peau	Insuline	IMC	Obésité	Fonction pedigree du diabète
1	50	6	148	72	35	0	33,6	4	0,627
2	31	1	85	66	29	0	26,6	3	0,351
3	32	8	183	64	0	0	23,3	2	0,672
4	21	1	89	66	23	94	28,1	3	0,167
5	33	0	137	40	35	168	43,1	6	2,288
6	30	5	116	74	0	0	25,6	3	0,201
7	26	3	78	50	32	88	31	4	0,248
8	29	10	115	0	0	0	35,3	5	0,134
9	53	2	197	70	45	543	30,5	4	0,158
10	54	8	125	96	0	0	0	1	0,232
11	30	4	110	92	0	0	37,6	5	0,193
12	34	10	168	74	0	0	38	5	0,537
13	57	10	129	80	0	0	37,1	3	1,441
14	59	1	189	60	23	846	30,1	4	0,398
15	51	5	166	72	19	175	25,8	3	0,587
...

Figure 1.2.14 Sélection de la deuxième ligne du tableau *Échantillon*

2. Sélectionner l'onglet **Affichage**. Dans le groupe **Fenêtre**, cliquer sur la flèche du menu déroulant de l'icône **Figer les volets** (voir la [Figure 1.2.15](#)).

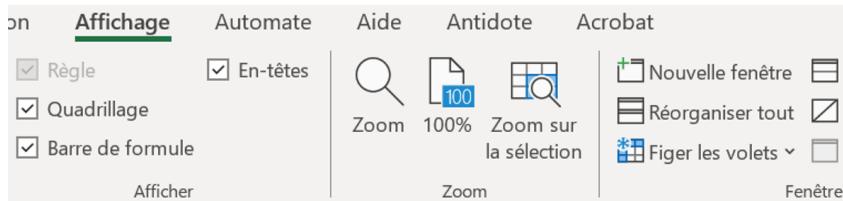


Figure 1.2.15 Sélection de l'onglet *Affichage* et de l'icône *Figer les volets*

3. Cliquer sur **Figer les volets** (voir [Figure 1.2.16](#)).

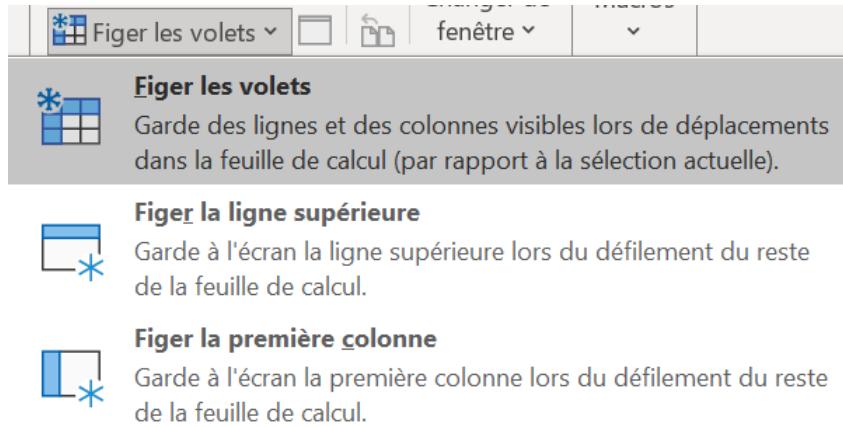


Figure 1.2.16 Sélection de l'option *Figer les volets*

Maintenant, lorsque l'on fait défiler le tableau, les titres des colonnes restent visibles. Il est possible de verrouiller à la fois des lignes et des colonnes, ainsi que de libérer les volets, si nécessaire.

1.2.4.2 Libérer les volets

Si l'on veut libérer les volets, on peut les supprimer.

1. Sélectionner l'onglet **Affichage** (voir la figure [Figure 1.2.15](#)).
2. Dans le groupe **Fenêtre**, cliquer sur la flèche du menu déroulant de l'icône **Figer les volets** (voir la [Figure 1.2.17](#)). Cliquer sur l'option **Libérer les volets**.

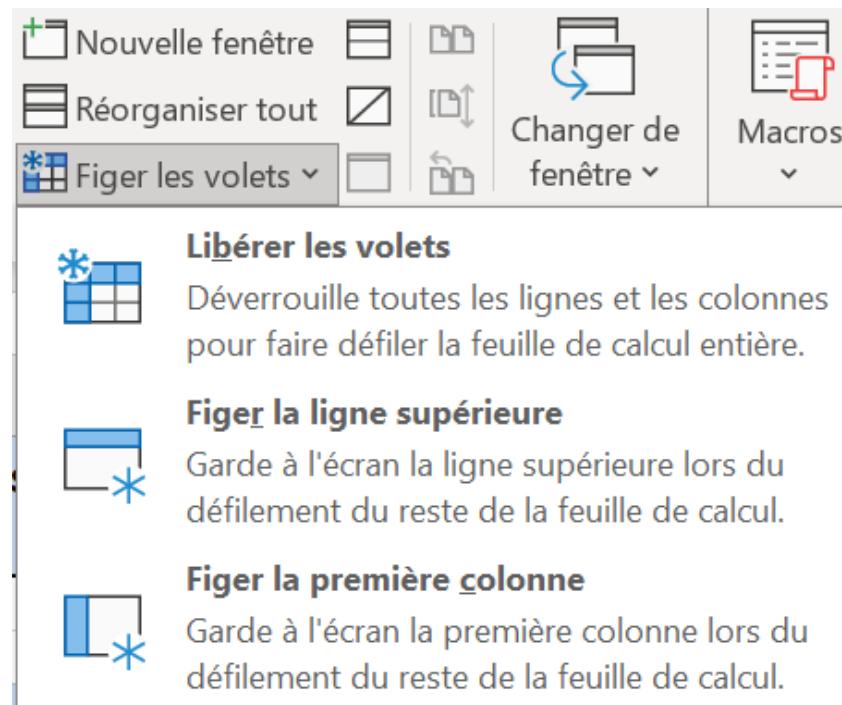


Figure 1.2.17 Sélection de l'option **Libérer les volets**

1.2.5 Filtrer et trier des données

Dans Excel, il est possible d'ajouter des filtres permettant de dépister des valeurs aberrantes ou manquantes, ainsi que de choisir ou masquer certaines modalités. Il est également possible de trier des séries statistiques dans un certain ordre voulu.

1.2.5.1 Ajouter des filtres

Les étapes qui mènent au filtrage de données sont présentées.

1. Sélectionner le tableau **Échantillon**.
2. Sélectionner l'onglet **Données** (voir la [Figure 1.2.18](#)). Cliquer sur l'icône **Filtrer** du groupe **Trier et filtrer**.

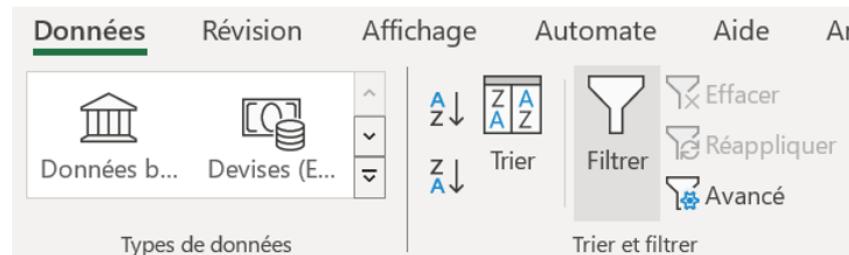


Figure 1.2.18 Sélection de l'icône **Filtrer**

3. Chaque variable affiche désormais un petit triangle dans une boîte grise à droite de son nom (voir la [Figure 1.2.19](#)).

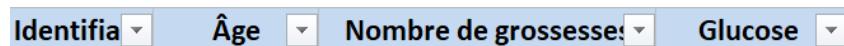


Figure 1.2.19 Triangle affichant toutes les modalités d'une variable qualitative et toutes les valeurs d'une variable quantitative

4. En cliquant sur ce triangle, toutes les modalités d'une variable qualitative et toutes les valeurs d'une variable quantitative sont affichées (voir la [Figure 1.2.20](#)). Il est possible de décocher certaines catégories pour les masquer et n'afficher que celles souhaitées.

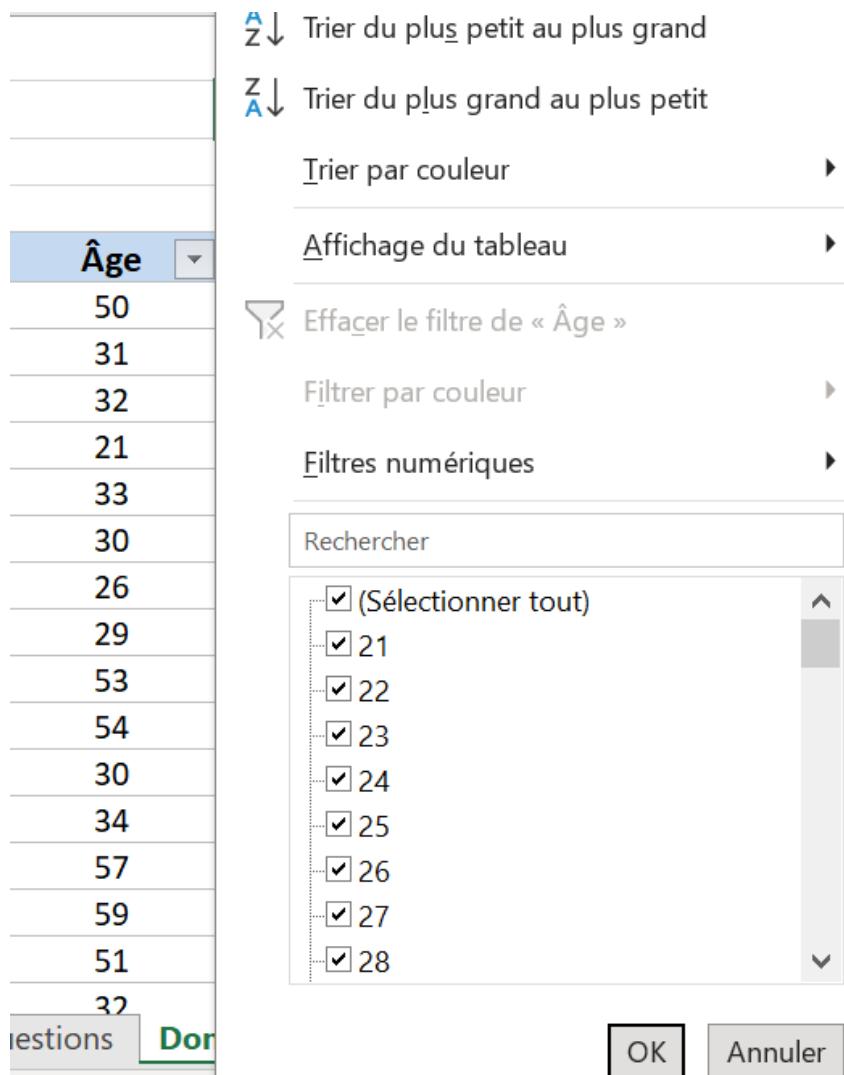


Figure 1.2.20 Toutes les valeurs de la variable *Âge*

1.2.5.2 Trier à partir des filtres

Il existe deux façons de trier les données d'un tableau. La première se fait en utilisant des filtres.

1. En cliquant sur le petit triangle de filtre de la variable *Âge*, une boîte de dialogue s'ouvre. La première option permet de trier les valeurs par

ordre croissant (plus petit au plus grand, A à Z pour les variables qualitatives), tandis que la deuxième option permet de trier les valeurs par ordre décroissant (du plus grand au plus petit, Z à A pour les variables qualitatives) (voir la Figure 1.2.21).



Figure 1.2.21 Options de tris

1.2.5.3 Trier à partir de l'onglet Données

La deuxième façon de trier des données passe par l'onglet **Données**.

1. Cliquer sur l'onglet **Données**. Cliquer l'icône **Trier** (voir la Figure 1.2.22).

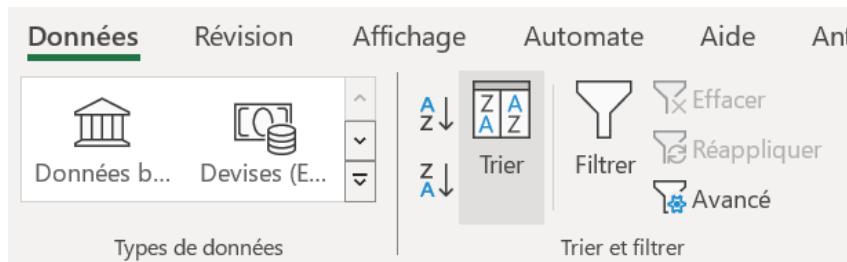


Figure 1.2.22 Options de tris

2. Une boîte de dialogue s'ouvre.

Il est possible de sélectionner la variable à trier et d'appliquer un tri croissant, un tri décroissant ou un tri personnalisé (voir la Figure 1.2.23).

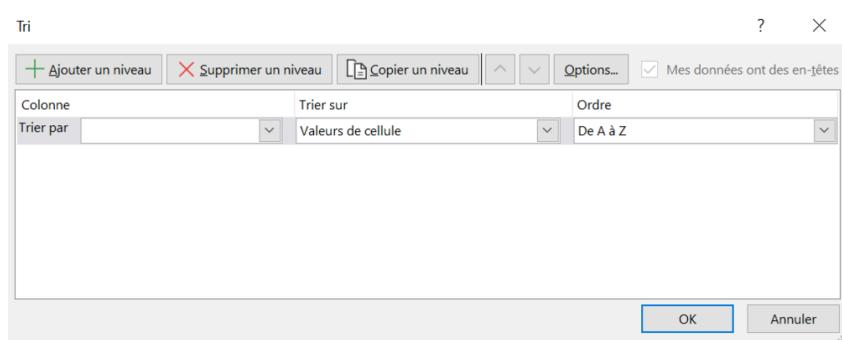


Figure 1.2.23 Boîte de dialogue pour le tri et options de variables à trier

3. Sauvegarder.

1.2.6 Ajouter et nommer une feuille de calcul

Lorsqu'on réalise plusieurs études à partir d'une même base de données, il est préférable de travailler dans un seul classeur et d'effectuer chaque étude dans une feuille de calcul distincte afin de faciliter la consultation.

1. Cliquer sur le symbole + (voir la Figure 1.2.24).



Figure 1.2.24 Ajouter une nouvelle feuille de calcul

2. Une feuille intitulée **Feuil1** s'ouvrira.



Figure 1.2.25 Une nouvelle feuille de calcul intitulée *Feuil1*

3. Déplacer la feuille en dernière place, après les feuilles *Questions* et *Données*, si Excel ne le fait pas par défaut.
4. Cliquer droit sur le nom de la feuille.
5. Sélectionner l'icône *Renommer* et taper les mots *Étude Atteint*.

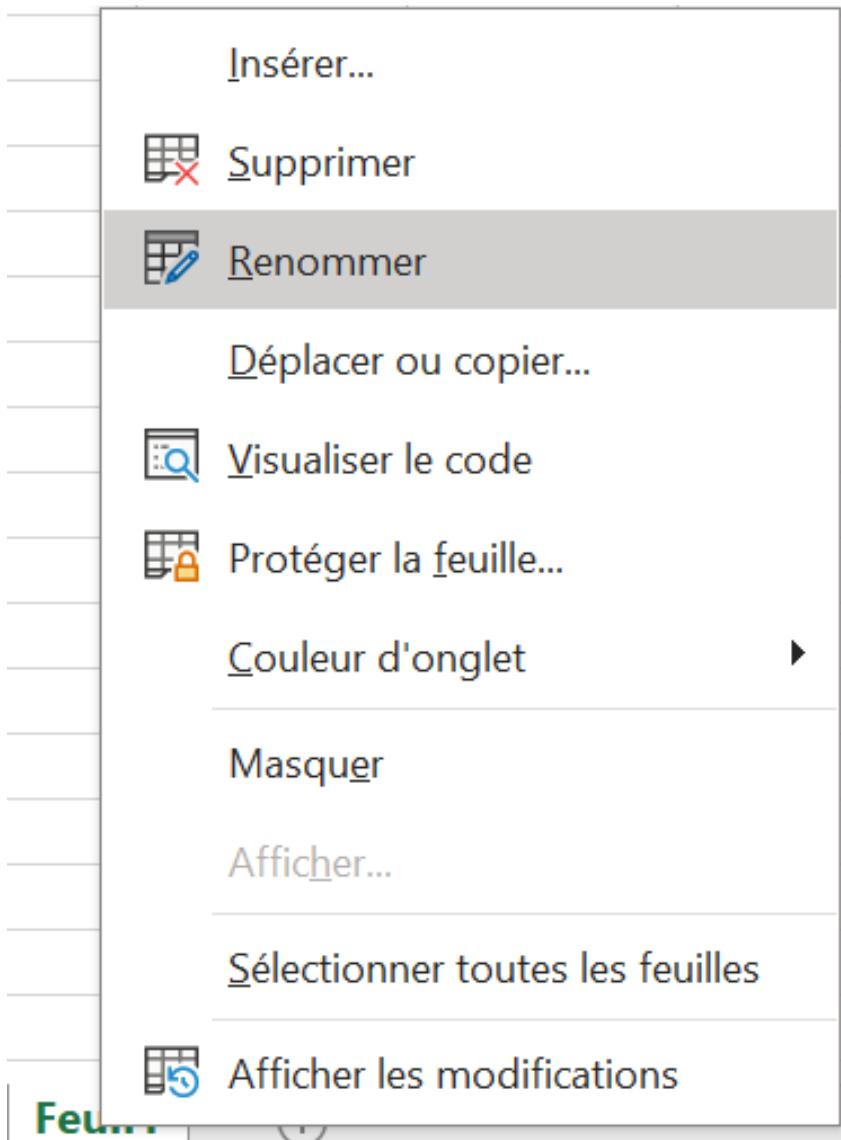


Figure 1.2.26 Sélection de l'icône *Renommer*

1.2.7 Remplissage automatique

1.2.7.1 Remplissage automatique d'une valeur

Il est possible de remplir une colonne avec la même valeur en utilisant la fonction de recopie automatique d'Excel. Par exemple, dans un problème de physique ou de chimie, il se peut que l'incertitude soit la même pour toutes les mesures et l'on aimerait la recopier sans taper manuellement la valeur plusieurs fois. Les étapes pour faire un remplissage automatique sont présentées ci-dessous.

1. Dans le classeur de travail, ajouter une feuille de travail ([Figure 1.2.24](#)).
2. Dans la cellule C5 de cette nouvelle feuille, taper la valeur 1 ([Figure 1.2.27](#)).

	A	B	C	D
1				
2				
3				
4				
5			1	
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				

Figure 1.2.27 Remplissage automatique d'une valeur à recopier

3. Sélectionner la cellule jusqu'à ce qu'elle soit encadrée d'une bordure verte ([Figure 1.2.27](#)).
4. Cliquer et maintenir enfoncé le bouton gauche de la souris sur le coin inférieur droit de la cellule (un petit carré vert). Glisser le pointeur vers le bas jusqu'à la cellule voulue, la cellule C17 dans ce cas ([Figure 1.2.27](#)).

5. Relâcher le bouton gauche de la souris. La valeur 1 apparaît dans toutes les cases sélectionnées, soit la plage C5:C17 ([Figure 1.2.27](#)). Le remplissage peut également être fait le long d'une ligne.

1.2.7.2 Remplissage automatique d'une suite de nombre

Il est possible de remplir une colonne avec une suite de nombres en utilisant la fonction de recopie automatique d'Excel. Les étapes sont présentées ci-dessous.

1. Dans les cellules E5:E7 de la même feuille de calcul, taper les valeurs 1, 2 et 3 respectivement ([Figure 1.2.28](#)).

	A	B	C	D	E	F
1						
2						
3						
4						
5				1	1	
6				1	2	
7				1	3	
8				1		
9				1		
10				1		
11				1		
12				1		
13				1		
14				1		
15				1		
16				1		
17				1		
18						
19						
20						
21						
22						
23						
24						

Figure 1.2.28 Remplissage automatique d'une suite de nombres

2. Sélectionner les cellules E5:E7 jusqu'à ce qu'elles soient encadrées d'une bordure verte ([Figure 1.2.28](#)).
3. Cliquer et maintenir enfoncé le bouton gauche de la souris sur le coin inférieur droit de la plage de cellules (un petit carré vert). Glisser le pointeur vers le bas jusqu'à la cellule voulue, la cellule E17 dans ce cas ([Figure 1.2.28](#)). Relâcher le bouton gauche de la souris une fois la cellule voulue atteinte.

En sélectionnant trois cellules qui ont trois nombres consécutifs, Excel comprend qu'en effectuant un remplissage automatique, on veut poursuivre la séquence et non recopier ces trois valeurs. Ainsi, il remplit les cellules avec les nombres 4 à 13 ([Figure 1.2.28](#)).

1.2.7.3 Remplissage automatique d'une valeur en doublecliquant

Il est possible de remplir une colonne adjacente à une autre en utilisant le doubleclique. Les étapes sont présentées ci-dessous.

1. Dans les cellules F5:F6 de la même feuille de calcul, taper la valeur 2 ([Figure 1.2.29](#)). On place ces valeurs dans la colonne adjacente à celles déjà remplies avec les valeurs 1 à 13.

	A	B	C	D	E	F	G
1							
2							
3							
4							
5			1		1	2	
6			1		2	2	
7			1		3		
8			1		4		
9			1		5		
10			1		6		
11			1		7		
12			1		8		
13			1		9		
14			1		10		
15			1		11		
16			1		12		
17			1		13		
18							
19							
20							

Figure 1.2.29 Remplissage automatique d'une valeur ou d'une formule avec le doubleclique

2. Sélectionner les cellules F5:F6 jusqu'à ce qu'elles soient encadrées d'une bordure verte ([Figure 1.2.29](#)).
3. Doublecliquer le bouton gauche de la souris sur le coin inférieur droit de la plage de cellules (un petit carré vert) (voir la [Figure 1.2.29](#)). La plage F5:F17 se remplit automatiquement. Le remplissage arrête à la ligne 17, car Excel remplit jusqu'à la même ligne que la colonne adjacente à la colonne F.

1.2.8 Tableau croisé dynamique

Lorsque l'on dispose d'une vaste base de données, il peut être difficile de mettre en évidence les faits saillants. Excel permet de créer des tableaux croisés dynamiques pour organiser, présenter, synthétiser et analyser des données. Pour ce faire, il est crucial que les données de chaque variable soient initialement organisées en colonnes dans un seul tableau maître. Le mot « croisé » fait référence à la possibilité de regrouper et de comparer plusieurs variables, tandis que le mot « dynamique » souligne la capacité du tableau à évoluer. En effet, si une valeur est modifiée dans la base de données, le tableau croisé dynamique peut être mis à jour. À partir d'un tableau croisé dynamique, il est possible de générer un tableau de fréquences pour visualiser la répartition des unités statistiques en fonction d'une variable.

La première enquête du laboratoire 2 porte sur la répartition des femmes d'origine pima de l'Arizona selon la présence de diabète. Ainsi, on génère un tableau croisé dynamique de cette distribution. Il existe deux méthodes pour le générer.

1.2.8.1 Générer un tableau croisé dynamique (variables qualitatives)

1. Dans la feuille ***Étude Atteint***, sélectionner la cellule B3, la cellule dans laquelle le tableau croisé dynamique sera inséré.
2. Cliquer sur l'onglet **Insertion** du ruban (voir la [Figure 1.2.30](#)).

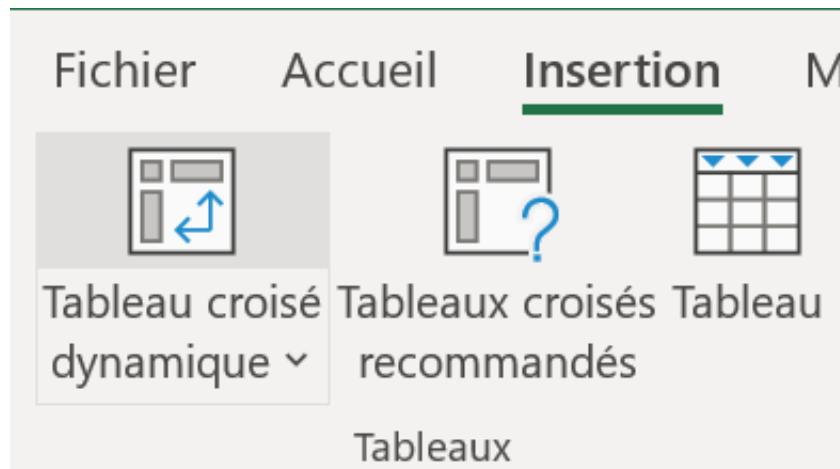


Figure 1.2.30 Sélection de l'onglet **Insertion**

3. Cliquer la première icône, soit **Tableau croisé dynamique** (voir la [Figure 1.2.30](#)).
4. Une boîte de dialogue apparaît à l'écran.

Dans la zone de saisie **Tableau/Plage**, écrire le nom donné au tableau principal, soit *Échantillon* (voir la [Figure 1.2.31](#))

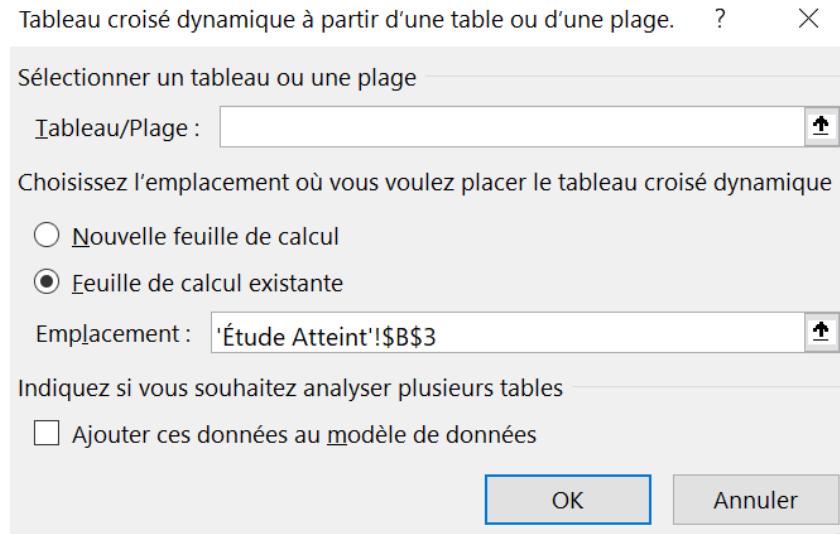


Figure 1.2.31 Boîte de dialogue pour l'insertion d'un tableau croisé dynamique et saisie du nom donné au tableau principal, soit *Échantillon*

5. Cliquer sur **OK**.

6. Un tableau croisé dynamique vide sera déposé dans la feuille ***Étude Atteint*** (voir Figure 1.2.32).

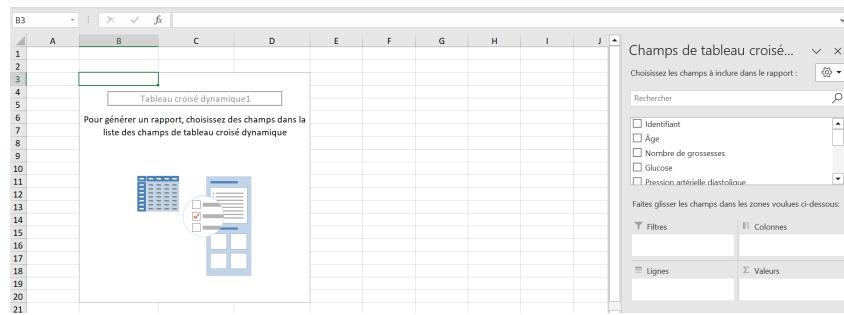


Figure 1.2.32 Tableau croisé dynamique généré

1. Dans la feuille ***Données***, sélectionner l'entièreté du tableau, soit la plage A5:K773. Il est possible de faire ceci en sélectionnant une cellule non vide du tableau ***Échantillon*** et en tapant la combinaison ***Ctrl***+***A***.
2. Cliquer sur l'onglet **Insertion** du ruban (voir la Figure 1.2.30).

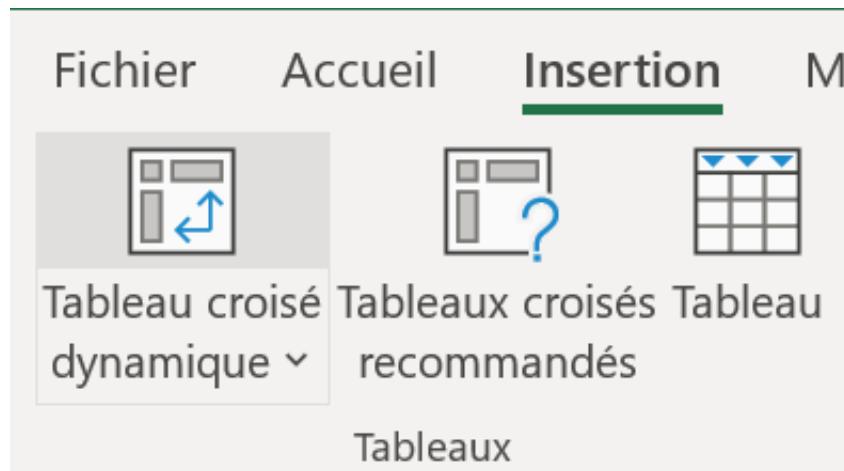


Figure 1.2.33 Sélection de l'onglet **Insertion**

3. Cliquer sur la première icône, soit **Tableau croisé dynamique** (voir la Figure 1.2.30).
 4. Une boîte de dialogue s'ouvre. La plage de données à partir de laquelle on veut créer un tableau croisé dynamique est bien sélectionnée.
- Cliquer sur l'option **Feuille de calcul existante** pour déterminer l'emplacement du tableau croisé dynamique (voir la Figure 1.2.34)

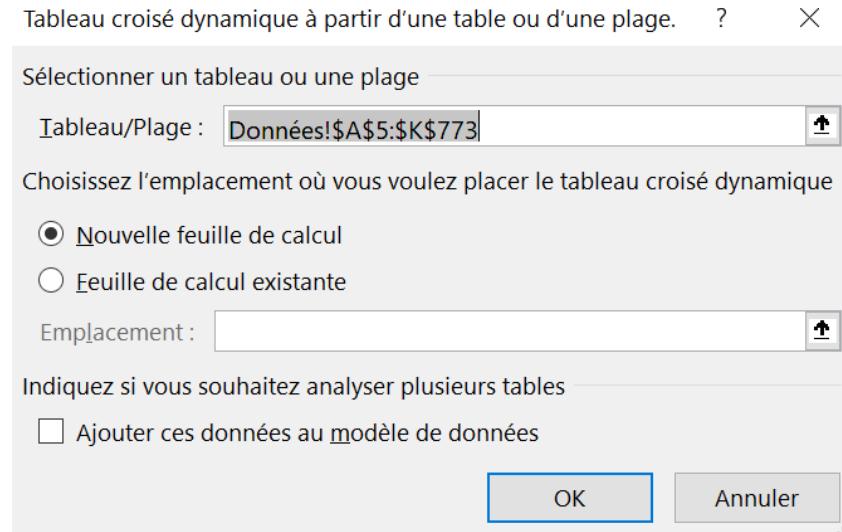


Figure 1.2.34 Sélection de la plage de données à partir de laquelle un tableau croisé dynamique sera généré et sélection de l'option ***Feuille de calcul existante*** et de la flèche pour l'emplacement du tableau croisé dynamique

5. Cliquer sur l'onglet de la feuille de calcul ***Étude Atteint***.
6. Une fois dans la feuille ***Étude Atteint***, cliquer sur la cellule **B3** de la feuille.
7. Cliquer l'onglet ***OK*** (voir la [Figure 1.2.35](#)).

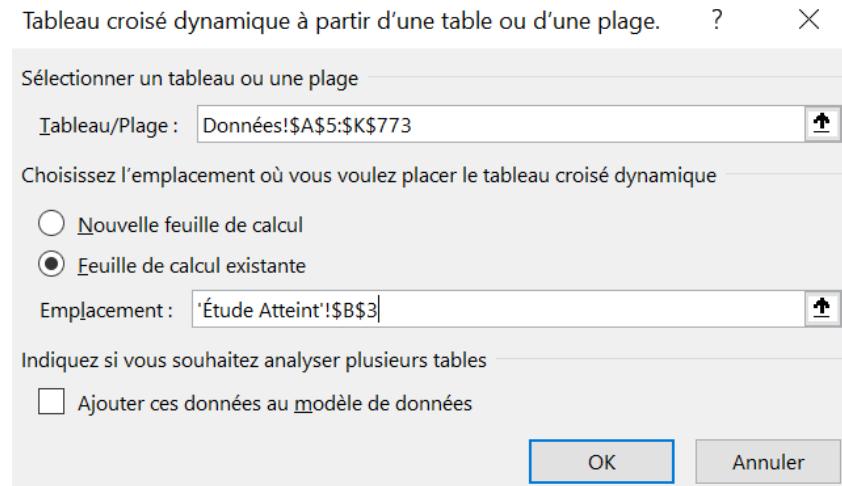


Figure 1.2.35 Confirmation de l'emplacement du tableau croisé dynamique

8. Un tableau croisé dynamique vide sera déposé dans la feuille ***Étude Atteint*** (voir [Figure 1.2.32](#)).

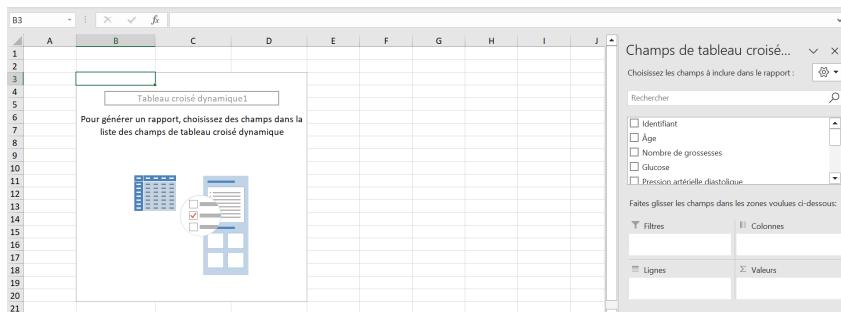


Figure 1.2.36 Tableau croisé dynamique généré

Une fois le tableau croisé dynamique généré, il faut le configurer. À la droite de la feuille de calcul, les champs du tableau croisé dynamique s'affichent. Les champs correspondent aux entêtes des colonnes du tableau source. On trouve cinq encadrés : le premier contient les champs. Les encadrés **Colonnes**, **Lignes** et **Valeurs** représentent des zones où l'on peut ajouter des champs. L'encadré **Valeurs** permet d'effectuer des calculs au sein du tableau croisé dynamique, tandis que l'encadré **Filtres** sert à appliquer des filtres aux données lors de la création du tableau.

1.2.8.2 Remplir un tableau croisé dynamique

Deuxième méthode. Avec 768 femmes étudiées, il est évident qu'un décompte manuel des femmes diabétiques est à la fois long, laborieux et sujet à de nombreuses erreurs. Les étapes pour créer le tableau croisé dynamique de la répartition des femmes pima selon la présence de diabète sont présentées ci-dessous.

1. Glisser et déposer la variable **Atteint** dans la zone de saisie **Lignes** (voir la [Figure 1.2.37](#)). Pour ce faire, il faut cliquer sur la variable **Atteint** avec le bouton de gauche de la souris, garder son doigt enfoncé et glisser la variable dans la zone **Lignes**.

Champs de tableau croisé... ▼ X

Choisissez les champs à inclure dans le rapport : 

Rechercher 

IVC
 Obésité
 Fonction pedigree du diabète
 Atteint
[Plus de tableaux...](#)

Faites glisser les champs dans les zones voulues ci-dessous:

▼ Filtres	 Colonnes
≡ Lignes	Σ Valeurs
Atteint	

Figure 1.2.37 Glissement de la variable *Atteint* dans la zone de saisie *Lignes*

Les modalités de la variable *Atteint* s'affichent dans la première colonne du tableau croisé dynamique (voir la Figure 1.2.38).

Étiquettes de lignes ▼	
0	
1	
Total général	

Figure 1.2.38 Première colonne du tableau croisé dynamique

2. Glisser et déposer la variable **Atteint** dans la zone de saisie **Valeurs** cette fois-ci (voir la [Figure 1.2.39](#)).

Par défaut, Excel effectue la somme des valeurs comme opération. Cependant, on veut compter le nombre de femmes dans chaque catégorie. Dans la zone de saisie **Valeurs**, cliquer sur la flèche du menu déroulant, puis sélectionner l'option **Paramètres des champs de valeurs** (voir la [Figure 1.2.39](#)) pour modifier le calcul.

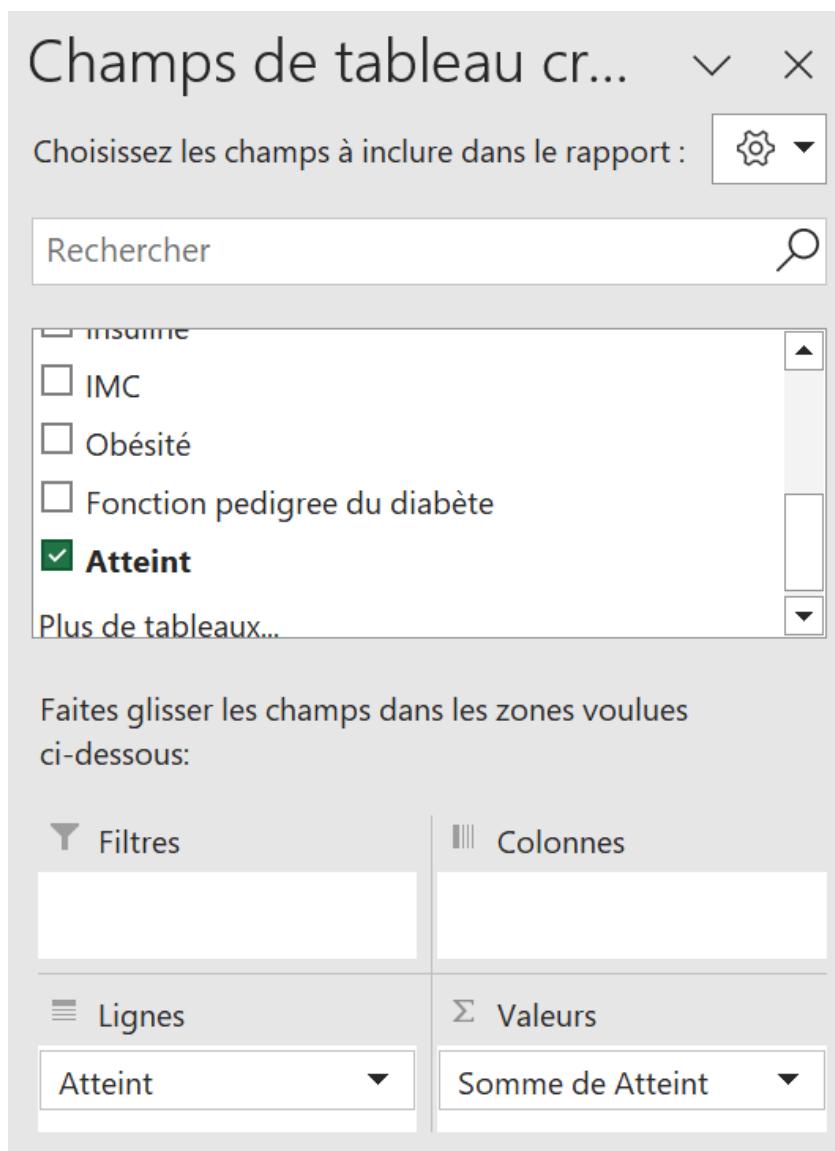


Figure 1.2.39 Glissement de la variable **Atteint** dans la zone de saisie **Valeurs**

Une nouvelle boîte de dialogue apparaît.

3. À l'onglet **Synthèse des valeurs par**, sélectionner l'option **Nombre** (voir la [Figure 1.2.40](#)).

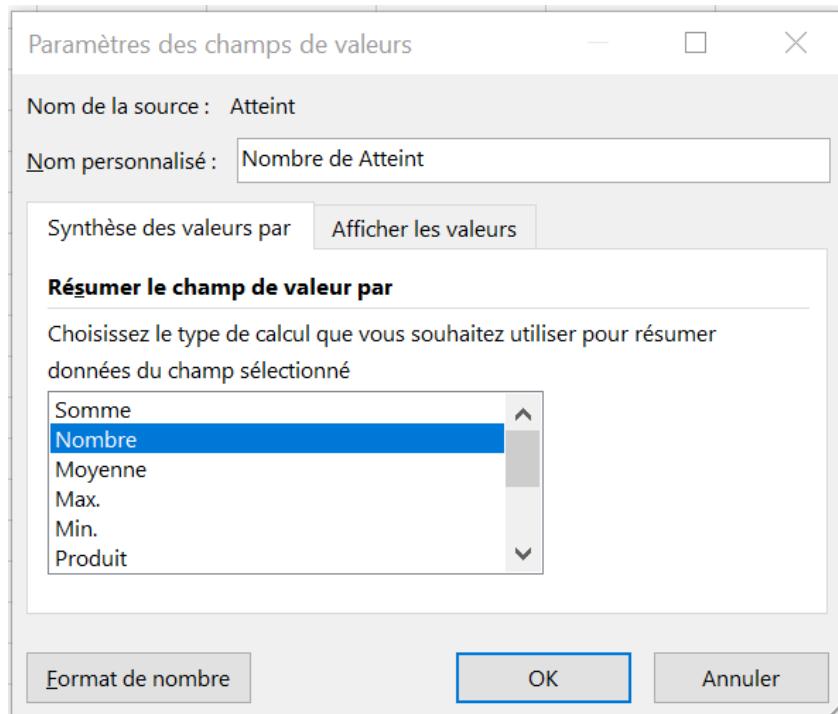


Figure 1.2.40 Sélection de l'option *Nombre* dans l'onglet *Synthèse des valeurs par*

4. Cliquer sur *OK*.

Cette action ajoute une deuxième colonne au tableau croisé dynamique, une colonne qui compte le nombre de femmes d'origine pima dans chaque catégorie de la variable *Atteint* (voir la Figure 1.2.41).

Étiquettes de lignes	Nombre de Atteint
0	500
1	268
Total général	768

Figure 1.2.41 Deuxième colonne du tableau croisé dynamique

5. Refaire les étapes 2 à 4 pour ajouter une troisième colonne au tableau croisé dynamique. On veut ajouter une colonne pour le pourcentage de femmes d'origine pima atteintes ou non du diabète.
6. À l'onglet *Synthèse des valeurs par*, sélectionner l'option *Nombre* (voir la Figure 1.2.40). Attention, il ne faut pas cliquer sur *OK* tout de suite!
7. Cliquer sur l'onglet *Afficher les valeurs*, l'onglet à droite de *Synthèse des valeurs par*. Cliquer sur la flèche du menu déroulant et sélectionner l'option *% du total général* (voir la figure Figure 1.2.42)

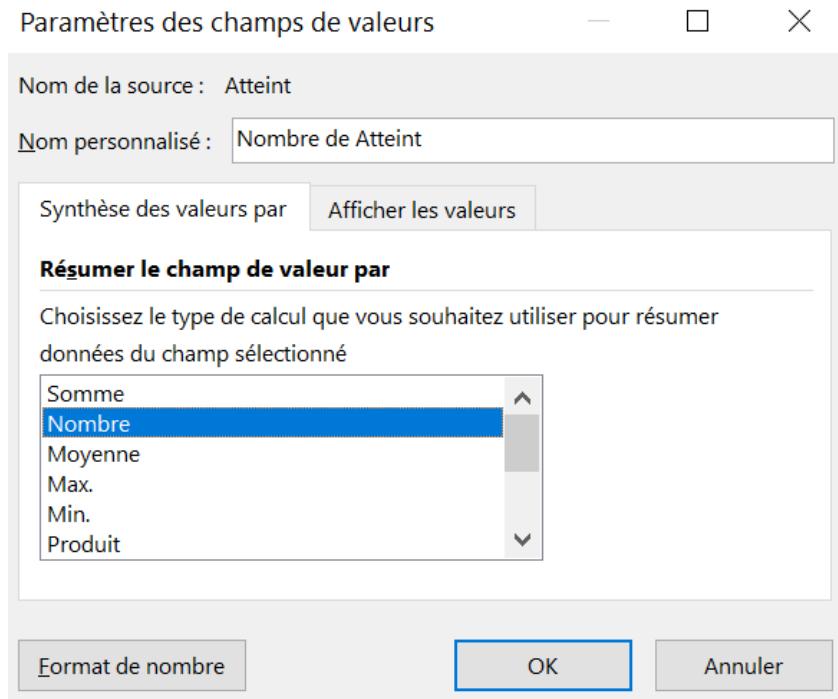


Figure 1.2.42 Sélection de l'onglet *Afficher les valeurs* et de l'option *% du total général*

8. Cliquer sur **OK**. La troisième colonne présente le pourcentage de femmes d'origine pima atteintes ou non du diabète.

Étiquettes de lignes	Nombre de Atteint	Nombre de Atteint2
0	500	65,10%
1	268	34,90%
Total général	768	100,00%

Figure 1.2.43 Les trois colonnes du tableau croisé dynamique final

9. Sauvegarder le travail.

1.2.8.3 Actualiser un tableau croisé dynamique

Si une valeur du tableau principal est modifiée, il est possible d'actualiser le tableau croisé dynamique pour refléter les changements effectués.

1. Cliquer avec le bouton de droite de la souris sur une des colonnes du tableau croisé dynamique. Un menu déroulant s'affiche. Cliquer l'option **Actualiser**.

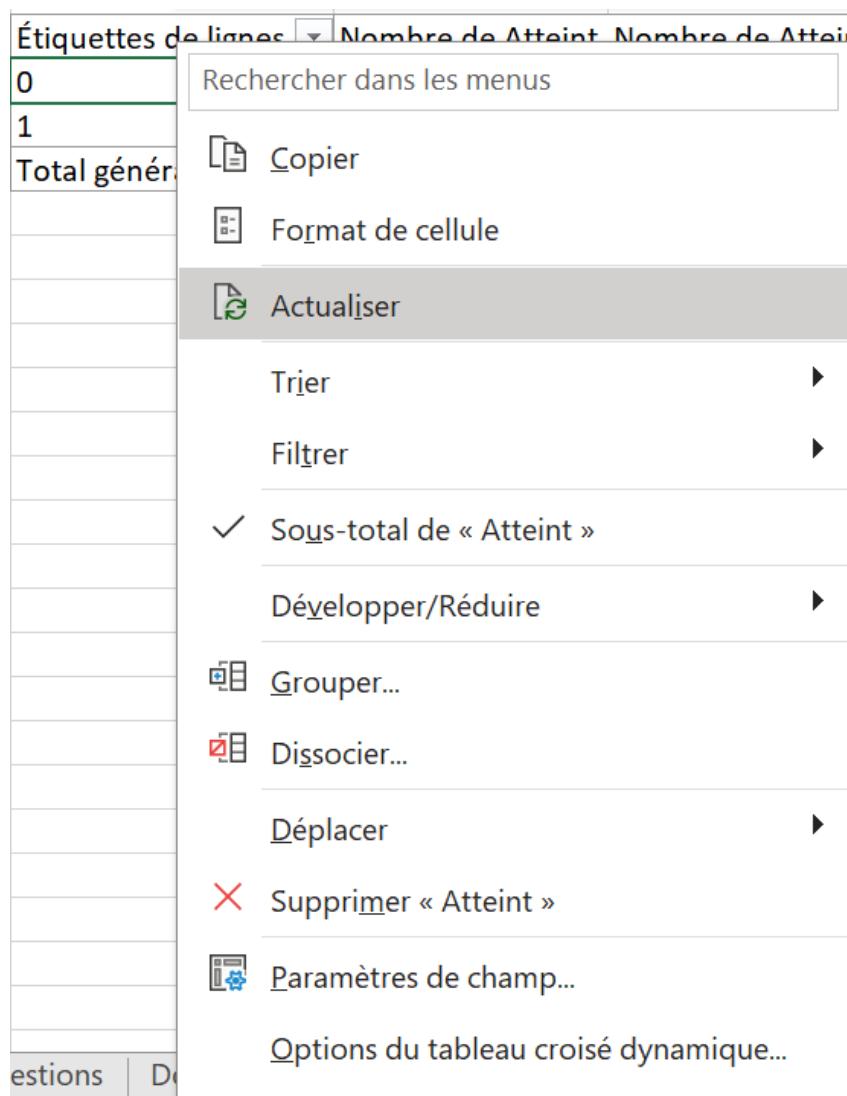


Figure 1.2.44 Sélection de l'option *Actualiser* d'un tableau croisé dynamique

2. Sauvegarder le travail.

Dans le prochain laboratoire seront abordées les étapes pour améliorer la présentation de ce tableau afin qu'il soit conforme aux normes statistiques

1.3 Réflexions

Avant d'entamer une collecte de données, toute personne chercheuse doit suivre un processus de recherche rigoureux et essentiel pour porter un jugement critique sur le sujet étudié. La première étape consiste à définir les objectifs de la recherche. En général, des hypothèses de recherche sont formulées avant de procéder à la collecte de données. Ce postlab vise à développer une compréhension intuitive des informations que révèlent les données avant d'entreprendre une étude quantitative plus approfondie. On rappelle que l'objectif de cette

étude est de prédire, à partir de ses indicateurs de santé, si une femme d'origine pima est atteinte ou non du diabète.

Travail à faire après le laboratoire

Objectifs

- Examiner les séries statistiques.
- Effectuer une revue de la littérature.
- Poser un regard critique sur les données.
- Formuler des hypothèses de recherche.

1. Effectuer les tris suivants dans le tableau *Échantillon*:

- Trier en ordre décroissant selon la colonne *Nombre de grossesses*;
- Trier en ordre croissant selon la colonne *Épaisseur peau*.

Dans la feuille *Étude atteint*, cliquer avec le bouton de droite sur le tableau croisé dynamique et cliquer sur *Afficher la liste de champs*. Glisser le champ *Nombre de grossesses* dans la zone filtre. Un nouvel élément apparaît au-dessus du tableau croisé dynamique. Explorer cet ajout et son effet sur le tableau. S'assurer de remettre la valeur du filtre sur (*Tous*) une fois l'exploration terminée.

2. En se basant sur les manipulations de l'exercice précédent, quelle(s) observation(s) peuvent être formulées à propos des différentes variables de l'étude?

3. Une hypothèse de recherche est un énoncé « provisoire à une ou plusieurs questions de recherche »¹. À la suite des observations faites à l'exercice Activité 1.3.2, formuler des hypothèses de recherche qui paraissent plausibles, sont appuyées par un contexte théorique et pourraient répondre aux observations formulées à l'exercice précédent.
4. Il peut être pertinent de comparer la santé physique d'un groupe à celle d'un autre. Identifier quelques groupes avec lesquels les femmes d'origine pima de l'Arizona pourraient être comparées, en précisant la pertinence de ces comparaisons.

¹MORNEAU, S. (2024). Probabilités et statistique en sciences de la nature. Les Éditions CEC. p.4

Chapitre 2

Variables qualitatives

Ce chapitre introduit le traitement et la présentation des variables qualitatives. Il aborde des techniques telles que la création de tableaux croisés dynamiques pour synthétiser les données qualitatives, la mise en forme de ces tableaux pour une présentation optimale et l'élaboration de graphiques adaptés à ces types de variables. Ces méthodes facilitent l'interprétation et la communication des informations issues des données qualitatives.

2.1 Prélab

Les variables qualitatives sont un type de variables pour lesquelles les modalités ne sont pas des nombres, mais plutôt des catégories. On retrouvera souvent parmi ces variables le sexe, la couleur, le niveau de satisfaction, etc. Elles se déclinent en deux catégories, soit les variables qualitatives nominales et ordinaires. Pour synthétiser visuellement l'information d'une variable qualitative, on utilise principalement un tableau de fréquences ainsi qu'un diagramme circulaire ou un diagramme à bandes.

Travail à faire avant le cours

Objectifs

- Créer les tableaux croisés dynamiques nécessaires pour le laboratoire
- Se conscientiser à l'importance de présenter des résultats de la bonne manière.

On continue de travailler à l'aide de la base de données du laboratoire d'[introduction 1](#). Le but de ce prélab est d'utiliser les connaissances et les outils acquis dans le premier laboratoire afin de préparer une feuille de calcul qui servira lors du prochain laboratoire.

1. Dans le fichier Excel du laboratoire, créer une nouvelle feuille de calcul intitulée **Étude Obésité**. Dans celle-ci, générer le tableau croisé dynamique de la répartition des femmes selon la variable **Obésité**.
2. Considérer les situations suivantes. Quelle(s) critique(s) peut-on poser à leur égard?
 - (a) Cette marque de dentifrice est recommandée par 80% des dentistes qui l'utilisent.
 - (b) Il y a 200 utilisateurs d'un produit *A* qui l'ont acheté à nouveau alors que 50 utilisateurs d'un produit compétitif *B* ont acheté à nouveau le produit *B*. Le produit *A* est donc un meilleur produit que le *B*.

3. Considérer les graphiques suivants. Expliquer en quoi l'information est trompeuse.

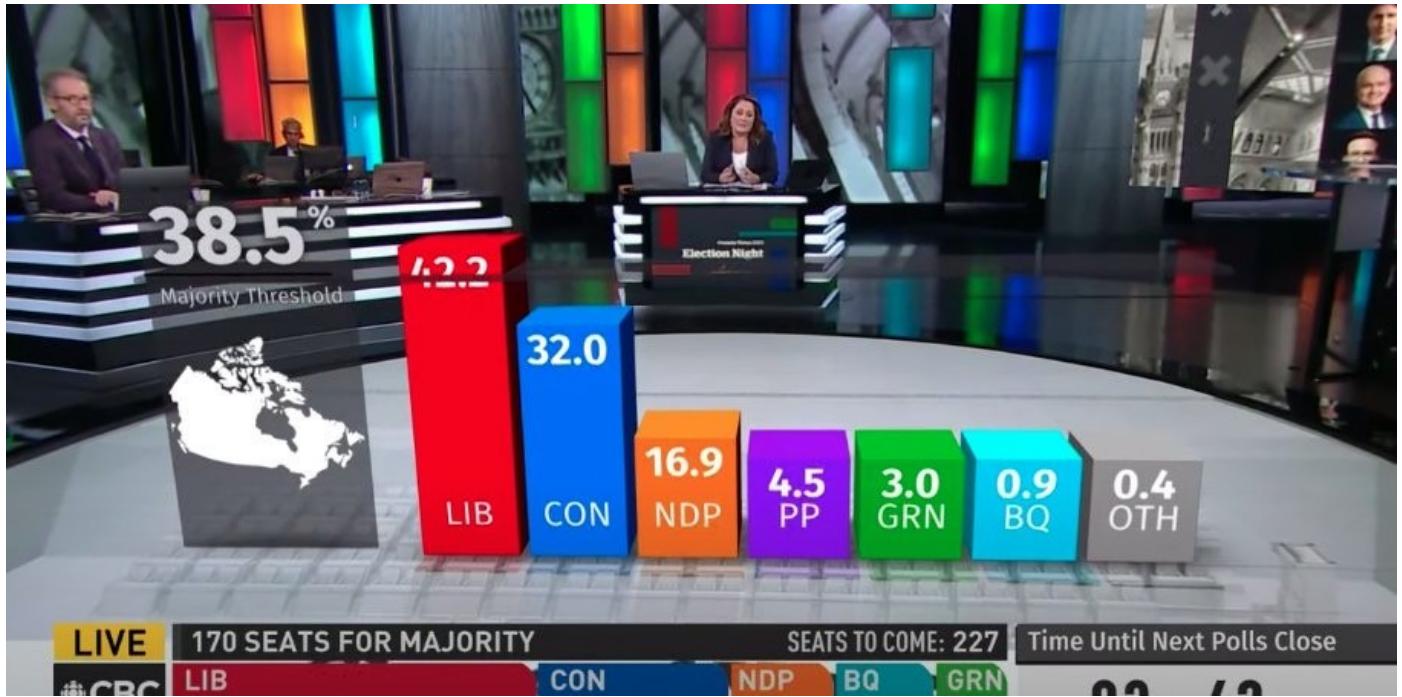


Figure 2.1.1 Diagramme à bandes des résultats de l'élection fédérale de 2021¹

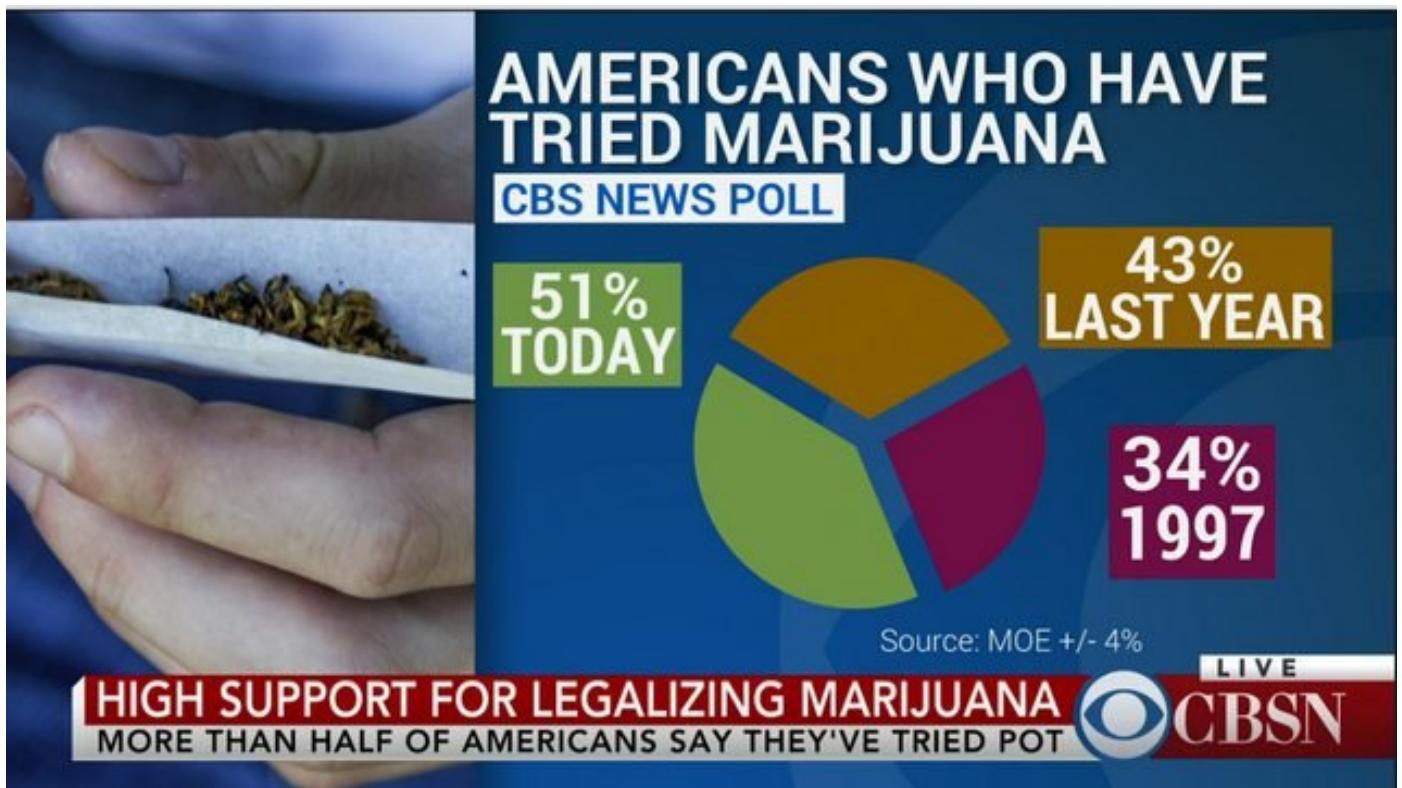


Figure 2.1.2 Diagramme circulaire à propos de la consommation de cannabis des Américains²

¹capture d'écran du reportage du réseau CBC sur le résultat de l'élection fédérale canadienne de 2021

²<https://www.painting-with-numbers.com/blog/getting-high-on-bad-data-visualization>, visité le 27 mars

4. Lequel des graphiques suivants semble le plus adéquat pour illustrer l'information? Dire pourquoi.

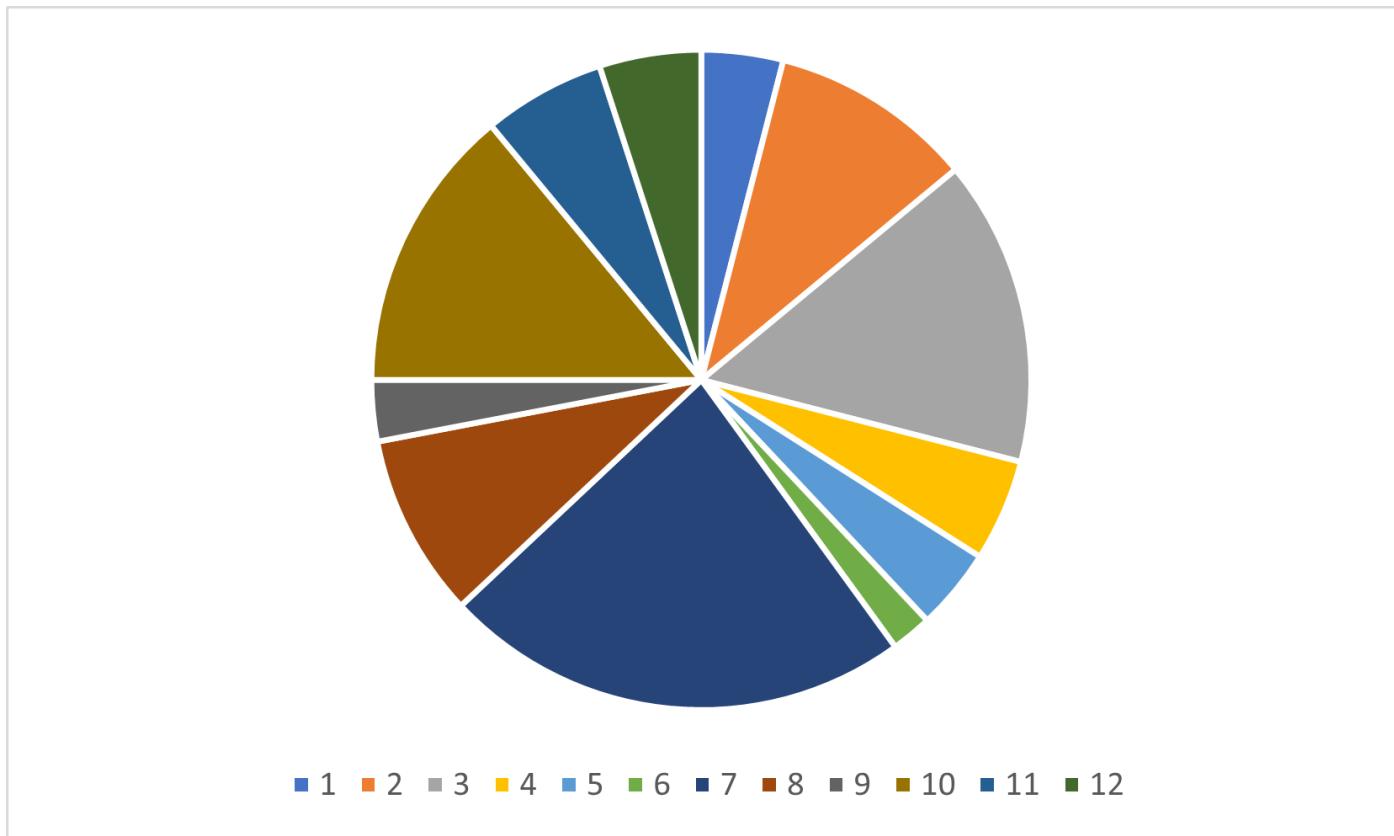


Figure 2.1.3 Répartition en pourcentage des élèves d'une classe de statistiques selon leur mois de naissance

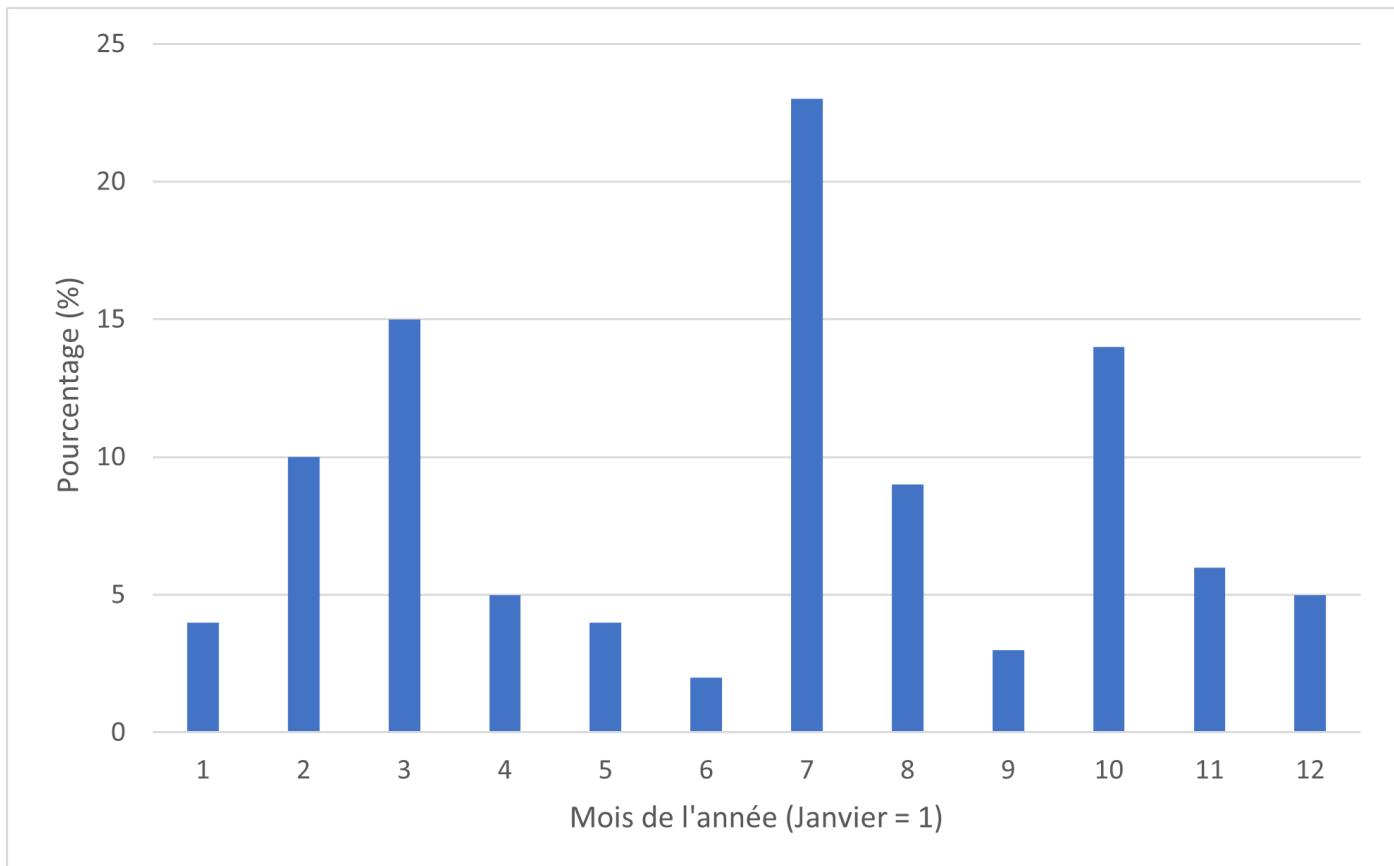


Figure 2.1.4 Répartition en pourcentage des élèves d'une classe de statistiques selon leur mois de naissance

2.2 Laboratoire

Dans ce laboratoire, on cherche à représenter la répartition du nombre de femmes d'origine pima vivant en Arizona qui sont atteintes du diabète, ainsi que leur répartition selon leur niveau d'obésité, tel que qualifié par Santé Canada en fonction de l'indice de masse corporelle.

Pour cela, on utilise les variables **Atteint** et **Obésité**, dont les tableaux croisés dynamiques ont été construits respectivement au [Laboratoire 1](#) et au [prélaboratoire](#) précédent. On rappelle que, pour la variable **Atteint**, le code 0 signifie l'absence de diabète et le code 1 signifie que la femme est atteinte du diabète. Pour la variable **Obésité**, l'échelle de 1 à 6 se traduit par:

1. Poids insuffisant;
2. Poids normal;
3. Excès de poids;
4. Obésité classe I;
5. Obésité classe II;
6. Obésité classe III.

Dans ce laboratoire, on introduit la notion de mise en forme d'un tableau à des fins de publication, la création d'un diagramme circulaire et d'un diagramme à bandes, ainsi que la mise en forme appropriée pour ces deux éléments graphiques.

2.2.1 Le tableau croisé dynamique et le tableau pour publication

Le tableau croisé dynamique construit par Excel à la fin du laboratoire 1 (répartition des femmes selon la présence de diabète) n'est pas adéquat si l'on souhaite le publier comme source d'information. Par exemple, il est nécessaire de préciser le nom des catégories, d'ajouter un titre significatif et la source des données. Voici une procédure pour construire un tableau plus propice au partage des informations.

Dans la feuille **Étude Atteint**, on sélectionne une cellule pour commencer le tableau, par exemple H8 dans la figure [Figure 2.2.1](#). On appuie ensuite sur **=**, on sélectionne la plage du tableau croisé dynamique correspondant aux données et l'on appuie sur **Enter**.

Il est également possible de faire la combinaison **Ctrl**+**C** sur la plage des données du tableau croisé dynamique et de faire un collage spécial à l'endroit souhaité. Pour cela, on peut faire la combinaison **Ctrl**+**V**, cliquer sur l'icône de collage dans le coin inférieur bas et sélectionner **Coller le lien (N)**.

Dans tous les cas, il est possible qu'il faille changer le format de la cellule pour **Pourcentage** (voir [Format de cellule en pourcentage](#)). On s'assure de garder deux chiffres significatifs après la virgule. L'animation ci-dessous permet de voir en trois étapes à quoi ressemble la progression de ces étapes.

A	B	C	D	E	F	G	H	I	J	K	L
1											
2											
3											
4											
5											
6											
7	Étiquettes de lignes	Nombre de Atteint	Nombre de Atteint2								
8	0	500	65,10%								
9	1	268	34,90%								
10	Total général	768	100,00%								
11											
12											
13											
14											

Figure 2.2.1 Création du tableau pour présentation - première méthode

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2														
3														
4														
5														
6														
7	Étiquettes de lignes	Nombre de Atteint	Nombre de Atteint2											
8	0	500	65,10%											
9	1	268	34,90%											
10	Total général	768	100,00%											
11														
12														
13														
14														
15														
16														
17														
18														
19														
20														
21														
22														
23														
24														

Figure 2.2.2 Création du tableau pour présentation - deuxième méthode

2.2.2 Mise en forme tableau

Une fois les données extraites du tableau croisé dynamique, on ajoute les étiquettes de colonnes et de lignes, le titre du tableau et la source des données. Pour le moment, on ne se soucie pas de la mise en forme de ces éléments. Dans un premier temps, pour le tableau concernant la variable **Atteint** du diabète, les étapes qui suivent devraient mener à un tableau de quatre lignes et trois colonnes.

Liste 2.2.3 Mise en forme du tableau pour publication

1. À gauche, on ajoute une colonne appelée «Présence du diabète» dont les lignes sont, du haut vers le bas : «Non»,«Oui»,«Total». La colonne «Présence du diabète» se trouve donc dans la colonne G de la feuille de calcul.
2. On ajoute le titre des autres colonnes, de gauche à droite : «Nombre de femmes» et «Pourcentage de femmes». On élargit les colonnes de manière à ce que le tout soit lisible.
3. Dans la dernière ligne, on a le total du nombre de femmes sous chaque colonne. Sous la colonne **Pourcentage de femmes** à la ligne **Total**, on s'assure qu'il est *toujours* écrit 100%, même si l'addition des nombres dans la colonne pourrait ne pas donner 100% en raison d'erreur d'arrondi. Dans de tels cas, on l'indiquera sous le tableau à l'aide de la note : «En raison de l'arrondissement des pourcentages, le total pourrait ne pas être exactement de 100,00%.»
4. Sous le tableau, on inscrit la source des données. Dans le cas de cette étude, la source est donnée dans la [Section 1.1](#). La mention «Source : National Institute of Diabetes and Digestive and Kidney Diseases (USA)» doit apparaître sous chacun des tableaux et des graphiques créés pour publication.

5. On titre le tableau. Le titre aura généralement la forme **Répartition d'un échantillon (ou d'une population) de [unités statistiques] selon [la variable], [le lieu], [période ou date]**.

La figure ci-dessous illustre le tableau pour publication une fois toutes les étapes effectuées.

Répartition d'un échantillon de femmes d'origine pima selon la présence de diabète, Arizona, année inconnue		
Présence de diabète	Nombre de femmes	Pourcentage de femmes
Non	500	65,10%
Oui	268	34,90%
Total	768	100%

Source : National Institute of Diabetes and Digestive and Kidney Diseases (USA)

Figure 2.2.4 Le tableau pour publication - après la mise en forme

Répéter les étapes ci-dessus avec le tableau croisé dynamique généré lors du prélaboratoire pour la variable **Obésité** dans la feuille **Étude Obésité**.

Ces tableaux sont appelés **tableau de fréquences** de la variable étudiée. S'il n'y a que la colonne du nombre d'individus, on parle alors de **tableau de fréquences absolues** et, s'il n'y a que la colonne des pourcentages, on dit **tableau de fréquences relatives**.

Habituellement, le titre, les entêtes de colonne, ainsi que le contenu de la ligne «Total» ont une mise en forme particulière, par exemple en gras.

2.2.3 Graphiques

Un tableau est une excellente manière de rassembler l'information d'une variable qualitative. Cela dit, illustrer la répartition des unités de manière graphique peut être aussi, et même davantage, utile. Pour une variable qualitative, il existe deux types de graphiques principaux: le diagramme circulaire et le diagramme à bandes. Dans le second cas, les bandes peuvent être horizontales ou verticales, mais, si la variable est ordinale, on préfère le diagramme à bandes verticales.

On construit le diagramme circulaire associé à la variable **Atteint**. On choisit de faire la répartition en pourcentage des effectifs.

1. Dans le tableau de fréquences de la feuille **Étude Atteint**, sélectionner les modalités de la variable aux cellules G8:G9 et, en maintenant la touche **Ctrl** enfoncée, les effectifs relatifs aux cellules I8:I9.
2. Sous l'onglet **Insertion**, cliquer sur le bouton correspondant au diagramme circulaire dans le ruban. Voir la figure ci-dessous.
3. Sous **Secteur 2D**, cliquer sur le premier type de graphique à gauche, appelé **Secteur**. Déplacer le graphique au besoin.
4. La figure [Figure 2.2.5](#) illustre les étapes précédentes sous forme d'animation.
5. On peut sélectionner un style prédéfini sous l'onglet création de graphique (apparaissant lorsque le graphique est sélectionné, voir la figure [Figure 2.2.6](#)) ou encore peaufiner les éléments graphiques selon ce qui est

attendu. Toutefois, en sciences, l'allure esthétique du graphique ne devrait pas prendre le dessus sur l'information transmise. On préférera un style relativement neutre sans trop de fioritures. Le graphique circulaire doit contenir les éléments suivants:

- Un titre représentatif, typiquement de la forme **Répartition d'un échantillon (ou d'une population) de [unités statistiques] selon [la variable], [le lieu], [période ou date]**;
- Une légende, pour distinguer les différents secteurs;
- Les étiquettes correspondant aux pourcentages ou au nombre d'effectifs sur le graphique, pour une information précise;
- La source, lorsque pertinente, dans le bas du graphique.

Si l'un ou plusieurs de ces éléments est manquant, on peut, lorsque le graphique est sélectionné, cliquer sur l'onglet **Création de graphique** et cliquer sur le bouton **Ajouter un élément graphique** situé dans la partie gauche du ruban. Il est aussi possible de cliquer sur le petit symbole de croix en haut à droite du graphique. La figure Figure 2.2.7 illustre ces deux options.

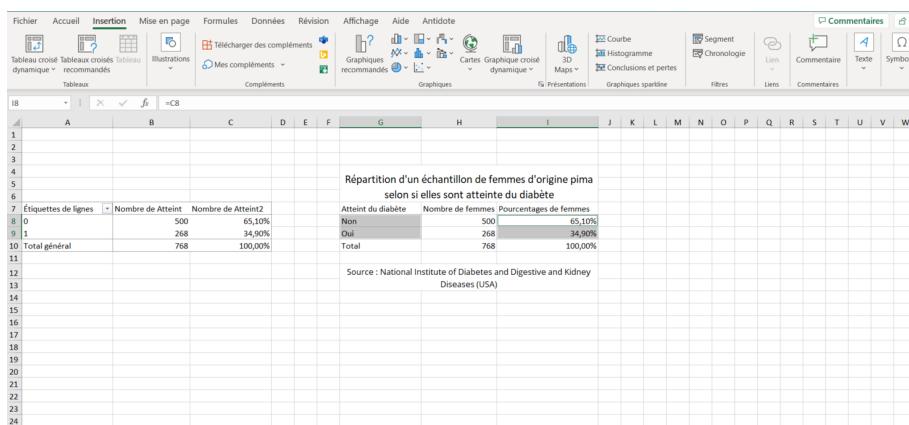


Figure 2.2.5 L'insertion d'un diagramme circulaire



Figure 2.2.6 Les styles prédéfinis d'Excel

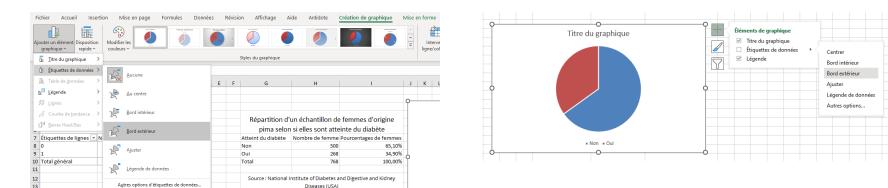


Figure 2.2.7 Ajout d'un élément graphique: deux options

On construit maintenant le diagramme à bandes pour la variable **Obésité**. Comme la variable est qualitative et a une échelle ordinale, on choisit les bandes verticales. On choisit d'utiliser les effectifs relatifs.

1. Dans le tableau de fréquences de la feuille **Étude Obésité**, sélectionner les modalités de la variable et, en maintenant la touche **Ctrl** enfoncee, les effectifs relatifs.

2. Sous l'onglet **Insertion**, cliquer sur le bouton correspondant à l'histogramme. Voir la figure [Figure 2.2.8](#) ci-dessous.
 - Sous **Histogramme 2D**, cliquer sur le premier type de graphique à gauche, appelé **Histogramme groupé**, pour un diagramme à bandes verticales.
 - Sous **Barres 2D**, cliquer sur le premier type de graphique à gauche, appelé **Barres groupées**, pour un diagramme à bandes horizontales.
3. On peut sélectionner un style prédéfini ou encore peaufiner les éléments graphiques selon ce qui est attendu. Toutefois, en sciences, l'allure esthétique du graphique ne devrait pas prendre le dessus sur l'information transmise. On préférera un style relativement neutre sans trop de fioritures. Le diagramme à bandes doit contenir les éléments suivants:
 - Un titre représentatif, typiquement de la forme *Répartition d'un échantillon (ou d'une population) de [unités statistiques] selon [la variable], [le lieu], [période ou date]*;
 - Des titres significatifs pour les axes;
 - Le nom des modalités identifié clairement, sous l'axe (pour un graphique à bandes verticales), à la gauche (pour un graphique à bandes horizontales) ou encore dans une légende (pour les deux types de graphiques).
 - L'axe des effectifs comprend une graduation bien faite. Une note est ajoutée s'il y a eu coupure de l'axe pour sauter certaines valeurs.
 - Les étiquettes correspondant aux pourcentages ou au nombre d'effectifs sur le graphique, pour une information précise.
 - La source, lorsque pertinente, dans le bas du graphique.

Si l'un ou plusieurs de ces éléments sont manquants, on peut, lorsque le graphique est sélectionné, cliquer sur l'onglet **Création de graphique** et cliquer sur le bouton **Ajouter un élément graphique**. Il est aussi possible de cliquer sur le petit symbole de croix en haut à droite du graphique. La figure [Figure 2.2.7](#) illustre ces deux options.

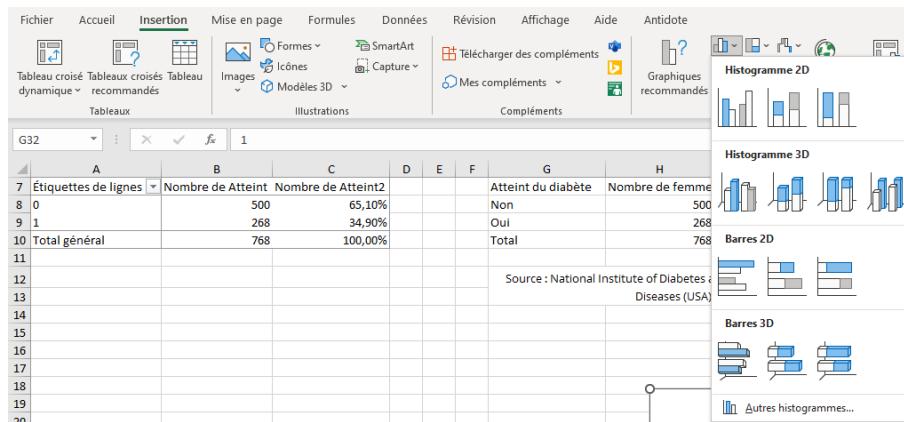
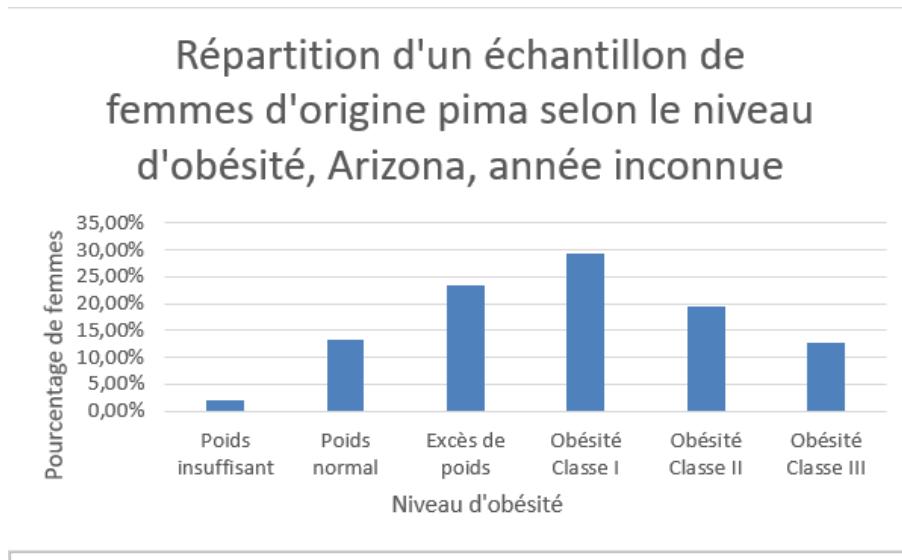


Figure 2.2.8 Insertion d'un diagramme à bandes



Source : National Institute of Diabetes and Digestive and Kidney Diseases (USA)

Figure 2.2.9 Diagramme à bandes verticales

2.3 Réflexions

L'information transmise par les tableaux de fréquences et les graphiques n'est que la première partie d'une série de renseignements qui permettront de bien caractériser la population étudiée et éventuellement de faire des inférences et de valider ou infirmer les hypothèses de recherche. Les questions qui suivent servent à regarder les problèmes qu'il y a à tirer des conclusions à partir d'une analyse sommaire des données comme la lecture de tableaux et de graphiques. Les prochains laboratoires permettront de tirer une multitude de détails des données disponibles et permettront de combler certaines lacunes soulevées ci-dessous.

Travail à faire après le laboratoire

Objectifs

- Examiner les variables **Atteint** et **Obésité**.
 - Formuler des conclusions
 - Poser un regard critique sur les données.
 - Formuler des hypothèses de recherche.
1. En deux courtes phrases, résumer la situation du diabète et du niveau d'obésité chez la population de femmes pimas.
 2. Selon la National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)¹, il y a 8,8% des femmes américaines qui sont atteintes de diabète (données de 2017-2020). Selon l'American Diabetes Association², il y a 6,9% des femmes mexicaines d'origine pima qui le sont aussi (données de 2006). En supposant que toutes ces valeurs et celles de la population étudiée sont restées les mêmes au fil du temps, quels constats peut-on faire au sujet des femmes pimas vivant en Arizona à la lumière de ces informations?
 3. Dans le feuille Excel **Étude Obésité**, construire le diagramme circulaire pour aller avec la variable **Obésité**. Que dire du résultat? Comparer avec le diagramme à bandes créés pour la même variable à la [Figure 2.2.9](#).
 4. Est-il correct de dire que les femmes pimas vivant en Arizona ont un niveau d'obésité moyen d'environ 4? Expliquer.
 5. Après avoir étudié les variables qualitatives de ce laboratoire, formuler deux hypothèses en lien avec ces variables et les autres présentes dans la base de données.

¹<https://www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics#factsstats>

²<https://ictnews.org/archive/mexico-vs-arizona-pima-indians>

Chapitre 3

Variables quantitatives

Ce chapitre introduit le traitement et la présentation des variables quantitatives. Il aborde des techniques telles que la création de tableaux croisés dynamiques pour synthétiser les données quantitatives, la mise en forme de ces tableaux pour une présentation optimale, et l'élaboration de graphiques adaptés à ces types de variables. Ces méthodes facilitent l'interprétation et la communication des informations issues des données quantitatives.

3.1 Prélab

Les variables quantitatives sont un type de variables pour lesquelles les modalités sont des nombres. On retrouvera souvent parmi ces variables le temps, la vitesse, la distance, la quantité de bactéries, le volume, la masse, etc. Elles se déclinent en deux catégories, soit les variables quantitatives discrètes et continues. Pour synthétiser visuellement l'information d'une variable quantitative, on utilise principalement un tableau de fréquences ainsi qu'un diagramme à bâtons, un histogramme ou un polygone de fréquences.

Travail à faire avant le cours

Objectifs

- Réinitialiser les filtres appliqués sur les données brutes.
- Créer le tableau croisé dynamique d'une variable quantitative discrète.
- Construire un diagramme à bâtons.

On continue de travailler à l'aide de la base de données du laboratoire d'[introduction 1](#). Le but de ce prélab est d'utiliser les connaissances acquises et les outils employés dans les deux premiers laboratoires afin de poursuivre l'étude des mesures diagnostiques des femmes d'origine pima.

1. Dans la feuille de calcul **Données**, enlever tous les filtres faits dans les travaux précédents en cliquant sur le bouton **Effacer** de la zone **Trier et Filtrer** (voir la [Figure 3.1.1](#)).

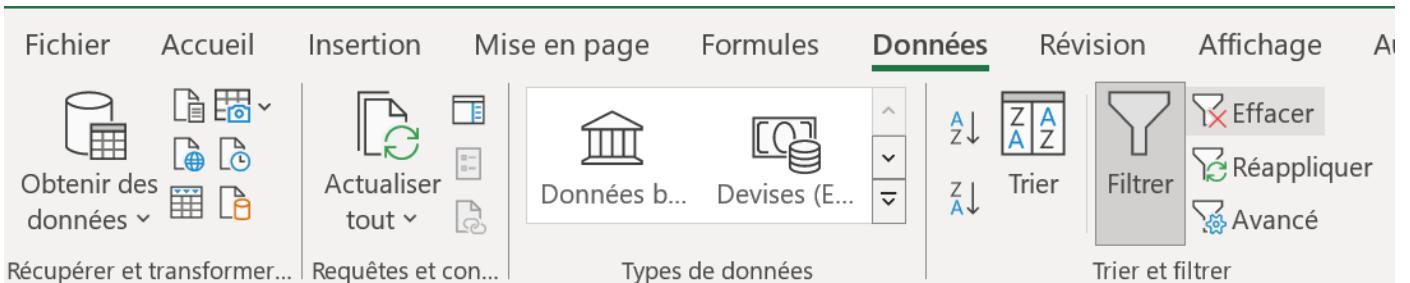


Figure 3.1.1 Bouton Effacer de la zone Trier et filtrer

2. Dans le fichier Excel des laboratoires 1 et 2, créer le tableau croisé dynamique des femmes d'origine pima selon le **Nombre de grossesses** en suivant les étapes ci-dessous.

- Ajouter une nouvelle feuille de calcul intitulée **Étude Nombre de grossesses**. Déplacer cette feuille en dernière position si Excel ne le fait pas automatiquement.
- Dans cette feuille, créer le tableau croisé dynamique de la répartition de l'échantillon de femmes d'origine pima selon le nombre de grossesses.
- Faire la mise en forme du tableau de fréquences en respectant toutes les normes de présentation abordées lors du laboratoire précédent.

3. Puisque la variable ***Nombre de grossesses*** est quantitative discrète, créer un diagramme à bâtons des femmes d'origine pima selon le ***Nombre de grossesses*** en suivant les étapes ci-dessous.

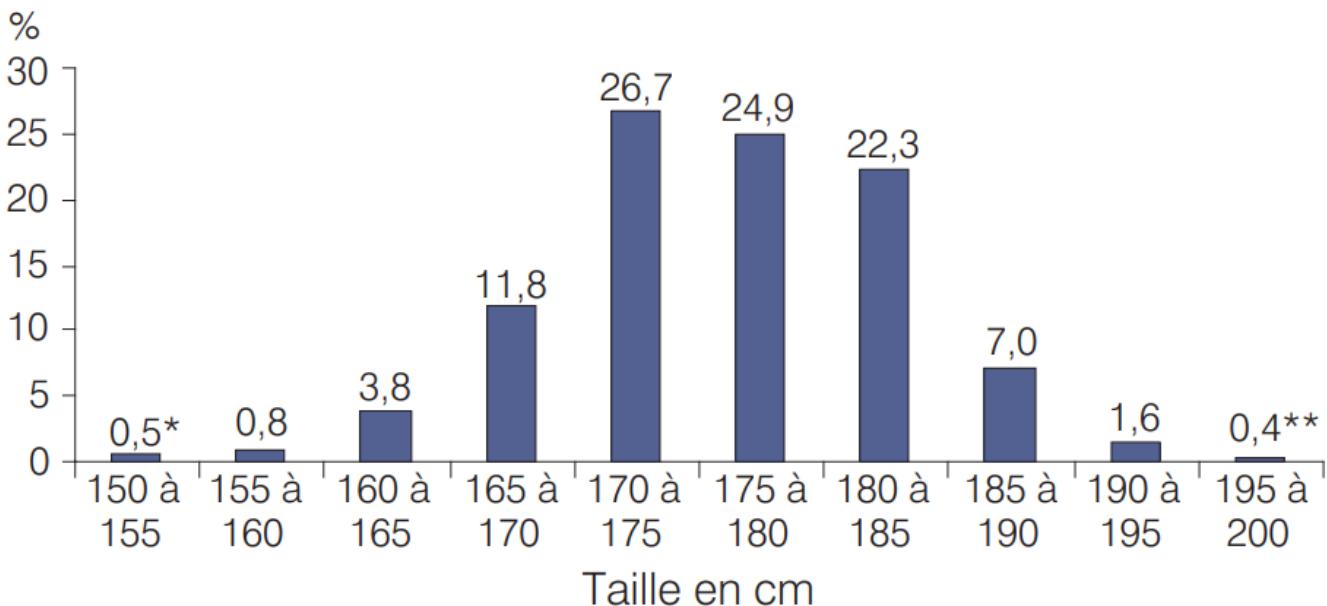
- Dans la feuille de calcul ***Étude Nombre de grossesses***, créer le diagramme à bâtons en sélectionnant l'icône ***Insérer un histogramme ou un graphique à barres*** à partir de l'onglet ***Insertion***. Sélectionner la première option de graphique proposée par Excel, appelée ***Histogramme groupé***.
- Faire la mise en forme du diagramme en respectant toutes les normes de présentation abordées lors des deux laboratoires précédents. Cliquer avec le bouton de droite sur un des bâtons du diagramme et sélectionner le dernier onglet ***Mettre en forme une série de données***. Une fenêtre s'ouvre sur le côté droit de la feuille de calcul. Dans la zone de saisie des ***Options des séries***, taper ***500 %*** comme ***Largeur des intervalles*** pour avoir la largeur maximale entre les bâtons et pour les amincir le plus possible (voir la [Figure 3.2.34](#) et la [Figure 3.2.35](#)).

4. Détailler les faits saillants du diagramme à bâtons fait à l'exercice [Activité 3.1.3](#).

5. À la lumière du graphique produit à l'exercice [Activité 3.1.3](#), quelles sont les limites d'un diagramme à bâtons?

6. Considérer le graphique suivant. Décéler tous les problèmes de ce graphique.

Répartition des hommes de 20 ans et plus en fonction de leur taille, Québec, 2005



Note : Les valeurs situées sous 150 cm et au-dessus de 200 cm sont exclues.

* Coefficient de variation entre 15 % et 25 %; interpréter avec prudence.

** Coefficient de variation supérieur à 25 %; estimation imprécise fournie à titre indicatif seulement

Source : Statistique Canada, *Enquête sur la santé dans les collectivités canadiennes*, cycle 3.1, 2005, fichier de partage.

Compilation : Institut de la statistique du Québec.

Figure 3.1.2 Histogramme de la répartition des hommes québécois de 20 ans et plus en fonction de leur taille en 2005

3.2 Laboratoire

Dans ce laboratoire, l'objectif est de représenter la répartition de femmes pima vivant en Arizona selon leur indice de masse corporelle, ainsi que la répartition de femmes pima vivant en Arizona par présence de diabète, en fonction de l'indice de masse corporelle. Il s'agit donc de présenter les étapes de l'étude d'une variable quantitative continue, ainsi que celles de l'étude simultanée d'une variable quantitative continue avec une variable qualitative.

On introduit la notion de filtrage de données aberrantes, la création de classes pour des variables quantitatives continues, la création d'un histogramme et d'un polygone de fréquences, la mise en forme appropriée pour ces deux graphiques, ainsi que le calcul de mesures descriptives pour des variables quantitatives.

Pour cela, on utilise les variables ***IMC*** et ***Atteint***.

3.2.1 L'étude d'une variable quantitative continue

Les étapes de l'analyse d'une variable quantitative continue sont les suivantes : filtrage de données aberrantes, création de classes pour les valeurs de la variable, groupement des données lors de la création du tableau croisé dynamique, mise en forme de ce tableau croisé dynamique, création d'un graphique approprié, tel qu'un histogramme ou un polygone de fréquences, calcul de mesures descriptives, et enfin, interprétation des résultats. Les étapes qui suivent mènent à l'analyse de l'indice de masse corporelle.

3.2.1.1 Filtrer les valeurs aberrantes du tableau principal

Avant de générer un tableau croisé dynamique impliquant une variable quantitative, il est important d'effectuer une enquête préliminaire des données dans le but de filtrer, si nécessaire, des données aberrantes.

En appliquant un filtre à la variable ***IMC***, on constate que certaines femmes ont enregistré un indice de masse corporelle de 0. Cette valeur étant impossible, il s'agit donc d'une valeur aberrante. On choisit de filtrer ces valeurs et de les exclure lors de la création du tableau de fréquences, de l'histogramme, ainsi que des calculs des mesures statistiques, car ces dernières pourraient fausser les interprétations. Les étapes suivantes mènent au filtrage du tableau principal ***Échantillon***.

1. Ouvrir le classeur ***Données_Diabète.xlsx*** sauvegardé avec le travail fait aux laboratoires 1 et 2.
2. Dans le feuille ***Données***, cliquer sur l'icône du filtre (petit triangle) à droite du titre de la colonne ***IMC*** (voir la Figure 3.2.1).

Figure 3.2.1 Icône de filtre à droite du titre de la colonne ***IMC***

Un menu déroulant s'affiche (voir la Figure 3.2.2).

3. Découcher le crochet à gauche de la valeur 0 (voir la Figure 3.2.2). Cliquer sur ***OK***.

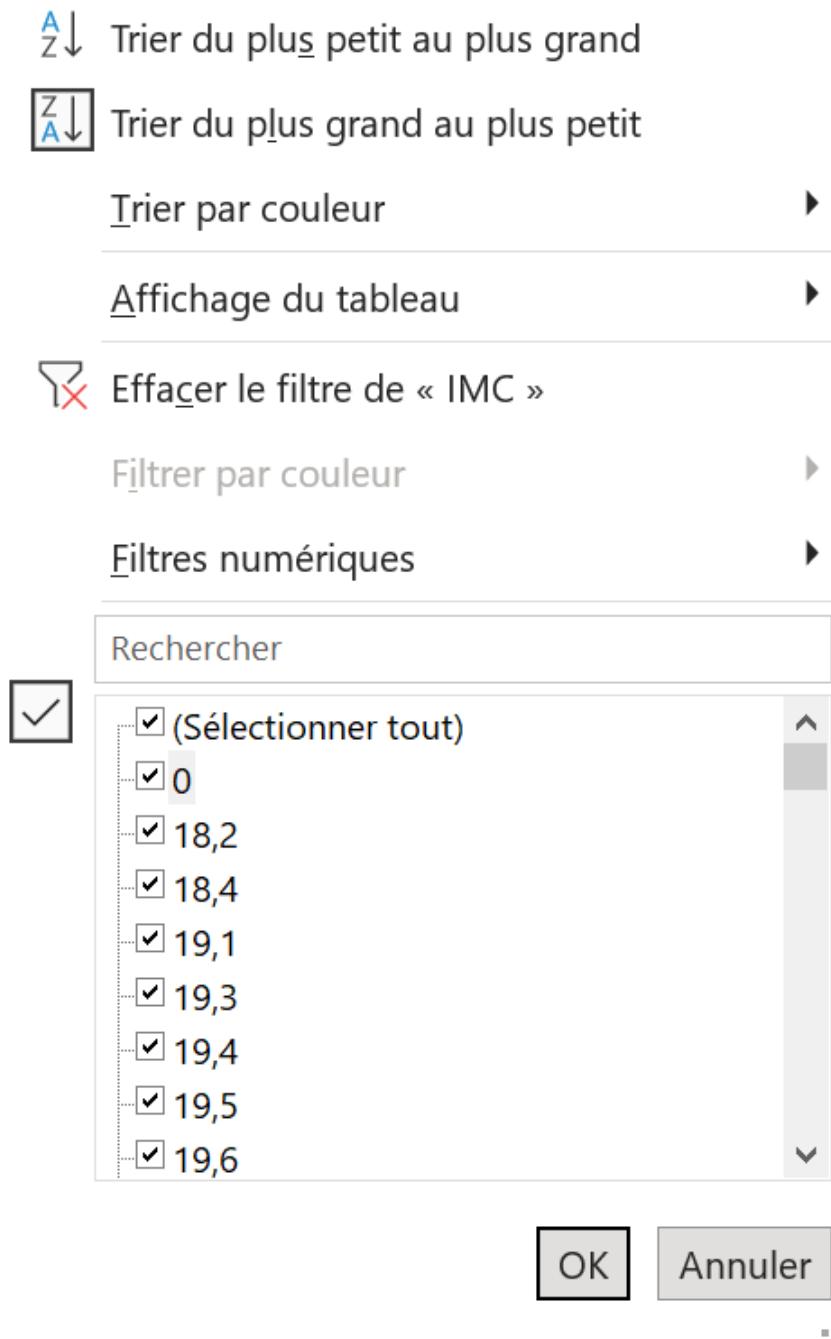


Figure 3.2.2 Filtrage des valeurs nulles de l'IMC

Remarque 3.2.3 Filtrer des données. Malheureusement, ce filtrage ne s'applique que sur les données du tableau principal **Échantillon** et non sur les tableaux croisés dynamiques. Toutefois, il est pratique de filtrer les données brutes lors de la réalisation d'une partie d'une étude afin de pouvoir comparer les résultats obtenus et leur conformité avec les données.

3.2.1.2 Tableau croisé dynamique d'une variable quantitative continue

Pour réaliser l'étude d'une variable quantitative continue, il est nécessaire de construire un tableau de fréquences en regroupant les valeurs en classes, car il y a trop de valeurs différentes pour permettre une synthèse efficace. Le tableau croisé dynamique généré par Excel n'est pas adéquat, puisqu'il ne regroupe pas les valeurs des variables par défaut. On commence par rappeler les étapes de création d'un tableau croisé dynamique, soit celui de la répartition d'un échantillon de femmes d'origine pima vivant en Arizona selon l'indice de masse corporelle.

1. Dans le classeur Excel, ajouter une nouvelle feuille de calcul intitulé **Étude IMC**. Déplacer cette feuille en dernière position si Excel ne le fait pas automatiquement.
2. Sélectionner la cellule **B3** dans cette feuille de calcul.
3. Insérer un tableau croisé dynamique vide tel que vu à la [Sous sous-section 1.2.8.1](#).
4. Glisser et déposer la variable **IMC** dans la zone de saisie **Lignes**, ainsi que deux fois dans la zone de saisie **Valeurs** (voir la [Figure 3.2.4](#)).

Choisissez les champs à inclure dans le rapport : ⚙️

Rechercher 🔍

Champs de tableau croisé

Insuline

IMC

Obésité

Fonction pedigree du diabète

Faites glisser les champs dans les zones voulues ci-dessous:

▼ Filtres

☰ Lignes

IMC

☰ Colonnes

Σ Valeurs

Σ Valeurs

Nombre de IMC

Nombre de IMC 2

Figure 3.2.4 Glissement de la variable *IMC* dans les zones de saisie *Lignes* et *Valeurs*

Dans la zone de saisie **Valeurs**, cliquer sur la flèche du menu déroulant du premier onglet (*Nombre de IMC*), puis sélectionner l'option **Paramètres des champs de valeurs** pour modifier le calcul. On veut le nombre de femmes pour la deuxième colonne (et non la somme comme Excel fait par défaut). Pour *Nombre de IMC 2*, on veut le pourcentage de femmes (voir la [Sous sous-section 1.2.8.2](#) pour référence).

5. Au final, le tableau croisé dynamique généré doit ressembler à la [Figure 3.2.5](#).

Étiquettes de lignes	Nombre de IMC	Nombre de IMC 2
0	11	1,43%
18,2	3	0,39%
18,4	1	0,13%
19,1	1	0,13%
19,3	1	0,13%
19,4	1	0,13%
19,5	2	0,26%
19,6	3	0,39%
19,9	1	0,13%
20	1	0,13%
20,1	1	0,13%
20,4	2	0,26%
20,8	2	0,26%
21	2	0,26%
21,1	4	0,52%
21,2	1	0,13%
21,7	1	0,13%
21,8	5	0,65%
21,9	3	0,39%
22,1	2	0,26%
22,2	2	0,26%
22,3	1	0,13%

Figure 3.2.5 Les trois colonnes du tableau croisé dynamique final de la répartition de l'échantillon de femmes d'origine pima selon l'indice de masse corporelle

La première ligne du tableau croisé dynamique révèle que onze femmes ont enregistré un indice de masse corporelle de 0. Il est possible de constater que le tableau croisé dynamique n'a pas filtré les données contenant une valeur nulle comme le tableau principal *Échantillon* l'a fait. Il faudra refaire le filtrage pour le tableau croisé dynamique.

3.2.1.3 Filtrer les valeurs aberrantes d'un tableau croisé dynamique

Même si l'on a filtré les valeurs aberrantes nulles du tableau principal, ces dernières apparaissent tout de même dans le tableau croisé dynamique. Il existe plusieurs façons de les exclure. On choisit de le faire avant de regrouper les valeurs en classes.

1. Dans une cellule de la première colonne du tableau croisé dynamique généré à la [Sous sous-section 3.2.1.2](#), cliquer avec le bouton de droite de la souris. Un menu contextuel s'affiche (voir la [Figure 3.2.6](#)).

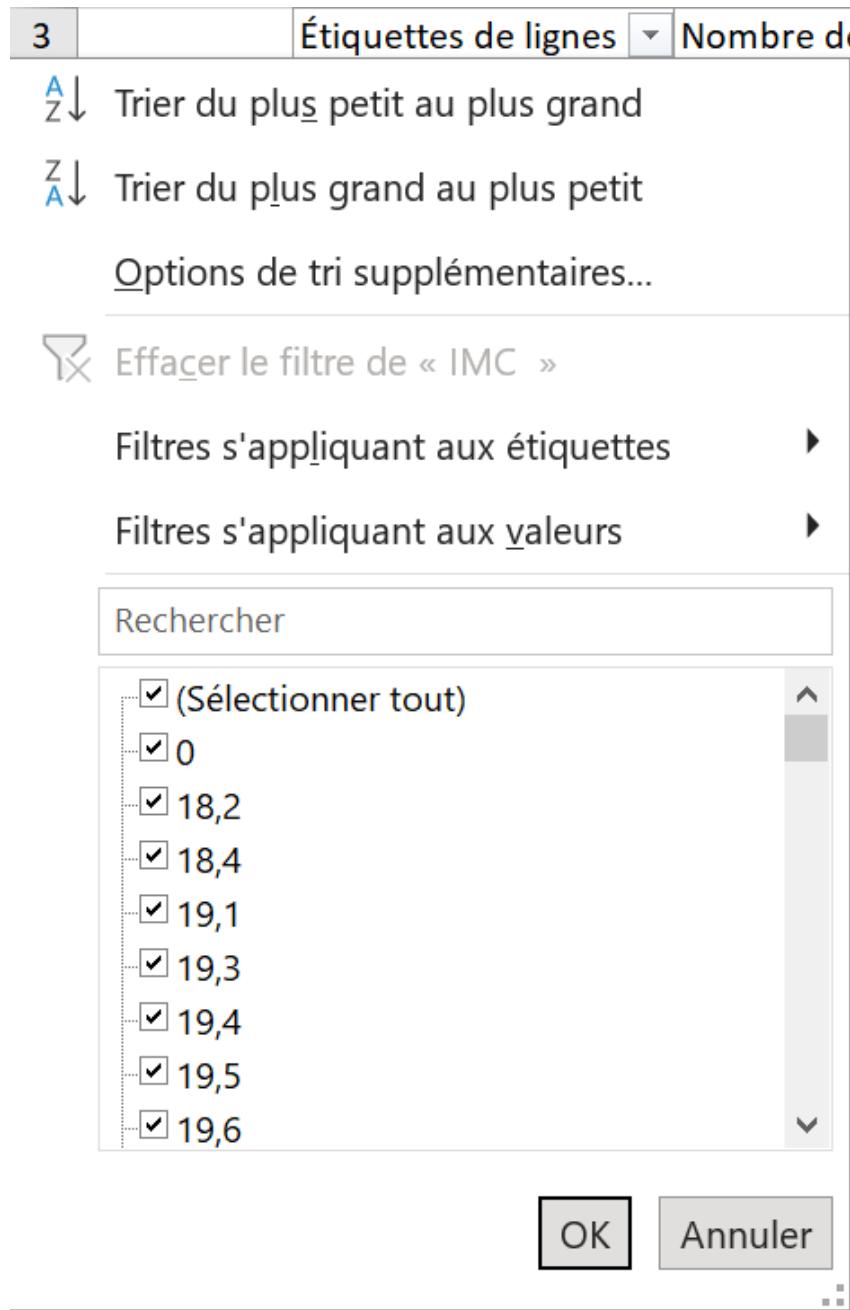


Figure 3.2.6 Affichage du menu contextuel de la première colonne du tableau croisé dynamique

2. Sélectionner l'option ***Filtres s'appliquant aux étiquettes*** suivie de l'option ***Est différent de*** (voir la [Figure 3.2.7](#)).

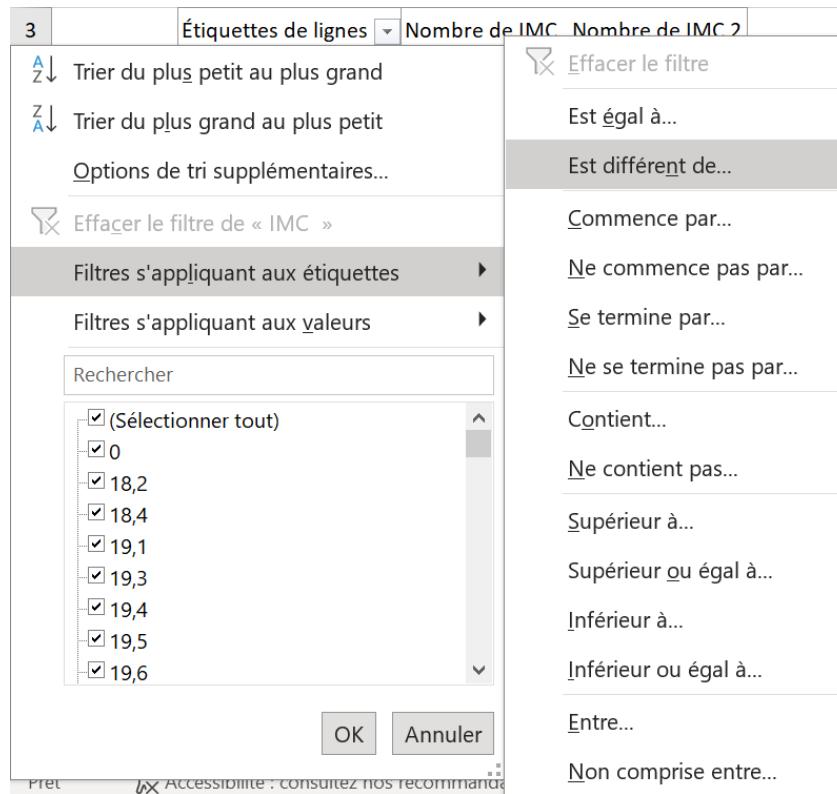


Figure 3.2.7 Filtrage s'appliquant aux étiquettes

3. Une boite de dialogue s'affiche à l'écran. Il faut choisir d'afficher les éléments pour lesquels l'étiquette est différente de 0. Ainsi, dans la zone de saisie à droite de l'option *est différent de*, il faut taper la valeur *0* (voir la [Figure 3.2.8](#)) et cliquer sur **OK**.

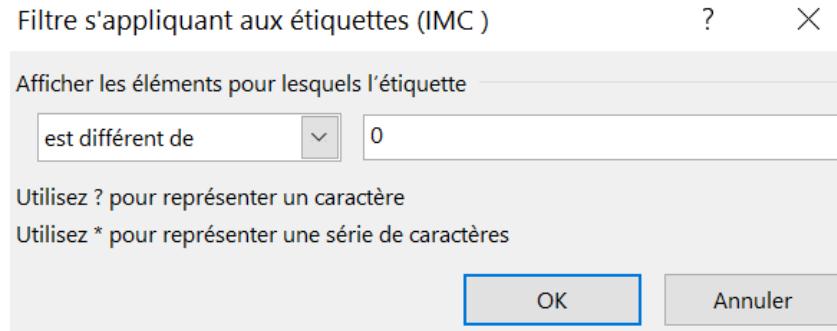


Figure 3.2.8 Afficher les éléments pour lesquels l'étiquette est différente de 0

Le tableau croisé dynamique résultat exclut désormais les valeurs pour lesquelles l'IMC vaut 0.

3.2.1.4 Créer des classes

Comme présenté, le tableau croisé dynamique de la [Figure 3.2.5](#) n'est pas optimal pour l'analyse de la répartition des femmes selon l'indice de masse corporelle. Un regroupement des données en classes est nécessaire. Avant

de procéder, il est important de déterminer le nombre optimal de classes et l'amplitude de ces dernières. Les choix par défaut d'Excel ne sont pas toujours adéquats. Il est donc recommandé de documenter clairement les décisions prises dans la feuille de calcul afin d'en conserver une trace. Les étapes suivantes mènent au groupement des données de la variable ***IMC***.

Choix d'amplitude.

1. Dans la cellule **G3** de la feuille de calcul **Étude IMC**, taper le titre *Calcul de l'amplitude* (voir la [Figure 3.2.9](#)).

Taper *Nombre de classes* dans la cellule **G5**, *Étendue de l'IMC* dans la cellule **G6**, *Valeur minimale de l'IMC* en **G7**, *Amplitude théorique* en **G8**, et finalement, *Amplitude choisie* dans la cellule **G9** (voir la [Figure 3.2.9](#)).

Calcul de l'amplitude

Nombre de classes

Étendue de l'IMC

Valeur minimale de l'IMC

Amplitude théorique

Amplitude choisie

Figure 3.2.9 Section pour documenter le calcul de l'amplitude

On souhaite calculer le nombre théorique de classes à l'aide de la formule de Sturges, soit $1 + \frac{10}{3} \log n$, où n est le nombre de données.

Dans Excel, on peut insérer des fonctions en saisissant directement le symbole **=** dans une cellule, puis en tapant manuellement la fonction. **ATTENTION :** il ne faut jamais oublier le symboler **=** avant d'insérer une fonction.

Il est également possible d'insérer une fonction à partir de l'onglet **Formules** (voir la [Figure 3.2.10](#)), puis en cliquant sur l'icône **Insérer une fonction**.



Figure 3.2.10 Insérer une fonction à partir de l'onglet **Formules**

En cliquant sur l'icône **Insérer une fonction**, il est possible d'explorer les différentes fonctions disponibles comme les fonctions statistiques (voir la [Figure 3.2.11](#)).

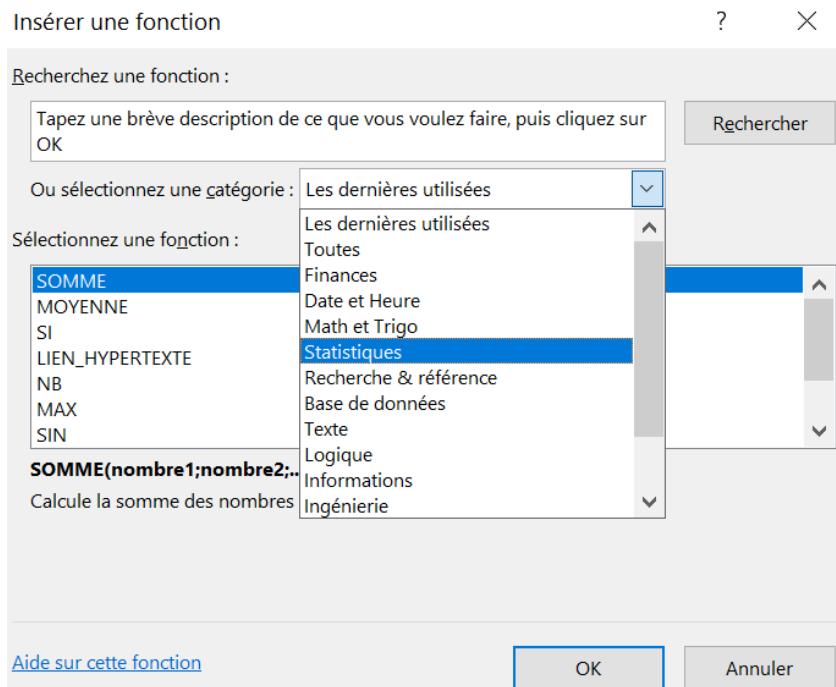


Figure 3.2.11 Exploration des fonctions statistiques d’Excel

2. Dans Excel, plusieurs fonctions de base existent pour faire des calculs. Celles-ci incluent la fonction **MIN** (renvoie la valeur minimale), la fonction **MAX** (renvoie la valeur maximale), la fonction **NB** (renvoie le nombre de données), etc. Malheureusement, lorsque l’on applique un filtre à une variable, comme le filtre appliqué à la variable **IMC**, on ne peut pas utiliser les fonctions de base d’Excel. Excel pallie ce problème avec la fonction **SOUS.TOTAL**.

La fonction **SOUS.TOTAL** d’Excel permet de calculer certaines mesures statistiques (pas toutes) sur un sous-ensemble de données lorsqu’un filtre a été appliqué sur une variable (comme fait à la [Sous sous-section 1.2.5.1](#)). Cela signifie qu’Excel peut faire certains calculs sur les données filtrées.

Dans la cellule **H5**, taper la formule $=1+10/3*\text{LOG}(\text{SOUS.TOTAL}(2;\text{IMC}))$ (voir la [Figure 3.2.12](#)).

La formule Excel **LOG** renvoie le logarithme de l’argument qui se trouve dans les parenthèses. Dans des formules Excel, on peut faire référence à des adresses de cellules ou au nom d’une plage de données. Il ne faut pas oublier le symbole ***** pour le produit.

Dans la fonction **SOUS.TOTAL**, le chiffre 2 fait référence au calcul du nombre de données. Les nombres 1 à 11 spécifient la fonction à utiliser pour calculer le sous-total. Le deuxième paramètre, soit **IMC**, fait référence au nom de la plage de données dont on souhaite calculer le sous-total. La fonction Excel **NB** renvoie le nombre de données de la plage de cellules sélectionnées. Si l’on mettait tout simplement $=\text{NB}(\text{IMC})$, ceci renverrait la valeur 768, soit le nombre total de cellules non vides de la variable **IMC**. Cependant, on veut exclure les femmes qui ont eu une valeur erronée de 0 comme indice de masse corporelle. La fonction **SOUS.TOTAL** permet d’exclure ces valeurs.

Calcul de l'amplitude

Nombre de classes	<code>=1+10/3*LOG(SOUS.TOTAL(2;IMC))</code>
Étendue de l'IMC	
Valeur minimale de l'IMC	
Amplitude théorique	
Amplitude choisie	

Figure 3.2.12 Formule pour déterminer le nombre théorique de classes

3. Une fois la formule entrée, taper **Enter**. On obtient environ 10,6.

Calcul de l'amplitude

Nombre de classes	10,596986
Étendue de l'IMC	
Valeur minimale de l'IMC	
Amplitude théorique	
Amplitude choisie	

Figure 3.2.13 Nombre théorique de classes selon la formule de Sturges

4. Dans la cellule **H6**, on souhaite déterminer l'étendue de l'indice de masse corporelle (valeur maximale moins la valeur minimale). Taper la formule `=SOUS.TOTAL(4;IMC)-SOUS.TOTAL(5;IMC)` (voir la Figure 3.2.14) suivie de la touche **Enter**.

La formule Excel **SOUS.TOTAL(4;IMC)** renvoie la plus grande valeur parmi la liste de valeurs (le maximum). La fonction Excel **SOUS.TOTAL(5;IMC)** renvoie la plus petite valeur parmi une liste de valeurs (le minimum).

Si l'on inscrit la formule `=MIN(IMC)`, ceci nous renvoie la valeur 0, le minimum de la série statistique non filtrée de l'indice de masse corporelle. Cependant, on veut exclure les femmes qui ont eu une valeur erronée de 0 comme indice de masse corporelle. On recherche plutôt la deuxième plus petite valeur.

Si l'on inscrit la formule `=MAX(IMC)`, on obtient la même valeur que `=SOUS.TOTAL(4;IMC)` puisqu'aucune valeur dans l'extrémité supérieure n'a été exclue.

Calcul de l'amplitude

Nombre de classes	10,596986
Étendue de l'IMC	=SOUS.TOTAL(4;IMC)-SOUS.TOTAL(5;IMC)
Valeur minimale de l'IMC	
Amplitude théorique	
Amplitude choisie	

Figure 3.2.14 Formule pour déterminer l'étendue de l'IMC

L'étendue vaut 48,9.

5. Dans la cellule H7, taper la formule =SOUS.TOTAL(5;IMC) suivie de la touche **Enter** (voir la Figure 3.2.15) pour déterminer la valeur minimale des données filtrées de l'indice de masse corporelle. Il est essentiel de connaître la valeur minimale pour s'assurer de l'inclure lorsque l'on regroupe les valeurs.

Calcul de l'amplitude

Nombre de classes	10,596986
Étendue de l'IMC	48,9
Valeur minimale de l'IMC	=SOUS.TOTAL(5;IMC)
Amplitude théorique	
Amplitude choisie	

Figure 3.2.15 Formule pour déterminer la valeur minimale de l'IMC (différente de 0)

La deuxième valeur la plus petite est 18,2.

6. On veut déterminer l'amplitude théorique de chaque classe, soit l'étendue divisée par le nombre théorique de classes. Dans la cellule H8, taper la formule =H6/H5 suivie de la touche **Enter** (voir la Figure 3.2.16). Dans des formules Excel, on peut faire référence à des cellules. La cellule H6 correspond à l'étendue et la cellule H5 correspond au nombre théorique de classes.

Calcul de l'amplitude

Nombre de classes	10,596986
Étendue de l'IMC	48,9
Valeur minimale de l'IMC	
Amplitude théorique	=H6/H5
Amplitude choisie	

Figure 3.2.16 Formule pour déterminer l'amplitude théorique des classes

7. On obtient une amplitude théorique d'environ 4,6. Ceci n'est pas un nombre entier. On choisit 5. On écrit donc 5 dans la cellule H9 (voir la Figure 3.2.17).

Calcul de l'amplitude

Nombre de classes	10,596986
Étendue de l'IMC	48,9
Valeur minimale de l'IMC	18,2
Amplitude théorique	4,6145195
Amplitude choisie	5

Figure 3.2.17 Détermination de l'amplitude choisie

Calcul de l'amplitude	
Nombre de classes	=1+10/3*LOG(SOUS.TOTAL(2;IMC))
Étendue de l'IMC	
Valeur minimale de l'IMC	
Amplitude théorique	
Amplitude choisie	

Figure 3.2.18 Étapes pour déterminer l'amplitude des classes

Grouper les valeurs en classes. Une fois l'amplitude des classes et la valeur de début déterminées, il est possible de forcer le regroupement voulu

des données.

Remarque 3.2.19 Afficher les éléments sans données. Lorsque l'on groupe les valeurs d'un tableau croisé dynamique, il est bien de s'assurer d'afficher les éléments sans données, car sinon, il se peut qu'un intervalle vide soit manquant sans qu'on s'en aperçoive.

Les étapes suivantes mènent à l'affichage des classes sans données.

1. Dans une cellule de la première colonne du tableau croisé dynamique de l'indice de masse corporelle, cliquer sur le bouton de droite de la souris (voir la [Figure 3.2.20](#)). Un menu déroulant s'affiche.

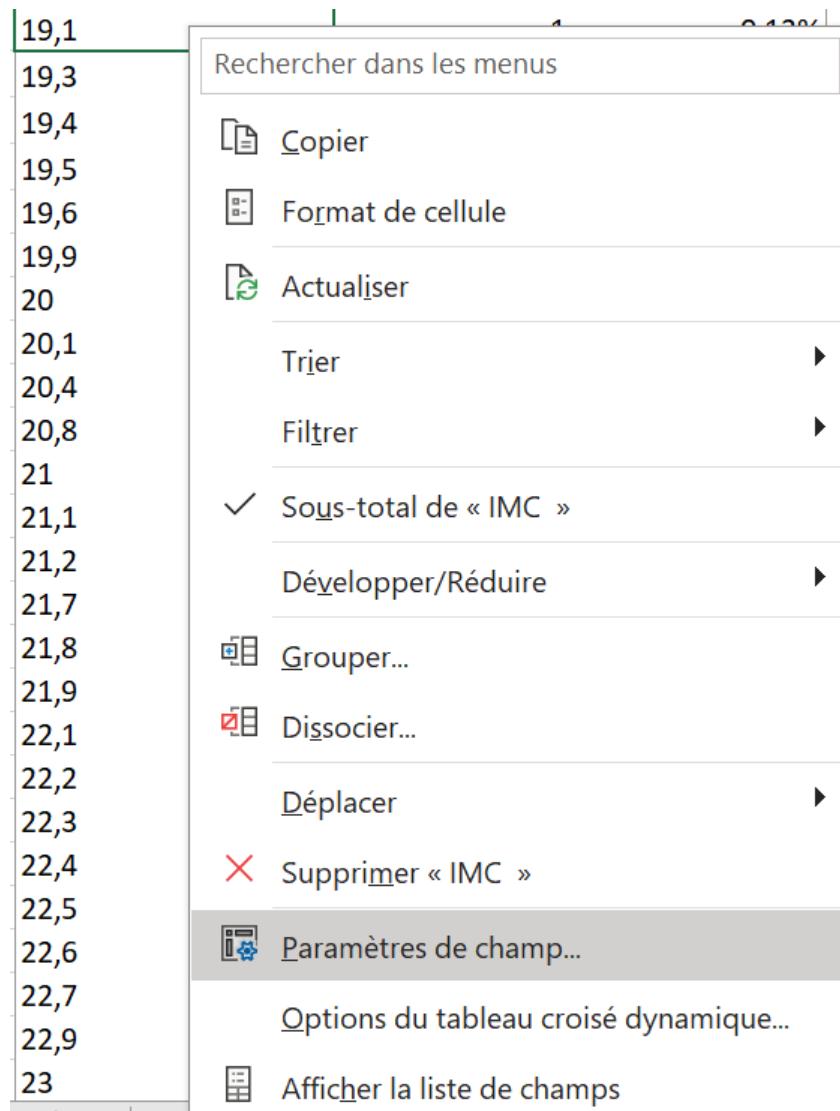


Figure 3.2.20 Menu déroulant de la première colonne d'un tableau croisé dynamique

2. Cliquer sur l'option **Paramètres de champ...** (voir la [Figure 3.2.20](#)).
3. Une boîte de dialogue s'affiche (voir la [Figure 3.2.21](#)). Cliquer sur l'onglet **Disposition et impression**. Cocher ensuite l'option **Afficher les éléments sans données**.

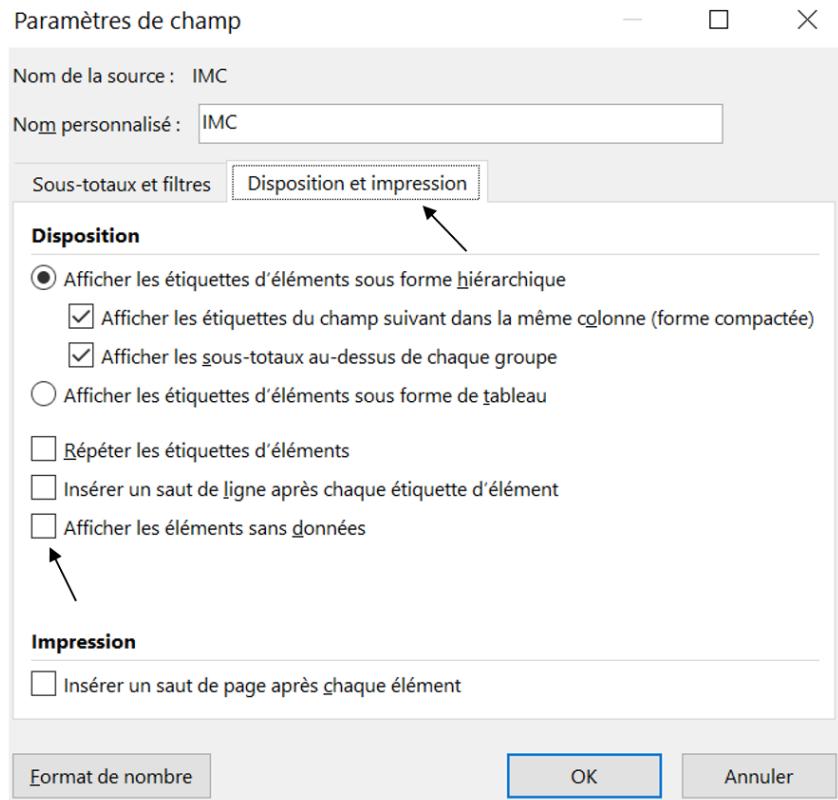


Figure 3.2.21 La boîte de dialogue pour afficher les éléments sans données

En cochant cette option, on s'assure que lorsque l'on groupera les valeurs de l'indice de masse corporelle en classes, Excel va afficher les classes qui ne contiennent aucun élément.

4. Cliquer sur **OK**.
5. Dans une cellule de la première colonne du tableau croisé dynamique, cliquer à nouveau sur le bouton de droite de la souris (voir la [Figure 3.2.22](#)). Un menu déroulant s'affiche.

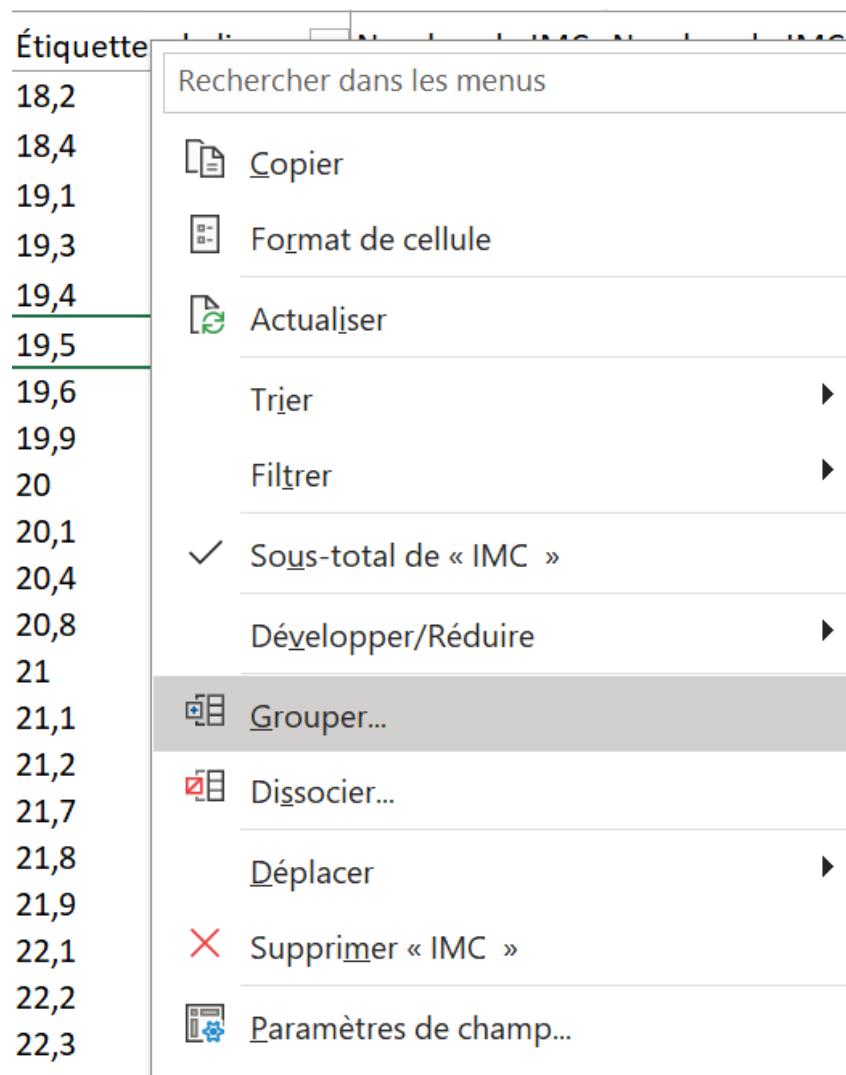


Figure 3.2.22 Menu déroulant de la première colonne d'un tableau croisé dynamique

6. Cliquer sur l'option **Groupier**.
7. Une boite de dialogue s'affiche permettant à l'utilisateur de choisir un groupement approprié pour les données (voir la [Figure 3.2.23](#)). Excel propose une borne inférieure (soit 0, la valeur minimale de la variable **IMC**, une borne supérieure (soit 67,1, la valeur maximale de l'IMC) et une amplitude pour les classes (soit 10).
Taper 15 comme valeur de début au lieu du choix suggéré d'Excel de 0 et taper 5 comme amplitude au lieu de 10 (voir la [Figure 3.2.23](#))

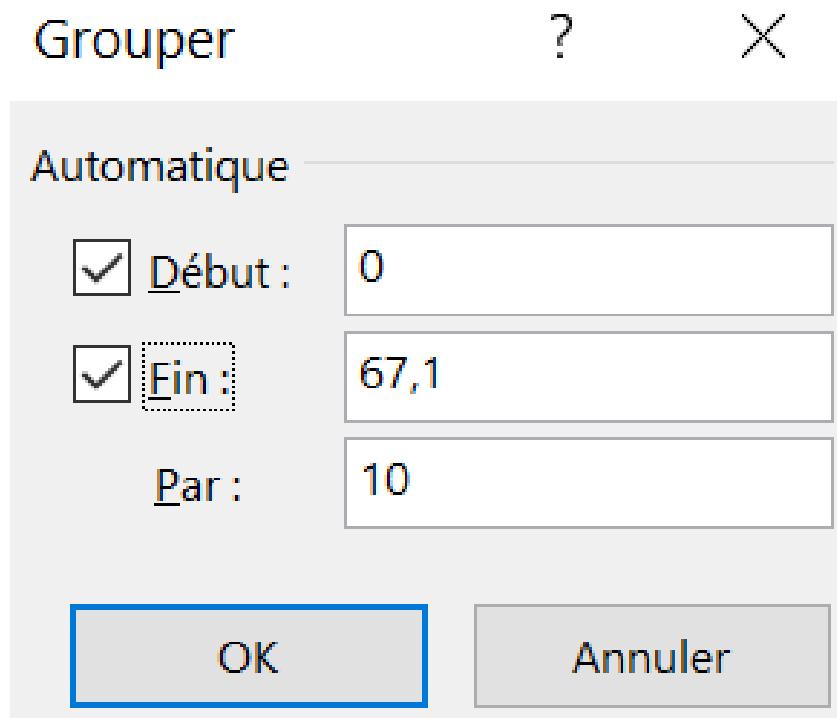


Figure 3.2.23 La boite de dialogue pour grouper les données d'un tableau croisé dynamique, ainsi que le choix de 15 comme valeur minimale et 5 comme amplitude des classes

En choisissant 15 comme valeur minimale de la première classe, on s'assure d'inclure la valeur minimale de 18,2. La valeur 15 est un choix logique puisque c'est un multiple de 5.

8. Cliquer sur **OK**. Le tableau croisé dynamique résultant groupe les valeurs de la variable **IMC** en classe d'amplitude valant 5 (voir la [Figure 3.2.24](#)). Puisque l'on a appliqué un filtre excluant les femmes une valeur nulle, le tableau commence à 15.

Étiquettes de lignes	Nombre de IMC	Nombre de IMC 2
<15		0,00%
15-20	13	1,72%
20-25	93	12,29%
25-30	179	23,65%
30-35	224	29,59%
35-40	150	19,82%
40-45	62	8,19%
45-50	27	3,57%
50-55	5	0,66%
55-60	3	0,40%
60-65		0,00%
65-70	1	0,13%
>70		0,00%
Total général	757	100,00%

Figure 3.2.24 Tableau croisé dynamique commençant à 15 avec des classes d'amplitude 5

On remarque que les trois dernières classes (excluant la classe > 70), celles de 55 à 60, 60 à 65 et 65 à 70, contiennent un très faible pourcentage de données. Dans ce cas, il est conseillé de créer une classe ouverte lorsque les premières ou les dernières classes ont peu de données (moins de 1% chacune) afin de faciliter l'interprétation des données.

9. Dans une cellule de la première colonne du tableau croisé dynamique, cliquer sur le bouton de droite de la souris. Sélectionner l'option **Grouper** à nouveau. Taper 50 comme valeur de fin, soit la limite inférieure de la première classe ouverte (voir la Figure 3.2.25).

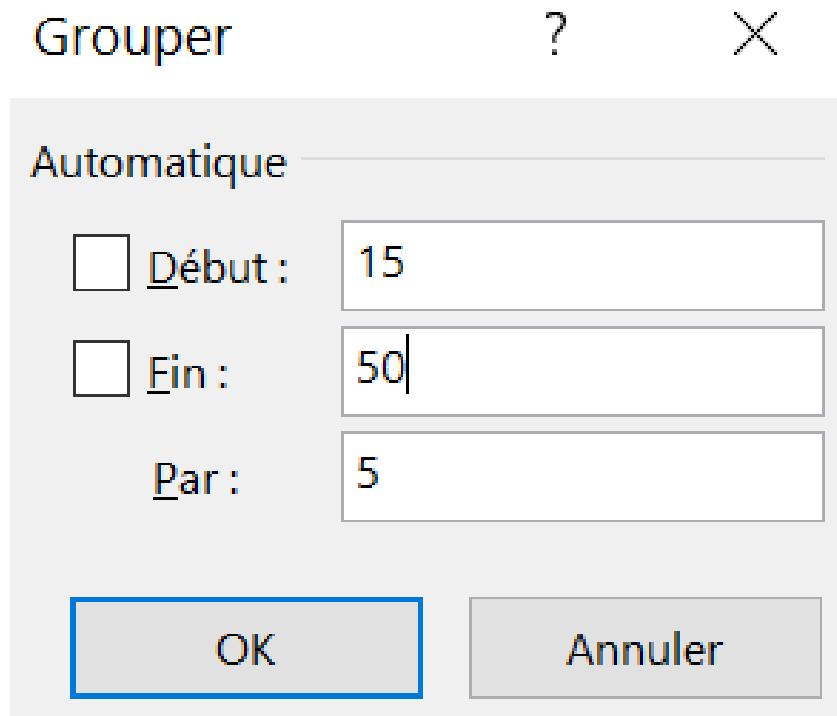


Figure 3.2.25 Choix de 50 comme fin pour créer une classe ouverte

Lorsque vient le temps de construire l'histogramme correspondant à ce tableau, on ne peut pas avoir une classe ouverte. Ainsi, on ferme la dernière classe en lui donnant la même amplitude que les autres classes. Ceci évite d'avoir beaucoup de bandes avec peu de données. Il ne faut cependant pas oublier d'écrire une note pour informer le lecteur de ce choix. Cliquer sur **OK**. Le tableau croisé dynamique résultant est présenté à la [Figure 3.2.26](#).

Étiquettes de lignes	Nombre de IMC	Nombre de IMC 2
<15		0,00%
15-20	13	1,72%
20-25	93	12,29%
25-30	179	23,65%
30-35	224	29,59%
35-40	150	19,82%
40-45	62	8,19%
45-50	28	3,70%
>50	8	1,06%
Total général	757	100,00%

Figure 3.2.26 Tableau croisé dynamique finale pour la variable **IMC**

Dans un espace sous le tableau croisé dynamique, faire la mise en forme du tableau de fréquences correspondant à la répartition des femmes selon

l'indice de masse corporelle. La version finale du tableau de fréquences est présentée à la [Figure 3.2.27](#) (voir la [Sous-section 2.2.2](#)).

Répartition d'un échantillon de femmes d'origine pima selon l'indice de masse corporelle, Arizona, année inconnue

IMC (kg/m²)	Nombre de femmes	Pourcentage de femmes
[15 ; 20[13	1,72%
[20 ; 25[93	12,29%
[25 ; 30[179	23,65%
[30 ; 35[224	29,59%
[35 ; 40[150	19,82%
[40 ; 45[62	8,19%
[45 ; 50[28	3,70%
50 et plus	8	1,06%
Total	757	100,00%

Source : National Institute of Diabetes and Digestive and Kidney Diseases (USA)

Figure 3.2.27 Version définitive du tableau de fréquences de la variable *IMC*

3.2.1.5 Créer un histogramme

Pour représenter la répartition d'un échantillon de femmes en fonction de l'indice de masse corporelle, soit une variable quantitative continue, l'histogramme est un choix de graphique approprié lorsque le nombre d'unités statistiques est important.

Dans le tableau de fréquences de la [Figure 3.2.27](#), la dernière classe est désormais fermée de 50 à 55. Cependant, trois données se situent entre 55 et 60, et une autre entre 65 et 70. Cela réduit légèrement le nombre de classes par rapport aux prévisions faites à la [Sous sous-section 3.2.1.4](#). Afin d'éviter un histogramme avec plusieurs bandes aux extrémités contenant peu de données, le graphique à construire devrait respecter les choix effectués lors de la création du tableau de fréquences correspondant.

Les étapes qui suivent mènent à la construction de l'histogramme représentant la répartition de l'échantillon de femmes d'origine pima selon leur indice de masse corporelle.

1. Sélectionner la plage de données représentant les fréquences relatives de l'indice de masse corporelle, soit la plage de cellules **D5:D12** (voir la [Figure 3.2.28](#)). Ne pas sélectionner les classes de l'indice de masse corporelle, ni les titres des colonnes, ni les données de la ligne *Total*.

A	B	C	D
1			
2			
3	Étiquettes de lignes <input checked="" type="checkbox"/>	Nombre de IMC	Nombre de IMC 2
4	<15		0,00%
5	15-20	13	1,72%
6	20-25	93	12,29%
7	25-30	179	23,65%
8	30-35	224	29,59%
9	35-40	150	19,82%
10	40-45	62	8,19%
11	45-50	28	3,70%
12	>50	8	1,06%
13	Total général	757	100,00%

Figure 3.2.28 Sélection des fréquences relatives de l'IMC

2. Copier la plage sélectionnée et coller ces valeurs dans la cellule **H19** (voir la [Figure 3.2.29](#)).
3. Ajouter une classe fictive nulle avant la première valeur et après la dernière valeur (0% dans les cellules **H18** et **H27**) (voir la [Figure 3.2.29](#)). Cette étape vise à faciliter la mise en forme de l'histogramme et à améliorer son apparence.
4. Il faut écrire les bornes inférieures des classes l'une à la suite de l'autre, commençant par la borne inférieure de la première classe et finissant par la borne supérieure de la dernière. Dans la cellule **G19**, écrire la borne inférieure de la première classe, soit 15. Dans la cellule **G20**, écrire la borne inférieure de la deuxième classe, soit 20. Dans la cellule **G21**, écrire 25 (voir la [Figure 3.2.29](#)).
5. Sélectionner la plage de cellules **G19:G21** (voir la [Figure 3.2.29](#)). La plage est encadré d'une bordure verte et un petit carré vert apparaît dans le coin inférieur droit. Approcher le curseur au-dessus du carré vert. Dès qu'une croix noire apparaît, double-cliquer (voir la [Figure 3.2.29](#) et la [Sous sous-section 1.2.7.2](#)).

	1,72%
	12,29%
	23,65%
	29,59%
	19,82%
	8,19%
	3,70%
	1,06%

Figure 3.2.29 Séquence de collage des fréquences relatives de l'IMC et inscription des bornes inférieures des classes

6. Sélectionner les valeurs des fréquences relatives de la colonne de droite incluant les 0% avant et après, soit la plage de cellules **G18:H27** (voir la Figure 3.2.30).

	0,00%
15	1,72%
20	12,29%
25	23,65%
30	29,59%
35	19,82%
40	8,19%
45	3,70%
50	1,06%
55	0,00%

Figure 3.2.30 Sélection des fréquences relatives

7. Cliquer sur l'onglet **Insertion**. Dans le groupe **Graphiques**, cliquer sur l'icône **Insérer un histogramme ou un graphique à barres** (voir la Figure 3.2.31).

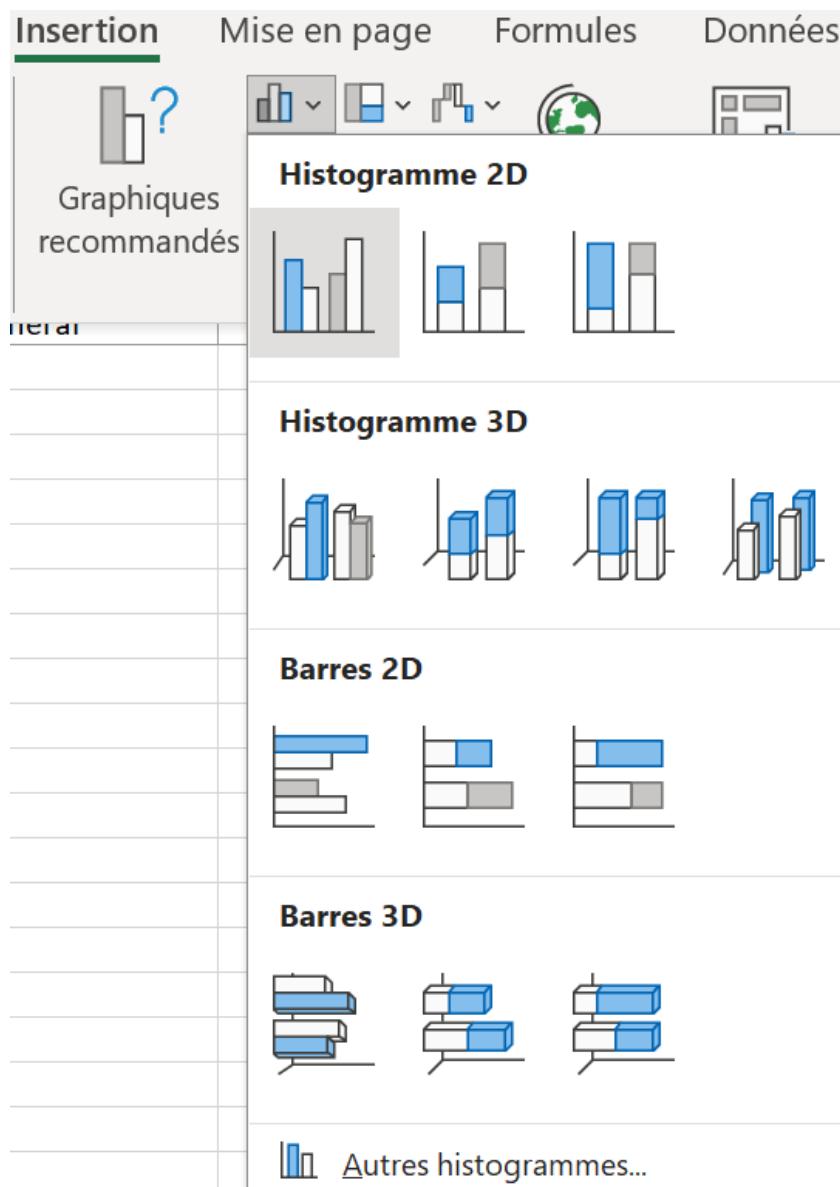


Figure 3.2.31 Sélection de l'icône *Insérer un histogramme ou un graphique à barres*

8. Dans la section **Histogramme 2D**, sélectionner la première option (voir la [Figure 3.2.31](#)). Le graphique ci-dessous s'affiche dans la feuille de calcul (voir la [Figure 3.2.32](#))

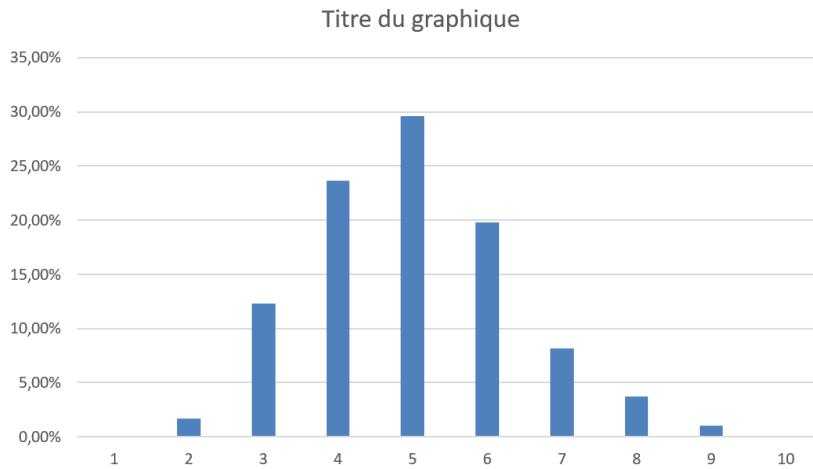


Figure 3.2.32 Graphique créé après la sélection d'insérer un *histogramme 2D*

Il est évident que l'allure de ce graphique ne correspond pas à celle d'un histogramme. Il faut effectuer sa mise en forme.

9. Ajouter un titre au graphique et un titre à chaque axe (voir la [Figure 2.2.7](#) et la [Figure 3.2.33](#)).

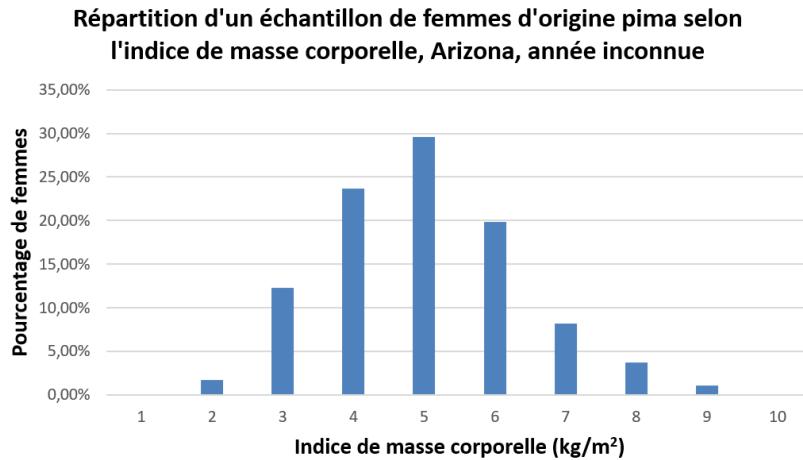


Figure 3.2.33 Ajout de titres

10. Les bandes d'un histogramme doivent être collées. Sur une des bandes, cliquer avec le bouton de droite de la souris et sélectionner l'option **Mettre en forme une série de données** (voir la [Figure 3.2.34](#)).

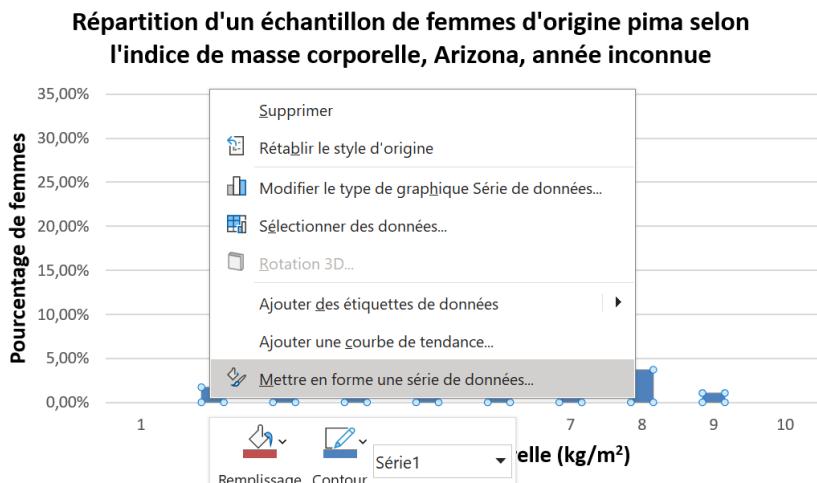


Figure 3.2.34 Sélection de l'option *Mettre en forme une série de données*

11. Une boîte de dialogue grise avec des options de séries s'affiche à la droite de la feuille de calcul (voir la [Figure 3.2.35](#)).
À l'option **Largeur des intervalles**, Excel met 219% par défaut. Pour un histogramme, on veut que la largeur entre les bandes soit de 0%. Effacer 219 et taper 0 (voir la [Figure 3.2.35](#)).

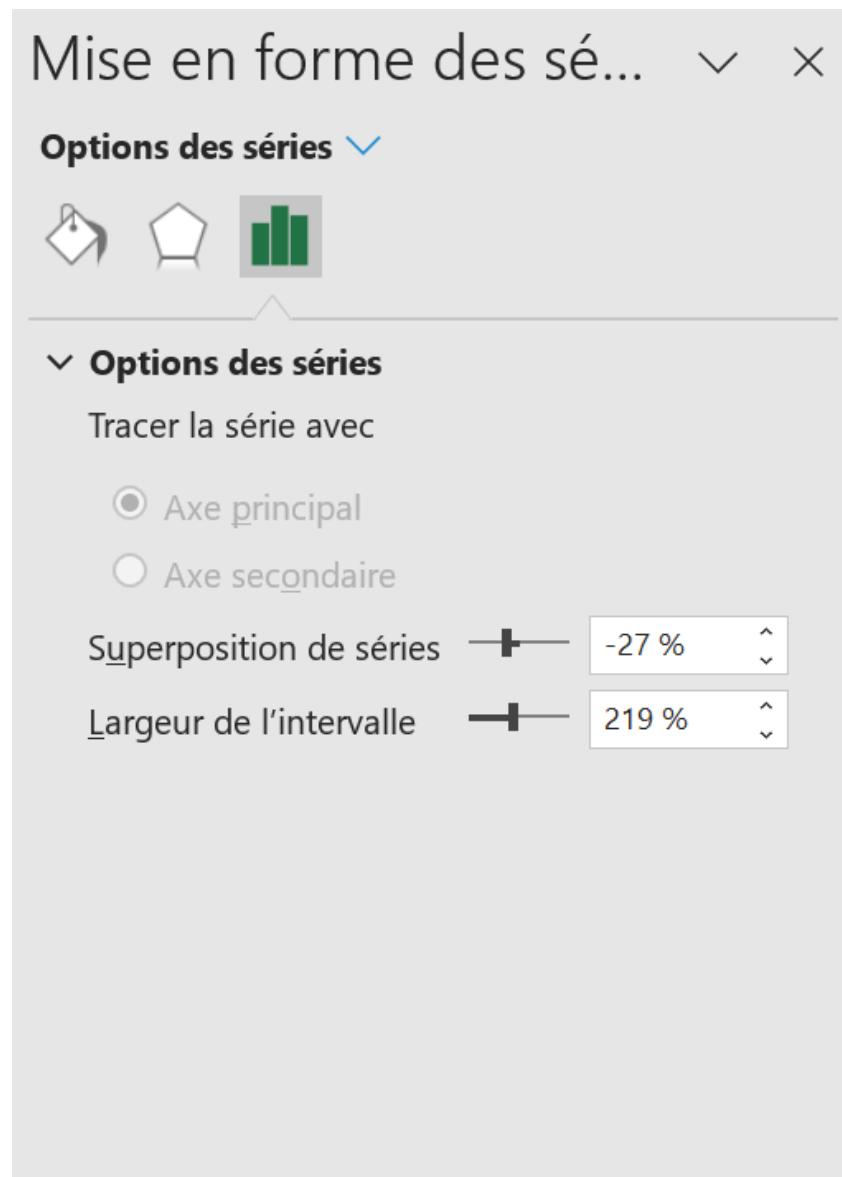


Figure 3.2.35 Mettre 0% comme largeur des intervalles entre les bandes au lieu de 219%

L'allure des bandes est désormais la suivante (voir la [Figure 3.2.36](#)).

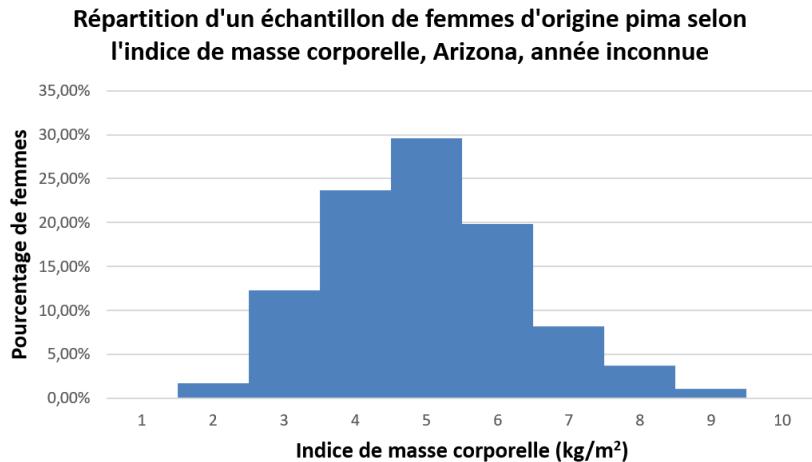


Figure 3.2.36 Histogramme avec 0% comme largeur des intervalles

12. Cliquer avec le bouton de droite sur une des bandes et sélectionner à nouveau l'option **Mettre en forme une série de données**. Sélectionner la première icône **Remplissage et couleur** qui ressemble à un pot de peinture (voir la [Figure 3.2.37](#)). Dans le menu **Bordure**, il est possible de modifier la couleur des bordures des bandes de l'histogramme. Choisir la couleur noire en trait plein (voir la [Figure 3.2.38](#) pour le résultat).

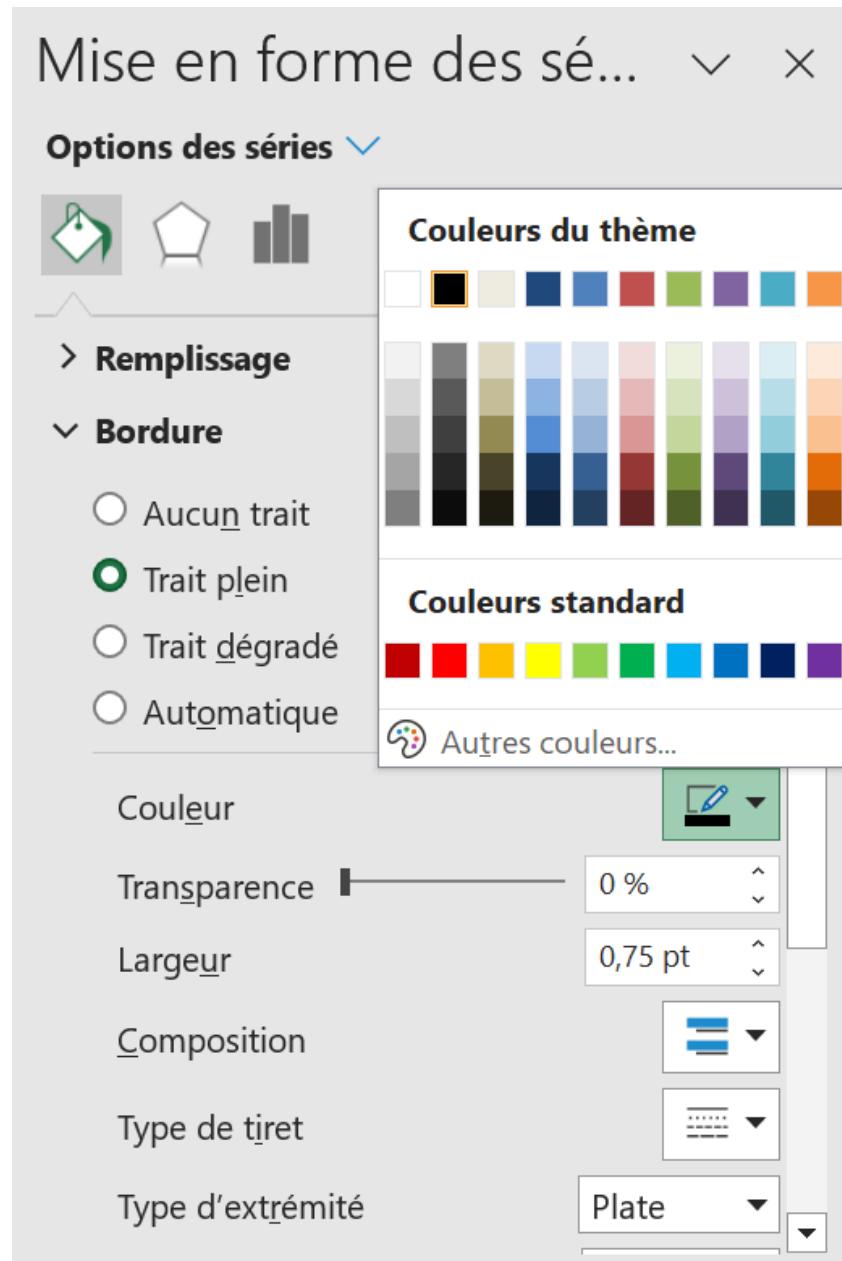


Figure 3.2.37 Sélection d'une bordure noire pour les bandes de l'histogramme

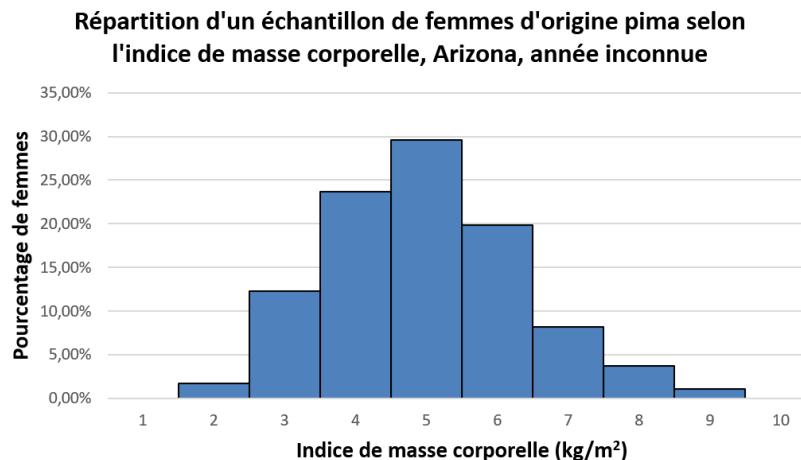


Figure 3.2.38 Bandes de l'histogramme avec une bordure noire

- Il faut ajuster les valeurs de l'axe horizontal pour les faire correspondre aux bornes des classes de l'indice de masse corporelle. Cliquer sur une des bandes de l'histogramme avec le bouton de droite de la souris et sélectionner l'option **Sélectionner des données...** (voir la Figure 3.2.39).

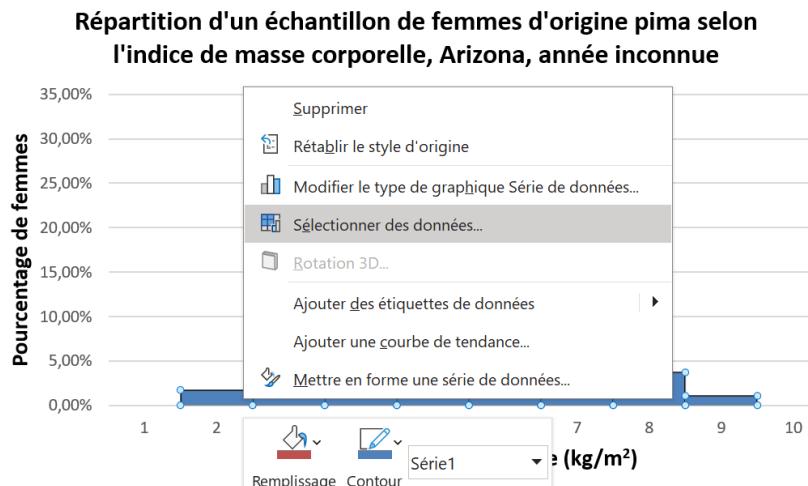


Figure 3.2.39 Sélection de l'option **Sélectionner des données...**

- Une boîte de dialogue s'affiche. Sous l'option **Étiquettes de l'axe horizontal (abscisse)** (menu droit de la boîte), cliquer l'icône **Modifier** (voir la Figure 3.2.40).

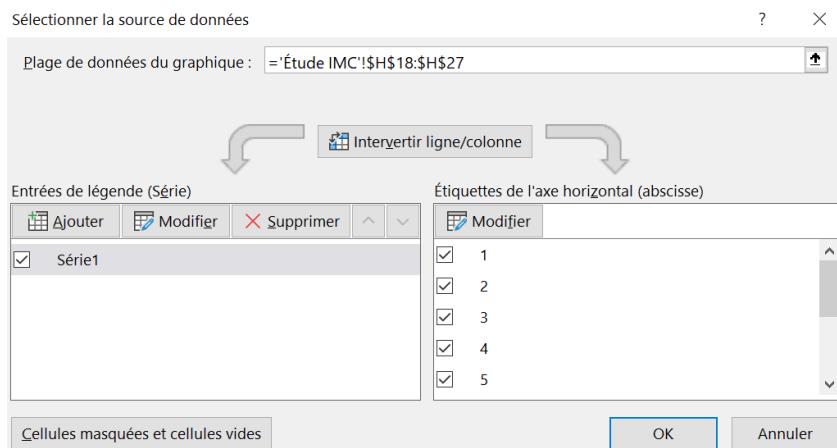


Figure 3.2.40 Sélection de l'icône *Modifier*

15. Une autre boîte de dialogue s'affiche et permet la sélection d'une plage de données (voir la [Figure 3.2.41](#)).

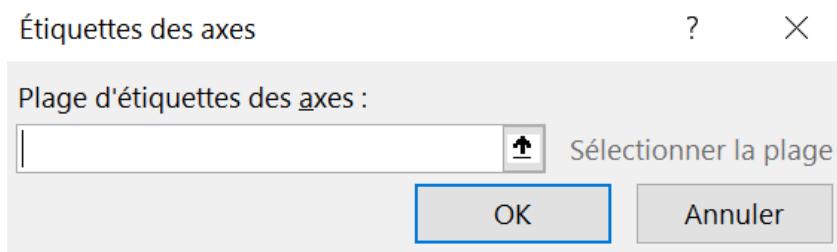


Figure 3.2.41 Boîte de dialogue pour la sélection des étiquettes de l'axe horizontal

16. Sélectionner la plage de cellules **G19:G27**, soit les valeurs 15 à 55 (voir la [Figure 3.2.42](#)).

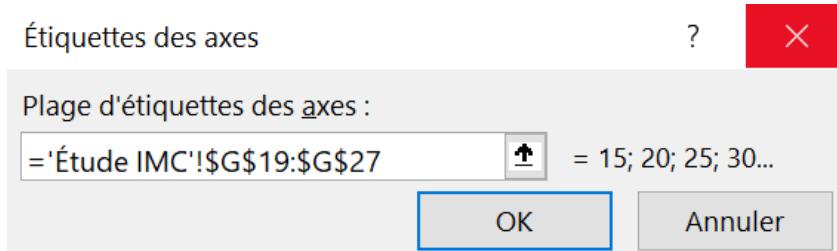


Figure 3.2.42 Sélection de la plage de cellules **G19:G27**

17. Cliquer sur *OK* deux fois. L'histogramme résultant ressemble à la [Figure 3.2.43](#).

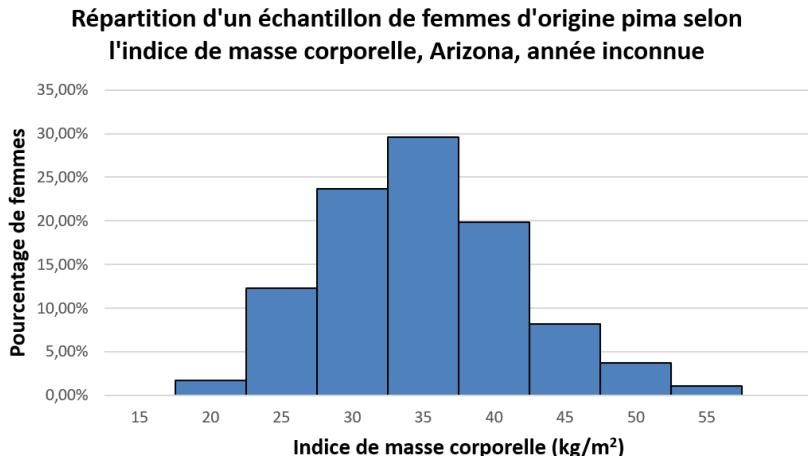


Figure 3.2.43 Histogramme avec les bonnes valeurs sur l'axe des abscisses

18. Les valeurs des étiquettes des abscisses ne sont pas bien alignées. Il faut les aligner à droite. Cliquer sur une des valeurs de l'abscisse jusqu'à ce que l'entièreté des valeurs soit comprise dans un encadré. Cliquer sur l'onglet **Accueil**. Dans le groupe **Alignement**, cliquer sur l'icône **Aligner à droite** (voir la [Figure 3.2.44](#)).

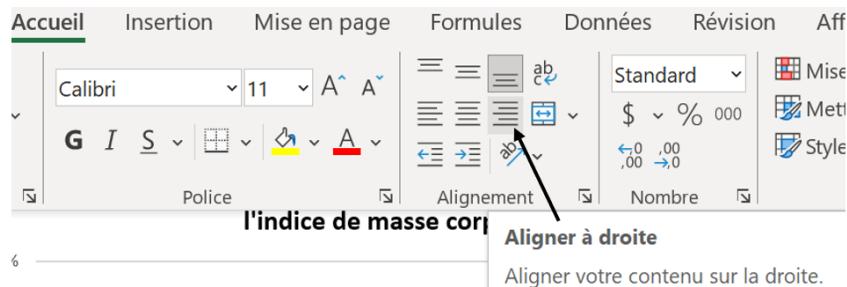


Figure 3.2.44 Sélection de l'option *Aligner à droite*

L'histogramme résultant ressemble à la [Figure 3.2.45](#).

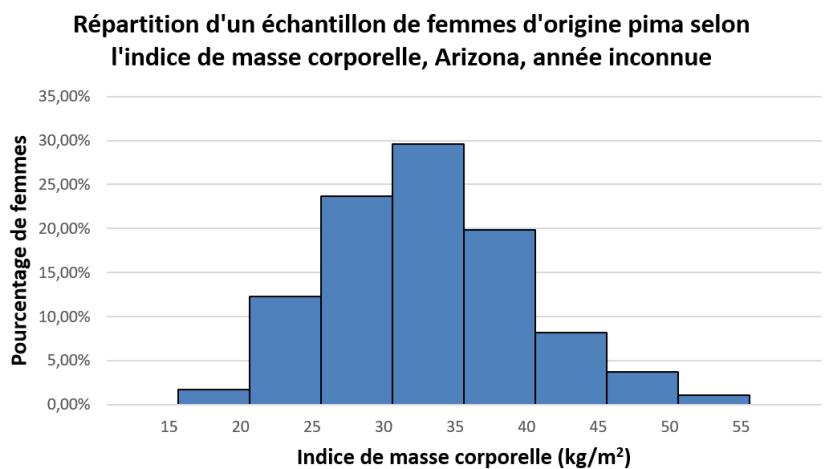


Figure 3.2.45 Alignement à droite des valeurs de l'axe des abscisses

19. Il est possible d'ajouter les étiquettes des fréquences relatives au-dessus des bandes. Cliquer avec le bouton de droite sur une des bandes de l'histogramme et sélectionner l'option *Ajouter des étiquettes de données* Figure 3.2.46.

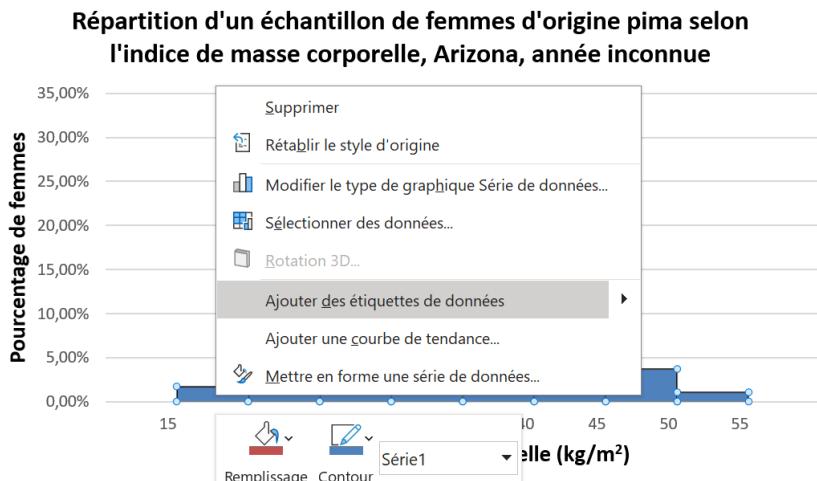


Figure 3.2.46 Sélection de l'option *Ajouter des étiquettes de données*

L'histogramme résultant ressemble à la Figure 3.2.47.

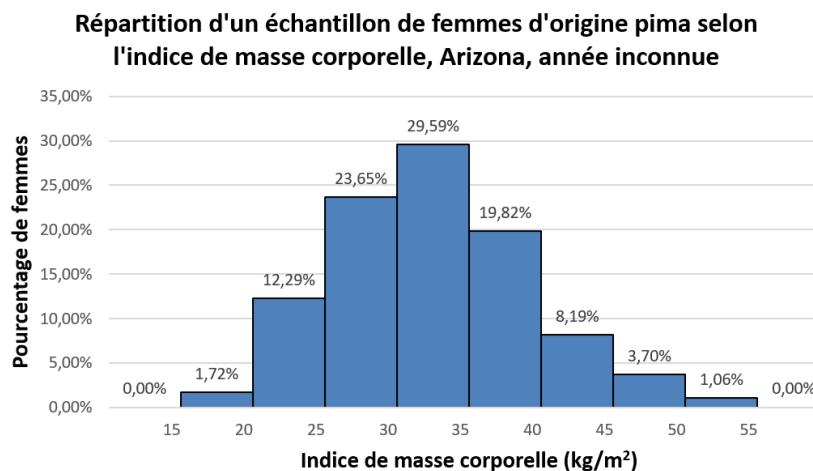
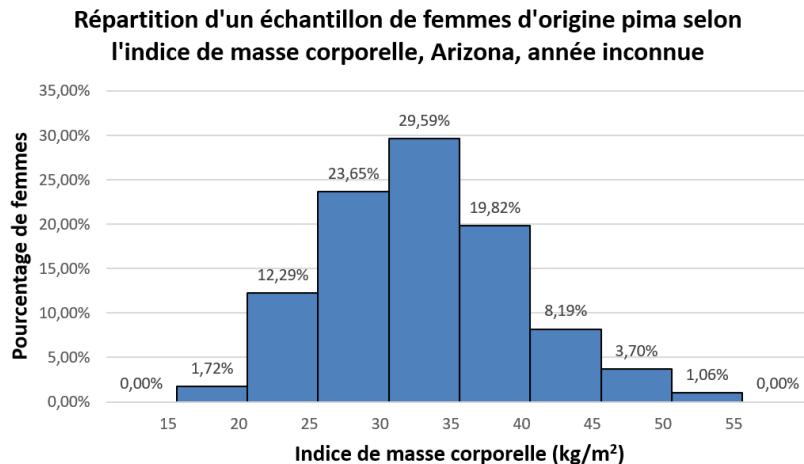


Figure 3.2.47 Ajout des étiquettes de données

20. Comme il y a un saut de valeurs entre la première classe et l'axe des ordonnées, il est recommandé de l'indiquer dans une note au bas du graphique.
21. L'étape finale consiste à mettre la source des données en dessous de l'histogramme (voir la Figure 3.2.48).



Source : National Institute of Diabetes and Digestive and Kidney Diseases (USA)

Figure 3.2.48 Version définitive de l'histogramme avec la source des données et la note de saut

3.2.1.6 Mesures descriptives

Avec Excel, il est possible de calculer les mesures descriptives d'une variable comme les mesures de tendance centrale (moyenne, médiane et mode), les mesures de dispersion (écart type et coefficient de variation) et les mesures de position (quantiles). Pour approfondir l'étude de l'indice de masse corporelle de l'échantillon de femmes d'origine pima, on calculera et interprétera plusieurs mesures. De plus, on a jugé bon de comparer les valeurs des mesures descriptives avec et sans le filtrage des valeurs nulles de l'indice de masse corporelle.

Mesures descriptives sans filtre. On commence par créer un tableau pour reporter toutes les mesures statistiques que l'on va calculer.

1. Dans la cellule **K2** de la feuille de calcul *Étude IMC*, taper le titre *Mesures descriptives de l'indice de masse corporelle d'un échantillon de femmes d'origine pima, Arizona* (voir la [Figure 3.2.49](#)).

Mesures descriptives de l'indice de masse corporelle d'un échantillon de femmes d'origine pima, Arizona

Figure 3.2.49 Titre du tableau de mesures descriptives

Puisque le titre est long, on va fusionner quelques cellules et centrer le texte.

2. Sélectionner la plage de cellules **K2:L3** (voir la [Figure 3.2.50](#)).

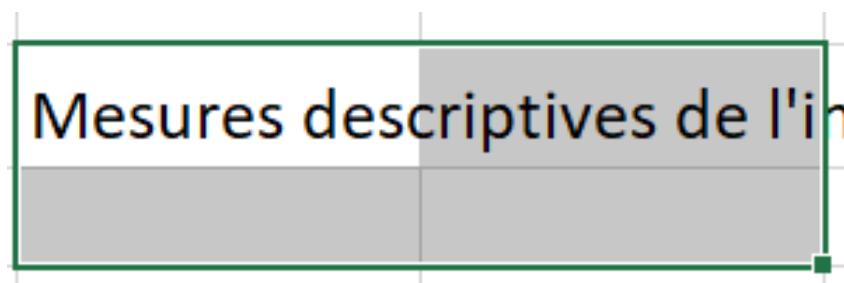


Figure 3.2.50 Sélection des cellules K2:L3

3. Cliquer sur l'onglet **Accueil**. Cliquer sur l'icône **Alignement** et sélectionner l'option **Fusionner et centrer** ainsi que l'option **Renvoyer à la ligne automatiquement** (voir la Figure 3.2.51).

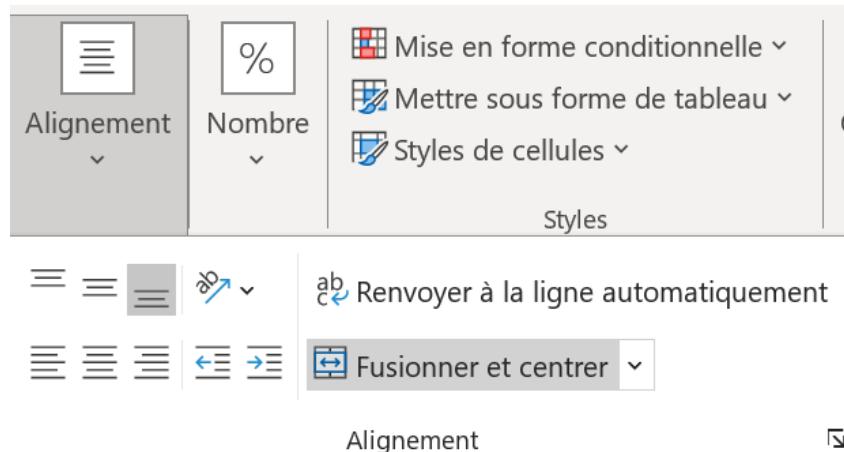


Figure 3.2.51 Sélection des options *Fusionner et centrer* et *Renvoyer à la ligne automatiquement*

4. Ajuster la taille des colonnes **K** et **L** ainsi que celle des lignes **2** et **3** pour que le titre soit bien visible en entier.
5. Dans les cellules **K4:K12**, taper le nom des mesures descriptives à calculer (voir la Figure 3.2.52).

Mesures descriptives de l'indice de masse corporelle d'un échantillon de femmes d'origine pima, Arizona	
Nombre de données	
Minimum	
Maximum	
Moyenne	
Médiane	
Mode	
Écart type corrigé	
Coefficient de variation	
Premier quartile	

Figure 3.2.52 Mesures descriptives à calculer

6. Ajouter une bordure noire à ce tableau (voir la [Figure 3.2.53](#)).

Mesures descriptives de l'indice de masse corporelle d'un échantillon de femmes d'origine pima, Arizona	
Nombre de données	
Minimum	
Maximum	
Moyenne	
Médiane	
Mode	
Écart type corrigé	
Coefficient de variation	
Premier quartile	

Figure 3.2.53 Bordure au tableau de mesures descriptives

7. Si l'on ne connaît pas la formule pour la fonction que l'on veut, on peut faire une recherche (voir la [Figure 3.2.10](#) et la [Figure 3.2.11](#)).

Pour les fonctions statistiques, les formules sont assez intuitives. Pour le nombre de données, la formule est **NB**. Dans la cellule **L4**, taper **=NB(IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.54](#)).

On voit maintenant l'avantage d'avoir nommé la variable **IMC** dans le laboratoire 1. Il n'est pas nécessaire de retourner à la feuille de calcul **Données** et de sélectionner la plage de données avec les valeurs de l'IMC.

8. Dans la cellule **L5**, taper **=MIN(IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.54](#)).

Sans filtre, on rappelle que le minimum de l'IMC est 0.

9. Dans la cellule **L6**, taper **=MAX(IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.54](#)).

10. Dans la cellule **L7**, taper **=MOYENNE(IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.54](#)).

11. Dans la cellule **L8**, taper **=MEDIANE(IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.54](#)).

12. Dans la cellule **L9**, taper **=MODE.SIMPLE(IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.54](#)).

13. Dans la cellule **L10**, taper **=ECARTYPE.STANDARD(IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.54](#)). La formule **ECARTYPE.PEARSON**

calcule l'écart type de données issues d'une population. La formule **ECARTTYPE.STANDARD** renvoie l'écart type corrigé de données provenant d'un échantillon.

14. Pour le calcul du coefficient de variation, Excel ne dispose pas d'une formule intégrée dans son logiciel. Cependant, comme le calcul repose sur la moyenne et l'écart type, il peut être facilement effectué manuellement. Dans la cellule **L11**, taper **=L10/L7** et appuyer sur la touche **[Enter]**. Il est possible de sélectionner les cellules **L10** et **L7** au lieu de les taper (voir la Figure 3.2.54).

Mesures descriptives de l'indice de masse corporelle d'un échantillon de femmes d'origine pima, Arizona	
Nombre de données	=NB(IMC)
Minimum	
Maximum	
Moyenne	
Médiane	
Mode	
Écart type corrigé	
Coefficient de variation	
Premier quartile	

Figure 3.2.54 Formules pour calculer les différentes mesures descriptives de l'IMC

15. Afficher le coefficient de variation en pourcentage. Sélectionner la cellule **L11**. Cliquer sur l'onglet **Accueil**. Dans le menu de l'icône **Nombre**, cliquer sur l'option **Style de pourcentage (%)** (voir la Figure 3.2.55).

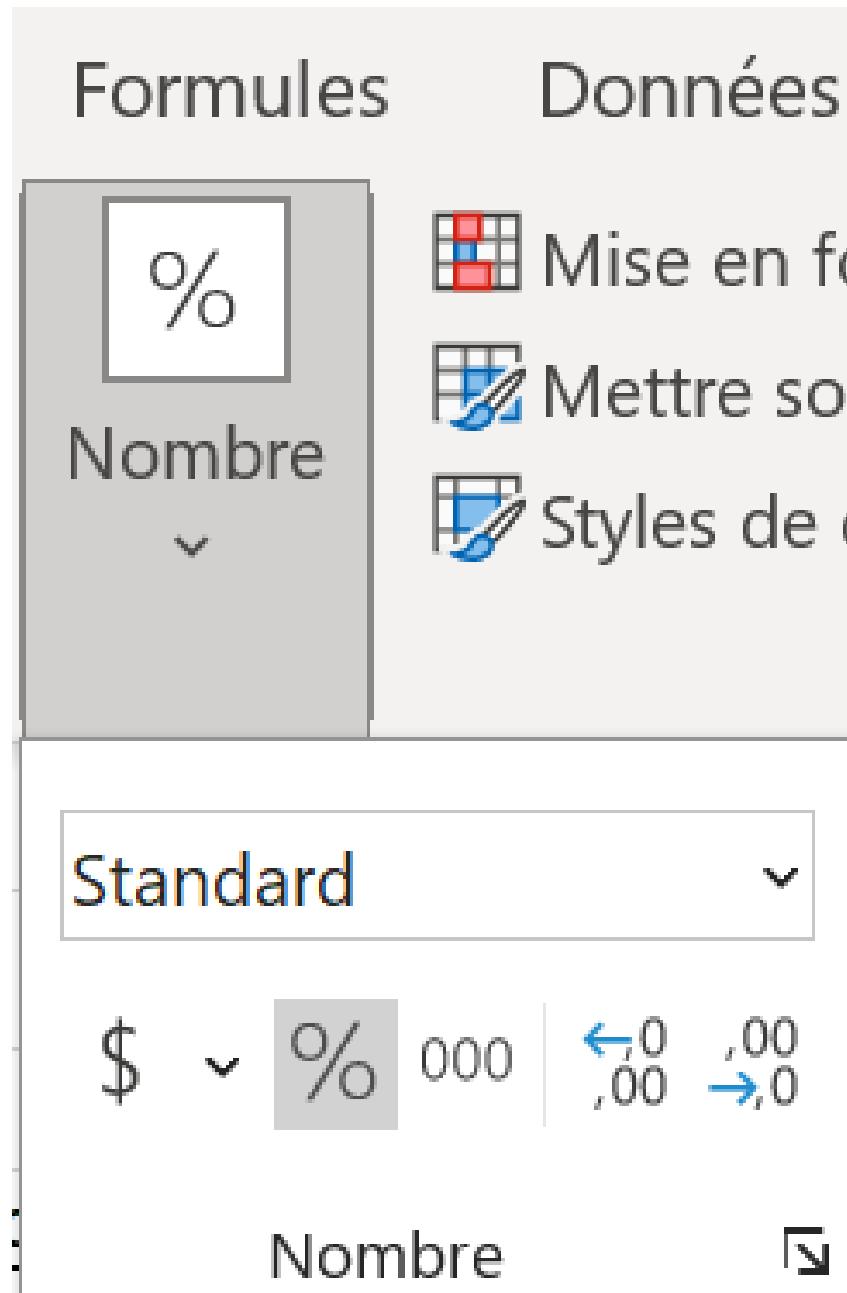


Figure 3.2.55 Afficher le coefficient de variation en pourcentage

16. Dans la cellule L12, taper =CENTILE.INCLURE(IMC;0,25) et appuyer sur la touche **Enter** (voir la [Figure 3.2.56](#)). Le centile recherché est le 25^e. Le deuxième paramètre à inscrire dans la formule Excel est le centile recherché en notation décimale. Pour les quartiles, Excel a une formule leur étant dédiée, soit QUARTILE.INCLURE.

Mesures descriptives de l'indice de masse corporelle d'un échantillon de femmes d'origine pima, Arizona	
Nombre de données	768
Minimum	0
Maximum	67,1
Moyenne	31,99
Médiane	32,00
Mode	32,00
Écart type corrigé	7,88
Coefficient de variation	
Premier quartile	=CENTILE.INCLURE(IMC;0,25)

Figure 3.2.56 Formule pour calculer le premier quartile de l'IMC

Les valeurs des mesures sans filtrage se retrouvent à la [Figure 3.2.57](#)

Mesures descriptives de l'indice de masse corporelle d'un échantillon de femmes d'origine pima, Arizona	
Nombre de données	768
Minimum	0
Maximum	67,1
Moyenne	31,99
Médiane	32,00
Mode	32,00
Écart type corrigé	7,88
Coefficient de variation	25%
Premier quartile	27,30

Figure 3.2.57 Mesures descriptives de l'IMC sans filtrage

17. Centrer chaque mesure descriptive dans sa cellule et l'afficher avec une décimale (sauf pour le nombre de données) à l'aide des fonctionnalités disponibles dans l'onglet **Accueil** (voir la [Figure 3.2.58](#)).

Mesures descriptives de l'indice de masse corporelle d'un échantillon de femmes d'origine pima, Arizona	
Nombre de données	768
Minimum	0,0
Maximum	67,1
Moyenne	32,0
Médiane	32,0
Mode	32,0
Écart type corrigé	7,9
Coefficient de variation	24,6%
Premier quartile	27,3

Figure 3.2.58 Formatage des mesures descriptives de l'IMC

Mesures descriptives avec filtre. Puisque la variable **IMC** du tableau **Échantillon** a été filtrée, on est en mesure d'effectuer des calculs avec les données filtrées.

1. Dans la feuille de calcul **Étude IMC**, copier la plage de cellules **K2:L12**.
2. Coller cette plage de cellules dans les cellules **N2:O12**. Supprimer les valeurs des mesures statistiques des cellules **O4:O12** en sélectionnant la plage et en cliquant sur la touche **[suppr]** du clavier. Modifier le titre pour qu'on lise *Mesures descriptives de l'indice de masse corporelle d'un échantillon filtré de femmes d'origine pima, Arizona*
3. Ajuster la taille des colonnes **N** et **O** ainsi que les lignes 2 et 3 au besoin.
4. Dans la cellule **O4**, taper **=Sous.Total(2;IMC)** et appuyer sur la touche **[Enter]** (voir la Figure 3.2.59).

Le chiffre 2 fait référence au calcul du nombre de données. Les nombres 1 à 11 spécifie la fonction à utiliser pour calculer le sous-total. Le deuxième paramètre, soit **IMC**, fait référence à la plage de données dont on souhaite calculer le sous-total.

Mesures descriptives de l'indice de masse corporelle d'un échantillon filtré de femmes d'origine pima, Arizona	
Nombre de données	=SOUS.TOTAL(2;IMC)
Minimum	
Maximum	
Moyenne	
Médiane	
Mode	
Écart type corrigé	
Coefficient de variation	
Premier quartile	

Figure 3.2.59 Formule pour calculer le nombre de données du sous-total de la variable **IMC** filtrée

5. Dans la cellule **05**, taper **=SOUS.TOTAL(5;IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.60](#)).
6. Dans la cellule **06**, taper **=SOUS.TOTAL(4;IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.60](#)).
7. Dans la cellule **07**, taper **=SOUS.TOTAL(1;IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.60](#)).
8. Malheureusement, Excel n'est pas en mesure de calculer la médiane du sous-total d'une plage de données. Il est donc nécessaire d'utiliser la fonction Excel **SI**, fonction qui permet d'appliquer des conditions lors de l'emploi d'une formule. Dans la cellule **08**, taper **=MEDIANE(SI(IMC <> 0 ;IMC))** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.60](#)).

SI(IMC <> 0 ;IMC) renvoie les valeurs de l'IMC différentes de 0. Excel fait ensuite le calcul de la médiane de ces valeurs.
9. Malheureusement, Excel n'est pas en mesure de calculer le mode du sous-total d'une plage de données. Il faut donc utiliser la fonction Excel **SI** encore une fois. Dans la cellule **09**, taper **=MODE.SIMPLE(SI(IMC <> 0 ;IMC))** (voir la [Figure 3.2.60](#)).
10. Dans la cellule **010**, taper **=SOUS.TOTAL(7;IMC)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.60](#)).
11. Dans la cellule **011**, taper **=010/07** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.60](#)).
12. La fonction **SOUS.TOTAL** ne permet pas de calculer des centiles. Il faut utiliser la fonction **SI**. Dans la cellule **012**, taper **=CENTILE.INCLURE(SI(IMC<>0;IMC);0,25)** et appuyer sur la touche **[Enter]** (voir la [Figure 3.2.60](#)).

Les valeurs des mesures descriptives de la variable IMC filtrée se retrouvent à la [Figure 3.2.60](#)

Mesures descriptives de l'indice de masse corporelle d'un échantillon filtré de femmes d'origine pima, Arizona	
Nombre de données	=SOUS.TOTAL(3;IMC)
Minimum	
Maximum	
Moyenne	
Médiane	
Mode	
Écart type corrigé	
Coefficient de variation	
Premier quartile	

Figure 3.2.60 Mesures descriptives de l'IMC filtrée

Pour une étude préliminaire d'un échantillon, le troisième quartile est une mesure pertinente pour examiner l'allure globale d'une distribution et l'étalement des données. Combiné avec le premier quartile, il permet de détecter une éventuelle asymétrie dans la distribution des données. Un tableau final des mesures descriptives d'une étude préliminaire de la variable IMC filtrée se retrouve à la [Figure 3.2.61](#).

Mesures descriptives de l'indice de masse corporelle d'un échantillon filtré de femmes d'origine pima, Arizona

Nombre de données	757
Minimum	18,2
Maximum	67,1
Moyenne	32,5
Médiane	32,3
Mode	32,0
Écart type corrigé	6,9
Coefficient de variation	21,3%
Premier quartile	27,5
Troisième quartile	36,6

Figure 3.2.61 Mesures descriptives d'une étude préliminaire de l'IMC filtrée

Le diagramme à quartiles, communément appelé la «boîte à moustaches», si le diagramme est placé horizontalement, illustre le premier quartile, la médiane, le troisième quartile, une valeur minimale qui est située à une distance d'une fois et demie l'écart interquartile en dessous du premier quartile, ainsi qu'une valeur maximale située à une distance d'une fois et demie l'écart interquartile au-dessus du troisième quartile. Ce graphique permet également de détecter d'un coup d'œil les asymétries possibles à l'aide de la longueur des moustaches, les deux lignes qui s'étendent des quartiles aux valeurs minimale et maximale.

Avec Excel, les détails pour tracer un diagramme à quartiles de la variable **IMC** sont présentés ci-dessous. Ces étapes ne sont pas à faire pour ce laboratoire. Elles sont simplement présentées pour démontrer la pertinence de ce type de graphique.

1. Dans la feuille **Données**, sélectionner la plage de données de la variable dont on souhaite tracer le diagramme à quartiles. Ici, c'est la variable **IMC**. S'assurer que les données de cette variable sont filtrées pour enlever les valeurs nulles. La sélection devrait être H6:H773, ou H17:H773 (quand les valeurs sont ordonnées en ordre croissant et n'incluent pas les valeurs nulles).
2. Cliquer sur l'onglet **Insertion** et sélectionner l'option **Boîte à moustaches** (voir la [Figure 3.2.62](#)).

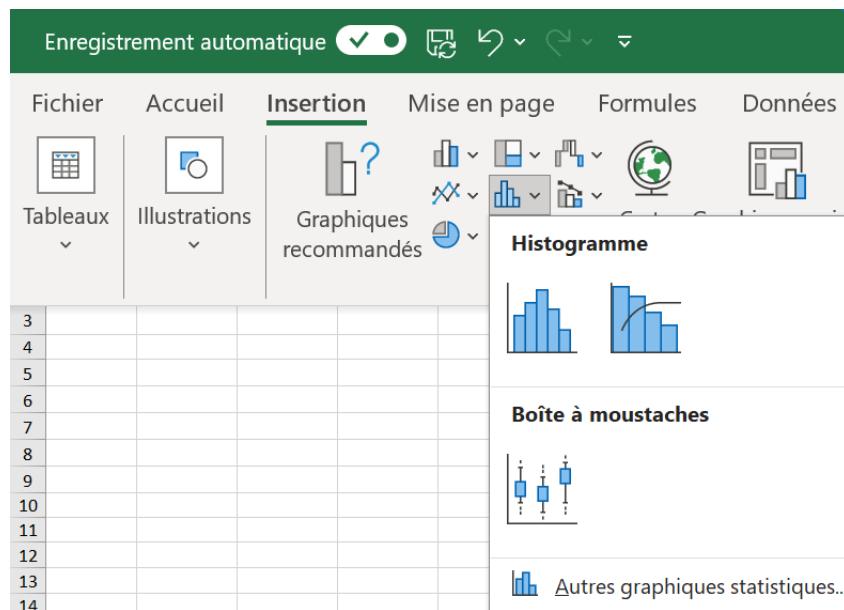


Figure 3.2.62 Insertion d'une boîte à moustaches

3. En cliquant avec le bouton de droite sur le graphique, déplacer ce dernier à la feuille **Étude IMC**, dans un endroit vide.

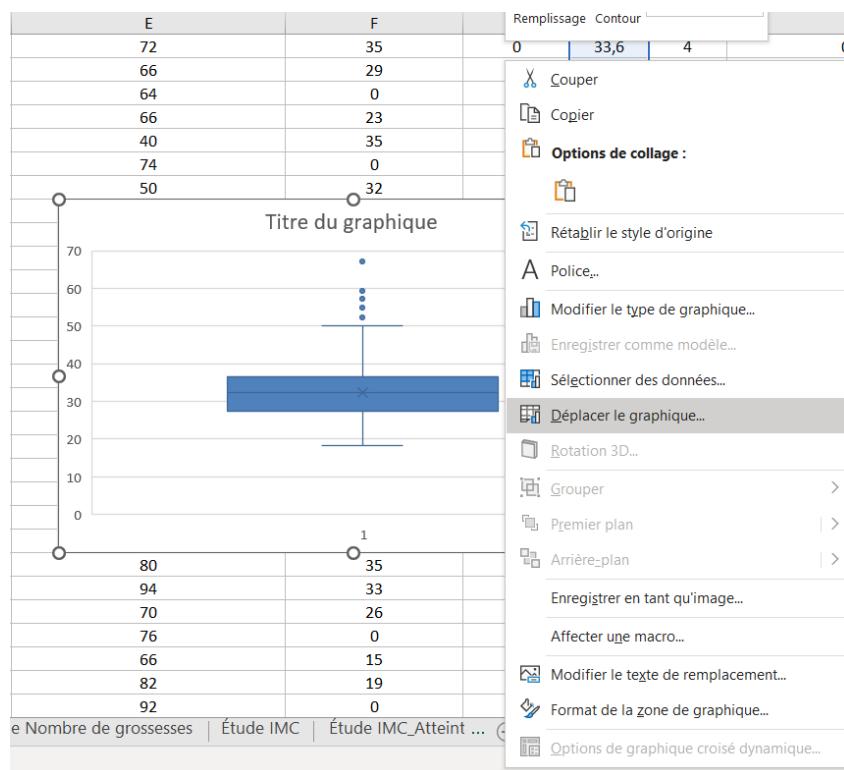


Figure 3.2.63 Déplacer la boîte à moustaches

4. Faire un clic avec le bouton de droite sur les données, cliquer sur **Mettre en forme une série de données**,
5. Faire la mise en forme du graphique. Il devrait ressembler au graphique

de la [Figure 3.2.64](#). À noter que les valeurs numériques ont été ajoutées manuellement pour montrer l'utilité de ce type de graphique.

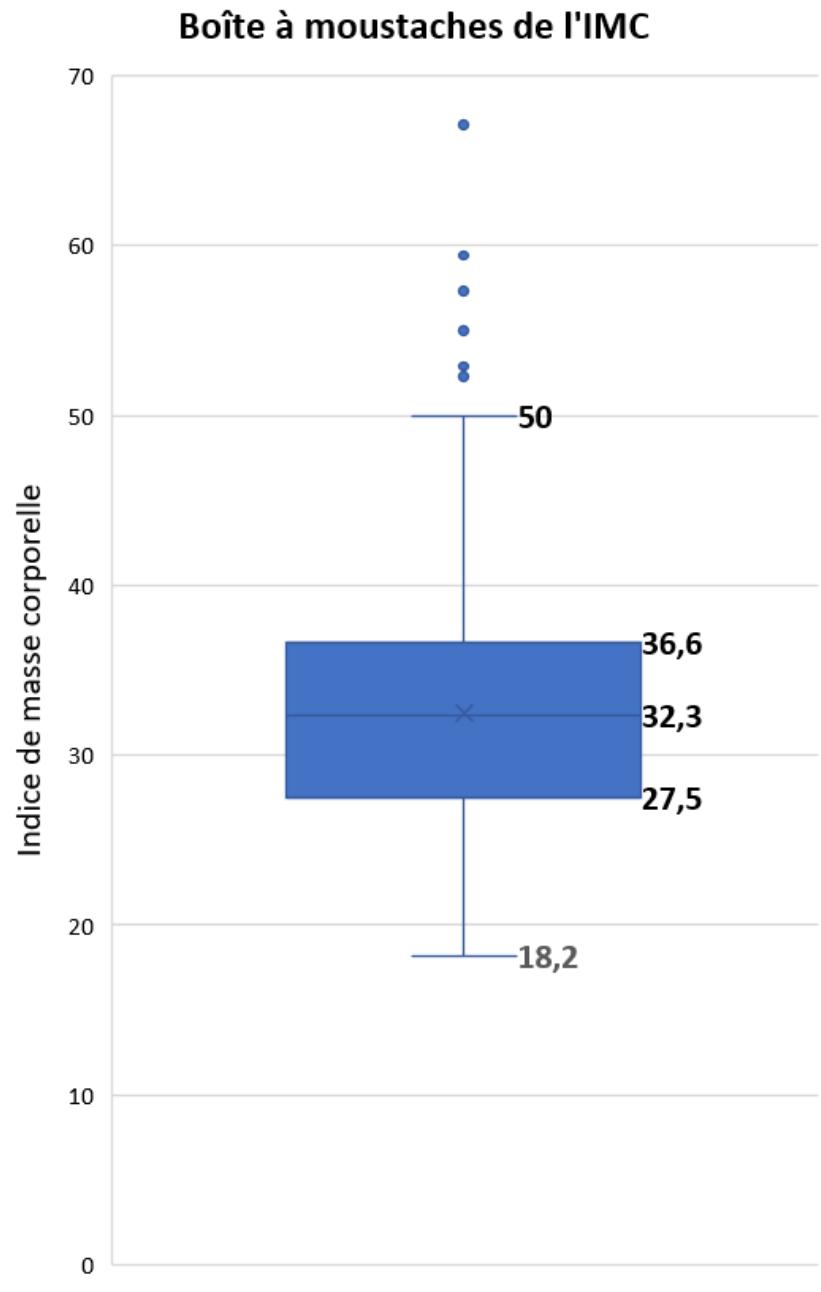


Figure 3.2.64 Boîte à moustaches de la variable **IMC**

Interprétation des mesures descriptives. Le calcul d'une mesure descriptive n'est pas complet sans son interprétation.

1. Dans la feuille de calcul **Étude IMC**, cliquer sur l'onglet **Insertion**.
2. Cliquer sur l'icône **Texte** suivie de l'option **Zone de texte** (voir la [Figure 3.2.65](#)).

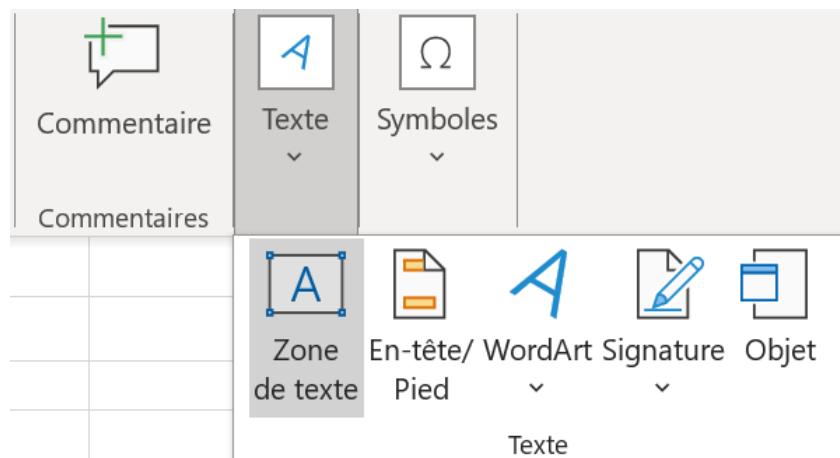


Figure 3.2.65 Insertion d'une zone de texte

3. Cliquer, en maintenant enfoncé le bouton de gauche de la souris, un endroit sous le tableau des mesures statistiques de l'IMC filtrée. Glisser le curseur vers la droite pour créer la zone de texte de taille désirée.
4. Interpréter chaque mesure statistique de l'indice de masse corporelle filtrée.

3.2.2 L'étude simultanée de deux variables dont l'une quantitative continue

Pour réaliser l'étude simultanée d'une variable quantitative continue et d'une variable qualitative, il est nécessaire de construire un tableau de fréquences à double entrée. Dans cette section, on va approfondir l'étude de l'indice de masse corporelle de l'échantillon de femmes d'origine pima, cette fois-ci par atteinte du diabète ou non.

3.2.2.1 Tableau croisé dynamique à double entrée

1. Dans le classeur Excel, ajouter une nouvelle feuille de calcul intitulé **Étude IMC_Atteinte**. Déplacer cette feuille en dernière position si Excel ne le fait pas automatiquement.
2. Sélectionner la cellule **B3** dans cette feuille de calcul.
3. Insérer un tableau croisé dynamique vide tel que vu à la [Sous sous-section 1.2.8.1](#).
4. Pour créer un tableau à double entrée, il faut placer une variable en ligne et une variable en colonne. Glisser et déposer la variable **IMC** dans la zone de saisie **Lignes**. Glisser et déposer la variable **Atteinte** dans la zone de saisie **Colonnes** (voir la [Figure 3.2.66](#)).

Champs de tableau cr... ▼ X

Choisissez les champs à inclure dans le rapport : ⚙️ ▾

Rechercher 🔍

IMC ▲

Obésité

Fonction pedigree du diabète

Atteint ▼

[Plus de tableaux...](#)

Faites glisser les champs dans les zones voulues ci-dessous:

▼ Filtres	☰ Colonnes
☰ Lignes	Σ Valeurs
IMC	Atteint

Figure 3.2.66 Glissement de la variable **IMC** dans la zone de saisie **Lignes** et de la variable **Atteinte** dans la zone **Colonnes**

Les classes de la variable **IMC** sont désormais en ligne, et les deux modalités de la variable **Atteinte** sont en colonnes. Il est possible de constater que l'indice de masse corporelle, variable quantitative continue, a été regroupé en classes avec le même groupement que fait à la [Sous sous-section 3.2.1.4](#) (voir la [Figure 3.2.67](#)).

Étiquettes de colonnes			
Étiquettes de lignes	0	1	Total général
<15			
15-20			
20-25			
25-30			
30-35			
35-40			
40-45			
45-50			
50-55			
55-60			
65-70			
Total général			

Figure 3.2.67 Tableau croisé dynamique vide de la variable ***IMC*** en ligne et la variable ***Atteinte*** en colonnes

5. Glisser et déposer la variable ***IMC*** dans la zone de saisie **Valeurs** (voir la [Figure 3.2.68](#)). On aurait pu choisir la variable ***Atteint***.

Champs de tableau cr... X

Choisissez les champs à inclure dans le rapport : ⚙️ ▾

Rechercher 🔍

IMC
 Obésité
 Fonction pedigree du diabète
 Atteint
[Plus de tableaux...](#)

Faites glisser les champs dans les zones voulues ci-dessous:

▼ Filtres	☰ Colonnes
	Atteint ▼
☰ Lignes	Σ Valeurs
IMC ▼	Nombre de IMC ▼

Figure 3.2.68 Variable **IMC** dans la zone de saisie **Valeurs**

6. Dans la zone de saisie **Valeurs**, cliquer sur la flèche du menu déroulant de la variable, puis sélectionner l'option **Paramètres des champs de valeurs** pour modifier le calcul.

Dans l'onglet **Synthèse des valeurs par**, s'assurer que le type de calcul sélectionné est **Nombre** puisque l'on veut compter le nombre de femmes dans chaque catégorie. Ensuite, cliquer sur l'onglet **Afficher les valeurs**, suivi de la flèche du menu déroulant et sélectionner l'option **% du total de la colonne** (voir la [Figure 3.2.69](#)).

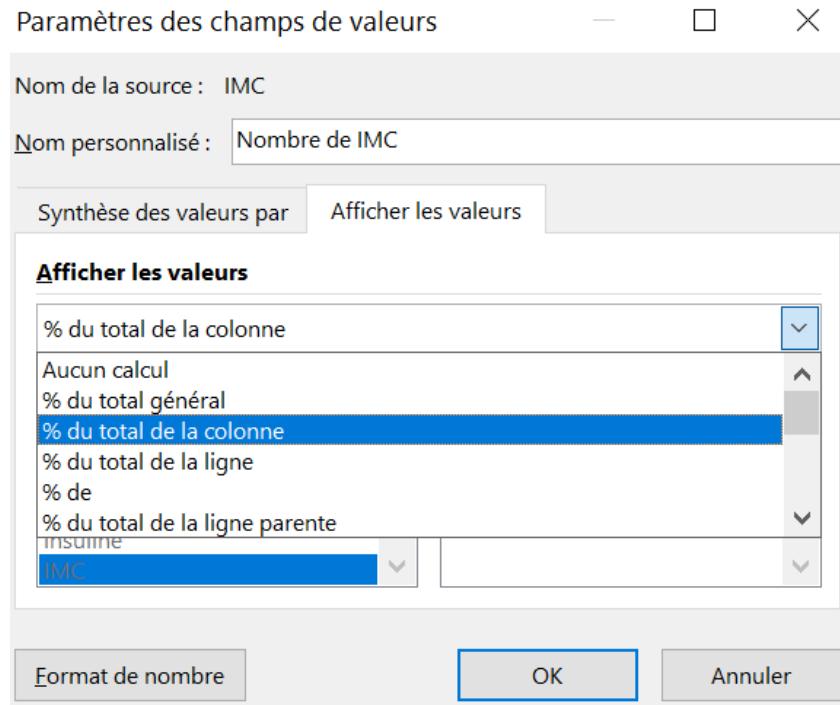


Figure 3.2.69 Sélection du pourcentage du total de la colonne

Le choix d'afficher les valeurs en pourcentage du total de la colonne est déterminé en fonction de l'analyse souhaitée. L'objectif est de mettre en évidence la répartition des femmes d'origine pima, **PAR** présence de diabète, selon l'indice de masse corporelle. On obtient le tableau croisé dynamique de la [Figure 3.2.70](#)

Nombre de IMC	Étiquettes de colonnes		
	0	1	Total général
Étiquettes de lignes			
<15	1,80%	0,75%	1,43%
15-20	2,60%	0,00%	1,69%
20-25	17,20%	2,61%	12,11%
25-30	27,80%	14,93%	23,31%
30-35	24,60%	37,69%	29,17%
35-40	17,40%	23,51%	19,53%
40-45	5,80%	12,31%	8,07%
45-50	2,40%	5,60%	3,52%
50-55	0,20%	1,49%	0,65%
55-60	0,20%	0,75%	0,39%
65-70	0,00%	0,37%	0,13%
Total général	100,00%	100,00%	100,00%

Figure 3.2.70 Les trois colonnes du tableau croisé dynamique final de la répartition des femmes selon l'indice de masse corporelle par présence de diabète

- Il faut filtrer les valeurs nulles encore une fois et créer une classe ouverte regroupant les trois dernières classes puisque celles-ci contiennent peu de données (revoir les étapes de la [Sous sous-section 3.2.1.3](#) et de la

Sous sous-section 3.2.1.4). L'allure du tableau croisé dynamique final est présentée à la Figure 3.2.71.

Nombre de IMC	Étiquettes de colonnes			Total général
	0	1		
Étiquettes de lignes				
15-20	2,65%	0,00%		1,72%
20-25	17,52%	2,63%		12,29%
25-30	28,31%	15,04%		23,65%
30-35	25,05%	37,97%		29,59%
35-40	17,72%	23,68%		19,82%
40-45	5,91%	12,41%		8,19%
45-50	2,44%	6,02%		3,70%
>50	0,41%	2,26%		1,06%
Total général	100,00%	100,00%		100,00%

Figure 3.2.71 Choix de 50 comme valeur de fin dans le groupement et filtrage des valeurs nulles

8. Copier et coller les classes et les pourcentages de chaque catégorie du tableau croisé dynamique dans l'endroit souhaité de la feuille de calcul et faire le formatage du tableau à double entrée. L'allure du tableau croisé dynamique final est présentée à la Figure 3.2.71.

Répartition d'un échantillon de femmes d'origine pima, par présence de diabète, selon l'indice de masse corporelle, Arizona, année inconnue			
IMC (kg/m ²)	Présence de diabète		Total
	Non	Oui	
[15 ; 20[2,65%	0,00%	1,72%
[20 ; 25[17,52%	2,63%	12,29%
[25 ; 30[28,31%	15,04%	23,65%
[30 ; 35[25,05%	37,97%	29,59%
[35 ; 40[17,72%	23,68%	19,82%
[40 ; 45[5,91%	12,41%	8,19%
[45 ; 50[2,44%	6,02%	3,70%
50 et plus	0,41%	2,26%	1,06%
Total	100,00%	100,00%	100,00%

Source : National Institute of Diabetes and Digestive and Kidney Diseases (USA)

Figure 3.2.72 Version définitive du tableau de fréquences de la variable *IMC* avec la variable *Atteinte*

3.2.2.2 Polygone de fréquences

Pour représenter simultanément deux variables, l'une quantitative et l'autre qualitative, le polygone de fréquences est l'option à privilégier. Sur l'axe des abscisses, on place la variable quantitative continue, chaque courbe représente ensuite les différentes modalités de la variable qualitative.

Les étapes qui suivent mènent à la construction du polygone de fréquences de la répartition de l'échantillon de femmes d'origine pima, par présence de diabète, selon l'indice de masse corporelle.

- Sélectionner la plage de données représentant les fréquences relatives de l'indice de masse corporelle par présence de diabète, soit la plage de cellules **C5:D12**. Ne pas sélectionner les classes de l'indice de masse corporelle, ni les titres des colonnes, ni les données de la ligne et de la colonne *Total général*.

	A	B	C	D	E
1					
2					
3		Nombre de IMC	Étiquettes de colc		
4		Étiquettes de lign	0	1	Total général
5	15-20		2,65%	0,00%	1,72%
6	20-25		17,52%	2,63%	12,29%
7	25-30		28,31%	15,04%	23,65%
8	30-35		25,05%	37,97%	29,59%
9	35-40		17,72%	23,68%	19,82%
10	40-45		5,91%	12,41%	8,19%
11	45-50		2,44%	6,02%	3,70%
12	>50		0,41%	2,26%	1,06%
13	Total général		100,00%	100,00%	100,00%

Figure 3.2.73 Sélection des fréquences relatives de l'IMC par présence de diabète

- Copier la plage sélectionnée et coller ces valeurs dans la cellule **H22** (voir la [Figure 3.2.74](#)).
- Ajouter une classe fictive nulle avant et après les fréquences relatives (0% avant les premiers pourcentages dans les cellules **H21** et **I21** et 0% après les derniers pourcentages dans les cellules **H30** et **I30**) (voir la [Figure 3.2.74](#)). Cette étape facilite la mise en forme du polygone de fréquences et améliore son apparence.
- Dans les cellules **H20** et **I20**, ajouter des titres aux colonnes pour chaque modalité, soit *Non* pour la colonne **H** et *Oui* pour la colonne **I** (voir la [Figure 3.2.74](#)).
- Dans un polygone de fréquences, chaque fréquence relative est associée au point milieu de sa classe. Il faut ainsi trouver les points milieux. Dans la cellule **G22**, écrire la valeur du point milieu de la première classe, soit 17,5. Dans la cellule **G23**, écrire la valeur du point milieu de la deuxième classe, soit 22,5 (voir la [Figure 3.2.74](#)).
- Sélectionner la plage de cellules **G22:G23** (voir la [Figure 3.2.74](#)). La plage est encadré d'une bordure verte et un petit carré vert apparaît dans le coin inférieur droit. Approcher le curseur au-dessus du carré vert. Dès qu'une croix noire apparaît, double-cliquer (voir la [Figure 3.2.74](#)).
- Dans la cellule **G21**, taper la valeur milieu de la première classe fictive, soit 12,5 (voir la [Figure 3.2.74](#)).

	2,65%	0,00%
	17,52%	2,63%
	28,31%	15,04%
	25,05%	37,97%
	17,72%	23,68%
	5,91%	12,41%
	2,44%	6,02%
	0,41%	2,26%

Figure 3.2.74 Séquence de collage des fréquences relatives de l'IMC par présence de diabète et insertion des points milieux des classes

8. Sélectionner la plage de cellules **G20:I30** (voir la [Figure 3.2.75](#)).

	Non	Oui
12,5	0%	0%
17,5	2,65%	0,00%
22,5	17,52%	2,63%
27,5	28,31%	15,04%
32,5	25,05%	37,97%
37,5	17,72%	23,68%
42,5	5,91%	12,41%
47,5	2,44%	6,02%
52,5	0,41%	2,26%
57,5	0%	0%

Figure 3.2.75 Sélection des cellules pour créer le polygone de fréquences

9. Cliquer sur l'onglet **Insertion**. Dans le groupe **Graphiques**, cliquer sur l'icône **Insérer un graphique en courbes ou en aires** (voir la [Figure 3.2.76](#)).

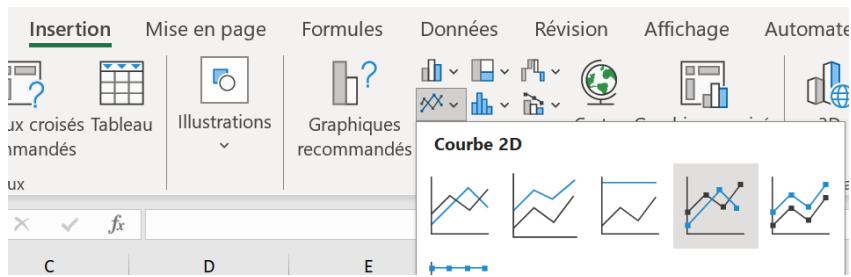


Figure 3.2.76 Sélection de l'icône *Insérer un graphique en courbes ou en aires*

10. Dans la section **Courbe 2D**, sélectionner la quatrième option, soit **Courbes avec marques** (voir la [Figure 3.2.76](#)). Le graphique ci-dessous s'affiche (voir la [Figure 3.2.77](#))

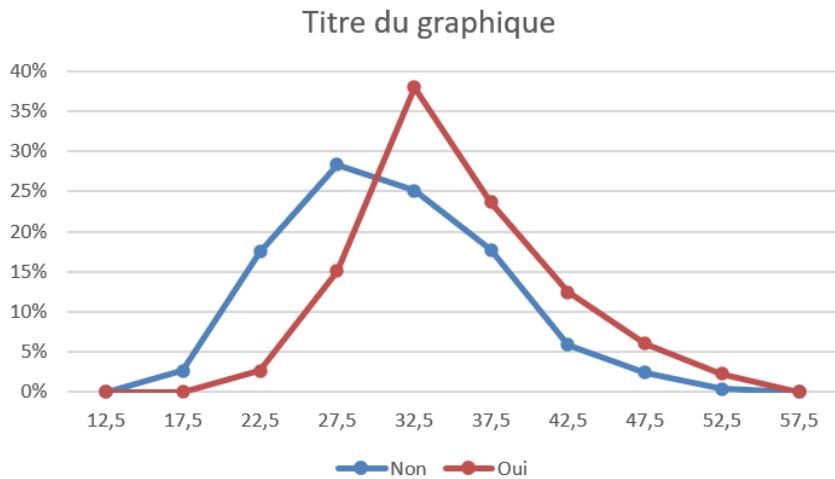


Figure 3.2.77 Graphique créé après la sélection de *insérer une courbe 2D*

Il faut effectuer la mise en forme de ce graphique.

11. Il est possible de déplacer la légende des modalités de la variable qualitative. Cliquer avec le bouton de droite de la souris sur une des modalités de la légende et sélectionner le dernier onglet **Format de la légende** (voir la [Figure 3.2.78](#))

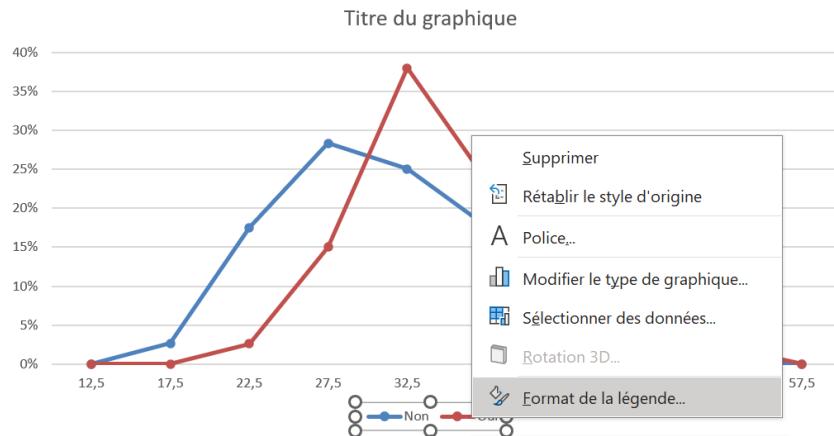


Figure 3.2.78 Sélection de l'onglet *Format de la légende* pour déplacer la légende

12. Une boîte de dialogue s'affiche à droite de la feuille de calcul. Changer l'option de l'emplacement *bas* pour *droite* afin que la légende s'affiche à droite du polygone de fréquences (voir la [Figure 3.2.79](#)).

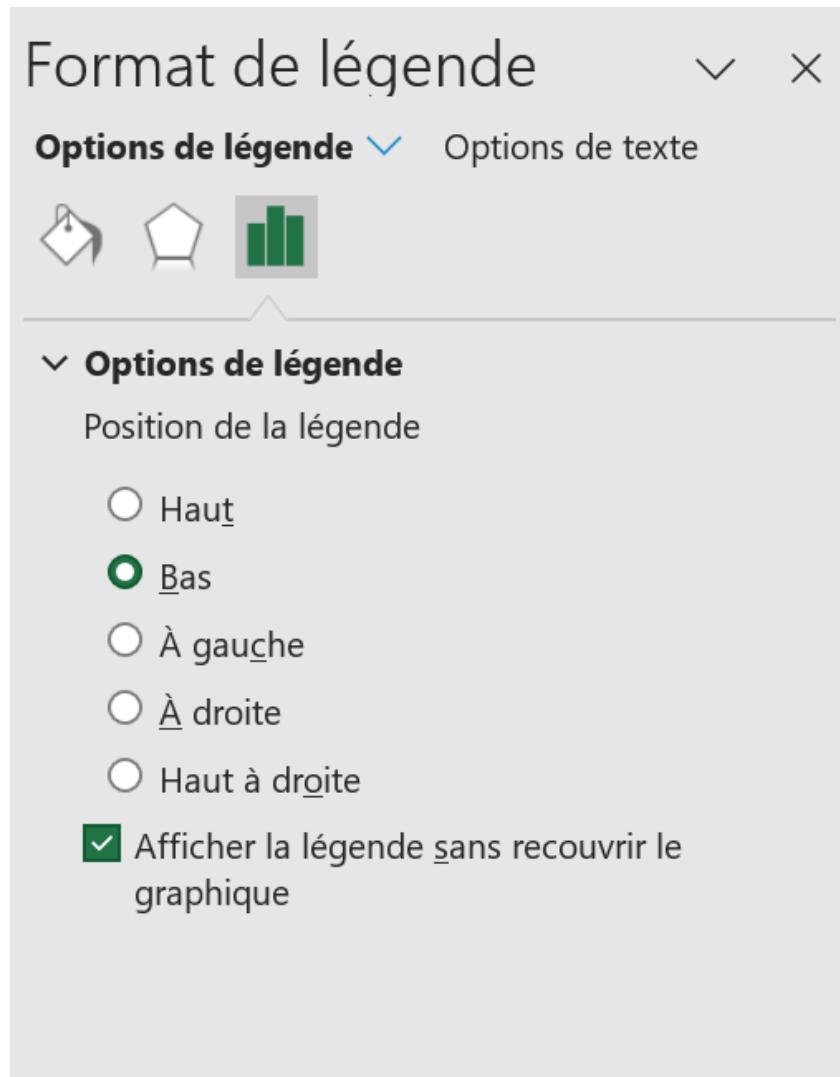


Figure 3.2.79 Changer l'emplacement de la légende

La légende est déplacée à droite (voir la Figure 3.2.80).

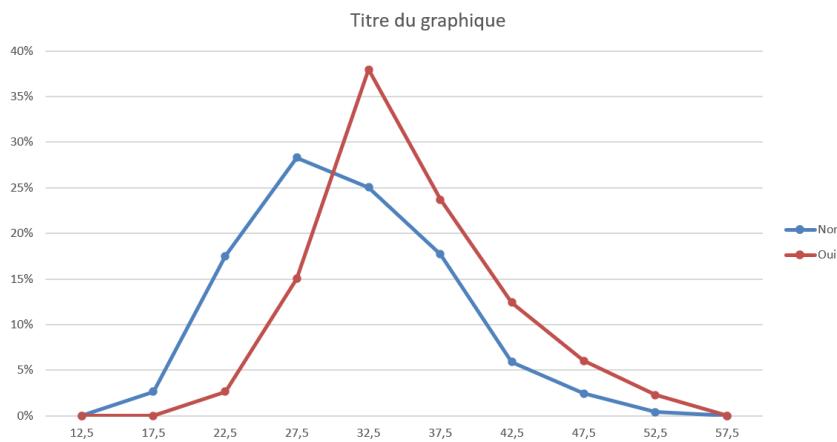


Figure 3.2.80 Position de la légende

13. Ajouter un titre au graphique, un titre aux axes et la source (voir la Figure 3.2.81).

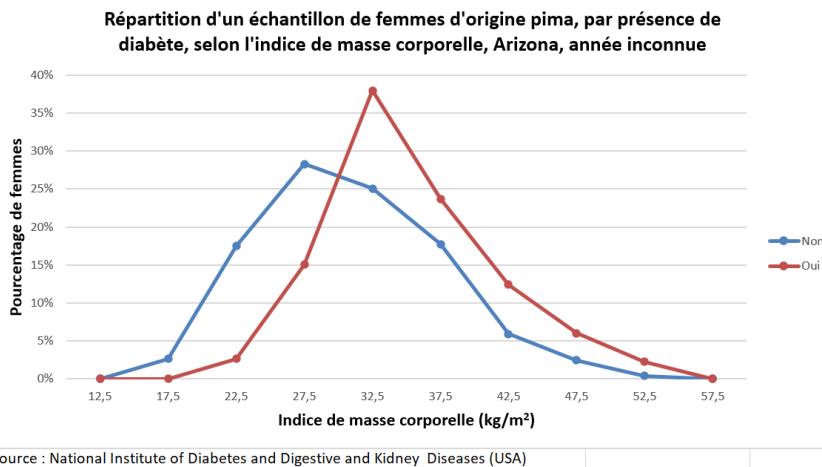


Figure 3.2.81 Ajout du titre du graphique, des titres des axes et de la source

14. Il est possible d'enlever les lignes horizontales grisées pour avoir un fond purement blanc (voir la Figure 3.2.82). Cliquer sur une des lignes horizontales. L'ensemble des lignes est désormais sélectionné. Appuyer sur la touche **suppr** du clavier.

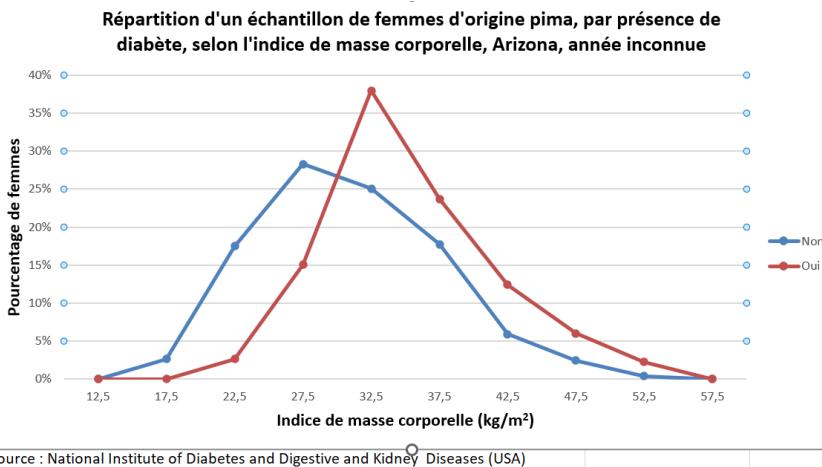


Figure 3.2.82 Enlever les lignes horizontales du quadrillage

15. Il est possible de lisser le polygone, c'est-à-dire ne pas avoir une courbe avec marques. Cliquer avec le bouton de droite de la souris sur le fond blanc du graphique. Sélectionner l'option **Modifier le type de graphique** (voir la Figure 3.2.83).

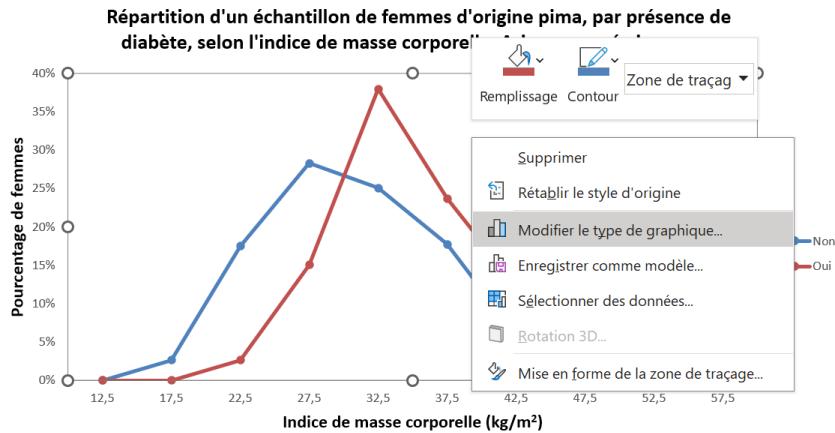


Figure 3.2.83 Sélection de l'option *Modifier le type de graphique*

16. Une boîte de dialogue s'affiche (voir la [Figure 3.2.84](#)). Sélectionner la première option parmi les sept options de courbes 2D, soit *Courbe*, et cliquer sur **OK**.

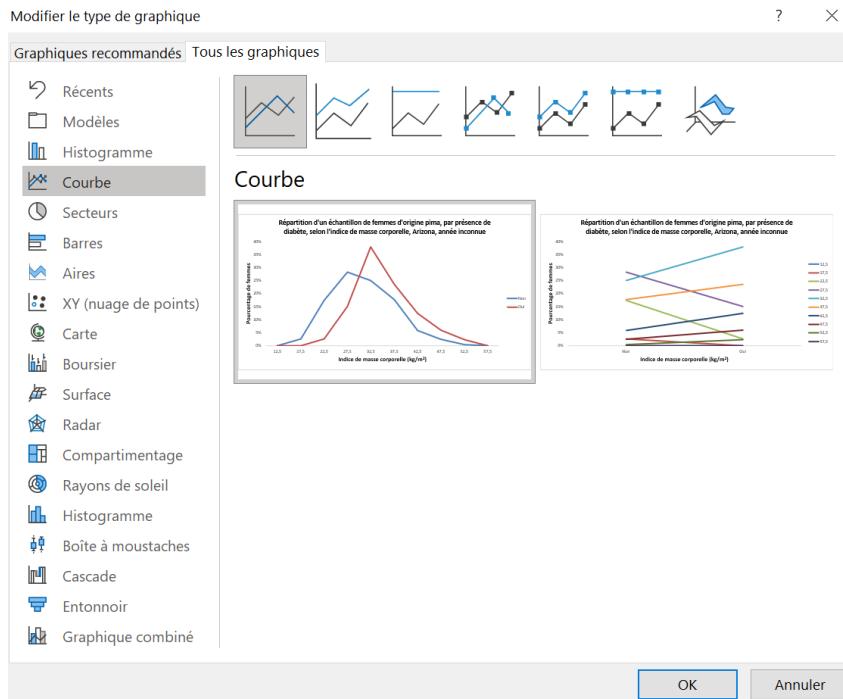
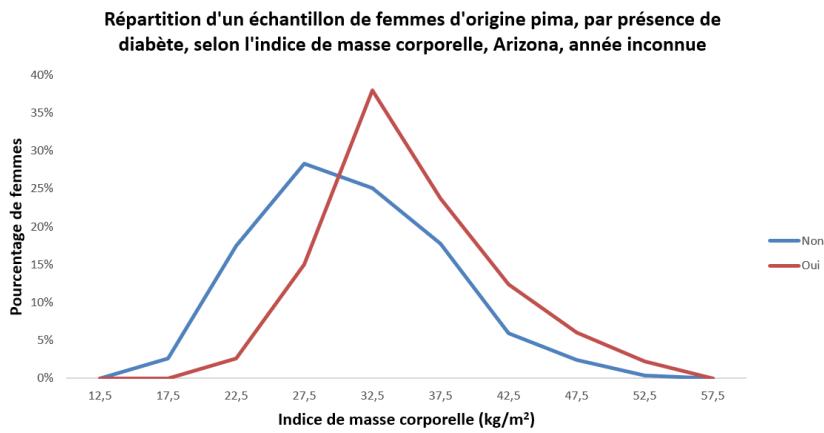


Figure 3.2.84 Sélection de la courbe sans marque

17. La courbe sans marque s'affiche (voir la [Figure 3.2.85](#)).



Source : National Institute of Diabetes and Digestive and Kidney Diseases (USA)

Figure 3.2.85 Polygone de fréquences sans marque

3.2.2.3 Mesures descriptives pour deux variables

Malheureusement, le logiciel Excel n'a pas de formule qui calcule les mesures descriptives du croisement de deux variables. Cependant, les tableaux croisés dynamiques permettent de faire certains calculs comme la moyenne et l'écart type.

1. Dans la feuille **Étude IMC_Atteinte**, sélectionner la cellule **I3**.
2. Insérer un tableau croisé dynamique vide tel que vu à la [Sous sous-section 1.2.8.1](#).
3. Glisser et déposer la variable **Atteint** dans la zone de saisie **Lignes**.
Glisser et déposer la variable **IMC** dans la zone de saisie **Valeurs** (voir la [Figure 3.2.86](#)).

Champs de tableau cr... ▾ ×

Choisissez les champs à inclure dans le rapport : ⚙️

Rechercher 🔍

IMC
 Obésité
 Fonction pedigree du diabète
 Atteint
[Plus de tableaux...](#)

Faites glisser les champs dans les zones voulues ci-dessous:

▼ Filtres	☰ Colonnes
≡ Lignes	Σ Valeurs
Atteint	

Figure 3.2.86 Glissement de la variable **Atteinte** dans la zone de saisie **Lignes** et de la variable **IMC** dans la zone **Valeurs**

4. Dans la zone de saisie **Valeurs**, cliquer sur la flèche du menu déroulant de la variable **IMC**, puis sélectionner l'option **Paramètres des champs de valeurs** pour modifier le calcul (voir la [Figure 3.2.86](#)).

Dans l'onglet **Synthèse des valeurs par**, sélectionner l'option de calcul **Moyenne** et cliquer sur **OK** (voir la [Figure 3.2.87](#)).

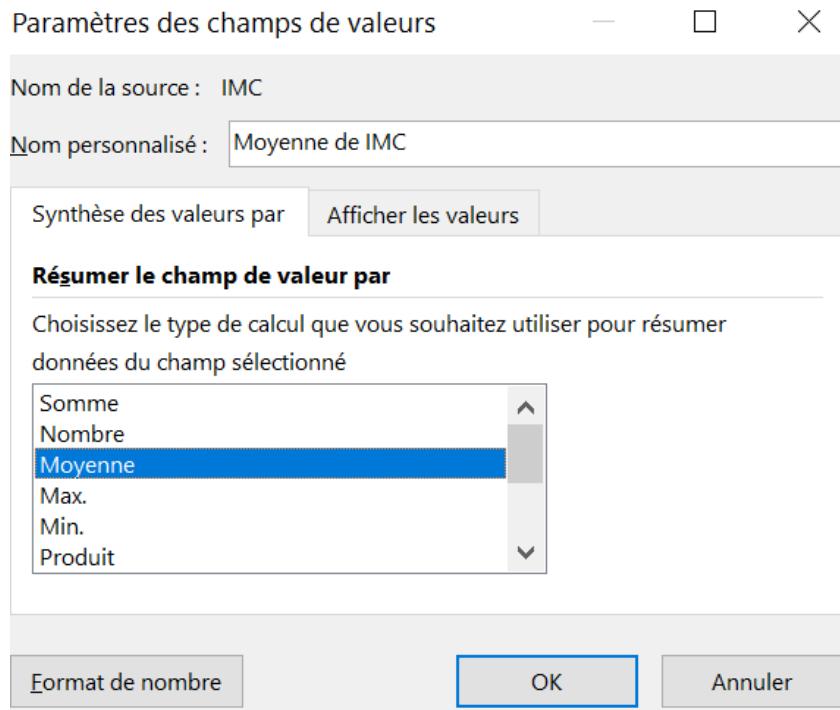


Figure 3.2.87 Sélection de l'option de calcul *Moyenne*

Ce choix de calcul permet d'obtenir la moyenne de l'indice de masse corporelle des femmes qui n'ont pas le diabète et la moyenne de celles qui l'ont.

Étiquettes de lignes	Moyenne de IMC
0	30,3042
1	35,14253731
Total général	31,99257813

Figure 3.2.88 Tableau croisé dynamique de la moyenne de l'indice de masse corporelle selon la présence de diabète

5. Glisser et déposer la variable **IMC** à nouveau dans la zone de saisie **Valeurs**.
6. Dans la zone de saisie **Valeurs**, cliquer sur la flèche du menu déroulant de la variable **IMC** nouvellement ajoutée, puis sélectionner l'option **Paramètres des champs de valeurs** pour modifier le calcul. Dans l'onglet **Synthèse des valeurs par**, sélectionner l'option de calcul **Ecartype**.

Le tableau croisé dynamique résultant est présenté à la [Figure 3.2.89](#).

Étiquettes de lignes	Moyenne de IMC	Écart type de IMC
0	30,3042	7,689855012
1	35,14253731	7,262967242
Total général	31,99257813	7,88416032

Figure 3.2.89 Tableau croisé dynamique de la moyenne et de l'écart type corrigé de l'indice de masse corporelle selon la présence de diabète

Remarque 3.2.90 Mesures descriptives de l'IMC par présence de diabète. Il faut rappeler que dans un tableau croisé dynamique, les calculs sont effectués sur des données non filtrées. Un filtrage est nécessaire pour obtenir les moyennes et les écarts types corrigés sans tenir compte des valeurs nulles de l'indice de masse corporelle.

7. Cliquer sur l'onglet **Insertion**.

Cliquer sur l'option **Filtres** et l'option **Segments** (voir la [Figure 3.2.91](#)). Un **segment** est un outil de filtrage qui permet de filtrer, entre autres, les données d'un tableau croisé dynamique.

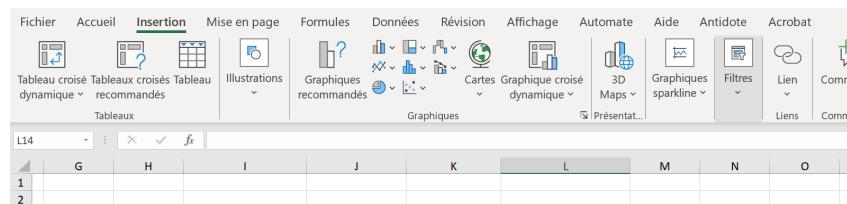


Figure 3.2.91 Introduction d'un segment pour filtrer les données du tableau croisé dynamique

8. Une boîte de dialogue s'affiche permettant de choisir la variable que l'on veut filtrer (voir la [Figure 3.2.92](#)).

Sélectionner la variable **IMC** et cliquer sur **OK**.

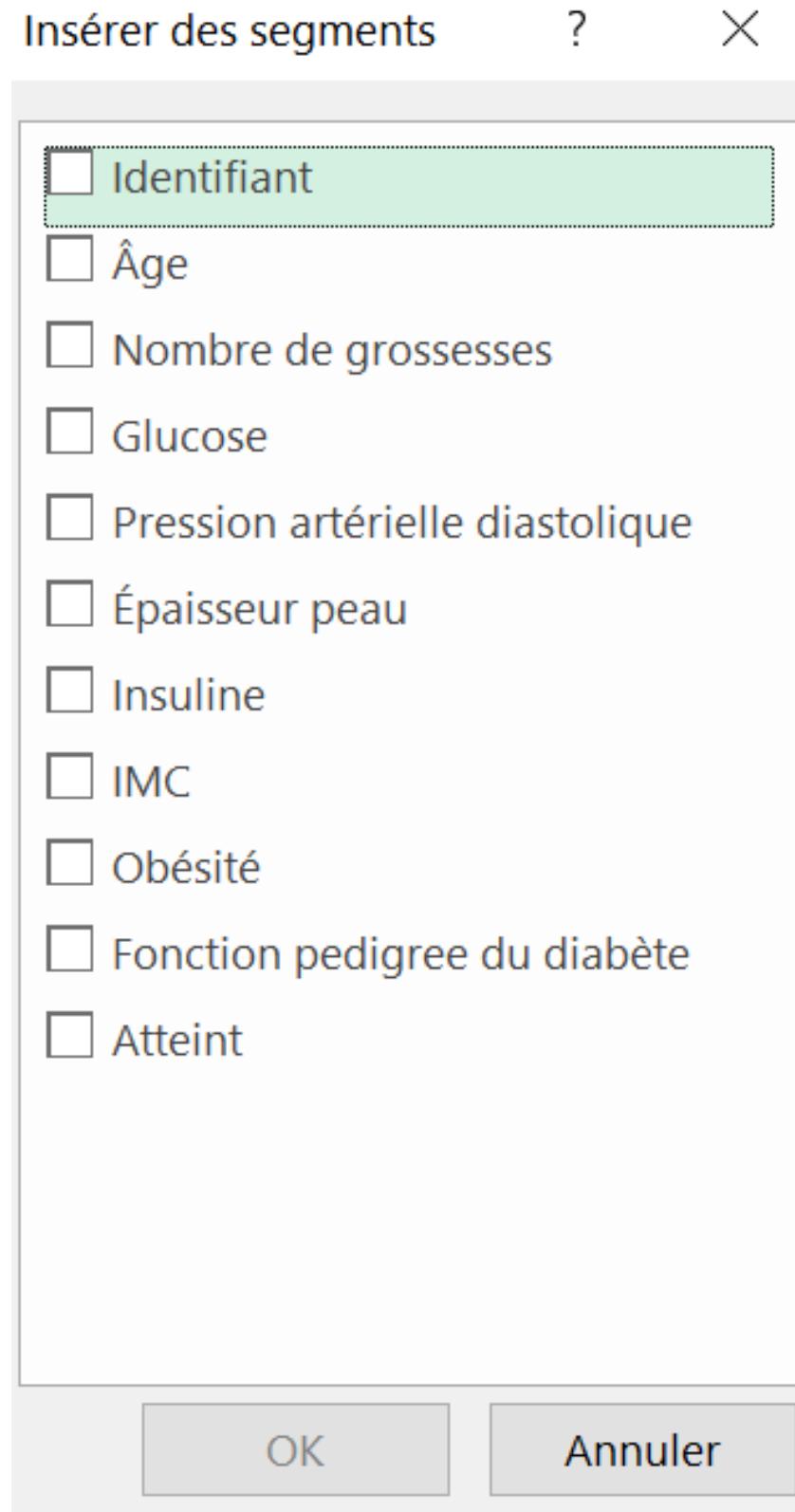


Figure 3.2.92 Sélection de l'option **IMC**

9. Une autre boîte de dialogue s'affiche permettant de filtrer des valeurs de

la variable **IMC** (voir la [Figure 3.2.93](#)).

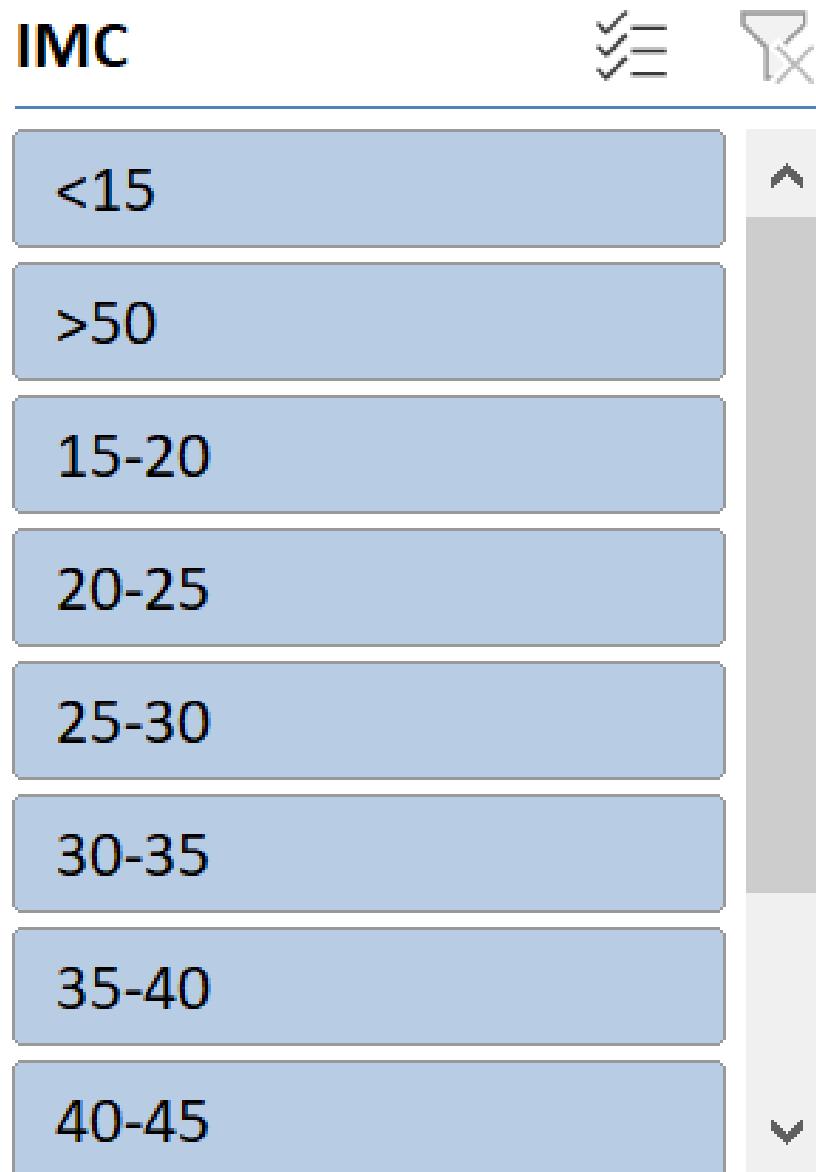


Figure 3.2.93 Boîte de dialogue permettant de filtrer la variable **IMC**

10. Cliquer sur l'icône avec les trois crochets du coin supérieur droit de la boîte de dialogue (voir la [Figure 3.2.94](#)).

Toutes les classes sont en bleues. Cela signifie que toutes les valeurs sont incluses dans les calculs du tableau croisé dynamique. Puisque l'on veut exclure les valeurs inférieures à 15 (soit les valeurs nulles), décliquer l'option < 15 (voir la [Figure 3.2.94](#)).

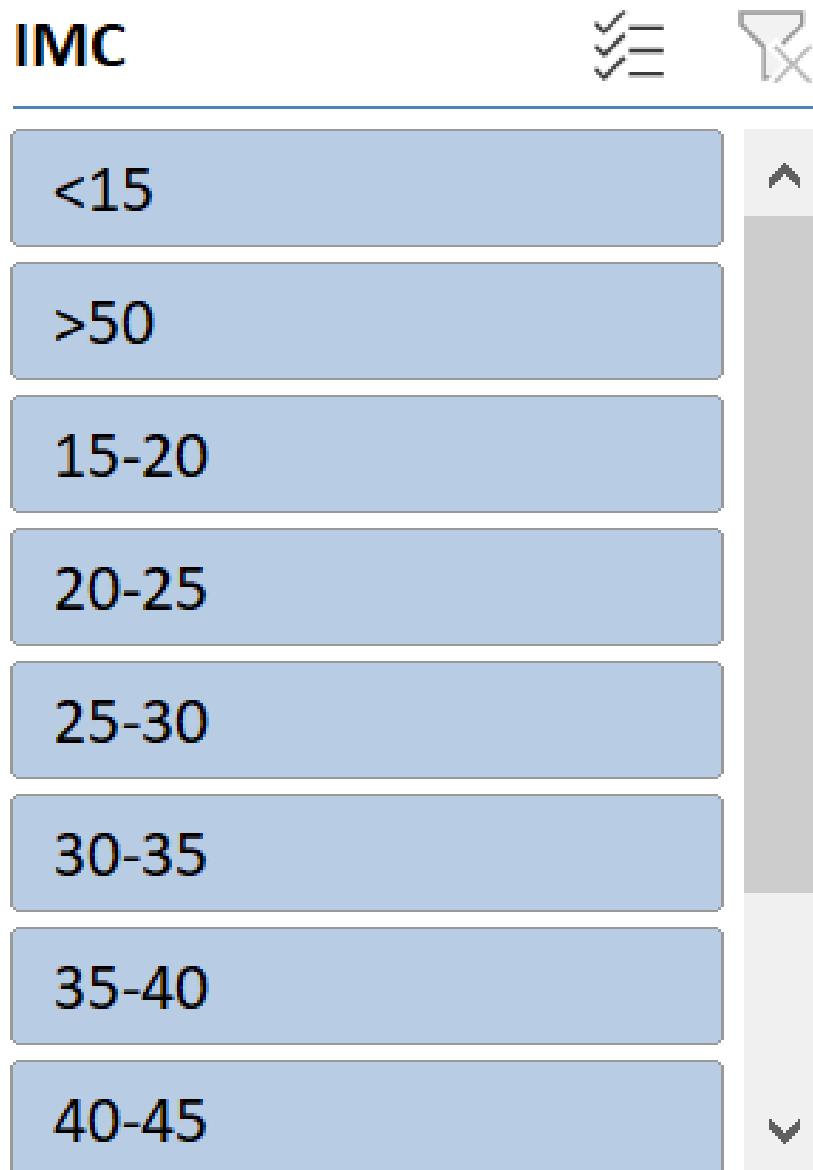


Figure 3.2.94 Filtrer les valeurs de la variable *IMC* pour le calcul des mesures descriptives

Le tableau croisé dynamique résultant est présenté à la [Figure 3.2.95](#).

Étiquettes de lignes	Moyenne de IMC	Écartype de IMC
0	30,85967413	6,560736884
1	35,40676692	6,614982371
Total général	32,45746367	6,924988332

Figure 3.2.95 Tableau croisé dynamique de la moyenne et de l'écart type corrigé de l'indice de masse corporelle filtré selon la présence de diabète

11. Copier et coller le tableau croisé dynamique dans les cellules **I10:K13**(voir la [Figure 2.2.1](#)).

Ajouter une colonne à droite de l'écart type corrigé pour le calcul du coefficient de variation. Dans la cellule **L10**, taper **=K11/J11** (voir la Figure 3.2.96).

Présence de diabète	Moyenne de l'IMC	Écart type de l'IMC	Coefficient de variation
Non	30,85967413	6,560736884	=K11/J11
Oui	35,40676692	6,614982371	
Total	32,45746367	6,924988332	

Figure 3.2.96 Calcul du coefficient de variation de l'indice de masse corporelle selon la présence de diabète

12. Copier la formule de la cellule **L11** dans les cellules **L12** et **L13** en faisant un remplissage automatique (voir la Sous sous-section 1.2.7.1 et Figure 3.2.97 pour le résultat).

Présence de diabète	Moyenne de l'IMC	Écart type de l'IMC	Coefficient de variation
Non	30,85967413	6,560736884	0,212599033
Oui	35,40676692	6,614982371	0,186828195
Total	32,45746367	6,924988332	0,213355806

Figure 3.2.97 Coefficient de variation de l'indice de masse corporelle selon la présence de diabète

13. Formater le tableau pour fin de présentation (voir la Figure 3.2.98).

Présence de diabète	Moyenne de l'IMC	Écart type de l'IMC	Coefficient de variation
Non	30,85967413	6,560736884	0,212599033
Oui	35,40676692	6,614982371	0,186828195
Total	32,45746367	6,924988332	0,213355806

Figure 3.2.98 Formatage du tableau de présentation des mesures descriptives de l'indice de masse corporelle selon la présence de diabète

Le tableau définitif est présenté à la Figure 3.2.99

Mesures descriptives de l'indice de masse corporelle d'un échantillon de femmes d'origine pima, selon la présence de diabète, Arizona, année inconnue			
Présence de diabète	Moyenne de l'IMC	Écart type de l'IMC	Coefficient de variation
Non	30,9	6,6	21,3%
Oui	35,4	6,6	18,7%
Total	32,5	6,9	21,3%

Figure 3.2.99 Tableau définitif des mesures descriptives de l'indice de masse corporelle selon la présence de diabète

3.3 Réflexions

Compte tenu du nombre de variables quantitatives dans la base de données, le travail réalisé lors du laboratoire 3 n'est que le point de départ de l'analyse. Il est essentiel de poursuivre l'étude des mesures diagnostiques des femmes d'origine pima vivant en Arizona afin d'extraire un maximum d'informations concernant leur situation médicale.

Travail à faire après le laboratoire

Objectifs

- Approfondir l'étude de la variable **IMC**.
 - Examiner les autres variables quantitatives continues (**Glucose**, **Pression artérielle diastolique**, **Épaisseur de la peau**, **Insuline**, **Fonction pedigree du diabète**).
 - Examiner une variable quantitative continue et une variable qualitative.
 - Formuler des conclusions.
 - Poser un regard critique sur les données.
 - Formuler des hypothèses de recherche.
1. En deux courtes phrases, résumer la situation du diabète et de l'indice de masse corporelle chez la population de femmes pimas.
 2. Quelles sont les limites des deux études réalisées dans le laboratoire 3? Quelles variables semblent reliées entre elles?
 3. À la [Sous sous-section 3.2.1.6](#), on a calculé des mesures descriptives avec et sans filtrage des valeurs nulles de l'indice de masse corporelle. Expliquer les effets de ne pas exclure les valeurs nulles lors des calculs. Quel impact cela a-t-il sur les mesures descriptives? Cet impact serait-il le même si la base de données contenait 20 unités statistiques au lieu de 768?
 4. Calculer la moyenne de l'indice de masse corporelle à partir du tableau de fréquences des données groupées en classe avec les techniques vues en classe (voir la [Figure 3.2.24](#)). La valeur obtenue est-elle la même que celle obtenue avec la formule Excel (la moyenne avec filtrage)? Sinon, expliquer la différence. Quelle valeur doit être privilégiée?
 5. Une donnée aberrante est une donnée située anormalement loin des autres observations. Une donnée est considérée comme aberrante si elle se situe à 1,5 fois l'écart interquartile ($1,5 \times (Q_3 - Q_1)$) en dessous du premier quartile ou à 1,5 fois l'écart interquartile au-dessus du troisième quartile. Dans la feuille de calcul **Étude IMC**, déterminer toutes les valeurs aberrantes de l'indice de masse corporelle en effectuant les calculs nécessaires dans des cellules vides. À la [Sous sous-section 3.2.1.1](#), quelles valeurs auraient dû être exclues en plus des valeurs nulles? Quel graphique permet de visualiser (d'identifier d'un coup d'œil) les données aberrantes d'une variable?
 6. Devrait-on toujours exclure les valeurs aberrantes lors du traitement des données (graphiques et calculs de mesures)? Donner des exemples où l'exclusion est justifiée et des exemples où elle pourrait masquer la réalité d'un phénomène.
 7. Le choix de l'amplitude et du nombre de classes d'un histogramme peut influencer l'interprétation du graphique. Dans une nouvelle feuille de calcul intitulée **Étude IMC 2**, reproduire l'histogramme de la répartition de l'échantillon des femmes d'origine pima selon l'indice de masse corporelle en regroupant les données en cinq classes. Il est préférable de refaire le tableau croisé dynamique. Il se peut que ceci affecte le regroupement fait à la [Sous sous-section 3.2.1.4](#). Interpréter l'allure du graphique. Comparer l'histogramme reproduit à l'histogramme fait à la [Sous sous-section 3.2.1.5](#).
 8. Choisir une variable quantitative continue (autre que l'IMC) à étudier. Ajouter une feuille de calcul dans le fichier Excel avec un nom approprié reflétant le contenu. Construire le tableau de fréquences ainsi que l'histogramme de la répartition de l'échantillon de femmes d'origine pima selon la variable choisie. Ne pas oublier de filtrer les données aberrantes, d'indiquer le choix de l'amplitude des classes et de regrouper les données. Interpréter le résultat.
 9. Calculer les mesures de tendance centrales, les mesures de dispersion et le troisième quintile de la variable choisie à la question précédente.
 10. Choisir une variable quantitative continue (autre que l'IMC) à étudier simultanément avec la variable qualitative **Atteint**. Ajouter une feuille de calcul dans le fichier Excel avec un nom approprié reflétant le contenu. Construire le tableau de fréquences à double entrée ainsi que le polygone de fréquences de la répartition de l'échantillon de femmes d'origine pima, par présence de diabète, selon la variable quantitative continue choisie. Ne pas oublier de filtrer les données aberrantes, d'indiquer le choix de l'amplitude des classes et de regrouper les données. Interpréter le résultat.
 11. Calculer et interpréter la moyenne, l'écart type corrigé et le coefficient de variation de la variable quantitative continue choisie à la question précédente par présence de diabète. Il faut faire ces calculs à l'aide d'un tableau croisé dynamique tel que présenté à la [Sous sous-section 3.2.2.3](#).
 12. Après avoir étudié certaines variables quantitatives, formuler quelques hypothèses de recherche en lien avec ces variables et les autres de la base de données.

Chapitre 4

Distribution d'échantillonnage

Text before the first section.

4.1 Prélab

BIXI Montréal est un organisme à but non lucratif créé en 2014 par la Ville de Montréal pour gérer le système de vélopartage à Montréal. Le réseau comprend plus de 11 000 vélos (dont 2 600 BIXI électriques) et plus de 900 stations sur le territoire montréalais, ainsi qu'à Laval, Longueuil, Boucherville, Terrebonne, Sainte-Julie, Westmount, Ville de Mont-Royal et Montréal-Est. Beaucoup plus qu'un simple mode de transport, BIXI est aujourd'hui un fabuleux raccourci qui permet de circuler librement dans la ville où et quand on le désire pour aller où l'inspiration et/ou le devoir nous mène.¹

L'organisme BIXI Montréal compile les données des trajets effectués par les utilisateurs depuis le tout début. Ces données sont disponibles pour le grand public.

¹Texte tiré de <https://bixi.com/fr/qui-sommes-nous/>, page consultée le 16 octobre 2024.

Travail à faire avant le cours

Objectifs

- Explorer le comportement de la moyenne échantillonnale pour une petite population.
- Faire l'étude descriptive de la population des trajets en BIXI.

Dans ce laboratoire, on explore le comportement de la moyenne d'un échantillon. Pour cela, on considère le temps de tous les trajets effectués en BIXI sur l'île de Montréal lors du mois d'août 2016, qui fait office de population. Les données proviennent du site web de [Bixi²](#) et ont été nettoyées un peu afin de les structurer et d'éliminer des variables non pertinentes pour l'objet de ce laboratoire. L'organisme a répertorié près de 700 000 trajets lors de cet unique mois. Le but de ce laboratoire est de voir comment la moyenne estimée par un échantillon varie selon l'échantillon qui est sondé.

En passant. Bien que convivial et pratique, le logiciel Excel possède des limites bien réelles. Par exemple il limite l'utilisateur à un peu plus d'un million de lignes et un peu plus de seize mille colonnes. Cela peut être suffisant pour bien des études, mais quand on considère qu'en 2023 il y a plus de dix millions de trajets effectués en Bixi, on comprend que pour les analyser, il faudrait une meilleure solution. Des logiciels statistiques plus avancés ainsi que des langages de programmation spécifiques n'ont souvent pas ces limitations et permettent d'évaluer des jeux de données de grandes taille. Des fonctions complémentaires à Excel sont aussi disponibles, mais ne seront pas discutées dans cet ouvrage.

1. L'un des objectifs du laboratoire est d'observer le comportement de la moyenne d'un échantillon par rapport à la véritable moyenne de la population en fonction de l'échantillon choisi. Afin de bien cerner ce contexte théorique, on propose un exemple simple illustrant certains des concepts qui seront par la suite approfondis.

On considère une population de 5 individus dont la hauteur (en cm) est donnée dans la table suivante:

Table 4.1.1 Grandeur en cm d'une population de 5 individus

Individu	Grandeur (cm)
x_1	185
x_2	175
x_3	155
x_4	165
x_5	195

(a) Calculer la moyenne et l'écart type de cette population.

(b) Il existe 31 échantillons différents de cette population. Par exemple, il y a $\binom{5}{2} = 10$ échantillons de taille 2. Chacun de ces échantillons produit une moyenne et un écart type qui diffère possiblement des valeurs calculées dans la question précédente. On s'intéresse au comportement de la moyenne échantillonnale en ce qui a trait à la taille de l'échantillon choisi. Compléter les données manquantes dans la table suivante. Les deux dernières lignes représentent la moyenne et l'écart type des moyennes des échantillons pour une taille donnée.

Faire les calculs à l'aide d'un logiciel ou d'une calculatrice.

²bixi.com/fr/donnees-ouvertes/

Table 4.1.2 Moyennes échantillonnales de tous les échantillons possibles

$n = 1$	$n = 2$	$n = 3$
Échantillons	Échantillons	Échantillons
\bar{x}	\bar{x}	\bar{x}
$\{x_1\}$	$\{x_1, x_2\}$	$\{x_1, x_2, x_3\}$
185	180	171, 67
$\{x_2\}$	$\{x_1, x_3\}$	$\{x_1, x_2, x_4\}$
175	170	175
$\{x_3\}$	$\{x_1, x_4\}$	$\{x_1, x_2, x_5\}$
155	175	185
$\{x_4\}$	$\{x_1, x_5\}$	$\{x_1, x_3, x_4\}$
165	190	
$\{x_5\}$	$\{x_2, x_3\}$	$\{x_1, x_3, x_5\}$
	$\{x_2, x_4\}$	$\{x_1, x_4, x_5\}$
	$\{x_2, x_5\}$	$\{x_2, x_3, x_4\}$
	$\{x_3, x_4\}$	$\{x_2, x_3, x_5\}$
	$\{x_3, x_5\}$	$\{x_2, x_4, x_5\}$
	$\{x_4, x_5\}$	$\{x_3, x_4, x_5\}$
Moyenne $\mu_{\bar{X}}$	Moyenne $\mu_{\bar{X}}$	Moyenne $\mu_{\bar{X}}$
Écart type $\sigma_{\bar{X}}$	Écart type $\sigma_{\bar{X}}$	Écart type $\sigma_{\bar{X}}$

Table 4.1.3 Moyennes échantillonnales de tous les échantillons possibles

$n = 4$	$n = 5$
Échantillons	Échantillons
\bar{x}	\bar{x}
$\{x_1, x_2, x_3, x_4\}$	$\{x_1, x_2, x_3, x_4, x_5\}$
170	175
$\{x_1, x_2, x_3, x_5\}$	
177, 5	
$\{x_1, x_2, x_4, x_5\}$	
$\{x_1, x_3, x_4, x_5\}$	
$\{x_2, x_3, x_4, x_5\}$	
Moyenne $\mu_{\bar{X}}$	Moyenne $\mu_{\bar{X}}$
Écart type $\sigma_{\bar{X}}$	Écart type $\sigma_{\bar{X}}$

(c) Formuler des observations sur le comportement de \bar{X} par rapport à la taille de l'échantillon.

2. On regarde les notes finales de tous les élèves qui suivent un cours de probabilités et statistique. On considère aussi la moyenne finale de tous les groupes de ce même cours. Laquelle de ces deux variables aléatoires devrait posséder la plus grande variabilité? Comparer la réponse à cette question avec les observations de la question 1.

3. Télécharger et ouvrir le fichier **Bixi_août_2016.xlsx** disponible à l'adresse suivante³. Ce fichier contient une feuille de calcul appelée «Trajets en BIXI août 2016», laquelle contient quatre variables et 688174 données, représentant l'ensemble de tous les trajets effectués en Bixi à Montréal durant le mois d'août 2016. On considère ces trajets comme la population à l'étude.
- (a) Dans la feuille de calcul **Trajets en BIXI août 2016**, nommer le tableau des données ainsi que ses colonnes, tel que montré à la [Sous-section 1.2.3](#). Créer ensuite une deuxième feuille de calculs intitulée «Échantillonnage_canevas». Déplacer cette feuille de calcul à la gauche de la première. Cela évitera plus tard de la sélectionner par erreur.
- (b) Faire l'étude descriptive de la population telle que décrite dans la [Sous sous-section 3.2.1.6](#). À la vue de ces mesures, est-ce qu'on peut qualifier la distribution d'à peu près normale? Expliquer.
4. Faire une lecture sommaire des sections [Adresse](#), [Indirect](#) et [Substitue](#) de l'annexe.

³github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Bixi_ao%C3%BBt_2016.xlsx?raw=true

4.2 Laboratoire

Il est rare que l'on ait accès à la population. Lorsque l'on calcule des mesures statistiques comme la moyenne, la médiane ou l'écart type à partir d'un échantillon, il est essentiel de comprendre que ces valeurs seraient fort probablement différentes si l'échantillon était lui-même différent. En fait, chacune des ces mesures statistiques est une variable aléatoire dont la distribution dépend de celle de la population et de ses paramètres.

Puisque l'on a accès aux données de la population, il est possible d'observer les variations de la moyenne échantillonnale, notée \bar{X} , selon l'échantillon choisi et de comparer ces variations à la véritable moyenne, que l'on peut calculer. Dans la pratique, ces données ne seraient probablement pas disponibles. Il faudrait donc estimer les paramètres de la distribution, comme la moyenne, à l'aide d'un échantillon. Toutefois, il faut être conscient des limites de ces estimations et comprendre comment se comporte la variable aléatoire \bar{X} pour en tirer de bonnes conclusions.

4.2.1 Préparation du fichier

L'échantillonnage est le procédé par lequel on sélectionne les unités statistiques d'une population afin qu'ils fassent partie d'un échantillon. Il existe plusieurs techniques d'échantillonnage, séparées en deux catégories: aléatoires et non aléatoire. Dans un monde idéal, il faudrait toujours que l'échantillonnage soit fait de manière aléatoire, mais, pour des raisons parfois d'impratibilité, ce n'est pas toujours possible. Comme l'étude des techniques d'échantillonnage n'est pas l'objectif de ce laboratoire, on procédera par l'une des méthodes les plus simples, soit l'échantillonnage aléatoire simple, avec remise. Cela signifie que chaque individu de la population a la même chance d'être choisi pour faire partie de l'échantillon et ce, pour chaque membre de l'échantillon (c'est ce que le «avec remise» signifie). Comme la taille de la population est grande comparativement à la taille des échantillons qui seront considérés, cette technique n'a pas tendance à donner des résultats trop différents de celle où l'on aurait pris les échantillons sans remise.

Pour comprendre le comportement de \bar{X} , on s'intéresse à deux paramètres, soit la taille des échantillons, noté n et le nombre d'échantillons tirés, noté Nb_E , permettant d'observer l'allure de la distribution.

On voudra étudier comment varie la distribution de \bar{X} en la regardant sous ces deux angles. Dans le premier cas, on fixe une taille d'échantillon $n = 10$ et l'on regarde la distribution de \bar{X} en considérant plusieurs ensembles d'échantillons de taille variée ($Nb_E \in \{5, 10, 100, 1000, 10000\}$). Ensuite, on étudie l'effet de l'augmentation de la taille des échantillons $n \in \{10, 30, 100, 1000, 10000\}$ sur la distribution lorsque le nombre d'échantillons tirés est fixé à $Nb_E = 100$.

Dans le coin supérieur gauche de la feuille de calculs *Échantillonnage_canevas*, reproduire la table ci-dessous.

	A	B	C	D	E
1	n	début	fin		
2	nombre d'échantillons	5			
3	taille d'un échantillon	10			
4					
5					

Figure 4.2.1 Table de construction de l'échantillonnage

Sous la table, par exemple dans la cellule A5, inscrire «Taille de la population». À droite de cette case, entrer la formule =NB(Durée_s). Cette formule devrait retourner 688174, soit la taille de la population.

Quelque part à la droite de la table, insérer un nouveau tableau en cliquant sur le bouton approprié du ruban sous l'onglet *insertion*. Le tableau doit contenir une colonne et moins de dix lignes. Renommer le nom de la colonne «Échantillonage1». Sous l'onglet *Formules*, cliquer sur le bouton *Gestionnaire de noms* et renommer le tableau créé «Échantillonnages».

Dans la première ligne du tableau, entrer la formule =ALEA.ENTRE.BORNES(2;\$B\$5+1). Si B5 n'est pas la cellule qui contient la taille de la population, remplacer par la cellule appropriée. Cette commande devrait remplir le tableau de nombres aléatoires. Ces nombres correspondent à des numéros de lignes du tableau *Données*. Ils indiquent quels trajets feront partie des échantillons.

Le tableau *Échantillonnages* va contenir autant de colonnes que le nombre d'échantillons que l'on souhaite avoir et autant de ligne que la taille de ces échantillons. Chaque colonne de ce tableau représentera ainsi un échantillon de la population. Puisque l'on veut étudier la variation de \bar{X} selon le nombre d'échantillons et la taille, il faut que les dimensions de ce tableau puissent s'ajuster. C'est ici que la table construite dans la plage A1:D3 va être utile. À l'aide de fonctions Excel, dont le fonctionnement est expliqué dans l'[Appendice E](#), on pourra calculer la plage que devra occuper le tableau.

Dans la cellule C2, on veut inscrire la première colonne du tableau. Entrer la formule =@COLONNE(Échantillonnages). Ceci devrait donner sous la forme d'un nombre la première colonne du tableau. Dans la cellule D2, on veut connaître la dernière colonne du tableau, en fonction du nombre qui est inscrit dans la cellule B2 (par défaut, 5, mais ce sera modifié plus tard). On veut aussi connaître cette colonne non pas sous sa forme numérique, mais sous son nom lettré. Entrer la formule =SUBSTITUE(ADRESSE(1;C2+B2-1;4);"1";""). Cela devrait donner la lettre correspondant à la cinquième colonne d'un tableau qui en contiendrait 5 si l'on commençait la où se trouve le tableau *Échantillonnages*.

A	B	C	D	E	F	G	H
1	n	début	fin				
2	nombre d'échantillons	5	7 K			Échantillonage1	
3	taille d'un échantillon	10	3	12		670327	
4						401166	
5	Taille de la population	688174					
6							
7							

Figure 4.2.2 La feuille Échantillonnage

On répète la procédure pour déterminer la dernière ligne du tableau. Dans la cellule C3, entrer la formule =@LIGNE(Échantillonages). Dans la cellule D3, entrer la formule =C3+B3-1. À ce stade-ci, la feuille de calculs devrait ressembler à celle de l'image ci-dessous, avec des valeurs différentes dans le tableau **Échantillonages**.

Avant de construire le premier ensemble d'échantillon, on va sauvegarder le travail effectué jusqu'à maintenant sous la forme d'un canevas que l'on pourra réutiliser pour les différentes études ci-dessous et dans le post laboratoire. Effectuer un clic-droit sur le nom de la feuille de calcul et cliquer sur «Protéger la feuille... ». Une fenêtre devrait apparaître. N'entrer pas de mot de passe et cliquer simplement sur «OK». Ceci empêche maintenant toute modification sur cette feuille de calculs.

Toujours à l'aide d'un clic droit sur le nom de la feuille, cliquer sur «Déplacer ou copier». Dans la fenêtre qui apparaît, cocher la case «Créer une copie», sélectionner le positionnement «(en dernier)» et appuyer sur «OK». Excel crée ainsi une copie du canevas qui constituera la feuille de calculs de la première étude. On renomme cette feuille de calcul «Échantillonage1». Cette nouvelle feuille de calculs est aussi verrouillée. Il est possible de la déverrouiller en effectuant un clic droit sur son nom et en cliquant sur «Ôter la protection de la feuille ...» permettant ainsi sa modification. Également, le tableau de cette copie s'est vu attribuer un nom par Excel afin de le distinguer du tableau de la feuille originale. Ce n'est pas un problème en soi, mais il faudra être conscient du nom de ce tableau pour la suite ou encore le renommer. Par simplicité, on continue d'y faire référence sous le nom **Échantillonages** dans ce qui suit.

4.2.2 Crédit du premier échantillon

Le premier cas de figure étudié est lorsque la taille de l'échantillon est égale à 10. On commence par tirer cinq échantillons au hasard dans la population. Comme ce sont ces paramètres qui avaient été placés dans la table de construction, on n'a pas à modifier ces valeurs cette fois. Voici la procédure pour sélectionner les échantillons.

1. Cliquer sur l'une des cellules du tableau **Échantillonages**.
2. Sous l'onglet «Création de tableau», cliquer sur le bouton du ruban appelé «Redimensionner le tableau».
3. Dans la fenêtre apparaissant, modifier la fin de la plage afin qu'elle corresponde avec ce qui est inscrit aux cellules D2 et D3. Appuyer sur «OK». Le nouveau tableau devrait apparaître, sans valeurs dans les colonnes à droites.
4. Cliquer sur une cellule de nouveau tableau et effectuer la combinaison **Ctrl**+**A** afin de sélectionner toutes les cellules du tableau, puis effectuer la combinaison **Ctrl**+**D** afin de propager les formules vers la droite.

La figure ci-dessous illustre ces étapes.

The screenshot shows a Microsoft Excel spreadsheet with a data table and a 'Redimensionner le tableau' (Resize Table) dialog box.

Data Table:

	A	B	C	D	E	F	G	H	I	J	K	L
1	n	début	fin									
2	nombre d'échantillons	5	7	K								
3	taille d'un échantillon	10	3	12								
4							Échantillonnage1					
5	Taille de la population	688174					134729					
6							228474					
7												
8												
9												
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												

Dialog Box:

Redimensionner le tableau
? OK Annuler

Sélectionnez la nouvelle plage de données pour votre tableau:
\$G\$2:\$G\$4

Remarque : les en-têtes doivent rester sur la même ligne et la plage de données générée doit chevaucher la plage du tableau d'origine.

Figure 4.2.3 Sélection des unités statistiques des échantillons

Les nombres dans la tableau représentent les lignes du tableau **Données** qui ont été sélectionnées pour faire partie des échantillons. Il faut maintenant aller lire dans les données la valeur de la variable à l'étude. Créer une copie de la feuille **Échantillonnage1** et nommer cette feuille «Échantillons1».

1. Dans cette nouvelle copie, inscrire sous la table de construction dans la colonne A le texte «Colonne de la variable étudiée».
2. Entrer `=@COLONNE(Durée__s)` dans la cellule de adjacente.
3. Dans le gestionnaire de noms, vérifier et changer le nom du tableau de cette nouvelle feuille pour «Échantillons1».
4. Renommer le titre de la première colonne de ce tableau «Échantillon».
5. Toujours avec le titre de la première colonne sélectionné, effectuer la combinaison **[Ctrl]+[Shift]+[→]** afin de sélectionner tous les titres de colonnes.
6. Effectuer la combinaison **[Ctrl]+[D]** pour propager à droite le titre **Échantillon**.
7. Ajouter un «1» au titre de la première colonne.
8. Dans n'importe quelle cellule de la première colonne du tableau **Échantillonnages**, remplacer la formule en place par `=@INDIRECT(ADRESSE(@Échantillonnages1;B7;;;"Trajet en BIXI août 2016"))`, en prenant le soin de remplacer le nom «Échantillonnages1» par le nom qu'Excel a donné au tableau de la feuille **Échantillonnage1**.
9. Sélectionner tout le tableau **Échantillons1** et propager à droite la nouvelle formule. Le tableau contient maintenant la durée en secondes des échantillons. La [Figure 4.2.6](#) illustre à quoi devrait ressembler la feuille de calculs **Échantillons1** jusqu'à présent.
10. On veut maintenant ajouter les moyennes sous le tableau. Pour cela:
 - (a) Sélectionner le tableau **Échantillons1**;
 - (b) Sous l'onglet **Création de tableau**, cocher la case **Ligne Total**. Il est aussi possible de cliquer sur l'icône dans le coin inférieur droit qui apparaît une fois le tableau sélectionné;

- (c) Dans les deux cas, cliquer sur cet icône, puis sous **Totaux**, cliquer sur **Moyenne**. La figure ci-dessous illustre ces étapes.

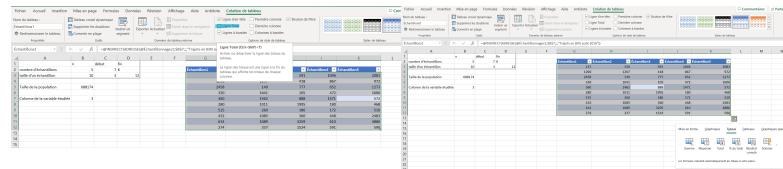


Figure 4.2.4 Ajout de la ligne des moyennes

- (d) Il est aussi possible d'utiliser la formule Moyenne sous la première colonne et de propage vers la droite par la suite. En fait, ceci est préférable s'il y a beaucoup de valeurs dans le tableau, car l'utilitaire ci-dessus nécessite davantage de ressources de la part de l'ordinateur.

	A	B	C	D	E	F	G	H	I	J	K	L
1	nombre d'échantillons	n	début	fin			Echantillon1	Echantillon2	Echantillon3	Echantillon4	Echantillon5	
2	5		7	K			708	411	515	1319	543	
3	taille d'un échantillon	10	3	12			1531	262	693	159	823	
4							1760	1179	283	131	1051	
5	Taille de la population	688174					1057	209	739	1314	483	
6							233	358	1029	870	371	
7	Colonne de la variable étudiée	3					765	227	813	186	834	
8							872	358	1626	394	316	
9							833	1465	547	714	693	
10							988	609	614	1001	951	
11							1180	684	1726	947	1395	
12												
13												
14												
15												

Figure 4.2.5 Les échantillons avec la durée en secondes

La figure ci-dessous montre le résultat final du tableau des échantillons, avec la ligne des moyennes ajoutée.

	A	B	C	D	E	F	G	H	I	J	K	L
1	nombre d'échantillons	n	début	fin		Echantillon1	Echantillon2	Echantillon3	Echantillon4	Echantillon5		
2	5		7	K		3210	701	1125	1837	711		
3	taille d'un échantillon	10	3	12		329	346	400	547	636		
4						1437	312	813	515	1418		
5	Taille de la population	688174				457	424	1626	915	938		
6						2247	1466	346	461	675		
7	Colonne de la variable étudiée	3				395	490	411	920	4217		
8						2171	1233	315	192	292		
9						532	186	207	1384	1947		
10						1650	317	1354	321	598		
11						485	894	601	748	167		
12							1291,3	636,9	708,5	856,9	1189,9	
13												
14												
15												

Figure 4.2.6 La feuille Échantillons1

4.2.3 Analyse de l'échantillon

Il est maintenant temps de procéder à l'analyse des données. Le lecteur aura sans doute remarqué qu'Excel recalcule les échantillons chaque fois qu'une nouvelle information est entrée dans une cellule, que celle-ci soit reliée au tableau des échantillons ou non. Toutefois, lorsque vient le temps d'analyser les échantillons, il est préférable que les données ne changent plus. Bien qu'il soit possible d'automatiser plusieurs étapes, certaines décisions doivent être prises en regard avec les données spécifiques à un ensemble d'échantillons. Par exemple, on peut penser à l'amplitude des classes dans l'histogramme, qui pourrait varier d'un tirage à un autre.

Copier les moyennes du tableau à un autre endroit dans la feuille **Échantillons1**, en s'assurant de faire un collage spécial (ToDo référence annexe?) avec les valeurs seulement. Incrire à gauche de cette ligne de moyennes le titre « Moyennes pour analyse ». Sous cette cellule, entrer la commande **=TRANSPOSE(plage)**, où **plage** est la plage de cellules où se retrouvent les moyennes pour analyse. Ceci va convertir le format horizontal en format vertical, nécessaire pour introduire un tableau croisé dynamique. Nommer cette plage verticale dans le gestionnaire de noms.

	Échantillon1	Échantillon2	Échantillon3	Échantillon4	Échantillon5	
1	1594	1147	1611	464	1213	
2	505	504	1276	590	1838	
3	1325	336	878	940	671	
4	1020	1894	731	479	1970	
5	730	925	2011	708	419	
6	474	1421	210	294	1142	
7	1076	1172	310	681	1517	
8	786	474	832	2631	1831	
9	715	189	980	2396	397	
10	362	951	738	522	752	
11	852,9	901,3	957,7	970,5	1175	
12						
13						
14						
15						
16						
17						
18	Moyennes pour analyse	1281,1	1279,8	589,5	865	922,5
19		1281,1				
20		1279,8				
21		589,5				
22		865				
23		922,5				
24						
25						

Figure 4.2.7 Le résultat de la transposition

Créer une nouvelle feuille de calculs appelée «Analyse1». Dans cette nouvelle feuille, insérer un tableau croisé dynamique à partir de la plage verticale contenant les moyennes, dans le but de construire une histogramme.

En suivant la procédure décrite dans la [Sous sous-section 3.2.1.4](#), faire le regroupement dans tableau croisé dynamique. Ensuite, produire l'histogramme pour représenter les données.

4.2.4 L'influence du nombre d'échantillons

On répète maintenant les étapes des sous-sections [Sous-section 4.2.2](#) et [Sous-section 4.2.3](#) afin de regarder le comportement lorsque le nombre d'échantillons tirés est de 10, 100, 1000 et 10000 trajets. On note que pour aller avec un plus grand nombre d'échantillons, par exemple 100000, il faudrait changer la manière de procéder puisqu'Excel est limité à un peu plus de 16000 colonnes.

Une fois les histogrammes pour les autres cas créés, on les combine dans un seul graphique afin d'observer l'impact du nombre d'échantillons sur le comportement de la moyenne \bar{X} . On pourrait utiliser le polygone de fréquences d'Excel pour afficher simultanément les courbes sur un même graphique, mais ce type de graphique ne permet pas de tracer des séries de données qui possèdent des abscisses différents, ce qui est fort probablement le cas ici.

On utilise plutôt le graphique « Nuages de points ». De plus, on utilise l'option « avec lissage » dans le but de voir le comportement limite s'approchant de la loi normale.

1. Créer une nouvelle feuille de calculs appelée « Polygones de fréquences ».
2. Recopier les valeurs ayant servi à construire les histogrammes, en prenant soin de changer les valeurs des abscisses pour le point milieu de chacune

des classe. Le résultat devrait ressembler à ce qui est illustré à la [Figure 4.2.8](#).

3. Sélectionner la première plage de données, incluant le titre $n = 5$.
4. Sous l'onglet **Insertion**, cliquer sur le bouton « Insérer un nuage de point (X,Y) ou un graphique en bulles » et choisir l'option « Nuage de points avec courbes lissées ».
5. Déplacer ce graphique afin qu'il ne couvre pas les séries de données.
6. Sélectionner le graphique. Sous l'onglet **Création de graphique**, cliquer sur **Sélectionner des données**. Il est également possible de faire un clic droit sur le graphique et de cliquer sur **Sélectionner des données**.
7. Dans la fenêtre qui s'ouvre, cliquer sur **Ajouter** sous « Entrées de légende (Séries).
8. Sélectionner les plages correspondant au titre, aux abscisses et aux ordonnées (les pourcentages). Appuyer sur **OK**.
9. Répéter afin d'ajouter toutes les séries de données.
10. Effectuer la mise en forme, selon la procédure décrite à la [Sous sous-section 3.2.2.2](#). La [Figure 4.2.9](#) illustre ces étapes.

	A	B	C	D	E	F	G	H	I	J	K
1	n=5		n=10		n=100		n=1000		n=10000		
2	400	0%	550	0,00%	425	0%	275	0%	350	0%	
3	600	20,00%	650	40,00%	575	14,00%	425	2,00%	450	1,73%	
4	800	20,00%	750	0,00%	725	20,00%	575	11,10%	550	5,97%	
5	1000	20,00%	850	20,00%	875	27,00%	725	25,40%	650	13,66%	
6	1200	40,00%	950	30,00%	1025	22,00%	875	29,30%	750	19,19%	
7	1400	0,00%	1050	10,00%	1175	11,00%	1025	17,30%	850	20,22%	
8			1150	0,00%	1325	4,00%	1175	8,80%	950	15,27%	
9					1475	2,00%	1325	3,30%	1050	10,05%	
10					1625	0,00%	1475	2,80%	1150	5,83%	
11						1625	0,00%	1250	3,35%		
12								1350	2,15%		
13									1450	1,11%	
14									1550	1,47%	
15									1650	0,00%	
16											
17											

Figure 4.2.8 Les données pour les courbes de distribution

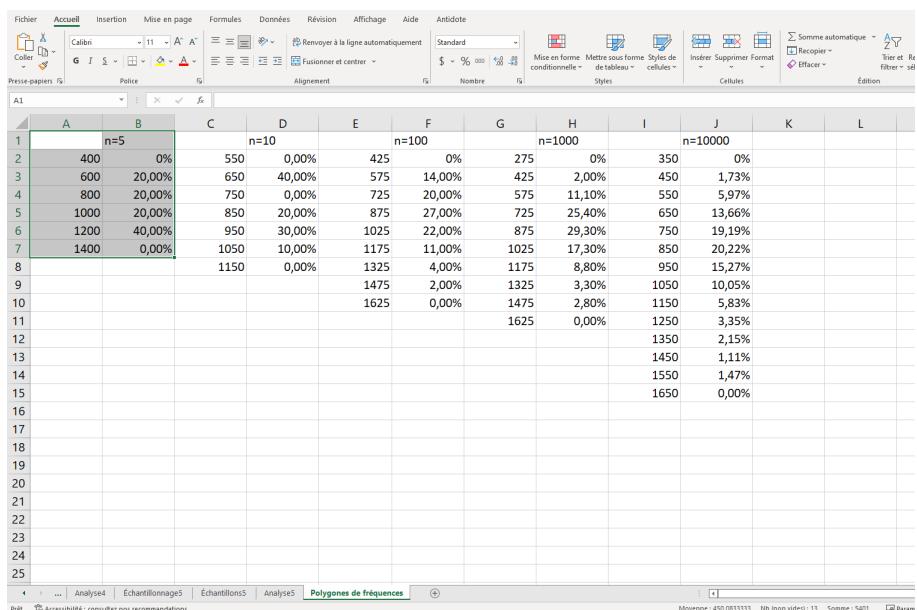


Figure 4.2.9 Création des polygones de fréquences lissés.

On termine cette l'analyse de l'impact du nombre d'échantillons pour une taille fixé en regardant le comportement de la moyenne des moyennes et de l'écart type des moyennes. Pour cela, on commence par calculer la moyenne et l'écart type de la population, chose qu'il n'est normalement pas possible de faire.

1. Créer une feuille de calcul appelée « Analyse quantitative taille 10 ».
2. Reproduire le résultat de la [Figure 4.2.10](#). Évidemment, les valeurs seront différentes. S'assurer de calculer les moyennes à partir des données qui ont été gelées et de prendre la formule pour un échantillon pour les écarts types, sauf pour celui de la population.

Échantillons	Moyenne	Écart type
Nb_E=5	987,58	295,49766
Nb_E=10	826,43	146,332263
Nb_E=100	897,76	216,891033
Nb_E=1000	880,3438	220,501633
Nb_E=10000	873,87568	221,132223
Population	874,956088	694,079566

Figure 4.2.10 Table de comparaison pour les moyennes et les écarts types

On observe que la moyenne des moyennes semble se rapprocher de la véritable moyenne, mais pas l'écart type. Ce dernier semble tout de même se rapprocher d'une valeur.

4.2.5 L'influence de la taille des échantillons

On regarde maintenant comment se comporte la moyenne échantillonnale lorsque le nombre d'échantillons reste fixe, mais que la taille de chacun des échantillons augmente. En répétant les étapes des sous-sections précédentes, faire l'analyse complète pour des tailles des échantillons de taille 10, 30, 100, 1000 et 10000. Fixer le nombre d'échantillons pour chaque cas à 100. À noter que le cas pour la taille 10 a déjà été effectué et devrait se trouver dans les feuilles de calculs *Échantillonnage3*, *Échantillons3* et *Analyse3*.

Encore une fois, on peut observer que la moyenne des moyennes se rapproche de la véritable moyenne, mais pas l'écart type. Toutefois, l'écart type semble devenir de plus en plus petit, en accord avec le théorème central limite, ce qui transpire aussi dans l'allure des courbes, qui s'écrasent de plus en plus vers la valeur centrale à mesure que n augmente.

4.3 Réflexions

Le travail accompli dans ce laboratoire théorique aura permis de voir que, même avec un échantillon de grande taille, une statistique comme la moyenne n'est qu'une approximation de la valeur réelle dans la population.

Travail à faire après le laboratoire

Objectifs

- Comparer certains résultats obtenus avec la théorie.
 - Poser un regard critique sur les données.
 - Formuler des hypothèses de recherche.
- Dans l'analyse de la moyenne échantillonnale en fonction de la taille de l'échantillon, pourquoi est-ce que l'écart-type varie en fonction de n ? Est-ce ce qui est attendu de la théorie?
 - Utiliser les écarts types calculés dans la sous-section [Sous-section 4.2.5](#) afin de tracer le graphique de l'écart type en fonction de la taille de l'échantillon. Ajouter une courbe de tendance de type « puissance » et comparer l'équation donnée par Excel avec celle établie par le théorème central limite. Commenter les différences.

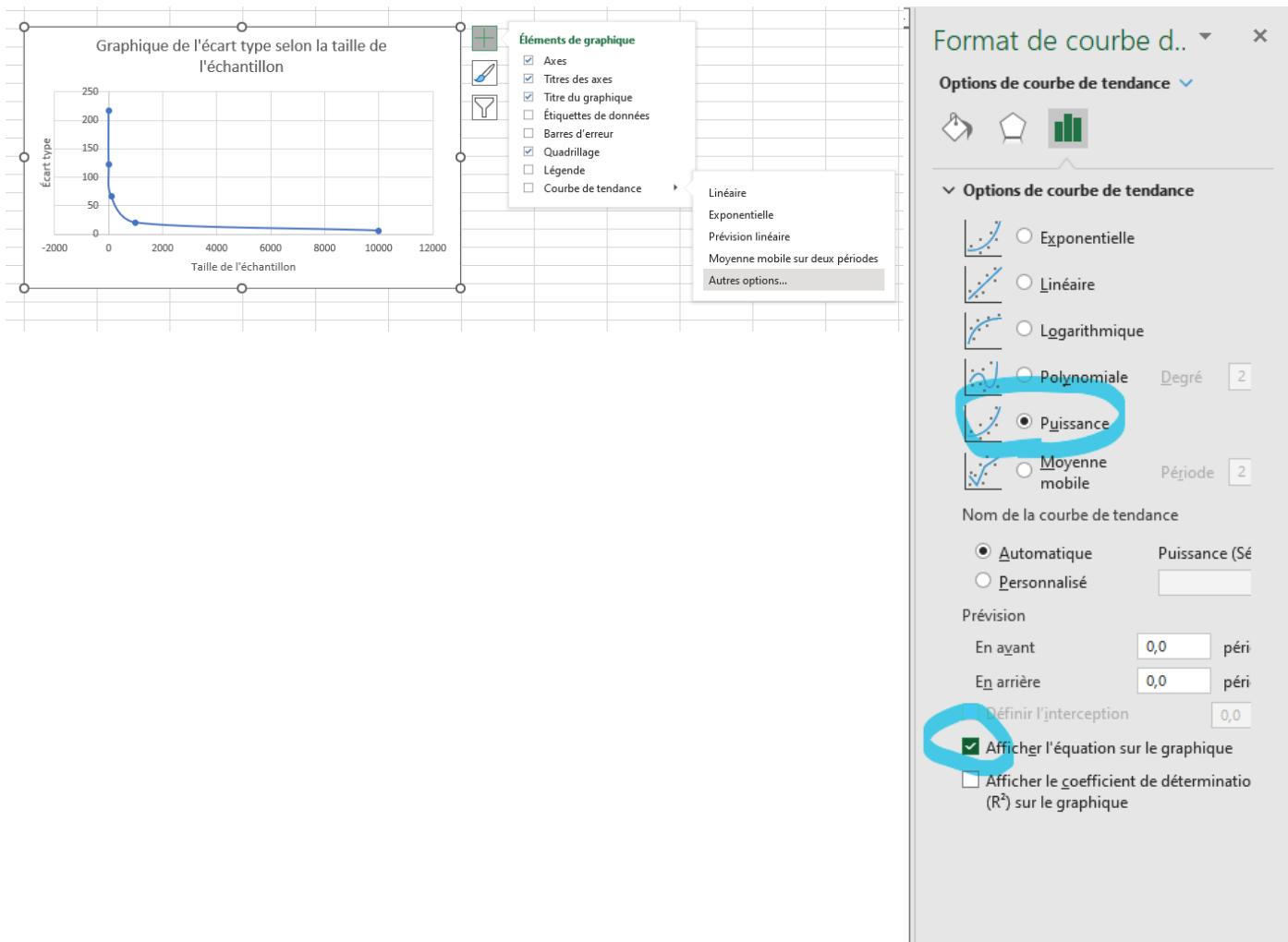


Figure 4.3.1 Insertion d'une courbe de tendance de type « puissance »

- Commenter l'allure des courbes à mesure que la valeur de n ou Nb_E augmente. Est-ce le comportement attendu? Justifier brièvement.
- La théorie de la loi normale stipule que 99.73% des données provenant d'une loi normale se situe entre -3σ et 3σ . Vérifier ce fait à l'aide des 10 000 échantillons de taille 10. Utiliser la véritable moyenne μ et $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{10}}$.
- Regarder la valeur minimale et la valeur maximale de la durée des trajets.
 - Est-ce que ces valeurs apparaissent suspectes? Expliquer pourquoi (au besoin, convertir en heures).
 - En tant qu'analyste, quelle démarche serait-il possible d'entreprendre afin de vérifier la validité des données?
- Le graphique ci-dessous illustre la distribution du maximum échantillonnale pour 100 échantillons de 10 000 trajets en BIXI. On peut voir que la forme de la distribution n'est pas la même que la cloche de la loi normale, contrairement à

celle pour la distribution des moyennes. Commenter cette remarque.

Répartition de 100 échantillons aléatoires de 10000 trajets en BIXI selon la durée maximale des trajets, Montréal, août 2016

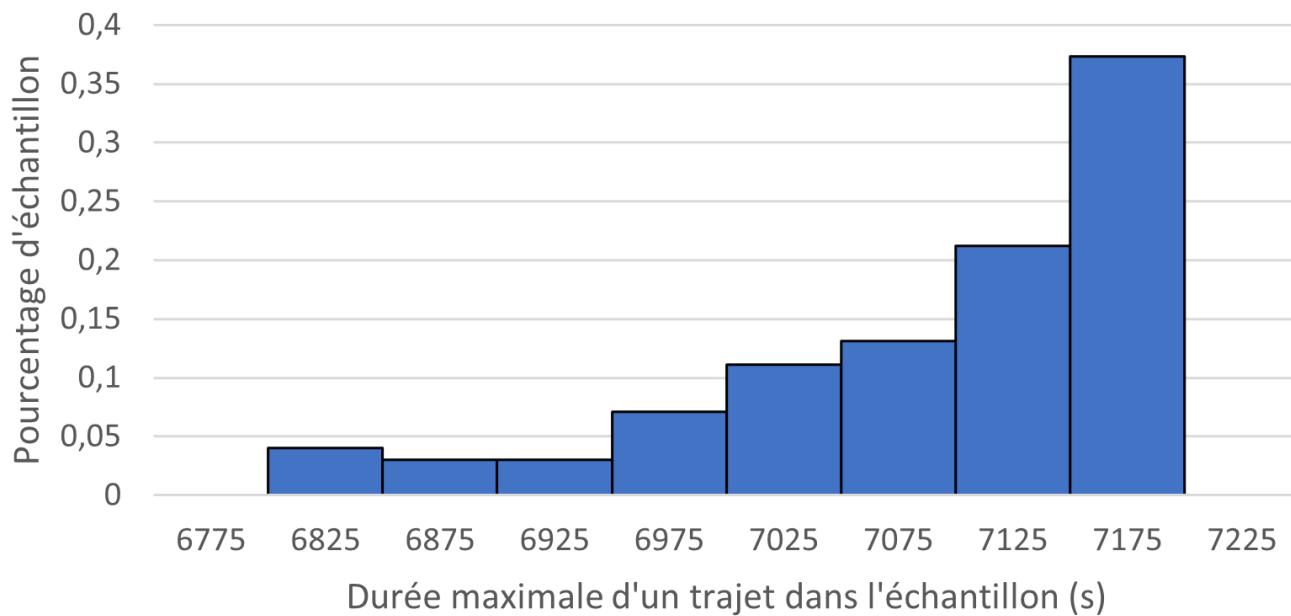


Figure 4.3.2

7. Générer 100 échantillons de taille 1000 et considérer la proportion des trajets qui proviennent de membres de BIXI. Tracer l'histogramme des proportions échantillonnaires.
8. Formulez quelques hypothèses utilisant certaines les variables de cette base de données.

Chapitre 5

Estimation par intervalle de confiance

Text before the first section.

5.1 Prélab

La moyenne ou la proportion échantillonnale constituent une première estimation du paramètre de la population. Toutefois, cette estimation, dite ponctuelle, n'offre pas beaucoup de détails quant à sa précision. Il est possible de définir un intervalle autour de la mesure échantillonnale selon une précision souhaitée. On peut alors quantifier le risque d'erreur, c'est-à-dire la proportion des intervalles ainsi construits qui ne contiendraient pas le véritable paramètre de la population.

Travail à faire avant le cours

Objectifs

- Explorer le concept de marge d'erreur.
- Introduire les notions de risque d'erreur et de niveau de confiance.

Dans le prélab, on poursuit l'analyse de la base de données BIXI en cherchant à définir un ensemble de valeurs possibles pour estimer la durée moyenne des trajets ainsi que la proportion de trajets effectués par des membres. Il convient encore une fois ici de rappeler qu'en temps normal, les données de la population ne seraient pas disponibles et que cette absence est la raison d'être de faire ces estimations. Le but ici est de faire ces estimations par intervalle de confiance afin de valider la théorie en comparant avec les valeurs réelles dans la population.

1. Selon le théorème central limite, la distribution de \bar{X} se rapproche d'une loi normale à mesure que la taille de l'échantillon augmente. Concrètement, on peut dire que \bar{X} suit une loi approximativement équivalente à $\mathcal{N}(\mu; \frac{\sigma^2}{n})$. Dans cet exercice, on considère des échantillons de taille $n = 30$ avec $\mu = 874,96$ et $\sigma = 694,08$.

(a) Calculer la probabilité que $\bar{X} \in [626,59; 1123,32]$.

(b) Dans le classeur associé à la base de données Bixi, générer à nouveau 1000 échantillons de taille 30 comme ceux se trouvant dans la feuille **Échantillons6**. Il est nécessaire de regénérer puisqu'Excel a sans doute changé les valeurs de l'échantillon depuis que les moyennes ont été fixées.

(c) Calculer les moyennes échantillonnelles (si ce n'est pas déjà fait), de même que l'écart type des 1000 échantillons.

(d) Déterminer quelle proportion de ces 1000 moyennes échantillonnelles se trouvent dans l'intervalle [626,59; 1123,32]. Comparer avec la valeur obtenue à l'exercice (a). Utiliser les [provisional cross-reference: commandes dans l'annexe] NB.SI ou NB au besoin.

2. Dans la question précédente, l'intervalle $[626, 59; 1123, 32]$ a été choisi afin que la probabilité demandée à la question [Tâche 5.1.1.a](#) soit celle obtenue. On explore la construction de cet intervalle.
- (a) Quelle est la valeur de $z_{\alpha/2}$ pour que la probabilité qu'une variable normale centrée réduite se retrouve dans l'intervalle $[-z_{\alpha/2}; z_{\alpha/2}]$ soit égale à 95%.
- (b) Vérifier que $[\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}] = [626, 59; 1123, 32]$.
- (c) Le terme $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ est appelé la **marge d'erreur**. Si on fixe $n = 30$, que devrait-être $z_{\alpha/2}$ pour avoir une marge d'erreur de 50 secondes?
- (d) Si cette fois on fixe $\alpha = 0.05$, déterminer la taille d'échantillon minimale nécessaire afin d'avoir une marge d'erreur de 50 secondes.
- (e) Discuter de la différence entre ces deux intervalles de confiance.

5.2 Laboratoire

Comme mentionné dans l'introduction, on n'a normalement pas accès aux valeurs de la population. On va alors tenter de reproduire le calcul d'un intervalle équivalent à celui de l'exercice (5.1.1.a), mais qui sera centré autour de \bar{x} et qui utilisera s plutôt que σ . Les intervalles auront la forme

$$\left[\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}; \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right], \quad (5.2.1)$$

où $\alpha = 0.05$. On reproduit ainsi un intervalle dont le **risque d'erreur** est de 5%. Ci-dessous, on analyse l'impact de prendre s comme approximation de σ .

5.2.1 Intervalles pour une moyenne

Créer une nouvelle feuille de calculs nommée « Intervalle1 » dans le fichier **Bixi**. Dans la plage A1:B4, inscrire dans la colonne A les étiquettes n =, α = et $z_{\alpha/2}$ = et entrer 30 et 0.05 dans B1 et B2 respectivement pour n et α . Calculer à l'aide des formules de la loi normale la valeur de $z_{\alpha/2}$.

On commence par construire un tableau qui contiendra les moyennes échantillonnelles, les écarts types échantillonaux, ainsi que les bornes inférieures et supérieures de l'intervalle de l'équation (5.2.1) des échantillons créés à l'exercice 5.1.1.b. Pour cela,

1. Reporter les moyennes pour analyse et les écarts types calculés à l'exercice 5.1.1.c;
2. Ensuite, calculer les bornes inférieures et supérieures de chaque intervalle selon la formule (5.2.1);
3. Ajouter une ligne (ou une colonne, selon la construction précédente) qui déterminera si μ se retrouver dans l'intervalle pour chaque valeurs \bar{x} . Pour cela, on utilise ensemble les commandes SI et ET. La commande SI retournera 1 si sa condition est remplie et 0 sinon. La condition en tant que telle sera donnée par la fonction ET, qui vérifiera si la véritable moyenne est dans l'intervalle. La formule ressemblera à `SI(ET(B4>=H2;B4<=I2);1;0)`. Dans cet exemple, la cellule B4 contient la vraie moyenne de la population et les cellules H2 et I2 sont respectivement les bornes inférieures et supérieures de l'intervalle de confiance;
4. Calculer la proportion des intervalles qui contiennent μ . Pour cela, il suffit de compter combien de 1 ont été retournés par les fonctions SI et de diviser par le nombre d'échantillons totaux, soit 1000 dans ce cas. Comparer avec le résultat « attendu » de 95%. Au besoin, recalculer la feuille avec **Shift**+**F9** si le calcul automatique est désactivé.

À présent, on s'intéresse au résultat des questions 5.1.2.c et 5.1.2.d du pré-laboratoire.

Créer une copie de la feuille **Intervalle1** et modifier la valeur de α afin qu'elle corresponde à la valeur de l'exercice 5.1.2.c. Vérifier que la proportion des intervalles contenant la moyenne est près du niveau de confiance attendu et vérifier également que la marge d'erreur est d'environ 50 secondes. en ajoutant une colonne ou une ligne «marge d'erreur».

Créer ensuite une série de nouvelles feuilles afin de tirer 100 échantillons de la taille appropriée, soit celle calculée à la question 5.1.2.d. Vérifier que la proportion des intervalles contenant la moyenne est près du niveau de confiance 95% et vérifier également que la marge d'erreur est d'environ 50 secondes, comme il est attendu.

En vertu de ces constructions, est-ce que la réponse à l'exercice Tâche 5.1.2.e est toujours la bonne?

5.2.2 Intervalles pour une proportion

À l'[Activité 4.3.7](#), on a considéré la proportion de trajets effectués en BIXI qui proviennent des membres de la plateforme. On va construire un intervalle de confiance pour ce paramètre. On note π la proportion théorique de la population. On cherche à étudier le comportement d'un intervalle de confiance construit autour d'une estimation ponctuelle p . Si l'on considère qu'un membre de BIXI est un succès, alors la variable X représentant le nombre de membres dans un échantillon aléatoire pris avec remise suit une loi binomiale de paramètres n, π . Sous les conditions $n \geq 30, n\pi \geq 5$ et $n(1 - \pi) \geq 5$, alors on peut montrer que la proportion $P = \frac{X}{n}$ se comporte approximativement comme une loi normale. On dit alors que

$$\frac{X}{n} \sim \mathcal{N}\left(\pi; \frac{\pi(1 - \pi)}{n}\right)$$

sous les conditions nommées plus haut.

Si p est l'estimation ponctuelle provenant d'un échantillon, on a alors

$$\pi \in \left[p - z_{\alpha/2} \sqrt{\frac{\pi(1 - \pi)}{n}}; p + z_{\alpha/2} \sqrt{\frac{\pi(1 - \pi)}{n}} \right]$$

à un niveau de confiance de $1 - \alpha$.

Puisque le but est d'estimer la proportion π inconnue, on remplace sa valeur dans l'intervalle par l'estimation

$$\pi \in \left[p - z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}}; p + z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}} \right], \quad (5.2.2)$$

ce qui constitue l'intervalle de confiance pour la proportion π .

1. Créer une nouvelle feuille de calculs appelée «Intervalle proportion1».
2. Comme on l'a fait à l'exercice [5.1.1.d](#), on commence par regarder la proportion des valeurs estimées p se retrouve dans l'intervalle théorique $\left[\pi - z_{\alpha/2} \sqrt{\frac{\pi(1 - \pi)}{n}}; \pi + z_{\alpha/2} \sqrt{\frac{\pi(1 - \pi)}{n}}\right]$. Le résultat devrait ressembler à celui de la figure [Figure 5.2.1](#).
3. Pour chacun des 100 échantillons générés à l'exercice [4.3.7](#), construire l'intervalle de confiance selon la formule [\(5.2.2\)](#) et vérifier dans combien de ces intervalles se trouve la véritable proportion π .

pi=	0,770488278
n=	1000
alpha=	0,05
Zalpha	1,959963985
Binf	0,744424729
Bsup	0,796551826
Pourcentage dans intervalle	97,00%

Figure 5.2.1 Proportion des estimations p dans l'intervalle de confiance

5.3 Réflexions

Travail à faire après le laboratoire

Objectifs

- -
 -
1. Dans la [Sous-section 5.2.1](#), on a voulu construire des intervalles de confiance avec une marge d'erreur de 50 secondes. Expliquer pourquoi les intervalles calculés n'ont pas exactement 50 comme marge d'erreur.
 2. Commenter la différence entre la proportion des échantillons pour lesquels \bar{x} est dans l'intervalle de confiance théorique autour de μ (calculée à [Tâche 5.1.1.d](#)) et entre la proportion des intervalles de confiance autour des différentes valeurs de \bar{x} qui contiennent μ (calculée dans le [Section 5.2](#)).

Chapitre 6

Tests d'hypothèses

Ce chapitre présente les tests d'hypothèse paramétriques, une branche de l'inférence statistique. L'objectif principal d'une étude scientifique est de répondre aux questions de recherche en généralisant les résultats obtenus à partir d'un échantillon à l'ensemble de la population. Les observations issues de l'échantillon révèlent-elles une modification du comportement au sein de la population, ou sont-elles simplement le fruit du hasard lié à l'échantillonnage?

6.1 Prélab

Dans une recherche scientifique, la première étape consiste à planifier l'étude. Cela inclut la revue du contexte théorique, la formulation des questions et des hypothèses de recherche, la définition des unités statistiques ainsi que des variables à étudier, le choix de la méthode d'échantillonnage et de l'outil de collecte de données.

Ensuite, la collecte de données est effectuée à l'aide de l'outil de collecte approprié. Une fois les données récoltées, il convient de les organiser et de les traiter.

Le but ultime d'une étude scientifique est de répondre aux questions de recherche en généralisant les résultats d'un échantillon à la population. Les résultats observés dans l'échantillon révèlent-ils un changement dans le comportement de la population, ou sont-ils simplement dus au hasard de l'échantillonnage? Dans les laboratoires 1, 2 et 3, il a été question de traitement et d'organisation de données. La prochaine étape de la recherche consiste à réaliser des tests d'hypothèse paramétriques, notamment des tests d'hypothèse sur une moyenne, des tests sur une proportion, des tests de comparaison entre deux moyennes et des tests de comparaison entre deux proportions. Ce sera l'occasion de se familiariser avec des canevas Excel fournis et de comprendre les subtilités des tests d'hypothèses paramétriques.

Travail à faire avant le cours

Objectifs

- Examiner des séries statistiques.
 - Effectuer une revue de la littérature.
 - Poser un regard critique sur des données.
 - Formuler des hypothèses de recherche.
 - Se questionner sur la généralisation des mesures échantillonnelles à la population.
1. Une enseignante veut effectuer une expérience en classe avec ses élèves. Dans sa main se trouve un petit sac noir. Elle indique aux élèves que le sac contient six billes, quatre billes bleues et deux billes vertes. Ensemble, ils vont réaliser une expérience aléatoire qui consiste à piger avec remise six billes du sac. On s'intéresse au nombre de billes bleues pigées. À tour de rôle, la personne enseignante fait piger une bille par un élève. À la fin du tirage, six billes vertes ont été pigées. Ce résultat peut être vu comme le résultat du prélèvement d'un échantillon aléatoire. Piger six billes vertes, est-ce un résultat attendu? Sur quoi faut-il se baser pour répondre à cette question? À quoi compare-t-on la valeur échantillonnelle?
2. Justifier la réponse de l'exercice 1 avec les calculs appropriés (cote z ou probabilité d'obtenir un tel résultat, soit piger six billes vertes).

3. À la lumière des réflexions des exercices 1 et 2, quelles sont les deux conclusions possibles? Laquelle vous semble la plus probable?
4. Une enseignante veut effectuer une expérience en classe avec ses élèves. Elle donne à chaque élève un dé à six faces. Elle souhaite savoir quels dés sont truqués (pipés). Pour le déterminer, les élèves doivent les lancer plusieurs fois et noter la distribution des résultats obtenus. Lors de l'élaboration d'un test d'hypothèse, il faut formuler une hypothèse de base et une hypothèse alternative. L'hypothèse de base, soit H_0 , communément appelée l'hypothèse nulle, est un énoncé à tester partant de l'idée qu'il n'y a pas de différence significative dans le paramètre mesuré ou de l'idée que la distribution obtenue se répartit selon un modèle théorique connu. L'hypothèse alternative est l'énoncé contraire à l'hypothèse nulle, soit qu'il y a une différence dans le paramètre mesuré. Dans ce scénario, pourquoi est-ce que l'hypothèse nulle serait que le dé n'est pas pipé. Pourquoi est-ce que l'hypothèse alternative serait que le dé est pipé? Expliquer.

5. La quantité moyenne quotidienne de précipitations totales reçues au courant du mois de mars à Montréal entre 1991 et 2020 est de 2,49 mm par jour. Cependant, ces dernières années, de nombreux Montréalais ont l'impression que les précipitations mensuelles sont en baisse^{(1) (2) (3)}. Pour vérifier cette hypothèse, on souhaite réaliser un test d'hypothèse sur une moyenne en se basant sur les données échantillonnelles d'Environnement Canada pour le mois de mars 2024 (voir la [Section A.7](#)) pour vérifier si la quantité moyenne de précipitations a diminué. On a obtenu une valeur de 1,93 mm par jour. Il se peut que l'échantillon prélevé a donné une moyenne un peu plus basse que le paramètre de la population. C'est ce qu'on appelle la variabilité échantillonnale. Chaque échantillon donne une moyenne différente et rarement égale à la moyenne de la population. Il se peut que la différence ne soit pas simplement due au hasard de l'échantillonnage. Sans faire de calculs, est-ce que la différence entre 2,49 mm (μ) et 1,93 mm (\bar{x}) paraît grande à première vue, et donc significative? Justifier ce choix.

¹<https://www.lapresse.ca/actualites/environnement/2022-11-13/la-premiere-neige-de-plus-en-plus-tardive.php>, page consultée le 23 novembre 2024

²<https://ici.radio-canada.ca/nouvelle/2038368/deficit-significatif-neige-quebec>, page consultée le 23 novembre 2024

³<https://lactualite.com/actualites/montreal-a-perdu-le-tiers-de-sa-neige-depuis-1863/>, page consultée le 23 novembre 2024

6.2 Laboratoire

Dans ce laboratoire, l'objectif est de poursuivre l'analyse d'une base de données, soit l'étape d'inférence statistique. On souhaite répondre à des questions de recherche bien formulées et valider les hypothèses de recherche à l'aide de tests paramétriques.

Quatre bases de données seront exploitées, chacune correspondant à un test d'hypothèse spécifique. Pour le test d'hypothèse sur une moyenne, on analysera les données des précipitations totales reçues au mois de mars 2024, présentées à la [Section A.7](#). Le test d'hypothèse sur une proportion s'appuiera sur les données démographiques des soldats de l'armée américaine disponibles à la [Section A.1](#). Pour le test sur deux moyennes indépendantes, les données diagnostiques des femmes d'origine pima d'Arizona (voir la [Section A.5](#)) seront utilisées. Enfin, concernant le test sur deux moyennes dépendantes, une étude sur la qualité de l'air à Montréal sera réalisée avec les données présentées à la [Section A.6](#).

6.2.1 Présentation des canevas

Les tests d'hypothèses paramétriques comportent plusieurs étapes : formuler les hypothèses nulle et alternative, préciser les informations du problème et fixer le seuil de signification, vérifier les conditions d'application du test, calculer l'écart type $\sigma_{\bar{X}}$, énoncer la règle de décision, calculer la statistique du test, calculer la cote z ou t observée de la statistique ou la valeur p , puis prendre une décision et conclure.

Le logiciel Excel réalise certaines de ces étapes, mais ne décompose pas toutes les phases d'un test. Ainsi, un canevas a été créé pour bien visualiser l'ensemble des étapes d'un test de manière plus détaillée. Dans cette sous-section, on explique les particularités des canevas.

La figure [Figure 6.2.4](#) présente la configuration du canevas pour un test d'hypothèse sur une moyenne, soit la feuille ***Une moyenne*** du fichier Excel ***Canevas_tests_hypotheses.xlsx***.

Test d'hypothèse sur une moyenne										
1	A	B	C	D	E	F	G	H	I	J
2	1) Hypothèses :						<input type="button" value="Effacer le contenu des cellules grises"/>			
3										
4										
5	2) Seuil et informations :			$\alpha =$	$\bar{x} =$					
6				$N =$	$n =$					
7				$\sigma =$	$s =$					
8										
9	3) Conditions d'application :			$Z_0 =$						
10										
11										
12										
13										
14	4) Écart type $\sigma_{\bar{X}} :$			$N =$	$20 * n =$					
15										
16										
17										
18										
19										
20	5) Règle de décision :			$\sigma_{\bar{X}} =$						
21										
22										

Figure 6.2.1 Configuration générale d'un canevas (quatre premières étapes d'un test)

26	A	B	C	D	E	F	G	H	I	J	K
27											
28											
29	5) Règle de décision :		$=$								
30											
31											
32											
33											
34	Méthode 1 (statistique du test) :		$=$								
35											
36											
37											
38	Méthode 2 (valeur p) :		On rejette H_0 si $p \leq$								
39											
40	Calcul de la statistique ou la valeur p :		$=$								
41											
42	Méthode 1 (valeur p) :		$valeurs p =$								
43											

Figure 6.2.2 Configuration générale d'un canevas (étapes 5 et 6 d'un test)

A	B	C	D	E	F	G	H	I	J	K
45	7) Décision et conclusion :									
46										
47										
48										
49										
50										

Figure 6.2.3 Configuration générale d'un canevas (dernière étape d'un test)

A	B	C	D	E	F	G	H	I
Test d'hypothèse sur une moyenne								
5	1) Hypothèses :	$H_0 : \mu =$						
6		$H_1 : \mu$						
7								
8								
9	2) Seuil et informations :	$\alpha =$		$\bar{x} =$				
10		$N =$			$n =$			
11		$\sigma =$			$s =$			
12								
13								
14	3) Conditions d'application :	Loi :						
15								
16		car						
17								
18								
19								
20	4) Écart type $\sigma_{\bar{X}}$:	$N =$		$20 * n =$				
21								
22		$\sigma_{\bar{X}} =$						

Figure 6.2.4 Configuration générale d'un canevas

- Les cellules ayant un fond bleu contiennent les titres de chaque étape du test d'hypothèse.
- Les cellules ayant un fond rose ne peuvent pas être modifiées. À cet effet, elles sont protégées pour empêcher l'utilisateur d'y apporter des changements. Certaines de ces cellules contiennent des formules; par exemple, la cellule C22 du canevas contient la formule permettant de calculer $\sigma_{\bar{X}}$. Cette cellule reste vide jusqu'à ce qu'on ait attribué une valeur numérique à tous les paramètres nécessaires à son calcul. Il arrive parfois que du contenu soit effacé par erreur lors de la saisie. La protection des cellules prévient ce type d'erreur.
- Les cellules ayant un fond gris et une bordure noire doivent être remplies. Certaines de ces cellules ont des menus déroulants permettant à l'utilisateur de choisir une option. Par exemple, la cellule D6 contient un menu déroulant obligeant l'utilisateur à choisir le symbole approprié ($<$, $>$ ou \neq) pour le type de test que ce dernier souhaite réaliser (voir la [Figure 6.2.5](#)).

A	B	C	D	E	F	G	H	I
Test d'hypothèse sur une moyenne								
5	1) Hypothèses :	$H_0 : \mu =$						
6		$H_1 : \mu$						
7								
8	2) Seuil et informations :	$\alpha =$						
9		$N =$						
10								

Figure 6.2.5 Menu déroulant de certaines cellules

Lorsqu'une cellule griseée contient un menu déroulant, l'utilisateur doit choisir une des options.

D'autres cellules grisesées sont vides et doivent être remplies par l'utilisateur. Parfois, il faut y inscrire des valeurs numériques; d'autres fois, il faut y inscrire du texte. Tel est le cas pour la cellule E5. L'utilisateur doit taper la valeur numérique de μ_0 qu'il souhaite tester.

Certaines des cellules grisées sont accompagnées d'une boîte offrant des instructions lorsque la cellule est sélectionnée. La figure Figure 6.2.6 présente un tel exemple pour la cellule C10.

	A	B	C	D
1				
2				
3				
4				
5	1) Hypothèses :		$H_0 : \mu =$	
6			$H_1 : \mu$	
7				
8				
9	2) Seuil et informations :		$\alpha =$	
10			$N =$	
11			$\sigma =$	
12				
13				
14	3) Conditions d'application :		Loi :	
15				

Figure 6.2.6 Exemple d'un message accompagnant la sélection d'une cellule grisée

- Lorsque l'on clique sur la cellule intitulée ***Effacer le contenu des cellules grises***(voir la Figure 6.2.5), un message d'avertissement s'affiche et demande à l'utilisateur si ce dernier veut vraiment effacer tout le contenu des cellules grises. En cliquant Oui, le contenu des cellules grises est effacé.

6.2.2 Test d'hypothèse sur une moyenne

Les tests d'hypothèses sur une moyenne permettent de décider si la moyenne d'une population a changé en se basant sur une moyenne échantillonnale, c'est-à-dire de déterminer si la différence entre les deux moyennes est statistiquement significative ou si elle n'est due qu'au hasard de l'échantillonnage.

D'après les données récoltées entre 1991 à 2020 par Environnement Canada à la station météorologique de l'Aéroport international Pierre-Elliott-Trudeau de Montréal¹, la ville de Montréal reçoit en moyenne 77,2 mm de précipitations totales au mois de mars, soit environ 2,49 mm par jour. Les précipitations totales sont la somme de la pluie totale et de l'équivalent en eau de la neige totale en millimètres. Cependant, ces dernières années, de nombreux Montréalais ont l'impression que les précipitations mensuelles sont en baisse^(2 3 4)). Pour vérifier cette hypothèse, on va la tester avec les données d'Environnement Canada pour le mois de mars 2024⁵.

¹Données tirées de https://climat.meteo.gc.ca/climate_normals/results_1991_2020_f.html?searchType=stnName_1991&txtStationName_1991=montreal&searchMethod=contains&txtCentralLatMin=0&txtCentralLatSec=0&txtCentralLongMin=0&txtCentralLongSec=0&stnID=123000000&dispBack=1, page consultée le 21 novembre 2024

²<https://www.lapresse.ca/actualites/environnement/2022-11-13/la-premiere-neige-de-plus-en-plus-tardive.php>, page consultée le 23 novembre 2024

³<https://ici.radio-canada.ca/nouvelle/2038368/deficit-significatif-neige-quebec>, page consultée le 23 novembre 2024

⁴<https://lactualite.com/actualites/montreal-a-perdu-le-tiers-de-sa-neige-depuis-1863/>, page consultée le 23 novembre 2024

⁵Données tirées de https://climate.weather.gc.ca/climate_data/daily_data_e.html?StationID=51157&timeframe=2&StartYear=1840&EndYear=2024&Day=8&Year=2024&Month=3#, page consultée le 21 novembre 2024

Calculer les mesures statistiques échantillonnelles. Pour réaliser un test d'hypothèse sur une moyenne, il faut préalablement avoir calculé les mesures statistiques de l'échantillon, c'est-à-dire la moyenne et l'écart type corrigé. Ceci peut se faire en utilisant les fonctions Excel MOYENNE et ÉCARTYPE.STANDARD.

1. Télécharger et ouvrir le classeur *Données_Précipitations.xlsx* à l'adresse suivante⁶.
2. Dans la feuille intitulée ***Données_Précipitations***, sélectionner le tableau dans son entiereté, soit la plage de cellules F5:G36.
3. En suivant les étapes présentées à la [Sous sous-section 1.2.3.1](#), attribuer au tableau le nom « **Échantillon** ».
4. En suivant les étapes présentées à la [Sous sous-section 1.2.3.3](#), attribuer des noms aux deux colonnes du tableau **Échantillon**.

Copier le cavenas *Une moyenne*. Pour effectuer un test d'hypothèse, il faut copier (importer) le canevas du fichier *Canevas_tests_hypotheses.xlsx* dans le fichier Excel de travail *Données_Précipitations.xlsx*.

1. Ouvrir les fichiers Excel *Données_Précipitations.xlsx* et *Canevas_tests_hypotheses.xlsx*.
2. Effectuer la procédure présentée au laboratoire 4 pour copier la feuille ***Une moyenne*** du fichier *Canevas_tests_hypotheses* au classeur ***Données_Précipitations***.

6.2.2.1 Étapes d'un test sur une moyenne

Les étapes pour réaliser un test d'hypothèse sur une moyenne sont présentées. On veut tester si la quantité moyenne de précipitations reçues par jour en mars à Montréal a diminué. La valeur historique de 1991 à 2020 est de 2,49 mm par jour. On va se baser sur un échantillon du mois de mars 2024. On choisit un seuil de signification de 5%.

1. Dans le classeur ***Données_Précipitations***, sélectionner la feuille ***Une moyenne***.
2. La première étape d'un test consiste à écrire la moyenne de référence de la population, soit μ_0 , ainsi qu'à définir le type de test que l'on souhaite réaliser, à savoir un test unilatéral à gauche, à droite ou bilatéral. Dans la cellule E5, taper $=77,2/31$, soit la quantité moyenne de précipitations reçues en mars de 1991 à 2020 divisée par le nombre de jours en mars. La cellule E6 se remplit automatiquement avec la même valeur (voir la [Figure 6.2.7](#)).

Puisque notre hypothèse de recherche est que la quantité moyenne de précipitations quotidiennes a diminué, on privilégie un test unilatéral à gauche. Dans la cellule D6 de la feuille ***Une moyenne***, il faut donc choisir le symbole < (voir la [Figure 6.2.7](#)).

⁶github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Pr%C3%A9cipitations.xlsx?raw=true

	A	B	C	D	E	F	G	H	I
Test d'hypothèse sur une moyenne									
1									
2									
3									
4									
5	1) Hypothèses :	$H_0 : \mu =$	<input type="text" value="=77,2/31"/>						
6	$H_1 : \mu$	<input type="text"/>							
7									
8									
9	2) Seuil et informations :	$\alpha =$	<input type="text"/>	$\bar{x} =$	<input type="text"/>				
10	$N =$	<input type="text"/>		$n =$	<input type="text"/>				
11	$\sigma =$	<input type="text"/>		$s =$	<input type="text"/>				
12									
13									

Figure 6.2.7 Détermination de la moyenne de référence μ_0 et choix de l'hypothèse alternative

3. Dans la deuxième étape du test, il faut remplir les cases grisées avec les informations demandées.

- Choisir un seuil de 0,05. Dans la cellule C9, taper $=0,05$ (voir la Figure 6.2.8).
- Puisque la taille de la population est le nombre de jours du mois de mars (31) multiplié par le nombre d'années (30), taper $=30*31$ dans la cellule C10 (voir la Figure 6.2.8).
- L'écart type de la population est inconnu. Ainsi, dans la cellule C11, taper **inconnu** (voir la Figure 6.2.8).
- Il faut trouver la moyenne échantillonnale, soit \bar{x} . Dans la cellule F9, taper $=MOYENNE(Précipitations)$ (voir la Figure 6.2.8). **Précipitations** est le nom attribué à la deuxième colonne du tableau **Échantillon**, soit la plage de cellules G5:G36.
- Dans la cellule F10, taper $=NB(Précipitations)$ pour indiquer la taille de l'échantillon (voir la Figure 6.2.8).
- Dans la cellule F11, taper $=ECARTTYPE.STANDARD(Précipitations)$ pour indiquer l'écart type corrigé s de l'échantillon (voir la Figure 6.2.8).
- Une fois les plages F9:F11 remplies, il est possible de constater que les cellules C20, E20 et C22 se remplissent automatiquement, les cellules étant préremplies avec les formules appropriées (voir la Figure 6.2.8).

	A	B	C	D	E	F	G	H	I
Test d'hypothèse sur une moyenne									
1									
2									
3									
4									
5	1) Hypothèses :	$H_0 : \mu =$	<input type="text" value="2,490322581"/>						
6	$H_1 : \mu$	<input type="text"/>	<input type="text" value="2,490322581"/>						
7									
8									
9	2) Seuil et informations :	$\alpha =$	<input type="text" value="=0,05"/>	$\bar{x} =$	<input type="text"/>				
10	$N =$	<input type="text"/>		$n =$	<input type="text"/>				
11	$\sigma =$	<input type="text"/>	La valeur du seuil de signification doit être entre 0 et 1.				$s =$	<input type="text"/>	
12									
13									
14	3) Conditions d'application :	Loi :	<input type="text"/>						
15									
16		car	<input type="text"/>						
17									
18									
19									
20	4) Écart type $\sigma_{\bar{x}}$:	$N =$	<input type="text"/>	$20*n =$	<input type="text"/>				
21									
22									

Figure 6.2.8 Remplissage des informations de l'étape 2

4. À la troisième étape, il faut vérifier les conditions d'application du test. Puisque la taille de l'échantillon est supérieure à 30, on utilise la loi normale. Dans la cellule C14, sélectionner **Normale**. Écrire un texte dans la cellule fusionnée C16 expliquant ce choix (voir la Figure 6.2.9).

A	B	C	D	E	F	G	H	I
Test d'hypothèse sur une moyenne								
5 1) Hypothèses :	H_0 :	μ	=	2,490322581				
	H_1 :	μ	<	2,490322581				
					Effacer le contenu des cellules grises			
9 2) Seuil et informations :	α =	0,05		\bar{x} =	1,9290323			
	N =	930		n =	31			
	σ =	inconnu		s =	3,297089			
14 3) Conditions d'application :	Loi :	Normale						
	car	étonnante la loi à utiliser. Il faut choisir une des deux options.						

Figure 6.2.9 Choix de la loi à utiliser pour le test

5. L'écart type de la distribution \bar{X} , soit $\sigma_{\bar{X}}$, est calculé à l'aide des informations se trouvant dans les cellules C10, C11, F10, F11, C20 et C21.
6. À la cinquième étape, il faut écrire la règle de décision, soit avec le calcul de la statistique du test ou avec la valeur p. Dans la cellule B29, sélectionner la statistique appropriée du test unilatéral à gauche utilisant la loi normale, soit $-z_\alpha$ (voir la [Figure 6.2.10](#)). Remarquer que la valeur de $-z_\alpha$ (environ $-1,645$) dans la cellule D29 et F31 est préremplie et correspond bien à la cote z pour un seuil de signification de 5%.
7. Dans la cellule E31, sélectionner les mots appropriés parmi les trois choix du menu déroulant. Puisque l'on effectue un test unilatéral à gauche, il faut choisir les mots « plus petite que » (voir la [Figure 6.2.10](#)).

La cellule F34 est préremplie avec la valeur du seuil de signification pour indiquer que l'on rejette l'hypothèse nulle si la valeur p trouvée est inférieure à 5%.

A	B	C	D	E	F	G	H	I
25	26 5) Règle de décision :							
27								
28								
29	Méthode 1 (statistique du test) :	$-z_\alpha$	=	-1,64485363				
30		z est négatif la						
31		z est négatif la						
32		z est négatif la						
33		z est négatif la						
34	Méthode 2 (valeur p) :		est		-1,6448536			
35								
		On rejette H_0 si la valeur p est inférieure ou égale à 0,05						

Figure 6.2.10 Remplissage des cellules pour l'étape 5

8. À la sixième étape, toutes les cases roses se remplissent automatiquement selon les informations fournies précédemment. On observe que la cote z_{obs} a une valeur d'environ $-0,948$. La valeur p calculée par Excel est d'environ 0,172 (voir la [Figure 6.2.11](#)).

A	B	C	D	E	F	G	H
25							
26 5) Règle de décision :							
27							
28							
29 Méthode 1 (statistique du test) :	-z _{0.5}	=	-1,64485363				
30							
31	On rejette H ₀ si	z _{obs}	est	plus petite que	-1,6448536		
32							
33							
34 Méthode 2 (valeur p) :			On rejette H ₀ si la valeur p est inférieure ou égale à	0,05			
35							
36							
37 6) Calcul de la statistique ou la valeur p :							
38							
39							
40 Méthode 1 (cote z ou t observée) :	z _{obs}	=	-0,9478459				
41							
42 Méthode 2 (valeur p) :	valeur p	=	0,171604				
43							

Figure 6.2.11 Remplissage automatique des cellules pour l'étape 6

9. Dans la cellule A47, écrire la décision et la conclusion du test en fonction de la règle de décision.

37							
38 6) Calcul de la statistique ou la valeur p :							
39							
40 Méthode 1 (cote z ou t observée) :	z _{obs}	=	-0,9478459				
41							
42 Méthode 2 (valeur p) :	valeur p	=	0,171604				
43							
44							
45 7) Décision et conclusion :							
46							
47							
48	Puisque la cote z observée de -0,948 n'est pas plus petite que la cote z critique, -z _{alpha} , de -1,645, on ne rejette pas l'hypothèse nulle						
49	H ₀ . Au seuil de 5%, les données échantillonnelles ne permettent pas de conclure que la quantité moyenne de précipitations quotidiennes						
50	en mars a diminué.						

Figure 6.2.12 Décision et conclusion du test

6.2.3 Test d'hypothèse sur une proportion

Les tests d'hypothèses sur une proportion permettent de décider si la proportion d'une population a changé en se basant sur une proportion échantillonnale, c'est-à-dire de déterminer si la différence entre les deux pourcentages est statistiquement significative ou si elle n'est due qu'au hasard de l'échantillonnage.

Selon un rapport démographique de la communauté militaire américaine⁷, en 2010, l'armée américaine comptait 30,5% de membres issus de minorités ethniques. Dans cette sous-section, on introduit les données d'un échantillon de 6068 membres de l'armée américaine, prélevé en 2011. On pourrait supposer que la proportion de membres issus de minorités ethniques augmente chaque année. C'est ce que l'on va vérifier. La base de données⁸ a été modifiée pour répondre aux objectifs de ce laboratoire. Dans le fichier, chacun des groupes ethniques est associé à un code nominal. Le code 1 signifie une personne blanche, les codes 2 à 7 correspondent aux différentes minorités ethniques.

Calculer la proportion échantillonnale. Pour réaliser un test d'hypothèse sur une proportion, il faut préalablement avoir calculé la proportion échantillonnale. Ceci peut se faire en générant un tableau croisé dynamique.

- Ouvrir le classeur **Données_Armée.xlsxm** à l'adresse suivante⁹.

⁷<https://download.militaryonesource.mil/12038/MOS/Reports/2020-demographics-report.pdf>, page consultée le 24 novembre 2024

⁸<https://www.openlab.psu.edu/ansur2/>, page consultée le 24 novembre 2024

⁹github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Arm%C3%A9e.xlsx?raw=true

2. Dans la feuille intitulée ***Données_Armée***, sélectionner le tableau dans son entiereté, soit la plage de cellules A5:I6073.
3. En suivant les étapes présentées à la [Sous sous-section 1.2.3.1](#), attribuer au tableau le nom « **Échantillon** ».
4. En suivant les étapes présentées à la [Sous sous-section 1.2.3.3](#), attribuer un nom à chacune des colonnes du tableau **Échantillon**.
5. En suivant les étapes présentées à la [Sous sous-section 1.2.8.1](#), générer, dans une nouvelle feuille de calcul intitulé **TCD_Ethnie**, un tableau croisé dynamique vide à partir du tableau **Échantillon**. Placer ce tableau dans la cellule A3.
6. Glisser et déposer la variable **Ethnie** dans la zone de saisie **Lignes** (voir la [Figure 6.2.13](#)).

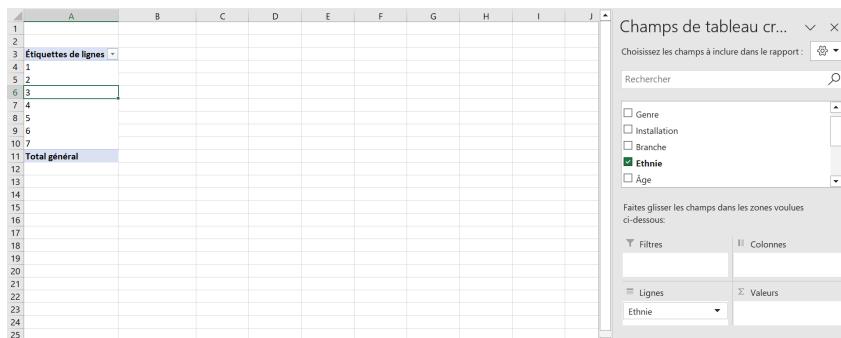


Figure 6.2.13 Tableau croisé dynamique de la répartition de l'échantillon de 6068 membres de l'armée américaine selon leur ethnie en 2011

7. Glisser et déposer la variable **Ethnie** dans la zone de saisie **Valeurs** (voir la [Figure 6.2.13](#)).
8. En suivant les étapes présentées à la [Sous sous-section 1.2.8.2](#), afficher le nombre de personnes en pourcentage (voir la [Figure 6.2.13](#)).
9. On s'intéresse au pourcentage total des codes 2 à 8. Il est possible de regrouper ces modalités pour trouver la proportion échantillonnale. Sélectionner les cellules A5:B10 (voir la [Figure 6.2.14](#)).

	A	B	C	D
1				
2				
3	Étiquettes de lignes ▾	Nombre de Ethnie		
4	1	62,49%		
5	2	21,39%		
6	3	11,19%		
7	4	3,10%		
8	5	0,81%		
9	6	0,97%		
10	7	0,05%		
11	Total général	100,00%		
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
~				

Figure 6.2.14 Tableau croisé dynamique final

10. Cliquer sur le bouton de droite de la souris et sélectionner l'onglet **Grouper** (voir la [Figure 6.2.14](#)). Le groupe 1 dans le tableau croisé dynamique regroupe tous les membres issus de minorités ethniques. La proportion échantillonnielle est de 37,51%.

Copier le cavenas *Une proportion (normale)*. Pour effectuer un test d'hypothèse, il faut copier le canevas dans le fichier de travail.

- Ouvrir les fichiers Excel **Données_Armée.xlsm** et **Canevas_tests_hypotheses.xlsm**.
- Effectuer la procédure présentée au laboratoire 4 pour copier la feuille **Une proportion** du fichier **Canevas_tests_hypotheses** au classeur **Données_Armée**. Placer cette feuille en dernière position.

6.2.3.1 Étapes d'un test sur une proportion

Les étapes pour réaliser un test d'hypothèse sur une proportion sont présentées. On veut tester si la proportion de membres de l'armée américaine issus de minorités ethniques a augmenté en 2011 versus 2010. On choisit un seuil de 5%.

- Dans le classeur **Données_Armée.xlsm**, sélectionner la feuille **Une proportion (normale)**.

2. La première étape d'un test consiste à écrire la proportion de référence de la population, ainsi qu'à définir le type de test que l'on souhaite réaliser, à savoir un test unilatéral à gauche, à droite ou bilatéral. Dans la cellule E5, taper la valeur 0,305. On a choisi d'écrire les mesures en décimales. La cellule E6 se remplit automatiquement avec la même valeur (voir la Figure 6.2.15).

Puisque notre hypothèse de recherche est que la proportion de membres de l'armée américaine issus de minorités ethniques a augmenté, on priviliege un test unilatéral à droite. Dans la cellule D6 de la feuille ***Une proportion (normale)***, il faut donc choisir le symbole > (voir la Figure 6.2.15).

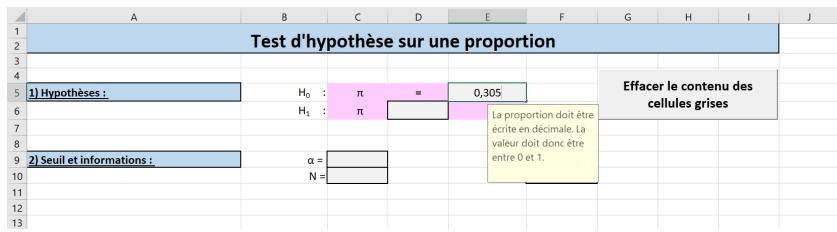


Figure 6.2.15 Détermination de la proportion de référence π et choix de l'hypothèse alternative

3. Dans la deuxième étape du test, il faut remplir les cases grisées avec les informations demandées.

- Choisir un seuil de 0,05. Dans la cellule C9, taper =0,05 (voir la Figure 6.2.16).
- Puisqu'on ne connaît pas la taille de la population de l'armée américaine, on suppose qu'elle est grande. Selon le rapport démographique de 2020, il y avait au moins 400000 membres en 2010. Ainsi, le test ne requiert pas de facteur de correction. Dans la cellule C10, taper le mot « grande » (voir la Figure 6.2.16).
- Il faut indiquer la proportion échantillonnale de P . On peut la copier de la feuille **TCD_Ethnie** ou simplement recopier la valeur manuellement. Dans la cellule F9 de la feuille ***Une proportion***, taper =0,3751 (voir la Figure 6.2.16).
- Dans la cellule F10, taper =6068 pour indiquer la taille de l'échantillon (voir la Figure 6.2.16).
- Une fois les plages F9:F10 remplies, il est possible de constater que les cellules C20, E20 et C22 se remplissent automatiquement, les cellules étant préremplies avec les formules appropriées (voir la Figure 6.2.16).

Test d'hypothèse sur une proportion										
1) Hypothèses :				$H_0 : \pi = 0,305$			Effacer le contenu des cellules grises			
H ₁ : $\pi > 0,305$										
2) Seuil et informations :				$\alpha = 0,05$	$p =$					
N =				$n =$						
3) Conditions d'application :				Loi : Normale						
car										
4) Écart type σ_P :				N =	$20 * n =$					
$\sigma_P =$										

Figure 6.2.16 Remplissage des informations de l'étape 2

4. À la troisième étape, il faut vérifier les conditions d'application du test. Puisque la taille de l'échantillon est supérieure à 30 et que les produits $n\pi$ et $n(1 - \pi)$ sont supérieurs à 5, on utilise la loi normale. Écrire un texte dans la cellule fusionnée C16 expliquant ce choix (voir la Figure 6.2.17).

Test d'hypothèse sur une proportion										
1) Hypothèses :				$H_0 : \pi = 0,305$			Effacer le contenu des cellules grises			
H ₁ : $\pi > 0,305$										
2) Seuil et informations :				$\alpha = 0,05$	$p = 0,3751$					
N = grande				$n = 6068$						
3) Conditions d'application :				Loi : Normale						
car				$n > 30, n * \pi > 5 \text{ et } n * (1 - \pi) > 5$						
4) Écart type σ_P :				N = grande	$20 * n =$			Justifier la loi choisie.		
$\sigma_P = 0,00591044$										

Figure 6.2.17 Choix de la loi à utiliser pour le test

5. L'écart type de la distribution P , soit σ_P dans la cellule C22, est calculé à l'aide des informations se trouvant dans les cellules E5, F10, C20, C21.
6. À la cinquième étape, il faut écrire la règle de décision, soit avec le calcul de la statistique du test ou avec la valeur p . Dans la cellule B29, sélectionner la statistique appropriée du test unilatéral à droite utilisant la loi normale, soit z_α . Remarquer que la valeur de z_α (environ 1,645) dans la cellule D29 et G31 est préremplie et correspond bien à la cote z pour un seuil de signification de 5% (voir la Figure 6.2.18).
7. Dans la cellule E31, sélectionner les mots appropriés parmi les trois choix du menu déroulant. Puisque l'on effectue un test unilatéral à droite, il faut choisir les mots « plus grande que » (voir la Figure 6.2.18).

A	B	C	D	E	F	G	H	I	J
26 5) Règle de décision :									
27									
28									
29 Méthode 1 (statistique du test) :	=		1,64485363						
30	-z _{alpha}	mer la							
31	z _{alpha}	le							
32	z _{alpha/2}	le							
33	appropriée du test.								
34 Méthode 2 (valeur p) :	On rejette H ₀ si la valeur p est inférieure ou égale à	0,05							
35									
36									
37									
38 6) Calcul de la statistique ou la valeur p :									
39									
40 Méthode 1 (cote z ou t observée) :	z _{obs}	=	11,8603767						
41									
42 Méthode 2 (valeur p) :	valeur p	=	0						
43									

Figure 6.2.18 Remplissage des cellules pour l'étape 5

8. À la sixième étape, toutes les cases roses se remplissent automatiquement selon les informations fournies précédemment (voir la Figure 6.2.19).
9. Dans la cellule A47, écrire la décision et la conclusion du test en fonction de la règle de décision (voir la Figure 6.2.19).

A	B	C	D	E	F	G
38 6) Calcul de la statistique ou la valeur p :						
39						
40 Méthode 1 (cote z ou t observée) :	z _{obs}	=	11,8603767			
41						
42 Méthode 2 (valeur p) :	valeur p	=	0			
43						
44						
45 7) Décision et conclusion :						
46						
47						
48						
49	Puisque la cote z observée de 11,86 est supérieure à la cote z_alpha de 1,645, on rejette l'hypothèse nulle H0. Au seuil de 5%, on peut conclure que la proportion de membres de l'armée américaine issus de minorités ethniques a augmenté en 2011.					
50						
51						

Figure 6.2.19 Décision et conclusion du test

6.2.4 Test d'hypothèse sur deux moyennes indépendantes

Les tests d'hypothèses sur deux moyennes indépendantes, aussi appelés des tests sur la comparaison ou la différence de deux moyennes, permettent de comparer les moyennes de deux populations à partir de deux moyennes échantillonnelles, c'est-à-dire de déterminer si la différence entre les deux moyennes est statistiquement significative ou si elle n'est due qu'au hasard de l'échantillonnage.

Pour réaliser un test sur deux moyennes indépendantes, il faut s'assurer que les deux échantillons sont sélectionnés indépendamment l'un de l'autre.

Dans cette sous-section, on revient sur l'une des questions de recherche du laboratoire 3, à savoir si les femmes d'origine pima d'Arizona atteintes de diabète ont un indice de masse corporelle moyen plus élevé que celui des femmes non atteintes de cette maladie. Les deux échantillons (femmes avec diabète et femmes sans diabète) sont bien indépendants.

Calculer les mesures statistiques échantillonnelles. Pour réaliser un test d'hypothèse sur deux moyennes, il faut préalablement avoir calculé les mesures statistiques de chaque échantillon, c'est-à-dire leur moyenne et leur écart type corrigé respectif. Heureusement, à la [Sous sous-section 3.2.2.3](#), les indices de masse corporelle moyens pour les femmes atteintes et non de diabète ont déjà été calculés, ainsi que les écarts types corrigés respectifs. Ils sont présentés à la [Figure 3.2.99](#). On est en mesure d'effectuer le test.

Remarque 6.2.20 Filtrage de données aberrantes. Lors d'un test d'hypothèse paramétrique, il est important de filtrer les données aberrantes

avant de calculer les mesures descriptives. Dans le cas des données diagnostiques des femmes d'origine pima, un filtrage est nécessaire si l'on veut les moyennes et les écarts types corrigés sans tenir compte des valeurs nulles de l'indice de masse corporelle.

Copier le cavenas *2 moyennes indépendantes*.

1. Ouvrir les fichiers Excel **Données_Diabète_pour_tests.xlsxm** à l'adresse suivante¹⁰, le fichier complété du laboratoire 3, et **Canevas_tests_hypotheses.xlsxm** à l'adresse suivante¹¹.
2. Effectuer la procédure présentée au laboratoire 4 pour copier la feuille *2 moyennes indépendantes* du fichier **Canevas_tests_hypotheses** au classeur **Données**.
3. Placer la feuille *2 moyennes indépendantes* en dernière position si Excel ne le fait pas par défaut.

6.2.4.1 Étapes d'un test sur deux moyennes indépendantes

Les étapes pour réaliser un test d'hypothèse sur deux moyennes indépendantes sont présentées.

1. Il faut définir X_1 et X_2 . Soit X_1 , les femmes d'origine pima d'Arizona atteintes de diabète, et soit X_2 , les femmes d'origine pima d'Arizona sans diabète. Pour se souvenir de ce choix, il est pertinent de taper dans la cellule C4 « Atteint », et « Non Atteint » dans la cellule E4 (voir la Figure 6.2.21).
2. La première étape d'un test consiste à choisir le type de test que l'on souhaite réaliser, à savoir un test unilatéral à gauche, à droite ou bilatéral. Puisque notre hypothèse de recherche est que les femmes d'origine pima atteintes de diabète ont un indice de masse corporelle moyen supérieur à celui des femmes non atteintes, on privilégie un test unilatéral à droite. Dans la cellule D6 du canevas, il faut donc choisir le symbole $>$ (voir la Figure 6.2.21).

Test d'hypothèse sur deux moyennes indépendantes									I	J
	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4										
5	1) Hypothèses :		Atteint		Non Atteint					
6		$H_0 : \mu_1 = \mu_2$								
7		$H_1 : \mu_1 > \mu_2$								
8										
9	2) Seuil et informations :		$\alpha =$		$\bar{x}_1 =$					
10			N =		n ₁ =					

Figure 6.2.21 Remplissage des cellules pour la première étape d'un test sur deux moyennes indépendantes

3. Dans la deuxième étape du test, il faut remplir les cases grisées avec les informations demandées.
 - Choisir un seuil de 0,01. Dans la cellule C9, taper $=0,01$ (voir la Figure 6.2.22).

¹⁰github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Diab%C3%A8te_pour_tests.xlsxm?raw=true

¹¹github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Canevas/Canevas_tests_hypotheses.xlsxm?raw=true

- Puisqu'on ne connaît pas la taille de la population des femmes d'origine pima vivant en Arizona, on suppose qu'elle est grande. Dans la cellule C10, taper le mot « grande » (voir la Figure 6.2.22).

Test d'hypothèse sur deux moyennes indépendantes								
			Atteint	Non Atteint				
1) Hypothèses :	$H_0 : \mu_1 = \mu_2$	$H_1 : \mu_1 > \mu_2$						Effacer le contenu des cellules grises
2) Seuil et informations :	$\alpha = 0,01$		$\bar{x}_1 =$					$\bar{x}_2 =$
	$N =$ grande		$n_1 =$					$n_2 =$
			$s_1 =$					$s_2 =$
3) Conditions d'application :	Loi :							

Figure 6.2.22 Remplissage du seuil de signification α et de la taille de la population N

- Pour les mesures descriptives échantillonnaires de X_1 et X_2 , on utilise les valeurs calculées qui se trouvent dans la feuille **Étude IMC-Atteinte** (voir la Figure 3.2.97). Dans la cellule F9 de la feuille **2 moyennes indépendantes**, taper le symbole [=]. Ensuite, sélectionner la feuille de calcul **Étude IMC-Atteinte** suivie de la cellule J5 et appuyer sur la touche **Enter** (voir la Figure 6.2.23).

Test d'hypothèse sur deux moyennes indépendantes								
			Atteint	Non Atteint				
1) Hypothèses :	$H_0 : \mu_1 = \mu_2$	$H_1 : \mu_1 > \mu_2$						Effacer le contenu des cellules grises
2) Seuil et informations :	$\alpha = 0,01$		$\bar{x}_1 =$	=				$\bar{x}_2 =$
	$N =$ grande		$n_1 =$					$n_2 =$
			$s_1 =$					$s_2 =$
3) Conditions d'application :	Loi :							

Figure 6.2.23 Insertion de la valeur de la moyenne échantillonnable des femmes atteintes de diabète \bar{x}_1

- Dans la cellule I9 de feuille **2 moyennes indépendantes**, taper le symbole [=]. Ensuite, sélectionner la feuille de calcul **Étude IMC-Atteinte**, suivie de la cellule J4 et appuyer sur la touche **Enter** (voir la Figure 6.2.24).
- Pour la taille de chaque échantillon, si cela n'a pas été fait dans la **Section 3.3**, ajouter une quatrième colonne au tableau croisé dynamique de la feuille **Étude IMC-Atteinte** calculant le nombre de femmes dans chaque catégorie. Suivre les étapes de la **Sous sous-section 1.2.8.2** (voir la Figure 6.2.24).
- Dans la cellule F10 de feuille **2 moyennes indépendantes**, taper le symbole [=]. Ensuite, sélectionner la feuille de calcul **Étude IMC-Atteinte**, suivie de la cellule L5 et appuyer sur la touche **Enter** (voir la Figure 6.2.24).
- Dans la cellule I10 de feuille **2 moyennes indépendantes**, taper le symbole [=]. Ensuite, sélectionner la feuille de calcul **Étude IMC-Atteinte**, suivie de la cellule L4 et appuyer sur la touche **Enter** (voir la Figure 6.2.24).
- Répéter les étapes précédentes pour recopier les valeurs de s_1 et s_2 dans les cellules F11 et I11 respectivement (voir la Figure 6.2.24).

- Une fois les plages F9:F11 et I9:I11 remplies, il est possible de constater que les cellules C20 et C21 se remplissent automatiquement, les cellules étant préremplies avec les formules appropriées. Dans la cellule C20 se trouve la formule pour s , une mesure nécessaire pour le calcul de l'écart type $\sigma_{\bar{X}_1 - \bar{X}_2}$ quand la taille d'un des échantillons est inférieur à 30. Il ne faut pas confondre cette mesure avec l'écart type corrigé d'un échantillon. Dans ce test, puisque les deux échantillons ont des tailles supérieures à 30, on n'a pas besoin de s (voir la Figure 6.2.24).

Test d'hypothèse sur deux moyennes indépendantes								
				Atteint	Non Atteint			
1) Hypothèses :	H ₀ :	μ_1	=	μ_2				
	H ₁ :	μ_1	>	μ_2				
Effacer le contenu des cellules grises								
2) Seuil et informations :	$\alpha =$	0,01		$\bar{x}_1 =$	35,406767			
	N =	grande		$n_1 =$		$\bar{x}_2 =$	30,859674	
				$s_1 =$		$n_2 =$		$s_2 =$
3) Conditions d'application :	Loi :							
	car							
4) Écart type $\sigma_{\bar{X}_1 - \bar{X}_2}$:	$s =$							
	$\sigma_{\bar{X}_1 - \bar{X}_2} =$							

Figure 6.2.24 Remplissage des mesures descriptives de chaque échantillon

- À la troisième étape, il faut vérifier les conditions d'application du test. Puisque les tailles d'échantillon n_1 et n_2 sont supérieures à 30, on utilise la loi normale. Dans la cellule C14, sélectionner **Normale**. Ecrire un texte dans la cellule fusionnée C16 expliquant ce choix (voir la Figure 6.2.25).

Test d'hypothèse sur deux moyennes indépendantes								
				Atteint	Non Atteint			
1) Hypothèses :	H ₀ :	μ_1	=	μ_2				
	H ₁ :	μ_1	>	μ_2				
Effacer le contenu des cellules grises								
2) Seuil et informations :	$\alpha =$	0,01		$\bar{x}_1 =$	35,406767			
	N =	grande		$n_1 =$	266		$\bar{x}_2 =$	30,859674
				$s_1 =$	6,6149824		$n_2 =$	491
							$s_2 =$	6,5607369
3) Conditions d'application :	Loi :							
	car							
4) Écart type $\sigma_{\bar{X}_1 - \bar{X}_2}$:	$s =$	non-applicable						
	$\sigma_{\bar{X}_1 - \bar{X}_2} =$	0,502163545						

Figure 6.2.25 Choix de la loi à utiliser pour le test

- L'écart type de la distribution $\bar{X}_1 - \bar{X}_2$, soit $\sigma_{\bar{X}_1 - \bar{X}_2}$, est calculé à l'aide de formule se trouvant dans les cellules C20 et C21.
- À la cinquième étape, il faut écrire la règle de décision, soit avec le calcul de la statistique du test ou avec la valeur p. Dans la cellule B27, sélectionner la statistique appropriée du test unilatéral à droite, soit z_α . Remarquer que la valeur de z_α (environ 2,33) dans la cellule D27 et G29 est préremplie et correspond bien à la cote z pour un seuil de signification de 1% (voir la Figure 6.2.26).

7. Dans la cellule E29, sélectionner les mots appropriés parmi les trois choix du menu déroulant. Puisque l'on effectue un test unilatéral à droite, il faut choisir les mots « plus grande que » (voir la Figure 6.2.26).

A	B	C	D	E	F	G	H	I	J
23									
24 5) Règle de décision :									
25									
26									
27 Méthode 1 (statistique du test) :									
28									
29									
30									
31									
32 Méthode 2 (valeur p) :									
33									
34									
35									

Figure 6.2.26 Remplissage des cellules pour l'étape 5

8. À la sixième étape, toutes les cases roses se remplissent automatiquement selon les informations fournies précédemment (voir la Figure 6.2.27). On observe que la cote z_{obs} est énorme avec une valeur d'environ 9,06. Le rejet de l'hypothèse nulle est presque assuré. De plus, pour renchérir sur ce point, la valeur p calculée par Excel est quasiment 0 (même en augmentant le nombre de décimales affichées).
9. Dans la cellule A45, écrire la décision et la conclusion du test en fonction de la règle de décision (voir la Figure 6.2.27).

36 6) Calcul de la statistique ou la valeur p :									
37									
38 Méthode 1 (cote z ou t observée) :									
39									
40 Méthode 2 (valeur p) :									
41									
42									
43 7) Décision et conclusion :									
44									
45									
46	Puisque la cote z observée de 9,06 est supérieure à la cote z_{alpha} de 2,33, on rejette l'hypothèse nulle H_0 . Au seuil de 1%, on peut								
47	conclure que l'indice de masse corporelle des femmes pimas atteintes de diabète est supérieur à celui des femmes pima non atteintes								
48	de diabète. Également, puisque la valeur p est quasiment 0 et donc inférieure à 0,01, l'hypothèse nulle est rejetée.								
49									
50									
51									

Figure 6.2.27 Décision et conclusion du test

6.2.4.2 Utilitaire d'analyse

Avec Excel, il est possible de réaliser des tests sur deux moyennes à l'aide de l'outil **Utilitaire d'analyse**. Ce dernier permet de calculer la cote t critique d'un test bilatéral ($t_{\alpha/2}$) et celle d'un test unilatéral (t_α), ainsi que la statistique du test (t_{obs}) et la valeur p du test (bilatéral et unilatéral).

Cet outil possède néanmoins un inconvénient qui rend son utilisation moins pertinente dans certains cas. Il faut préalablement filtrer les données par échantillon. Dans le cas de l'indice de masse corporelle des femmes atteintes et non du diabète, il incombe de séparer les valeurs par présence de diabète avant d'utiliser l'utilitaire d'analyse. Excel ne permet pas d'inscrire les valeurs des moyennes de chaque échantillon.

Les étapes menant aux calculs des statistiques pertinentes avec l'utilitaire d'analyse sont présentées ci-dessous pour effectuer le test sur deux moyennes indépendantes présenté à la [Sous sous-section 6.2.4.1](#).

1. Dans la feuille **Données** du classeur **Données_Diabète**, filtrer les données de la variable **Atteint** pour ne faire ressortir que les valeurs concernant les femmes atteintes de diabète (voir la [Sous sous-section 1.2.5.1](#)).

S'assurer que le filtre sur la variable ***IMC*** est toujours en place, c'est-à-dire que les valeurs nulles ne sont pas prises en compte.

2. Copier les valeurs de la colonne ***IMC*** et les coller dans la cellule D2 d'une nouvelle feuille. Taper le titre ***Atteint*** dans la cellule D1.
3. Dans la feuille ***Données***, filtrer les données de la variable ***Atteint*** pour ne faire ressortir que les valeurs concernant les femmes non atteintes de diabète (voir la [Sous sous-section 1.2.5.1](#)).
4. Copier les valeurs de la colonne ***IMC*** et les coller dans la cellule E2 d'une nouvelle feuille ajoutée. Taper le titre ***Atteint*** dans la cellule E1 (voir la [Figure 6.2.28](#)).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	R23														
1				Atteint	Non Atteint										
2				67,1	57,3										
3				59,4	52,3										
4				55	47,9										
5				53,2	46,8										
6				52,9	46,7										
7				52,3	46,6										
8				50	46,3										
9				49,7	46,2										
10				49,6	46,1										
11				49,3	45,3										
12				48,8	45,3										
13				48,3	45,3										
14				47,9	45,2										
15				46,8	45										
16				46,2	44,6										
17				46,1	44,5										
18				45,8	44,2										
19				45,7	43,5										
20				45,6	43,5										
21				45,6	43,4										
22				45,5	43,3										
23				45,4	42,9										
24				44,5	42,8										
25				44,2	42,7										

Figure 6.2.28 Données brutes de l'IMC par présence de diabète

5. Cliquer sur l'onglet ***Données***. Dans le groupe ***Analyse***, cliquer sur l'onglet ***Utilitaire d'analyse*** (voir la [Figure 6.2.29](#)). Il faut s'assurer que l'outil ***Utilitaire d'analyse*** soit visible (METTRE RÉFÉRENCE ANNEXE).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	R23														
1				Atteint	Non Atteint										
2				67,1	57,3										
3				59,4	52,3										
4				55	47,9										
5				53,2	46,8										
6				52,9	46,7										
7				52,3	46,6										
8				50	46,3										
9				49,7	46,2										
10				49,6	46,1										
11				49,3	45,3										
12				48,8	45,3										
13				48,3	45,3										
14				47,9	45,2										
15				46,8	45										
16				46,2	44,6										
17				46,1	44,5										
18				45,8	44,2										
19				45,7	43,5										
20				45,6	43,5										
21				45,6	43,4										
22				45,5	43,3										
23				45,4	42,9										
24				44,5	42,8										
25				44,2	42,7								

Figure 6.2.29 Sélection de l'onglet ***Utilitaire d'analyse***

6. Sélectionner l'option ***Test d'égalité des espérances: deux observations de variances différentes*** (voir la [Figure 6.2.29](#)).
7. Une boîte de dialogue s'affiche. Il faut entrer la plage de données brutes des deux variables. Pour la variable 1, soit l'indice de masse corporelle des femmes atteintes de diabète, sélectionner la plage de cellules D2:D267. Pour la variable 2, soit l'indice de masse corporelle des femmes non atteintes de diabète, sélectionner la plage de cellules E2:E492 (voir la [Figure 6.2.30](#)).

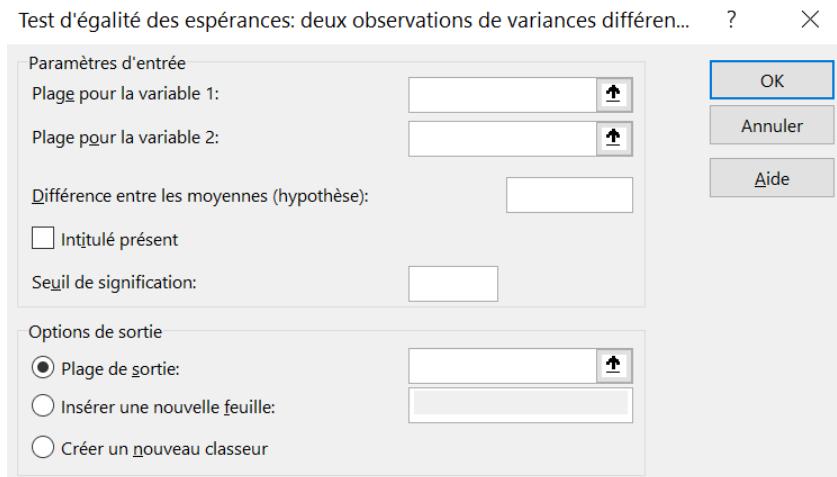


Figure 6.2.30 Utilisation de l'outil *Utilitaire d'analyse*

8. Il y a une zone pour écrire le seuil de signification voulu. Taper $0,01$ (voir la [Figure 6.2.30](#)).
9. Il y a une zone pour indiquer la plage de sortie. Sélectionner la cellule $H2$ ou taper $\$H\2 (voir la [Figure 6.2.30](#)).
10. Cliquer sur **OK**.
11. Le rapport généré s'affiche. Excel utilise la loi de Student pour tout test sur deux moyennes indépendantes, et ce, même si la loi normale peut être appliquée. Le rapport indique la statistique du test (t_{obs}), la cote t critique du test unilatéral et bilatéral, ainsi que la valeur p du test unilatéral et bilatéral (voir la [Figure 6.2.31](#)).

Test d'égalité des espérances: deux observations de variances différentes		
	Atteint	Non Atteint
Moyenne	35,4067669	30,8596741
Variance	43,7579918	43,0432685
Observations	266	491
Différence hypothétique des moyennes	0	
Degré de liberté	540	
Statistique t	9,05500374	
P($T \leq t$) unilatéral	1,2405E-18	
Valeur critique de t (unilatéral)	2,33327322	
P($T \leq t$) bilatéral	2,481E-18	
Valeur critique de t (bilatéral)	2,5849644	

Figure 6.2.31 Rapport généré par l'*Utilitaire d'analyse*

On constate que la statistique du test (t_{obs}) obtenue avec l'utilitaire d'analyse est bien identique à celle obtenue avec le canevas, comme il est attendu. Les valeurs p sont sensiblement les mêmes, proches de 0.

6.2.5 Test d'hypothèse sur deux moyennes dépendantes

Les tests d'hypothèse sur deux moyennes dépendantes, aussi appelés des tests sur des données appariées, permettent de comparer les moyennes de deux séries

de mesures provenant d'un même échantillon. C'est le cas lorsque l'on mesure en paires des observations sur les mêmes unités statistiques. Un tel exemple est l'étude de l'impact de fortes précipitations sur la qualité de l'air, mesurée avant et après l'événement aux mêmes stations météorologiques.

Pour réaliser un test sur deux moyennes dépendantes, il faut s'assurer que les mesures sont bien liées.

Dans cette section, on compare la qualité de l'air le jour précédent et le jour suivant le grand déluge qu'a connu la ville de Montréal le 9 août 2024, en s'appuyant sur les mesures des particules fines (les $PM_{2,5}$). Les mesures proviennent des mêmes onze stations actives dans la région montréalaise. Ces deux séries de données sont bien dépendantes, puisque chaque station constitue une paire de mesures correspondantes. On contrôle les variables qui ne sont pas mesurées dans l'étude en les maintenant constantes. Dans l'étude de la qualité de l'air, l'emplacement des mesures et l'heure de la collecte sont maintenus fixes, avec les mesures prises aux mêmes onze stations à 13h.

6.2.5.1 Tableau croisé dynamique de la quantité de polluant par station

Le tableau des mesures de la qualité d'air contient 933881 observations distinctes. Pour effectuer le test d'hypothèse comparant les mesures des particules fines (les $PM_{2,5}$) des onze stations le jour précédent et suivant le 9 août à 13h, il est nécessaire de créer un tableau croisé dynamique afin d'extraire les paires de mesures requises. Les étapes sont présentées.

1. Ouvrir le fichier Excel **Données_Polluant.xlsm** à l'adresse suivante¹².
2. Dans la feuille **Données_Qualité_Air**, sélectionner le tableau dans son entiereté, soit la plage de cellules A1:E933881.
3. En suivant les étapes présentées à la [Sous sous-section 1.2.3.1](#), attribuer au tableau le nom « **Échantillon** ».
4. En suivant les étapes présentées à la [Sous sous-section 1.2.3.3](#), attribuer des noms à toutes les colonnes du tableau **Échantillon**.
5. En suivant les étapes présentées à la [Sous sous-section 1.2.8.1](#), générer, dans une nouvelle feuille de calcul qu'on renomme **TCD_Polluant**, un tableau croisé dynamique vide à partir du tableau **Échantillon**. Placer ce tableau dans la cellule A3.
6. Glisser et déposer la variable **Station** dans la zone de saisie **Lignes** (voir la [Figure 6.2.32](#)).

¹²github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Polluant.xlsm?raw=true

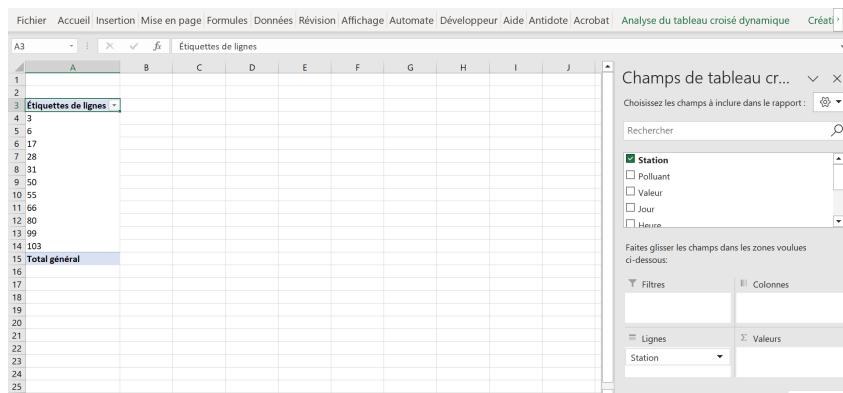


Figure 6.2.32 Tableau croisé dynamique avec la variable *Station* en lignes et la variable *Jour* en colonnes

7. Glisser et déposer la variable *Jour* dans la zone de saisie *Colonnes* (voir la Figure 6.2.32).
8. On ne veut afficher que les valeurs du mois d'août 2024. Cliquer sur la flèche du menu déroulant de l'*Étiquettes de colonnes*. Sélectionner seulement l'année 2024. Cliquer sur *OK* (voir la Figure 6.2.33).

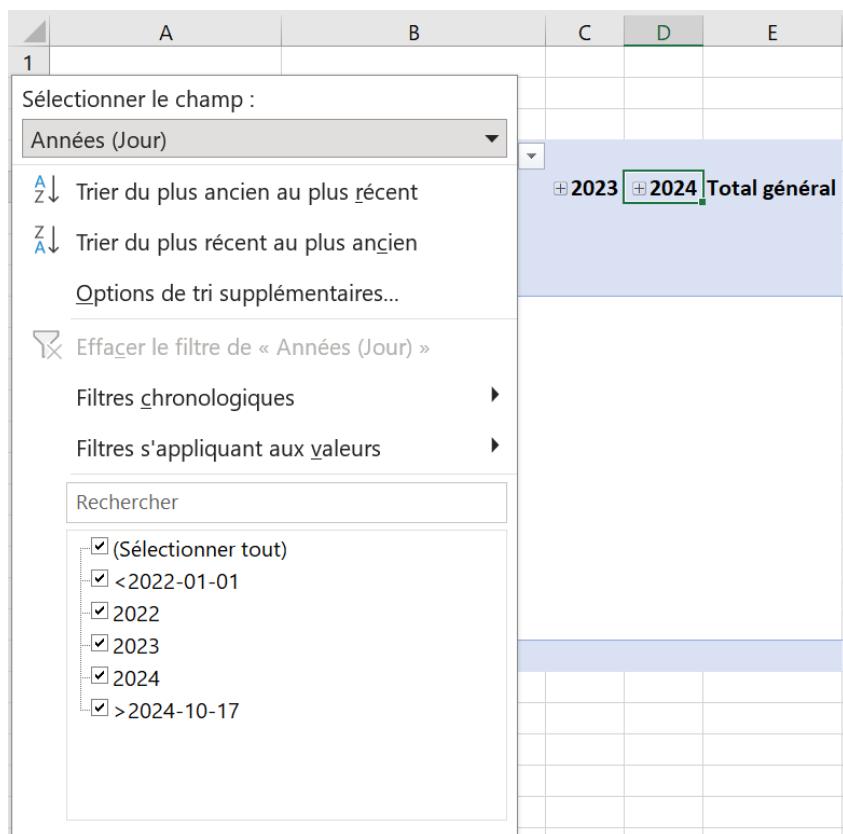


Figure 6.2.33 Filtrage de la variable *Jour* pour afficher l'année 2024

9. Cliquer sur le symbole $[+]$ à côté de l'étiquette 2024 de la cellule B4 pour développer le champ 2024 (voir la Figure 6.2.33 et la Figure 6.2.34 pour le résultat final).

	A	B	C	D	E	F
1						
2						
3						
4		Étiquettes de colonnes				
5		2024				Total général
6		+ Trimestre1	+ Trimestre2	+ Trimestre3	+ Trimestre4	
7	Étiquettes de lignes					
8	3					
9	6					
10	17					
11	28					
12	31					
13	50					
14	55					
15	66					
16	80					
17	99					
18	103					
19	Total général					

Figure 6.2.34 Développer la valeur 2024

10. Sélectionner la cellule B5, soit la cellule où l'étiquette **Trimestre 1** apparaît (voir la [Figure 6.2.35](#)).

	A	B	C	D	E	F
1						
2						
3						
4		Sélectionner le champ :				
5		Trimestres (Jour)				
6		<input type="checkbox"/> Trier du plus ancien au plus récent				
7		<input type="checkbox"/> Trier du plus récent au plus ancien				
8		<input type="checkbox"/> Options de tri supplémentaires...				
9		<input type="checkbox"/> Effacer le filtre de « Trimestres (Jour) »				
10		<input type="checkbox"/> Filtres chronologiques				
11		<input type="checkbox"/> Filtres s'appliquant aux valeurs				
12		Rechercher				
13		<input checked="" type="checkbox"/> (Sélectionner tout)				
14		<input checked="" type="checkbox"/> <2022-01-01				
15		<input checked="" type="checkbox"/> Trimestre1				
16		<input checked="" type="checkbox"/> Trimestre2				
17		<input checked="" type="checkbox"/> Trimestre3				
18		<input checked="" type="checkbox"/> Trimestre4				
19		<input checked="" type="checkbox"/> >2024-10-17				

Figure 6.2.35 Filtrage de la variable **Jour** pour n'afficher que le mois d'août 2024

11. Cliquer à nouveau sur la flèche du menu déroulant de l'**Étiquettes de colonnes** (voir la [Figure 6.2.35](#)).
12. Ne cocher que la case pour **Trimestre 3** (voir la [Figure 6.2.35](#)).
13. Cliquer sur le symbole à côté de l'étiquette **Trimestre 3** de la cellule B5 (voir la [Figure 6.2.35](#)).
14. Sélectionner la cellule B6, soit la cellule où l'étiquette **Juillet** apparaît (voir la [Figure 6.2.35](#)).

15. Cliquer à nouveau sur la flèche du menu déroulant de l'**Étiquettes de colonnes** (voir la [Figure 6.2.35](#)).
16. Ne cocher que la case pour **Août** (voir la [Figure 6.2.35](#)).
17. Cliquer sur le symbole **[+]** à côté de l'étiquette **Août** de la cellule B7 (voir la [Figure 6.2.36](#) pour le résultat final).

3		Etiquettes de colonnes	▼							
4		2024								
5		Trimestre3								
6		août								
7	Etiquettes de lignes	▼	2024-08-01	2024-08-02	2024-08-03	2024-08-04	2024-08-05	2024-08-06	2024-08-07	2024-08-08
8	3									
9	6									
10	17									
11	28									
12	31									
13	50									
14	55									
15	66									
16	80									
17	99									
18	103									
19	Total général									

Figure 6.2.36 Développer la catégorie *août*

18. On ne veut que le polluant **PM25**. Il faut donc imposer un filtre. Glisser et déposer la variable **Polluant** dans la zone de saisie **Filtres**. Un filtre apparaît dans les cellules A1:B1 (voir la [Figure 6.2.37](#)).

Figure 6.2.37 Ajout de filtres pour le type de polluant et l'heure de la journée et insertion des valeurs de *PM_{2,5}*

19. Cliquer sur la flèche du menu déroulant de la cellule B1. Sélectionner **PM** (voir la [Figure 6.2.37](#)).
20. On souhaite choisir une prise de mesure à une heure précise, soit 13h. Glisser et déposer la variable **Heure** dans la zone de saisie **Filtres**. Un filtre apparaît dans les cellules A2:B2 (voir la [Figure 6.2.37](#)).
21. Cliquer sur la flèche du menu déroulant de la cellule B2. Sélectionner **13** pour 13h (voir la [Figure 6.2.37](#)).
22. Glisser et déposer la variable **Valeur** dans la zone de saisie **Valeurs**. Puisqu'il n'y a qu'une mesure de *PM_{2,5}* par heure par station, on conserve le calcul de **Somme de Valeurs** (voir la [Figure 6.2.37](#)). On ne veut pas le compte.
23. Au final, on ne veut que les valeurs du 8 août et du 10 août (voir la [Figure 6.2.38](#)).

A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1 Polluant	PM													
2 Heure	13													
3														
4 Somme de Valeur	Étiquettes de colonnes													
5	2024													
6	-Trimestre3													
7	août													
8 Étiquettes de lignes		2024-08-01	2024-08-02	2024-08-03	2024-08-04	2024-08-05	2024-08-06	2024-08-07	2024-08-08	2024-08-09	2024-08-10	2024-08-11	2024-08-12	2024-08-13
9 8		12	27	24	20	11	8	6	9	3	7	5	5	23
10 6		13	30	24	22	13	14	7	10	3	8	6	6	24
11 17		17	34	29	23	14	12	7	12	4	8	6	6	22
12 28		13	31	21	22	16	13	8	11	3	11	7	15	23
13 31		15	31	25	24	14	15	10	13	5	11	6	7	
14 50		15	30	24	24	13	11	6	12	4	9	5	6	21
15 55		13	31	26	22	13	8	6	11	4	7	5	5	24
16 66		13	32	22	23	14	11	7	12	4	10	5	5	18
17 80		13	29	22	22	13	12	7	10	3	10	5	6	21
18 99		12	30	20	23	13	8	7	11	8	5	5	5	20
19 103		13	29	22	22	15	12	10	3	9	5	6	5	19
20 Total général		149	334	259	247	149	124	71	121	36	98	60	67	215

Figure 6.2.38 Sélection des colonnes du 8 août et du 10 août 2024

24. Copier et coller les deux colonnes en dessous du tableau croisé dynamique. Créer un tableau respectant toute la mise en forme vue à la [Sous-section 2.2.2](#) (voir la [Figure 6.2.39](#)).

Répartition des onze stations montréalaises selon la quantité de PM25, Montréal, août 2024			
Station	PM25 8 août	PM25 10 août	Différence
3	9	7	-2
6	10	8	-2
17	12	8	-4
28	11	11	0
31	13	11	-2
50	12	9	-3
55	11	7	-4
66	12	10	-2
80	10	10	0
99	11	8	-3
103	10	9	-1

Figure 6.2.39 Tableau final des mesures pairees des particules fines avant et après le 9 août

25. Donner au tableau créé à la [Figure 6.2.39](#) le nom **DonnéesTest** (voir la [Sous-section 1.2.3](#)).
26. Sélectionner le tableau **DonnéesTest** et générer des noms pour les quatre colonnes qui le composent (voir la [Sous sous-section 1.2.3.3](#)).

Copier le cavenas 2 moyennes dépendantes. Avant d'effectuer le test sur deux moyennes dépendantes, il faut copier le canevas **2 moyennes dépendantes** dans le classeur **Données_Polluant**.

- Ouvrir le fichier **Canevas_tests_hypotheses.xlsxm** à l'adresse suivante¹³.
- Effectuer la procédure présentée au laboratoire 4 pour copier la feuille **2 moyennes dépendantes** du fichier **Canevas_tests_hypotheses.xlsxm** au classeur **Données_Polluant.xlsxm**.

¹³github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Canevas/Canevas_tests_hypotheses.xlsxm?raw=true

6.2.5.2 Étapes d'un test sur deux moyennes dépendantes

Les étapes pour réaliser un test d'hypothèse sur deux moyennes dépendantes sont présentées.

1. La variable X est définie comme les mesures de particules fines du 8 août 2024 et Y , les mesures du 10 août 2024. Soit $D = Y - X$, la différence. La quatrième colonne de la [Figure 6.2.39](#) présente le calcul de la variable aléatoire D .
2. La première étape d'un test consiste à choisir le type de test que l'on souhaite réaliser, à savoir un test unilatéral à gauche, à droite ou bilatéral. Puisque l'impact d'intenses précipitations n'est pas clair (voir la [Section 6.1](#)), il est prudent d'être conservateur et de choisir un test bilatéral. Dans la cellule D6 de la feuille **2 moyennes dépendantes** du classeur **Données_Polluant.xlsx**, il faut donc choisir le symbole \neq (voir la [Figure 6.2.40](#)).

Test d'hypothèse sur deux moyennes dépendantes (sur des données appariées)								
1) Hypothèses : $H_0 : \mu_D = 0$ $H_1 : \mu_D \neq 0$ Sélectionner le symbole voulu. Effacer le contenu des cellules grises								
2) Seuil et informations : $\alpha =$ $N =$ $\bar{x}_D =$ $n =$ $s_D =$								
1	2	3	4	5	6	7	8	9
11	12	13						

Figure 6.2.40 Remplissage des cellules pour la première étape d'un test sur deux moyennes dépendantes

3. Dans la deuxième étape du test, il faut remplir les cases grisées avec les informations demandées.
 - Choisir un seuil de 0,01. Dans la cellule C9, taper $=0,01$ (voir la [Figure 6.2.41](#)).
 - Au moment de l'enregistrement du fichier, le nombre total de données était de 933881 observations. Prendre ce nombre comme la taille de la population. Taper « 933881 » (voir la [Figure 6.2.41](#)).

Test d'hypothèse sur deux moyennes dépendantes (sur des données appariées)								
1) Hypothèses : $H_0 : \mu_D = 0$ $H_1 : \mu_D \neq 0$ Sélectionner le symbole voulu. Effacer le contenu des cellules grises								
2) Seuil et informations : $\alpha =$ $N =$ $\bar{x}_D =$ $n =$ $s_D =$								
1	2	3	4	5	6	7	8	9
14	15	16	17	18				

Figure 6.2.41 Remplissage du seuil de signification α et des mesures descriptives

- Dans la cellule F9, taper $=MOYENNE(Différence)$, *Différence* étant le nom de la quatrième colonne de la [Figure 6.2.39](#). Appuyer sur la touche **Enter** du clavier (voir la [Figure 6.2.41](#)).

- Dans la cellule F10, taper =NB(Différence) pour calculer automatiquement le nombre de mesures pairées. Appuyer sur la touche **Enter** (voir la Figure 6.2.41).
 - Dans la cellule F11, taper =ECARTYPE.STANDARD(Différence) pour calculer l'écart type des mesures. Appuyer sur la touche **Enter** (voir la Figure 6.2.41).
 - Une fois les plages F9:F11 remplies, il est possible de constater que les cellules C20, D27, G29 et F32 se remplissent automatiquement, les cellules étant préremplies avec les formules appropriées.
4. À la troisième étape, il faut vérifier les conditions d'application du test. Puisque l'on effectue le test avec un petit échantillon (inférieur à 30), il faut supposer que les différences sont normalement distribuées pour poursuivre le test. On utilise la loi de Student. Écrire un texte dans la cellule fusionnée C16 expliquant ce choix (voir la Figure 6.2.42).

14	3) Conditions d'application :	Loi : Student
15		
16		car n<30 et on suppose que les différences sont normalement distribuées
17		
18		Justifier la loi choisie.
19		

Figure 6.2.42 Choix de la loi à utiliser pour le test

5. L'écart type de la distribution D , soit $\sigma_{\bar{D}_1}$, est calculé à l'aide de formule se trouvant dans la cellule C20 (voir la Figure 6.2.43).

20	4) Écart type $\sigma_{\bar{D}}$:	$\sigma_{\bar{D}} =$	0,41460925
21			
22			
23			
24	5) Règle de décision :		
25			
26			
27	Méthode 1 (statistique du test) :	=	3,16927267
28		t_{α/2} to to to/z_{α/2}	
29		s est que 3,16927267	
30		appropriée pour le test	
31			
32	Méthode 2 (valeur p) :	On rejette H_0 si la valeur p est inférieure ou égale à	0,01
33			
34			
35			

Figure 6.2.43 Remplissage des cellules pour les étapes 4 et 5

6. À la cinquième étape, il faut écrire la règle de décision, soit avec le calcul de la statistique du test ou avec la valeur p. Dans la cellule B27, sélectionner la statistique appropriée du test bilatéral, soit $t_{\alpha/2}$. Remarquer que la valeur de $t_{\alpha/2}$ (environ 3,17) dans les cellules D27 et G29 est préremplie et correspond bien à la cote t pour un seuil de signification de 1% et un degré de liberté de 10 (voir la Figure 6.2.43).
7. Dans la cellule E29, sélectionner les mots appropriés parmi les trois choix du menu déroulant. Puisque l'on effectue un test bilatéral, il faut choisir les mots « est différente de » (voir la Figure 6.2.43).
8. À la sixième étape, toutes les cases roses se remplissent automatiquement selon les informations fournies précédemment (voir la Figure 6.2.44). On observe que la cote $|t_{obs}|$ est grande avec une valeur d'environ 5,04. Le rejet de l'hypothèse nulle est presque assuré. De plus, pour renchérir sur ce point, la valeur p calculée par Excel est d'environ 0,0005.

9. Dans la cellule A45, écrire la décision et la conclusion du test en fonction de la règle de décision (voir la Figure 6.2.44).

A	B	C	D	E	F	G
36 6) Calcul de la statistique ou la valeur p :						
37						
38 Méthode 1 (cote z ou t observée) :	tobs	=	5,04308361			
39						
40 Méthode 2 (valeur p) :	valeur p	=	0,00050434			
41						
42						
43 7) Décision et conclusion :						
44						
45						
46	Puisque la valeur absolue de la cote t observée est supérieure à la cote t critique, on rejette l'hypothèse nulle H0. Au seuil de 1%, on					
47	peut conclure qu'il y a eu une différence significative dans la qualité de l'air avant et après la forte accumulation de précipitations du					
48	9 août.					

Figure 6.2.44 Décision et conclusion du test

Il serait intéressant de vérifier si cette conclusion s'applique à tous les types de polluants et non seulement aux particules fines.

Il est possible d'utiliser l'utilitaire d'analyse, comme présenté à la [Sous-sous-section 6.2.4.2](#). Il suffit de sélectionner l'option **Test d'égalité des espérances: observations pairées**, l'option pour les tests d'hypothèses sur deux moyennes dépendantes.

6.3 Réflexions

Ce laboratoire aura permis d'explorer les concepts liés aux tests d'hypothèses paramétriques. Toutefois, une étude plus approfondie est utile pour en saisir toutes les subtilités.

Travail à faire après le laboratoire

Objectifs

- Poser un regard critique sur des données.
 - Formuler des hypothèses de recherche.
 - Choisir le type de tests adéquatement et comprendre les distinctions entre chaque type de tests.
 - Effectuer un test d'hypothèse paramétriques pour vérifier une hypothèse.
 - Interpréter adéquatement la conclusion d'un test.
 - Comprendre les effets d'un changement de seuil de signification.
 - Explorer les caractéristiques des erreurs de type I et de type II.
1. Avant de faire une collecte de données, un chercheur pose ses hypothèses de recherche. Ce dernier effectue une étude sur l'efficacité de certains désinfectants sur la croissance de bactéries. Ce dernier croit que le désinfectant commercial Lysol sera plus efficace que le vinaigre contre la croissance bactérienne. Une fois les données récoltées, le chercheur constate que la moyenne du diamètre d'inhibition des bactéries avec Lysol est plus petite que la moyenne avec le vinaigre. Il décide de manipuler ses données pour aller dans le sens de son hypothèse de recherche de départ. Discuter des impacts de ce choix dans ce contexte et dans d'autres contextes. Que devrait faire le chercheur avant de réaliser son test d'hypothèse?
 2. Dans le test sur deux moyennes dépendantes de la [Sous-section 6.2.5](#), à partir de quelle valeur du seuil de signification la conclusion du test change-t-elle? Donner la valeur au centième près. Expliquer en considérant les valeurs des statistiques et de la valeur p.
 3. Un chercheur peut commettre deux types d'erreurs lorsque vient le temps de conclure: une erreur de type I (un faux positif), c'est-à-dire rejeter H_0 alors que H_0 est vraie, ou une erreur de type II (un faux négatif), soit ne pas rejeter H_0 alors que H_0 est faux. Supposer qu'un professionnel de la santé veut tester si un patient a une maladie. Dans quel contexte est-ce que l'erreur de type I (faux positif) est plus grave? Dans quel contexte est-ce que l'erreur de type II (faux négatif) est plus grave? Donner des exemples concrets.
 4. Parmi toutes les bases de données disponibles (vin, armée, collisions, etc.), faire deux tests d'hypothèses paramétriques de type différent en utilisant les canevas. Les hypothèses doivent être appuyées par une source crédible et rigoureuse.
 5. Pour trois des quatre tests effectués dans le laboratoire 6, les cotes z ou t observées sont loin des valeurs critiques et les valeurs p sont très petites. Donner une ou deux raisons qui pourraient expliquer la puissance de ces rejets de l'hypothèse nulle. Les feuilles des fichiers Excel remis doivent être bien identifiées.

Chapitre 7

Tests du khi-deux

Il existe des types de tests qui n'impliquent pas de paramètres comme la moyenne et la proportion. On les appelle des tests non paramétriques. Ce chapitre présente plus particulièrement les tests d'indépendance du khi-deux. Ces derniers permettent de déterminer s'il existe un lien entre deux variables à partir d'un échantillon. Les variables peuvent être qualitatives ou quantitatives.

7.1 Prélab

Les tests d'ajustement du khi-deux permettent de déterminer si une variable se répartit selon une certaine distribution théorique.

Les tests d'indépendance du khi-deux permettent de déterminer l'existence ou l'absence d'un lien entre deux variables. Ces dernières peuvent être qualitatives ou quantitatives, mais elles doivent être à échelle nominale ou ordinaire pour pouvoir effectuer le test.

Dans ce prochain laboratoire, la base nationale des collisions automobiles de la [Section A.4](#) sera analysée.

7.1.1 Travail à faire avant le cours

Objectifs

- Examiner les séries statistiques.
 - Effectuer une revue de la littérature.
 - Poser un regard critique sur les données.
 - Formuler des hypothèses de recherche.
 - Explorer le lien entre deux variables.
1. Dans un test d'ajustement du khi-deux, l'hypothèse nulle postule que la variable se distribue selon un certain modèle théorique, alors que l'hypothèse alternative affirme que la variable ne se distribue pas selon ce modèle théorique. Le ministère de la Santé et des Services Sociaux du Québec veut dresser un portrait de la situation des urgences. Il s'intéresse à la répartition des personnes en attente d'un médecin à l'urgence en fonction des jours de la semaine. Il croit qu'il y a proportionnellement moins de personnes qui se présentent aux urgences en début de semaine (du lundi au jeudi) qu'en fin de semaine (du vendredi au dimanche). Quelles seraient l'hypothèse nulle et l'hypothèse alternative de ce test d'ajustement du khi-deux? Expliquer les choix.
 2. Dans un test d'indépendance, l'hypothèse nulle postule l'indépendance entre deux variables, alors que l'hypothèse alternative affirme une dépendance. Expliquer pourquoi ça ne peut pas être l'inverse. Utiliser les concepts de probabilité pour le faire.
 3. Ouvrir la base de données présentée à la [Section A.4](#). En suivant les étapes présentées à la [Sous sous-section 1.2.3.1](#), attribuer au tableau le nom « Échantillon ». En suivant les étapes présentées à la [Sous sous-section 1.2.3.3](#), attribuer des noms aux deux colonnes du tableau **Échantillon**.
 4. Dans la base de données présentée à la [Section A.4](#), explorer les variables. Déterminer quelques paires de variables qui pourraient fort probablement avoir un lien entre elles. Déterminer quelques paires de variables dont une dépendance est moins claire et mériterait d'être vérifiée.

7.2 Laboratoire

Dans ce laboratoire, l'objectif est de poursuivre l'analyse d'une base de données en vérifiant s'il existe des liens entre deux variables à échelle nominale ou ordinale. Est-ce que des différences observées au niveau d'un échantillon sont assez significatives pour être généralisées à la population? Ceci se fait à l'aide d'un test d'indépendance du khi-deux. Les étapes d'un tel test sont présentées dans ce laboratoire.

On souhaite étudier la base nationale des collisions automobiles de 2019. Plus spécifiquement, on s'intéresse aux liens possibles entre les variables présentées à la [Section A.4](#). Comme dans le laboratoire 6 sur les tests d'hypothèse paramétriques, un canevas sera utilisé pour les tests d'indépendance du khi-deux.

Deux tests d'indépendance sont présentés. Un premier vérifiant s'il existe un lien ou non entre la gravité d'une collision et l'âge de la personne conductrice impliquée dans la collision; et un second vérifiant s'il existe un lien entre la gravité d'une collision et le type de mesures de sécurité. La gravité d'une collision est une variable qualitative nominale avec quatre modalités : une collision provoquant au moins une perte de vie (code 1 dans le fichier Excel), une collision provoquant une blessure non mortelle mais pas de perte de vie (code 2), gravité inconnue (code U) et la juridiction ne fournit pas cet élément d'information (code X). Le type de dispositifs de sécurité est une variable qualitative nominale avec plusieurs modalités dont aucun dispositif de sécurité (codes 1, 12 ou 13 selon le type de véhicule), des ceintures (code 2), des sièges de bébé (codes 5 ou 6), des ports de casque (code 9), etc.

Puisque la base de données représente l'ensemble de toutes les collisions de l'année 2019, on sélectionne un échantillon aléatoire pour faire les tests d'indépendance, soit les données du mois de décembre.

7.2.1 Lien entre la gravité d'un accident et l'âge d'un conducteur

Les étapes pour réaliser un test d'indépendance du khi-deux entre une variable qualitative et une variable quantitative sont présentées. On souhaite vérifier s'il existe un lien entre la gravité d'un accident et l'âge de la personne conductrice. On choisit un seuil de signification de 5%.

Copier le cavenas *Test d'indépendance*. Pour effectuer un test d'indépendance, il faut copier le canevas dans le fichier de travail.

1. Ouvrir le fichier Excel **Données_Collisions.xlsx** à l'adresse [suivante](#)¹ et le ficher > à l'adresse [suivante](#)².
2. Effectuer la procédure présentée au laboratoire 4 pour copier la feuille **Test d'indépendance** du fichier **Canevas_tests_independance** au classeur **Données_Collisions**.

7.2.1.1 Étapes d'un test d'indépendance avec au moins une variable quantitative

Lorsque l'on fait un test d'indépendance avec une variable quantitative continue, l'étape de la construction du tableau des effectifs observés peut être laborieuse. Il faut grouper les valeurs du tableau croisé dynamique généré dans des classes. De plus, il faut s'assurer que les effectifs théoriques sont assez grands.

¹github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Collisions.xlsx?raw=true

²github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Canevas/Canevas_tests_independance.xlsx?raw=true

1. Dans le classeur **Données_Collisions**, sélectionner la feuille **Test d'indépendance**. Renommer cette feuille « Test Khi-Deux Age»
2. La première étape d'un test d'indépendance consiste à définir ses variables et écrire les hypothèses nulle et alternative. Dans la cellule C4, taper « La gravité d'une collision ». Dans La cellule C5, taper « L'âge d'une personne conductrice » (voir la [Figure 7.2.1](#)).

Pour tout test d'indépendance, l'hypothèse nulle H_0 est que les deux variables sont indépendantes. L'hypothèse alternative H_1 est son contraire, soit que les deux variables sont dépendantes.

Dans l'encadré gris de la cellule D8, vis-à-vis H_0 , taper « La gravité d'une collision et l'âge d'une personne conductrice sont indépendants» (voir la [Figure 7.2.1](#)). Dans l'encadré gris de la cellule D11, vis-à-vis H_1 , taper « La gravité d'une collision et l'âge d'une personne conductrice sont dépendants» (voir la [Figure 7.2.1](#))

A	B	C	D	E	F	G	H	I	J
1									
2									
3									
4				X:	La gravité d'une collision				
5				Y:					
6									
7									
8	1) Hypothèses :			H ₀ :					
9									
10									
11				H ₁ :					
12									
13									

Figure 7.2.1 Remplissage des encadrés de la première étape d'un test d'indépendance entre la gravité d'une collision et l'âge d'une personne conductrice

3. La deuxième étape consiste à construire le tableau des effectifs observés et le tableau des effectifs théoriques.

- Il faut commencer par générer un tableau croisé dynamique des effectifs observés. Dans la feuille **Données_Collisions**, sélectionner le tableau « Données ».
- En suivant les étapes présentées à la [Sous sous-section 1.2.8.1](#), générer, dans une nouvelle feuille de calcul qu'on renomme **TCD_GraviteAge**, un tableau croisé dynamique vide à partir du tableau **Données**. Placer ce tableau dans la cellule A3 (voir la [Figure 7.2.2](#)).

Figure 7.2.2 Générer un tableau croisé dynamique vide

- Glisser et déposer la variable **Âge** dans la zone de saisie **Lignes** (voir la [Figure 7.2.3](#)).

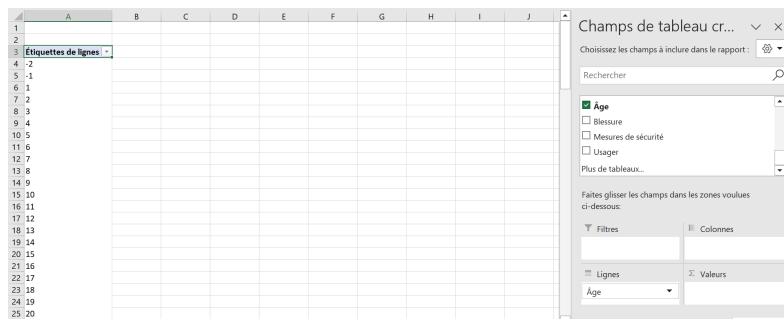


Figure 7.2.3 Tableau croisé dynamique avec la variable *Âge* en lignes et la variable *Gravité* en colonnes

- Glisser et déposer la variable *Gravité* dans la zone de saisie *Colonnes* (voir la Figure 7.2.3).
- On veut maintenant grouper les valeurs de la variable *Âge* de 0 à 100 ans avec une amplitude de 10 ans. Il est important de se fier à la table de Sturges lorsque l'on fixe une amplitude. Cependant, on ne veut pas non plus avoir trop de classes avec peu d'effectifs observés pouvant mener au non-respect de la condition d'application du test d'indépendance de khi-deux.

Dans la colonne des étiquettes de lignes de la variable *Âge*, cliquer avec le bouton de droite de la souris une des valeurs de l'âge. Il importe peu laquelle. Sélectionner l'option *Grouper* (voir la Figure 7.2.4). Une boîte de dialogue s'affiche

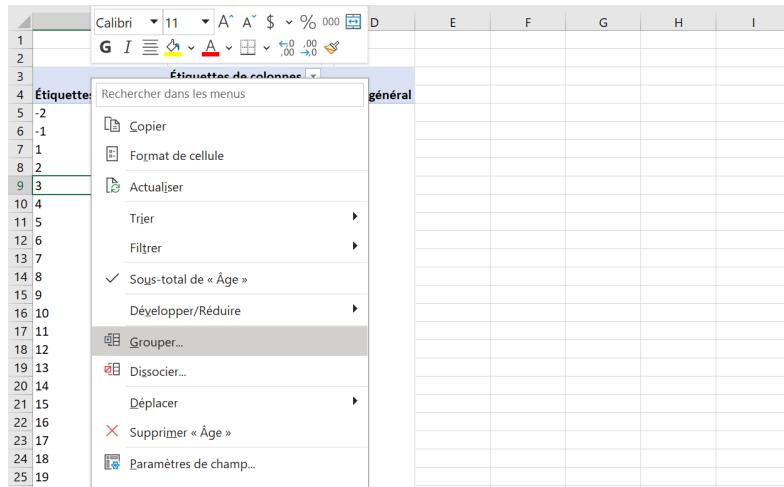


Figure 7.2.4 Grouper en classes la variable *Âge*

- Taper 0 comme valeur de début et 10 comme amplitude (voir la Figure 7.2.4). On choisit 0 comme valeur initiale, car on ne veut pas tenir compte des valeurs inférieures à 0, c'est-à-dire les modalités -2 et -1 qui indiquent des informations inconnues par rapport à l'âge des personnes. Une fois le regroupement fait, on peut filtrer les valeurs inférieures à 0.
- Pour filtrer les valeurs inférieures à 0, cliquer sur le petit triangle du menu déroulant de l'étiquette de lignes (voir la Figure 7.2.4).
- Sélectionner l'option *Filtres s'appliquant aux étiquettes*, suivie de l'option *Supérieur ou égal à* (voir la Figure 7.2.4).

- Une boîte de dialogue s'affiche. Dans la zone de saisie à droite l'option *est supérieur ou égal à*, taper 0 (voir la Figure 7.2.4). Cliquer sur *OK*.
À ce stade, le tableau croisé dynamique devrait ressembler à la Figure 7.2.5

		Étiquettes de colonnes	
		1 2 Total général	
Étiquettes de lignes		0-9	10-19
0-9		0	0
10-19		0	0
20-29		0	0
30-39		0	0
40-49		0	0
50-59		0	0
60-69		0	0
70-79		0	0
80-89		0	0
90-99		0	0
Total général		0	0

Figure 7.2.5 Tableau croisé dynamique avec les variables *Âge* et *Gravité*

- Glisser et déposer la variable *Âge* dans la zone de saisie *Valeurs* (voir la Figure 7.2.6). S'assurer que le calcul fait est bien le compte (*Nombre*) et non la somme. Si c'est la somme, changer pour *Nombre* avec les étapes vues à la Sous sous-section 1.2.8.2.

1 Mois	B	C	D	E	F	G	H	I
2	(Tous)							
3 Nombre d'Age	Étiquettes de colonnes:	1	2 Total général					
4	Étiquettes de lignes:	1	157	11414	11571			
5 0-9		405	27084	27491				
6 10-19		790	53145	53935				
7 20-29		622	43036	43658				
8 30-39		554	37004	37500				
9 40-49		634	35734	36358				
10 50-59		502	23901	24403				
11 60-69		303	12956	13259				
12 70-79		143	4946	5087				
13 80-89		32	581	613				
14 90-99		4134	250141	254275				
15 Total général								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								

Figure 7.2.6 Remplissage du tableau croisé dynamique des effectifs observés par *Âge* et *Gravité d'accident* et filtrage pour la période et l'usager voulu

- On ne veut inclure que les données du mois de décembre (code 12) et les usagers qui sont des conducteurs (code 1). On appliquer des filtres pour faire cela.

Glisser et déposer la variable *Mois* dans la zone de saisie *Filtres*. Le filtre apparaît dans les cellules A1:B1 (voir la Figure 7.2.6). Cliquer sur la flèche du menu déroulant de la cellule B1. Sélectionner 12 pour le mois de décembre (voir la Figure 7.2.6). Cliquer sur *Ok*.

Glisser et déposer la variable *Usager* dans la zone de saisie *Filtres*. Le filtre apparaît dans les cellules A2:B2 (voir la Figure 7.2.6). Cliquer sur la flèche du menu déroulant de la cellule B2. Sélectionner 1 pour le conducteur (voir la Figure 7.2.6). Cliquer sur *Ok*.

À ce stade, le tableau croisé dynamique devrait ressembler à la Figure 7.2.7.

Nombre de Âge Étiquettes de lignes	Étiquettes de colonnes	Total général		
		1	2	Total général
0-9			2	2
10-19		11	902	913
20-29		40	3203	3243
30-39		44	2857	2901
40-49		54	2497	2551
50-59		44	2338	2382
60-69		29	1576	1605
70-79		13	795	808
80-89		5	281	286
90-99		1	30	31
Total général		241	14481	14722

Figure 7.2.7 Tableau croisé dynamique avec les variables *Âge* et *Gravité* filtrées

Remarque 7.2.8 Valeurs aberrantes? Selon le tableau de la Figure 7.2.7, deux conducteurs auraient entre 0 et 10 ans. En creusant la base de données, il est fort probable qu'il y a eu une erreur de saisie. Les données seront tout de même conservées, puisque deux unités statistiques sur 14722 ne feront pas une grande différence.

Remarque 7.2.9 Deux classes avec peu de valeurs. La première et la dernière classes ont peu d'individus. Il est fort probable que si on gardait les regroupements tels quels, les effectifs théoriques de ces classes seraient inférieurs à 5. Ainsi, on va grouper les deux premières classes et les deux dernières classes pour s'assurer que la condition d'application du test soit respectée.

- Pour regrouper des classes, le processus est un peu contre-intuitif. Il faut indiquer à Excel les classes qui ne seront pas regroupées. Dans ce cas, ce sont toutes les classes entre 20 ans et 79 ans. De cette façon, Excel regroupera en une classe les données avant la classe débutant à 20 ans et en une autre classe les données après la classe terminant à 79 ans.

Cliquer avec le bouton de droite sur n'importe quelle classe d'âge de la première colonne du tableau croisé dynamique. Sélectionner l'option *Grouper* (voir la Figure 7.2.10).

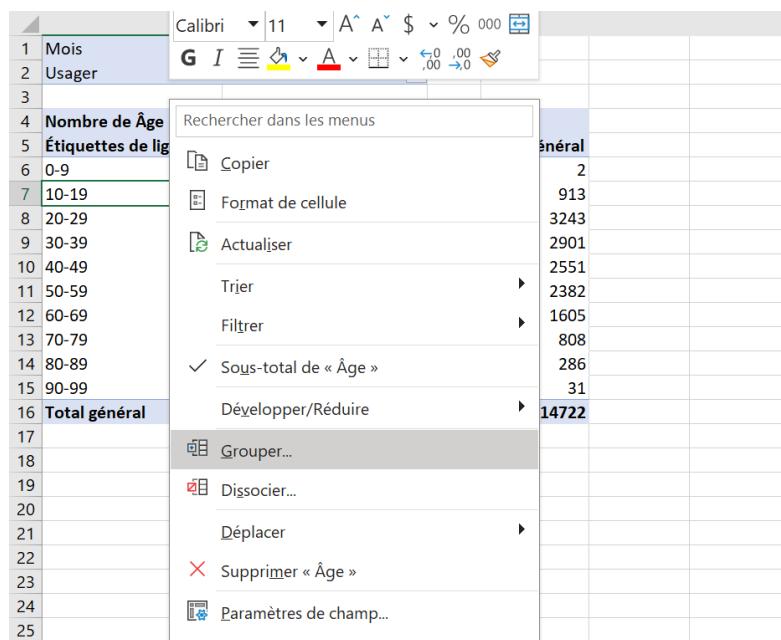


Figure 7.2.10 Groupage des deux premières et des deux dernières classes

Taper 20 comme début et 79 comme fin (voir la [Figure 7.2.10](#)). Le tableau croisé dynamique des effectifs observés est présenté à la [Figure 7.2.11](#).

Nombre de Âge	Étiquettes de colonnes ▾		
	1	2	Total général
Étiquettes de lignes ▾			
<20	11	904	915
20-29	40	3203	3243
30-39	44	2857	2901
40-49	54	2497	2551
50-59	44	2338	2382
60-69	29	1576	1605
70-79	13	795	808
>80	6	311	317
Total général	241	14481	14722

Figure 7.2.11 Tableau croisé dynamique des effectifs observés

On va maintenant coller une partie de ce tableau dans la feuille *Test Khi-Deux Age* et faire sa mise en forme.

- Dans la feuille *Test Khi-Deux Age*, cliquer sur la cellule B20 et taper le symbole $=$ (voir la [Figure 7.2.12](#)).

	A	B
12		
13		
14		
15	2) Tableaux des effectifs observés et théoriques :	
16		
17		
18		
19		
20	=	
21		
22		
23		
24		
25		
26		
27		
28		
29		
30		
31		
32		
33		
34	Condition d'application respectée?	
35		

Données_Collisions TCD_GraviteAge **Test Khi-Deux Age** +

Figure 7.2.12 Copiage d'une partie du tableau croisé dynamique dans le canevas **Test Khi-Deux Age**

- Cliquer sur l'onglet de la feuille **TCD_GraviteAge** et sélectionner la plage A6:D14 (voir la [Figure 7.2.12](#)).
- Taper la touche **Enter** du clavier. Le tableau suivant apparaît dans la feuille **Test Khi-Deux Age** (voir la [Figure 7.2.13](#)).

15	2) Tableaux des effectifs observés et théoriques :			
16				
17				
18				
19				
20	<20	11	904	915
21	20-29	40	3203	3243
22	30-39	44	2857	2901
23	40-49	54	2497	2551
24	50-59	44	2338	2382
25	60-69	29	1576	1605
26	70-79	13	795	808
27	>80	6	311	317
28	Total général	241	14481	14722

Figure 7.2.13 Tableau des effectifs observés

- Faire une mise en forme élémentaire du tableau. Le tableau final des effectifs observés doit ressembler au tableau de la [Figure 7.2.14](#). Les couleurs et bordures importent peu ici. Si ceci était un tableau de présentation pour un rapport final, il faudrait peaufiner la mise en page ainsi que l'écriture des classes d'âge. Par exemple, au lieu de voir « < 20 », il faudrait voir « Moins de 20 ans ».

Âge	Gravité de la collision		Total
	Au moins une perte de vie	Blessure non mortelle	
<20	11	904	915
20-29	40	3203	3243
30-39	44	2857	2901
40-49	54	2497	2551
50-59	44	2338	2382
60-69	29	1576	1605
70-79	13	795	808
>80	6	311	317
Total général	241	14481	14722

Figure 7.2.14 Tableau final des effectifs observés

- Il faut maintenant créer le tableau des effectifs théoriques. Dans la cellule G17, taper « Tableau des effectifs théoriques » (voir la [Figure 7.2.15](#)).

A	B	C	D	E	F	G	H	I	J	K
2) Tableaux des effectifs observés et théoriques :										
Tableau des effectifs observés										
Tableau des effectifs théoriques										

Figure 7.2.15 Crédit du tableau des effectifs théoriques

- Copier la plage de cellules A18:D28, soit le tableau des effectifs observés. Faire un collage spécial en ne collant que les valeurs dans la cellule G18 (voir la [Figure 7.2.15](#)).
- Sélectionner la plage de cellules H20:I27 et la supprimer en appuyant sur la touche **[Suppr]** du clavier (voir la [Figure 7.2.15](#)).
- Formatter les titres au besoin. Ce n'est cependant pas vraiment nécessaire.
- La formule pour calculer un effectif théorique est le total de la colonne multiplié par le total de la ligne, le tout divisé par la taille de l'échantillon. Dans la cellule H20, taper =H\$28*\$J20/\$J\$28 (voir la [Figure 7.2.16](#)).

A	B	C	D	E	F	G	H	I	J
2) Tableaux des effectifs observés et théoriques :									
Tableau des effectifs observés									
Tableau des effectifs théoriques									

Figure 7.2.16 Calcul des effectifs théoriques

Le symbole **[\$]** devant le nombre 28 permet de fixer la ligne à 28 lorsque la formule sera recopiée dans les cellules avoisinantes. Ainsi, le total de la ligne sera toujours celui de la ligne 28. Le symbole **[\$]** devant la lettre J fige la valeur du total de la colonne.

- Sélectionner la cellule H20. Placer le curseur dans le coin inférieur droit jusqu'à ce que la croix noire de recopie (**[+]**) apparaisse. Double-cliquer sur le coin inférieur droit. La formule se recopie jusqu'à la cellule H27 (voir la [Figure 7.2.16](#)). Il est également possible de garder enfonce le curseur de la souris et de glisser la formule vers le bas.

- La plage de cellules H20:I27 devrait toujours être sélectionnée. Placer le curseur dans le coin inférieur droit de la cellule H27 jusqu'à ce que la croix noire de recopie (+) apparaisse. Cliquer avec le bouton de gauche de la souris en gardant le bouton enfoncé et glisser le curseur vers la droite pour remplir les cellules I20:I27 (voir la Figure 7.2.16). Il est possible de voir que les effectifs observés et théoriques sont assez similaires.
- Finalement, il ne reste qu'à vérifier si la condition d'application d'un test d'indépendance du khi-deux est respectée pour procéder à la prochaine étape. Tous les effectifs théoriques des cellules H20:I27 sont supérieurs à 5. Ainsi, dans la case C34, taper « Oui, $T_{ij} > 5$ » (voir la Figure 7.2.17).

A Tableau des effectifs observés				B				C Gravité de la collision				D				E				F				G Tableau des effectifs théoriques				H Gravité de la collision				I				J			
	Age		Au moins une perte de vie	Blessure non mortelle	Total					Age		Au moins une perte de vie	Blessure non mortelle	Total																									
20	<20		11	904	915					20		14,97860345	900,0213965	915																									
21	20-29		40	3203	3243					20-29		53,0880944	3189,911901	3243																									
22	30-39		44	2857	2901					30-39		47,48953946	2853,510461	2901																									
23	40-49		54	2497	2551					40-49		41,7600000	2509,239981	2551																									
24	50-59		44	2383	2422					50-59		36,94747915	2343,510461	2422																									
25	60-69		29	1576	1605					60-69		26,27394376	1578,776956	1605																									
26	70-79		13	795	808					70-79		13,2270072	794,7729928	808																									
27	>80		6	311	317					>80		5,189308518	311,8106915	317																									
28	Total général		241	14481	14722					Total général		241	14481	14722																									
29																																							
30																																							
31																																							
32																																							
33																																							
34	Condition d'application respectée?									Oui, $T_{ij} > 5$																													
35																																							

Figure 7.2.17 Condition d'application vérifiée

4. À la troisième étape, on procède de deux façons pour en arriver à trouver une règle de décision. On utilise la technique avec la valeur p et la technique avec le khi-deux observé χ^2_{obs} . Il faut remplir les cases grisées, soit la valeur p (l'aire à droite du khi-deux observé), le degré de liberté et le seuil de signification α . Les cases roses se rempliront automatiquement, car les formules pour le calcul des khi-deux respectifs sont inscrites. La fonction Excel **LOI.KHIDEUX.INVERSE.DROITE** est utilisée. Elle nécessite deux paramètres, soit l'aire à droite d'une valeur de khi-deux et le degré de liberté.

La fonction Excel **CHISQ.TEST** permet de calculer la valeur p . Il faut inscrire la plage des effectifs observés ainsi que la plage des effectifs théoriques.

- Dans la cellule B40, taper =CHISQ.TEST(B20:C27;H20:I27) (voir la Figure 7.2.18). Il est possible de sélectionner les plages de cellules au lieu de taper la formule.

La valeur p s'approche de 0,2294, une probabilité assez élevée.

A	B
37	
38 3) Calcul du khi-deux observé et du khi-deux critique :	
39	
40 Valeur p (aire à droite du khi-deux observé)	=CHISQ.TEST(B20:C27;H20:I27) CHISQ.TEST(plage_réelle; plage_prévue)
41 Degré de liberté v :	
42 Khi-deux observé χ^2_{obs} =	
43	
44	
45 Seuil de signification α :	
46 Khi-deux critique χ^2_c :	
47	
48	
49 4) Règle de décision :	
50	
51 On rejette H_0 si le khi-deux observé est supérieur à	
52	
53	

Figure 7.2.18 Remplissage des cases pour les étapes 3 et 4

- Dans la cellule B41, taper $=(8-1)*(2-1)$, soit le calcul du degré de liberté d'un test d'indépendance. On le calcule en multipliant le nombre de modalités de la première variable moins 1 et le nombre de modalités de la deuxième variable moins 1 (voir la [Figure 7.2.18](#)).
 - La cellule B42 se remplit automatiquement. La valeur du χ^2_{obs} est d'environ 9,34.
 - Dans la cellule B45, taper $=0,05$ pour un seuil de 5% (voir la [Figure 7.2.18](#)).
 - La cellule B46 se remplit automatiquement. La valeur du χ^2_c est d'environ 14,07.
5. Dans la cellule A56, écrire un texte qui permet de conclure s'il existe un lien entre les variables **Gravité** et **Âge** (voir la [Figure 7.2.19](#)).

54 5) Décision et conclusion :				
55				
56				
57 Puisque le khi-deux observé (9,34) est inférieur au khi-deux critique (14,07), on ne rejette pas H_0 . Au seuil de 5%, les données échantillonnelles ne permettent				
58 pas de conclure qu'il existe un lien entre la gravité d'une collision et l'âge d'une personne conductrice. Les deux variables sont vraisemblablement				
59 indépendantes.				
60				

Figure 7.2.19 Décision et conclusion du test d'indépendance entre la gravité et l'âge

7.2.2 Lien entre la gravité d'un accident et le type de mesures de sécurité utilisées

Les étapes pour réaliser un test d'indépendance du khi-deux entre deux variables qualitatives sont présentées dans cette sous-section. On souhaite vérifier s'il existe un lien entre la gravité d'un accident et le type de mesures de sécurité utilisées. Il est attendu que la gravité d'un accident dépende des mesures de sécurité utilisées. Selon la Sûreté du Québec, le port de la ceinture et plusieurs autres dispositifs de sécurité comme les sièges pour bébé diminuent la gravité des blessures³. On va tester cette hypothèse au seuil de signification de 5%.

³<https://www.sq.gouv.qc.ca/communiques/sur-la-route-la-securite-et-moi-ca-clique/#:~:text=Les%20recherches%20d%C3%A9montrent%20que%20%3A,70%20bless%C3%A9s%20graves%20chaque%20ann%C3%A9e.>, page consultée le 17 décembre 2024

7.2.2.1 Étapes d'un test d'indépendance avec deux variables qualitatives

Les étapes d'un test d'indépendance avec deux variables qualitatives sont présentées ci-dessous.

1. Ouvrir le fichier **Canevas_tests_independance.xlsx** à l'adresse suivante⁴.
2. Effectuer la procédure présentée au laboratoire 4 pour copier la feuille *Test d'indépendance* du fichier **Canevas_tests_independance** au classeur **Données_Collisions**. Une fois fait, fermer le fichier **Canevas_tests_independance.xlsx**.
3. Dans le classeur **Données_Collisions**, sélectionner la feuille *Test d'indépendance*. Renommer cette feuille « Test Khi-Deux Sécurité» et la placer en dernière position.
4. La première étape d'un test d'indépendance consiste à définir ses variables et écrire les hypothèses nulle et alternative. Dans la cellule C4, taper « La gravité d'une collision ». Dans La cellule C5, taper « Les mesures de sécurité » (voir la Figure 7.2.20).

Pour tout test d'indépendance, l'hypothèse nulle H_0 est que les deux variables sont indépendantes. L'hypothèse alternative H_1 est son contraire, soit que les deux variables sont dépendantes.

Dans l'encadré gris de la cellule D8, vis-à-vis H_0 , taper « La gravité d'une collision et les mesures de sécurité utilisées sont indépendantes » (voir la Figure 7.2.20). Dans l'encadré gris de la cellule D11, vis-à-vis H_1 , taper « La gravité d'une collision et les mesures de sécurité utilisées sont dépendantes » (voir la Figure 7.2.20)

A	B	C	D	E	F	G	H	I	J
1									
2									
3									
4		X :	La gravité d'un accident						
5		Y :							
6									
7									
8	1] Hypothèses :			H ₀ :					
9									
10				H ₁ :					
11									
12									
13									
14									
15	2] Tableaux des effectifs observés et théoriques :								
16									
17									
18									
19									
20									
21									

Figure 7.2.20 Remplissage des encadrés de la première étape d'un test d'indépendance entre la gravité d'une collision et les mesures de sécurité utilisées

5. La deuxième étape consiste à construire le tableau des effectifs observés et le tableau des effectifs théoriques.
 - Il faut commencer par générer un tableau croisé dynamique des effectifs observés. Dans la feuille **Données_Collisions**, sélectionner le tableau « Données ».
 - En suivant les étapes présentées à la Sous sous-section 1.2.8.1, générer, dans une nouvelle feuille de calcul qu'on renomme **TCD_GraviteSécurité**,

⁴github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Canevas/Canevas_tests_independance.xlsm?raw=true

un tableau croisé dynamique vide à partir du tableau *Données*. Placer ce tableau dans la cellule A3 (voir la Figure 7.2.21).

- Glisser et déposer la variable *Mesures de sécurité* dans la zone de saisie *Lignes* (voir la Figure 7.2.21).
- Glisser et déposer la variable *Gravité* dans la zone de saisie *Colonnes* (voir la Figure 7.2.21).
- Glisser et déposer la variable *Mesures de sécurité* dans la zone de saisie *Valeurs* (voir la Figure 7.2.21). S'assurer que le calcul fait est bien le compte (*Nombre*) et non la somme. Si c'est la somme, changer pour *Nombre* avec les étapes vues à la Sous sous-section 1.2.8.2.

Figure 7.2.21 Tableau croisé dynamique avec la variable *Mesures de sécurité* en lignes et la variable *Mesures de sécurité* en colonnes

- On ne veut inclure que les données du mois de décembre (code 12). On appliquer des filtres pour faire cela.

Glisser et déposer la variable *Mois* dans la zone de saisie *Filtres*. Le filtre apparaît dans les cellules A1:B1 (voir la Figure 7.2.21). Cliquer sur la flèche du menu déroulant de la cellule B1. Sélectionner 12 pour le mois de décembre (voir la Figure 7.2.21). Cliquer sur *Ok*.

À ce stade, le tableau croisé dynamique devrait ressembler à la Figure 7.2.22.

	A	B	C	D
1	Mois	12		
2				
3	Nombre de Mesures de sécurité	Étiquettes de colonnes		
4	Étiquettes de lignes		1	2 Total général
5	1		38	492 530
6	10		1	8 9
7	12		3	295 298
8	13		1	164 165
9	2		282	16417 16699
10	9		3	63 66
11	NN		32	2434 2466
12	QQ		12	269 281
13	UU		58	2683 2741
14	Total général		430	22825 23255

Figure 7.2.22 Tableau croisé dynamique avec les variables *Mesures de sécurité* et *Gravité* filtrées

Ce tableau croisé dynamique présente deux difficultés pour la poursuite d'un test d'indépendance du khi-deux. Premièrement, on ne

veut pas inclure les modalités NN, QQ et UU, car elles n'offrent aucune information. On rappelle que ce sont les codes pour « l'élément d'information est sans objet », « autre situation que les précédentes » et « Inconnu ». Deuxièmement, les modalités 9, 10 et 13 ont peu d'effectifs observés. En soi, ce n'est pas un problème. Cependant, le calcul d'effectifs théoriques impliquant ces modalités vont engendrer des valeurs inférieures à 5. Ainsi, pour éviter d'avoir des effectifs théoriques inférieurs à 5, on doit regrouper des modalités à échelle nominale.

Remarque 7.2.23 Regrouper les valeurs d'une variable à échelle nominale. Pour regrouper les valeurs d'une variable à échelle nominale, il est suggéré d'y aller avec la logique. Dans le cas des dispositifs de sécurité, les modalités 1, 12 et 13 sont des catégories dans lesquelles aucun dispositif de sécurité n'a été utilisé. Il y a donc un lien logique à regrouper ces catégories. Les modalités 2, 9 et 10 sont des catégories dans lesquelles une mesure de sécurité a été utilisée (ceinture, port de casque et port de vêtements réfléchissants respectivement). Il est donc logique de regrouper ces trois modalités.

- On veut maintenant grouper certaines modalités de la variable ***Mesures de sécurité***. Puisque les valeurs des modalités sont des codes numériques non consécutifs, il sera plus facile de les grouper s'ils sont placés l'un à la suite de l'autre dans l'ordre de groupement. Excel permet de déplacer des modalités.

Dans la colonne des étiquettes de lignes de la variable ***Mesures de sécurité***, cliquer avec le bouton de gauche de la souris sur la valeur 12 telle que cette dernière est encadrée d'une bordure verte. Ensuite, cliquer avec le bouton de droite de la souris sur la valeur 12. Sélectionner l'option **Déplacer** (voir la Figure 7.2.24), suivi de l'option **Déplacer “12” vers le haut**.

A	B	C	D	E	F	G	H
1 Mois	12						
2							
3 Nombre de Mesures de sécurité	Étiquettes de colonnes						
4 Étiquettes de lignes		1	2	Total général			
5 1		38	492	530			
6 10		1	8	9			
7 12		3	295	298			
8 13		1	164	165			
9 2		282	16417	16699			
10 9		3	63	66			
11 NN		32	2434	2466			
12 QQ		12	269	281			
13 UU		58	2683	2741			
14 Total général		430	22825	23255			
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
~							

Figure 7.2.24 Déplacer les modalités de la variable ***Mesures de sécurité***

- On va faire de même pour les modalités 10 et 13. Dans la colonne des étiquettes de lignes de la variable ***Mesures de sécurité***, cliquer avec le bouton de droite de la souris sur la valeur 10. Sélectionner l'option **Déplacer** (voir la Figure 7.2.24), suivi de l'option **Déplacer “10” vers le bas**. Au final, les modalités 1, 12 et 13

sont l'une à la suite de l'autre ainsi que les modalités 10, 2 et 9 (voir la Figure 7.2.25).

	A	B	C	D
1	Mois	12		
2				
3	Nombre de Mesures de sécurité	Étiquettes de colonnes		
4	Étiquettes de lignes		1	2 Total général
5	1		38	492 530
6	12		3	295 298
7	13		1	164 165
8	10		1	8 9
9	2		282	16417 16699
10	9		3	63 66
11	NN		32	2434 2466
12	QQ		12	269 281
13	UU		58	2683 2741
14	Total général		430	22825 23255

Figure 7.2.25 Séquence de modalités dans l'ordre voulu

- On va grouper les modalités 1, 12 et 13 ainsi que les modalités 2, 9 et 10.

Dans la colonne des étiquettes de lignes de la variable **Mesures de sécurité**, lorsque la flèche noire pointant vers la droite apparaît, sélectionner les trois lignes des modalités 1, 12 et 13 (voir la Figure 7.2.26).

	A	B	C	D	E
1	Mois	12			
2					
3	Nombre de Mesures de sécurité	Étiquettes de colonnes			
4	Étiquettes de lignes		1	2 Total général	
5	1		38	492 530	
6	12		3	295 298	
7	13		1	164 165	
8	10		1	8 9	
9	2		282	16417 16699	
10	9		3	63 66	
11	NN		32	2434 2466	
12	QQ		12	269 281	
13	UU		58	2683 2741	
14	Total général		430	22825 23255	
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					

Figure 7.2.26 Groupement de modalités

- Une fois les lignes grisées, cliquer avec le bouton de droite de la souris et sélectionner l'option **Grouper** (voir la Figure 7.2.26). Excel groupe les trois modalités dans un groupe intitulé **Groupe 1** (voir la Figure 7.2.27).

A	B	C	D
1 Mois	12		
3 Nombre de Mesures de sécurité	Étiquettes de colonnes		
4 Étiquettes de lignes	1 2 Total général		
5 Groupe1	42 951 993		
6 1	38 492 530		
7 12	3 295 298		
8 13	1 164 165		
9 10	1 8 9		
10 10	1 8 9		
11 2	282 16417 16699		
12 2	282 16417 16699		
13 9	3 63 66		
14 9	3 63 66		
15 NN	32 2434 2466		
16 NN	32 2434 2466		
17 QQ	12 269 281		
18 QQ	12 269 281		
19 UU	58 2683 2741		
20 UU	58 2683 2741		
21 Total général	430 22825 23255		

Figure 7.2.27 Groupement des modalités et création de Groupe 1

- À gauche du nom **Groupe 1**, il y a un icône -. Cliquer cette icône pour réduire les modalités de ce groupe, c'est-à-dire pour rendre la lecture moins encombrante (voir la [Figure 7.2.26](#)).
- Répéter l'étape précédente pour les modalités 10, 2, 9, *NN*, *QQ* ainsi que *UU* (voir la [Figure 7.2.26](#)).
- Dans la colonne des étiquettes de lignes de la variable **Mesures de sécurité**, lorsque la flèche noire pointant vers la droite apparaît, sélectionner les trois lignes des modalités 10, 2 et 9 (voir la [Figure 7.2.26](#)).
- Cliquer avec le bouton de droite de la souris et sélectionner l'option **Grouper** (voir la [Figure 7.2.26](#)). Le **Groupe 2** est créé. Cliquer sur l'icône à gauche de **Groupe 2** pour réduire les modalités de ce groupe.

Le résultat des groupements est présenté à la [Figure 7.2.28](#).

A	B	C	D
1 Mois	12		
3 Nombre de Mesures de sécurité	Étiquettes de colonnes		
4 Étiquettes de lignes	1 2 Total général		
5 Groupe1	42 951 993		
6 Groupe2	286 16488 16774		
7 NN	32 2434 2466		
8 QQ	12 269 281		
9 UU	58 2683 2741		
10 Total général	430 22825 23255		

Figure 7.2.28 Groupement final des modalités

- Pour filtrer les modalités *NN*, *QQ* et *UU*, cliquer sur le petit triangle du menu déroulant de l'étiquette de lignes (voir la [Figure 7.2.29](#)).

The screenshot shows a dynamic pivot table with columns A, B, C, D, and E. Column A contains rows from 1 to 25, with row 1 labeled 'Mois' and row 10 labeled 'Total général'. Column B contains row 12. The filter pane is open, showing the following settings:

- Sélectionner le champ :** Mesures de sécurité2
- Trier de A à Z :** Groupe1, Groupe2, NN, QQ, UU
- Trier de Z à A :** Total général, NN, QQ, UU, Groupe1, Groupe2
- Options de tri supplémentaires...**
- Effacer le filtre de « Mesures de sécurité2 »**
- Filtres s'appliquant aux étiquettes**
- Filtres s'appliquant aux valeurs**
- Rechercher** (checkboxes checked for Groupe1, Groupe2, NN, QQ, UU)

Figure 7.2.29 Filtrage des modalités *NN*, *QQ* et *UU*

- Décliquer les modalités *NN*, *QQ* et *UU*. Cliquer sur *OK* (voir la Figure 7.2.29).
- À ce stade, le tableau croisé dynamique devrait ressembler à la Figure 7.2.30

The screenshot shows a dynamic pivot table with columns A, B, C, and D. Column A contains rows from 1 to 7, with row 1 labeled 'Mois' and row 7 labeled 'Total général'. Column B contains row 12. The filter pane is open, showing the following settings:

- Sélectionner le champ :** Mesures de sécurité2
- Trier de A à Z :** Groupe1, Groupe2, NN, QQ, UU
- Trier de Z à A :** Total général, NN, QQ, UU, Groupe1, Groupe2
- Options de tri supplémentaires...**
- Effacer le filtre de « Mesures de sécurité2 »**
- Filtres s'appliquant aux étiquettes**
- Filtres s'appliquant aux valeurs**
- Rechercher** (checkboxes checked for Groupe1, Groupe2, NN, QQ, UU)

A	B	C	D
1 Mois	12		
2			
3 Nombre de Mesures de sécurité	Étiquettes de colonnes		
4 Étiquettes de lignes		1	2 Total général
5 Groupe1		42	951 993
6 Groupe2		286	16488 16774
7 Total général		328	17439 17767

Figure 7.2.30 Tableau croisé dynamique avec les variables *Mesures de sécurité* et *Gravité*

- Dans la feuille *Test Khi-Deux Sécurité*, cliquer sur la cellule B20 et taper le symbole [=] (voir la Figure 7.2.31).

A	B	C	D	E	F
9					
10					
11	H ₁ :	La gravité d'un accident et les mesures de sécurité			
12					
13					
14					
15	2) Tableaux des effectifs observés et théoriques :				
16					
17					
18					
19					
20	=				
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					

Figure 7.2.31 Copiage d'une partie du tableau croisé dynamique dans le canevas *Test Khi-Deux Sécurité*

- Cliquer sur l'onglet de la feuille **TCD_GraviteSécurité** et sélectionner la plage A5:D7 (voir la [Figure 7.2.31](#)).
- Taper la touche **Enter** du clavier. Le tableau suivant apparaît dans la feuille **Test Khi-Deux Sécurité** (voir la [Figure 7.2.32](#)).

A	B	C	D
12			
13			
14			
15	2) Tableaux des effectifs observés et théoriques :		
16			
17			
18			
19			
20	Groupe1	42	951
21	Groupe2	286	16488
22	Total général	328	17439
23			17767

Figure 7.2.32 Tableau des effectifs observés

- Faire une mise en forme élémentaire du tableau. Le tableau final des effectifs observés doit ressembler au tableau de la [Figure 7.2.33](#). Les couleurs et bordures importent peu ici. Si ceci était un tableau de présentation pour un rapport final, il faudrait peaufiner la mise en page.

A	B	C	D
12			
13			
14			
15	2) Tableaux des effectifs observés et théoriques :		
16			
17	Tableau des effectifs observés		
18	Mesures de sécurité	Gravité de la collision	
19		Au moins une perte de vie	Total
20	Aucun dispositif de sécurité	42	993
21	Certains dispositifs de sécurité	286	16774
22	Total général	328	17767

Figure 7.2.33 Tableau final des effectifs observés

- Il faut maintenant créer le tableau des effectifs théoriques. Dans la cellule G17, taper « Tableau des effectifs théoriques » (voir la [Figure 7.2.34](#)).

A	B	C	D	E	F	G	H	I	J	K
12										
13										
14										
15	2) Tableaux des effectifs observés et théoriques :									
16										
17	Tableau des effectifs observés									
18	Mesures de sécurité		Gravité de la collision							
19			Au moins une perte de vie	Blessure non mortelle	Total					
20	Aucun dispositif de sécurité	42		951	993					
21	Certains dispositifs de sécurité	286		16488	16774					
22	Total général	328		17439	17767					
23										
24										
25										
26										
27										
28										
29										

Figure 7.2.34 Création du tableau des effectifs théoriques

- Copier la plage de cellules A18:D22, soit le tableau des effectifs observés. Faire un collage spécial en ne collant que les valeurs dans la cellule G18 (voir la [Figure 7.2.34](#)).
 - Sélectionner la plage de cellules H20:I21 et la supprimer en appuyant sur la touche **[Suppr]** du clavier (voir la [Figure 7.2.34](#)).
 - Formatter les titres au besoin. Ce n'est pas vraiment nécessaire.
 - La formule pour calculer un effectif théorique est le total de la colonne multiplié par le total de la ligne divisé par la taille de l'échantillon. Dans la cellule H20, taper =H\$22*I\$20/\$J\$22 (voir la [Figure 7.2.34](#)).
 - Sélectionner la cellule H20. Placer le curseur dans le coin inférieur droit jusqu'à ce que la croix noire de recopie (+) apparaisse. Cliquer avec le bouton de gauche de la souris en gardant le bouton enfoncé et glisser le curseur vers le bas jusqu'à la cellule H21. Relâcher le bouton de gauche de la souris. Avec la plage H20:H21 sélectionnée, placer encore une fois le curseur dans le coin inférieur droit et glisser vers la droite pour remplir les cellules I20:I21 (voir la [Figure 7.2.34](#)).
 - Finalement, il ne reste qu'à vérifier si la condition d'application d'un test d'indépendance du khi-deux est respectée pour procéder à la prochaine étape. Tous les effectifs théoriques des cellules H20:I21 sont supérieurs à 5. Ainsi, dans la case C34, taper taper « Oui, $T_{ij} > 5$ ».
6. À la troisième étape, on procède de deux façons pour en arriver à trouver une règle de décision. On utilise la valeur p et le khi-deux observé χ^2_{obs} . Il faut remplir les cases grisées, soit la valeur p (l'aire à droite du khi-deux observé), le degré de liberté et le seuil de signification α . Les cases roses se rempliront automatiquement, car les formules pour le calcul des khi-deux respectifs sont inscrites. La fonction Excel **LOI.KHIDEUX.INVERSE.DROITE** est utilisée. Elle nécessite deux paramètres, soit l'aire à droite d'une valeur de khi-deux et le degré de liberté.
- La fonction Excel **CHISQ.TEST** permet de calculer la valeur p . Il faut inscrire la plage des effectifs observés ainsi que la plage des effectifs théoriques.
- Dans la cellule B40, taper =CHISQ.TEST(B20:C21;H20:I21) (voir la [Figure 7.2.35](#)). Il est possible de sélectionner les plages de cellules au lieu de taper la formule.
- La valeur p s'approche de $9,33 \times 10^{-9}$, une probabilité très petite.

37	
38	3) Calcul du khi-deux observé et du khi-deux critique :
39	
40	Valeur p (aire à droite du khi-deux observé) =CHISQ.TEST(B20:C21;H20:I21)
41	Degré de liberté v : CHISQ.TEST(plage_réelle; plage_prévue)
42	Khi-deux observé χ^2_{obs} =
43	
44	
45	Seuil de signification α :
46	Khi-deux critique χ^2_c :
47	
48	
49	4) Règle de décision :
50	
51	On rejette H_0 si le khi-deux observé est supérieur à
52	
53	

Figure 7.2.35 Remplissage des cases pour les étapes 3 et 4

- Dans la cellule B41, taper $=(2-1)*(2-1)$, soit le calcul du degré de liberté d'un test d'indépendance. On le calcule en multipliant le nombre de modalités de la première variable moins 1 et le nombre de modalités de la deuxième variable moins 1 (voir la [Figure 7.2.35](#)).
 - La cellule B42 se remplit automatiquement. La valeur du χ^2_{obs} est d'environ 32,98.
 - Dans la cellule B45, taper $=0,05$ pour un seuil de 5% (voir la [Figure 7.2.35](#)).
 - La cellule B46 se remplit automatiquement. La valeur du χ^2_c est d'environ 3,84.
7. Dans la cellule A56, écrire un texte qui permet de conclure s'il existe un lien entre les variables **Gravité** et **Mesures de sécurité** (voir la [Figure 7.2.36](#)).

54	5) Décision et conclusion :
55	
56	
57	Puisque le khi-deux observé (32,98) est supérieur au khi-deux critique (3,84), on rejette H_0 . Au seuil de 5%, on peut conclure que la gravité d'un accident et les mesures de sécurité utilisées sont dépendantes. La valeur p très petite et inférieure au seuil de signification de 5% permet aussi de constater la dépendance.
58	
59	
60	
61	

Figure 7.2.36 Décision et conclusion du test d'indépendance entre la gravité et le type de mesures de sécurité

7.3 Réflexions

Ce laboratoire aura permis d'explorer les concepts liés aux tests d'indépendance. Toutefois, une étude plus approfondie est utile pour en saisir toutes les subtilités

Travail à faire après le laboratoire

Objectifs

- Poser un regard critique sur les données.
 - Formuler des hypothèses de recherche.
 - Choisir le type de tests adéquatement et comprendre les distinctions entre chaque type de tests.
 - Effectuer un test d'indépendance pour vérifier s'il existe un lien ou non entre deux variables.
 - Interpréter adéquatement la conclusion d'un test.
1. Les cinq étapes d'un test d'indépendance permettent de vérifier s'il existe ou non un lien entre deux variables X et Y . Pourquoi n'est-il pas possible de simplement vérifier si $P(X_i) = P(X_i|Y_j)$, où $P(X_i)$ est la probabilité de l'événement X_i ? Pourquoi faut-il calculer des effectifs théoriques et les comparer aux effectifs observés?
 2. La conclusion d'un test d'indépendance peut parfois être obtenue avec un test paramétrique. Lequel? Expliquer.
 3. À la [Sous-section 7.2.1](#), on a fait un test d'indépendance vérifiant s'il existe un lien entre l'âge d'un conducteur et la gravité d'une collision. La variable *Âge* a été groupée en tranche de 10 ans. Ceci a mené au non-rejet de l'hypothèse nulle, c'est-à-dire qu'on n'a pas pu conclure qu'il existait un lien entre l'âge et la gravité. Choisir un groupement d'âge différent de celui fait à la [Sous-section 7.2.1](#) et montrer que la conclusion peut changer.
 4. Parmi toutes les bases de données disponibles dans l'[Appendice A](#), formuler quelques hypothèses de recherche et réaliser deux tests d'indépendance pour valider ces hypothèses. Utiliser les canevas fournis. Au moins un des tests doit incorporer une variable quantitative continue.

Chapitre 8

Corrélation et régression

Text before the first section.

8.1 Prélab

Dans ce laboratoire, on poursuit le travail des deux laboratoires précédents à propos de la recherche d'un lien entre deux variables. En particulier, lorsque les deux variables sont quantitatives continues, on peut s'intéresser au type de lien qu'il peut y avoir entre deux variables. Ce lien peut être linéaire, polynomiale, exponentielle, logarithmique et ainsi de suite. Par simplicité, on commence par considérer uniquement les variables dont le lien semble être linéaire. Afin de tirer les bonnes conclusions, il est primordial que cette linéarité soit présente. La principale manière de constater ce type de lien est en illustrant les variables sur un nuage de points, aussi appelé diagramme de dispersion.

Travail à faire avant le cours

Objectifs

- Déterminer visuellement si la nature du lien entre deux variables est linéaire.
 - Introduire le calcul de la droite de régression.
 - Préparer le fichier de base de données pour le laboratoire.
1. Parmi les images suivantes, déterminer celle ou celles qui ont le plus l'allure d'un lien linéaire.

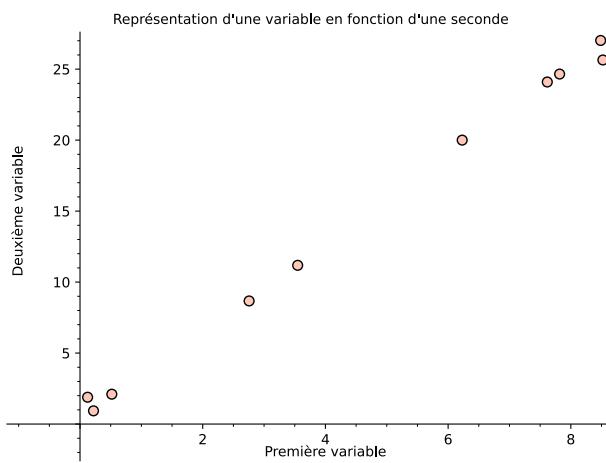


Figure 8.1.1 Un premier lien entre deux variables

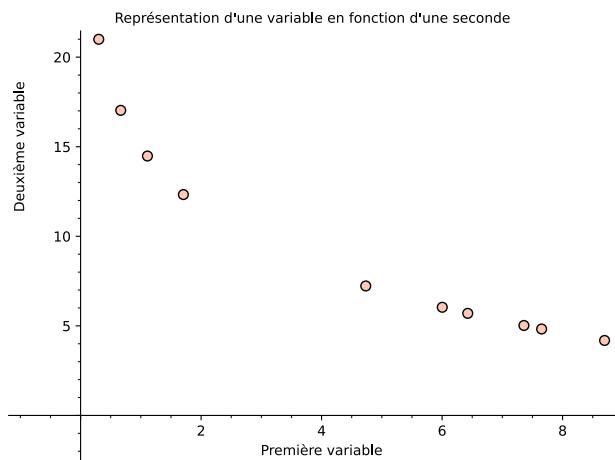


Figure 8.1.2 Un deuxième lien entre deux variables

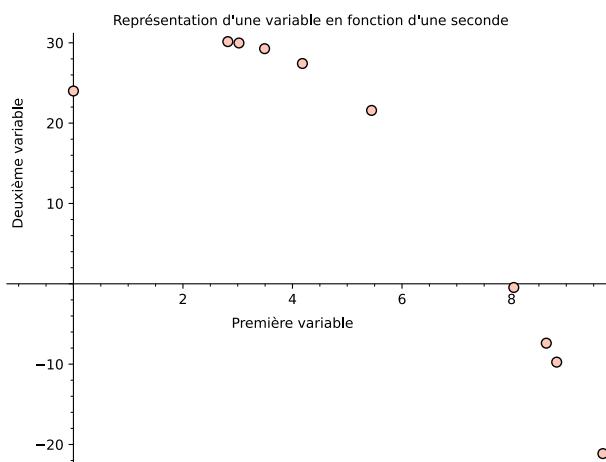


Figure 8.1.3 Un troisième lien entre deux variables

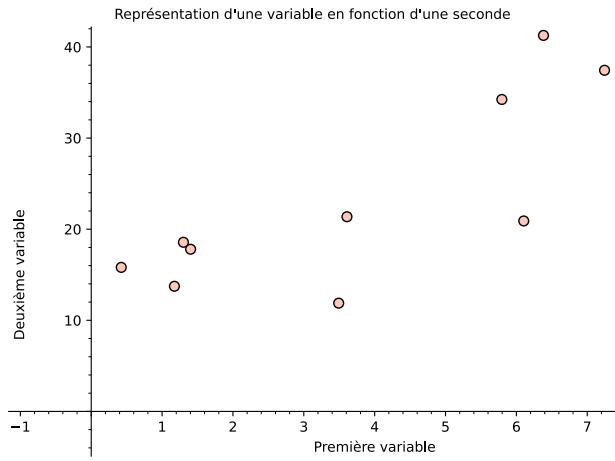


Figure 8.1.4 Un quatrième lien entre deux variables

2. Par deux points il ne peut passer qu'une seule droite. Étant donnés un ensemble de points, il est impossible de penser qu'une relation de la forme $y = ax + b$ pourra passer par tous les points. Il existe plusieurs manières de définir « la meilleure droite » $y = ax + b$ représentant un ensemble de points. La plus commune est celle obtenue en appliquant la méthode des moindres carrés.

Cette méthode consiste à calculer la différence entre chacune des valeurs dépendantes des données et la valeur dépendante de l'équation d'une droite de paramètres a, b et d'additionner le carré de toutes ces différences. On cherche les valeurs de a et b qui minimise cette somme.

- On considère les points $A(1; 1)$, $B(2; 3)$ et $C(4; 4)$. Écrire les trois termes de la somme des moindres carrés pour ces trois points.
- Dans un cours de calcul différentiel à plusieurs variables, on apprend que pour optimiser une fonction multivariée, il faut que les dérivées par rapport à chacune de ses variables soient égales à 0. Dans le cas des moindres carrés, deux variables sont présentes. En traitant à tour de rôle l'une variable comme étant constante, calculer la dérivée

de la somme obtenue à la partie précédente par rapport à l'autre variable pour obtenir deux équations linéaires en a, b .

- (c) En posant chacune de ces équations égales à 0, montrer que la droite des moindres carrés de ce problème est

$$y = \frac{13}{14}x + \frac{1}{2}.$$

3. Le fichier `13jeux_de_donnees.xlsx` disponible à l'adresse [adresse](#)¹ contient 13 ensembles de couples de données ayant été créés artificiellement afin de produire le résultat des exercices ci-dessous. Le but de cet exercice est de comprendre l'importance d'observer les données avant d'en tirer quelconque conclusion.
- (a) Dans une nouvelle feuille de calculs, créer un tableau croisé dynamique contenant dans la zone de saisie **Colonne** les entrées **Valeurs_X** et **Valeurs_Y**. Ajouter aussi un filtre avec l'entrée **Jeu x**
 - (b) Positionner le filtre en haut du tableau sur **Jeu 1**.
 - (c) À l'aide de fonctions Excel, calculer la moyenne des valeurs X, la moyenne des valeurs Y, l'écart type des valeurs X, l'écart type des valeurs Y ainsi que le coefficient de corrélation entre les valeurs X et Y.
 - (d) Insérer un nuage de points avec les valeurs du tableau croisé dynamique.
 - (e) Observer l'effet qu'a le changement du jeu de données dans le filtre sur les mesures statistiques calculées et sur l'allure du nuage de points. Commenter brièvement.
4. Télécharger le fichier `Données_Cepheides.xlsx` disponible à l'adresse [adresse](#)². Ce fichier contient différentes données de 421 étoiles de type « céphéides ». Dans le prochain laboratoire, on cherche à déterminer s'il existe des liens linéaires entre certaines de ces variables.
- (a) Dans une feuille appelée « RP », tracer le nuage de points représentant le rayon des étoiles en fonction de leur période de pulsation. Effectuer la mise en forme appropriée.
 - (b) Dans une feuille appelée «MappP », tracer le nuage de points représentant la magnitude apparente des étoiles en fonction de leur période de pulsation. Effectuer la mise en forme appropriée.
 - (c) Dans une feuille appelée «MabsP », tracer le nuage de points représentant la magnitude absolue des étoiles en fonction de leur période de pulsation. Effectuer la mise en forme appropriée.
 - (d) Sur chacune des trois feuilles, faire l'étude descriptive de la population telle que décrite dans la [Sous sous-section 3.2.1.6](#) pour les variables impliquées.

¹github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/13jeux_de_donnees.xlsx?raw=true

²github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Cepheides.xlsx?raw=true

8.2 Laboratoire

Dans ce laboratoire, on s'intéresse au type de lien qui peut exister entre deux variables quantitatives. Plus spécifiquement, on cherche à déterminer l'existence d'un lien linéaire entre des variables X, Y , et donc de l'existence de coefficients a, b pour lesquels $Y \approx aX + b$. Cette droite est appelée la droite de régression. Cette droite est appelée la droite de régression. Plus la dépendance entre X, Y est forte, plus cette approximation sera bonne. La force de ce lien (linéaire) est donnée par le coefficient de corrélation, aussi appelé coefficient de Pearson. Il est noté r dans le cas d'un échantillon et ρ (lettre grecque « rho ») pour une population.

La première étape d'une étude sur la nature du lien entre deux variables doit toujours être l'observation des données afin de vérifier qualitativement la plausibilité de l'existence d'un lien linéaire entre les variables. L'équation de la droite de régression ainsi que le coefficient de corrélation peuvent presque toujours être calculés, même si les données ne sont pas reliées linéairement. Il incombe à l'analyste de déterminer la pertinence et l'interprétation adéquate de ces outils.

Pour présenter ces concepts, on utilise une base de données portant sur 421 étoiles de type « céphéide ». Une céphéide est une étoile variable, c'est-à-dire dont la luminosité change dû à un mouvement périodique de dilatation et de contraction. La première à avoir été découverte, par l'astronome amateur John Goodricke en 1784, fait partie de la constellation Céphée, qui a par la suite prêté son nom à cette nouvelle catégorie d'étoile. Les données datent de 1997 et proviennent de l'[Observatoire-David-Dunlap](#)¹ à Richmond Hill, en Ontario. Cet observatoire a été désigné comme un [lieu historique national](#)² en 2019. Les variables présentes dans le fichier sont

- ID: Numéro de l'échantillon
- Étoile: Nom de l'étoile
- Période: Période de pulsation (en jours)
- Mag App: Magnitude apparente moyenne (sans unités)
- Couleur BV: Indice de couleur B-V moyen (sans unités)
- Excès: Excès de couleur (sans unités)
- Amplitude V: Amplitude de la luminosité observée au travers d'un filtre V (~500nm) (sans unités)
- Mag Abs: Magnitude Absolue (sans unités)
- Dist: Distance entre l'étoile et le soleil (parsecs)
- VR MOY: Vitesse radiale moyenne (km/s)
- RAYON: Rayon solaire (x 6,957x10⁸ m)

8.2.1 Corrélation et régression linéaire

Dans le prélaboratoire, on a demandé de tracer le nuage de points de la période de pulsation en fonction du rayon. À l'observation du graphique, il est plausible de conclure à l'existence d'un lien linéaire entre le rayon d'une céphéide et sa période de pulsation. On peut donc aller de l'avant avec l'analyse de la relation linéaire entre ces variables.

¹www.astro.utoronto.ca/DDO/research/cepheids/cepheids.html

²parcs.canada.ca/culture/designation/lieu-site/david-dunlap

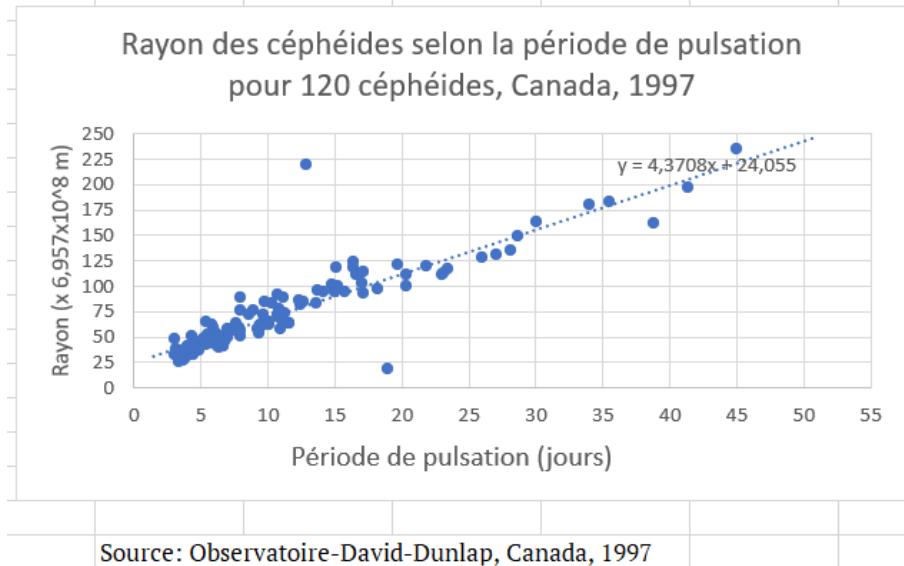


Figure 8.2.1 Le nuage de points créé sur Excel. Source: Observatoire-David-Dunlap, Canada, 1997

Dans un premier temps, on calcule le coefficient de corrélation linéaire. La formule `COEFFICIENT.CORRELATION(matrice1;matrice2)` permet d'obtenir ce coefficient, où `matrice1` et `matrice2` sont les plages contenant les données des deux variables. La formule mathématique du coefficient de corrélation est

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right). \quad (8.2.1)$$

C'est une mesure de la variabilité conjointe des variables, normalisée pour être entre -1 et 1 .

Le signe du coefficient de corrélation donne le sens de la variabilité de la variable Y lorsque X varie: une variation positive de X entraîne une variation positive de Y lorsque $r > 0$ et une variation positive de X entraîne une variation négative de Y lorsque $r < 0$. La magnitude du coefficient de corrélation quant à elle quantifie la force du lien linéaire entre les variables. Traditionnellement, la force est obtenue selon la valeur de $|r|$ dans la table suivante.

Table 8.2.2 Interprétation de la force du lien linéaire

$ r $	Interprétation du lien
$[0, 9; 1]$	Très fort à parfait
$[0, 6; 0, 9[$	Fort
$[0, 3; 0, 6[$	Moyen
$[0; 0, 3[$	Aucun à faible

On calcule le coefficient de corrélation entre la période de pulsation et le rayon des céphéides et on l'ajoute sous les études descriptives. Puisque les écarts types des variables font partie de la [formule mathématique \(8.2.1\)](#) du coefficient de variation, ce dernier est sensible aux valeurs extrêmes. Il est donc important de bien analyser les données lors d'une étape préliminaire.

Une fois la nature linéaire du lien confirmée, il est possible de déterminer l'équation de la meilleure droite qui relie les deux variables. L'exercice [Activité 8.1.2](#) donne une idée de la procédure mathématique à suivre pour la trouver. Un ouvrage statistique théorique peut montrer que la pente est donnée par $a = r \frac{s_y}{s_x}$ et l'ordonnée à l'origine par $b = \bar{y} - a\bar{x}$. Sur Excel, on

peut afficher directement sur le graphique la droite et son équation. Pour cela, on sélectionne le graphique et on clique sur la petite croix en haut à droite, puis sur *linéaire*. La case *Courbe de tendance* aurait aussi pu faire le travail, puisque son comportement par défaut est la relation linéaire. Pour ajouter l'équation, on clique sur *Autres options* ou on fait un double-clic sur la droite créée précédemment et on coche la case *Afficher l'équation sur le graphique*.

La fonction DROITEREG permet aussi d'obtenir directement dans les cellules les valeurs de a, b , mais elle a comme inconvénient qu'il faut que le nombre de valeurs pour chacune des variables soit le même. Dans le cas de la base de données des céphéides, le rayon de plusieurs étoiles est manquant. Il faut alors filtrer dans la formule pour exclure les entrées vides. La formule DROITEREG((FILTRE(Rayon;Rayon<1;>0));FILTRE(PÉRIODE;Rayon<>0)) permet d'accomplir cela, où PÉRIODE et RAYON sont les plages nommées correspondant à ces variables dans le tableau des données. Dans les deux cas, si l'on note P la période et R le rayon, on obtient

$$R = 4,370849753P + 24,05523133. \quad (8.2.2)$$

Avec cette droite, on est en mesure d'estimer le rayon d'une étoile dont la période de pulsation est connue ou à l'inverse, de déterminer la période de pulsation d'une céphéide de rayon donné. Par exemple, la céphéide FF_AQL dont l'identifiant est 4 possède une période égale à environ 4,47 jours. En entrant cette valeur dans l'équation, on estime un rayon de $43,59292972591(\times 6,957 \times 10^8)$ mètres. Selon [Wikipédia](#)³, la composante numérique du rayon est de 39.

Une autre mesure de la force du lien est donnée par le coefficient de détermination, habituellement noté R^2 (sans lien avec le rayon mentionné précédemment). Ce dernier quantifie la proportion des variations de la variable dépendante qui sont expliquées par la variable indépendante. Il est donné en pourcentage et lorsqu'il n'y a qu'une seule variable indépendante, on a toujours

$$R^2 = r^2 \times 100\%.$$

Sur excel, la commande COEFFICIENT.DETERMINATION(matrice1;matrice2) permet de l'obtenir, mais ne donne pas le résultat en pourcentage. Il faut manuellement modifier le format de cellule. En calculant le coefficient de détermination entre la période de pulsation et le rayon d'une céphéide, on trouve qu'une variation du rayon d'une étoile de type céphéide explique environ 81% de la variation de sa période de pulsation.

8.2.2 Absence de lien linéaire apparent et autres cas

Dans la feuille **MappP**, on trace le nuage de points représentant la magnitude apparente et la période de pulsation. Le résultat est illustré à la figure ci-dessous.

³en.wikipedia.org/wiki/FF_Aquilae#cite_note-turner-5

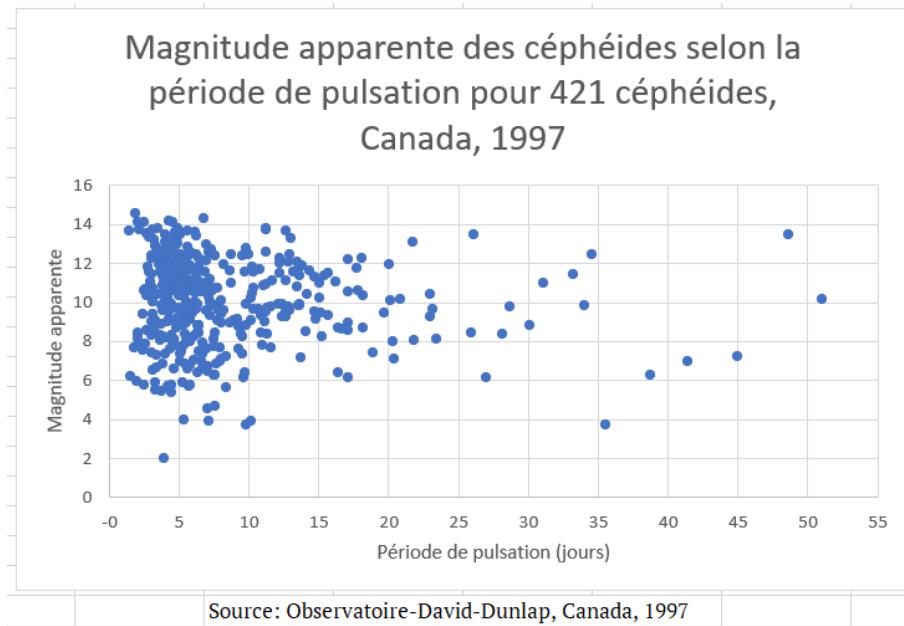


Figure 8.2.3 Nuage de points représentant la magnitude apparente et la période

Le calcul des coefficients de corrélation et de détermination montre d'autant plus que ces variables ne sont pas reliées de manière linéaire. Ceci s'explique probablement en partie du fait que la luminosité apparente dépend de la distance de l'étoile. Au contraire, la magnitude absolue ramène cette valeur sur une même échelle.

Dans la feuille ***MabsP***, on trace le nuage de points représentant la magnitude absolue et la période de pulsation. Le résultat est illustré à la figure ci-dessous.

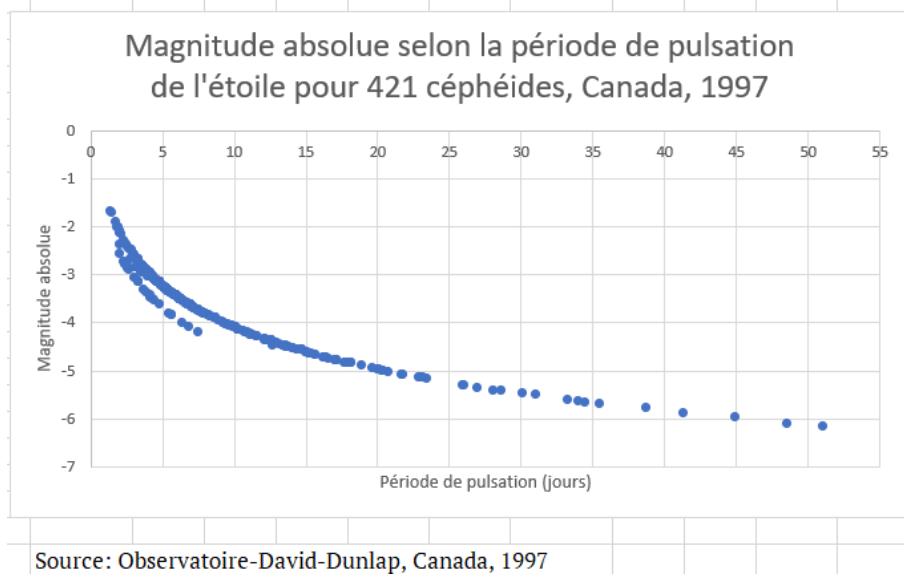


Figure 8.2.4 Nuage de points représentant la magnitude absolue et la période

Deux choses ressortent de ce graphique. Premièrement, la relation ne semble pas linéaire, mais il y a clairement un lien entre les deux variables. Deuxièmement, bien que la relation semble très forte, il semble aussi y avoir deux

types de relations. En effet, on peut observer dans le graphique produit deux amas principaux de points. Ces deux amas paraissent suivre une relation de même type, mais il y a possiblement une troisième variable qui influence le lien entre la magnitude absolue et la période de pulsation, ou des erreurs dans les données disponibles.

On commence par s'attarder à la première observation. Bien qu'elle ne soit pas linéaire, la relation liant les variables ressemble à un courbe bien connue, soit un logarithme. En cliquant sur l'outil croix du graphique, puis sur la flèche au niveau de l'option **courbe de tendance** suivi de **autres options**, on peut modifier le type de fonction utilisée pour approximer le lien. L'option logarithme est présente. On peut également afficher l'équation sur le graphique, qui sera de la forme $a \ln(x) + b$.

Une autre option est de modifier l'axe de la période pour que l'échelle soit logarithmique. Voici la marche à suivre pour y arriver.

1. Faire un clic à l'aide du bouton de droite de la souris sur l'axe de la période, ou encore effectuer un double-clic;
2. Dans le menu apparaissant, cliquer sur le petit icône ressemblant à un histogramme et appelé **Options d'axe**.
3. Cocher la boîte appelée **Échelle logarithmique**.

Ces deux options sont représentées dans la figure ci-dessous.

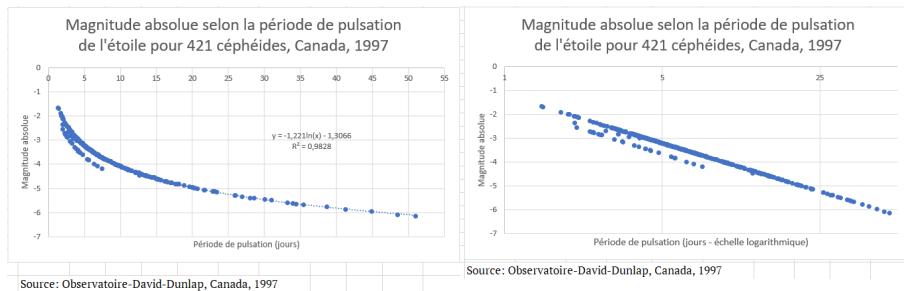


Figure 8.2.5 Comparaison entre la courbe de tendance logarithmique et l'échelle logarithmique

Pour la deuxième observation, déterminer la nature de la cause relèverait davantage de la physique que des statistiques. Il existe deux types de céphéides, pour lesquels la période et la magnitude absolue suivent en effet des relations différentes, mais toutes deux logarithmiques, mais une analyse des données a montré que la presque totalité des étoiles de cette base de données sont de type *I*.

Le calcul du coefficient de détermination linéaire pour ces deux variables est relativement fort. Par contre, dans les options de la courbe de tendance, il est possible de faire calculer le coefficient de détermination de la relation logarithmique, qui est encore plus fort. Afficher ce coefficient sur le graphique.

8.3 Réflexions

Ce laboratoire a montré comment déterminer la présence d'un lien linéaire entre deux variables, par le biais de l'observation (qualitativement) du nuage de points et par le calcul des coefficients de corrélation et de détermination (quantitativement). Ensemble, ces informations permettent de conclure à la présence ou non d'une relation linéaire ainsi que d'en quantifier la force.

De plus, on a vu qu'il est possible d'être en présence d'une relation qui n'est pas linéaire, mais qui peut être tout de même intéressante et très forte.

Travail à faire après le laboratoire

Objectifs

- Calculer des informations manquantes à l'aide de la régression linéaire.
 - Poser un regard critique sur les données.
 - Effectuer un test d'hypothèse paramétrique pour vérifier une hypothèse.
 - Interpréter adéquatement la conclusion d'un test.
 - Construire des intervalles de confiance pour estimer une variable à l'aide d'une régression.
1. Dans le tableau des données, on remarque que plusieurs étoiles n'ont pas de valeurs pour le rayon. On peut utiliser l'équation de la droite de régression (8.2.2) pour approximer ces données manquantes.
 - (a) Dans le tableau des données, filtrer la colonne **Rayon** afin d'afficher uniquement les étoiles dont la valeur du rayon est absente.
 - (b) Dans une nouvelle feuille de calcul, copier, le nom et la période de ces étoiles.
 - (c) À l'aide de l'équation (8.2.2), déterminer les rayons manquants dans cette feuille de calculs.
 - (d) Vérifier la précision de l'approximation en cherchant les véritables valeurs des rayons des étoiles AX_Cir (ID 207), R_Cru (ID 209), SU_Cyg (ID 225) et S_Sgr (ID 397). Citer les sources utilisées.
 2. Puisque le coefficient de corrélation r est une approximation du véritable facteur de corrélation (ρ) entre les variables et qu'il dépend du hasard de l'échantillon, ainsi que de sa taille, il est possible de faire un test d'hypothèse pour évaluer si le lien est significatif. L'hypothèse nulle est $H_0 : \rho = 0$, où ρ est la vraie valeur du facteur de corrélation entre les variables, et l'hypothèse alternative est $H_1 : \rho \neq 0$. La règle de décision stipule de rejeter H_0 lorsque $|T_{obs}| > t_{n-2;\alpha/2}$ où $t_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. Si l'on préfère utiliser la valeur P , qui dans ce cas vaut $2P(T_{n-2} > |t_{obs}|)$, on rejette H_0 si celle-ci est inférieure à α .
 - (a) Faire le test d'hypothèse au seuil de 5% en utilisant la première méthode.
 - (b) Faire le test d'hypothèse au seuil de 1% en utilisant la valeur P .
 - (c) Interpréter adéquatement ce test d'hypothèses.
 3. Lorsque l'on utilise l'équation de la droite de régression pour estimer une variable à l'aide d'une autre, ce que l'on fait est en réalité un calcul de moyenne conditionnelle, à savoir si les variables X, Y sont linéairement corrélées, alors l'équation $y = ax + b$ signifie aussi qu'en moyenne lorsque $X = x$, la variable Y sera égale à y . On peut utiliser ces informations pour construire un intervalle de confiance pour estimer une valeur de Y pour une valeur de X donnée, offrant ainsi plus de contrôle sur l'estimation. L'intervalle pour un niveau de confiance de $(a - \alpha)\%$ est de la forme

$$[ax + b - E; ax + b + E],$$

où la marge d'erreur E vaut, si s_x, s_y sont les écarts types estimés des variables X, Y ,

$$E = t_{n-2;\alpha/2} s_y \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}}.$$

- (a) Dans une nouvelle feuille de calculs appelée **IntervalleR**, faire une copie des couples de données période-rayon pour les couples dont le rayon n'est pas vide.
- (b) **EstimationEBinfBsupP dans intervalle**
- (c) À côté de ce tableau, calculer la moyenne des rayons, les écarts type pour les rayons et la période et le nombre de données dans le tableau. Incrire aussi la valeur $\alpha = 0,05$ et calculer la cote t nécessaire.
- (d) Dans le tableau, calculer les entrées de la colonne **Estimation** à l'aide de l'équation de la droite de régression obtenue pendant le laboratoire.
- (e) Remplir la colonne **E** en calculant les marges d'erreur.
- (f) Calculer les bornes inférieures et supérieures pour les intervalles de confiance dans les colonnes suivantes.

- (g) Dans la colonne ***P dans intervalle***, vérifier si la véritable période de chacune des étoiles se trouve dans l'intervalle de confiance construit.
- (h) Quelle est la proportion d'intervalles qui contiennent leur période? Commenter brièvement.

Appendice A

Bases de données

Les bases de données utilisées dans les laboratoires sont présentées.

A.1 Armée américaine

Selon un rapport démographique de la communauté militaire américaine¹, en 2010, l'armée américaine comptait 30,5% de membres issus de minorités ethniques. On introduit les données d'un échantillon de 6068 membres de l'armée américaine, prélevé en 2011². On pourrait supposer que la proportion de membres issus de minorité ethnique augmente à chaque année. C'est ce que l'on va vérifier en effectuant un test d'hypothèse sur une proportion. La base de données, modifiée pour répondre aux objectifs de ce laboratoire, a été obtenue du site³. La base de données à cette [adresse⁴](#) cette [adresse⁵](#) comprend les variables suivantes.

- l'identifiant du soldat;
- le genre du soldat;
 - Homme;
 - Femme;
- l'installation du soldat;
 - Camp Atterbury;
 - Camp Shelby;
 - Fort Bliss;
 - For Bragg;
 - Camp Drum;
 - Camp Gordon;
 - Camp Hood;
 - Camp Huachuca;
 - Camp Lee;
 - Camp McCoy;
 - Camp Rucker;

¹<https://download.militaryonesource.mil/12038/MOS/Reports/2020-demographics-report.pdf>, document consulté le 30 novembre 2024

²<https://www.openlab.psu.edu/ansur2/>, page consultée le 30 novembre 2024

³<https://www.openlab.psu.edu/ansur2/>

⁴github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Arm%C3%A9e.xls?raw=true

⁵github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Arm%C3%A9e.xls?raw=true

- Camp Stewart;
- la branche de service;
 - l'armée régulière;
 - garde nationale de l'armée;
 - réserve de l'armée
- l'ethnicité du soldat;
 1. blanc;
 2. noir;
 3. hispanique;
 4. asiatique (Cambodge, Chine, Corée du Sud, Japon, Laos, Myanmar, Taïwan, Vietnam);
 5. autochtone;
 6. îles du Pacifique;
 7. autre (Asie de l'est ou Inde, Moyen-Orient ou îles des Caraïbes);
- l'âge du soldat (en années);
- la grandeur du soldat (en pouces);
- le poids du soldat (en livres);
- la main dominante du soldat;
 - droite;
 - gauche;
 - aucune préférence.

A.2

[adresse¹](#)

A.3

[adresse¹](#)

A.4 Base canadienne de données sur les collisions

La base de données nationale sur les collisions s'agit « d'une base de données de toutes les collisions automobiles ayant fait l'objet d'un rapport de police et étant survenues sur les routes publiques du Canada en 2019 »¹. Le fichier Excel avec la base de données a été nettoyé pour faciliter le traitement statistique.

¹github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Bixi_ao%C3%BBt_2016.xlsx?raw=true

¹github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Cepheides.xlsx?raw=true

¹<https://ouvert.canada.ca/data/fr/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a>, page consultée le 30 novembre 2024

Les modalités de chaque variable peuvent être consultées sur le site web du gouvernement du Canada². Cette base de données comprend les variables suivantes.

- l'année de la collision;
- le mois de la collision;
- le jour de la semaine de la collision;
- l'heure de la collision;
- la sévérité de la collision;
- le nombre de véhicules en cause;
- le type de collision;
- l'endroit routier de la collision;
- les conditions météorologiques au moment de la collision;
- l'état de la chaussée;
- le segment de route où la collision a eu lieu;
- la signalisation;
- le type de véhicule impliqué dans la collision;
- l'année de modèle du véhicule;
- le sexe de la personne impliquée dans la collision;
- l'âge de la personne impliquée dans la collision;
 - Pour l'âge, les modalités *NN* et *UU* ont été remplacées par *-2* et *-1* respectivement afin de permettre des groupements dans un tableau croisé dynamique.
- le type de blessure;
- les mesures de sécurité utilisées;
- le type d'usager.

La base de données est en lien avec les individus impliqués dans une collision et non pas le nombre de collisions. Une collision est donc enregistrée plusieurs fois selon le nombre de personnes impliquées. Par exemple, si deux véhicules sont entrés en collision, chacun avec un conducteur et un passager, on retrouvera cette collision quatre fois dans la base de données.

La base de données se trouve à l'adresse suivante³.

²https://opencanada.blob.core.windows.net/opengovprod/resources/21eb7966-38da-4814-a80e-521bce6c4c27/data_dictionary.pdf?se=2024-12-03T16%3A24%3A17Z&sp=r&sv=2019-07-07&sr=b&sig=0FuNKQEYD1awb0RPtN42Qur6v0BaaDDL2Z69SprCqFM%3D, page consultée le 30 novembre 2024

³github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Collisions.xlsx?raw=true

A.5 Diabète

Les trois premiers laboratoires sont consacrés à l'étude de la maladie du diabète, en mettant l'accent sur les femmes d'origine pima. Les Pimas sont un peuple autochtone de la région du Sonora, au Mexique, et de l'État de l'Arizona, aux États-Unis. Les données utilisées proviennent du National Institute of Diabetes and Digestive and Kidney Diseases (l'Institut National du Diabète et des Maladies Digestives et Rénales des États-Unis). Les femmes ont été suivies sur une longue période de temps, soit de 1965 à 1995. Ce type de recherche, appelé étude longitudinale, vise à suivre l'évolution d'une variable dans le temps. L'année exacte de la collecte de données pour ce laboratoire n'est toutefois pas connue.

Selon Santé Canada, le diabète est défini comme « une maladie chronique qui se développe lorsque le corps ne produit pas l'insuline dont il a besoin pour transformer le sucre en énergie ou qu'il ne l'utilise pas efficacement »¹. Cette maladie se divise en trois catégories : le diabète de type 1, le diabète de type 2 et le diabète gestationnel. Le diabète de type 1 se caractérise par une production insuffisante d'insuline par le corps. Le diabète de type 2 est lié à une résistance du corps à l'insuline. Le diabète gestationnel, quant à lui, se manifeste par une élévation récente du taux de glucose dans le sang d'une femme enceinte.

Dans l'étude présentée, on se concentre sur le diabète de type 2, une forme de diabète fréquemment associée à l'inactivité physique, à l'obésité, à l'âge avancé d'un individu, ainsi qu'à des antécédents familiaux de diabète de type 2. Plusieurs communautés autochtones rencontrent des obstacles pour accéder à des soins de santé adéquats en raison de ressources limitées, de faibles revenus et de conditions économiques précaires.

Le fichier Excel à télécharger, ouvrir et enregistrer est *Données_Diabète.xlsx* disponible en cliquant [ici](#)².

La base de données de l'enquête comprend les mesures diagnostiques d'un échantillon de 768 femmes d'origine pima de l'Arizona. Les variables à l'étude sont :

- l'**identifiant** des participantes. L'éthique en matière de recherche exige l'anonymat des personnes participantes. Ainsi, un numéro est attribué à chaque individu afin d'éviter de divulguer leur identité;
- l'**âge** des participantes en année;
- la concentration de **glucose** plasmatique après deux heures lors d'un test de tolérance au glucose par voie orale en mg/dL. Un taux élevé de glucose est un signe précoce du diabète de type 2. Après ce test, une valeur considérée saine est inférieure à 140 mg/dL. Une valeur comprise entre 140 et 199 mg/dL est considérée comme un prédiabète. Une valeur de 200 mg/dL ou plus indique un diabète;
- la **pression artérielle diastolique** en mm Hg. La pression diastolique indique la pression dans les artères lorsque le cœur se repose entre deux battements. Une valeur comprise entre 60 et 80 est considérée normale. Une valeur entre 80 et 90 est qualifiée de préhypertension. Une valeur supérieure à 90 est classifiée comme hypertension;

¹<https://www.canada.ca/fr/sante-publique/services/maladies-chroniques/diabete.html>, page consulté le 26 août 2024

²github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Diab%C3%A8te.xlsx?raw=true

- l'épaisseur du pli cutané du triceps en mm;
 - l'**insuline** dans le sang en $\mu\text{U}/\text{mL}$. L'insuline est une hormone produite par le pancréas. Elle joue un rôle crucial dans le maintien de l'équilibre énergétique du corps et la régulation du taux de glucose sanguin. Après un repas riche en glucides, le taux de glucose peut augmenter rapidement; l'insuline intervient alors pour l'abaisser. En cas de production insuffisante d'insuline ou si le corps devient résistant à son action, le glucose reste en excès dans le sang, ce qui peut entraîner des maladies comme le diabète de type 1 ou le diabète de type 2. Dans l'étude des femmes d'origine pima, l'accent est mis sur le risque de développement du diabète de type 2;
 - l'**indice de masse corporelle** (IMC) en kg/m^2 . Comme l'indiquent ses unités, l'IMC est la valeur obtenue en divisant la masse d'un individu par sa taille au carré. Il s'agit d'un indicateur permettant d'estimer le surpoids d'une personne. Selon Statistique Canada, l'IMC fournit « un moyen de classer le poids corporel en fonction de risque pour la santé »³. Cependant, l'IMC n'est pas sans faille. Il ne fournit aucune information concernant la distribution de la matière grasse dans le corps;
 - les stades d'**obésité** selon Santé Canada (catégorie de l'IMC) :
 1. Poids insuffisant ($< 18,5$)
 2. Poids normal ($18,5 - 24,9$)
 3. Excès de poids ($25,0 - 29,9$)
 4. Obésité classe I ($30,0 - 34,9$)
 5. Obésité classe II ($35 - 39,9$)
 6. Obésité classe III ($\geq 40,0$)
 - la **fonction pedigree du diabète**. C'est un score mesurant le risque familial du diabète. Cette valeur mesure entre 0,08 et 2,42;
 - l'**atteinte** au diabète (avoir ou non le diabète).
1. , start=0 Non
2. , start=0 Oui

A.6 Polluants

Le 9 août 2024, la ville de Montréal a reçu une quantité de pluie record entre 158 et 173 mm en raison de la tempête tropicale Debby, entraînant d'importantes inondations et des dommages infrastructurels. Une question intéressante à examiner est la suivante : comment des précipitations intenses influencent-elles la qualité de l'air? Certaines études démontrent que « les précipitations intenses réduisent temporairement la concentration de particules fines ($PM_{2,5}$) et de dioxyde d'azote (NO_2) dans l'air en raison du lavage atmosphérique » (SOURCE? Wang, Y., et al. (2014), Eslami, A., et al. (2020), Kim, B. M., Park, J. H. (2001)). Cependant, l'impact d'intenses précipitations sur la qualité de l'air peut varier selon des facteurs locaux tels que différentes sources de polluants et les conditions météorologiques. Dans les zones urbaines à forte circulation ou avec plusieurs secteurs industriels, les améliorations de la qualité

³<https://www150.statcan.gc.ca/n1/pub/82-229-x/2009001/status/abm-fra.htm>

de l'air peuvent être temporaires si les polluants s'accumulent rapidement après l'arrêt de la pluie. Afin d'étudier cette question, les données de la qualité de l'air de Montréal pour l'année 2024 ont été recueillies¹. Cette base de données, disponible à cette [adresse](#)² comprend les variables suivantes.

- la station de laquelle les mesures ont été prélevées;
- le nom du polluant mesuré
 - l'ozone troposphérique, O_3 ;
 - le diazote de carbone, NO_2 ;
 - le monoxyde de carbone, CO ;
 - le dioxyde de soufre, SO_2 ;
 - les particules fines, $PM_{2,5}$.
- la valeur du polluant, tous en mg/m³;
- la journée du prélèvement;
- l'heure du prélèvement.

La base de données utilisée est disponible à l'adresse [suivante](#)³.

A.7 Précipitations

D'après les données récoltées entre 1991 à 2020 par Environnement Canada à la station météorologique de l'Aéroport international Pierre-Elliott-Trudeau de Montréal¹, la ville de Montréal reçoit en moyenne 77,2 mm de précipitations totales au mois de mars, soit environ 2,49 mm par jour. Les précipitations totales sont la somme de la pluie totale et de l'équivalent en eau de la neige totale en millimètres.

Les données échantillonnelles d'Environnement Canada pour le mois de mars 2024². La seule variable à l'étude est la quantité de précipitations totales en mm. La base de données utilisée est disponible à l'adresse [suivante](#)³.

A.8

La base de données sur l'analyse chimique de certains vins se trouve à l'adresse [suivante](#)¹.

¹<https://donnees.montreal.ca/dataset/rsqa-iqa-historique>, page consultée le 14 octobre 2024

²github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Polluant.xls?raw=true

³github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Pr%C3%A9cipitations.xls?raw=true

¹Données tirées de https://climat.meteo.gc.ca/climate_normals/results_1991_2020_f.html?searchType=stnName_1991&txtStationName_1991=montreal&searchMethod=contains&txtCentralLatMin=0&txtCentralLatSec=0&txtCentralLongMin=0&txtCentralLongSec=0&stnID=123000000&dispBack=1, page consultée le 21 novembre 2024

²Données tirées de https://climate.weather.gc.ca/climate_data/daily_data_e.html?StationID=51157&timeframe=2&StartYear=1840&EndYear=2024&Day=8&Year=2024&Month=3#, page consultée le 21 novembre 2024

³github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Pr%C3%A9cipitations.xls?raw=true

¹github.com/JeanSebastienTurcotte/LabosStats/blob/main/assets/Base%20de%20donn%C3%A9es/Donn%C3%A9es_Vins.xlsx?raw=true

Appendice B

Importer des données

Il existe plusieurs formats de base de données et plusieurs manières de les importer. On ne considère que l'importation des données provenant d'un fichier `csv`. Un fichier `csv` est un fichier texte où chaque ligne du texte correspond à une ligne du tableau et les virgules correspondent aux séparations entre les colonnes. Les portions de texte séparées par une virgule correspondent ainsi aux contenus des cellules du tableau. Une ligne est une suite ordonnée de caractères terminée par un caractère de fin de ligne.

Excel est en mesure d'ouvrir directement un fichier `csv`, mais il est préférable de ne pas travailler dans le fichier des données brutes, pour ne pas accidentellement les modifier ou les corrompre. Les étapes pour importer un fichier `csv` dans Excel et le convertir en format `xlsx` sont présentées ci-dessous.

1. Ouvrir un classeur vide d'Excel.
2. Cliquer sur l'onglet **Données** (voir la [Figure B.0.1](#)).

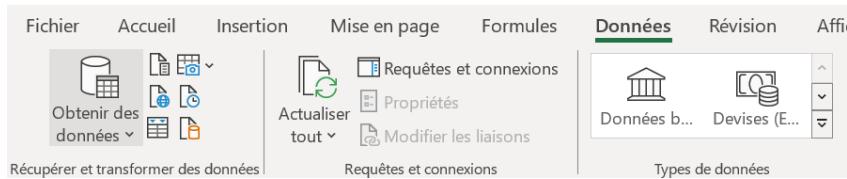


Figure B.0.1 Sélection de l'onglet **Données**

3. Dans le groupe **Récupérer et transformer des données**, cliquer sur la flèche du menu déroulant de l'icône **Obtenir des données** (voir la [Figure B.0.1](#)).
4. Sélectionner l'option **À partir d'un fichier**, suivi de l'option **À partir d'un fichier texte/CSV** (voir la [Figure B.0.2](#)).

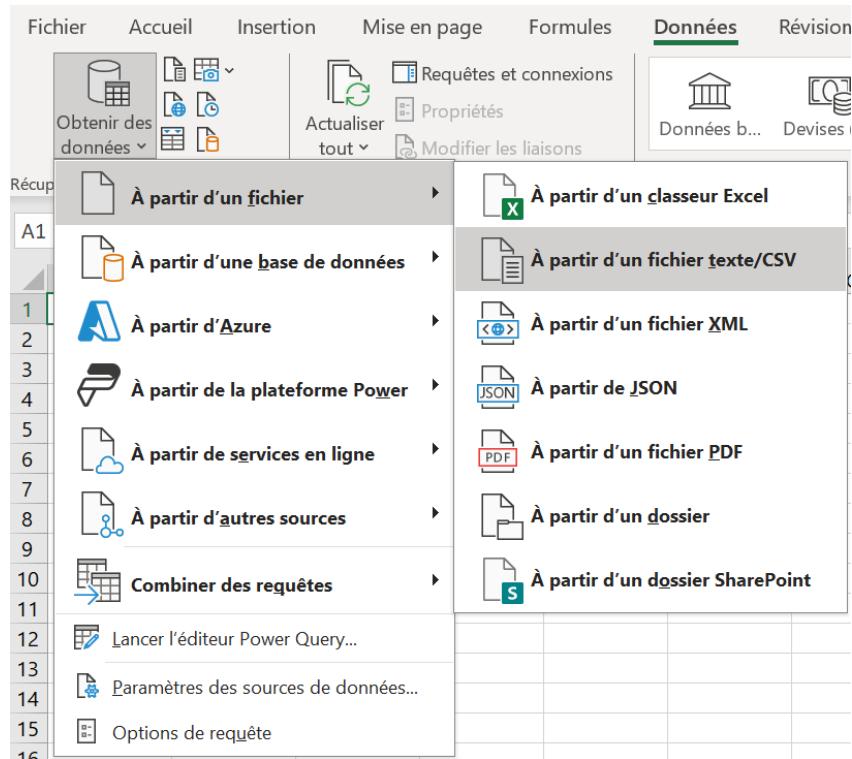


Figure B.0.2 Sélection de l'onglet *À partir d'un fichier texte/CSV*

5. Accéder au fichier “diabète.csv” sauvegardé sur l’ordinateur. Sélectionner-le et cliquer **Importer** (voir la [Figure B.0.3](#)).

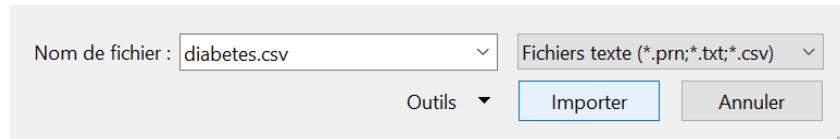


Figure B.0.3 Importer un fichier de format csv

6. Une boîte de dialogue s’ouvrira. Il faut s’assurer que le délimiteur est bien **Virgule**. Une fois fait, cliquer sur l’option **Charger** (voir la [Figure B.0.4](#)).

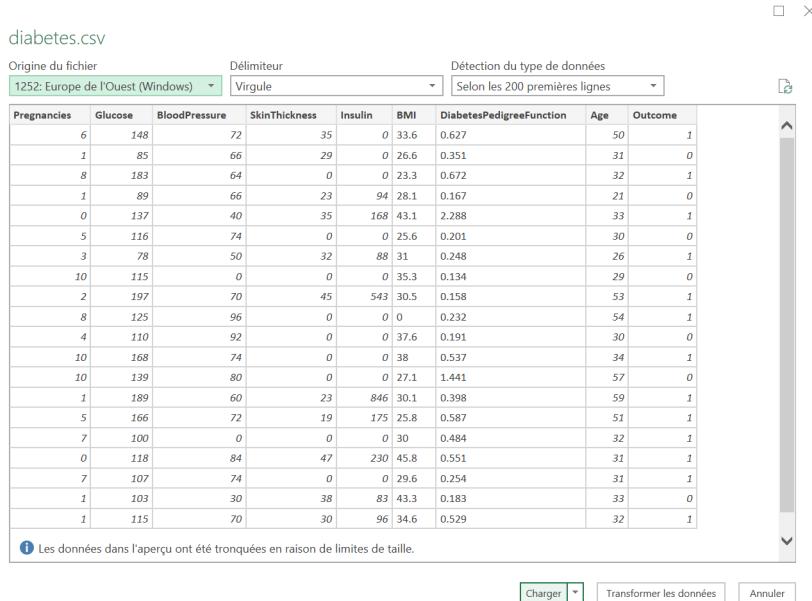


Figure B.0.4 Charger les données en format `xlsx`

7. Un tableau adapté pour travailler dans Excel est maintenant importé (voir la [Figure B.0.5](#)).

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0 33.6	0.627		50	1
3	1	85	66	29	0 26.6	0.351		31	0
4	8	183	64	0	0 23.3	0.672		32	1
5	1	89	66	23	94 28.1	0.167		21	0
6	0	137	40	35	168 43.1	2.288		33	1
7	5	116	74	0	0 25.6	0.201		30	0
8	3	78	50	32	88 31	0.248		26	1
9	10	115	74	0	0 35.3	0.134		34	1
10	2	197	70	45	543 30.5	0.158		53	1
11	8	125	96	0	0 0	0.232		54	1
12	4	110	92	0	0 37.6	0.191		30	0
13	10	168	74	0	0 38	0.537		34	1
14	10	139	80	0	0 27.1	1.441		57	0
15	1	189	60	23	846 30.1	0.398		59	1
16	5	166	72	19	175 25.8	0.587		51	1
17	7	100	0	0	0 30	0.484		32	1
18	0	118	84	47	230 45.8	0.551		31	1
19	7	107	74	0	0 29.6	0.254		31	1
20	1	103	30	38	83 43.3	0.183		33	0
21	1	115	70	30	96 34.6	0.529		32	1

Figure B.0.5 Tableau résultant dans Excel

Appendice C

Manipuler la feuille de calcul

Une feuille de calcul est l'endroit principal où se déroule le travail dans Excel. Bien que l'on puisse tenter de tout prévoir, il est fréquent qu'il devienne nécessaire de déplacer, ajouter ou supprimer des éléments. On distingue trois types d'éléments: la cellule unique, la ligne et la colonne.

Dans un premier temps, pour simplement effacer le contenu d'une cellule, on peut appuyer sur la touche **backspace**. Toutefois, si plus d'une cellule sont sélectionnées, seul le contenu de la cellule active sera supprimé. Pour effacer le contenu de toutes les cellules de la sélection, il vaut mieux utiliser la touche **suppr** (**del** en anglais).

Pour ajouter une cellule à un endroit, il faut faire un clic droit où l'on veut la nouvelle cellule et ensuite décider de la manière de faire l'insertion. On peut

- décaler la cellule active vers la droite,
- décaler la cellule active vers le bas,
- décaler la ligne entière vers le bas,
- décaler la colonne entière vers la droite.

Pour insérer une nouvelle ligne ou une nouvelle colonne, on peut effectuer un clic droit sur l'étiquette de la ligne ou de la colonne et cliquer sur **Insérer** ou encore utiliser l'option décrite ci-dessus dans l'ajout d'une cellule.

D'une manière semblable, on peut supprimer une cellule. Ceci est une opération différente de supprimer son contenu. La suppression d'une cellule requiert de décider ce que l'on doit faire avec l'espace qu'occupait la cellule. Il est possible de

- supprimer un cellule et ramener les cellules vers la gauche,
- supprimer un cellule et ramener les cellules active vers le haut,
- supprimer toute la ligne et ramener les lignes vers le haut,
- supprimer toute la colonne et ramener les colonnes vers la gauche.

Parfois, certaines lignes ou certaines colonnes ne sont utiles que pour faire des calculs et on ne souhaite pas les afficher. Il est possible de masquer une ligne ou une colonne. Pour cacher une ligne ou une colonne:

- On peut faire un clic droit sur l'étiquette de la ligne ou de la colonne et appuyer sur **Masquer**;
- Aller sous l'onglet **Accueil** et cliquer sur **Format**, puis **Masquer & afficher** et finalement, **Masquer les lignes** ou **Masquer les colonnes**

Pour afficher une ligne ou une colonne, on peut sélectionner les lignes ou les colonnes adjacentes et

- faire un clic droit et appuyer sur **Afficher**;
- aller sous l'onglet **Accueil** et cliquer sur **Format**, puis **Masquer & afficher** et finalement, **Afficher les lignes** ou **Afficher les colonnes**.

On peut aussi grouper des lignes ou des colonnes voisines afin de les masquer ou de les afficher. Pour cela, on sélectionne les lignes ou les colonnes que l'on souhaite grouper et on effectue la combinaison **Alt**+**Shift**+**→**. Pour dissocier les lignes ou les colonnes, on les sélectionne et on effectue la combinaison **Alt**+**Shift**+**←**.

Appendice D

Divers

ToDo

D.1 Mise en forme

Plusieurs aspects de la mise en forme et du format de l'affichage sont disponibles dans le ruban de l'onglet *Accueil*. En voici une liste non exhaustive.

Liste D.1.1 Éléments de mise en forme

Format de cellule en pourcentage	Pour faire en sorte que le contenu d'une cellule s'affiche en pourcentage, on change le menu déroulant à <i>Pourcentage</i> ou encore on clique sur le bouton juste en-dessous. Aussi, voir le raccourci Format de cellule en pourcentage .
---	---

Appendice E

Fonctions utiles

On introduit des fonctions Excel qui peuvent être utiles, mais qui ne sont pas nécessairement des fonctions statistiques.

E.1 Adresse

La première est la fonction **ADRESSE**, qui permet de retourner sous forme de texte le nom d'une cellule. Par exemple, si l'on tape =ADRESSE(2;1) dans une cellule quelconque, le résultat sera \$A\$2, puisque cette cellule est dans la deuxième ligne et dans la première colonne.

La forme générale de la formule et de ses arguments est **ADRESSE(no_lig, no_col, [no_abs], [a1], [feuille_texte])**, où les arguments entre crochets sont facultatifs. Ces arguments de la fonction ADRESSE sont:

no_lig	Le numéro de la ligne de la cellule visée;
no_col	Le numéro de la colonne de la cellule visée;
no_abs	Le type de référence absolue souhaitée;
a1	Format de l'adresse;
feuille_texte	Si spécifié, le nom de la feuille de travail donnée en argument fera partie de l'adresse.

Pour ce qui est du type de référence, la table suivante présente les options. Pour en savoir davantage sur le fonctionnement des références absolues, voir l'annexe [provisional cross-reference: annexe-references].

Table E.1.1 Type de référence associé à l'argument no<underscore>abs

no_abs	Type de référence
1 ou omis	Absolue
2	Ligne absolue, colonne relative
3	Ligne relative, colonne absolue
4	Relative

Pour l'argument **a1**, il ne sera pas utilisé dans ce manuel, mais permet de passer de la forme par défaut A3 (**a1=VRAI** ou omis) à une adresse dont le format est de la forme L1C3 (**a1=FAUX**).

La fonction ADRESSE seule n'est pas très utile. C'est en la combinant avec d'autres fonctions qu'elle peut être exploitée efficacement.

E.2 Indirect

L'une de ces fonctions est appelée INDIRECT. Cette fonction renvoie le contenu d'une cellule située à l'adresse donnée. Par exemple, si l'on tape =INDIRECT("A2") dans une cellule quelconque, le résultat sera le contenu de la cellule A2. La figure suivante illustre cela.

	A	B	C	D	E
1	1				
2	2				
3	1				
4	3				
5	1			=INDIRECT("A2")	
6	4				
7	1				
8	5				
9	1				
10	6				
11	1				
12	7				
13					

Figure E.2.1 Utilisation simple de la fonction INDIRECT

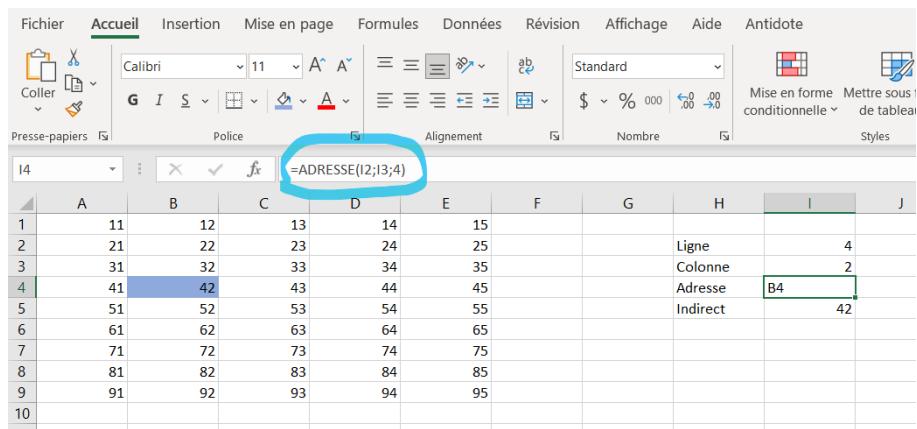
La fonction INDIRECT nécessite deux arguments, dont un facultatif. La syntaxe est INDIRECT(réf_texte, [a1]). Les arguments sont

- ref_texte** Une référence à une cellule sous forme de texte;
- a1** Paramètre booléen pour spécifier le type de référence (VRAI ou omis pour A3 ou FAUX pour L1C3).

La combinaison des fonctions ADRESSE et INDIRECT est naturelle et permet d'obtenir le contenu d'une cellule à l'aide d'une formule complexe. Ces deux fonctions seront utilisées ensemble dans la prochaine sous-section, mais voici une illustration simple de leur utilisation combinée.

Dans la figure Figure E.2.2, la plage A1:E5 contient des nombres arbitraires. Dans la cellule I2, on a écrit le numéro d'une ligne, et dans la cellule I3, celui d'une colonne. Dans la cellule I4, on a utilisé la fonction ADRESSE pour obtenir l'adresse de la cellule se trouvant à la ligne quatrième ligne deuxième

colonne. Finalement, on obtient le contenu de la cellule B4 en utilisant la fonction INDIRECT dans la cellule I5.



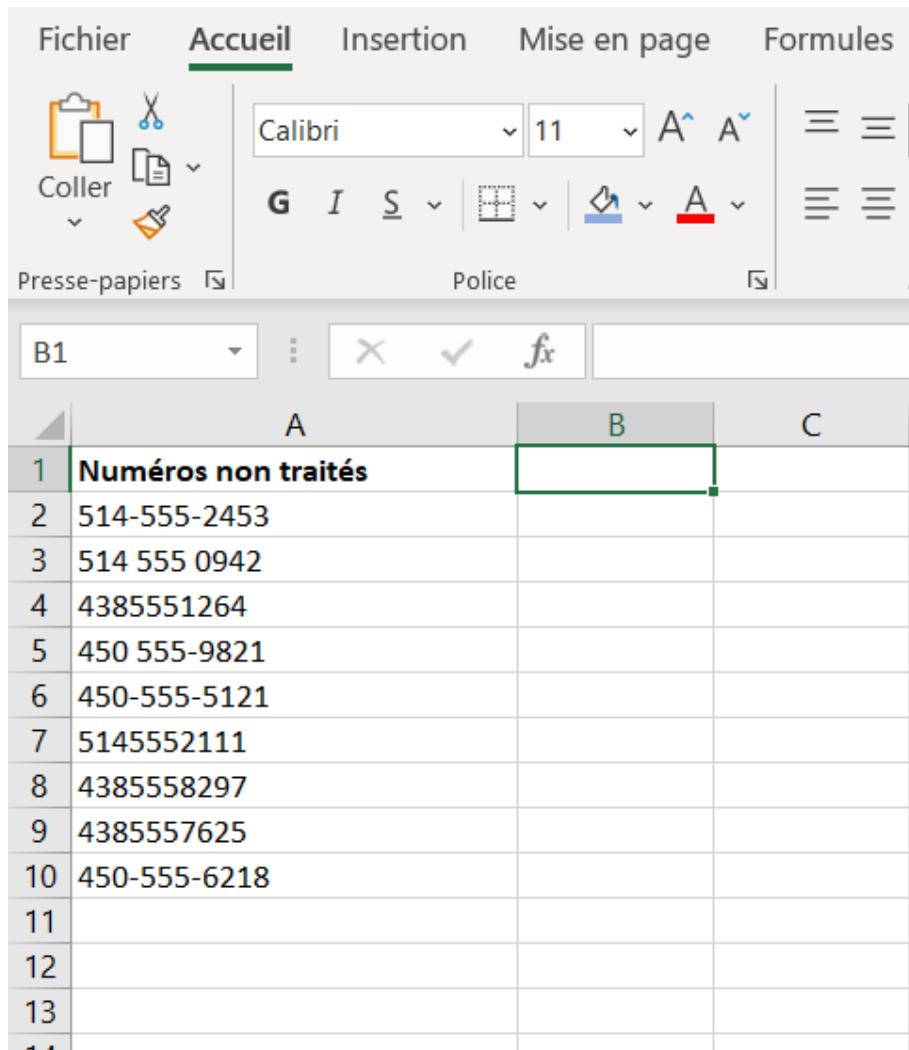
The screenshot shows a Microsoft Excel spreadsheet with the following details:

- Cell I4:** Contains the formula `=ADRESSE(I2;I3;4)`.
- Cell I5:** Contains the formula `=INDIRECT(I4)`, which is highlighted with a green border.
- Table Data:** A 10x5 grid of numbers from 11 to 55. Row 1: 11, 12, 13, 14, 15. Row 2: 21, 22, 23, 24, 25. Row 3: 31, 32, 33, 34, 35. Row 4: 41, 42, 43, 44, 45. Row 5: 51, 52, 53, 54, 55. Row 6: 61, 62, 63, 64, 65. Row 7: 71, 72, 73, 74, 75. Row 8: 81, 82, 83, 84, 85. Row 9: 91, 92, 93, 94, 95. Row 10: 101, 102, 103, 104, 105.
- Formula Bar:** Shows the formula `=ADRESSE(I2;I3;4)`.
- Excel ribbon:** Shows the Accueil tab selected. Other tabs include Insertion, Mise en page, Formules, Données, Révision, Affichage, Aide, and Antidote.
- Contextual ribbon:** Shows options for Presse-papiers, Police, Alignement, Nombre, and Styles.
- Table Tools ribbon:** Shows options for Mise en forme conditionnelle, Mettre sous forme de tableau, and Styles.

Figure E.2.2 Utilisation de INDIRECT et ADRESSE

E.3 Substitue

Une autre fonction pouvant se combiner avec ADRESSE est la fonction SUBSTITUE. Seule, cette fonction permet de remplacer du texte et est déjà en soi très utile selon le domaine. Excel est bon pour prolonger une suite logique avec l'outil croix lorsque celle-ci est numérique, mais si on veut glisser en modifiant du texte, il a besoin d'aide. Dans la table ci-dessous, on illustre une situation où un gestionnaire a une liste de numéros de téléphone dont il aimeraient uniformiser le format. Certains des numéros ont des tirets, d'autres ont des espaces entre l'indicatif régional et à l'intérieur du numéro local. La commande Substitue peut enlever les tirets et les espaces. Elle aurait aussi pu ajouter des tirets ou des espaces, voire même une combinaison des deux, selon la préférence de l'utilisateur.



	A	B	C
1	Numéros non traités		
2	514-555-2453		
3	514 555 0942		
4	4385551264		
5	450 555-9821		
6	450-555-5121		
7	5145552111		
8	4385558297		
9	4385557625		
10	450-555-6218		
11			
12			
13			
14			

Une liste de numéros de téléphone

Figure E.3.1 Liste de numéros de téléphone

Par exemple, pour supprimer les tirets du premier numéro, on peut utiliser la commande =SUBSTITUE(A2;"-";""). Le premier argument est l'emplacement du texte que l'on souhaite modifier, le deuxième est le texte à modifier (ici le tiret) et le troisième est ce par quoi on veut le remplacer (ici on le remplace par une chaîne vide, ce qui a pour effet de supprimer le tiret). Pour supprimer les espaces du numéro de la cellule A3, on pourrait modifier la formule précédente, mais ce ne serait pas efficace, compte tenu du fait qu'il existe des numéros avec à la fois des tirets et des espaces, comme celui dans la cellule A5. Il est possible d'imbriquer des formules l'une dans l'autre afin de modifier les deux caractères. En remplaçant la formule dans la cellule B1 par =SUBSTITUE(SUBSTITUE(A2;"-";"");" ";""), on pourra utiliser l'outil croix pour descendre le long de la liste et corriger le format des numéros. La figure ci-dessous présente le résultat final.

	A	B
1	Numéros non traités	Suppression des espaces et des tirets
2	514-555-2453	5145552453
3	514 555 0942	5145550942
4	4385551264	4385551264
5	450 555-9821	4505559821
6	450-555-5121	4505555121
7	5145552111	5145552111
8	4385558297	4385558297
9	4385557625	4385557625
10	450-555-6218	4505556218
11		

Une liste de numéros de téléphone dans le même format

Figure E.3.2 Liste de numéros de téléphone dans le même format

On imagine que l'on doive appliquer une formule 100 fois à l'aide de l'outil croix. Si l'on veut glisser verticalement, il peut être facile de compter jusqu'où on doit aller, mais s'il faut glisser horizontalement, le calcul est plus ardu. La commande `SUBSTITUE` avec la commande `ADRESSE` permet de déterminer le nom de la $n^{\text{ième}}$ colonne. Dans un premier temps, on peut avoir l'adresse de la $n^{\text{ième}}$ colonne avec la commande `=ADRESSE(1; n; 4)` (On pourrait remplacer 1 par n'importe quel nombre, cet argument correspond au numéro de ligne). Si l'on souhaite uniquement avoir le nom de la colonne, alors il suffit de remplacer le numéro de la ligne par une chaîne de caractères vide. Ainsi, la commande finale sera `=SUBSTITUE(ADRESSE(1; n; 4); "1"; "")`. En mettant les valeurs de n dans la colonne A et en appliquant cette formule dans la colonne B, on obtient le résultat de l'image ci-dessous. On remarque entre autre chose que le nombre de colonnes atteint son maximum à 16 385.

The screenshot shows a Microsoft Excel interface with the following details:

- Menu Bar:** Fichier, Accueil, Insertion, Mise en page, Formules, Données, Révision.
- Toolbar:** Includes icons for Coller (Paste), Presse-papiers (Clipboard), Calibri font, font size 11, bold, italic, underline, alignment options, and a formula bar.
- Formula Bar:** Displays the formula `=SUBSTITUE(ADRESSE(1;A2;4);"1";"")`.
- Table:** A 15x2 table with columns labeled A and B.
 - Row 1: Col A contains "n", Col B contains "Nom de la n^{ième} colonne".
 - Row 2: Col A contains "1", Col B contains "A".
 - Row 3: Col A contains "5", Col B contains "E".
 - Row 4: Col A contains "26", Col B contains "Z".
 - Row 5: Col A contains "42", Col B contains "AP".
 - Row 6: Col A contains "100", Col B contains "CV".
 - Row 7: Col A contains "1024", Col B contains "AMJ".
 - Row 8: Col A contains "16384", Col B contains "XFD".
 - Row 9: Col A contains "16385", Col B contains "#VALEUR!".
 - Row 10: Col A is empty, Col B has a small icon.
 - Rows 11 to 15: Both columns are empty.

Une feuille Excel montrant comment obtenir le nom de la $n^{\text{ième}}$ colonne.

Figure E.3.3 Nom de la $n^{\text{ième}}$ colonne

Appendice F

Les macros

F.1 Sauvegarder un fichier avec l'extension `xlsm`

Ce laboratoire utilise des canevas à remplir, créés pour effectuer des tests d'hypothèses paramétriques. Ces canevas contiennent des macros, une série d'instructions ou d'actions qu'un utilisateur peut exécuter automatiquement pour automatiser une tâche répétitive comme effacer le contenu d'une plage de cellules. Pour exécuter des macros, il faut sauvegarder les fichiers Excel avec l'extension `xlsm` au lieu de `xlsx`.

Les étapes pour sauvegarder un fichier Excel avec l'extension `xlsm` sont présentées ci-dessous.

1. Créer une copie du fichier Excel *Données_Diabète.xlsx*. Nommer ce fichier *Test_Diabète.xlsx*.
2. Ouvrir le fichier *Test_Diabète.xlsx*.
3. Cliquer sur l'onglet **Fichier** du ruban.
4. Sélectionner l'onglet **Enregistrer une copie**.
5. Cliquer sur la flèche du menu déroulant et sélectionner le deuxième choix de type de fichier, soit l'extension de fichier `xlsm` (voir la [Figure F.1.1](#)). Ceci permet la sauvegarde d'un fichier Excel prenant en charge les macros.

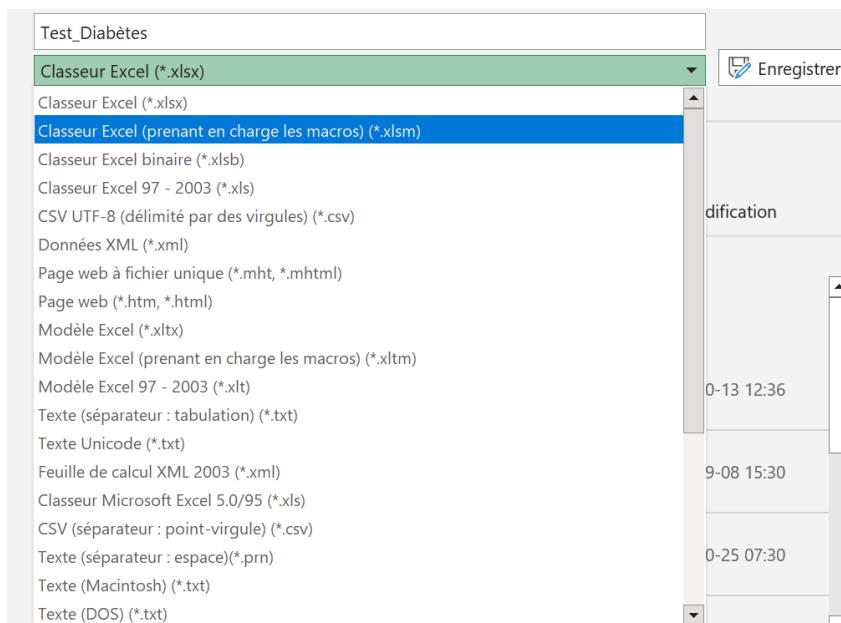


Figure F.1.1 Sélection de l'extension de fichier **xlsm**

6. Sauvegarder le fichier dans le répertoire désiré.

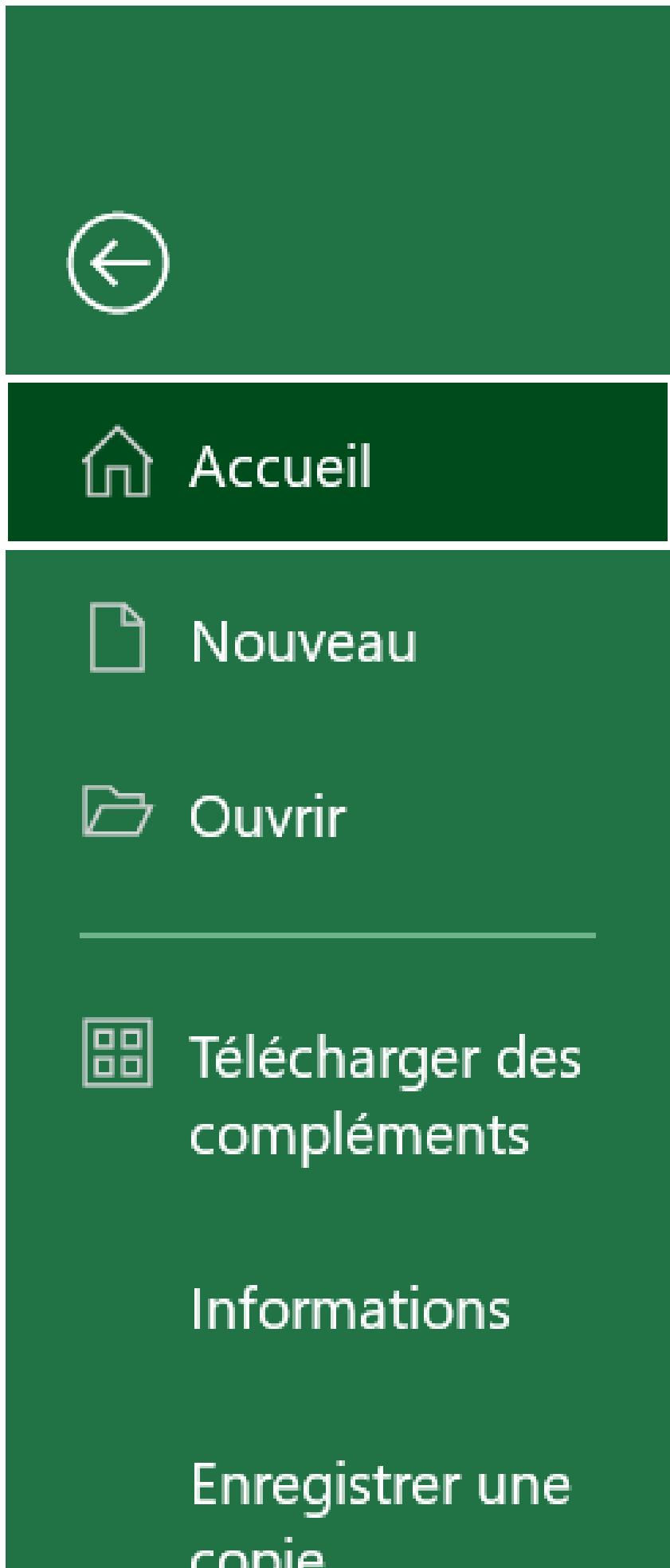
Le fichier **Test_Diabètes.xlsm** est sauvegardé. Répéter les étapes avec le fichier **Données_Polluant.xlsx**.

F.2 Afficher l'onglet Développeur

Puisque ce laboratoire utilise des canevas qui contiennent des macros, il faut afficher l'onglet **Développeur** dans le ruban d'Excel.

Les étapes pour afficher l'onglet sont présentées ci-dessous.

1. Ouvrir le fichier **Test_Diabètes.xlsm**.
2. Cliquer sur l'onglet **Fichier** du ruban.
3. Sélectionner l'option **Autres** (voir la [Figure F.2.1](#)).



4. Des options s'affichent. Sélectionner l'onglet *Options* (voir la Figure F.2.2).

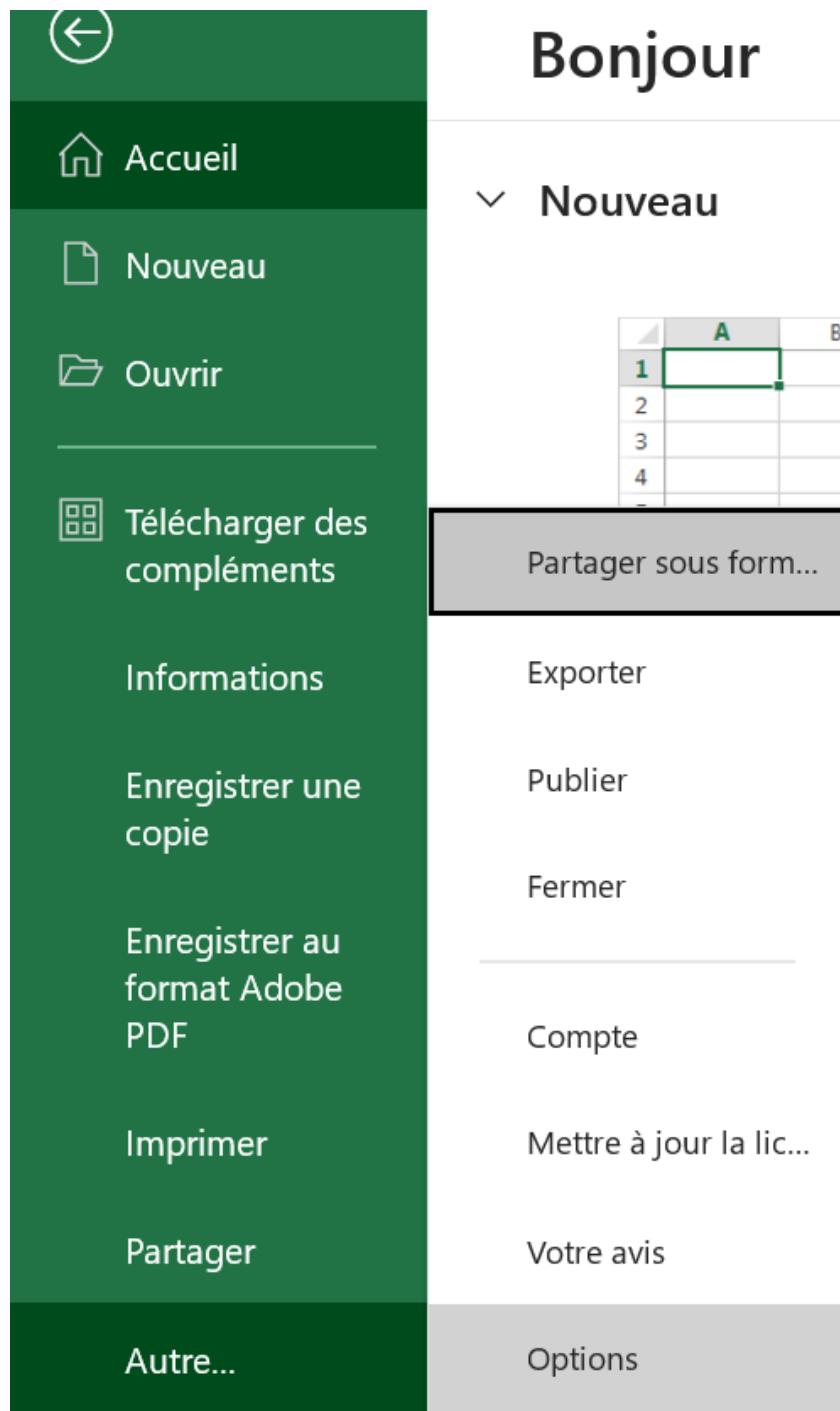


Figure F.2.2 Sélection de l'onglet *Options*

5. Une boîte de dialogue s'affiche présentant des onglets dans sa partie gauche. Sélectionner l'onglet *Personnaliser le ruban* (voir la Figure F.2.3).
6. À droite de la boîte de dialogue, certains onglets principaux sont cochés. Il faut glisser la barre déroulante vers le bas et trouver l'icône *Développeur*

(voir la [Figure F.2.3](#)).

7. Cocher l'encadré à gauche de **Développeur** (voir la [Figure F.2.3](#)). Ceci fait en sorte que l'onglet **Développeur** apparaît parmi les onglets principaux d'Excel.

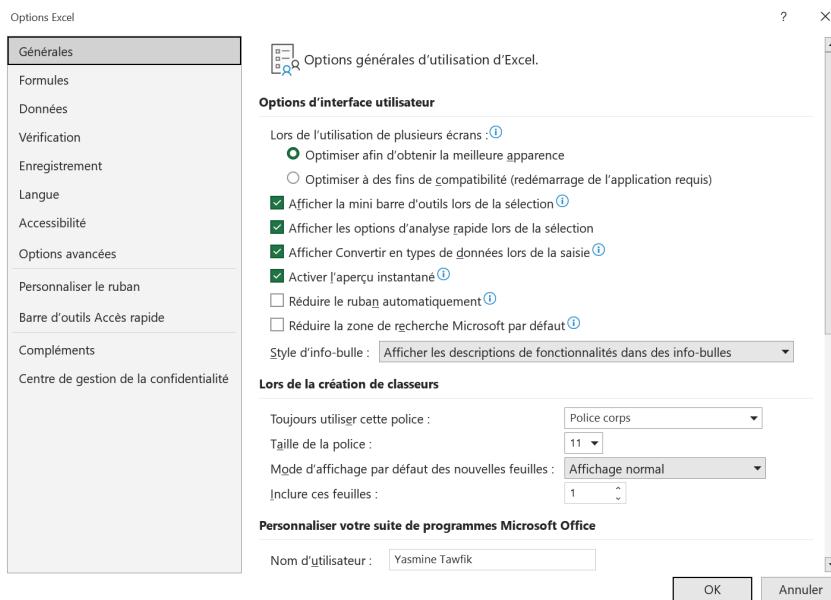


Figure F.2.3 Sélection des onglets **Personnaliser le ruban** et **Développeur**

8. Cliquer sur **OK**.

Appendice G

Raccourcis pratiques

Cette annexe présente quelques raccourcis claviers qui permettent de sauver du temps dans Excel. Lorsqu'il est mention d'une combinaison, par exemple la combinaison [Shift]+[Enter], cela signifie qu'il faut maintenir la touche qui précède le symbole + enfoncee et ensuite appuyer sur la ou les touches qui suivent.

G.1 Utilité

Si l'on est en mode édition d'une cellule et que l'on veut le quitter, il suffit d'appuyer sur la touche [Esc] (ou [échap]).

Deux des raccourcis claviers les plus utilisés sont sans aucun doute les fameux copier-coller. Dans Excel, la fonction coller est un peu différente, car on peut coller de plusieurs façon. Puisqu'une cellule peut contenir des formules, référençant potentiellement d'autres cellules, il faut déterminer lors du collage si l'on veut copier ces formules ou si l'on souhaite copier uniquement les valeurs des cellules.

Pour comprendre l'importance de la distinction entre ces deux méthodes, on considère l'exemple suivant. Dans la première colonne, on a écrit les entiers de 1 à 5. Dans la deuxième colonne, on a inscrit la formule qui double l'entrée de la colonne A. Si l'on veut copier le contenu de la colonne B, on peut avoir (minimalemen) deux intentions:

1. Copier la formule, dans le but de continuer la multiplication par 2;
2. Copier les valeurs, dans le but de les retranscrire ailleurs pour réutilisation.

Si l'on copie les valeurs à l'aide de la combinaison [CTRL]+[C], par défaut Excel va aussi copier les formules. Pour coller les valeurs seules, on peut appliquer la combinaison [CTRL]+[Alt]+[V] afin d'ouvrir le menu de collage et appuyer sur [V] à nouveau pour sélectionner l'option *Valeurs*.

G.2 Navigation

Ces premiers raccourcis sont simples, mais très utiles. Lorsqu'une cellule est sélectionnée, on peut naviguer vers les cellules voisines en cliquant sur les flèches du clavier [←],[→],[↑],[↓]. De manière équivalente, on peut se déplacer vers le bas avec la touche [Enter] ou vers le haut avec la combinaison [Shift]+[Enter] et

vers la droite avec la touche **Tab** ou vers la gauche avec la combinaison **Shift** + **Tab**.

Remarque G.2.1 Navigation dans un ensemble de cellules sélectionnées. Si un ensemble de cellules avait été préalablement sélectionné, alors la navigation avec **Enter** et **Tab** restera dans la sélection, changeant de colonne ou de ligne au besoin.

Un clic sur le bouton **Home** change la cellule active pour la première cellule de la ligne alors que la combinaison **Ctrl**+**Home** ramène à la première cellule de la feuille, soit celle en haut à gauche (normalement identifiée A1).

G.3 Sélection variée

Il est souvent utile de sélectionner un ensemble de cellules, parfois rapprochées. Les quatre raccourcis suivants permettent de sélectionner la cellule active ainsi que tous ses voisins non vides dans l'une des quatre directions:

- La combinaison **Ctrl**+**Shift**+**←** permet de sélectionner la cellule active et tout ce qui est non vide à sa gauche.
- La combinaison **Ctrl**+**Shift**+**→** permet de sélectionner la cellule active et tout ce qui est non vide à sa droite.
- La combinaison **Ctrl**+**Shift**+**↑** permet de sélectionner la cellule active et tout ce qui est non vide au-dessus.
- La combinaison **Ctrl**+**Shift**+**↓** permet de sélectionner la cellule active et tout ce qui est non vide en-dessous.

À noter qu'une utilisation additionnelle de ces combinaisons sélectionnera toutes les cellules voisines de la sélection actuelle, y compris les cellules vides, jusqu'à la prochaine cellule non vide ou jusqu'à la fin de la feuille.

Si l'on souhaite plutôt sélectionner la ligne ou la colonne en totalité, alors en plus de pouvoir cliquer sur l'étiquette de la dite ligne ou colonne, on peut également utiliser les combinaisons **Shift**+**Espace** et **Ctrl**+**Espace** pour sélectionner respectivement la ligne ou la colonne de la cellule active.

Remarque G.3.1 Sélection dans un tableau. Si la cellule active fait partie d'un tableau, alors l'application des raccourcis ci-dessus est d'abord limitée au tableau.

On peut sélectionner toutes les données d'un tableau à l'aide de la combinaison **Ctrl**+**A**. À l'extérieur d'un tableau, cette combinaison sélectionne l'entièreté de la feuille de calcul.

G.4 Mise en forme

Pour la mise en forme, il existe plusieurs raccourcis qui sont prédéfinis. En voici une liste des principaux qui pourraient être utilisés.

Liste G.4.1

Format de cellule en pourcentage	Pour que le format de cellule passe à Pourcentage , on effectue la combinaison Ctrl + Shift + % .
---	---

Colophon

This book was authored in PreTeXt.