Statistiques à deux variables

Matthieu Barreau – 23 avril 2016

1 Qu'est ce que ça veut bien vouloir dire ?

Les statistiques en général permettent de donner des repères, de pouvoir évaluer en un coup d'oeil un gros ensemble de données.

Exemple *A quoi servent les statistiques ?*

Par exemple, si vous regardez les accidents de la route, la liste de tous les accidents s'étant produit entre 2010 et 2015 n'est sans aucun doute pas le meilleur moyen de savoir s'il y a eu plus d'accidents en 2015 qu'en 2014. Le gouvernement aurait également besoin de savoir si les réformes qu'il fait réduisent le nombre d'accidents.
En revanche, si on regroupe les accidents par années et que l'on affiche dans un tableau chaque année, on trouve quelque chose de beaucoup plus facile à lire :

Année	2010	2011	2012	2013	2014	2015
Nombre d'accidents	67288	65024	60437	56812	58191	56109

Faire des statistiques, c'est donc juste « simplifier un grand ensemble de données ».

Les deux années précédentes, vous avez étudié les **statistiques à une variable**, c'est à dire qu'un seul paramètre pouvait changer.

Exemple *Statistique à une variable*

Le lancé d'un dé à 6 faces est une **expérience aléatoire** dont seul le résultat était variable.
On pourrait aussi piocher une carte au hasard, ou bien jouer au loto...

Ces statistiques à une variable sont pratiques pour comprendre le vocabulaire et comment les statistiques marchent mais les situations sont un peu limitées...

Les statistiques à **deux variables** sont beaucoup plus intéressantes car souvent plus proches de la réalité et pourtant pas beaucoup plus difficiles.

Définition

Une **série statistique à deux variables** possède **deux .....**. On note souvent une donnée sous forme de couple : (1^{er}caractère; 2^{nd}caractère).

Remarque

Si un des caractères est **le temps**, on dit que la série est **.....**.

Donnons quelques exemples pour mieux comprendre cette notion.

Exemples

Nous pouvons reprendre notre exemple des accidents de la route. Cette série peut être vue comme une série statistique à deux variables : le nombre d'accident en fonction de l'année. Comme elle dépend du temps, elle est chronologique.

Année	2010	2011	2012
Nombre d'accidents	67288	65024	60437
Couple	(2010;67288)	(2011;65024)	(.....;.....)

Année	2013	2014	2015
Nombre d'accidents	56812	58191	56109
Couple	.....	.....	.....

D'autres exemples pourraient être le temps qu'il fait en fonction du jour, le salaire moyen en fonction de l'année...
Bref, il y a plein d'exemples comme vous pouvez le voir.

Remarque

Qui est x ? Et qui est y ?

Le choix de la variable x comme l'année est très important. En effet, si on avait choisi y = année dans l'exemple précédent, cela voudrait dire que l'année dépend du nombre d'accidents, autrement dit, si on a 50000 accidents alors on est en 2017... ce qui est absurde !

La plupart du temps, nous allons utiliser la calculette / l'ordinateur pour résoudre des problèmes de statistiques à deux variables car c'est souvent trop laborieux de le faire soit-même.

2 Représentation en nuage de points

Faire des tableaux c'est une bonne idée, mais on préfère toujours faire des graphiques, c'est plus facile à lire et on comprend souvent mieux.

Pour représenter une série statistique à deux variables, l'outil que l'on utilise s'appelle le ..... Ce beau nom un peu poétique veut dire que nous allons dessiner un point M pour chaque couple de donnée $(x; y)$.

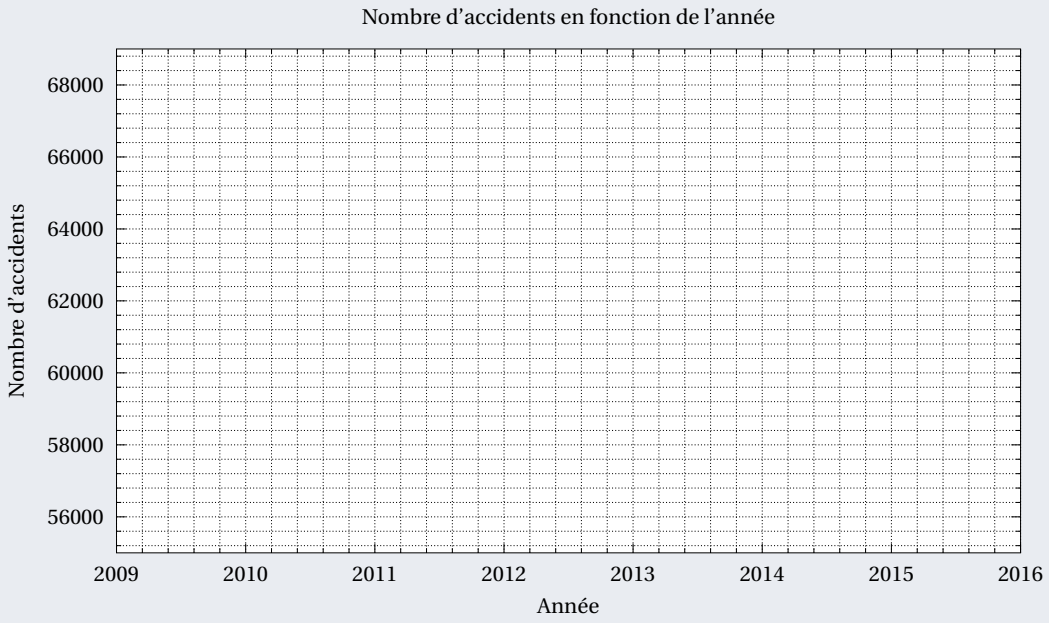
Remarque

Nous allons donc obtenir plein de points qui ne seront pas reliés par une droite !

Exemple

Nuage de points

Il faut ajouter au graphique suivant le nuage de points de l'exemple sur les accidents de la route.



Le nuage de point est une première étape qui va nous permettre d'analyser le graphique. On pourrait essayer de tracer une courbe qui passe par un maximum de points par exemple, c'est à dire rechercher une fonction qui s'approche le plus de nos données. Cette fonction est utile si nous voulons **extrapoler**, donc prédire le futur dans le cas d'une série chronologique. Une autre utilité serait d'avoir une fonction, car c'est alors beaucoup plus simple pour faire des calculs.

Le problème, c'est que l'on connait déjà plusieurs fonctions : les droites, les paraboles et les hyperboles. Le plus simple restant les droites, nous pouvons faire ce qui s'appelle ....

Définition

L'ajustement affine, c'est la détermination des paramètres a et b de l'équation de la droite $y = ax + b$ telle que sa courbe passe **au plus près de l'ensemble des points**.

Remarque

Dessiner une telle droite peut être pratique pour voir des **tendances**. C'est à dire savoir si c'est plutôt en augmentation ou en diminution.
Dans l'exemple précédent, on voit que le nombre d'accidents à tendance à diminuer.

Il est cependant difficile de tracer une telle droite, le mieux est d'utiliser un tableur ou sa calculatrice. La seule indication que nous avons est que le **point moyen** appartient à la droite. Mais qu'est ce que le point moyen ?

Définition

Le **point moyen** est un point de coordonnées $(\bar{x}; \bar{y})$ avec $\bar{x} = \frac{x_1+x_2+x_3+\ldots}{\text{Nombre de points}}$ la moyenne des x et $\bar{y} = \frac{y_1+y_2+y_3+\ldots}{\text{Nombre de points}}$ la moyenne des y. Ce point s'appelle généralement G.

Exemple

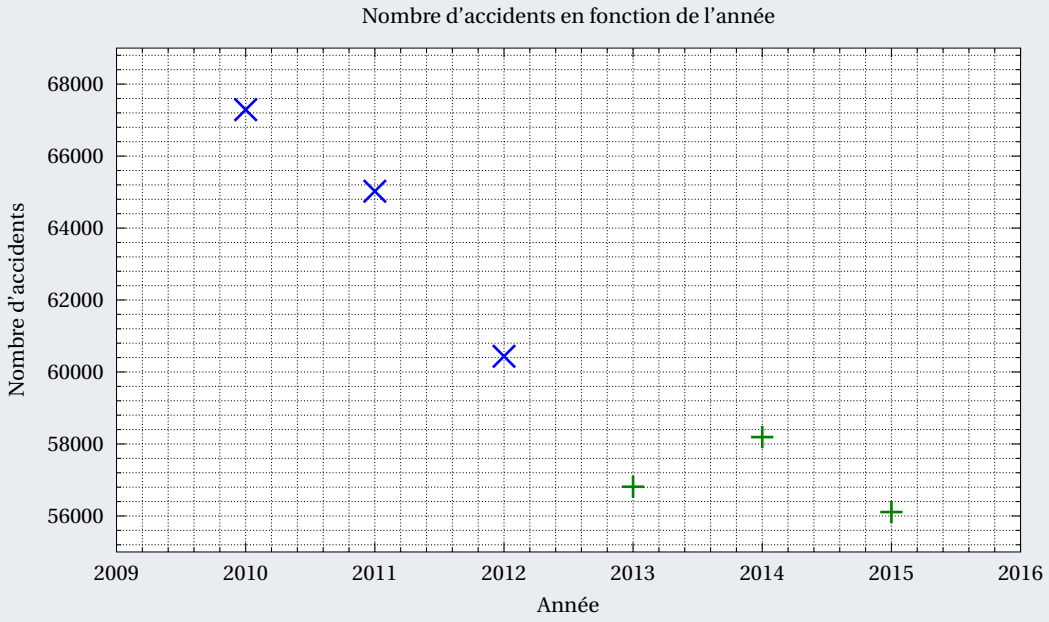
Nous commençons par découper la série S des accidents de la route en deux séries S_1 et S_2 que l'on trie par nombre d'accidents croissants, on obtient les tableaux suivants :

Année	2010	2011	.....	2014	.....	2015
Nombre d'accidents	67288	65024	.....	58191	.....	.....
Série	Série 1			Série 2		

Ensuite, nous pouvons calculer les deux points moyens :

$$\begin{cases} G_1 = \left(\dfrac{+\quad +}{3}; \dfrac{+\quad +}{3}\right) \\ \quad = (\quad ; \quad) \\ G_2 = \left(\dfrac{+\quad +}{3}; \dfrac{+\quad +}{3}\right) \\ \quad = (\quad ; \quad) \end{cases}$$

Placer ces deux points sur le nuage de points suivant et tracer la droite G_1G_2, cette droite s'appelle la **droite de Mayer** :



3 Un exemple complet

Un responsable de ventes de magasin analyse l'évolution de son chiffre d'affaires sur la dernière période. Il relève pour cela le montant des frais de publicité engagés sur la même période. Il dresse le tableau suivant :

Frais de publicité x_i (million d'euros)	10	6	6.5	11.5	11	8	7	6.5	11	9
Chiffre d'affaire y_i (million d'euros)	250	220	228	262	268	244	240	222	259	246

1 Est-ce une série statistique à deux variables ? Si oui, quelles sont ces deux variables ? Quelle variable est x ? Et laquelle est y ? Est-ce chronologique ?

.....

.....

.....

.....

.....

.....

2 Quelle est la moyenne des frais de publicité ? Et la moyenne du chiffre d'affaire ?

.....

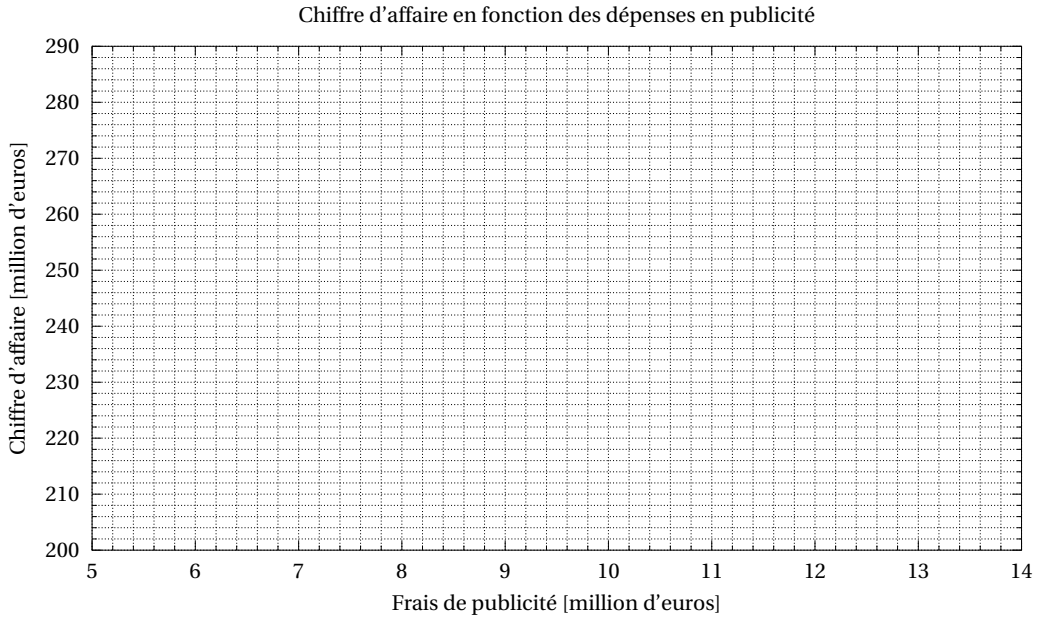
.....

.....

.....

.....

3 Représentez le nuage de point associé aux 10 couples du tableau.



4 Remplissez le tableau trié suivant (selon les frais de publicité croissant).

Frais de publicité x_i (million d'euros)	6	6.5			8
Chiffre d'affaire y_i (million d'euros)	220				
Frais de publicité x_i (million d'euros)	9		11		11.5
Chiffre d'affaire y_i (million d'euros)					262

5 Quelles sont les coordonnées du point G_1, point moyen des 5 premiers couples du tableau ci-dessus ? Et les coordonnées de G_2 point moyen des 5 derniers couples du tableau ci-dessus ? Dessinez les deux points sur le nuage de point.

.....

.....

.....

.....

.....

6 Tracez la droite $\mathscr{D}$ passant par G_1 et G_2.

7 Calculez les paramètres a et b dans l'équation $y = ax + b$ de $\mathscr{D}$ en se servant des formules $a = \frac{y_{G_2} - y_{G_1}}{x_{G_2} - x_{G_1}}$ et $b = y_{G_2} - a \times x_{G_2}$.

.....

.....
.....
.....
.....

8 Imaginons que nous dépensons 13 millions en publicité, quel est le chiffre d'affaire attendu ? (utilisez l'équation précédente et la lecture graphique)

.....
.....
.....
.....
.....

9 Utilisez un tableur et notez l'équation de la droite $\mathscr{D}_2$. Faites de même avec votre calculatrice.

.....
.....
.....
.....

4 Sources

Pour écrire ce court document, j'ai utilisé le livre « Maths - Collection Perspectives » pour Terminale Professionnelle groupement C et le cours « séries statistiques à deux variables » de T. Tchangaï.