

L'étude d'une distribution statistique à deux variables

Introduction

Notions de base

La réalisation d'une étude sur la corrélation existant entre deux variables

Établir les bases de l'étude

Prise des données

Construction d'un tableau de distribution statistique

Construction d'un graphique de type « nuage de points »

Les types de corrélation

Calcul du coefficient de corrélation

Interprétation de la corrélation

Production d'un modèle de prédiction mathématique

Introduction

Qui n'a pas un jour essayé d'établir des rapprochements entre deux éléments d'une même situation? Existe-t-il un lien entre la mesure du tour de cuisse d'un haltérophile et la charge qu'il peut soulever? Est-ce qu'il y a vraiment un lien entre l'âge des conducteurs et les accidents impliquant la vitesse? Tout le monde voit des liens entre différentes variables et chacun a ses hypothèses personnelles sur plusieurs sujets.

Dans ce module, tu apprendras comment les scientifiques et les mathématiciens font pour dépasser le stade des hypothèses et affirmer, hors de tous doutes, que deux variables sont étroitement liées ensemble.

Pour établir s'il existe des liens entre deux variables, tu devras passer aux travers de ce document afin d'apprendre la manière la plus simple qui existe pour déterminer efficacement s'il existe une relation entre deux variables.

Notions de base

Dans toutes situations, il existe ce qu'on appelle des variables, c'est-à-dire des éléments de la situation qui peuvent changer. Par exemple, les scientifiques ont clairement établi la relation, au golf, entre la vitesse d'impact du bois numéro 1 avec la balle et la distance parcourue par cette dernière. Pour établir cette relation, plusieurs éléments sont restés constants afin de ne pas influencer la relation. Le type de balles utilisé et le bâton utilisé sont restés les mêmes pour

tous les essais. Le « swing » du bâton était aussi le même, car il était reproduit par une machine. Cette situation comportait aussi deux variables. La vitesse du « swing du bâton » était modifiée par l'expérimentateur et il en résultait un changement sur la distance parcourue par la balle de golf.

En résumé, lorsqu'on fait une étude statistique à deux variables :

Nous devons identifier les deux variables que nous voulons étudier. Une **variable** sera dite **dépendante** et l'autre sera dite **indépendante**. Dans certaines situations, il n'est pas évident de déterminer quelle variable influence l'autre variable, dans ces cas particuliers ce sera à l'expérimentateur de déterminer quelle variable sera indépendante et quelle variable sera dépendante. Dans le cas de notre exemple du golf, il est évident que la distance parcourue par la balle est dépendante de la vitesse avec laquelle elle sera frappée par le bâton.

Afin d'être certain qu'aucune autre variable n'interfère dans notre étude, il faut dresser la liste des autres éléments qui pourraient influencer notre relation et s'assurer que ces éléments restent constants. Dans l'exemple du golf, on s'est assuré de toujours utiliser le même bâton, le même type de balles, le même swing, l'expérience a été réalisée dans les mêmes conditions de vent et à la même température.

Lorsqu'on fait l'étude d'une relation statistique, il est important de posséder plusieurs couples de variables. Plus votre étude comportera de données (couples de variables) plus vos résultats seront exacts. Comme les scientifiques cherchent à créer des modèles exacts, ils analysent souvent plusieurs milliers de données. Il faut dire que l'arrivée de l'informatique a beaucoup facilité l'analyse d'un nombre important de données. La relation établie au golf repose sur des milliers de balles frappées. À plus petite échelle, nos analyses en classe reposeront sur autant de données que notre montage expérimental et le temps dont nous disposerons nous permettront de recueillir. Cette année, il sera peu fréquent pour nous de travailler avec plus d'une quinzaine de données. L'ensemble des données que vous utiliserez porte un nom, c'est votre **distribution statistique à deux variables**. Habituellement, on représente une distribution statistique sous forme de tableau ou sous forme de **graphique de type nuage de points**.

Lorsque les variables que l'on étudie sont chiffrées, pour l'exemple du golf la vitesse et la distance sont des mesures représentées par des valeurs numériques, on dit que nos **variables** sont **quantitatives**. Le lien qui peut exister entre des variables quantitatives se nomme **corrélation**. Une corrélation peut être nulle, donc aucun lien, jusqu'à parfaite, donc un lien présent pour 100% des cas étudiés. Les statisticiens ont développé différentes méthodes pour obtenir une valeur numérique nous renseignant sur la force de notre corrélation. Cette valeur numérique se nomme **coefficient de corrélation**.

La réalisation d'une étude sur la corrélation existant entre deux variables

Voici les étapes d'une étude statistique sur deux variables :

1- Établir les bases de l'étude

C'est à cette étape que le chercheur identifie son sujet d'étude d'après un questionnement, des observations ou comme suite à une demande. En premier lieu, le chercheur identifie les deux variables qu'il veut mettre en relation. Par la suite, il liste les autres variables qui, d'après lui et selon les connaissances déjà établies dans ce domaine, pourraient influencer sa prise de données. Il trouve ensuite des solutions pour garder constantes ces variables qui pourraient interférer. En réalité, cette étape est très ardue. Il n'est pas rare qu'un chercheur doive recommencer son expérience parce qu'il découvre en cours d'expérience que ses résultats sont influencés par une variable à laquelle il n'avait pas pensé. Certaines recherches n'aboutiront jamais, car nos connaissances ne sont pas encore assez développées pour être certain que toutes les variables autres que celles que l'on veut étudier sont maintenues constantes. D'autres changeront d'orientations parce que la découverte d'une autre variable s'avère plus intéressante à étudier qu'une des deux variables de départ. La science est une suite d'essais et d'erreurs ponctuée par des embûches et d'agréables surprises dont la récompense ultime est la découverte.

2- Prise des données

C'est à cette étape que le chercheur construit une expérience qui lui permettra de contrôler certaines variables et de prendre des mesures pour les deux variables qu'il a choisi d'étudier. La minutie avec laquelle il va prendre les données influence beaucoup la qualité de son étude. Plus ses mesures seront précises plus les résultats de sa recherche seront près de la réalité.

3- Construction d'un tableau de distribution statistique

Le tableau est devenu rapidement le moyen privilégié par les scientifiques pour prendre en notes leurs données. Il est important de bien connaître toutes les règles de construction d'un tableau afin qu'il corresponde aux standards établis par la communauté scientifique :

Chaque tableau doit être accompagné d'un titre complet. Le titre doit décrire la relation entre les deux variables étudiées ainsi que les conditions expérimentales importantes qui doivent être respectées pour obtenir les mêmes résultats. Si une étude comporte plusieurs tableaux,

chacun des tableaux doit être numéroté, cela facilite les renvois lors de l'écriture d'un rapport de laboratoire. La rédaction d'un bon titre est souvent l'étape la plus difficile à maîtriser dans l'apprentissage des règles de construction d'un tableau. **Un bon titre devrait répondre aux questions : qui, quoi, quand, où et comment. Toute information susceptible d'influencer les résultats de votre étude devrait figurer dans votre titre. Le titre ne devrait pas être beaucoup plus large que le tableau, il est préférable de l'écrire sur plusieurs lignes.**

Les données sont rarement représentées à l'horizontale, cette représentation est utilisée seulement dans les cas où le tableau ne comporte que très peu de variables. **Les données sont centrées dans leur colonne et occupent environ le tiers de la largeur de la colonne. Le tableau devrait être conçu de façon à être imprimé sur une seule page. Il est parfois utile d'utiliser le mode paysage ou, dans les cas où l'étude compte une grande quantité de données, de séparer nos données en plusieurs tableaux qui pourront alors prendre plusieurs pages. De plus, les données d'une même colonne doivent compter le même nombre de décimales, car ces décimales nous informent sur la précision de l'appareil ayant servi à prendre ;les mesures.**

Chaque colonne dans laquelle sont inscrites les données doit indiquer les unités de mesure de ces données **entre parenthèses.**

Quand les données ne proviennent pas d'une expérience, mais plutôt de la littérature, par exemple un atlas ou un organisme tel que Statistique Canada, vous devez alors écrire sous le graphique et en petits caractères la source de vos données.

Exemple de tableau :

Tableau 1

Relation entre la vitesse de la tête d'un bois no 1 et la distance parcourue par une balle de golf frappée par ce bâton. (vent de face de 10 km/h, température: 23°C)

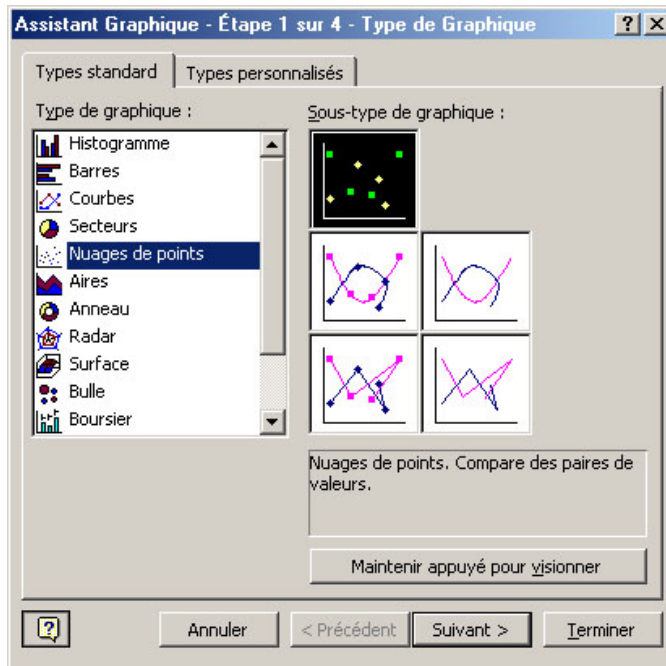
Vitesse de la tête du bois 1 (milles / h)	Distance parcourue par la balle (verges)
100	240
110	270
94	220
118	310
90	200
98	223
112	288

4- Construction d'un graphique de type « nuage de points »

Le graphique nuage de points est un graphique cartésien qui représente point par point chacun des couples de votre distribution statistique. C'est le moyen par excellence développé par les mathématiciens afin de lire et d'interpréter rapidement une distribution statistique. Dans ce type de graphique, les points ne sont pas reliés entre eux. Lorsque les points tendent à former une ligne nous dirons que la **corrélation** est **de type linéaire**, plus les points s'approcheront de cette ligne imaginaire plus la corrélation entre les deux variables sera fortes.

Je vous conseille de construire ces types de graphique dans Excel, même si vous avez une bonne connaissance de TI-intercative. Excel est très puissant pour gérer de grandes distributions statistiques et il est le logiciel le plus répandu dans les laboratoires informatiques des cégeps et des universités. De plus, il vous sera très utile pour réaliser des automatisations et rendre vos graphiques dépendants de vos tableaux, ce qui vous permettra, au cours de l'année, de sauver bien du temps lors de certaines expériences ou missions. Il est impératif que vous maîtrisiez parfaitement Excel pour tracer ces types de graphiques, car vous aurez à construire de tels graphiques presque chaque semaine, et ce, tout au cours de l'année.

Excel vous offre plusieurs types de nuages de points, voici celui qui est utilisé en sciences :



La construction d'un graphique en sciences répond elle aussi à plusieurs critères établis par la communauté scientifique, j'ai numéroté les critères de façon à pouvoir rapidement vous indiquer, lors de la correction de vos graphiques, quels sont ceux qui ne sont pas maîtrisés :

Grandeur :

C1 : Chaque graphique doit occuper une page entière.

C2 : Les points doivent occuper au moins les deux tiers de la zone du graphique, à moins que cela nuise à l'observation de la tendance générale des points. **On peut tolérer que le graphique prenne un peu moins des 2 tiers de l'espace si cela permet de garder dans la zone visible l'ordonnée à l'origine du plan cartésien.**

Titre : **C3** : Il doit être disposé en haut du graphique et au centre de ce dernier.

C4 : Il respecte les mêmes conditions que le titre d'un tableau **et par le fait même, le titre du tableau peut être utilisé intégralement pour le graphique.**

Axes :

C5 : On doit utiliser seulement le premier quadrant, à moins que les valeurs négatives aient un sens réel.

C6 : Chaque axe doit porter un titre, une lettre représentant votre variable doit aussi y figurer en italique. Certaines variables sont déjà fixées par des conventions internationales, par exemple *t* pour le temps, *T* pour la température, *v* pour la vitesse, etc. À la suite de

cette identification, on inscrit, entre parenthèses, le symbole des unités de mesure de la variable.

Ex. vitesse v d'un mobile (m/s)

Échelles :

C7 : Le graphique doit être facile à lire. On évite la surabondance de chiffres et de lignes et on s'arrange plutôt pour que la graduation soit facilement devinable. **Une telle graduation inclut des lignes qui ne sont pas accompagnées d'une valeur numérique. Par contre, cette dernière doit pouvoir être devinée instantanément.** Ce critère est celui qui cause toujours le plus de problèmes lors de la construction d'un graphique, je te conseille de me montrer souvent tes graphiques afin que je t'aide à comprendre leurs faiblesses. Profites en pendant que tu es en formatif.

Quadrillage :

C8 : On évite le quadrillage dans un graphique nuage de points.

Autres lignes :

C9 : S'il est nécessaire de tracer dans le graphique d'autres lignes que la courbe de tendance (ex. pour trouver l'ordonnée à l'origine ou pour déterminer une aire sous la courbe.), elles doivent être tracées en pointillées pour montrer qu'elles ne sont pas trouvées grâce à des valeurs expérimentales. Tu comprendras plus loin ce qu'est une courbe de tendance.

Légende :

C10 : La légende ne doit pas apparaître dans le graphique à moins qu'il y ait plus d'une relation illustrée dans le même graphique.

Exemple : si dans notre situation du golf on avait voulu illustrer plus d'une sorte de bâtons dans le même graphique.

Équation :

Lorsqu'on applique un modèle mathématique à notre graphique, la règle du modèle doit être intégrée dans la zone du graphique, utiliser les variables définies dans les titres d'axe et être disposée de façon à ne pas cacher de points.

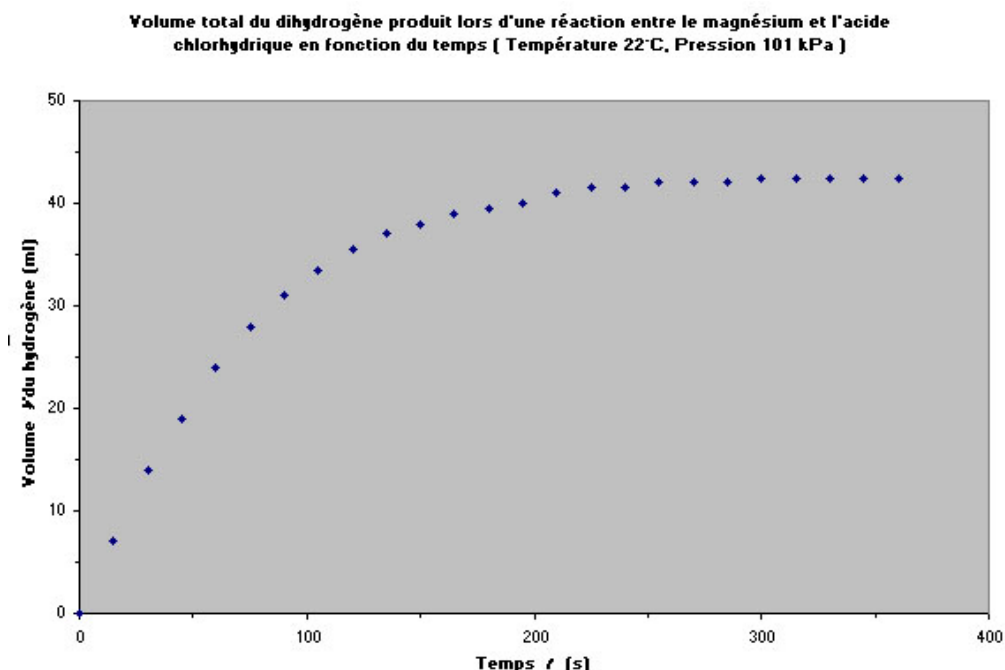
Le coefficient de détermination :

Lorsqu'on applique un coefficient de détermination à notre graphique, il doit être intégré dans la zone du graphique, et être disposé de façon à ne pas cacher de points.

Fond du graphique :

Lorsque l'on veut imprimer le graphique, afin d'économiser l'encre, il est important de le mettre en blanc.

Voici un exemple de graphique nuage de points bien conçu, mis à part le fond qui n'a pas été mis en blanc pour l'impression :



5- Les types de corrélation

Il existe deux grands types de corrélation :

La **corrélation positive** : Dans ce type de corrélation les variables varient dans le même sens, lorsque la valeur de « x » d'un couple augmente le « y » qui lui est associé augmente aussi.

La **corrélation négative** : Dans ce type de corrélation, les variables varient en sens inverse, lorsque la valeur d'une variable diminue l'autre augmente et vice et versa.

Chacun de ces types de corrélation comporte plusieurs **intensités** : nulle, faible, moyenne, forte, très forte et parfaite. Dans une corrélation nulle, les points semblent disposés au hasard dans le graphique et il est difficile d'établir le sens de la corrélation, c'est-à-dire de déterminer si elle est positive ou négative. Dans une corrélation parfaite, les points formeront une ligne ou une courbe très bien définie.

Tu trouveras une très bonne représentation des deux sens ainsi que des diverses intensités de la corrélation à la page 60 de ton manuel réflexion tome 2.

6- Calcul du coefficient de corrélation linéaire (r)

Comme il est très difficile à l'œil de différencier l'intensité d'une corrélation linéaire, les mathématiciens ont développé un calcul à partir de la distribution statistique pour nous renseigner sur le niveau de corrélation. Le résultat de ce calcul se nomme coefficient de corrélation. Ce coefficient peut prendre une valeur entre -1 et 1 . Le coefficient est composé d'un signe (+ ou -) et d'un chiffre de 0 à 1. Le signe nous indique le sens de la corrélation « positive » ou « négative ». Le nombre quant à lui nous renseignera sur l'intensité de la corrélation.

Valeur de r	Intensité du lien
Près de 0	Indique un lien nul entre les deux variables étudiées
Près de $-0,5$ ou $0,5$	Indique un lien faible entre les deux variables étudiées
Près de $-0,75$ ou $0,75$	Indique un lien moyen entre les deux variables étudiées
Près de $-0,87$ ou $0,87$	Indique un lien fort entre les deux variables étudiées
Près de -1 ou 1	Indique un lien très fort entre les deux variables étudiées
Égale à 1	Indique un lien parfait entre les deux variables étudiées

Le calcul du coefficient de corrélation à partir de la formule s'avère compliqué, nous demanderons donc à Excel de réaliser le calcul sans que nous ayons à en comprendre les mécanismes.

CALCUL DU COEFFICIENT DE CORRÉLATION PAR EXCEL

- 1) Sélectionner la cellule dans laquelle vous voulez voir apparaître le coefficient de corrélation.
- 2) Dans le menu sélectionner « Insertion » puis « fonction »
- 3) Dans « catégories de fonction » sélectionnez tous, vous pourrez sélectionner dans la fenêtre de droite « coefficient de corrélation »
- 4) Dans « Matrice 1 » sélectionnez les « x » de votre distribution statistique.
- 5) Dans « Matrice 2 » sélectionnez les « y » de votre distribution statistique.
- 6) Cliquez sur « OK »

7- Interprétation de la corrélation

L'interprétation de la corrélation se base sur deux éléments. Le premier élément est objectif, il s'agit bien sûr de notre coefficient de corrélation qui nous renseigne sur la force du lien existant entre les deux variables et qui nous renseigne aussi sur le sens de la corrélation. Le second élément est lui beaucoup plus subjectif, il fait appel au bon sens de la personne qui mène la recherche et nécessite qu'on tienne compte des implications de la découverte.

Exemple :

On découvre une corrélation entre un médicament et la guérison du cancer dont le coefficient est de 0,45. En se basant strictement sur le coefficient, c'est une corrélation faible donc le lien est quasi inexistant et on ne devrait pas produire ce médicament, car il ne sera pas efficace. Mais pour une personne dont la vie est en jeu, ce faible lien est encourageant et souhaite avoir accès à ce médicament en espérant qu'il pourrait être un des rares cas qui arriverait à guérir.

À l'inverse, si on découvrait qu'un certain médicament servant à soulager les allergies graves pourrait provoquer le cancer selon une corrélation de 0,3. En se basant simplement sur le coefficient de corrélation, on dirait que la corrélation est très faible et qu'il n'existe pas vraiment de lien entre ce médicament et le cancer, mais comme cette corrélation implique des vies humaines elle prend une plus grande importance et probablement que le médicament ne sera pas lancé sur le marché avant une étude sérieuse des conséquences possibles, même si le lien statistique est presque inexistant.

Donc, le coefficient est très objectif et renseigne sur l'intensité réelle du lien entre les deux variables, mais parfois il est important de regarder plus loin que les nombres pour bien analyser la valeur et les implications d'une découverte.

Production d'un modèle de prédiction mathématique

Lorsque la corrélation entre deux variables est très forte ou parfaite, il devient possible de bâtir des outils mathématiques pour expliquer et prédire un phénomène. Depuis le troisième secondaire, tu as étudié certaines relations mathématiques par exemple la droite et la parabole. Il arrive que lors d'une étude entre deux variables que nous remarquons que le lien entre les deux variables se comporte selon une relation mathématique définie. Dans un tel cas, une équation mathématique représente notre corrélation et il est possible de s'en servir pour faire des prédictions.

Excel permet de faire ressortir la tendance des points dans un nuage de points sous la forme d'une droite ou d'une courbe. La représentation graphique de la tendance des points se nomme **courbe de régression**. Voici la démarche à suivre pour faire apparaître dans un graphique la tendance de vos points.

Démarche pour tracer la courbe de régression sur Excel :

- 1) Après avoir tracé le graphique de nuage de points, sélectionner ce dernier.
- 2) Dans le menu, sélectionner « graphique » et « ajouter une courbe de tendance »
- 3) Choisir le type de régression que semblent suivre vos points, si vous hésitez entre deux types, c'est le coefficient de détermination (R^2) calculé par l'ordinateur qui vous permettra de choisir la meilleure régression. Vous devez garder la courbe présentant le meilleur coefficient de détermination. Le coefficient de détermination sera abordé plus loin dans le texte.

Lorsque vous désirez une parabole il vous faut sélectionner « polynomiales de degré 2 ». Plus votre coefficient de détermination sera élevé, plus votre courbe de régression représentera bien la corrélation qui existe entre vos deux variables.

Une fois la courbe de tendance ajoutée, vos points n'ont plus aucune importance, les résultats de votre expérience sont maintenant représentés par la courbe.

Plusieurs relations existant autour de nous ont une courbe de tendance représentée par une droite nommée « **droite de régression** », ne t'étonne donc pas de la voir apparaître souvent dans plusieurs études que tu auras à réaliser cette année.

Excel est capable de réaliser pour vous les calculs permettant d'obtenir l'équation mathématique représentant votre courbe de tendance. Au même endroit où vous cochez une case pour obtenir votre coefficient de détermination vous pouvez aussi cocher une autre case, « afficher l'équation sur le graphique » pour obtenir l'équation de votre courbe de tendance.

LE COEFFICIENT DE DÉTERMINATION (R^2)

Il peut être intéressant de savoir si la droite de régression calculée par Excel s'ajuste bien à nos points. Le coefficient de détermination est défini comme un indice de qualité de l'ajustement de la droite aux points expérimentaux. Le coefficient varie entre 0 (aucun ajustement linéaire) et 1 (ajustement linéaire parfait). Lorsque tous les points se situent exactement sur la droite on obtient un coefficient de détermination de 1.

Démarche de calcul du coefficient de détermination :

- 1) sélectionner « graphique » dans le menu
- 2) puis « ajouter une courbe de tendance »
- 3) dans l'onglet options, cocher « afficher le coefficient de détermination (R^2) ».

L'équation mathématique de votre courbe de régression

Lorsque par le coefficient de détermination vous avez trouvé quelle courbe de tendance représentait mieux la tendance de vos points, vous pouvez faire afficher par Excel l'équation mathématique de cette courbe. Il s'agit de cocher l'affichage de l'équation au même endroit où vous avez demandé à Excel d'afficher le coefficient de détermination. Excel affiche toujours l'équation en utilisant les variables « x » et « y », vous devrez donc remplacer ces deux variables par celles que vous avez fixées dans votre graphique.

Notez bien, dans Excel, si vous modifiez des points dans votre tableau, le graphique, les taux de corrélation et de détermination ainsi que l'équation s'ajusteront automatiquement. À partir, du moment où vous remplacez les variables « x » et « y » de votre équation par vos variables expérimentales, l'équation perdra sa capacité de s'ajuster automatiquement à tout changement, cette étape doit donc être faite en tout dernier lieu, quand tous les points du tableau sont définitifs.