Première partie

Statistique descriptive

1 statistiques à une variable

1.1 vocabulaire, représentation

La statistique est l'étude des **populations**, dont les éléments sont des **individus** ; le plus souvent on n'étudie pas toute la population, mais seulement un **échantillon** de celle-ci. L'**effectif** d'un échantillon est le nombre d'individus qui le composent.

Plus précisément, on étudie certains **caractères** des individus, caractères qui peuvent être **qualitatifs** (par exemple le prénom, la nationalité, ...) ou **quantitatifs** (l'âge, la taille, les revenus mensuels...). Les caractères quantitatifs peuvent être **discrets** (la pointure de chaussures, le nombre de personnes au foyer, ...) ou **continus** (la taille, la superficie d'une région, ...).

Pour faciliter l'étude, en particulier des caractères continus, on peut regrouper les valeurs en **classes**, c'est à dire en intervalles deux à deux disjoints. La longueur d'un tel intervalle est appelé **amplitude** de la classe.

Par exemple, pour décrire la taille d'un adulte, on pourra considérer les intervalles $[0;100[$, $[100,110[$, ..., $[190,200[$, $[200,+\infty[$, la première classe est d'amplitude 100, la dernière d'amplitude infinie alors que toutes les autres sont d'amplitude 10.

Une **série statistique** est un ensemble de couples (x_i, n_i), où les x_i sont les valeurs prises du caractère et les n_i le nombre de fois où la valeur x_i apparaît.

L'effectif total de l'échantillon est donc $n = \sum_i n_i$.

On appelle **fréquence** d'appparition de x_i le nombre $f_i = n_i/n$.

exemple 1 : sur un échantillon de mille pièces tirées de la production journalière d'une usine, on compte le nombre de défauts constatés :

nombre de défauts	0	1	2	3	4
effectifs	570	215	140	60	15
fréquences	0.57	0.215	0.140	0.06	0.015

Ici les valeurs sont donc $x_0 = 0$, $x_1 = 1$, ..., $x_4 = 4$ d'effectifs respectifs $n_0 = 570$, $n_1 = 215$, $n_2 = 140$, $n_3 = 60$, $n_4 = 15$.

L'effectif total est de 1000, ce qui permet de calculer facilement les fréquences situées sur la dernière ligne du tableau.

On peut imaginer de multiples représentations graphiques pour une série statistique : diagramme en batons, camemberts...Une seule présente une petite difficulté : l'histogramme, utilisé pour représenter par une suite de rectangle des résultats regroupés en classes.

exemple 2 : un technicien mesurant des tiges métalliques obtient les valeurs suivantes :

longueur (mm)	$[330;340[$	$[340;343[$	$[343;345[$	$[345;350[$	$[350;360[$
effectifs	57	195	204	30	14
fréquences	$\frac{57}{500} \simeq 0.11$	$\frac{195}{500} \simeq 0.39$	$\frac{204}{500} \simeq 0.41$	$\frac{30}{500} \simeq 0.06$	$\frac{14}{500} \simeq 0.03$

Pour tracer l'histogramme on place en abscisse les différentes classes, ici $[330;340[$, $[340;343[$, $[343;345[$, $[345;350[$ et $[350;360[$.

Pour chaque classe on calcule alors la hauteur du rectangle correspondant : c'est l'effectif divisé par l'amplitude de la classe. Ici, on trouve donc respectivement 5.7, 65, 102, 6 et 14.

Alors l'aire de chaque rectangle est proportionnelle à l'effectif de chaque classe. Attention, c'est bien **l'aire**, et non la hauteur, qui est proportionnelle à l'effectif !

1.2 caractéristiques de position

le mode est la valeur la plus fréquente d'une série statistique ; pour une série répartie en classe on parle de **classe modale**. Le mode n'est pas forcément unique.

Dans l'exemple 1, le mode est 0 ; dans l'exemple 2, la classe modale est l'intervalle $[343;345[$.

la médiane est la valeur Me telle que la moitié des individus de la série ont un caractère inférieur ou égal à Me et l'autre moitié un caractère supérieur ou égal.

Quand les données sont regroupées en classes on parle de classe modale.

Dans l'exemple 1, la médiane est 0 ; dans l'exemple 2, la classe médiane est $[340; 343[$.

la moyenne d'une série statistique (x_i, n_i) est le nombre

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{n} = f_1x_1 + f_2x_2 + \ldots + f_px_p$$

Dans l'exemple 1, la moyenne est donc $0.215 \times 1 + 0.14 \times 2 + 0.06 \times 3 + 0.015 \times 4 = 0.735$.

Dans le cas où les valeurs sont regroupées en classes, on peut déterminer une valeur approchée de la moyenne en remplaçant pour le calcul chaque classe par son milieu : dans l'exemple 2, la moyenne est $0.114 \times 335 + 0.39 \times 341.5 + 0.408 \times 344 + 0.06 \times 347.5 + 0.028 \times 355 = 342.517$. Ce calcul n'est satisfaisant que si les classes sont « bien choisies », i.e d'amplitudes et d'effectifs comparables.

1.3 caractéristiques de dispersion

On souhaite estimer si les valeurs d'une série statistique sont regroupées ou non autour de la valeur moyenne.

La caractéristique de dispersion la plus élémentaire et la plus facile à calculer est **l'étendue**, différence entre la plus grande et la plus petite des valeurs. On peut aussi considérer la moyenne des écarts à la moyenne de chaque valeur.

Mais on préfère utiliser **la variance** et **l'écart-type** qui pour chaque valeur :

la variance de la série statistique est

$$\frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \ldots + n_p(x_p - \bar{x})^2}{n} =$$

$$f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + \ldots + f_p(x_p - \bar{x})^2.$$

L'écart-type de la série statistique, noté σ',
est la racine carrée de sa variance.

Ainsi, la variance est simplement « la moyenne des carrés des écarts à la moyenne ».

Dans l'exemple 1, on trouve une étendue de 4, une variance égale à $0.57 \times (0.735)^2 + 0.215 \times (1 - 0.735)^2 + 0.14 \times (2 - 0.735)^2 + 0.06 \times (3 - 0.735)^2 + 0.015 \times (4 - 0.735)^2 \simeq 1.015$ et donc $\sigma' \simeq 1.01$.

remarque 1 : on peut développer l'expression donnant la variance et obtenir après calcul la formule $\sigma'^2 = \overline{x^2} - \bar{x}^2$.

remarque 2 : la formule donnant la variance fait bien intervenir un n, et pas un $n - 1$, cf 10.2 pour plus de précisions.

2 statistiques à deux variables

Une **série statistique double** est une série de n mesures de deux quantités x et y : $x_1, x_2, \ldots, x_n$ et $y_1, y_2, \ldots, y_n$.

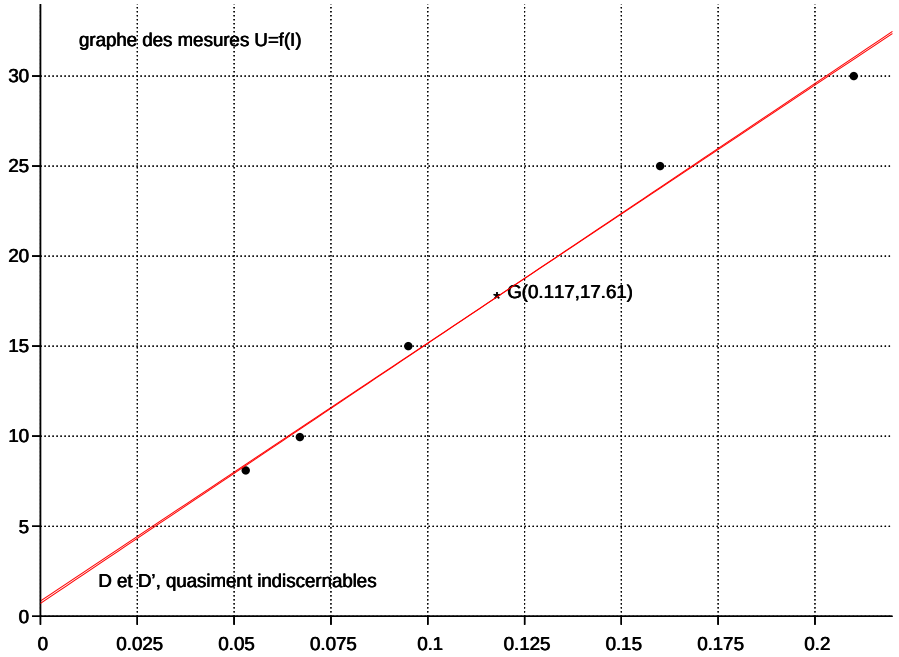
On s'intéresse surtout à la question de savoir si les mesures des deux valeurs sont indépendantes ou non. Si elles ne sont pas indépendantes, comment sont-elles reliées ?

2.1 Droite de régression linéaire

exemple 1 : On mesure simultanément le courant et l'intensité aux bornes d'une résistance. On obtient les valeurs :

intensité en ampères	0.053	0.067	0.095	0.16	0.21
tension en volts	8.1	9.95	15	25	30

On peut représenter ces mesures par un **nuage de points** $M_i(x_i, y_i)$; $G(\bar{x}, \bar{y})$ est appelé **point moyen**. Ici, on trouve pour point moyen $G(\bar{x} = 0.117, \bar{y} = 17.61)$.



Sur cet exemple, que constate-t-on ? Les mesures semblent indiquer qu'il y a une relation linéaire entre les valeurs x du courant et y de la tension, i.e que l'on peut écrire de manière « presque » exacte $y = ax + b$. Mais comment choisir les « meilleurs » a et b ?

On peut bien sûr tracer à la main une droite qui passe au plus près des points, puis déterminer par lecture graphique son coefficient directeur a et son ordonnée à l'origine b. Cette méthode est tout à fait valable, en particulier pour des valeurs obtenues en TP et tracées à la main !

Mais nous allons voir une méthode calculatoire plus systématique (et plus utilisable lors d'un traitement informatique des données) : la **méthode des moindres carrés**.

On commence par determiner les caractéristiques de chacune des séries : les moyennes $\bar{x} = 0.117$ et $\bar{y} = 17.61$ déjà calculées, et les écarts-types $\sigma_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = 0.0593$ et $\sigma_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}} = 8.53$.

On définit alors la **covariance** de la série des (x_i, y_i) par la formule

$$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

On peut développer l'expression définissant la covariance $\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^n (x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y}) = \frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x} \times \frac{1}{n}\sum_{i=1}^n y_i - \bar{y} \times \frac{1}{n}\sum_{i=1}^n x_i + \bar{x}\bar{y} = \frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y}$, et obtenir ainsi une autre expression de la covariance comme moyenne des produits moins produit des moyennes :

$$\sigma_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \quad - \quad \bar{x}\bar{y}.$$

En utilisant l'une ou l'autre de ces formules, on trouve ici $\sigma_{x,y} = 0.503$.

On appelle alors **droite de régression de y en x** la droite $D : y = ax + b$ passant par G et de coefficient directeur

$$a = \frac{\sigma_{x,y}}{\sigma_x^2}.$$

C'est la droite D pour laquelle la somme des $M_i P_i^2$ est minimale, les P_i étant les points de D d'abscisse x_i.

Ici, $D : y = 143.27x - 0.847$.

De même la **droite de régression de x en y** $D' : x = a'y + b'$ passant par G et de coefficient $a' = \frac{\sigma_{x,y}}{\sigma_y^2}$ minimise la somme $\sum_i M_i Q_i^2$ où les Q_i sont les points de D' d'ordonnée y_i.

Ici $D' : x = 6.927 \times 10^{-3}\, y - 0.00498$, soit $y = 144.36x + 0.719$.

On constate sur cet exemple que les deux droites sont quasiment indiscernables, et la loi théorique $U = RI$ (soit ici $y = Rx$) semble a peu près vérifiée, avec une valeur de R proche de 143 ou 144 Ω.

exemple 2 : dans un SAV, on note pour chaque appareil défectueux l'heure d'arrivée et le temps d'atelier nécessaire à la réparation. Dans ce cas, il est probable que le graphique ressemble à un nuage de points d'apparence aléatoire, car les deux caractéristiques n'ont probablement aucun lien entre elles. Les droites D et D' ne coïncideront pas du tout.

Un outil numérique permet d'estimer si deux variables sont liées ou pas par une relation linéaire :

le **coefficient de corrélation** $r = \frac{\sigma_{x,y}}{\sigma_x \sigma_y}$.

r est toujours compris entre -1 et 1. S'il vaut ± 1, les droites D et D' sont confondues, et plus il est proche de ± 1, plus les points (x_i, y_i) semblent alignés : on dit qu'il y a une bonne corrélation linéaire entre les quantités x et y.

Ici, $r = 0.994$. En pratique, on commence à considérer une valeur $|r| > 0.7$ comme significative d'une corrélation linéaire.

remarque 1 : le fait que r soit proche de 0 n'indique pas qu'il n'y a aucune corrélation entre les variables, mais seulement qu'il n'y a pas de corrélation **linéaire**. En effet, on rencontre souvent des relations du type $y = 1/(ax+b)$, $y = kc^x$, $y = kx^c$, ... et tout ce qui précède est inadapté pour traiter ces corrélations non linéaires.

Mais il suffit d'un changement de variable pour se ramener au cas linéaire : voir les exercices.

remarque 2 : une bonne corrélation ($|r|$ proche de 1) ne signifie pas qu'il existe une relation de cause à effet entre deux phénomènes ; une étude physique plus approfondie sera nécessaire pour le savoir : r n'est qu'un indice pour le technicien.

On peut illustrer cela par un troisième exemple : si à Grenoble on note, chaque jour de l'hiver, la hauteur de neige tombée et la température de l'air, on observera une corrélation : il neige très peu les jours très froids. Peut-on en déduire que le froid empêche la neige de tomber ? En fait, il n'y a pas de relation physique directe : en altitude, ou dans d'autres régions du globe, il peut neiger avec des températures très froides. Mais en France, ce sont les anticyclones sibériens qui amènent le froid vif...et l'air sec. Il y a bien corrélation entre les phénomènes, mais pas de lien de cause à effet.

Complément : démonstration de la formule des moindres carrés

On considère un nuage de points $M_i(x_i, y_i)$ et une droite $D : y = ax + b$.

Une manière d'exprimer le fait que la droite D passe au plus près des points M_i est de demander que le produit des carrés des écarts d'ordonnée $y_i - (ax_i + b)$ soit le plus petit possible : on souhaite trouver a et b tels que la quantité $\sum_{i=1}^n ((ax_i + b) - y_i)^2$ soit minimale.

Il s'agit d'une application de deux variables (a, b) positive et à valeurs réelles. De plus cette application est dérivable : si elle admet un minimum, on sait qu'en celui-ci les dérivées partielles doivent s'annuler. On peut calculer ces dérivées partielles :

$\frac{\partial(\sum_{i=1}^n (ax_i + b - y_i)^2)}{\partial a} = 2\sum_{i=1}^n x_i(ax_i + b - y_i)$ et $\frac{\partial(\sum_{i=1}^n (ax_i + b - y_i)^2)}{\partial b} = 2\sum_{i=1}^n (ax_i + b - y_i)$.

Ainsi la condition nécessaire de minimum s'exprime par les deux équations $\sum_{i=1}^n x_i(ax_i + b - y_i) = 0$ et $\sum_{i=1}^n (ax_i + b - y_i) = 0$, soit encore $a\sum_{i=1}^n x_i^2 + b\sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i = 0$ et $a\sum_{i=1}^n x_i + nb - \sum_{i=1}^n y_i = 0$. En divisant par n les deux équations on obtient le système :

$$\begin{cases} a\frac{1}{n}\sum_{i=1}^n x_i^2 + b\bar{x} &= \frac{1}{n}\sum_{i=1}^n x_i y_i \\ a\bar{x} + b &= \bar{y} \end{cases}$$

que l'on résoud en enlevant $\bar{x}$ fois la deuxième ligne à la première :

$$\begin{cases} a(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2) &= \frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y} \\ a\bar{x} + b &= \bar{y} \end{cases}$$

Mais on reconnait dans la première équations les expressions de la variance $\overline{x^2} - \bar{x}^2$ et de la covariance déjà étudiées. Ainsi, $a = \frac{\sigma_{x,y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{\sigma_{x,y}}{\sigma_x^2}$, et la deuxième équation exprime bien le fait que la droite D passe par le point moyen du nuage de points.

Des considérations intuitives montrent (la démonstration rigoureuse dépasse le niveau de ce cours) qu'il doit bien exister une droite réalisant ce minimum, et on vient de prouver que ce n'est possible que pour les valeurs de a et b définies plus haut.

2.2 Régression linéaire passant par l'origine

Dans le cas de la résistance, si l'on connaît préalablement à l'expérience la loi d'Ohm, on sait que la relation à chercher est du type $U = RI$, et l'on peut souhaiter simplement déterminer le meilleur coefficient R correspondant aux mesures, avec une relation sans terme constant.

Si l'on recherche une relation de la forme $y = ax$, on prendra alors la valeur

$$a = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\overline{xy}}{\overline{x^2}}.$$

Dans cet exemple, $\sum_i x_i^2 = 0.086023$ et $\sum_i x_i y_i = 12.82095$, d'où $a \simeq 149.04\Omega$.

Le choix entre ces deux méthodes dépendra des circonstances : dans cette expérience, le fait qu'une régression linéaire « classique » ne donne pas $b = 0$ peut s'expliquer par le fait que la résistance n'est pas une résistance « pure » (le terme b ayant alors une signification physique) ou bien par le fait qu'une ou plusieurs mesures sont peu précises.