EDITORIAL

Constitution du G.R.E.S. au 1er mai 1999

ANGELIQUE	Françoise	LEGTA de NANCY
DESESQUELLES	René	LEGTA d'AMIENS
FAGES	Jean	ENFA TOULOUSE
GAUMET	Jean-Pascal	LEGTA LE ROBILLARD
PARNAUDEAU	Jean-Marie	LEGTA de VENOURS
PAVY	Jacques	LEGTA LE ROBILLARD
PRADIN	Jean	LEGTA de MOULINS
QUET	Guillaume	LEGTA d'AUBENAS
URDAMPILLETTA	Vincent	LEGTA de SURGERES
VARLOT	Chantal	LEGTA de CHALONS SUR MARNE

EDITORIAL

En écrivant l'éditorial de ce bulletin n°8, j'ai sous les yeux le bulletin n°1 qui est daté d'octobre 1995. Le groupe a donc édité, en moyenne, 2 bulletins par an.
C'est peu diront les collègues qui, isolés dans leur établissement, ont à faire face aux difficultés quotidiennes de l'enseignement de la Statistique et qui aimeraient plus d'information, davantage d'idées pédagogiques, un plus grand nombre d'exercices corrigés pour être encore plus efficaces auprès de leurs élèves.
C'est beaucoup diront les membres du GRES et le rédacteur en chef qui s'engagent les uns à fournir les articles avant tel jour dernier délai, l'autre à « boucler » le bulletin pour la reprographie pour telle date sans faute.

J'en profite pour signaler à nos fidèles lecteurs que les bulletins sont entièrement réalisés par les membres du GRES : écriture des articles, frappe, mise en page, réalisation de la maquette fournie à l'entreprise qui réalise la duplication. Ceci demande un important travail qui s'ajoute évidemment à un service complet d'enseignant puisque, pour l'instant, même dans le cadre de l'opération Pygmalion, le travail des membres du GRES est bénévole. Ceci explique en partie, la variation de l'effectif du groupe qui après s'être maintenu longtemps à 14 atteint actuellement le nombre fatidique de 10.

L'opération Pygmalion qui nous fournit le budget de fonctionnement (déplacements, édition et expédition des bulletins, matériel bibliographique) touche à sa fin. Que va faire le GRES ? poursuivre dans les mêmes conditions, dans le cadre d'une nouvelle opération de la DGER ? Je ne crois pas cela possible. Je pense que la DGER doit considérer que ce travail de réflexion et d'animation pédagogique, le travail de préparation des stages de formation continue fait partie intégrante du service de l'enseignant. En conséquence, la DGER doit attribuer à chaque enseignant volontaire et choisi pour accomplir ce travail sur la base d'un contrat passé entre l'ENFA (ou tout autre établissement d'enseignement supérieur qui pilote l'action), l'établissement d'exercice et l'enseignant concerné, un certain nombre d'heures de décharge effective. En effet, ce qui nous fait le plus cruellement défaut c'est le **temps**. Nous avons des idées, bon disons quelques idées, pour poursuivre dans l'esprit de ce qu'a fait le GRES jusqu'à maintenant, mais nous voulons avoir du temps pour mettre en œuvre ces idées correctement.

Mais j'y pense, n'est-il pas présomptueux de poursuivre une action dont on ne sait pas, ou pas assez, si elle est utile aux enseignants ?
Dans l'éditorial du n°1, j'écrivais :

*« … le GRES souhaite : *mener une réflexion sur l'enseignement des Probabilités et de la Statistique dans l'enseignement secondaire et supérieur court agricole,*

élaborer des outils pédagogiques à destination des professeurs de l'enseignement agricole,

concevoir et proposer des actions de formation continue,

mener une réflexion sur les programmes des filières de l'enseignement secondaire et supérieur court agricole.… »

J’ai la fierté de constater que nous avons, une fois n’est pas coutume, atteint (partiellement, soyons modestes) les objectifs que nous nous étions fixés et j’en remercie tous les membres du groupe qui n’ont pas ménagé leurs efforts.
Nous avons, bien-entendu, quelques remontées favorables lors des stages de formation continue ou lors de rencontres avec certaines ou certains d’entre vous, mais la grande déception du groupe est le peu de contacts établis directement avec les collègues qui lisent les bulletins. Je vous assure que tenir une rubrique « Courrier des lecteurs » est un important travail d’imagination. Nous avons besoin de savoir comment vous percevez ce que nous écrivons, pourquoi vous n’êtes pas d’accord avec tel ou tel article, comment vous avez présenté la théorie de l’échantillonnage ou les tests paramétriques ou autre. Toujours dans ce même éditorial du n°1, j’écrivais encore :

« …il [ce bulletin] *doit être aussi une tribune où vous pourrez vous exprimer, faire part à tous vos collègues de vos problèmes pédagogiques, de vos solutions, de vos outils. Il doit surtout être ce que vous souhaitez qu’il soit et pour cela nous comptons beaucoup sur votre courrier,… »*

Là, permettez-moi de vous dire que vous n’avez pas atteint les objectifs que nous vous avions fixés. Vous pouvez encore vous rattraper*, vous devez vous rattraper* pour nous donner la force de conviction qui nous permettra de défendre et de faire aboutir nos demandes auprès des services de la DGER.

D11 : SERIES STATISTIQUES A DEUX VARIABLES

Que disent le programme et les recommandations pédagogiques ?

Le programme : **Statistique descriptive**

Séries statistiques à deux variables : nuage de points ; ajustement affine (méthode des moindres carrés) ; ajustements, qui par un changement de variable, se ramènent à un ajustement affine ; régression, coefficient de corrélation.

Les recommandations pédagogiques : Des situations de la vie économique, des sciences et techniques seront exploitées pour des études de régression. On distinguera variable explicative et variable expliquée. On veillera à attirer l'attention des étudiants sur l'étude des résidus (on vérifiera que leur somme est nulle). La représentation graphique des résidus permettra de vérifier le bien-fondé du modèle d'ajustement envisagé : cette représentation ne doit laisser apparaître aucune tendance. On pourra déterminer le coefficient de détermination et on en donnera une interprétation.

Précisions de l'inspection.

Dans le programme et les recommandations pédagogiques qui l'accompagnent seul le mot « régression » peut poser problème et donner lieu à des interprétations différentes. Faut-il s'en tenir à la recherche de la courbe d'ajustement (dite aussi courbe de régression ou courbe d'estimation), au calcul des résidus et à leur interprétation, au calcul de l'estimation de la variable expliquée pour une valeur donnée de la variable explicative ? Ou bien faut-il aller plus loin, préciser les hypothèses du modèle linéaire et aborder les problèmes d'inférence statistique qui lui sont liés ? L'ambiguïté du mot « régression » a posé question à l'équipe du GRES qui a sollicité l'avis de l'inspection.

A l'occasion d'une réunion de travail, le GRES a invité Monsieur l'Inspecteur principal PACULL. Concernant le mot « régression » figurant dans le libellé du programme D11, Monsieur PACULL a précisé que l'idée qui avait prévalu lors de l'élaboration du programme était celle correspondant à la première interprétation. Il a insisté sur le fait que dans l'étude de séries statistiques à deux variables, les variables sont statistiques et non aléatoires, le hasard n'intervient pas, on n'a pas une situation probabiliste. Dès lors il n'est pas possible de faire de l'inférence, en faire constituerait une erreur.

Avec les moyens de calcul dont on dispose aujourd'hui, que faire sur les séries statistiques à deux variables ?

Compte tenu des moyens de calcul dont disposent les étudiants, il serait désuet de leur demander de construire le traditionnel tableau de calcul pour déterminer par la méthode des moindres carrés le coefficient de corrélation linéaire, le coefficient directeur et l'ordonnée à l'origine de la droite d'ajustement (ceci ne veut évidemment pas dire qu'il ne faille pas traiter manuellement un ou deux exemples d'école).

Tous possèdent des calculatrices qui fournissent directement ces résultats. Ils ne comprendraient pas que l'on n'utilise pas ces possibilités qui soulagent des calculs longs et fastidieux que nécessite la méthode des moindres carrés.
En outre toutes les calculatrices de la dernière génération fournissent non seulement les résultats pour les ajustement affines mais aussi pour les ajustements qui peuvent se ramener au modèle linéaire ou multilinéaire : ajustement exponentiel, ajustement puissance et ajustements polynomiaux (jusqu'au degré 4 pour certaines). De plus elles proposent un mini tableur (mode liste) qui permet de faire très simplement les opérations élémentaires de calcul matriciel et avec lequel le calcul des résidus devient facile. Dès maintenant certaines calculatrices vont jusqu'à fournir la série des résidus. On n'est plus limité par le nombre des observations.

Les présupposés théoriques et les pièges à éviter.

Concernant les résidus, les résultats théoriques utilisés sont établis dans le cas du modèle linéaire :

- Leur somme est nulle.
- Ils se répartissent à peu près équitablement entre résidus positifs et résidus négatifs et sans qu'aucune tendance n'apparaisse.

Etablis dans le cadre du modèle linéaire ces résultats ne s'appliquent que dans le cas d'un ajustement affine.

- Dans le cas de l'ajustement exponentiel d'une série (x_i,y_i), l'ajustement affine s'applique à la série transformée (x_i,z_i) avec $z_i = \ln(y_i)$ et bien entendu les résultats concernant les résidus s'appliquent à cette série mais pas à la série (x_i,y_i).
  En particulier la somme des résidus $e_i = y_i - \hat{y}_1$ de la série (x_i,y_i) n'est pas nulle. Certaines calculatrices fournissent ces résidus.
- Dans le cas de l'ajustement puissance d'une série (x_i,y_i), l'ajustement affine s'applique à la série transformée (u_i,z_i) avec $u_i = \ln(x_i)$ et $z_i = \ln(y_i)$ et les résultats concernant les résidus s'appliquent à cette série mais pas à la série (x_i,y_i).
  En particulier la somme des résidus $e_i = y_i - \hat{y}_i$ de la série (x_i,y_i) n'est pas nulle.

  Certaines calculatrices fournissent ces résidus.

La plupart des calculatrices fournissent directement un coefficient de corrélation linéaire r dans le cas d'un ajustement exponentiel ou d'un ajustement puissance. Il faut bien comprendre que :

- Dans le cas d'un ajustement exponentiel, le coefficient de corrélation linéaire fourni par la calculatrice est celui correspondant à l'ajustement affine de la série transformée (x_i,z_i). C'est le coefficient de corrélation linéaire entre les variables *X* et *Z*.
- Dans le cas d'un ajustement puissance, le coefficient de corrélation linéaire fourni par la calculatrice est celui correspondant à l'ajustement affine de la série transformée (u_i,z_i). C'est le coefficient de corrélation linéaire entre les variables *U* et *Z*.

Activité - Etude de la croissance d'une tige de tomate. (D'après un document de biologie)

On mesure l'allongement *X* de la tige d'une tomate, exprimé en mm/j, en fonction de la température diurne *T*, exprimée en °C. Le tableau suivant fournit le relevé des valeurs du couple de variables statistiques (*T*, *X*).

t_i	5	7	10	13	15	18	20	22	25	28	30
x_i	1	2	3	6	8	10	11	15	17	20	23

1- Construire le nuage de points représentant cette série statistique double. Que suggère l'examen du nuage ?

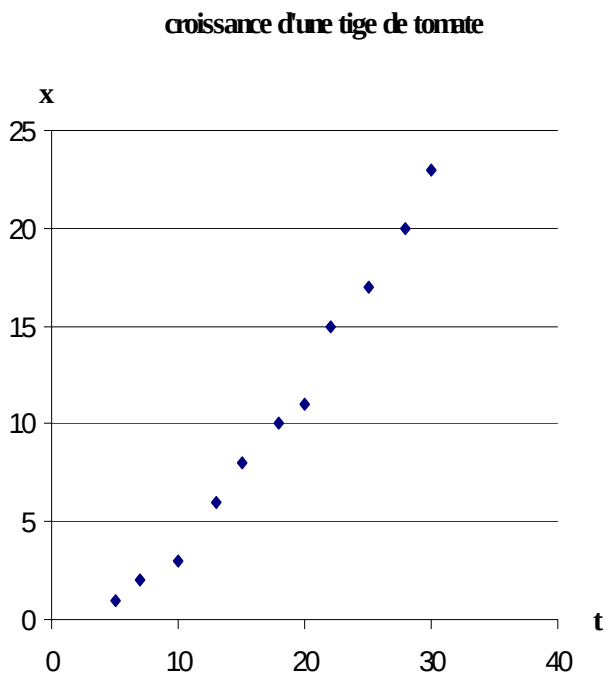
2- En utilisant une calculatrice :
a. donner une équation de la droite d'ajustement de *X* en *T*, obtenue par la méthode des moindres carrés.
b. construire la droite.
c. donner le coefficient de corrélation linéaire r_1 entre *X* et *T*.
d. donner le coefficient de détermination r_1^2 et interpréter ce coefficient.

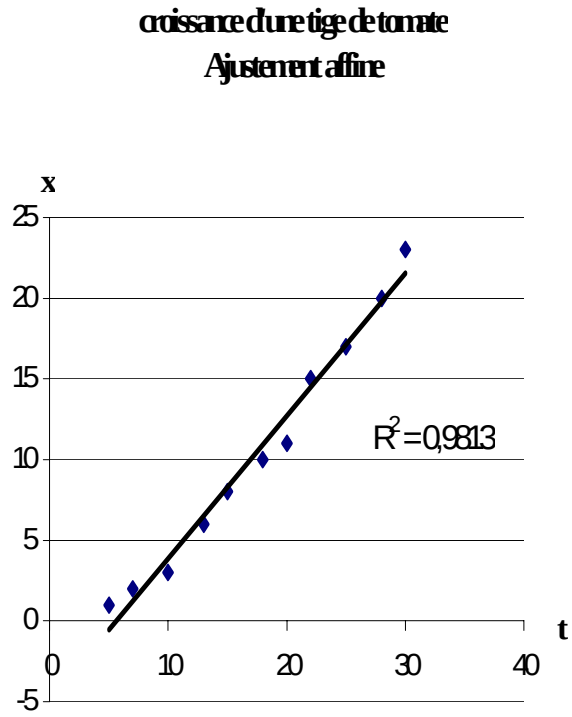
3- *a*. Calculer les écarts résiduels $e_i = x_i - \hat{x}_i$ où $\hat{x}_i$ est la valeur estimée correspondant à t_i, calculée en utilisant l'équation de la droite d'ajustement obtenue à la question 2- *a*.
b. Représenter les résidus en fonction de la variable explicative puis de la variable expliquée.
c. Que peut-on dire de la répartition des résidus et de l'ajustement affine ? Quelle remarque peut-on faire sur les coefficients de corrélation et de détermination entre *X* et *T* ?

4- **Ajustement exponentiel**. On pose $z = \ln x$:
a. Calculer les z_i. Construire le nuage de points représentant la série (t_i,z_i).
b. En utilisant une calculatrice : donner une équation de la droite d'ajustement de *Z* en *T*, obtenue par la méthode des moindres carrés.
c. Calculer les écarts résiduels $e'_i = z_i - \hat{z}_i$ où $\hat{z}_i$ est la valeur estimée correspondant à t_i, calculée en utilisant l'équation de la droite d'ajustement obtenue à la question 4- *b*.
***d*.** Représenter les résidus en fonction de la variable explicative *T*. Que peut-on dire de la répartition des résidus ? L'ajustement affine de la série (t_i,z_i) est-il envisageable ?

5- **Ajustement puissance**. On pose $u = \ln t$:
a. Calculer les u_i. Construire le nuage de points représentant la série (u_i,z_i).
b. En utilisant une calculatrice : donner une équation de la droite d'ajustement de *Z* en *U*, obtenue par la méthode des moindres carrés.
c. Calculer les écarts résiduels $e'_i = z_i - \hat{z}_i$ où $\hat{z}_i$ est la valeur estimée correspondant à u_i, calculée en utilisant l'équation de la droite d'ajustement obtenue à la question 5- *b*.
d. Représenter les résidus en fonction de la variable explicative *U*. Que peut-on dire de la répartition des résidus ? L'ajustement affine de la série (u_i,z_i) est-il envisageable ?
e. donner les coefficients de corrélation linéaire et de détermination entre les variables statistiques *Z* et *U* . Que peut-on dire de l'ajustement puissance ?
f. Déduire de l'équation de la droite d'ajustement de *Z* en *U* une relation de la forme $\hat{x} = f(t)$ liant $\hat{x}$ et *t*. Construire la courbe d'équation $\hat{x} = f(t)$ dans le même repère que le nuage de points représentant la série (t_i,x_i).

Eléments de correction.





Nuage de points représentant la série statistique (t_i,x_i).

Au vu du nuage de points on peut légitimement envisager un ajustement affine de la série (t_i,x_i)..

Remarque :

Le choix des unités, pour le repère dans lequel on construit le nuage de points, n'est pas neutre. Si celles-ci sont convenablement choisies on constate, dans l'exemple traité, que le nuage est incurvé vers le haut, ce qui conduit à penser que le modèle linéaire n'est peut-être pas le plus approprié. Mais un choix inadéquat des unités peut écraser le nuage, masquer cette courbure et induire que le seul modèle pertinent est l'ajustement affine. **Seul le diagramme des résidus va permettre de démasquer cette erreur**, en faisant apparaître que les résidus ont une distribution tendancieuse.

Ajustement affine de la série (t_i,x_i).

Résultats calculatrice : a = 0,883 *résultats numériques à 10^{-3} près*

b = -4,941

Equation de la droite d'ajustement : $\hat{x} = 0{,}883.t - 4{,}941$

Coefficient de corrélation linéaire : r_1 = 0,991

Coefficient de détermination : r_1^2 = 0,981

Interprétation du coefficient de détermination : 98% de la variabilité totale de X est expliquée par l'ajustement affine.

L'ajustement par une droite d'équation $\hat{x} = a.t + b$ semble, de ce point de vue, parfaitement justifié.

Toutefois l'étude de l'ajustement affine serait incomplète sans celle des résidus

ableau de calcul des résidus :

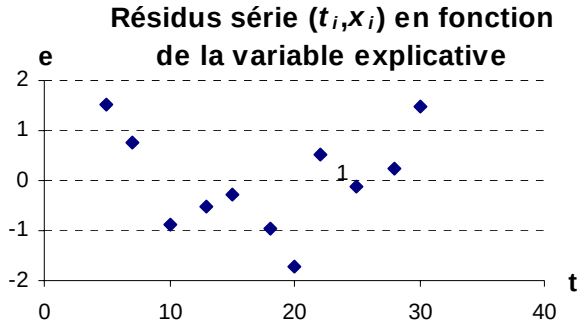
t_i	x_i	résidus série (t_i,x_i)	$z_i = \ln(x_i)$	résidus série (t_i,z_i)	$u_i = \ln(t_i)$	résidus série (u_i,z_i)
5	1	1,5	0	-0,54	1,609	-0,04
7	2	0,8	0,693	-0,08	1,946	0,06
10	3	-0,9	1,099	-0,03	2,303	-0,15
13	6	-0,5	1,792	0,32	2,565	0,08
15	8	-0,3	2,079	0,37	2,708	0,12
18	10	-0,9	2,303	0,25	2,89	0,02
20	11	-1,7	2,398	0,11	2,996	-0,06
22	15	0,5	2,708	0,19	3,091	0,08
25	17	-0,1	2,833	-0,04	3,219	-0,02
28	20	0,2	2,996	-0,23	3,332	-0,05
30	23	1,5	3,135	-0,32	3,401	-0,03
		Somme des résidus 0		Somme des résidus 0		Somme des résidus 0

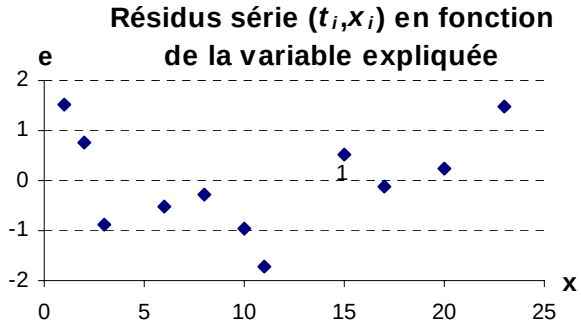
Etude des résidus : $e_i = x_i - \hat{x}_i$

$= x_i - (a.t_i + b)$

Calcul des résidus voir tableau ci-dessus.
On vérifie que la somme des résidus est nulle aux arrondis près.
L'examen des diagrammes des résidus montre que les points qui les représentent sont mal répartis par rapport aux axes des abscisses. Les résidus sont négatifs au centre positifs ailleurs.

Représentations des résidus :





Cela conduit à chercher un modèle plus pertinent pour traduire la relation entre *X* et *Y*.

La forme des nuages, représentant respectivement la série et les résidus, suggère qu'une courbe d'équation soit $\hat{y} = b.a^x$ soit $\hat{y} = b.x^a$ peut fournir un meilleur modèle d'ajustement.

Remarque : l'ajustement affine n'est manifestement pas le modèle le mieux adapté et pourtant coefficients de corrélation linéaire et de détermination prennent des valeurs très proches de 1 (respectivement r_1 = 0,99 et r_1^2 = 0,98). Cet exemple montre que ces coefficients ne permettent pas de juger de la pertinence du modèle retenu.
Le coefficient de détermination permet de mesurer la qualité d'un ajustement affine mais il ne permet pas de juger de sa validité.

Ajustement exponentiel.

$$\hat{x} = b.a^t \Leftrightarrow \ln \hat{x} = \ln b + t.\ln a$$

$$\Leftrightarrow \begin{cases} \ln \hat{x} = z,\ \ln a = A,\ \ln b = B \\ z = A.t + B \end{cases}$$

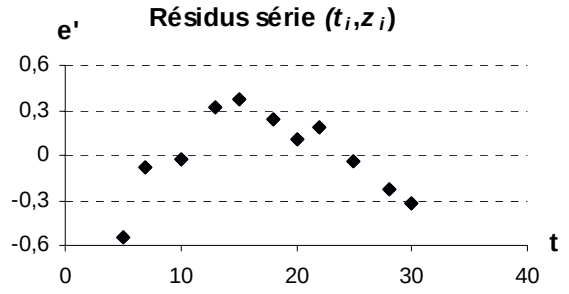
Si les points de coordonnées (t_i,x_i) sont ajustés par la courbe d'équation $\hat{x} = b.a^t$ alors les points de coordonnées (t_i,z_i) sont ajustés par la droite d'équation $\hat{z} = A.t + B$ et réciproquement.
Pour répondre à la question

- on calcule les valeurs $z_i = \ln(x_i)$, on construit le nuage de points représentant la série (t_i,z_i),
- selon l'aspect du nuage de points on envisage ou non un ajustement affine de la série (t_i,z_i).
- on effectue l'ajustement affine de la série (t_i,z_i)
  Résultats calculatrice : A = 0,117
  B = -0,042
  Equation de la droite d'ajustement : $\hat{z} = 0{,}117.t - 0{,}042$
- on calcule les résidus (voir tableau de calcul) et on les représente

Etude des résidus :

$$e'_i = z_i - \hat{z}_i$$
$$= z_i - (A.t_i + B)$$

On vérifie que la somme des résidus est nulle aux arrondis près.



Le diagramme des résidus montre que le modèle exponentiel n'est pas adapté.
On ne retient pas l'ajustement exponentiel.

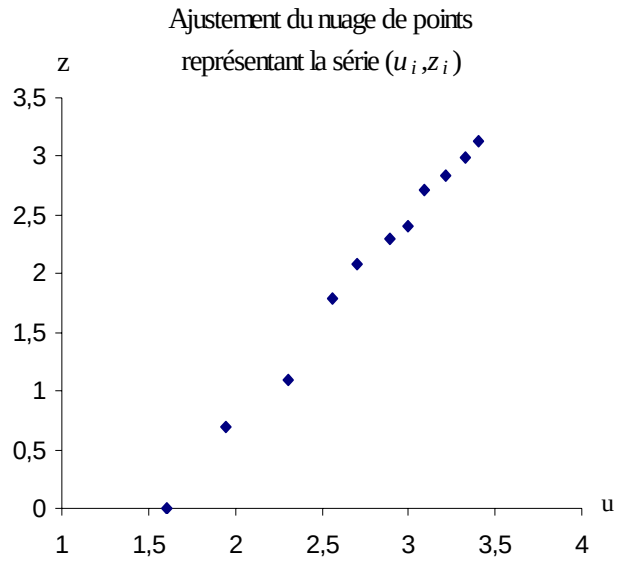
Ajustement puissance.

$$\hat{x} = b.t^a \Leftrightarrow \ln \hat{x} = \ln b + a.\ln t$$

$$\Leftrightarrow \begin{cases} \ln t = u,\ \ln \hat{x} = \hat{z},\ \ln b = B \\ \hat{z} = a.u + B \end{cases}$$

Si les points de coordonnées (t_i,x_i) sont ajustés par la courbe d'équation $\hat{x} = b.t^a$ alors les points de coordonnées (u_i,z_i) sont ajustés par la droite d'équation $\hat{z} = a.u + B$ et réciproquement.
Pour répondre à la question

- on calcule les $u_i = \ln(t_i)$ (voir tableau de calcul), on construit le nuage de points représentant la série (u_i,z_i),
- selon l'aspect du nuage de points on envisage ou non un ajustement affine de la série (u_i,z_i).



Les points du nuage sont approximativement alignés. L'ajustement par une droite d'équation $\hat{z} = a.u + B$ est tout à fait envisageable.

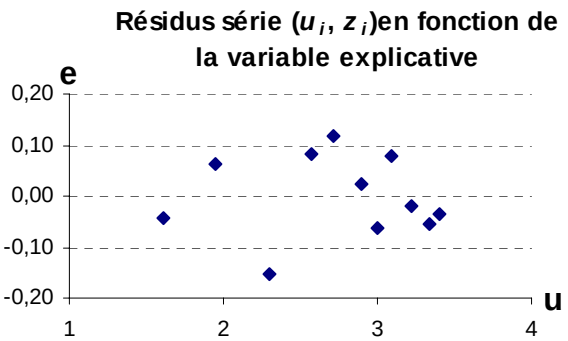
Résultats calculatrice : a = 1,745
B = -2,765
Equation de la droite d'ajustement : $\hat{z} = 1{,}745.u - 2{,}765$

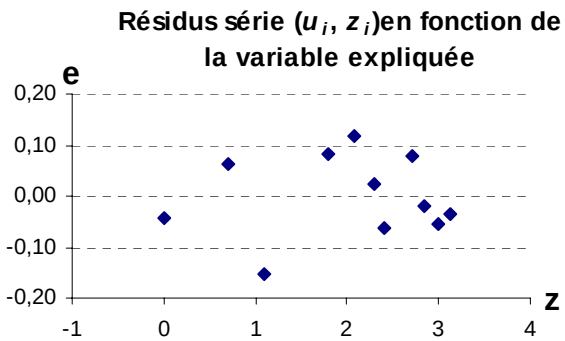
Pour étudier la validité du modèle on calcule les résidus, on construit leurs représentations graphiques en fonction de la variable explicative ou de la variable expliquée.

Etude des résidus.

Résidus : $e_i'' = z_i - \hat{z}_i$
$= z_i - (a.u_i + B)$

Calcul des résidus voir tableau. On vérifie que la somme des résidus est nulle aux arrondis près.
Représentation des résidus :





A première vue, il semble que les résidus soient convenablement répartis de part et d'autre des axes des abscisses et ne fassent pas apparaître de tendance. Toutefois un examen plus attentif montre que cette apparence tient au seul résidu –0,15 correspondant au point de coordonnées u_i = 2,303 z_i = 1,099. Les nuages des résidus, privés de ce point, font apparaître nettement une tendance.

L'ajustement puissance, lui non plus, ne semble pas très bien adapté.

Dans le cadre du D11.
Toutefois des trois ajustements au programme du D 11, c'est celui pour lequel les résidus sont les mieux répartis (aucune tendance perceptible) et, en restant dans le cadre du programme, c'est celui que nous retiendrons.
Nous allons maintenant évaluer la qualité de cet ajustement et pour ceci calculer le coefficient de détermination.
Résultats calculatrice : Coefficient de corrélation linéaire entre *U* et *Z* : r_3 = 0,997

Coefficient de détermination entre *U* et *Z* : r_3^2 = 0,994

Plus de 99% de la variabilité totale de *Z* est expliquée par l'ajustement affine.

Conclusion.
Des trois ajustements au programme du D11, l'ajustement puissance est le plus approprié pour établir la relation exprimant l'allongement d'une tige de tomate en fonction de la température.

Equation de la courbe d'ajustement .
Une équation de la droite d'ajustement des moindres carrées de z en u est :

$\hat{z} = 1{,}745.u - 2{,}765$ *avec* $u = \ln t$ *et* $\hat{z} = \ln \hat{x}$

on remplace *u* et $\hat{z}$ par leur expression en fonction de *t* et $\hat{x}$: $\ln \hat{x} = 1{,}745.\ln t - 2{,}765$

propriété de la fonction logarithme népérien : pour tout réel strictement positif *x*, pour tout rationnel α, $\alpha \ln x = \ln x^{\alpha}$: $\ln \hat{x} = \ln t^{1{,}745} - 2{,}765$

passage aux exponentielles : $e^{\ln \hat{x}} = e^{\ln t^{1{,}745} - 2{,}765}$

propriété de la fonction exponentielle : pour tout réel strictement positif *x*, $e^{\ln x} = x$: $\hat{x} = e^{\ln t^{1{,}745} - 2{,}765}$

propriété de la fonction exponentielle (propriété des puissances d'un nombre) $e^{a+b} = e^a.e^b$: $\hat{x} = e^{\ln t^{1{,}745}}.e^{-2{,}765}$

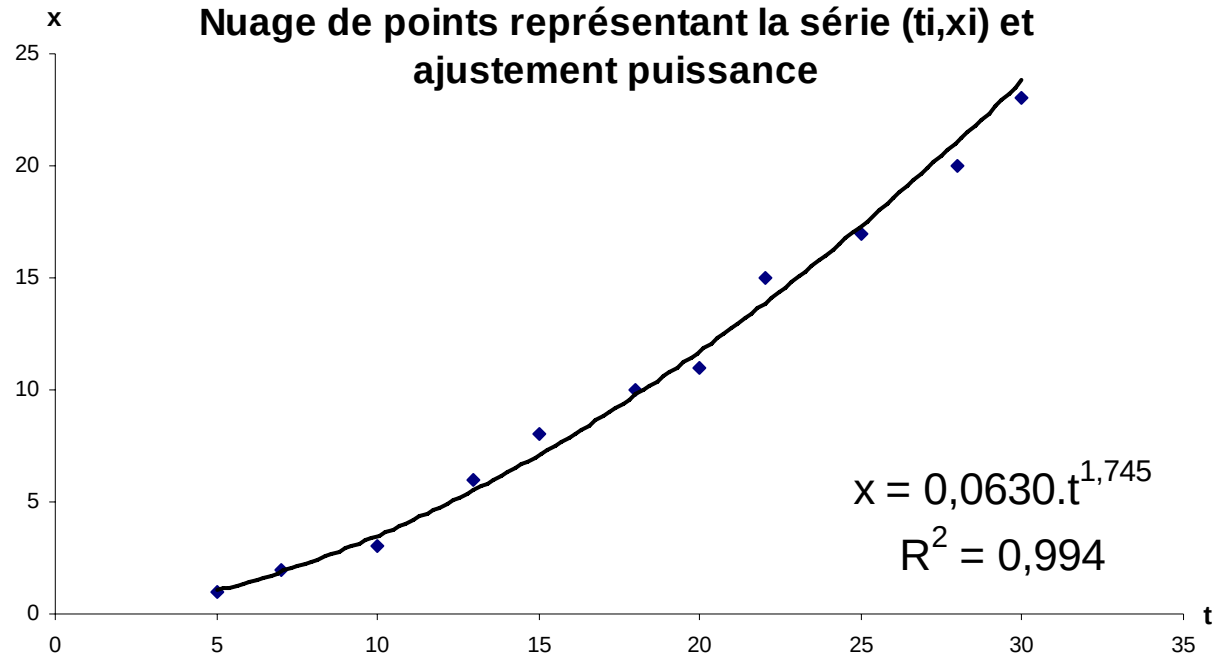
propriété de la fonction exponentielle : pour tout réel strictement positif *x*, $e^{\ln x} = x$: $\hat{x} = e^{-2{,}765}.t^{1{,}745}$ or

$e^{-2{,}765} = 0{,}063$

$$\hat{x} = 0.063 \times t^{1{,}745}$$

Tableau de valeurs

t	5	10	15	20	25	30
$\hat{x}$	1,0	3,5	7,1	11,7	17,3	23,8



Au-delà du D11.

Les calculatrices proposent des ajustements par des polynômes, alors pourquoi ne pas aller un peu plus loin, au moins entre profs ?

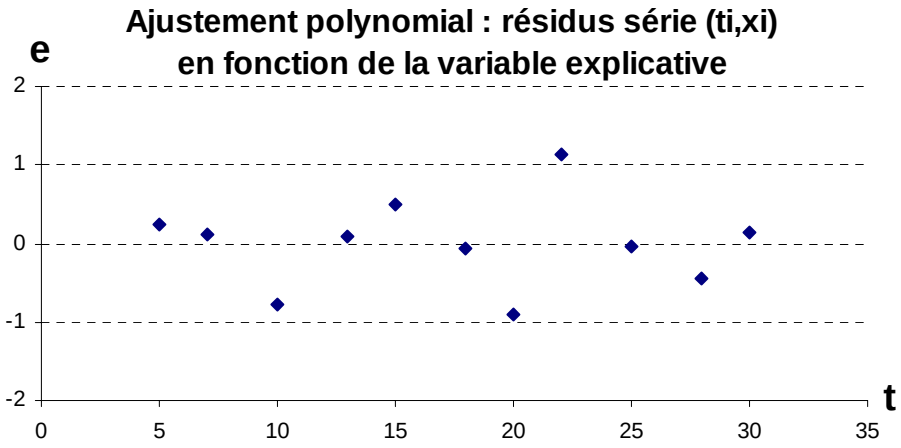
Ajustement par un polynôme du second degré .

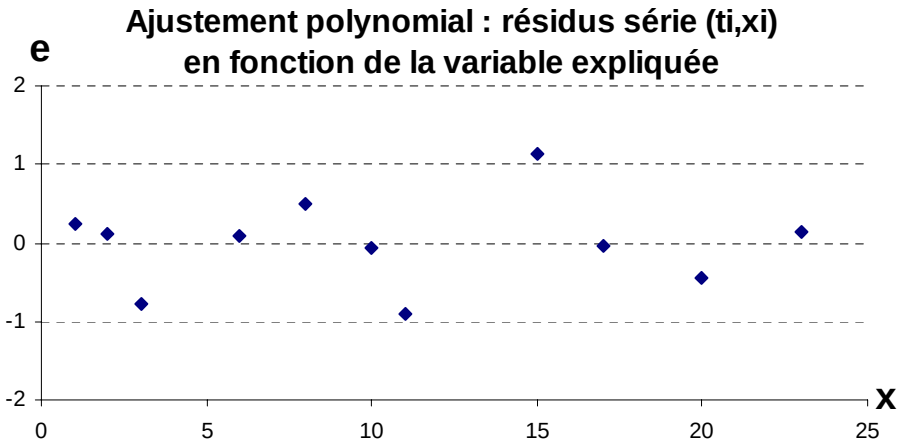
$\hat{x} = a.t^2 + b.t + c$

Résultats calculatrice : $a = 1.41.10^{-2}$, b = 0,391, c = -1,54

Calcul des résidus :

ti	5	7	10	13	15	18	20	22	25	28	30
xi	1	2	3	6	8	10	11	15	17	20	23
résidus	0,23	0,11	-0,78	0,08	0,51	-0,06	-0,91	1,13	-0,03	-0,44	0,15





Des quatre ajustements étudiés, l'ajustement quadratique est celui qui conduit à la meilleure distribution des résidus (voir représentations graphiques) ce qui ne signifie pas qu'il n'y en ait pas de plus pertinent.

C'est celui que l'on retient.

Pour évaluer la qualité de cet ajustement on calcule le coefficient de détermination en utilisant sa définition : $r^2 = \dfrac{\sum(\hat{x}_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$

A vos calculatrices !
Si on effectue le calcul du coefficient de détermination avec la calculatrice CASIO *fx-6910G,* celle-ci fournit le résultat avec 9 décimales : 0,994137663.
On retient $r^2 = 0{,}994$
99,4 % de la variabilité totale de X est expliquée par l'ajustement par un polynôme du second degré.

Voici, toujours avec la calculatrice CASIO *fx-6910G*, la liste des données et des commandes à saisir pour calculer les coefficients de l'ajustement polynomial et le coefficient de détermination :

En mode STAT : t_i en liste 1, x_i en liste 2, puis CALC SET 2Var X : List1 2Var Y : List2 2Var F : 1 QUIT CALC REG X^2.

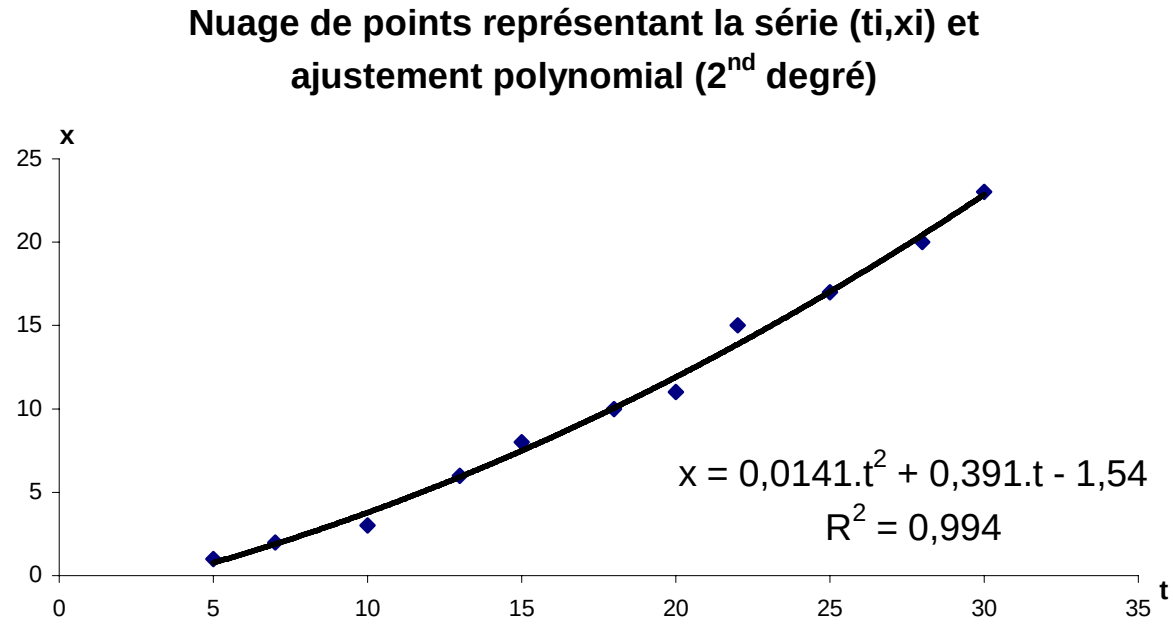
En mode RUN : OPTN LIST Sum ((VARS STAT GRPH a OPTN LIST List 1 ^ 2 + VARS STAT GRPH b OPTN LIST List 1 + VARS STAT GRPH c – OPTN LIST Mean(List 2)) ^ 2) / Sum ((List 2 – Mean(List 2)) ^ 2)

Ce qui donne à l'écran de la calculatrice :
Sum((aList 1^2+bList 1+c–Mean(List 2))^2)/Sum((List 2– Mean(List 2))^2)

Certaines calculatrices (TI 83) donnent le coefficient de détermination dans le cas d'un ajustement quadratique.

Avec le tableur EXCEL pour obtenir ces résultats on peut utiliser :

- soit le grapheur : après avoir représenté la série à deux variables par un nuage de points, on sélectionne ce nuage en cliquant sur un de ses points, puis ou on fait un clic droit ou on rentre dans le menu « graphique » et on utilise la commande « Ajouter une courbe de tendance ». Une boite de dialogue se présente avec deux onglets, l'onglet « Type » permet de choisir l'ajustement, l'onglet « Option » permet d'afficher sur le graphique l'équation de la courbe d'ajustement et le coefficient de détermination.
- soit l'outil d'analyse « Régression linéaire » de l'utilitaire d'analyse.



-

= - = - = - = - = - = - = - = -

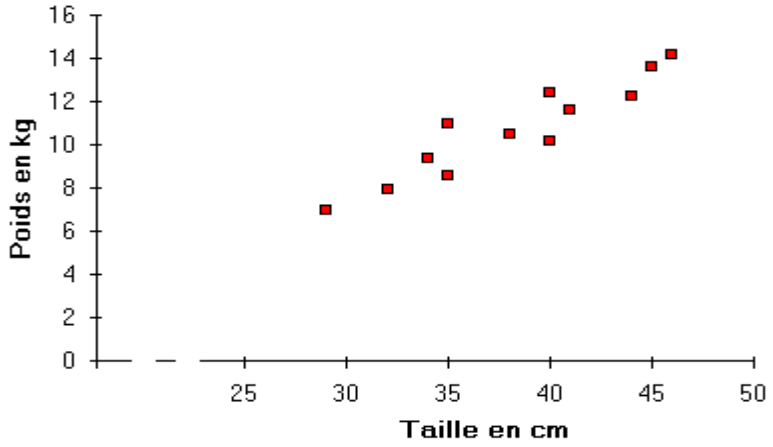
LES PETITS HOMMES VERTS

Notre première étude sur les petits hommes verts date de mai 1985.
Nous nous proposions de déterminer la nature de la relation entre : la taille X mesurée en cm et le poids Y exprimé en kg de ces petits hommes verts.
Nous disposions de données mesurées sur un échantillon de 12 petits hommes verts.

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	29	32	35	34	40	38	35	41	44	40	45	46
y_i	7	8	8.6	9.4	10.2	10.5	11	11.6	12.3	12.4	13.6	14.2

La première tâche fut bien sûr de représenter le nuage des points de coordonnées (x_i , y_i).

* Représentation graphique des données:



Compte tenu du profil du nuage, rien ne semblait s'opposer à un ajustement affine du nuage.
En d'autres termes , nous proposions la modélisation suivante:
Si $\mathbf{x_i}$ est la taille d'un petit homme vert

* son poids théorique vaut $y_i^* = \alpha \quad x_i \quad + \quad \beta$
* la différence entre le poids réel $\mathbf{y_i}$ de cet individu et le poids théorique est l'erreur notée $\boldsymbol{\varepsilon_i}$
* la répartition des erreurs ne dépend pas de $\mathbf{x_i}$
* Cette répartition est supposée Normale $\mathbf{N(0,\sigma)}$.

Modèle de paramètres α **et** β.

$$Y = \alpha \quad x \quad + \quad \beta \quad + \quad \varepsilon$$

α **et** β sont les paramètres du modèle.

La taille	**x**	est la variable ***indépendante (explicative)***et ***non aléatoire***
Le poids	**Y**	est la variable ***aléatoire dépendante(expliquée).***
L'erreur	ε	est une variable aléatoire distribuée selon la loi Normale $\mathbf{N(0,\sigma)}$.
Les tailles	$\mathbf{x_i}$	sont 12 valeurs de taille **x** connues sans erreur de mesure.
Les erreurs	$\boldsymbol{\varepsilon_i}$	sont 12 réalisations mutuellement indépendantes de la variable ε
Chaque poids	$\mathbf{y_i}$	est une réalisation de la variable $Y_i = \alpha \quad x_i \quad + \quad \beta \quad + \quad \varepsilon$

α, β et σ **sont les paramètres du modèle.**

Nous devions alors :

1. Estimer les paramètres α **et** β du modèle.
2. Apprécier l'adéquation du modèle aux données recensées et à d'éventuelles données supplémentaires.

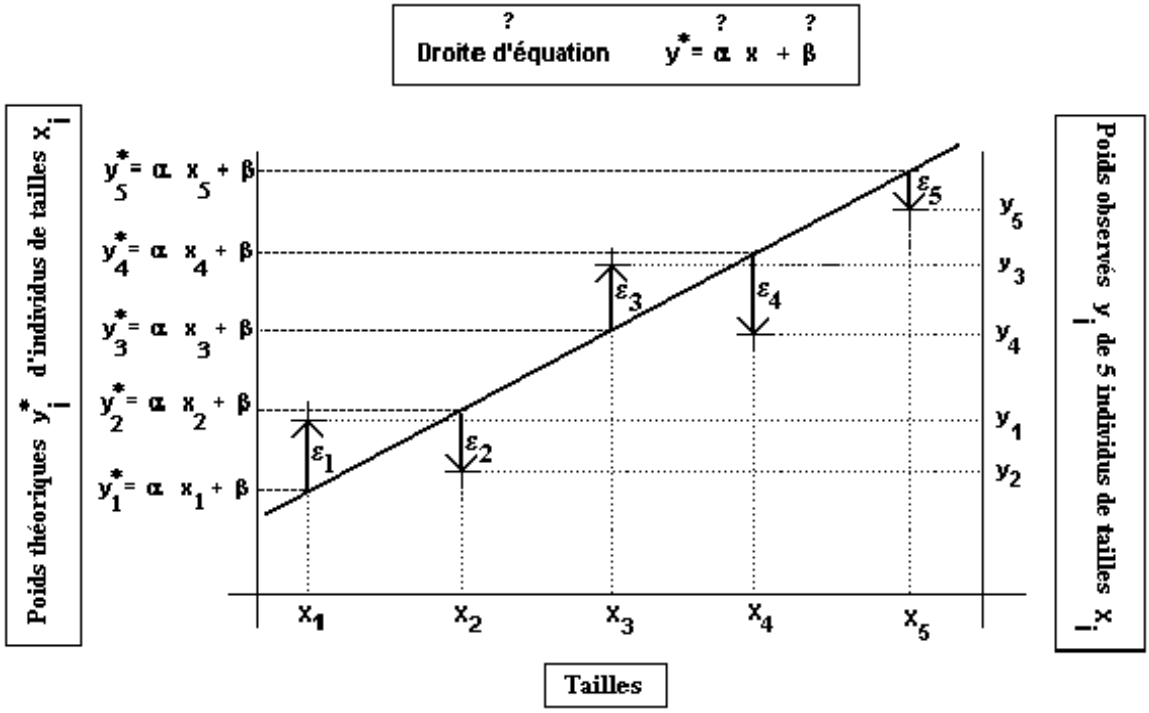
La "*meilleure*" méthode d'ajustement est la **méthode des moindres carrés.**

Droite d'ajustement selon la méthode des moindres carrés

Pour simplifier la présentation de cette méthode, nous allons dans les calculs et illustrations qui suivent, nous limiter à un nuage de 5 points.

A. Résidus et Erreurs : des notions à bien différencier.

1. Le schéma ci-dessous permet de comprendre la notion d'erreur.



Les paramètres α **et** β sont connus par le dieu des petits hommes verts, mais inaccessibles à nous, pauvres mortels.

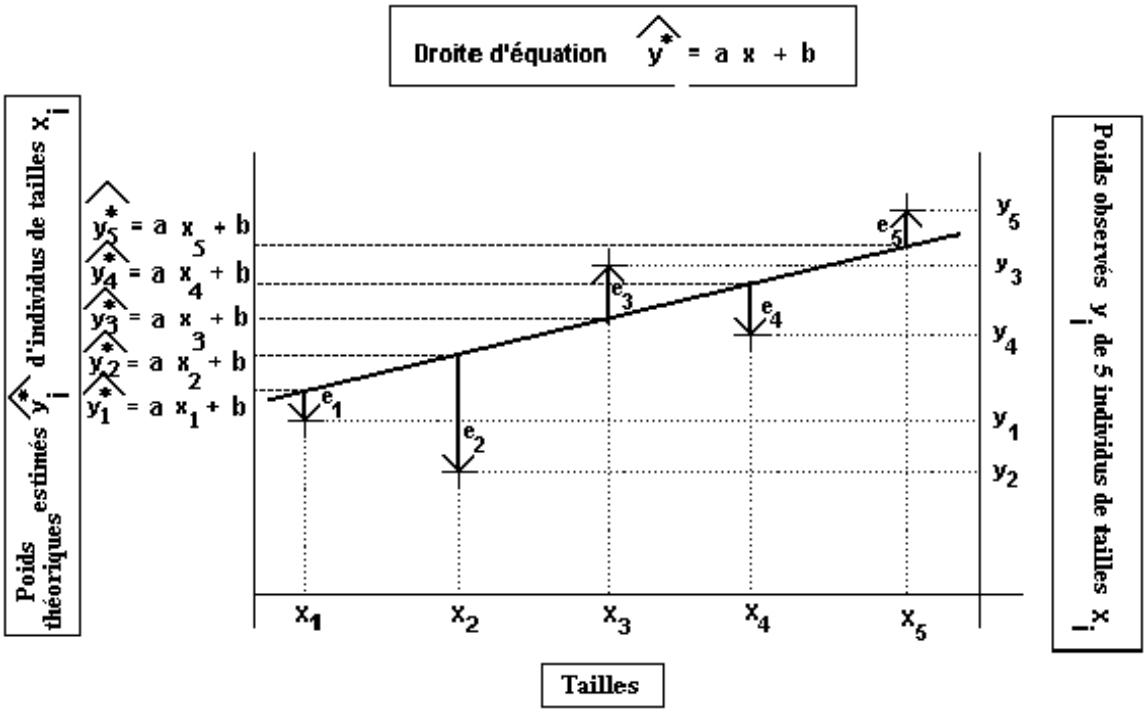
α et β nous sont à jamais inconnus

Par conséquent, les cinq erreurs sont inconnues

ε_1 , ε_2 , ε_3 , ε_4 , ε_5 **nous sont à jamais inconnues**

2. Le schéma ci-dessous permet de comprendre la notion de résidu.

Considérons une droite d'**ajustement empirique** d'équation **y = a x + b**



Les observations faites nous permettent de proposer deux nombres **a et b** comme estimations empiriques des paramètres **α et β**	$\hat{y}^*_1$, $\hat{y}^*_2$, $\hat{y}^*_3$, $\hat{y}^*_4$, $\hat{y}^*_5$ sont les estimations empiriques des poids théoriques d'individus de tailles respectives $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$, $\mathbf{x_4}$, $\mathbf{x_5}$

e_1 , e_2 , e_3 , e_4 , e_5
sont
les **résidus** relatifs à l'ajustement proposé **pour l'échantillon observé**

B. Droite d'ajustement.

Quelles sont les contraintes auxquelles doit satisfaire la droite d'ajustement selon la méthode des moindres carrés ?

1. La moyenne des résidus e_i doit être nulle. ***(justesse)***

$$e_1 + e_2 + e_3 + e_4 + e_5 = 0$$

2. La variabilité des résidus e_i doit être minimale. ***(précision)***

$e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2$ doit être minimale

des contraintes somme toute très naturelles !!

1. Analysons la première contrainte,

Ajoutons membre à membre les 5 égalités suivantes :

$$\begin{aligned} y_1 &= a\,x_1 + b + e_1 \\ y_2 &= a\,x_2 + b + e_2 \\ y_3 &= a\,x_3 + b + e_3 \\ y_4 &= a\,x_4 + b + e_4 \\ y_5 &= a\,x_5 + b + e_5 \end{aligned}$$

Nous obtenons :

$$y_1 + y_2 + y_3 + y_4 + y_5 = a\,(x_1 + x_2 + x_3 + x_4 + x_5) + 5\,b + e_1 + e_2 + e_3 + e_4 + e_5$$

Divisons chaque membre par 5 :

$$\bar{y} = a\,\bar{x} + b + \bar{e}$$

Comme $\bar{e}$ doit valoir 0, il en résulte:

$$\boxed{\bar{y} = a\,\bar{x} + b}$$

2. <u>Analysons la seconde contrainte</u>

$$\left.\begin{aligned} y_1 &= a\,x_1 + b + e_1 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_1 - \bar{y}) - a\,(x_1 - \bar{x}) = e_1$$

$$\left.\begin{aligned} y_2 &= a\,x_2 + b + e_2 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_2 - \bar{y}) - a\,(x_2 - \bar{x}) = e_2$$

$$\left.\begin{aligned} y_3 &= a\,x_3 + b + e_3 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_3 - \bar{y}) - a\,(x_3 - \bar{x}) = e_3$$

$$\left.\begin{aligned} y_4 &= a\,x_4 + b + e_4 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_4 - \bar{y}) - a\,(x_4 - \bar{x}) = e_4$$

$$\left.\begin{aligned} y_5 &= a\,x_5 + b + e_5 \\ \bar{y} &= a\,\bar{x} + b \end{aligned}\right\} \text{ donc } (y_5 - \bar{y}) - a\,(x_5 - \bar{x}) = e_5$$

Il en résulte :

$$\begin{aligned} e_1^2 &= (y_1 - \bar{y})^2 - 2a\,(x_1 - \bar{x})(y_1 - \bar{y}) + a^2\,(x_1 - \bar{x})^2 \\ e_2^2 &= (y_2 - \bar{y})^2 - 2a\,(x_2 - \bar{x})(y_2 - \bar{y}) + a^2\,(x_2 - \bar{x})^2 \\ e_3^2 &= (y_3 - \bar{y})^2 - 2a\,(x_3 - \bar{x})(y_3 - \bar{y}) + a^2\,(x_3 - \bar{x})^2 \\ e_4^2 &= (y_4 - \bar{y})^2 - 2a\,(x_4 - \bar{x})(y_4 - \bar{y}) + a^2\,(x_4 - \bar{x})^2 \\ e_5^2 &= (y_5 - \bar{y})^2 - 2a\,(x_5 - \bar{x})(y_5 - \bar{y}) + a^2\,(x_5 - \bar{x})^2 \end{aligned}$$

Notons σ_x et σ_y les écart-types respectifs des séries statistiques (x_i) et (y_i)

$$\sigma_x^2 = \frac{\sum_{i=1}^{5} (x_i - \bar{x})^2}{5} \qquad \text{et} \qquad \sigma_y^2 = \frac{\sum_{i=1}^{5} (y_i - \bar{y})^2}{5}$$

Notons σ_{xy} la covariance de la série double :

$$\sigma_{xy} = \frac{\sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y})}{5}$$

La variabilité des résidus vaut donc :

$$e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2 = 5\sigma_x^2\ a^2\ -10\ \sigma_{xy}\ a\ +\ 5\ \sigma_y^2$$

La variabilité des résidus est une fonction du second degré en a qu'il s'agit de minimiser en choisissant judicieusement a.

Etudions la fonction f définie par : $f(a) = 5\sigma_x^2\ a^2\ -10\ \sigma_{xy}\ a\ +\ 5\ \sigma_y^2$

Dérivée de f : $f'(a) = 10\sigma_x^2\ a\ -10\ \sigma_{xy}$

Signe de la dérivée de f :

$f'(a) > 0$ si et seulement si $a > \dfrac{\sigma_{xy}}{\sigma_x^2}$

$f'(a) = 0$ si et seulement si $a = \dfrac{\sigma_{xy}}{\sigma_x^2}$

Désignons par $\hat{\alpha}$ la quantité $\dfrac{\sigma_{xy}}{\sigma_x^2}$

Variations de f :

a	$-\infty$	$\hat{\alpha} = \dfrac{\sigma_{xy}}{\sigma_x^2}$	$+\infty$
f '(a)	$-$	0	$+$
	$+\infty$		$+\infty$
f (a)	↘	$f(\hat{\alpha})$	↗

La droite d'ajustement selon la méthode des moindres carrés a donc pour équation :

$$\mathbf{y} = \hat{\alpha}\ \mathbf{x} + \hat{\beta}$$

où $\hat{\alpha} = \dfrac{\sigma_{xy}}{\sigma_x^2}$ et $\hat{\beta} = \bar{y} - \hat{\alpha}\ \bar{x}$

En posant $\mathbf{r} = \dfrac{\sigma_{xy}}{\sigma_x\ \sigma_y}$, nous obtenons $\hat{\alpha} = \mathbf{r}\dfrac{\sigma_y}{\sigma_x}$.

La variabilité des résidus relatifs à cet ajustement vaut

$$f(\hat{\alpha}) = 5\sigma_x^2\ \hat{\alpha}^2\ -10\ \sigma_{xy}\ \hat{\alpha} + 5\ \sigma_y^2.$$

Exprimons cette variabilité en fonction de **r** :

$$f(\hat{\alpha}) = f\left(\mathbf{r}\frac{\sigma_y}{\sigma_x}\right) = 5\sigma_x^2\ \mathbf{r}^2\ \frac{\sigma_y^2}{\sigma_x^2}\ -10\ \sigma_{xy}\ \mathbf{r}\frac{\sigma_y}{\sigma_x} + 5\ \sigma_y^2$$

soit $$f(\hat{\alpha}) = 5\ \mathbf{r}^2\ \sigma_y^2\ -10\ \ \mathbf{r}^2\ \sigma_y^2\ +\ 5\ \sigma_y^2$$

soit $f\left(\hat{\alpha}\right) = 5 \quad \sigma_y^2 \quad \left(1 - \mathbf{r}^2\right)$

La variabilité des résidus relatifs à l'ajustement selon la méthode des moindres carrés vaut donc :

$$\boxed{e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2 = 5 \quad \sigma_y^2 \quad \left(1 - \mathbf{r}^2\right)} \qquad \text{Formule n° 1}$$

Quelque remarques :

* La valeur de r ne dépend pas des unités choisies pour mesurer les x_i et y_i
  r est donc un coefficient.
* Compte tenu de la positivité de f : **r est compris entre -1 et 1**.
* Si $\mathbf{r}^2 = 1$, la variabilité des résidus est nulle : les points du nuage sont alors alignés.
* La variabilité des résidus est d'autant plus faible que r² est proche de 1 (à σ_y constant)

La valeur de r² permet d'apprécier la qualité de l'ajustement.

* Le signe de r est celui de la pente de la droite : **r est du signe de $\hat{\alpha}$**

r est appelé coefficient de corrélation linéaire de la série (x_i, y_i).

C. Décomposition de la variabilité totale.

* Nous désignerons par variabilité résiduelle la somme des carrés des résidus notée $\mathbf{SC_{Res}}$

$$\mathbf{SC_{Res}} = \sum_{i=1}^{5} \mathbf{e_i^2} = \sum_{i=1}^{5} \left(\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}^* \right)^2$$

* Nous désignerons par variabilité totale, celle des (y_i)

$$\mathbf{SC_{Tot}} = \sum_{i=1}^{5} \left(\mathbf{y_i} - \bar{\mathbf{y}}\right)^2$$

* Nous désignerons par variabilité expliquée la quantité notée $\mathbf{SC_{Exp}}$ définie par

$$\mathbf{SC_{Exp}} = \mathbf{SC_{Tot}} - \mathbf{SC_{Res}}$$

Compte tenu de la formule n°1

$$\mathbf{SC_{Exp}} = \mathbf{r}^2 \sum_{i=1}^{5} \left(\mathbf{y_i} - \bar{\mathbf{y}}\right)^2$$

Ainsi

$$\boxed{\mathbf{r}^2 = \frac{\mathbf{SC_{Exp}}}{\mathbf{SC_{Tot}}} = \frac{\textbf{Variabilité Expliquée}}{\textbf{Variabilité Totale}}}$$

Dans la mesure où $\mathbf{r}^2$ mesure la part de variabilité expliquée par l'ajustement

$\mathbf{r}^2$ est appelé **coefficient de détermination**

* Autre expression de $\mathbf{SC_{Exp}}$:

$$\mathbf{SC_{Exp}} = 5 \quad \mathbf{r}^2 \quad \sigma_y^2 \quad = 5 \quad \hat{\alpha}^2 \quad \sigma_x^2$$

Comme $\hat{y}_i^* = \hat{\alpha} \quad x_i + \hat{\beta}$, $\hat{\alpha}^2 \, \sigma_x^2$ est la variance de la série statistique $(\hat{y}_i^*)$

N'oublions pas d'autre part, que la moyenne des (y_i) et celle des $(\hat{y}_i^*)$ sont égales:

Il en résulte $$\boxed{SC_{Exp} = \sum_{i=1}^{5} \left(\hat{y}_i^{*} - \bar{y} \right)^2}$$

* Formule de décomposition de la variabilité

$$\sum_{i=1}^{5}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{5}\left(\hat{y}^*_i - \bar{y}\right)^2 + \sum_{i=1}^{5}\left(y_i - \hat{y}^*_i\right)^2$$

$$SC_{Tot} = SC_{Exp} + SC_{Res}$$

C. analyse des résultats obtenus pour les douze points.

$\hat{\alpha} = +0.389$

$\hat{\beta} = -4.14$

L'équation de la droite de régression de Y en x est

$$Y = +0.389\ x - 4.14$$

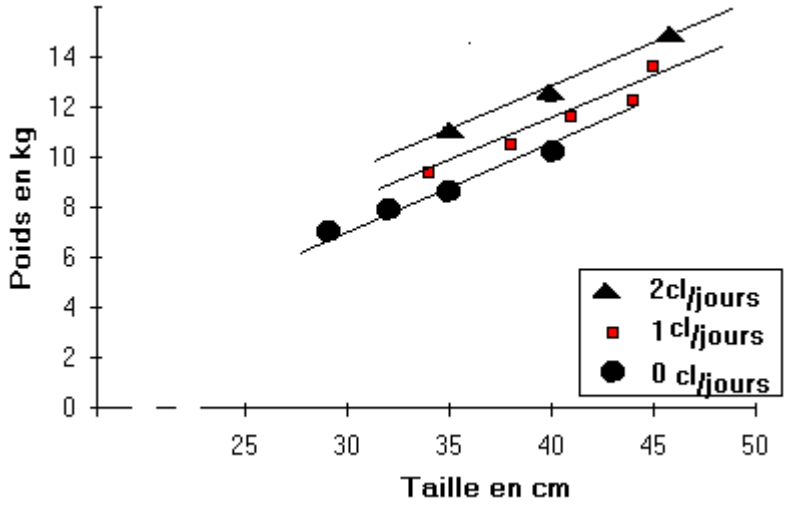
Variabilité expliquée:	47.839
Variabilité résiduelle:	6.528
Variabilité totale:	54.367

Le coefficient de détermination vaut : $r^2 = \frac{47.839}{54.367} = 0{,}88$

Ces résultats semblaient corrects !
Quelle ne fut pas notre surprise, lorsque quelques mois plus tard, furent retrouvés des renseignements égarés :

* à savoir les consommations journalières en eau de chaque individu.

Le nuage de points prenait alors une toute autre signification:



En un seul instant, notre modèle n°1 était anéanti :
Qu'allait-on faire, pour tenir compte de ces nouveaux renseignements?

- = - = - = - = - = - = - = - = -

LE COIN DU DEBUTANT

Nous vous proposons dans ce bulletin les corrigés de deux exercices de statistique proposés lors d'épreuves terminales dans des filières de BTSA.

SESSION 1997
France métropolitaine - Réunion - Mayotte
BTSA toutes options rénovées

EXERCICE **2 (** 7 **points)**
Une coopérative agricole commercialise des sacs de grains d'une masse nominale de 50 kg. En réalité la masse d'un sac exprimée en kg, est distribuée suivant la loi normale de moyenne μ = 50 et d'écart-type σ inconnu.
1°) De nombreuses mesures ont montré que 95% des sacs ont une masse comprise entre 48 kg et 52 kg.
Déterminer σ.
2°) On constitue à la livraison des lots de 40 sacs. Chaque lot peut être considéré comme un échantillon de taille 40 tiré au hasard dans la production.
On désigne par $\overline{X}$ la variable aléatoire prenant pour valeur la masse moyenne d'un sac dans un lot de 40 sacs. Justifier que la distribution de $\overline{X}$ est la loi normale de moyenne 50 et d'écart-type 0, 161 à 10^{-3} près.
3°) a - Quelle est la probabilité que la masse moyenne d'un lot de 40 sacs soit inférieure à 49,5 kg ?
b - Quelle est la probabilité que cette masse moyenne soit comprise entre 49,6 Kg et 50,4 kg ?
4°) Le contrat liant le fournisseur et un client stipule que la masse moyenne d'un lot doit être au moins de 50 kg.
En supposant que l'écart-type σ reste constant, quelle devra être la nouvelle moyenne μ' des sacs au conditionnement pour remplir ce contrat avec une probabilité de 0,99 ?

Remarque : d'un point de vue statistique la première question est discutable. Mal interprétée par des étudiants, elle peut être prise, à tort évidemment, pour une méthode d'estimation des paramètres d'une loi normale et induire ainsi de fausses idées sur la théorie de l'estimation. En fait cette question prend prétexte de la loi normale pour proposer un problème algébrique classique de mise en équation débouchant sur la résolution d'une équation du premier degré.

Proposition de corrigé.

1°) Déterminer σ.
Notons X la variable aléatoire qui prend pour valeur la masse, exprimée en kg, d'un sac de grains tiré au hasard.
X est distribuée selon la loi normale $N(\mu,\sigma)$ de moyenne μ = 50 et d'écart-type inconnu σ.
Donc la variable aléatoire U, définie par $U = \frac{X - \mu}{\sigma}$ soit $U = \frac{X - 50}{\sigma}$, est de loi normale $N(0 ; 1)$,

$P(48 \le X \le 52) = 0{,}95$

$x_1 = 48$ donc $u_1 = \dfrac{48-50}{\sigma}$ soit $u_1 = -\dfrac{2}{\sigma}$

$x_2 = 52$ donc $u_2 = \dfrac{52-50}{\sigma}$ soit $u_2 = \dfrac{2}{\sigma}$

$P(48 \le X \le 52) = 0{,}95 \quad \Leftrightarrow \quad P(\dfrac{48-50}{\sigma} \le U \le \dfrac{52-50}{\sigma}) = 0{,}95$

$\Leftrightarrow \quad P(-\dfrac{2}{\sigma} \le U \le \dfrac{2}{\sigma}) = 0{,}95$

$\Leftrightarrow \quad \Phi(\dfrac{2}{\sigma}) - \Phi(-\dfrac{2}{\sigma}) = 0{,}95$

$\Leftrightarrow \quad 2 \times \Phi(\dfrac{2}{\sigma}) - 1 = 0{,}95$

$\Leftrightarrow \quad \Phi(\dfrac{2}{\sigma}) = 0{,}975$

La lecture de la table de la fonction de répartition de la loi normale centrée réduite fournit $\Phi(1{,}96) = 0{,}95$ d'où $\dfrac{2}{\sigma} = 1{,}96$ soit

$\sigma = 1{,}02$.

2°) Loi de probabilité de $\overline{X}$.

On utilise le théorème fondamental d'échantillonnage de la moyenne :

Soit $\overline{X}$ la moyenne empirique de l'échantillon aléatoire $(X_1, X_2, \ldots, X_n)$ de taille n de la variable aléatoire X :

si la variable aléatoire X est distribuée selon la loi normale $N(\mu, \sigma)$, alors la moyenne $\overline{X}$ est distribuée selon la loi normale $N(\mu, \dfrac{\sigma}{\sqrt{n}})$.

Dans notre exemple :

$\overline{X}$ est distribuée selon la loi $N(50\,;\dfrac{1.02}{\sqrt{40}})$ soit $N(50\,;0{,}161)$.

3°) a) Probabilité que la masse moyenne d'un lot soit inférieure à 49,5 kg ?

D'après la question 2, la variable aléatoire U, définie par $U = \dfrac{\overline{X} - 50}{0{,}161}$, est de loi normale $N(0\,;1)$.

si $\overline{x} = 49{,}5$ alors $u = \dfrac{49{,}5 - 50}{0{,}161}$ soit $u = -3{,}11$

$P(\overline{X} \le 49{,}5) = P(U \le -3{,}11)$

$P(\overline{X} \le 49{,}5) = \Phi(-3{,}11)$

$P(\overline{X} \leq 49{,}5) = 1 - \Phi(3{,}11)$

$P(\overline{X} \leq 49{,}5) = 1\text{-}0{,}9990$

$P(\overline{X} \leq 49{,}5) = 0{,}0010$

La probabilité que la masse moyenne d'un lot soit inférieure à 49,5 est égale à 0,0010.

b) Probabilité que la masse moyenne soit comprise entre 49,6 Kg et 50,4 kg ?

si $\overline{x} = 49{,}6$ alors $u = \dfrac{49{,}6 - 50}{0{,}161}$ soit $u = -2{,}48$ à 10^{-2} près

si $\overline{x} = 50{,}4$ alors $u = \dfrac{50{,}4 - 50}{0{,}161}$ soit $u = 2{,}48$ à 10^{-2} près

$P(49{,}6 \leq \overline{X} \leq 50{,}4) = \Phi(2{,}48) - \Phi(-2{,}48)$

$P(49{,}6 \leq \overline{X} \leq 50{,}4) = 2 \times \Phi(2{,}48) - 1$

$P(49{,}6 \leq \overline{X} \leq 50{,}4) = 2 \times 0{,}9934 - 1$

$P(49{,}6 \leq \overline{X} \leq 50{,}4) = 0{,}9868$

La probabilité que la masse moyenne soit comprise entre 49,6 Kg et 50,4 Kg est égale à 0,9868

4°) Quelle devra être la nouvelle moyenne μ' des sacs au conditionnement ?

$\overline{X}$ est distribuée selon la loi $N(\mu' ; 0{,}161)$ donc la variable aléatoire *U*, définie par $U = \dfrac{\overline{X} - \mu'}{0{,}161}$, est de loi normale $N(0 ; 1)$,

$P(\overline{X} \geq 50) = 0{,}99$, l'aire sous la courbe normale après 50 est égale à 0,99 donc $\mu' \geq 50$,

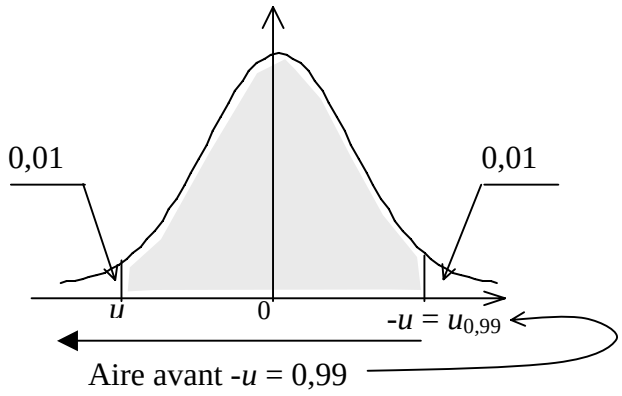
à $\overline{x} = 50$ correspond $u = \dfrac{50 - \mu'}{0{,}161}$, $u < 0$

$P(\overline{X} \geq 50) = 0{,}99 \Leftrightarrow P(\overline{X} \leq 50) = 0{,}01$

$\Leftrightarrow P(U \leq u) = 0{,}01$

$\Leftrightarrow \Phi(u) = 0{,}01$

$\Leftrightarrow \Phi(-u) = 0{,}99$



- u est le fractile $u_{0,99}$ de la loi normale centrée réduite.
On lit dans la table de la fonction de répartition de la loi normale centrée réduite : $u_{0,99} = 2,33$
$-u = 2,33$ soit $u = -2,33$
or $u = \dfrac{50 - \mu'}{0,161}$ donc $\dfrac{50 - \mu'}{0,161} = -2,33$
$\mu' = 50 + 2,33 \times 0,161$
$\mu' = 50,38$, à 10^{-2} près.

La nouvelle moyenne μ' des sacs au conditionnement est égale à 50,38

SESSION 1998
France métropolitaine - Réunion - Mayotte
BTSA toutes options rénovées

Exercice 2 (7 points)

Une entreprise commercialise des boîtes de lait dont la contenance nominale inscrite sur l'emballage est de 1 litre. On suppose que le volume de lait, exprimé en litres, versé dans une boîte est une variable aléatoire X distribuée suivant une loi normale de moyenne μ et d'écart-type donné égal à 0,006. La moyenne est obtenue par réglage de la machine sur la valeur "μ".

l) On règle la machine sur la valeur $\mu = 1,01$.

a) Quel est le pourcentage de boîtes dont le volume de lait sera inférieur à un litre ?

b) Pour vérifier que la machine est bien réglée, on prélève au hasard, toutes les heures, un échantillon de 9 boîtes. On calcule la moyenne $\bar{x}$ des volumes de lait contenu dans les 9 boîtes de cet échantillon.

On désigne par $\overline{X}$ la variable aléatoire qui, à chaque échantillon, associe sa moyenne $\bar{x}$.

Quelle est la loi de probabilité de $\overline{X}$?

Donner l'espérance mathématique et l'écart-type de $\overline{X}$.

Quelle est la probabilité que la moyenne sur un échantillon de 9 boîtes soit inférieure à 1,005 litres ?

2) Les volumes, exprimés en litres, de 9 boîtes d'un échantillon prélevé au hasard sont les suivants : 0,998 1,012 1,005 0,995 1,014 1,007 1,006 1,000 1,008.

Donner une estimation de la moyenne μ par intervalle de confiance au niveau 0,95 en justifiant la démarche et les résultats.

Remarque : La première question de l'énoncé nourrit la confusion entre pourcentage et probabilité. Il aurait été préférable de libeller la question sous la forme suivante : "Quelle est la probabilité qu'une boîte prélevée au hasard ait un volume de lait inférieur à un litre ?".

Proposition de corrigé.

1) a) On règle la machine sur la valeur μ= 1,01, X est distribuée selon la loi normale $N(\mu\,;\,\sigma)$ où μ=1,01 et σ=0,006.
X est distribuée selon la loi normale $N(1{,}01\,;\,0{,}006)$ donc la variable aléatoire *U*, définie par $U = \frac{X-\mu}{\sigma}$ soit $U = \frac{X-1{,}01}{0{,}006}$, est de loi normale $N(0\,;\,1)$

Calcul de la probabilité $P(X \leq 1)$ qu'une boîte ait un volume inférieure à un litre.

si $x = 1$ alors $u = \frac{1-1{,}01}{0{,}006}$ soit $u = -1{,}67$ à 10^{-2} près

$P(X \leq 1) = P(U \leq -1{,}67)$
$P(X \leq 1) = \Phi(-1{,}67)$

$P(X \leq 1) = 1 - \Phi(1{,}67)$

$P(X \leq 1) = 1-0{,}9525$

$P(X \leq 1) = 0{,}0475$

La probabilité qu'une boîte ait un volume inférieur à un litre est égale à 0,0475

b) Loi de probabilité de $\overline{X}$.

On utilise le théorème fondamental d'échantillonnage de la moyenne.

Soit $\overline{X}$ la moyenne empirique de l'échantillon aléatoire $(X_1,X_2,\ldots,X_n)$ de taille *n* de la variable aléatoire *X* :
si la variable aléatoire *X* est distribuée selon la loi normale $N(\mu,\sigma)$, alors la moyenne $\overline{X}$ est distribuée selon la loi normale $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Dans notre exemple :

$\overline{X}$ est distribuée selon la loi $N(1{,}01\,;\,\frac{0{,}006}{\sqrt{9}})$ soit $N(1{,}01\,;\,0{,}002)$.

Espérance mathématique et écart-type de $\overline{X}$: $E(\overline{X}) = 1{,}01$ $\sigma(\overline{X}) = 0{,}002$

Calcul de la probabilité $P(\overline{X} \leq 1{,}005)$ que la moyenne sur un échantillon de 9 boîtes soit inférieure à 1,005 litres :

D'après ce qui précède, la variable aléatoire *U*, définie par $U = \frac{\overline{X}-\mu}{\sigma(\overline{X})}$ soit $U = \frac{\overline{X}-1{,}01}{0{,}002}$,

est de loi normale $N(0\,;\,1)$.

Si $\overline{x} = 1{,}05$ alors $u = \frac{1{,}005-1{,}01}{0{,}002}$ soit $u = -2{,}5$

$P(\overline{X} \leq 1{,}005) = P(U \leq -2{,}5)$
$P(\overline{X} \leq 1{,}005) = \Phi(-2{,}5)$
$P(\overline{X} \leq 1{,}005) = 1 - \Phi(2{,}5)$
$P(\overline{X} \leq 1{,}005) = 1-0{,}9938$

$P(\overline{X} \leq 1{,}005) = 0{,}0062$

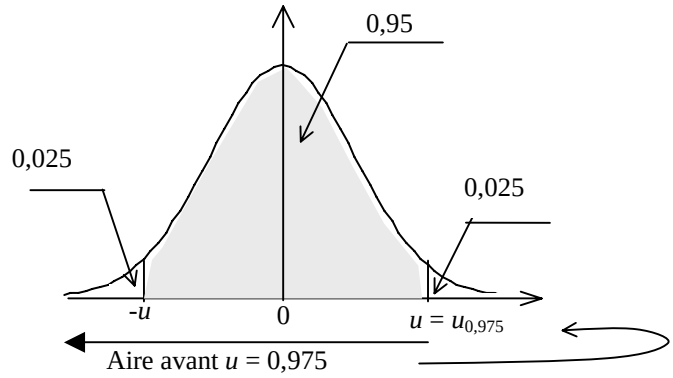
La probabilité que la moyenne sur un échantillon de 9 boîtes soit inférieure à 1,005 litres est égale à 0,0062

2) Intervalle de confiance à 95% de la moyenne.

La variance σ^2 est connue donc $\overline{X}$ est distribuée selon la loi normale $N(\mu, \frac{\sigma}{\sqrt{n}})$, et la variable aléatoire *U*, définie par $\frac{(\overline{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$, est distribuée selon la loi normale *N*(0 ; 1).

Détermination de l'intervalle de confiance aléatoire au niveau 95% :

$P(-u \leq U \leq u) = 0{,}95$ pour u = 1,96



u est le fractile $u_{0,975}$ de la loi normale centrée réduite.
On lit dans la table de la fonction de répartition de la loi normale centrée réduite : $u_{0,975} = 1{,}96$

$P(-1{,}96 \leq U \leq 1{,}96) = 0{,}95 \quad \Leftrightarrow \quad P(1{,}96\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$

$\Leftrightarrow \quad P(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}) = 0{,}95$

D'où l'intervalle de confiance aléatoire à 95% :

$[\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} \ ; \ \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}]$ soit $[\overline{X} - 1{,}96 \times 0{,}002 \ ; \ \overline{X} + 1{,}96 \times 0{,}002]$

$[\overline{X} - 0{,}004 \ ; \ \overline{X} + 0{,}004]$

Détermination d'un intervalle de confiance à 95% :

L'échantillon prélevé a pour moyenne $\overline{x} = 1{,}005$

$\overline{x} = 1{,}005$ est la valeur observée de la variable aléatoire $\overline{X}$ sur l'échantillon d'où l'intervalle de confiance correspondant à l'échantillon prélevé : [1,001 ; 1,009]

- = - = - = - = - = - = - = - = -

AUTOUR D'UN TEST DU KHI 2

L'objet de cet article est, au travers d'un exercice somme toute banal, de se poser quelques questions sur nos pratiques et aussi d'apporter quelques réponses.

Voici un texte d'exercice extrait d'un manuel [1] de DEUG, mais qui est tout à fait dans le cadre du programme de certaines filières de BTSA.

Un caractère A est présent ou non pour chaque individu d'une certaine population.
Une hypothèse est émise : la proportion "théorique" des porteurs du caractère A est de 75%.

> *1°) On examine un échantillon de 80 individus. On y trouve 50 porteurs du caractère et 30 non porteurs.*
> ***En utilisant un test du Khi2**, précisez si dans ces conditions, l'hypothèse émise doit être rejetée ou non au risque 5%.*
>
> *2°) Soit un échantillon de 80 individus, on désigne par n le nombre d'individus porteurs du caractère A.*
> *Entre quelles valeurs doit être compris le nombre n pour que l'hypothèse ne soit pas rejetée au risque 5%.*

Une solution rapide et mal rédigée ; mais là n'est pas la question :
Dans la première question, la mise en œuvre du test du Khi2 conduit à décider, au seuil de 5%, le rejet de l'hypothèse (les effectifs théoriques sont respectivement 60 et 20, le Khi2 calculé est de 6,67 et la valeur lue dans la table pour le risque donné est de 3,84).
Dans la deuxième question, les effectifs théoriques sont respectivement n et 80-n, le Khi2 calculé est égal à $\frac{(n-60)^2}{15}$. Il s'agit donc de résoudre l'inéquation $\frac{(n-60)^2}{15} < 3{,}84$ et on trouve que n doit être compris entre 53 inclus et 67 inclus.

Pourquoi ne pas ajouter la question suivante :

> 3°) Compte tenu des moyens de calculs actuels, pouvait-on envisager une autre méthode pour répondre à la question 2 ?

La population est supposée infinie (ou l'échantillon prélevé est supposé simple et aléatoire).Soit X la variable aléatoire qui à chaque échantillon de taille 80 associe le nombre de porteurs du caractère A. Sous l'hypothèse que la proportion de porteurs du caractère A dans la population est de 0,75, a loi de probabilité de X est la loi binomiale de paramètres n = 80 et p = 0,75.
Cette distribution ne pose aucun problème de calcul avec un tableur ou avec une calculatrice un tant soit peu récente.
En répartissant le risque "de façon symétrique en probabilité", le problème consiste donc chercher le plus grand entier a et le plus petit entier b tels que : $P(X \leq a) \leq 0{,}025$ et $P(X \geq b) \leq 0{,}025$.

Ces deux résolutions ne posent aucun problème puisque l'on dispose de l'ensemble des valeurs possibles des probabilités.
On trouve alors a = 51 et b = 68.

Deux questions se posent alors :

Pourquoi ne trouve -t-on pas le même résultat ?

Pouvait-on résoudre la première question avec cette méthode ?

La réponse à la deuxième question, pour un étudiant, est non parce que le texte disait "**En utilisant un test du Khi2**". Par contre, si le texte n'avait pas précisé le type de test à utiliser, alors on aurait très bien pu utiliser cette méthode.
Pour la première question, la réponse est plus facile d'un point de vue théorique, mais plus difficile d'un point de vue pédagogique.
En effet notons D^2 la variable aléatoire de décision utilisée lors de la résolution de la première question, cette variable s'écrit :

$D^2 = \frac{\left(n_1 - 80\times 0{,}75\right)^2}{80\times 0{,}75} + \frac{\left(n_2 - 80\times 0{,}25\right)^2}{80\times 0{,}25}$ où n_1 et n_2 désignent respectivement le nombre de porteurs du caractère A et le nombre de non porteurs du caractère.
La notation D^2 est donnée afin d'insister sur l'idée que ce que l'on calcule, c'est un "écart" entre les valeurs d'un tableau et les valeurs correspondantes du tableau que l'on devrait obtenir si l'hypothèse est vraie ; en fait une somme de carrés d'écarts. La variable D^2 est aussi notée K (voir bulletin du GRES N° 6 page 40).
Du fait que n_1 et n_2 sont des entiers positifs de somme égale à 80, la variable D^2 est une variable discrète, (si vous faites les calculs, D^2 ne prend que 61 valeurs distinctes) elle n'est donc sûrement pas distribuée selon une loi du χ^2 à 1 degré de liberté dans le cas de cet exercice. Cette affirmation reste vraie dans le cas où l'on aurait k modalités au lieu de 2.
En résolvant la première et la deuxième question, on a donc utilisé une approximation de la loi de D^2 et ce, sans aucune connaissance (et donc aucune maîtrise) sur cette approximation. D'ailleurs, les manuels eux-mêmes restent très évasifs sur cette question.
Lors de la résolution de la deuxième question, on a montré que $D^2 = \frac{\left(n - 60\right)^2}{15}$, on peut remarquer que $\frac{\left(n - 60\right)^2}{15} = \left(\frac{n - 60}{\sqrt{15}}\right)^2$.
La variable X définie précédemment a pour loi la loi binomiale de paramètres n = 80 et p = 0,75 et donc a pour moyenne np = 60 et pour écart-type $\sigma = \sqrt{15}$. D^2 est donc le carré d'une variable aléatoire de loi normale centrée réduite, là encore il s'agit d'une approximation. Au passage, en posant $U = \frac{\left(n - 60\right)}{\sqrt{15}}$, on voit que l'on aurait très bien pu utiliser la loi normale (centrée réduite) pour faire ce test.

On peut remarquer aussi que si le nombre de porteurs du caractère A est connu, alors le nombre de non porteurs est lui aussi connu. Par suite le test pourrait très bien porter seulement sur le nombre de porteurs du caractère A.
Dans ce cas, la loi binomiale est parfaitement adaptée et de plus cette loi est connue des étudiants (enfin, mieux connue que les lois du Khi 2!).

[1] Mathématiques pour les sciences de la vie P. Trousset et J.F. Morin Mac Graw Hill 1991 (p. 290-291).
[2] Bulletin du GRES n°3 : Quand deux tests se rejoignent (p. 17-19).

STATISTIQUE ET GEOMETRIE

Dans cet article, nous allons essayer d'établir des liens entre les statistiques d'une part, et la géométrie d'autre part. En effet, on lit souvent que des statisticiens célèbres ont montré (ou ont eu l'intuition) de résultats de statistique par des considérations de géométrie ; de plus, certains d'entre nous doivent ou devront enseigner ce qu'il est convenu d'appeler "l'analyse des données" (modules D4* en BTSA).
En restant modeste, essayons de voir ce qu'il en est pour les paramètres les plus classiques (moyenne, variance...).
Certains résultats seront généralisés aux variables aléatoires réelles, ces résultats seront donnés en italiques.

1. QUELQUES NOTATIONS :

Pour la partie statistique :

On considère une variable statistique, notée X dans la suite. On suppose que l'on dispose de n observations (n valeurs) pour X, notées $x_1, x_2, ..., x_n$. On note respectivement $\overline{x}$ et s^2 la moyenne et la variance de la série $(x_1, x_2, ..., x_n)$.
En fait, ici nous ne travaillerons pas avec la variable X, mais seulement avec la série des n observations $(x_1, x_2, ..., x_n)$ de cette variable.

Pour la partie géométrie :

On considère l'espace vectoriel euclidien muni du produit scalaire usuel, c'est-à-dire si $\vec{u}(a_1, a_2, ..., a_n)$ et $\vec{v}(b_1, b_2, ..., b_n)$ sont deux vecteurs de cet espace vectoriel alors on a : $\vec{u} \cdot \vec{v} = \sum_{i=1}^{n} a_i b_i$ et $\left\|\vec{u}\right\|^2 = \sum_{i=1}^{n} a_i^2$.

On note $\vec{1}$ le vecteur de coordonnées (1, 1, ..., 1).
De même, on considère l'espace affine euclidien $\mathbb{R}^n$ associé à cet espace vectoriel dont l'origine est notée O.

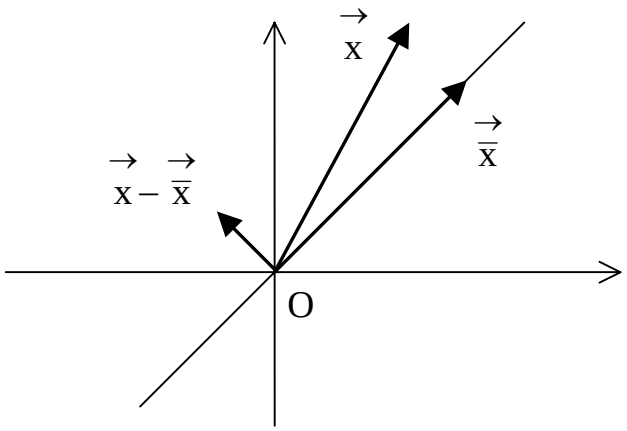
2. A PROPOS DE LA MOYENNE :

Considérons le vecteur $\vec{x}(x_1, x_2, ..., x_n)$, ce vecteur peut être considéré comme le vecteur « image » de la distribution statistique $(x_1, x_2, ..., x_n)$.

Par définition, le vecteur $\vec{\bar{x}}(\bar{x}, \bar{x}, ..., \bar{x})$ est le vecteur « moyenne ».

Par exemple, dans $\mathbb{R}^2$:

Considérons le vecteur $\vec{x}(2\,;4)$. On définit alors les vecteurs $\vec{\bar{x}}(3\,;3)$ et $\vec{x}-\vec{\bar{x}}(-1\,;1)$.



Résultat 1 :

$$\vec{x}\,.\,\vec{1}=\sum_{i=1}^{n}x_i \qquad \text{soit} \qquad \vec{x}\,.\,\vec{1}=n\bar{x}$$

L'application qui à tout vecteur $\vec{u}$ fait correspondre le réel $\vec{u}\,.\,\vec{1}$ est une application linéaire (en fait une forme linéaire).

L'opération, qui à tout n-uplet $(x_1, x_2, ..., x_n)$ associe sa moyenne $\bar{x}$, peut donc être considérée comme le produit scalaire $\vec{x}\,.\,\vec{1}$ (à la constante multiplicative $\frac{1}{n}$ près).

Cette opération a donc les mêmes propriétés que le produit scalaire. *On retrouve ainsi la propriété de « linéarité » de l'espérance mathématique, c'est-à-dire, en termes de variables aléatoires :*

Si a et b sont deux nombres réels alors E(aX + b) = aE(X) + b.

Les différentes moyennes pondérées peuvent aussi être vues sous la forme précédente (voir annexe 1).

On déduit aussi du résultat 1 que les vecteurs $\vec{x}$ et $\vec{1}$ sont orthogonaux si et seulement si $\bar{x}$ = 0.

Résultat 2 :

$$\left(\vec{x}-\vec{\bar{x}}\right).\vec{1}=\sum_{i=1}^{n}(x_i-\bar{x})$$

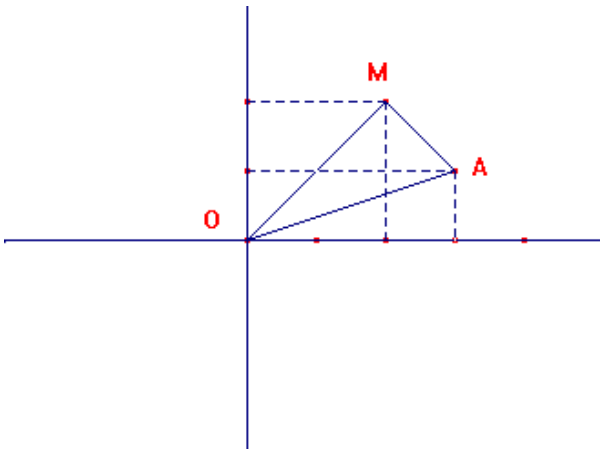
$\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-n\bar{x}=0$. On en déduit $\left(\vec{x}-\vec{\bar{x}}\right).\vec{1}=0$.

Les vecteurs $\vec{x}-\vec{\bar{x}}$ et $\vec{1}$ sont donc orthogonaux.

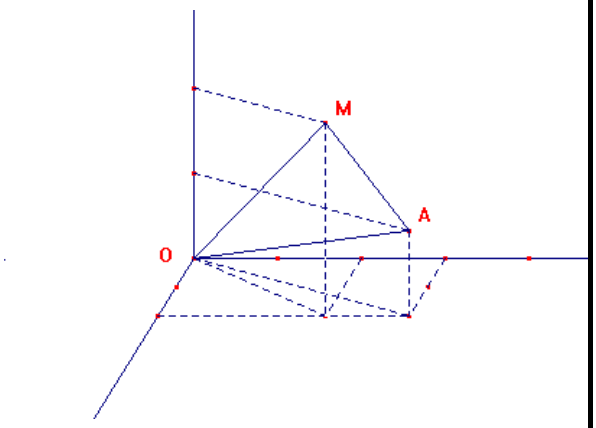
Par suite, les vecteurs $\vec{x}-\vec{\bar{x}}$ et $\vec{\bar{x}}$ sont orthogonaux (car $\vec{\bar{x}}=\bar{x}.\vec{1}$).

Le vecteur $\vec{\bar{x}}$ est le projeté orthogonal du vecteur $\vec{x}$ sur le sous-espace $\mathbb{R}\vec{1}$, c'est-à-dire la droite vectorielle engendrée par le vecteur $\vec{1}$.
En termes savants, c'est le théorème de la projection orthogonale.

Quelques représentations graphiques :



A est le point de coordonnées (3;1).
Le vecteur $\vec{x}$ est le vecteur $\overrightarrow{OA}$.
Le point M est le point de coordonnées (2;2).
Le vecteur $\vec{\bar{x}}$ est le vecteur $\overrightarrow{OM}$.
Le vecteur $\vec{x}$ - $\vec{\bar{x}}$ est le vecteur $\overrightarrow{MA}$.



A est le point de coordonnées (3;2;1).
Le vecteur $\vec{x}$ est le vecteur $\overrightarrow{OA}$.
Le point M est le point de coordonnées (2;2;2).
Le vecteur $\vec{\bar{x}}$ est le vecteur $\overrightarrow{OM}$.
Le vecteur $\vec{x}$ - $\vec{\bar{x}}$ est le vecteur $\overrightarrow{MA}$.

3. A PROPOS DE LA VARIANCE :

Résultat 3 :

$$\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$ c'est-à-dire $$\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = ns^2$$

La variance est donc égale au carré d'une norme divisé par n. Elle en possède donc les mêmes propriétés.

Par exemple, en termes de variables aléatoires :

Si a et b sont deux nombres réels alors $V(aX + b) = a^2V(X)$.

Au passage, on peut remarquer que $\left\| \vec{x} - \vec{\bar{x}} \right\|^2$ est la somme des carrés des écarts à la moyenne, elle est souvent notée SCE (notamment dans le cadre de l'analyse de la variance ou de la régression).

<u>Résultat 4</u> :

$$\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \vec{x}^2 - \vec{\bar{x}}.\vec{x}$$

En effet : $\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \left(\vec{x} - \vec{\bar{x}} \right).\left(\vec{x} - \vec{\bar{x}} \right)$

$\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \left(\vec{x} - \vec{\bar{x}} \right).\vec{x}$ car les vecteurs $\vec{x} - \vec{\bar{x}}$ et $\vec{\bar{x}}$ sont orthogonaux.

D'où, $\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \vec{x}^2 - \vec{\bar{x}}.\vec{x}$

Il en résulte : $s^2 = \frac{1}{n}\left(\vec{x}^2 - \vec{\bar{x}}.\vec{x} \right)$ c'est-à-dire $s^2 = \frac{1}{n}\left(\vec{x}^2 - \bar{x}.\vec{1}.\vec{x} \right)$

ou encore $s^2 = \frac{1}{n}\vec{x}^2 - \bar{x}^2$

On retrouve la formule de KŒNIG-HUYGENS, c'est-à-dire, *en termes de variables aléatoires :*

$$V(X) = E(X^2) - [E(X)]^2$$

En remarquant que l'égalité $\left\| \vec{x} - \vec{\bar{x}} \right\|^2 = \vec{x}^2 - \vec{\bar{x}}.\vec{x}$ peut s'écrire

$$\left\|\vec{x}-\vec{\bar{x}}\right\|^2=\left\|\vec{x}\right\|^2-\left\|\vec{\bar{x}}\right\|^2 \quad \text{ou encore} \quad \left\|\vec{x}\right\|^2=\left\|\vec{\bar{x}}\right\|^2+\left\|\vec{x}-\vec{\bar{x}}\right\|^2$$

on constate que la formule de KŒNIG-HUYGENS n'est rien d'autre que le théorème de PYTHAGORE exprimé dans $\mathbb{R}^n$.

Plaçons nous dans l'espace affine $\mathbb{R}^n$ rapporté à un repère orthonormal $(O, \vec{e_1}, \vec{e_2}, ..., \vec{e_n})$.
Étant donné un point $M(x_1, x_2,...,x_n)$ de $\mathbb{R}^n$, on note, pour tout i de {1, 2, ..., n}, $A_i(x_i, x_i, \ldots,x_i)$ le projeté orthogonal de M sur le sous-espace $\mathbb{R}\vec{e_i}$.
L'isobarycentre G des points A_i a pour coordonnées $(\bar{x}, \bar{x}, \ldots, \bar{x})$.

En appliquant la formule de LEIBNIZ (voir annexe 2),

$$\sum_{i=1}^{n}\alpha_i MA_i^2=\sum_{i=1}^{n}\alpha_i MG^2+\sum_{i=1}^{n}\alpha_i GA_i^2$$

avec M = O et pour tout i de {1, 2, ..., n} $\alpha_i=\alpha$ où α est un nombre réel non nul, on obtient :

$\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}\bar{x}^2+\sum_{i=1}^{n}(\bar{x}-x_i)^2$, c'est-à-dire $\left\|\vec{x}\right\|^2=\left\|\vec{\bar{x}}\right\|^2+\left\|\vec{x}-\vec{\bar{x}}\right\|^2$.

On retrouve donc la formule de KŒNIG-HUYGENS.

Ce résultat reste vrai pour une moyenne pondérée.

4. <u>A PROPOS DE LA SOMME DES CARRÉS DES ÉCARTS</u> :

La formule de LEIBNIZ permet de retrouver le résultat de statistique suivant :

La moyenne $\bar{x}$ est l'unique réel qui minimise la somme des carrés des écarts à un réel *a*.
Soit, exprimé autrement, $\sum_{i=1}^{n}(x_i-\bar{x})^2 = \inf_{a\in\mathbb{R}}\left(\sum_{i=1}^{n}(x_i-a)^2\right)$.

En effet, dans le cas général :

$\sum_{i=1}^{n} \alpha_i MA_i^2$ est minimum si et seulement si $\sum_{i=1}^{n} \alpha_i MG^2 = 0$

Les α_i sont tous strictement positifs, donc

$\sum_{i=1}^{n} \alpha_i MA_i^2$ est minimum si et seulement si M = G.

Remarquons aussi que si on appelle inertie du système de points pondérés (A_i, α_i) par rapport à un point M de $\mathbb{R}^n$, le nombre réel, noté I(M), défini par I(M) = $\sum_{i=1}^{n} \alpha_i MA_i^2$ alors l'inertie est minimum si et seulement si M = G.
Soit en termes de physiciens, « c'est par rapport au centre de gravité du système que les masses ont le plus petit moment d'inertie ».

Pour vous persuader de l'intérêt de la géométrie en statistique, vous trouverez en annexe 3 une preuve du résultat suivant :

Pour une série statistique donnée (x_i), l'écart-type est supérieur ou égal à l'écart absolu moyen : $\sigma \geq \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$.

5. <u>UNE PREMIÈRE CONCLUSION</u> : (en attendant une éventuelle suite)

Suite à cette modeste introduction, on peut se poser quelques questions pour aller plus loin, en voici trois.

a) Dans ce qui précède, nous avons considéré un paramètre de tendance centrale, la moyenne, et un paramètre de dispersion, l'écart-type.
L'idée géométrique sous-jacente étant : trouver le nombre réel *a* tel que la distance euclidienne classique entre le point $M(x_1, x_2, ..., x_n)$ et le point A(*a*, *a*, ..., *a*) soit minimale.

Plus généralement, on peut se poser la question suivante :

> Comment résumer une série statistique par un paramètre de tendance centrale et un paramètre de dispersion ?

Cette question peut aussi s'énoncer sous la forme :

Soit d une distance associée à $\mathbb{R}^n$, déterminer le réel *a* tel que la distance entre le point $M(x_1, x_2, ..., x_n)$ et le point A(*a*, *a*, ..., *a*) soit minimale.

Et dans ce cas, la valeur *a* ainsi déterminée est le paramètre de tendance centrale associé à la distance et d(A,M) caractérise la dispersion.

Par exemple, pour la distance euclidienne classique, on obtient la moyenne et l'écart-type (l'écart-type est égal à d(A,M) divisé par $\sqrt{n}$).

On voit bien que si l'on prend une autre distance, on aura a priori d'autres paramètres ; est ce que le réel *a* sera unique ?
Pour une série statistique donnée, quelle(s) distance(s) choisir ? Pourquoi choisir telle distance plutôt qu'une autre ?

b) Pour une distance donnée associée à $\mathbb{R}^n$, l'idée de minimiser ou maximiser l'inertie (par rapport à ...) est une des idées directrices de l'analyse des données.

c) Enfin,

Qu'en est-il dans le cas de deux ou de plusieurs variables ?
Comment interpréter géométriquement la covariance ?
En quoi le calcul matriciel permet de poser les problèmes sous une autre forme et de simplifier des preuves ?
Faut-il travailler dans l'espace des variables ou dans l'espace des individus ?
...

Bref, il reste encore beaucoup de questions à élucider ...

<u>ANNEXES</u> :

1°) Soit $(\alpha_1, \alpha_2, ..., \alpha_n)$ n réels <u>positifs</u>, on appelle moyenne pondérée de la série $(x_1, x_2, ..., x_n)$ le nombre réel $\frac{1}{\sum \alpha_i} \sum_{i=1}^{n} \alpha_i x_i$.

Les résultats énoncés précédemment pour la moyenne arithmétique restent vrais, il suffit de considérer (à la place du vecteur $\vec{1}$) le vecteur $\vec{p}(\alpha_1,\alpha_2,...,\alpha_n)$, (p comme poids).

2°) Soit $(A_i ; \alpha_i)$, $i \in \{1, 2, ..., n\}$ un système de points pondérés, on appelle fonction scalaire de LEIBNIZ la fonction qui à tout point M de l'espace associe le nombre réel $\sum_{i=1}^{n} \alpha_i MA_i^2$. Le résultat que nous utilisons est le suivant :

Si G désigne le barycentre de ce système de points, alors on a pour tout point M de l'espace, on a $\sum_{i=1}^{n} \alpha_i MA_i^2 = \sum_{i=1}^{n} \alpha_i MG^2 + \sum_{i=1}^{n} \alpha_i GA_i^2$.

3°) L'inégalité de CAUCHY-SCHWARZ s'écrit $|\vec{u}.\vec{v}| \leq \|\vec{u}\| \times \|\vec{v}\|$ soit encore

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leq \sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2} .$$

D'où, en prenant $a_i = \frac{1}{\sqrt{n}} |x_i|$ et $b_i = \frac{1}{\sqrt{n}}$, on a $\frac{1}{n} \sum_{i=1}^{n} |x_i| \leq \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$ (1)

Considérons une série statistique $(x_1, x_2, ..., x_n)$. Quitte à faire un changement de variable, on peut supposer que la moyenne est nulle. L'inégalité (1) signifie que l'écart absolu moyen est inférieur ou égal à l'écart-type.

- = - = - = - = - = - = - = -

EXCEL ET LE TEST DU χ^2

Parmi les nombreuses fonctions statistiques implantées dans EXCEL figure une fonction appelée TEST.KHIDEUX. Cette fonction a-t-elle un rapport avec les tests de Khi deux qui figurent au programme de certaines filières BTSA ? C'est ce que nous allons découvrir dans cet article.
Tout d'abord lisons attentivement les informations fournies dans l'aide en ligne :

> " *Renvoie le test d'indépendance. TEST.KHIDEUX renvoie la valeur de la distribution khi-deux (χ2) pour la statistique et les degrés de liberté appropriés. Utilisez les tests χ2 pour déterminer si les résultats prévus sont vérifiés par une expérimentation.*"

L'aide indique ensuite la syntaxe à utiliser :

> "*TEST.KHIDEUX(plage_réelle;plage_attendue)*
> *plage_réelle représente la plage de données contenant les observations à comparer aux valeurs prévues.*
> *plage_attendue représente la plage de données contenant le rapport du produit des totaux de ligne et de colonne avec le total général.*"

Une remarque donne quelques renseignements supplémentaires :

> "*Le test χ2 calcule d'abord une statistique χ2 puis additionne les différences entre les valeurs réelles et les valeurs prévues. L'équation de cette fonction est TEST.KHIDEUX=p(X>χ2), où :*
>
> $$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{\left(A_{ij}-E_{ij}\right)^2}{E_{ij}}$$
>
> *et où :*
> *Aij est la fréquence réelle dans la i-ème ligne et la j-ème colonne.*
> *Eij est la fréquence prévue dans la i-ème ligne et la j-ème colonne.*
> *l est le nombre de lignes.*
> *c est le nombre de colonnes.*
>
> *TEST.KHIDEUX renvoie la probabilité pour une statistique χ2 et des degrés de liberté, où df = (l - 1)(c - 1).*"

Voilà, avec ces explications on est censé être informé et comprendre le fonctionnement de cette fonction KHIDEUX ! ! !
N'oublions pas toutefois le petit exemple qui termine l'aide et qui sert à éclairer ces explications que certains pourraient trouver peu lumineuses :

"*Exemple*

	A	B	C
1	**Réel**		
2		**Hommes**	**Femmes**
3	**D'accord**	58	35
4	**Sans opinion**	11	25
5	**Pas d'accord**	10	23
6			
7	**Prévu**	**Hommes**	**Femmes**
8			
9	**D'accord**	45,35	47,65
10	**Sans opinion**	17,56	18,44
11	**Pas d'accord**	16,09	16,91

La statistique χ2 pour les données ci-dessus est de 16,16957 avec 2 degrés de liberté.
TEST.KHIDEUX(B3:C5,B9:C11) égale 0,000308."

Et bien il va falloir se débrouiller avec ça ! Expliquons donc ce qui précède et tâchons d'être plus clair que l'aide, ce qui a priori ne constitue pas une prouesse pédagogique.

On devine que cette fonction peut être utilisée lorsqu'on veut mettre en place un test d'indépendance.
Dans l'exemple fourni, il semble que l'on souhaite tester l'indépendance des caractères Sexe et Avis dans une population déterminée, à partir d'un échantillon de 162 personnes qui se répartissent, en fonction de leur avis et de leur sexe, comme indiqué dans le tableau intitulé **Réel** (qui constitue, en fait un "*tableau de contingence*").

Le tableau intitulé **Prévu** est le tableau dit des "*effectifs attendus*" ou "*effectifs théoriques*", il est construit sous l'hypothèse

H_0 : "les caractères Sexe et Avis sont stochastiquement indépendants".

L'aide nous rappelle que ce tableau s'obtient, à partir des effectifs marginaux et de l'effectif total du tableau Réel, en effectuant, pour tout i et pour tout j, le calcul : $\frac{n_{i.} * n_{.j}}{n}$ où $n_{i.} = \sum_k n_{ik}$ et $n_{.j} = \sum_k n_{kj}$ sont les effectifs marginaux et n l'effectif total.

En effet, ce tableau donne la répartition des 162 personnes (79 hommes, 83 femmes ; 93 D'accord, 36 Sans opinion et 33 Pas d'accord) lorsqu'on suppose que les caractères Sexe et Avis sont indépendants.
Ceci signifie que pour chaque ligne, c'est-à-dire pour chaque Avis, le pourcentage d'hommes est égal au pourcentage de femmes qui est donc égal au pourcentage de personnes de l'échantillon (sexes confondus) qui ont cet avis.
On doit donc avoir :
$\frac{a}{79} = \frac{d}{83} = \frac{93}{162}$ et la même chose pour les 2 autres lignes. On trouve

	A	B	C	D
1	**Réel**			
2		**Hommes**	**Femmes**	
3	**D'accord**	58	35	93
4	**Sans opinion**	11	25	36
5	**Pas d'accord**	10	23	33
6		79	83	162
7	**Prévu**			
8		**Hommes**	**Femmes**	
9	**D'accord**	a	d	93
10	**Sans opinion**	b	e	36
11	**Pas d'accord**	c	f	33
12		79	83	162
13				

donc : a = $\frac{79\times 93}{162}$, b = $\frac{83\times 93}{162}$ et ainsi de suite.

Avec EXCEL, la construction de ce tableau constitue un très bon exercice d'utilisation des références mixtes. Si, dans la colonne D, on inscrit en D3 la formule =SOMME(B3:C3) que l'on recopie jusqu'en D6 ; si, dans la ligne 6, on inscrit en B6 la formule =SOMME(B3:B5) que l'on recopie en C6 alors, il suffit d'inscrire en B9 la formule =$D3*B$6/$D$6 de la recopier en C9 puis de recopier cette plage jusqu'à la ligne 11 pour obtenir le tableau des effectifs théoriques.

On constate que ce tableau diffère du tableau réel, il s'agit de mesurer "l'écart" entre les deux tableaux pour savoir si cet écart est acceptable (dû aux fluctuations d'échantillonnage) ou bien si cet écart est trop important et nous conduit à refuser l'hypothèse d'indépendance des caractères.
L'instrument de mesure utilisé est la distance du χ^2, elle est obtenue en calculant la somme des carrés des différences des cellules homologues des deux tableaux divisés (ces carrés) par les valeurs des cellules homologues du tableau Prévu. C'est à dire la formule donnée par l'aide en considérant que A_{ij} représente les cellules du tableau Réel et E_{ij} celles du tableau Prévu, ce qui donne ici :

$$\frac{(58-45{,}35)^2}{45{,}35}+\frac{(35-47{,}65)^2}{47{,}65}+\frac{(11-17{,}56)^2}{17{,}56}+\frac{(25-18{,}44)^2}{18{,}44}+\frac{(10-16{,}09)^2}{16{,}09}+\frac{(23-16{,}91)^2}{16{,}91}$$

Si à chaque échantillon de taille 162, on associe le résultat de ce calcul, on obtient une variable aléatoire dont on admet qu'elle suit *approximativement* la loi de χ^2 à (3-1)*(2-1) degrés de liberté (3 et 2 sont les nombres de modalités des 2 caractères étudiés Avis et Sexe).

Effectuons automatiquement le calcul ci-dessus en utilisant la fonction SOMME.XMY2 et les calculs sur les matrices. Une seule formule suffit :
=SOMME.XMY2($B$3:$C$5/RACINE($B$9:$C$11);RACINE($B$9:$C$11)).
Cette formule renvoie la valeur 16,16957507.

Remarques : La fonction SOMME.XMY2, (*somme x moins y au carré*), calcule la somme des carrés des différences entre les éléments de même position de deux matrices (ou plages de cellules) de mêmes dimensions.

Les opérations sur les matrices s'effectuent sur chaque terme des matrices ainsi :
RACINE($B$9:$C$11) renvoie la matrice formée des racines carrées des éléments de la matrice $B$9:$C$11.
$B$3:$C$5/RACINE($B$9:$C$11) divise chaque élément de la plage $B$3:$C$5 par la racine carrée des éléments homologues de la plage $B$9:$C$11.
Si bien que la formule écrite réalise le calcul :

$$\sum\sum\left(\frac{A_{ij}}{\sqrt{E_{ij}}}-\sqrt{E_{ij}}\right)^2$$

, elle fournit donc la valeur du χ^2 calculée sur notre échantillon.

En utilisant la fonction LOI.KHIDEUX(x;degrés_liberté) on obtient la probabilité pour qu'une variable aléatoire suivant la loi de χ^2 à 2 d.d.l. prenne des valeurs supérieures ou égales à la valeur calculée ci-dessus, 16,16957507 :
=LOI.KHIDEUX(SOMME.XMY2($B$3:$C$5/RACINE($B$9:$C$11);RACINE($B$9:$C$11);2))
cette formule renvoie la valeur : 0,000308192.

C'est donc le résultat fourni par la fonction TEST.KHIDEUX !
Ceci signifie que, si les caractères Sexe et Avis sont indépendants, sur 10000 échantillons de 162 personnes il n'y en a que 3 qui donnent une valeur calculée du χ^2 aussi importante.
On peut donc conclure, au vu de cet échantillon, **à une différence de point de vue des hommes et des femmes** au risque de se tromper de 0,03%.

Remarque : *les programmeurs de la fonction TEST.KHIDEUX auraient pu programmer le calcul du tableau "Prévu" et éviter ainsi à l'utilisateur de faire ce calcul. Ceux qui connaissent un peu le VBA pourront reprogrammer la fonction KHIDEUX de telle sorte qu'elle n'ait plus besoin que d'un argument, le tableau Réel.*

En guise d'entraînement, on va corriger un exercice issu d'un sujet de BTSA, il s'agit du sujet Remplacement 1996 France Métropolitaine Options : Productions animales Formation hippique.

Exercice 3

Le tableau ci-dessous donne les résultats concernant la fertilité de trois étalons en insémination artificielle :

Etalons	Nombre de chaleurs fécondées	Nombre de chaleurs Non fécondées
Un atout	20	18
Arthy	20	16
Vainqueur	44	51

Peut-on conclure à une différence de fertilité par chaleur entre ces trois étalons, au seuil de 5%. On utilisera un test du KHI-2.

Eléments de correction en utilisant EXCEL :

Commençons par formaliser le problème. On peut imaginer une population théorique de juments inséminées artificiellement avec la semence de 3 étalons : Un atout, Arthy et Vainqueur. Sur cette population, on étudie les 2 caractères qualitatifs :

- résultat de l'insémination,
- identité du donneur.

Le premier caractère ne possède que 2 modalités : fécondée ou non fécondée (il s'agit de la jument),

Le deuxième caractère possède 3 modalités qui sont les noms des 3 étalons.
L'étude a pour but de conclure à la non indépendance ou à l'indépendance des deux caractères au vu des résultats constatés sur un échantillon aléatoire simple de 169 juments et fournis dans le tableau de contingence ci-dessus.

Pour cela on met en place un test de Khideux :

* Hypothèses :
  H_o : les 2 caractères sont indépendants
  H_1 : les 2 caractères ne sont pas indépendants.

* risque de première espèce : $\alpha = 0,05$.

* Variable aléatoire de décision : X^2 qui associe à chaque échantillon de taille 169 la distance entre le tableau de contingence et le tableau théorique (voir plus haut).
Sous l'hypothèse nulle, cette variable aléatoire suit approximativement la loi de χ^2 à (3-1)(2-1), c'est-à-dire 2, degrés de liberté.

* Règle de décision : Si la valeur du χ^2 calculée à partir de l'échantillon est supérieure ou égale à la valeur lue sur une table des lois de χ^2 pour $\alpha = 0,05$ et 2 d.d.l., alors on rejette l'hypothèse nulle et l'on conclut à la non indépendance des caractères (*ce qui signifie qu'il y a une différence de fertilité par chaleur entre ces trois étalons*). Si la valeur est inférieure il n'y a pas lieu de rejeter l'indépendance des 2 caractères (*ce qui signifie que l'on n'a pas pu mettre en évidence une différence de fertilité entre les trois étalons*).

Nous allons faire les calculs avec EXCEL. On calcule, grâce à la fonction TEST.KHIDEUX, la probabilité pour que, sous l'hypothèse d'indépendance, un échantillon de taille 169 donne un écart entre le tableau constaté et le tableau théorique supérieur à celui que nous obtenons avec notre échantillon.
Si cette probabilité est inférieure à 0,05 nous rejetons l'hypothèse d'indépendance sinon il n'y a pas lieu de la rejeter.
Voici quelle peut être la feuille de calcul :

	A	B	C	D
1	Etalons	Nombre de chaleurs fécondées	Nombre de chaleurs non fécondées	
2	Un atout	20	18	① 38
3	Arthy	20	16	36
4	Vainqueur	44	51	95
5		② 84	85	169
6				
7	Un atout	③ 18,887574	19,112426	
8	Arthy	17,8934911	18,1065089	
9	Vainqueur	47,2189349	47,7810651	
10				
11				
12		TEST.KHIDEUX : ④ 0,5887177		
13				

Formules : D2 : ① =SOMME(B2 :C2)

Recopiée vers le bas jusqu'en D5

B5 : ② =SOMME(B2 :B4)
Recopiée vers la droite jusqu'en C5

B7 : ③ =$D2*B$5/$D$5
Recopiée à droite jusqu'en C7 puis,
le tout recopié vers le bas jusqu'en ligne 9.

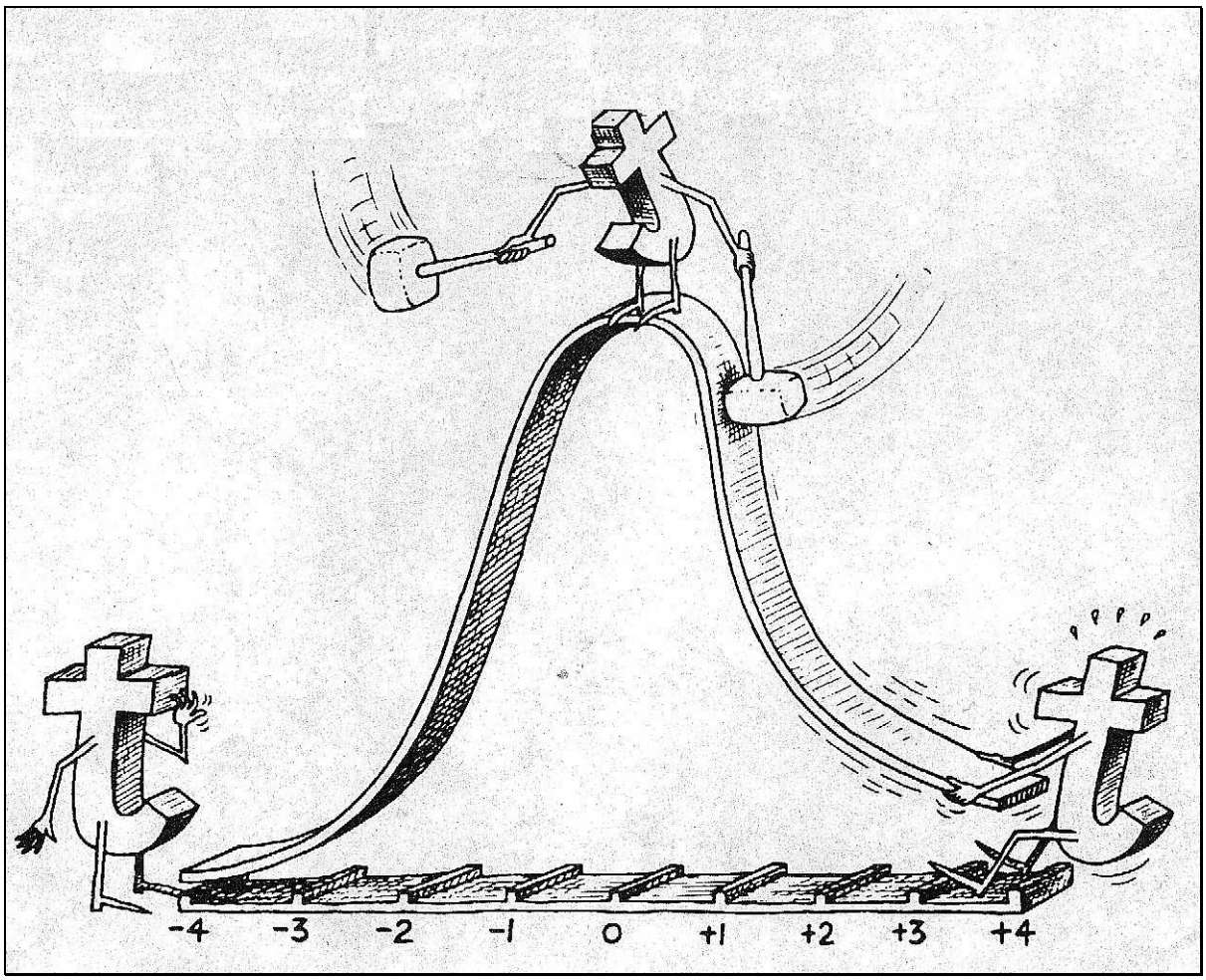
C12 : ④ =TEST.KHIDEUX($B$2:$C$4;$B$7:$C$9)

Notre conclusion : Puisque nous obtenons une probabilité nettement supérieure à 0,05 nous pouvons affirmer que :
au vu de cet échantillon, et au seuil de 5%, il n'y a pas lieu de conclure à une différence de fertilité par chaleur entre ces trois étalons.

Cette même fonction TEST.KHIDEUX peut être utilisée pour mettre en œuvre certains tests d'ajustement. Nous donnerons quelques exemples dans le prochain bulletin.

- = - = - = - = - = - = - = -

La Loi de STUDENT a encore frappé !



Normalisation (*à la soviétique*) d'une loi de Student

(*D'après un dessin illustrant les tables de l'ITCF*)

GEOMETRIE EUCLIDIENNE ET STATISTIQUE

Position du problème

On considère une variable aléatoire X définie sur une population. On suppose que X est de loi normale, $N(\mu,\sigma)$.

On prélève dans cette population un échantillon aléatoire et simple de taille n. Il en résulte n variables aléatoires, $X_1, X_2, ...X_n$ indépendantes et de même loi que X.

On note : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Nous allons démontrer que la variable aléatoire $\frac{nS^2}{\sigma^2}$ suit la loi de χ^2 à (n – 1) degrés de liberté.

Rappel

Par définition une variable aléatoire suit la loi de χ^2 à ν degrés de liberté si elle est la somme des carrés de ν variables aléatoires normales, centrées, réduites et indépendantes.

Principe

On va travailler dans l'espace euclidien $\mathbb{R}^n$ muni de son produit scalaire usuel.

On considère le vecteur aléatoire $V = \begin{pmatrix} \frac{X_1 - \overline{X}}{\sigma} \\ \frac{X_2 - \overline{X}}{\sigma} \\ . \\ . \\ \frac{X_n - \overline{X}}{\sigma} \end{pmatrix}$ et le vecteur $I = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}$

$$V.I = \frac{1}{\sigma}\left((X_1 - \overline{X}) + (X_2 - \overline{X}) + ... + (X_n - \overline{X})\right) = \frac{1}{\sigma}\left(X_1 + X_2 + ... + X_n - n\overline{X}\right) = 0$$

V est donc orthogonal à I. Il en résulte que V appartient à l'hyperplan (sous-espace vectoriel de dimension n-1) de $\mathbb{R}^n$ orthogonal à I.

D'autre part $\|V\|^2 = \frac{1}{\sigma^2}\left((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + ... + (X_n - \overline{X})^2\right) = \frac{nS^2}{\sigma^2}$

La démonstration consiste à utiliser une base orthonormale, $(e_i)_{1\le i\le n}$, de $\mathbb{R}^n$ telle que $e_1 = \frac{1}{\sqrt{n}} I$. On sait qu'une telle base existe et, dans une telle base, V a pour coordonnées $(0, X'_2, X'_3, ..., X'_n)$ et $\|V\|^2 = \frac{nS^2}{\sigma^2} = \sum_{i=2}^{n} X'^2_i$.

Il restera à démontrer que les $X'_i \ (2 \le i \le n)$ sont de loi N(0, 1) et indépendantes.

Quelques rappels d'algèbre linéaire

Pour obtenir une matrice de changement de base on place en <u>colonnes</u> les coordonnées, dans la première base, des vecteurs de la deuxième base. Cette matrice permet de calculer les coordonnées d'un vecteur dans la première base en fonction de celles de ce vecteur dans la deuxième base.

Ce qui nous intéresse ici est la réciproque : calculer les nouvelles coordonnées en fonction des anciennes. Nous avons donc besoin de la matrice inverse de la matrice de changement de base.

Lors du passage d'une base orthonormale à une autre base orthonormale, la matrice de passage est une matrice orthogonale, son inverse est alors égale à sa transposée. Nous utiliserons donc une matrice dont les <u>lignes</u> seront les coordonnées des vecteurs de la nouvelle base.

Exemple de matrice convenant ici :

Considérons la matrice suivante :

$$\begin{pmatrix} 1 & 1 & 1 & 1 & ... & ... & 1 & 1 \\ 1 & -1 & 0 & 0 & ... & ... & 0 & 0 \\ 1 & 1 & -2 & 0 & ... & ... & 0 & 0 \\ 1 & 1 & 1 & -3 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 1 & 1 & 1 & ... & ... & 1 & -n+2 & 0 \\ 1 & 1 & 1 & ... & ... & 1 & 1 & -n+1 \end{pmatrix}$$

Ses vecteurs-lignes sont 2 à 2 orthogonaux.. pour obtenir une matrice répondant à notre problème il suffit de multiplier chacun d'eux par l'inverse de sa norme.

Démonstration

- Soit une base orthonormale, $(e_i)_{1\le i\le n}$, de $\mathbb{R}^n$ telle que $e_1 = \frac{1}{\sqrt{n}} I$

Pour $1 \le i \le n$ et $1 \le j \le n$ on notera a_{ij} la $j^{ième}$ coordonnée de e_i.

Pour tout i, $1 \le i \le n$, on a : $\|e_i\| = 1$ d'où $\sum_{j=1}^{n} a_{ij}^2 = 1.$ (1)

D'autre part, pour $i \neq k$, $e_i.e_k = 0$ d'où : $\sum_{j=1}^{n} a_{ij}.a_{kj} = 0.$ (2)

En particulier, pour tout i, $2 \le i \le n$, $e_i.e_1 = 0$ d'où : $\sum_{j=1}^{n} a_{ij} = 0.$ (3)

- Avec les notations définies précédemment on obtient :

$$\begin{pmatrix} X_1' \\ X_2' \\ . \\ . \\ X_n' \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \dots & \dots & \frac{1}{\sqrt{n}} \\ a_{21} & a_{22} & \dots & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} \frac{X_1 - \overline{X}}{\sigma} \\ \frac{X_2 - \overline{X}}{\sigma} \\ . \\ . \\ \frac{X_n - \overline{X}}{\sigma} \end{pmatrix}$$

$X_1' = \frac{1}{\sqrt{n}} V.I = 0$, d'où, puisque la nouvelle base est aussi orthonormale, l'égalité :

$\|V\|^2 = \sum_{i=2}^{n} X_i'^2$ c'est-à-dire $\frac{nS^2}{\sigma^2} = \sum_{i=2}^{n} X_i'^2$.

- D'autre part pour $2 \le i \le n$ on a :

$X_i' = \frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}(X_j - \overline{X}) = \frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}X_j - \frac{\overline{X}}{\sigma}\sum_{j=1}^{n} a_{ij} = \frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}X_j$ d'après (3).

X_i' est donc de loi normale comme combinaison linéaire de variables aléatoires normales indépendantes. De plus :

$E(X_i') = \frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}E(X_j) = \frac{\mu}{\sigma}\sum_{j=1}^{n} a_{ij} = 0$ d'après (3).

et, les X_i étant indépendantes,

$V(X_i') = \frac{1}{\sigma^2}\sum_{j=1}^{n} a_{ij}^2 V(X_j) = \frac{\sigma^2}{\sigma^2}\sum_{j=1}^{n} a_{ij}^2 = 1$ d'après (1).

Il en résulte que, pour tout i, $2 \le i \le n$, X_i' est de loi normale centrée et réduite.

- Il reste à montrer que les $X_i^{'}$, pour $2 \le i \le n$, sont indépendantes.

Toute combinaison linéaire des $X_i^{'}$ est une combinaison linéaire des X_i et est donc de loi normale. La définition donnée en annexe permet donc de dire que le vecteur aléatoire $\left(X_2^{'}, X_3^{'}, ...X_n^{'}\right)$ est gaussien. Pour que les $X_i^{'}$ $\left(2 \le i \le n\right)$ soient indépendantes il suffit, d'après le théorème donné en annexe, qu'elles soient 2 à 2 non corrélées.

Calculons donc, pour h différent de k (h et k compris entre 2 et n), la covariance de $X_h^{'}$ et $X_k^{'}$:

$$\mathrm{Cov}(X_h^{'}, X_k^{'}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{hi}.a_{kj}\mathrm{Cov}(X_i, X_j)$$

Or, pour $i \neq j$, $\mathrm{Cov}(X_i, X_j) = 0$ puisque les X_i sont indépendantes et, pour tout i compris entre 1 et n, $\mathrm{Cov}(X_i, X_i) = V(X_i) = \sigma^2$ .Il en résulte :

$$\mathrm{Cov}(X_h^{'}, X_k^{'}) = \frac{1}{\sigma^2} \sum_{j=1}^{n} a_{hj}.a_{kj}.\sigma^2 = \sum_{j=1}^{n} a_{hj}a_{kj} = 0 \text{ d'après (2).}$$

Les $X_i^{'}$ $\left(2 \le i \le n\right)$ sont donc indépendantes.

Conclusion : La variable aléatoire $\frac{nS^2}{\sigma^2}$ est la somme des carrés de n - 1 variables aléatoires normales centrées réduites et indépendantes, elle est donc de loi de χ^2 à n - 1 degrés de liberté.

***Prolongements possibles* :**

- On peut également montrer que $\frac{nS^2}{\sigma^2}$ est indépendante de $\overline{X}$:

$$Cov(X_i^{'}, \overline{X}) = Cov(\frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}X_j, \frac{1}{n}\sum_{k=1}^{n} X_k) = \frac{1}{\sigma n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ij}Cov(X_j, X_k)$$

$$Cov(X_i^{'}, \overline{X}) = \frac{1}{\sigma n}\sum_{j=1}^{n} a_{ij}Cov(X_j, X_j) = \frac{\sigma^2}{\sigma n}\sum_{j=1}^{n} a_{ij} = 0 \text{ d'après (3).}$$

Il en résulte que, pour tout i compris entre 2 et n, $X_i^{'}$ et $\overline{X}$ sont non corrélées et donc indépendantes d'après le théorème cité précédemment. On en déduit que $\frac{nS^2}{\sigma^2}$ qui est la somme des carrés de ces variables est indépendante de $\overline{X}$.

- En analyse de variance à un facteur :

On démontre de la même manière que, lorsque les moyennes des p populations sont égales, la variable aléatoire $\frac{SCE_F}{\sigma^2}$ suit la loi de χ^2 à p - 1 ddl (le même type de matrice est utilisable avec n = p lorsque les échantillons sont de même taille sinon on doit avoir $a_{1j} = \sqrt{\frac{n_j}{n}}$, les n_j étant les effectifs des échantillons et n l'effectif total).

- La matrice donnée en exemple (ou toute autre construite de la même manière)

est utilisée pour les comparaisons multiples de moyennes par une méthode appelée méthode des contrastes lorsque les effectifs des échantillons sont égaux.

Mais tout cela sera pour une autre fois si certains en manifestent le souhait.

Annexe

Définition : Une variable aléatoire définie sur un espace de probabilité et à valeurs dans $\mathbb{R}^n$ est dite normale ou gaussienne si toute combinaison linéaire de ses coordonnées est de loi normale. (remarque : une variable aléatoire constante est considérée comme étant de loi normale d'écart type 0 ou dégénérée)

Théorème : Pour que la suite des coordonnées d'un vecteur normal soit indépendante, il faut et il suffit que la covariance de ce vecteur (matrice des covariances de ses coordonnées) soit une matrice diagonale.

- = - = - = - = - = - = - = -

COURRIER

Heureusement que la boîte à lettres du GRES n'a pas été plus remplie que d'ordinaire car le responsable du courrier n'avait que peu de temps à consacrer à la rubrique pour ce numéro 8.

Signalons toutefois 2 lettres :

La première nous est envoyée par notre collègue du *LEGTA de ST LO Thère*, *Bénédicte GIMER*, qui nous demande des informations, des documents, des exemples d'utilisation des **cartes de contrôle EWMA** (cartes de contrôle à moyennes mobiles avec pondération exponentielle). Elle dispose déjà de la norme AFNOR NFX-06-031-3 mais elle souhaiterait avoir des titres d'ouvrages ou de documents, des exemples d'utilisation etc…
A notre grande honte, nous n'avons pas encore pu lui fournir de renseignements intéressants sur ce type de carte de contrôle. Même des recherches sur Internet en utilisant plusieurs moteurs de recherche ne nous ont pas permis de lui donner satisfaction. Nous faisons donc appel aux nombreux lecteurs du bulletin de France et d'Amérique (*en effet en plus de nos collègues des Antilles qui reçoivent régulièrement le bulletin du GRES, nous sommes en relation avec l'Association Mathématique du Québec qui reçoit et apprécie -c'est la responsable qui nous l'a écrit- notre bulletin en échange du bulletin de l'AMQ que nous tenons à votre disposition*) : si vous avez des informations sur ces cartes de contrôle EWMA contactez-nous[*] ou contactez directement Bénédicte GIMER.

Je profite de cette occasion pour signaler à tous les professeurs de Mathématiques l'ouverture, dernièrement, d'une **conférence, sur la messagerie Educagri, à destination des professeurs de Mathématiques-Informatique**. Chaque enseignant de Mathématiques de l'enseignement agricole public peut demander (par l'intermédiaire du Délégué Régional Informatique) à se faire ouvrir une boîte à lettre électronique et à être inscrit à la conférence qui est en fait un forum où chacun peut poser des questions, répondre à des questions, exprimer ses états d'âme quant à la Mathématique, l'Informatique, leur enseignement …

Inscrivez-vous si ce n'est déjà fait !

La deuxième contribution épistolaire nous vient d'un soit disant lecteur qui souhaite rester âne à Nîmes. *Alors moi je dis* qu'il donne son nom !, c'est facile de mettre le feu et de ne laisser aucune trace, même pas un briquet (*sauf peut être quand il s'agit de paillotes !*).
Ce soit disant lecteur donc nous envoie le texte d'un exercice qu'il a trouvé dans un manuel scolaire[*]. Il a utilisé cet exercice avec ses étudiants de BTS pour leur faire comprendre la différence entre corrélation (ou co-relation) et causalité. Nous vous livrons cet énoncé et ne doutons pas de l'utilisation pertinente que vous en ferez pour illustrer les abus du "***post hoc, ergo propter hoc***" (*pages rouges du Petit Larousse illustré : A la suite de cela, donc à cause de cela*) .

On donne les deux chroniques suivantes:

[*] Vous pouvez écrire, à Jean PRADIN, LEGTA MOULINS NEUVILLE, 03000 NEUVY
ou à Jean FAGES, ENFA, B.P. 87, 31326 CASTANET - TOLOSAN
ou à tout membre du groupe dans son établissement.

[*] STATISTIQUE de J. LAMAT collection Techniques et Vulgarisation 1968 p. 213

Années	1924	1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936
Postes de radio (pour 10000 habit.)	14	20	23	25	27	31	36	46	55	63	70	76	81
Nombre de maladies mentales pour 1000 habitants	8	8	9	10	11	11	12	16	18	19	20	21	22

Calculez le coefficient de corrélation linéaire.
Peut-on en induire une relation causale ?

Vous trouverez d'autres exemples de corrélation pour lesquelles des conclusions hâtives seraient malheureuses dans l'ouvrage :

LES STATISTIQUES Une nouvelle approche
Donald H. Sanders et François Allard
Mc Graw Hill
p.37

Je vous transcris, pour finir, la définition du mot corrélation du *Petit Larousse illustré* :

"Dépendance réciproque de deux phénomènes qui varient simultanément, qui sont fonction l'un de l'autre, qui évoquent ou manifestent un lien de cause à effet."

- = - = - = - = - = - = - = -