2.5 La méthode du maximum de vraisemblance

Une des techniques parmi les plus populaires pour trouver des estimateurs est celle dite du *maximum de vraisemblance*. Si $X_1, \ldots, X_n$ est une famille de variables aléatoires indépendantes issues de la même densité $f(x;\theta)$ (où θ est le paramètre de la densité f). La fonction de vraisemblance[3] est définie par :

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

Justification de cette formule dans le cas discret

Il faut penser à $L(\theta; x_1, \ldots, x_n)$ comme étant la probabilité que l'échantillon $x_1, \ldots, x_n$ se produise. Dans le cas discret on a

$$L(\theta; x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1 \text{ et } X_2 = x_2 \text{ et } \ldots \text{ et } X_n = x_n)$$

Comme les X_i sont indépendants, on a $L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) = \prod_{i=1}^{n} f(x_i;\theta)$.

L'estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance[4] de θ, noté $\widehat{\theta}_{\text{MLE}}$, est l'estimateur pour lequel la fonction de vraisemblance est la plus grande. Autrement dit, il s'agit de la valeur du paramètre pour laquelle l'échantillon observé est le plus vraisemblable.

Pour chercher cet estimateur, on doit trouver le maximum de la fonction $L(\theta; x_1, \ldots, x_n)$. C'est une optimisation classique : un symptôme d'un maximum est un zéro de la dérivée.

En pratique, il peut être plus facile de trouver le maximum du logarithme de la fonction de vraisemblance (en effet, la fonction ln est croissante et transforme les produits en somme ; ainsi $\ln(L(\theta; x_1, \ldots, x_n)) = \sum_{i=1}^{n} \ln(f(x_i;\theta)))$.

Exemple : les MLEs de la loi normale

On considère la loi normale de paramètres μ et σ dont la densité est donnée en page 16. En exercice, on trouvera[5] les estimateurs du maximum de vraisemblance suivants.

$$\widehat{\mu}_{\text{MLE}} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{et} \quad \widehat{\sigma}_{\text{MLE}} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \widehat{\mu}_{\text{MLE}})^2}{n}}$$

Théorème Si f est une fonction d'un estimateur θ, alors $\widehat{f(\theta)}_{\text{MLE}} = f\left(\widehat{\theta}_{\text{MLE}}\right)$.

Idée de la preuve

Le MLE $\widehat{\theta}_{\text{MLE}}$ est la valeur pour laquelle la fonction de vraisemblance est maximale. On nomme $\nu = f(\theta)$ un nouveau paramètre. On doit exprimer la fonction de vraisemblance en fonction de ce nouveau paramètre ; on a $\theta = f^{-1}(\nu)$ (on crée une fonction réciproque de f en choisissant un domaine de définition). Ainsi $L(\theta) = L(f^{-1}(\nu))$. Cette fonction admet un maximum lorsque $f^{-1}(\widehat{\nu}_{\text{MLE}}) = \widehat{\theta}_{\text{MLE}}$, c'est-à-dire $\widehat{\nu}_{\text{MLE}} = f(\widehat{\theta}_{\text{MLE}})$.

3. En anglais, vraisemblance se dit likelihood. Ainsi la fonction de vraisemblance est notée L.

4. En anglais, estimateur du maximum de vraisemblance se dit **maximum likelihood estimator** (MLE).

5. Lorsqu'on cherche à optimiser une fonction à plusieurs variables, on cherche les valeurs des variables qui annulent le *gradient* (c'est le vecteur dont la i-ème composante est la dérivée de la fonction par rapport à la i-ième variable uniquement).

2.6 Qualités d'un estimateur

Le biais d'un estimateur

Soit $\widehat{\theta}$ un estimateur d'un paramètre θ. Le *biais* de $\widehat{\theta}$ est défini par :

$$\mathrm{b}_\theta\big(\widehat{\theta}\big) = E\Big(\widehat{\theta} - \theta\Big) \overset{\substack{\text{linéarité de} \\ \text{l'espérance}}}{=} E\big(\widehat{\theta}\big) - \theta$$

Un estimateur $\widehat{\theta}$ est dit *sans biais* lorsque $\mathrm{b}_\theta\big(\widehat{\theta}\big) = 0$. Autrement dit, lorsqu'en moyenne l'estimateur est égal au paramètre que l'on cherche à estimer.

Exemple : les estimateurs les plus célèbres

On suppose qu'on a n variables aléatoires X_i indépendantes et suivant toutes une même loi d'espérance μ et de variance σ^2. Alors l'espérance μ peut être estimée sans biais par l'estimateur $\overline{X}$ et la variance σ^2 peut être estimée sans biais par l'estimateur S^2. Ces estimateurs sont définis comme suit :

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{et} \quad S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$$

Dans le cas d'une loi normale $\mathcal{N}(\mu, \sigma^2)$, on remarque que $\overline{X}$ n'est rien d'autre que l'estimateur de vraisemblance de μ, tandis que $S^2 = \frac{n}{n-1}\widehat{\sigma}^2_{\mathrm{MLE}}$.

1. **L'estimateur de la moyenne est sans biais**

   En effet, on a :

   $$E\big(\overline{X}\big) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) \overset{\substack{\text{linéarité de} \\ \text{l'espérance}}}{=} \frac{\sum_{i=1}^n E(X_i)}{n} = \frac{\sum_{i=1}^n \mu}{n} = \frac{n\mu}{n} = \mu$$

   Donc l'estimateur $\overline{X}$ de l'espérance μ est sans biais, car $\mathrm{b}_\mu\big(\overline{X}\big) = E\big(\overline{X}\big) - \mu = 0$.

2. **L'estimateur de la variance est sans biais**

   En effet, on a :

   $$\begin{aligned}
   E(S^2) &= E\left(\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}\right) \overset{\text{lin.}}{=} \frac{1}{n-1}\sum_{i=1}^n E\left((X_i - \overline{X})^2\right) \\
   &= \frac{1}{n-1}\sum_{i=1}^n E\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right) \\
   &\overset{\text{lin.}}{=} \frac{1}{n-1}\left(\sum_{i=1}^n E(X_i^2) - 2E\Big(\textstyle\sum_{i=1}^n X_i \cdot \overline{X}\Big) + \sum_{i=1}^n E(\overline{X}^2)\right) \\
   &= \frac{1}{n-1}\left(\sum_{i=1}^n E(X_i^2) - 2nE\Big(\frac{\sum_{i=1}^n X_i}{n} \cdot \overline{X}\Big) + nE(\overline{X}^2)\right) \\
   &= \frac{1}{n-1}\left(\sum_{i=1}^n E(X_i^2) - 2nE(\overline{X} \cdot \overline{X}) + nE(\overline{X}^2)\right) \\
   &= \frac{1}{n-1}\left(\sum_{i=1}^n E(X_i^2) - nE(\overline{X}^2)\right)
   \end{aligned}$$

Or, d'après la seconde formule de la variance, on a $V(X) = E(X^2) - E(X)^2$. Ainsi, on a la formule suivante pour n'importe quelle variable aléatoire X :

$$E(X^2) = V(X) + E(X)^2 \quad \bigstar$$

Cela permet de reprendre le calcul précédent :

$$\begin{aligned} E\big(S^2\big) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)\right) \\ &\overset{\bigstar}{=} \frac{1}{n-1}\left(\sum_{i=1}^{n}\Big(V(X_i) + E(X_i)^2\Big) - n\Big(V(\overline{X}) + E(\overline{X})^2\Big)\right) \end{aligned}$$

Puisque l'espérance de chaque X_i est égale à μ, que la variance de chaque X_i est égale à σ^2 et que l'estimateur $\overline{X}$ est sans biais, on a :

$$\begin{aligned} E\big(S^2\big) &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\Big(\sigma^2 + \mu^2\Big) - n\Big(V(\overline{X}) + \mu^2\Big)\right) \\ &= \frac{1}{n-1}\Big(n\Big(\sigma^2 + \mu^2\Big) - n\Big(V(\overline{X}) + \mu^2\Big)\Big) = \frac{1}{n-1}\left(n\sigma^2 - nV(\overline{X})\right) \end{aligned}$$

Or, grâce au théorème de la page 24 qui dit que $V(\overline{X}) = \frac{\sigma^2}{n}$ (parce que les variables aléatoires sont indépendantes), on peut finalement montrer que S^2 est sans biais, car :

$$E\big(S^2\big) = \frac{1}{n-1}\left(n\sigma^2 - nV(\overline{X})\right) = \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

L'erreur quadratique moyenne de l'erreur d'un estimateur

Soit $\widehat{\theta}$ un estimateur d'un paramètre θ. L'*erreur quadratique moyenne* de $\widehat{\theta}$ est définie[6] par :

$$\mathrm{MSE}_\theta\big(\widehat{\theta}\big) = E\big((\widehat{\theta} - \theta)^2\big) \overset{\bigstar}{=} V\big(\widehat{\theta}\big) + \Big(\mathrm{b}_\theta\big(\widehat{\theta}\big)\Big)^2$$

Où $V\big(\widehat{\theta}\big)$ est la *variance* de $\widehat{\theta}$ (rappelons qu'un estimateur est une variable aléatoire).

Théorème (sans preuve)

On suppose qu'on a n variables aléatoires X_i indépendantes et qui suivent toutes une loi normale $\mathcal{N}(\mu, \sigma^2)$ d'espérance μ et de variance σ^2. On a :

$$\mathrm{MSE}_\mu\big(\overline{X}\big) = V(\overline{X}) = \frac{\sigma^2}{n} \quad \text{et} \quad \mathrm{MSE}_{\sigma^2}\big(S^2\big) = V(S^2) = \frac{2\sigma^4}{n-1}$$

6. En anglais, on parle de **m**ean **s**quared **e**rror, abrégée MSE.

Chapitre 3

Tests d'hypothèses

«Un statisticien est une personne dont l'ambition principale est d'avoir tort dans 5% des cas.»
Anonyme

Les *tests d'hypothèses* sont utiles pour vérifier si une affirmation sur un modèle théorique correspondant à une expérience aléatoire est cohérente avec les mesures effectuées.

Dans une étude statistique, on peut se demander si les mesures observées peuvent correspondre à une certaine réalité. Par exemple, est-ce que la série de jets d'une pièce de monnaie $(P; P; F; P; F; P)$ peut correspondre à une pièce bien équilibrée ?

Le but d'un test d'hypothèses est de confronter deux hypothèses entre elles : l'*hypothèse nulle* H_0 et l'*hypohèse alternative* H_1. Les *hypothèses* sont des énoncés qui concernent un paramètre d'une population. La confrontation s'effectue à l'aide d'un estimateur du paramètre en question, appelé *statistique de test.*

Dans l'exemple des jets de pièce de monnaie (voir page 38), le paramètre en question est la probabilité p que la pièce tombe sur pile. Son estimateur est donné par $\widehat{P} = \overline{X}$. L'hypothèse nulle sera "la pièce est bien équilibrée $(p = \frac{1}{2})$" et l'hypothèse alternative sera "la pièce n'est pas bien équilibrée $(p \neq \frac{1}{2})$".

On va traiter plusieurs cas :

1. Tests d'hypothèses sur une moyenne ou une proportion.
   (a) Variance connue ou variance inconnue.
   Selon le fait que la variance est connue ou inconnue, on utilisera une loi différente pour effectuer les calculs.
   (b) Tests symétriques ou asymétriques.
   Dans le cours, on présente les tests symétriques, les tests asymétriques seront vus en exercices.

   Les tests d'hypothèses sur une proportion sont des tests où la variance est connue, parce que des lois de Bernoulli et binomiale apparaissent.
2. Test d'hypothèses de comparaison de moyenne : données appariées ou non.
3. Les tests du chi-carré
   a) pour l'adéquation à une loi ; b) pour la comparaison d'échantillon ;
   c) pour l'indépendance.

Il existe d'autres tests d'hypothèses. Par exemple, les tests permettant d'inférer la variance σ^2 d'une population. Ces tests utilisent d'autres lois de probabilité, comme, par exemple, la distribution de Fisher à deux paramètres.

3.1 Tests d'hypothèses symétriques sur une moyenne

3.1.1 Tests d'hypothèses symétriques sur une moyenne, variance connue

Considérons des variables aléatoires $X_1, \ldots, X_n$ indépendantes qui suivent une même loi d'espérance μ (c'est la moyenne théorique qu'on cherche à tester) et de variance σ^2 (supposée connue).

Dans ce modèle de tests d'hypothèses, on teste l'hypothèse «la moyenne théorique vaut μ_0» à partir des n mesures effectuées.

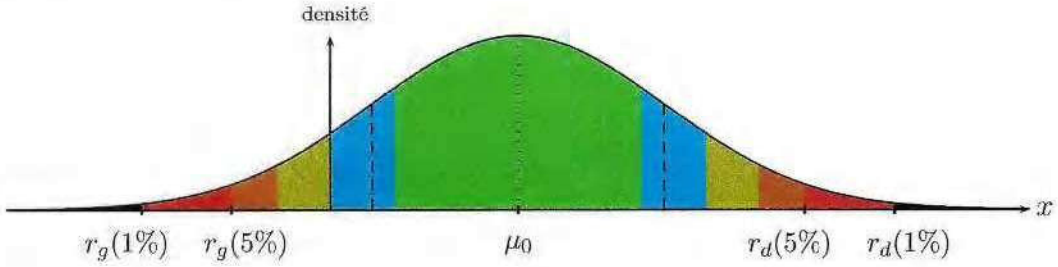
On se trouve face à deux alternatives, appelées *hypothèses*, qui sont

hypothèse nulle		*hypothèse alternative*
$H_0 : \mu = \mu_0$	et	$H_1 : \mu \neq \mu_0$

Comme dans un raisonnement par l'absurde, on suppose qu'on se trouve sous l'hypothèse H_0 et on regarde si les données mesurées permettent d'en tirer une contradiction.

Sous l'hypothèse H_0, les variables aléatoires X_i sont d'espérance μ_0.

Par le théorème de la limite centrale, $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ suit approximativement une loi normale $\mathcal{N}(\mu_0, \frac{\sigma^2}{n})$. En traitillés, on voit les bornes de l'intervalle $\left[\mu_0 - \frac{\sigma}{\sqrt{n}}, \mu_0 + \frac{\sigma}{\sqrt{n}}\right]$; on remarque ainsi que lorsque n grandit la courbe se resserre.



On estime $\overline{X}$ par la moyenne des mesures effectuées $\overline{x} = \frac{x_1+\cdots+x_n}{n}$ (attention à bien faire la différence entre les majuscules et les minuscules). L'estimation $\overline{x}$ donnée par les mesures va donc se trouver quelque part sous la loi.

Il y a une probabilité de 60% que $\overline{x}$ tombe dans la zone ■, il y a une probabilité de 20% que $\overline{x}$ tombe dans la zone ■, il y a une probabilité de 10% que $\overline{x}$ tombe dans la zone ■, il y a une probabilité de 5% que $\overline{x}$ tombe dans la zone ■, il y a une probabilité de 4% que $\overline{x}$ tombe dans la zone ■, il y a une probabilité de 1% que $\overline{x}$ tombe dans la zone ■.

Plus on se trouve dans une zone éloignée de μ_0 donc dans l'ordre : ■, ■, ■, ■, ■, ■ ; plus les mesures sont en contradiction avec l'hypothèse H_0. On décide ainsi du critère suivant.

On *rejette l'hypothèse* H_0 *au seuil de signification* 5% si $\overline{x}$ se trouve dans la zone ■ ou ■.
On *rejette l'hypothèse* H_0 *au seuil de signification* 1% si $\overline{x}$ se trouve dans la zone ■.

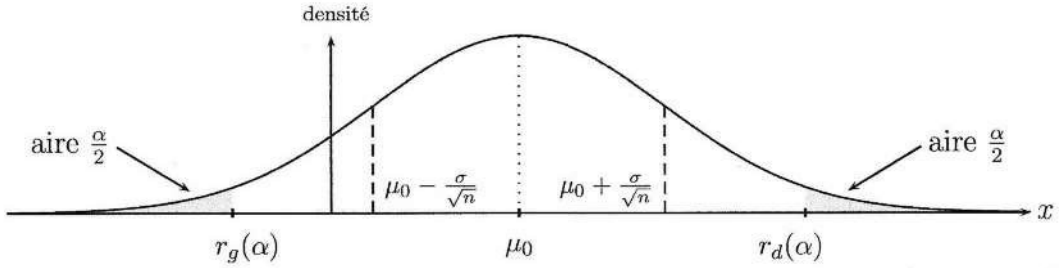
Le *seuil de signification* du test est la probabilité α que l'on a de rejeter l'hypothèse H_0 sachant que H_0 est vraie.

$$\mathbb{P}(\text{ rejeter } H_0 \mid H_0 \text{ est vraie }) = \alpha$$

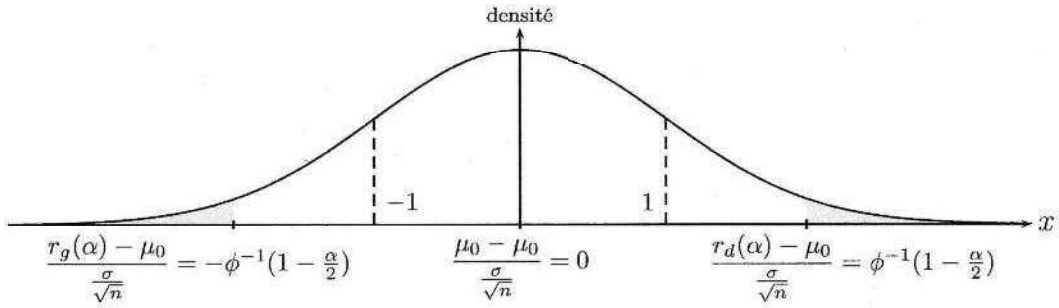
La distribution ci-dessous est la distribution sous l'hypothèse H_0, donc α se voit sur le dessin : c'est l'aire sous la distribution dans la zone de rejet de H_0 (en grisé).

On montre dans le cours OS que si on a une variable N qui suit une loi normale $\mathcal{N}(\mu_0, \frac{\sigma^2}{n})$, alors la variable $\frac{N-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ suit la loi normale $\mathcal{N}(0,1)$. On dit que N a été *centrée-réduite.*

Si la densité de $\overline{X}$ est



Alors la densité de $\dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ est (l'échelle a changé)



où $\phi(x) = \mathbb{P}(Z \leqslant x)$ est la fonction de répartition de la variable aléatoire Z qui suit la loi normale $\mathcal{N}(0,1)$. Les valeurs de ϕ et ϕ^{-1} se trouvent dans la table de la page 30.

On rejette H_0 au seuil de signification α si

pour la variable aléatoire $\overline{X}$	pour la variable aléatoire $\dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
$\overline{x} < r_g(\alpha)$ ou $\overline{x} > r_d(\alpha)$	$\left\lvert \dfrac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right\rvert > \phi^{-1}\left(1 - \frac{\alpha}{2}\right)$

En particulier, si $\alpha = 5\%$, le critère est $\left\lvert \dfrac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right\rvert > \phi^{-1}(0.975) \overset{\text{table}}{\cong} 1.96$.

De même, si $\alpha = 1\%$, le critère est $\left\lvert \dfrac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right\rvert > \phi^{-1}(0.995) \overset{\text{table}}{\cong} 2.58$.

Exemple

Le test "Gauche-Droite" de Jean Piaget a pour but de vérifier l'acquisition par l'enfant des notions gauche-droite à différents points de vue et d'évaluer ainsi son niveau de socialisation et subjectivisation. Les enfants de 7 ans obtiennent en moyenne 12 comme résultat avec un écart type de 3.4. On applique ces tests à 25 enfants gauchers de 7 ans choisis au hasard et on obtient un résultat moyen de 13.4.

À partir de ces tests, peut-on affirmer qu'il y a une différence significative entre les gauchers et les droitiers ?

Réponse

Ici, on met en doute le fait que la moyenne théorique des gauchers est égale à celle des droitiers. On suppose par ailleurs que l'écart type théorique est le même pour les deux populations[1]. L'approximation par le théorème de la limite centrale est de bonne qualité car $n = 25$. On peut donc effectuer le test avec les alternatives suivantes.

$$\begin{array}{ccc} \textit{hypothèse nulle} & & \textit{hypothèse alternative} \\ H_0 : \mu = 12 & \text{et} & H_1 : \mu \neq 12 \end{array}$$

Prétendre que $\mu = 12$ revient à prétendre que les gauchers suivent le même modèle que celui de l'ensemble de la population (donc a fortiori celui des droitiers).

Les données livrent une moyenne $\overline{x} = 13.4$. On a ainsi

$$\left| \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| = \left| \frac{13.4 - 12}{\frac{3.4}{\sqrt{25}}} \right| \cong 2.059$$

On voit qu'on rejette H_0 au seuil 5%, mais qu'on ne peut pas rejeter H_0 au seuil 1% (en fait $\overline{x}$ est tombé dans la zone ■, mais pas dans la zone ■).

Remarques

1. En supposant que H_0 soit vraie, et qu'un grand nombre d'examinateurs aient fait passer ces tests à d'autres groupes de 25 gauchers, alors 1% des examinateurs auraient eu une moyenne inférieur à 10.25 ou supérieur à 13.75 et 5% des examinateurs auraient eu une moyenne inférieur à 10.7 ou supérieur à 13.3.
2. On ne fixe pas le seuil de signification après avoir fait le test, mais avant. Ceci afin d'éviter de régler le seuil pour que le test donne le résultat escompté après avoir effectué les mesures.
3. Il est habituel de prendre 1% pour confirmer H_0 et 5% pour infirmer H_0, mais ces seuils sont arbitraires. En cas de doute sur la conclusion du test, il est important d'aller regarder de plus près comment les mesures ont été effectuées. On peut aussi refaire l'expérience sur un autre échantillon. Si lors des mesures, une erreur a été effectuée, le test ne sert plus à rien.
4. Pour qu'un test d'hypothèses soit utile, il faut que les mesures aient été effectuées sur un échantillon représentatif de la population.

1. Si le lecteur n'est pas d'accord avec cette hypothèse, il peut utiliser le test de Student mis au point par William Gosset (voir section suivante).

Les différents types d'erreurs

Il y a deux probabilités α et β qui décrivent deux types de risque qu'on rencontre en effectuant un test d'hypothèses. On a déjà parlé de α qui est aussi appelé le *seuil de signification d'un test d'hypothèse*, mais il y a aussi β qui est défini ci-dessous.

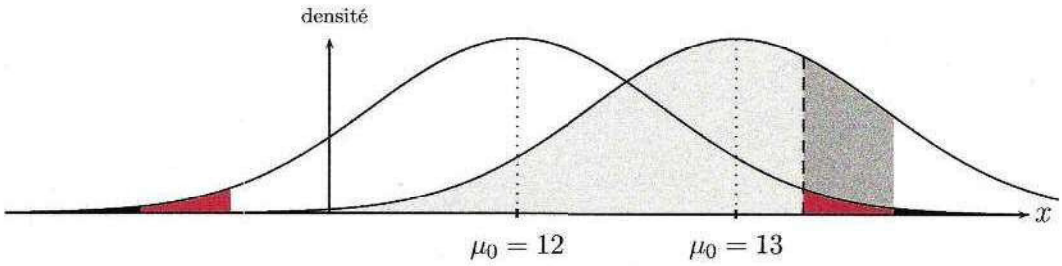
$$\begin{aligned} \alpha &= \mathbb{P}(\text{ rejeter } H_0 \mid H_0 \text{ est vraie }) \\ \beta &= \mathbb{P}(\text{ ne pas rejeter } H_0 \mid H_0 \text{ est fausse }) \end{aligned}$$

Comme il est difficile de calculer β, dans la pratique on calcule plutôt des β_{μ_i} comme définis ci-dessous.

	H_0 est vraie	H_0 est fausse		
	$\mu = \mu_0$	$\mu = \mu_1$	$\ldots$	$\mu = \mu_k$
On rejette H_0	Mauvaise décision Probabilité α	Bonne décision Probabilité $1 - \beta_{\mu_1}$	$\ldots$	Bonne décision Probabilité $1 - \beta_{\mu_k}$
On ne rejette pas H_0	Bonne décision Probabilité $1 - \alpha$	Mauvaise décision Probabilité β_{μ_1}	$\ldots$	Mauvaise décision Probabilité β_{μ_k}

Le seuil α est une probabilité appelée *risque de première espèce* et les nombres β_{μ_i}, tout comme β, sont appelées *risques de deuxième espèce.*

Calcul d'un risque de deuxième espèce (suite de l'exemple précédent)



Pour un seuil $\alpha = 5\%$, le risque de deuxième espèce avec $H_1 : \mu = 13$ donne $\beta_{13} \cong 69\%$ (c'est la proportion de la zone en gris clair déterminée par les zones ■ et ■). Pour un seuil $\alpha = 1\%$, le risque de deuxième espèce avec $H_1 : \mu = 13$ donne $\beta_{13} \cong 87\%$ (c'est la proportion de la zone en gris clair et foncé déterminée par la zone ■).

Autrement dit, au seuil $\alpha = 5\%$. Si H_0 est vrai, la bonne décision est de ne pas rejeter l'hypothèse ; sa probabilité est de 95%. La mauvaise décision a une probabilité de 5%.

Par contre (au même seuil α), si au lieu de H_0, c'est $H_1 : \mu = 13$ qui est vrai, alors la bonne décision est de rejeter H_0, la probabilité est maintenant de 31%. La mauvaise décision a une probabilité de 69%.

On le voit sur cet exemple, si le risque de premier espèce diminue, alors le risque de deuxième espèce augmente.

Pour diminuer le risque de deuxième espèce, il faudrait augmenter le seuil α et par conséquent le risque de première espèce.

C'est un compromis qu'il faut savoir accepter.

3.1.2 Test d'hypothèses symétriques sur une moyenne, variance inconnue

On suppose que les variables aléatoires $X_1, \ldots, X_n$ qui correspondent aux n valeurs observées sont indépendantes et suivent une loi normale $\mathcal{N}(\mu, \sigma^2)$ où les paramètres μ et σ sont inconnus.

Sous ces hypothèses, l'estimateur $\overline{X}$ suit une loi normale de paramètres $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

On teste l'hypothèse nulle $H_0 : \mu = \mu_0$ contre l'hypothèse alternative $H_1 : \mu \neq \mu_0$.

Comme avant, on se place sous l'hypothèse nulle et on contemple. Malheureusement, on ne peut pas utiliser la loi $\mathcal{N}(\mu_0, \frac{\sigma^2}{n})$ pour calculer les probabilités puisque σ est inconnu. Heureusement, William Gosset, brasseur et mathématicien britannique trouva une loi qui permit de travailler sous l'hypothèse H_0. Il publia son résultat sous le pseudonyme de Student.

Théorème (sans preuve)

Si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes qui suivent une loi normale $\mathcal{N}(\mu, \sigma^2)$, alors la variable aléatoire[2] $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ suit une distribution de Student avec $n-1$ degrés de liberté.

Théorème

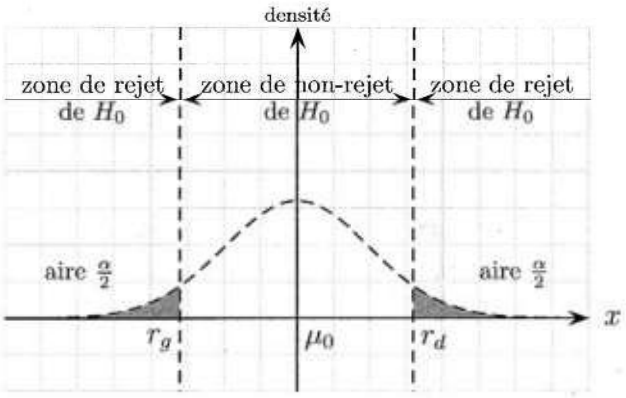
On rejette l'hypothèse H_0 au seuil de signification α si[3]

$$\overline{x} \notin \left[\mu_0 - \phi_{n-1}^{-1}\left(1-\tfrac{\alpha}{2}\right)\cdot\tfrac{s}{\sqrt{n}}\ ,\ \mu_0 + \phi_{n-1}^{-1}\left(1-\tfrac{\alpha}{2}\right)\cdot\tfrac{s}{\sqrt{n}}\right] \quad \text{ou} \quad \left|\tfrac{\overline{x}-\mu_0}{s/\sqrt{n}}\right| > \phi_{n-1}^{-1}\left(1-\tfrac{\alpha}{2}\right)$$

où $\phi_{n-1}(x)$ est la fonction de répartition de la loi de Student à $n-1$ degrés de liberté.

Preuve

On se fixe le seuil de signification α. Sous l'hypothèse H_0, on sait que $\overline{X}$ suit une distribution normale $\mathcal{N}(\mu_0; \frac{\sigma^2}{n})$ de la forme suivante :



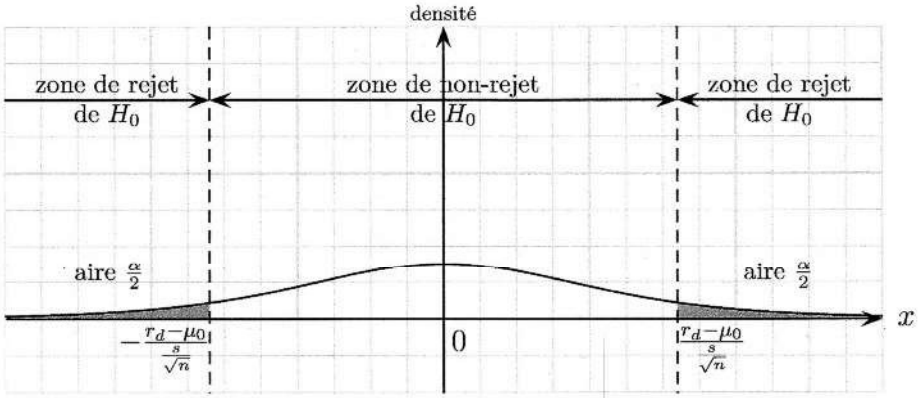
2. Ce n'est pas la variable centrée réduite de $\overline{X}$, car on a remplacé σ par son estimateur sans biais S.

3. Dans l'énoncé de ce théorème, s est l'estimation de l'estimateur S, tout comme $\overline{x}$ est l'estimation de l'estimateur $\overline{X}$.

Malheureusement, comme σ est inconnu, on ne peut pas utiliser cette loi pour calculer la valeur exacte de r_d sous l'hypothèse H_0. Mais grâce à Gosset, on sait que la variable aléatoire $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ suit une distribution de Student avec $n-1$ degrés de liberté.

$$1-\frac{\alpha}{2} = \mathbb{P}\left(\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \leqslant \frac{r_d-\mu}{\frac{s}{\sqrt{n}}}\right) \overset{\text{on se place sous l'hypothèse } H_0}{\overset{\downarrow}{=}} \mathbb{P}\left(\frac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}} \leqslant \frac{r_d-\mu_0}{\frac{s}{\sqrt{n}}}\right) = \phi_{n-1}\left(\frac{r_d-\mu_0}{\frac{s}{\sqrt{n}}}\right)$$

Voici la représentation de cette loi de Student :



Donc, on a :

$$\boxed{\frac{r_d-\mu_0}{\frac{s}{\sqrt{n}}} = \phi_{n-1}^{-1}\left(1-\tfrac{\alpha}{2}\right)} \iff r_d-\mu_0 = \phi_{n-1}^{-1}\left(1-\tfrac{\alpha}{2}\right)\cdot\frac{s}{\sqrt{n}} \iff \boxed{r_d = \mu_0 + \phi_{n-1}^{-1}\left(1-\tfrac{\alpha}{2}\right)\cdot\frac{s}{\sqrt{n}}}$$

Donc, on rejette l'hypothèse H_0 au seuil de signification α si l'estimation $\overline{x}$ est plus grande que r_d ou plus petite que $r_g = \mu_0 - \phi_{n-1}^{-1}\left(1-\frac{\alpha}{2}\right)\cdot\frac{s}{\sqrt{n}}$. □

3.1.3 Résumé et autres statistiques de tests sur les moyennes

Dans les tableaux ci-dessous, on note $z_\alpha = \phi^{-1}(\alpha)$ où ϕ est la fonction de répartition de la loi normale centrée réduite et $t_{\nu,\alpha} = \phi_\nu^{-1}(\alpha)$ où ϕ_ν est la fonction de répartition de la loi de Student à ν degrés de liberté.

Tests d'inférence d'une espérance (rappel)

On considère n variables aléatoires indépendantes $X_1, X_2, \ldots, X_n$ qui suivent toute la même loi normale $\mathcal{N}(\mu, \sigma^2)$.

H_0	Statistique de test	H_1	Région de rejet
$\mu = \mu_0$	$Z = \dfrac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$ σ connu	$\mu \neq \mu_0$ $\mu < \mu_0$ $\mu > \mu_0$	$Z \leqslant z_{\frac{\alpha}{2}}$ ou $Z \geqslant z_{1-\frac{\alpha}{2}}$ $Z \leqslant z_\alpha$ $Z \geqslant z_{1-\alpha}$
$\mu = \mu_0$	$T = \dfrac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$ σ inconnu	$\mu \neq \mu_0$ $\mu < \mu_0$ $\mu > \mu_0$	$T \leqslant t_{n-1,\frac{\alpha}{2}}$ ou $T \geqslant t_{n-1,1-\frac{\alpha}{2}}$ $T \leqslant t_{n-1,\alpha}$ $T \geqslant t_{n-1,1-\alpha}$

Tests de comparaison de deux espérances

On considère m variables aléatoires indépendantes $X_1, X_2, \ldots, X_m$ qui suivent toute la même loi normale $\mathcal{N}(\mu_X, \sigma_X^2)$ et n variables aléatoires indépendantes $Y_1, Y_2, \ldots, Y_n$ qui suivent toute la même loi normale $\mathcal{N}(\mu_Y, \sigma_Y^2)$. On suppose en outre que ces familles de variables sont indépendantes.

H_0	Statistique de test	H_1	Région de rejet
$\mu_X - \mu_Y = d_0$	$Z = \dfrac{(\overline{X} - \overline{Y}) - d_0}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}$ σ_X et σ_Y connus	$\mu_X - \mu_Y \neq d_0$ $\mu_X - \mu_Y < d_0$ $\mu_X - \mu_Y > d_0$	$Z \leqslant z_{\frac{\alpha}{2}}$ ou $Z \geqslant z_{1-\frac{\alpha}{2}}$ $Z \leqslant z_\alpha$ $Z \geqslant z_{1-\alpha}$
$\mu_X - \mu_Y = d_0$	$T = \dfrac{(\overline{X} - \overline{Y}) - d_0}{\sqrt{\frac{S_p^2}{m} + \frac{S_p^2}{n}}}$ $\sigma_X = \sigma_Y$ inconnus $S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$	$\mu_X - \mu_Y \neq d_0$ $\mu_X - \mu_Y < d_0$ $\mu_X - \mu_Y > d_0$	$T \leqslant t_{m+n-2,\frac{\alpha}{2}}$ ou $T \geqslant t_{m+n-2,1-\frac{\alpha}{2}}$ $T \leqslant t_{m+n-2,\alpha}$ $T \geqslant t_{m+n-2,1-\alpha}$
$\mu_X - \mu_Y = d_0$	$T = \dfrac{(\overline{X} - \overline{Y}) - d_0}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}$ $\sigma_X \neq \sigma_Y$ inconnus $\hat{\nu} = \dfrac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^2}{\frac{\left(\frac{S_X^2}{m}\right)^2}{m-1} + \frac{\left(\frac{S_Y^2}{n}\right)^2}{n-1}}$	$\mu_X - \mu_Y \neq d_0$ $\mu_X - \mu_Y < d_0$ $\mu_X - \mu_Y > d_0$	$T \leqslant t_{\hat{\nu},\frac{\alpha}{2}}$ ou $T \geqslant t_{\hat{\nu},1-\frac{\alpha}{2}}$ $T \leqslant t_{\hat{\nu},\alpha}$ $T \geqslant t_{\hat{\nu},1-\alpha}$

Tests de comparaison de deux espérances d'échantillons appariés

On considère n variables aléatoires indépendantes X_1, X_2, ..., X_n qui suivent toute la même loi normale $\mathcal{N}(\mu_X, \sigma_X^2)$ et n variables aléatoires indépendantes Y_1, Y_2, ..., Y_n qui suivent toute la même loi normale $\mathcal{N}(\mu_Y, \sigma_Y^2)$. On suppose de plus que pour chaque i, il y a une dépendance entre X_i et Y_i : on dit que X_i et Y_i sont appariés.

On sait (c'est un théorème) que les variables aléatoires $D_i = Y_i - X_i$ suivent aussi une loi normale $\mathcal{N}(\mu_D, \sigma_D^2)$ où $\mu_D = \mu_Y - \mu_X$. On est ainsi ramené au test d'inférence de l'espérance de D_i qui se trouve sur la page précédente (celui où la variance n'est pas supposée connue).

H_0	Statistique de test	H_1	Région de rejet
$\underbrace{\mu_Y - \mu_X}_{=\mu_D} = d_0$	$Z = \dfrac{\overline{D} - d_0}{\sqrt{\frac{\sigma_D^2}{n}}}$ σ_D connu	$\mu_Y - \mu_X \neq d_0$ $\mu_Y - \mu_X < d_0$ $\mu_Y - \mu_X > d_0$	$Z \leqslant z_{\frac{\alpha}{2}}$ ou $Z \geqslant z_{1-\frac{\alpha}{2}}$ $Z \leqslant z_\alpha$ $Z \geqslant z_{1-\alpha}$
$\underbrace{\mu_Y - \mu_X}_{=\mu_D} = d_0$	$T = \dfrac{\overline{D} - d_0}{\sqrt{\frac{S_D^2}{n}}}$ σ_D inconnu	$\mu_Y - \mu_X \neq d_0$ $\mu_Y - \mu_X < d_0$ $\mu_Y - \mu_X > d_0$	$T \leqslant t_{n-1,\frac{\alpha}{2}}$ ou $T \geqslant t_{n-1,1-\frac{\alpha}{2}}$ $T \leqslant t_{n-1,\alpha}$ $T \geqslant t_{n-1,1-\alpha}$

3.2 Test du chi-carré : adéquation à une loi

Considérons des variables aléatoires $X_1, \ldots, X_n$ indépendantes qui suivent une même loi qu'une variable aléatoire X. On cherche à tester la densité de probabilité (ou distribution) de X.

On découpe les valeurs possibles pour X en k classes $A_1, \ldots, A_k$. On note N_i les variables aléatoires qui comptent le nombre de mesures qui tombent dans la classe A_i.

tableau des probabilités théoriques

A_1	$\cdots$	A_k	total
$\mathbb{P}(X \in A_1)$	$\cdots$	$\mathbb{P}(X \in A_k)$	1

Dans ce modèle de test d'hypothèses, on teste l'hypothèse «la variable X a une certaine densité de probabilité» à partir des n mesures effectuées. À l'aide du découpage en classes, cette hypothèse est reformulée par «les probabilités théoriques $\mathbb{P}(X \in A_i)$ valent p_i».

On se trouve face à deux alternatives qui sont

hypothèse nulle $H_0 : \mathbb{P}(X \in A_i) = p_i$ pour tout i et *hypothèse alternative* $H_1 :$ il existe i tel que $\mathbb{P}(X \in A_i) \neq p_i$

Comme dans un raisonnement par l'absurde, on suppose qu'on se trouve sous l'hypothèse H_0 et on regarde si les données mesurées permettent d'en tirer une contradiction.

Sous l'hypothèse H_0, on peut établir le tableau des effectifs théoriques.

effectifs mesurés

A_1	$\cdots$	A_k	total
n_1	$\cdots$	n_k	n

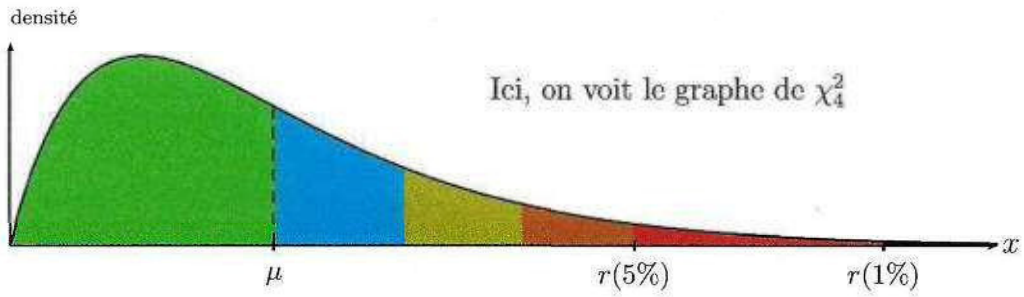
effectifs théoriques

A_1	$\cdots$	A_k	total
np_1	$\cdots$	np_k	n

Les mathématiciens ont montré que si les effectifs théoriques np_i sont de taille au moins 5, alors la variable aléatoire D^2 suivante suit approximativement une loi du chi-carré à $k-1$ degrés de liberté.

$$D^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

On a $k-1$ degrés de liberté, car les variables aléatoires N_1 à N_k sont liées entre-elles par la condition : $\sum_i N_i = n$.



On estime D^2 à l'aide des mesures effectuées $d^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$ (attention à bien faire la différence entre les majuscules et les minuscules). L'estimation d^2 donnée par les mesures va donc se trouver quelque part sous la loi.

Il y a une probabilité de 60% que d^2 tombe dans la zone ■, il y a une probabilité de 20% que d^2 tombe dans la zone ■, il y a une probabilité de 10% que d^2 tombe dans la zone ■, il y a une probabilité de 5% que d^2 tombe dans la zone ■, il y a une probabilité de 4% que d^2 tombe dans la zone ■, il y a une probabilité de 1% que d^2 tombe dans la zone ■.

Plus on se trouve dans une zone éloignée de 0 donc dans l'ordre : ■, ■, ■, ■, ■, ■; plus les mesures sont en contradiction avec l'hypothèse H_0. On décide ainsi du critère suivant.

On *rejette l'hypothèse H_0 au seuil de signification* 5% si d^2 se trouve dans la zone ■ ou ■.
On *rejette l'hypothèse H_0 au seuil de signification* 1% si d^2 se trouve dans la zone ■.

On rejette H_0 au seuil de signification α si

$$d^2 > \phi_{k-1}^{-1}(1-\alpha)$$

où $\phi_{k-1}(x) = \mathbb{P}(D^2 \leqslant x)$ est la fonction de répartition de la variable aléatoire D^2 qui suit la loi du chi-carré χ^2_{k-1} avec $k-1$ degrés de liberté. Les valeurs de cette fonction ϕ_{k-1} sont données par la table de la page 32.

En particulier, si $\alpha = 5\%$, le critère est $d^2 > \phi_{k-1}^{-1}(0.95) \overset{\text{table}}{=} 11.07$ si $k = 6$.

De même, si $\alpha = 1\%$, le critère est $d^2 > \phi_{k-1}^{-1}(0.99) \overset{\text{table}}{=} 15.09$ si $k = 6$.

Remarque sur les degrés de liberté

Si lors du calcul des probabilités p_i, on doit utiliser k estimateurs, il faut encore enlever k degrés de liberté.

Ce peut être le cas lorsque, par exemple, on calcule les p_i avec une loi normale où il faut d'abord estimer l'espérance et la variance.

Remarque sur le risque de deuxième espèce

Dans le cas d'un test du χ^2, le risque de deuxième espèce ne se calcule pas aussi facilement que dans le cas d'un test sur un moyenne. Cela dépasse le cadre de ce cours.

Exemples

Reprenons l'exemple du début du cours où on a lancé un dé 60 fois. Les classes choisies sont $A_i = \{i\}$ de sorte que A_i corresponde à l'apparition de la face i.

1. Testons l'alternative suivante.

   H_0 : le dé est bien équilibré, c'est-à-dire $p_i = \frac{1}{6}$.
   H_1 : le dé n'est pas bien équilibré.

   On peut donc construire les tableaux suivants.

   effectifs mesurés

face	1	2	3	4	5	6	total
n_i	13	13	11	8	8	7	60

   effectifs théoriques

face	1	2	3	4	5	6	total
np_i	10	10	10	10	10	10	60

   Pour que l'approximation soit bonne, il faut que $np_i \geqslant 5$ pour chaque i. C'est le cas ici. On a

   $$d^2 = \sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} = \underbrace{\frac{(13-10)^2}{10}}_{\text{pour la face 1}} + \underbrace{\frac{(13-10)^2}{10}}_{\text{pour la face 2}} + \cdots + \underbrace{\frac{(7-10)^2}{10}}_{\text{pour la face 6}} = 3.6$$

   Comme $d^2 < \phi_5^{-1}(0.95) \cong 11.07$, on ne peut pas rejeter l'hypothèse H_0 au seuil 5%. Il n'y a pas assez de preuves pour dire que le dé n'est pas bien équilibré.

2. Imaginons qu'on pense qu'il y a une bille de plomb vers le sommet adjacent aux faces 4, 5 et 6 qui fait que le dé montre pour le 75% des lancés les faces 1, 2 ou 3.

   On en déduit que $p_1 = p_2 = p_3 = \frac{3}{12}$ et que $p_4 = p_5 = p_6 = \frac{1}{12}$ (le total des p_i vaut bien 1, et $p_1 + p_2 + p_3 = 3(p_4 + p_5 + p_6)$ tout comme 75% vaut $3 \cdot 25\%$).

   Testons l'alternative suivante.

   H_0 : le dé est truqué comme ci-dessus.
   H_1 : le dé n'est pas truqué comme ci-dessus.

   On peut donc construire les tableaux suivants.

   effectifs mesurés

face	1	2	3	4	5	6	total
n_i	13	13	11	8	8	7	60

   effectifs théoriques

face	1	2	3	4	5	6	total
np_i	15	15	15	5	5	5	60

   Pour que l'approximation soit bonne, il faut que $np_i \geqslant 5$ pour chaque i. C'est le cas ici. On a

   $$d^2 = \sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} = \underbrace{\frac{(13-15)^2}{15}}_{\text{pour la face 1}} + \underbrace{\frac{(13-15)^2}{15}}_{\text{pour la face 2}} + \cdots + \underbrace{\frac{(7-5)^2}{5}}_{\text{pour la face 6}} = 6$$

   Comme $d^2 < \phi_5^{-1}(0.95) \cong 11.07$, on ne peut pas rejeter l'hypothèse H_0 au seuil 5%. Il n'y a pas assez de preuves pour dire que le dé n'est pas truqué de cette façon.

Les résultats des 60 jets ne sont pas assez différents de ce qu'on aurait pu obtenir avec un dé bien équilibré ou un dé truqué de la façon ci-dessus. La valeur mesurée d^2 indiquerait que la situation la plus probable est que le dé est bien équilibré (la p-valeur est meilleure sur le premier exemple ($\cong 60.8\%$) que sur le deuxième ($\cong 30.6\%$)).

Néanmoins, le mieux serait de lancer ce dé encore 60 fois et de refaire ces tests sur les 120 lancers obtenus.

3.3 Test du chi-carré : comparaison d'échantillons

Pour chaque échantillon i, $1 \leqslant i \leqslant k$, considérons des variables aléatoires $X_1^{(i)}, \ldots, X_{n^{(i)}}^{(i)}$ indépendantes qui suivent une même loi qu'une variable aléatoire $X^{(i)}$. On cherche à tester si les lois $X^{(i)}$ suivent toutes une même loi qu'une variable aléatoire X.

On découpe les valeurs possibles pour les échantillons en l classes $A_1, \ldots, A_l$ (les mêmes pour chaque échantillon). On note $N_j^{(i)}$ les variables aléatoires qui comptent le nombre de mesures de l'échantillon i qui tombent dans la classe A_j. On note aussi $N_j = \sum_i N_j^{(i)}$.

tableau des probabilités théoriques

	A_1	$\cdots$	A_l	totaux
échantillon 1	$\mathbb{P}(X^{(1)} \in A_1)$	$\cdots$	$\mathbb{P}(X^{(1)} \in A_l)$	1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
échantillon k	$\mathbb{P}(X^{(k)} \in A_1)$	$\cdots$	$\mathbb{P}(X^{(k)} \in A_l)$	1
moyenne	$\frac{1}{k}\sum_i \mathbb{P}(X^{(i)} \in A_1)$	$\cdots$	$\frac{1}{k}\sum_i \mathbb{P}(X^{(i)} \in A_l)$	1

Dans ce modèle de tests d'hypothèses, on teste l'hypothèse «les échantillons suivent tous la même loi» à partir des mesures effectuées sur chaque échantillon. Cette hypothèse est reformulée par «les $X^{(i)}$ suivent une même loi X pour tout i».

On se trouve face à deux alternatives qui sont

$$H_0 : \mathbb{P}(X^{(i)} \in A_j) = \mathbb{P}(X \in A_j) \text{ pour tout } i$$

$$H_1 : \text{ il existe } i \text{ tel que } \mathbb{P}(X^{(i)} \in A_j) \neq \mathbb{P}(X \in A_j)$$

Comme dans un raisonnement par l'absurde, on suppose qu'on se trouve sous l'hypothèse H_0 et on regarde si les données mesurées permettent d'en tirer une contradiction.

Sous l'hypothèse H_0, les moyennes ci-dessus deviennent

$$\frac{1}{k}\sum_i \mathbb{P}(X^{(i)} \in A_j) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{P}(X \in A_j) = \mathbb{P}(X \in A_j)$$

et on peut établir le tableau des effectifs théoriques en notant pour simplifier n pour le nombre total de mesures $n = \sum_i n^{(i)}$ et $p_j = \mathbb{P}(X \in A_j)$.

effectifs mesurés

	A_1	$\cdots$	A_l	totaux
éch. 1	$n_1^{(1)}$	$\cdots$	$n_l^{(1)}$	$n^{(1)}$
$\vdots$	$\vdots$	$n_j^{(i)}$	$\vdots$	$\vdots$
éch. k	$n_1^{(k)}$	$\cdots$	$n_l^{(k)}$	$n^{(k)}$
totaux	n_1	$\cdots$	n_l	n

effectifs théoriques

	A_1	$\cdots$	A_l	totaux
éch. 1	$n^{(1)}p_1$	$\cdots$	$n^{(1)}p_l$	$n^{(1)}$
$\vdots$	$\vdots$	$n^{(i)}p_j$	$\vdots$	$\vdots$
éch. k	$n^{(k)}p_1$	$\cdots$	$n^{(k)}p_l$	$n^{(k)}$
totaux	np_1	$\cdots$	np_l	n

Les mathématiciens ont montré que si les effectifs théoriques $n^{(i)}p_j$ sont de taille au moins 5, alors la variable aléatoire D^2 suivante suit approximativement une loi du χ^2 à $kl-k$ degrés de liberté (on enlève à kl un degré par échantillon, car les variables aléatoires $N_1^{(i)}$ à $N_l^{(i)}$ sont liées entre-elles par la condition : $\sum_j N_j^{(i)} = n^{(i)}$).

$$D^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(N_j^{(i)} - n^{(i)}p_j)^2}{n^{(i)}p_j}$$

Lorsque que les paramètres p_j sont inconnus : on les estime[4] par $p_j = \frac{n_j}{n}$. Ces estimateurs sont naturels, ils font en sorte que les totaux des deux tableaux soient identiques.

Dans ce cas, il faut recalculer les degrés de liberté, il faut encore enlever $(l-1)$ degrés de liberté (les estimateurs p_1 à p_l sont liés entre-eux par la condition $\sum_j p_j = 1$). Ainsi le nombre de degré de liberté vaut $kl - k - (l-1) = kl - k - l + 1 = k(l-1) - (l-1) = (k-1)(l-1)$.

On estime D^2 à l'aide des mesures effectuées $d^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_j^{(i)} - n^{(i)} p_j)^2}{n^{(i)} p_j}$ (attention à bien faire la différence entre les majuscules et les minuscules). L'estimation d^2 donnée par les mesures va donc se trouver quelque part sous la loi.

L'allure de la loi de probabilité et le critère de rejet sont les mêmes que pour le test d'adéquation (attention aux degrés de liberté qui changent).

Exemple

Cet exemple est basé sur les relevés concernant les déchets urbains du Jura en 2010, 2000 et 1994. Les déchets sont subdivisés en cinq catégories : les déchets urbains combustibles ; les déchets compostables ; le papier et le carton ; le verre ; l'aluminium, le fer blanc, la ferraille.

Ici les échantillons correspondent aux trois années susmentionnées, et on va tester si la façon dont les types de déchets sont répartis évolue avec les années (si c'est le cas, on peut éventuellement affirmer que les habitants ont pris conscience de l'importance de trier les déchets).

H_0 : la répartition des déchets dans les cinq catégories est la même pour les trois ans.
H_1 : il y a deux ans au moins pour lesquels les déchets sont répartis différemment.

Voici les tableaux des effectifs mesurés et celui des effectifs théoriques (calculés à partir de ces mesures sous l'hypothèse H_0). Les unités sont en kilogrammes par habitant et par année. Les erreurs d'arrondis ont été lissées (pour que les totaux jouent).

effectifs mesurés

	comb.	comp.	papier	verre	fer	totaux
1994	284	19	28	33	6	370
2000	265	75	52	47	23	462
2010	252	117	75	48	13	505
totaux	801	211	155	128	42	1337

effectifs théoriques

	comb.	comp.	papier	verre	fer	totaux
1994	222	58	43	35	12	370
2000	277	73	54	44	14	462
2010	302	80	58	49	16	505
totaux	801	211	155	128	42	1337

Pour que l'approximation par la loi du chi-carré soit bonne, il faut que les effectifs théoriques soient plus grands ou égaux à 5. C'est le cas ici, et on a

$$d^2 = \sum_{i=1}^{3} \sum_{j=1}^{5} \frac{(n_{i;j} - np_iq_j)^2}{np_iq_j} = \frac{(284-222)^2}{222} + \frac{(19-58)^2}{58} + \cdots + \frac{(13-16)^2}{16} \cong 89$$

(sans arrondir les effectifs théoriques, on trouve environ 88.952)

Les paramètres p_j ont été estimés, on se retrouve avec $(3-1)(5-1) = 8$ degrés de liberté.

Comme $d^2 > \phi_2^{-1}(0.99) \cong 20.090$, on rejette l'hypothèse H_0 au seuil de signification de 1% (c'est-à-dire avec une probabilité de 1% de chance de rejeter à tort).

Les données laissent à penser que les citoyens trient de mieux en mieux leurs déchets.

4. Il s'agit des estimateurs du maximum de vraisemblance , notés $\widehat{p_j}_{\text{MLE}}$.

3.4 Test du chi-carré : indépendance

Considérons des couples de variables aléatoires $(X_1;Y_1)$, ..., $(X_n;Y_n)$ indépendantes qui suivent les mêmes lois que le couple de variables aléatoires $(X;Y)$. On cherche à tester l'indépendance entre X et Y.

On découpe les valeurs possibles pour X en k classes A_1, ..., A_k, et celles pour Y en l classes B_1, ..., B_l. On note $N_{i;j}$ les variables aléatoires qui comptent le nombre de couples de mesures qui tombent dans $(A_i;B_j)$. On note aussi $N_{i;\heartsuit} = \sum_j N_{i;j}$ et $N_{\heartsuit;j} = \sum_i N_{i;j}$.

tableau des probabilités théoriques

	B_1	$\cdots$	B_l	totaux
A_1	$\mathbb{P}\big((X \in A_1) \cap (Y \in B_1)\big)$	$\cdots$	$\mathbb{P}\big((X \in A_1) \cap (Y \in B_l)\big)$	$\mathbb{P}(X \in A_1)$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_k	$\mathbb{P}\big((X \in A_k) \cap (Y \in B_1)\big)$	$\cdots$	$\mathbb{P}\big((X \in A_k) \cap (Y \in B_l)\big)$	$\mathbb{P}(X \in A_k)$
totaux	$\mathbb{P}(Y \in B_1)$	$\cdots$	$\mathbb{P}(Y \in B_l)$	1

Dans ce modèle de tests d'hypothèses, on teste l'hypothèse «les variables X et Y sont indépendantes» à partir des n mesures effectuées. En utilisant le découpage en classes et la théorie des probabilités, cette hypothèse est reformulée par «les probabilités théoriques satisfont $\mathbb{P}\big((X \in A_i) \cap (Y \in B_j)\big) = \mathbb{P}(X \in A_i) \cdot \mathbb{P}(Y \in B_j)$».

On se trouve face à deux alternatives qui sont

$H_0 : \mathbb{P}\big((X \in A_i) \cap (Y \in B_j)\big) = \mathbb{P}(X \in A_i) \cdot \mathbb{P}(Y \in B_j)$ pour tout i et j

H_1 : il existe i et j tels que $\mathbb{P}\big((X \in A_i) \cap (Y \in B_j)\big) \neq \mathbb{P}(X \in A_i) \cdot \mathbb{P}(Y \in B_j)$

Comme dans un raisonnement par l'absurde, on suppose qu'on se trouve sous l'hypothèse H_0 et on regarde si les données mesurées permettent d'en tirer une contradiction.

Sous l'hypothèse H_0, on peut établir le tableau des effectifs théoriques en notant pour simplifier $p_i = \mathbb{P}(X \in A_i)$ et $q_j = \mathbb{P}(X \in B_j)$.

effectifs mesurés

	B_1	$\cdots$	B_l	totaux
A_1	$n_{1;1}$	$\cdots$	$n_{1;l}$	$n_{1;\heartsuit}$
$\vdots$	$\vdots$	$n_{i;j}$	$\vdots$	$\vdots$
A_k	$n_{k;1}$	$\cdots$	$n_{k;l}$	$n_{k;\heartsuit}$
totaux	$n_{\heartsuit;1}$	$\cdots$	$n_{\heartsuit;l}$	n

effectifs théoriques

	B_1	$\cdots$	B_l	totaux
A_1	np_1q_1	$\cdots$	np_1q_l	np_1
$\vdots$	$\vdots$	np_iq_j	$\vdots$	$\vdots$
A_k	np_kq_1	$\cdots$	np_kq_l	np_k
totaux	nq_1	$\cdots$	nq_l	n

Les mathématiciens ont montré que si les effectifs théoriques np_iq_j sont de taille au moins 5, alors la variable aléatoire D^2 suivante suit approximativement une loi du χ^2 à $kl-1$ degrés de liberté (on enlève à kl un degré de liberté, car les variables aléatoires $N_{i;j}$ sont liées entre-elles par la condition : $\sum_{i,j} N_{i;j} = n$).

$$D^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(N_{i;j} - np_iq_j)^2}{np_iq_j}$$

Lorsque que les paramètres p_i et q_j sont inconnus : on les estime[5] par $p_i = \frac{n_{i;\heartsuit}}{n}$ et $q_j = \frac{n_{\heartsuit;j}}{n}$. Ces estimateurs sont naturels, ils font en sorte que les totaux des deux tableaux soient identiques.

Dans ce cas, il faut recalculer les degrés de liberté, il faut encore enlever $(k-1)+(l-1)$ degrés de liberté (les estimateurs p_i et q_j sont liés entre-eux par les conditions $\sum_i p_i = 1$ et $\sum_j q_j = 1$). Ainsi le nombre de degré de liberté vaut $kl - 1 - (l-1) - (k-1) = kl - k - l + 1 = k(l-1) - (l-1) = (k-1)(l-1)$.

On estime D^2 à l'aide des mesures effectuées $d^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{i;j} - np_iq_j)^2}{np_iq_j}$ (attention à bien faire la différence entre les majuscules et les minuscules). L'estimation d^2 donnée par les mesures va donc se trouver quelque part sous la loi.

L'allure de la loi de probabilité et le critère de rejet sont les mêmes que pour le test d'adéquation (attention aux degrés de liberté qui changent).

Exemple

Cet exemple est tiré du livre «Statistics» (2e édition) écrit par Freeman, Pisani, Purves et Adhikari, aux éditions Norton, international student edition.

Une étude basée sur 2237 américains âgés de 25 à 34 a permis de montrer que les femmes sont plus souvent droitières que les hommes.

L'étude consistait à faire un test d'hypothèse sur l'indépendance entre le sexe d'une personne et le fait qu'elle soit droitière ou gauchère (ou ambidextre).

On teste l'alternative suivante.

H_0 : le sexe d'une personne est indépendant du fait qu'elle soit gauchère ou droitière.
H_1 : il n'y a pas indépendance.

Voici les données observées (effectifs mesurés) et les effectifs théoriques (calculés à partir de ces mesures sous l'hypothèse H_0).

effectifs mesurés

	femmes	hommes	totaux
droitiers	1070	934	2004
gauchers	92	113	205
ambidextres	8	20	28
totaux	1170	1067	2237

effectifs théoriques

	femmes	hommes	totaux
droitiers	1048	956	2004
gauchers	107	98	205
ambidextres	15	13	28
totaux	1170	1067	2237

Pour que l'approximation par la loi du chi-carré soit bonne, il faut que les effectifs théoriques soient plus grands ou égaux à 5. C'est le cas ici, et on a

$$d^2 = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(n_{i;j} - np_iq_j)^2}{np_iq_j} = \frac{(1070-1048)^2}{1048} + \frac{(934-956)^2}{956} + \cdots + \frac{(20-13)^2}{13} \cong 12$$

(sans arrondir les effectifs théoriques, on trouve environ 11.806)

Les paramètres p_i et q_j ont été estimés, on se retrouve avec $(3-1)(2-1) = 2$ degrés de liberté.

Comme $d^2 > \phi_2^{-1}(0.99) \cong 9.210$, on rejette l'hypothèse H_0 au seuil de signification de 1% (c'est-à-dire avec une probabilité de 1% de chance de rejeter à tort).

Les données laissent penser que les femmes sont plus souvent droitières que les hommes.

5. Il s'agit des estimateurs du maximum de vraisemblance, notés $\widehat{p_i}_{\text{MLE}}$ et $\widehat{q_j}_{\text{MLE}}$.

3.5 La p-valeur associée à un test d'hypothèse

La *p-valeur* est la probabilité sous l'hypothèse nulle H_0 que la statistique de test[6] soit au moins aussi "extrême" que la valeur observée à partir des données.

La signification du mot "extrême" dépend de la façon dont la probabilité indiquée par le seuil de signification α est définie dans le test d'hypothèse en question.

Exemple d'un test bilatéral avec une distribution symétrique

Rappelons que les distributions des lois normales, de Student et binomiales sont des distributions symétriques.

Dans le cas des tests bilatéraux où la statistique de test T possède une distribution symétrique, le seuil de signification α satisfait la condition suivante :

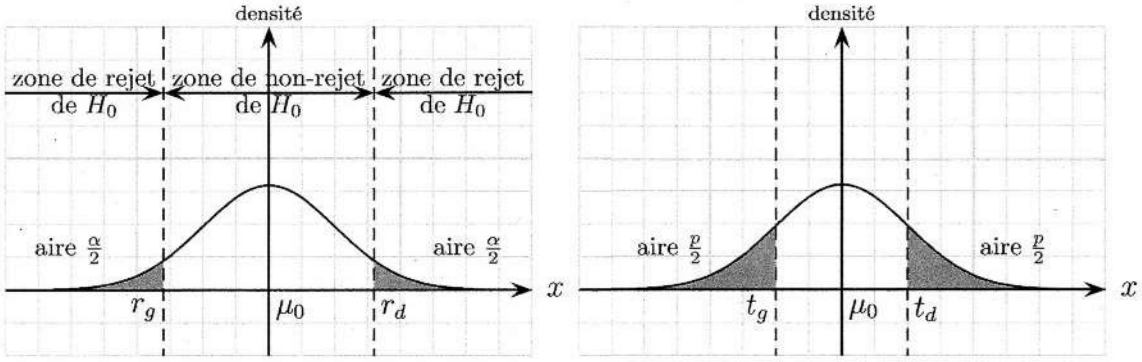
$$\alpha = \underbrace{\mathbb{P}(T < r_g)}_{\frac{\alpha}{2}} + \underbrace{\mathbb{P}(T > r_d)}_{\frac{\alpha}{2}} \quad \text{avec} \quad r_g = F^{-1}(\tfrac{\alpha}{2}) \quad \text{et} \quad r_d = F^{-1}(1 - \tfrac{\alpha}{2})$$

où $F(x)$ est la fonction de répartition associée à la statistique de test T. Les nombres r_g et r_d sont les limites de la zone de rejet de H_0.

De façon similaire, on définit la p-valeur par $\quad p = \underbrace{\mathbb{P}(T < t_g)}_{\frac{p}{2}} + \underbrace{\mathbb{P}(T > t_d)}_{\frac{p}{2}}$

où t_g est la valeur à gauche qui est obtenue par symétrie de t (observation de la statistique de test T) par rapport à l'axe de symétrie de la distribution de T et où t_d est la valeur symétrique de t à droite (l'une des deux est égale à t !).

Description graphique d'un tel test d'hypothèse (pour cette situation, H_0 ne peut pas être rejetée au seuil de signification α).



Utilité de la p-valeur

Par construction de la p-valeur, on rejette H_0 au seuil de signification α lorsque $p < \alpha$.
En pratique, on peut aussi utiliser le tableau suivant :

p-valeur	évidence contre H_0	seuil de signification associé
p-valeur $\geqslant 0.10$	négligeable	$10\% < \alpha$
$0.10 > p$-valeur $\geqslant 0.05$	faible	$5\% < \alpha \leqslant 10\%$
$0.05 > p$-valeur $\geqslant 0.01$	modérée	$1\% < \alpha \leqslant 5\%$
$0.01 > p$-valeur $\geqslant 0.001$	forte	$0.1\% < \alpha \leqslant 1\%$
$0.001 > p$-valeur	très forte	$\alpha \leqslant 0.1\%$

6. Les *statistiques de test* sont les variables aléatoires utilisées pour les tests d'hypothèses.

Chapitre 4

Intervalles de confiance

Dans une étude statistique, on peut se demander si les mesures observées peuvent correspondre à une certaine réalité. Par exemple, est-ce que notre série de jets d'une pièce de monnaie $(P; P; F; P; F; P)$ peut correspondre à une pièce bien équilibrée ?

Pour répondre à cette question, on cherche à estimer la probabilité p que la pièce montre pile. Si on arrive à établir que $p = \frac{1}{2}$, alors la pièce est parfaitement équilibrée. Malheureusement, il est extrêmement difficile, voire impossible, de montrer que $p = \frac{1}{2}$.

On peut tenter d'estimer p à l'aide d'un estimateur $\widehat{P}$, mais la probabilité $\mathbf{P}(\widehat{P} = p)$ que $\widehat{P}$ soit exactement égal au paramètre p est, en général, extrêmement petite, voire nulle[1]. Par conséquent, il est nécessaire de s'intéresser à la probabilité

$$\mathbf{P}\big(\widehat{P} \geqslant p - r \text{ et } \widehat{P} \leqslant p + r\big) = \mathbf{P}\big(\widehat{P} - r \leqslant p \leqslant \widehat{P} + r\big) = \mathbf{P}\big(\big[\widehat{P} - r, \widehat{P} + r\big] \ni p\big)$$

Cette probabilité s'approche de 1 au fur et à mesure que le nombre positif r devient grand. Néanmoins, plus r grandit, plus l'approximation de p par un intervalle perd en précision. Il faut donc trouver des compromis.

Définition

Un *intervalle de confiance* est un intervalle $[G, D]$ qui contient le paramètre θ à estimer avec une certaine probabilité β (G et D sont des variables aléatoires).

$$\mathbf{P}([G, D] \supset \theta) = \beta$$

On dit que β est le *seuil de confiance* ou encore la *probabilité de couverture*, tandis que $\alpha = 1 - \beta$ est appelé le *risque d'erreur*.

Construction d'un intervalle de confiance

Pour construire un intervalle de confiance pour le paramètre θ, on commence par choisir le seuil de confiance β désiré (ce qui est équivalent à choisir un risque d'erreur $\alpha = 1 - \beta$). Ensuite, il faut réussir à déterminer le nombre positif r correspondant en utilisant la distribution théorique de l'estimateur $\hat{\theta}$.

Comme d'habitude, on choisit un seuil de confiance de 95% dans une démarche infirmative, et un seuil de confiance de 99% dans une démarche confirmative.

1. De plus, la probabilité pour que deux estimations $\widehat{p}_1$ et $\widehat{p}_2$ de l'estimateur $\widehat{P}$ soient égales est aussi, en général, extrêmement faible.

4.1 L'intervalle de confiance sur une moyenne, variance connue

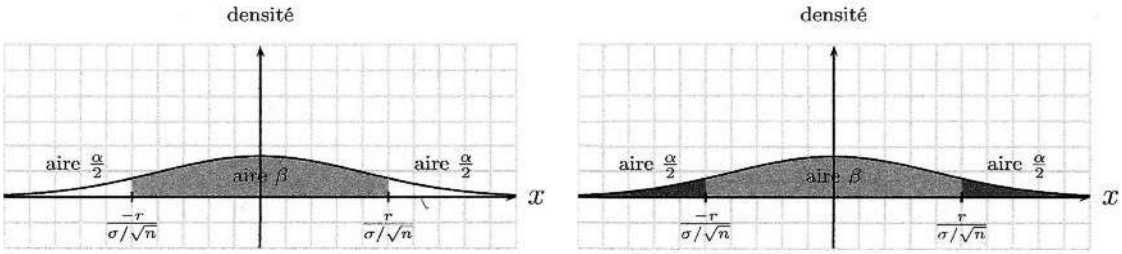
On considère n variables aléatoires indépendantes, notées X_1, X_2, ..., X_n, qui suivent toutes une loi normale $\mathcal{N}(\mu, \sigma^2)$ où seul le paramètre σ est connu. Pour estimer μ, on prend le MLE $\overline{X}$ qui est aussi sans biais. Sous ces hypothèses, $\overline{X}$ suit une loi normale de paramètres $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. Afin de déterminer l'intervalle de confiance, on va devoir centrer-réduire cet estimateur de sorte à pouvoir utiliser la table concernant la fonction de répartition ϕ de la loi normale $\mathcal{N}(0,1)$. On trouvera l'intervalle de confiance suivant :

$$\left[\overline{X} - \phi^{-1}(1-\tfrac{\alpha}{2}) \cdot \tfrac{\sigma}{\sqrt{n}} \ , \ \overline{X} + \phi^{-1}(1-\tfrac{\alpha}{2}) \cdot \tfrac{\sigma}{\sqrt{n}}\right]$$

En effet, on cherche $r > 0$ tel que :

$$\begin{aligned} 1-\alpha \ = \ \beta \ &= \ \mathbb{P}\big([\overline{X}-r, \overline{X}+r] \ni \mu\big) \ = \ \mathbb{P}\big(\overline{X}-r \leqslant \mu \leqslant \overline{X}+r\big) \\ &= \ \mathbb{P}\big(\overline{X} \geqslant \mu - r \ \text{ et } \ \overline{X} \leqslant \mu + r\big) \ = \ \mathbb{P}\big(\mu - r \leqslant \overline{X} \leqslant \mu + r\big) \\ &= \ \mathbb{P}\big(-r \leqslant \overline{X} - \mu \leqslant r\big) \ = \ \mathbb{P}\big(\tfrac{-r}{\frac{\sigma}{\sqrt{n}}} \leqslant \tfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leqslant \tfrac{r}{\frac{\sigma}{\sqrt{n}}}\big) \end{aligned}$$

On s'est maintenant ramené à la loi normale centrée réduite Z.



On peut ensuite utiliser la fonction de répartition ϕ de la loi normale centrée réduite.

$$\begin{aligned} 1-\alpha \ &= \ \mathbb{P}\big(\tfrac{-r}{\sigma/\sqrt{n}} \leqslant Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big) \ = \ \mathbb{P}\big(Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big) - \mathbb{P}\big(Z < \tfrac{-r}{\sigma/\sqrt{n}}\big) \\ &= \ \mathbb{P}\big(Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big) - \underbrace{\Big(1 - \mathbb{P}\big(Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big)\Big)}_{} \ = \ 2\mathbb{P}\big(Z \leqslant \tfrac{r}{\sigma/\sqrt{n}}\big) - 1 \\ &= \ 2\phi\big(\tfrac{r}{\sigma/\sqrt{n}}\big) - 1 \qquad \text{car } \mathbf{P}\big(Z < \tfrac{-r}{\sigma/\sqrt{n}}\big) = \mathbf{P}\big(Z > \tfrac{r}{\sigma/\sqrt{n}}\big) \text{ par symétrie} \end{aligned}$$

Finalement, on isole r :

$$1-\alpha \ = \ 2\phi\big(\tfrac{r}{\sigma/\sqrt{n}}\big) - 1 \iff 2\phi\big(\tfrac{r}{\sigma/\sqrt{n}}\big) \ = \ 2-\alpha \iff \phi\big(\tfrac{r}{\sigma/\sqrt{n}}\big) = 1 - \tfrac{\alpha}{2}$$

$$\iff \tfrac{r}{\sigma/\sqrt{n}} = \phi^{-1}(1-\tfrac{\alpha}{2}) \iff r = \phi^{-1}(1-\tfrac{\alpha}{2}) \cdot \tfrac{\sigma}{\sqrt{n}}$$

4.2 L'intervalle de confiance sur une moyenne, variance inconnue

On considère n variables aléatoires indépendantes, notées X_1, X_2, ..., X_n, qui suivent toutes une loi normale $\mathcal{N}(\mu, \sigma^2)$ où le paramètre σ est aussi inconnu. Pour estimer μ, on prend le MLE $\overline{X}$ qui est aussi sans biais. Sous ces hypothèses, $\overline{X}$ suit une loi normale de paramètres $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Malheureusement, comme on ne connaît pas σ, ce n'est pas très utile. Mais, en utilisant l'estimateur sans biais de σ, noté S, on peut utiliser le résultat de William Gosset : la variable aléatoire[2] $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ suit une distribution de Student avec $n-1$ degrés de liberté.

Afin de déterminer l'intervalle de confiance, on procède comme ci-dessus de sorte à pouvoir utiliser la table concernant la fonction de répartition ϕ_{n-1} de la loi de Student à $n-1$ degrés de liberté.

On trouvera l'intervalle de confiance suivant :

$$\left[\overline{X} - \phi_{n-1}^{-1}(1-\tfrac{\alpha}{2}) \cdot \tfrac{S}{\sqrt{n}} \ , \ \overline{X} + \phi_{n-1}^{-1}(1-\tfrac{\alpha}{2}) \cdot \tfrac{S}{\sqrt{n}}\right]$$

2. Ce n'est pas la variable centrée réduite de $\overline{X}$, car on a remplacé σ par son estimateur sans biais S.

Chapitre 5

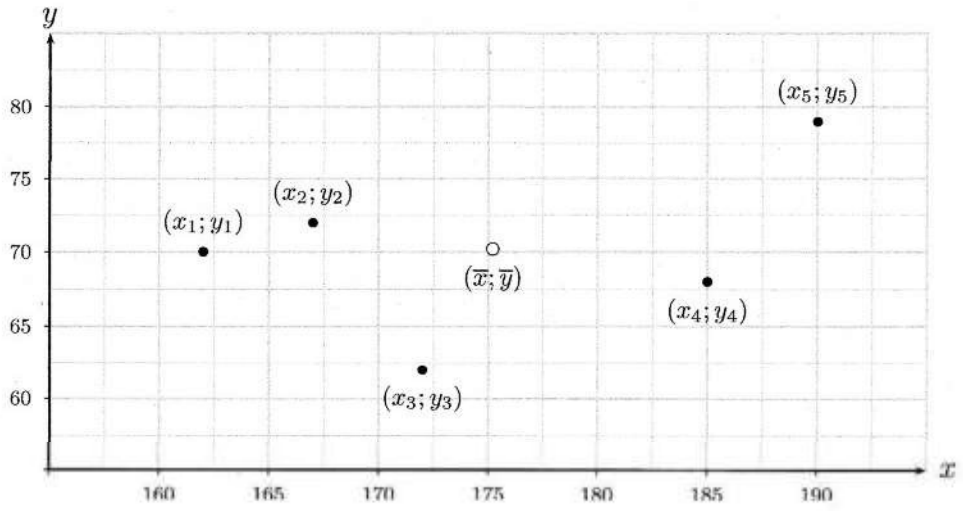
Régression linéaire

Supposons que l'on se donne deux caractéristiques X et Y sur une même population. Par exemple, on pourrait mesurer la taille en centimètres pour X et le poids en kilogrammes pour Y des lycéens de première année. Pour cela, on mesure les caractéristiques X et Y sur un échantillon aléatoire de taille n. On obtient ainsi des observations à deux coordonnées (x_i, y_i) pour $i \in \{1, 2, \ldots, n\}$. Pour cet exemple, prenons $n = 5$.

élève i	1	2	3	4	5
taille en cm x_i	162	167	172	185	190
poids en kg y_i	70	72	62	68	79

moyennes
$\overline{x} = 175.2$
$\overline{y} = 70.2$

La méthode la plus simple pour observer la relation entre X et Y est de représenter ces points dans le plan où l'axe horizontal représente la caractéristique X et l'axe vertical la caractéristique Y. Une telle représentation est appelée un *diagramme de dispersion.*



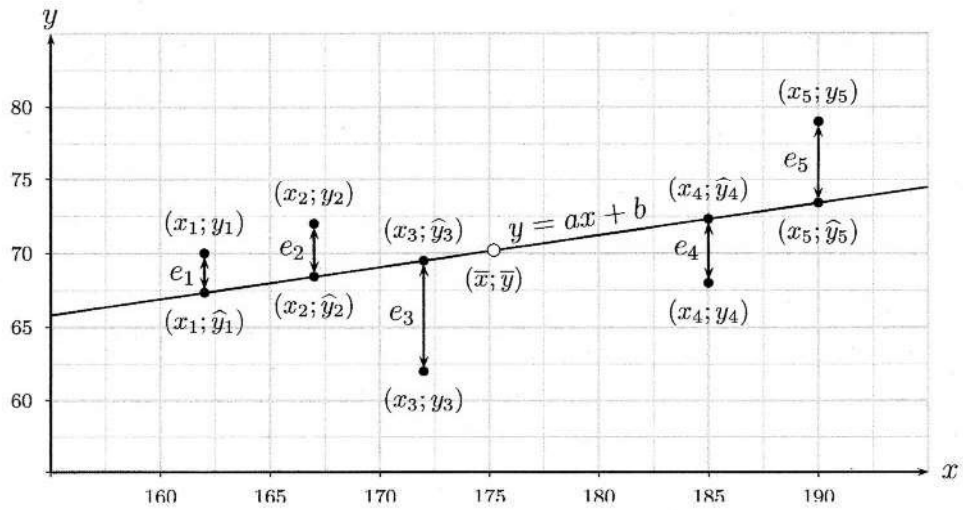
Si la relation entre X et Y est exacte, alors on devrait pouvoir trouver, pour une mesure x_i donnée, l'UNIQUE valeur pour y_i. Ainsi, Y serait une FONCTION de X $(y = f(x))$.

Malheureusement (ou heureusement), il se trouve que dans la plupart des cas, la relation n'est pas exacte (par exemple deux individus de même taille n'ont pas exactement le même poids). Néanmoins, même s'il n'y a pas de relation exacte, il se pourrait qu'il y ait une relation théorique et que, dans chaque mesure, il y ait une part aléatoire.

Dans un tel contexte, on dit que Y est la *variable dépendante*, et X est la *variable indépendante.*

Pour commencer, on va regarder s'il y a une chance pour que la relation (exacte ou non) entre X et Y soit affine (le graphe d'une fonction affine est une droite).

On va donc essayer de faire passer une droite "au mieux" parmi les points (x_i, y_i).



Les notations de la régression linéaire

y_i	y_i est la mesure effective associée à x_i.
$y = ax + b$	Il s'agit du modèle affine théorique entre les caractéristiques X et Y. Il faut déterminer les valeurs des bons paramètres a et b.
$\widehat{y_i} = ax_i + b$	$\widehat{y_i}$ est l'approximation théorique par le modèle affine associée à x_i.
$e_i = y_i - \widehat{y_i}$	e_i est l'erreur entre la mesure effective y_i et son approximation théorique $\widehat{y_i}$ associée à la i-ième mesure. Les e_i sont appelés les *résidus* associés au modèle.

On a ainsi.

$$e_i = y_i - \widehat{y_i} \iff y_i = \widehat{y_i} + e_i \iff y_i = \overbrace{ax_i + b}^{\text{modèle linéaire}} + e_i$$

5.1 La droite des moindres carrés

Dans ce cas le modèle théorique $y = ax + b$ est construit de manière à ce que la somme des carrés des résidus soit la plus petite possible.

Autrement dit, on veut a et b tels que la somme $\sum_{i=1}^{n} e_i^2$ soit minimale.

Pourquoi les carrés

1. L'élévation au carré néglige les signes. Ainsi une erreur négative ne sera pas compensée par une erreur positive.
2. L'élévation au carré réduit les petits écarts (car $(\frac{1}{2})^2 = \frac{1}{4}$; $(\frac{1}{4})^2 = \frac{1}{16}$) et amplifie les grands écarts (car $2^2 = 4$; $4^2 = 16$).

Bien sûr, il existe d'autres méthodes, comme la méthode des moindres valeurs absolues où l'on cherche les paramètres a et b qui minimisent $\sum_{i=1}^{n} |e_i|$. Mais les calculs associés à cette méthode sont plus compliqués.

Un résultat pratique

Les deux sommes suivantes sont égales.

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \sum_{i=1}^{n} x_i y_i \;-\; n\,\overline{x}\,\overline{y}$$

Preuve

On commence par développer la somme de gauche (les sommes ayant toujours les mêmes indices, on les notera juste $\sum$). À la fin du calcul, on utilise $n\overline{x} = \sum x_i$ et $n\overline{y} = \sum y_i$.

$$\begin{aligned}
\sum(x_i-\overline{x})(y_i-\overline{y}) &= \sum(x_iy_i - \overline{x}y_i - x_i\overline{y} + \overline{x}\,\overline{y}) \\
&\overset{\sum_1}{=} \sum x_iy_i \;-\; \sum \overline{x}y_i \;-\; \sum x_i\overline{y} \;+\; \sum \overline{x}\,\overline{y} \\
&\overset{\sum_2}{=} \sum x_iy_i \;-\; \overline{x}\sum y_i \;-\; \overline{y}\sum x_i \;+\; \overline{x}\,\overline{y}\sum 1 \\
&\overset{\sum_3}{=} \sum x_iy_i \;-\; \overline{x}\,n\,\overline{y} \;-\; \overline{y}\,n\,\overline{x} \;+\; n\,\overline{x}\,\overline{y} \\
&= \sum x_iy_i \;-\; n\,\overline{x}\,\overline{y}
\end{aligned}$$

Notations

On va ainsi noter[1] chacune de ces deux sommes σ_{XY}.

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \sigma_{XY} = \sum_{i=1}^{n} x_iy_i \;-\; n\,\overline{x}\,\overline{y}$$

En remplaçant Y par X ou X par Y, on a les formules suivantes.

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 = \sigma_{XX} = \sum_{i=1}^{n} x_i^2 \;-\; n\,\overline{x}^{\,2} \quad \text{et} \quad \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sigma_{YY} = \sum_{i=1}^{n} y_i^2 \;-\; n\,\overline{y}^{\,2}$$

Théorème des moindres carrés

Les valeurs des paramètres a et b pour la droite des moindres carrés $y = ax + b$ sont :

$$\boxed{a = \frac{\sigma_{XY}}{\sigma_{XX}}} \quad \text{et} \quad \boxed{b = \overline{y} - a\,\overline{x}}$$

Ces formules ne sont valables que s'il existe au moins deux x_i qui ont des valeurs différentes (sinon, il y a une division par zéro dans la formule pour a).

Conséquences graphiques

Dans la section 5.6, on montre deux propriétés graphiquement intéressantes.

1. la droite des moindres carrés $y = ax + b$ passe par $(\overline{x};\overline{y})$, qui est le *centre de gravité* des points $(x_i;y_i)$. En effet, on a la relation $\overline{y} = a\overline{x} + b$.
2. la droite des moindres carrés $y = ax + b$ est telle que la somme des résidus est nulle. En effet, la relation $\sum \widehat{y}_i = \sum y_i$ est équivalente à $\sum e_i = 0$.

Ainsi, la droite de régression est agréable à regarder.

---

1. Par rapport aux notations en probabilités, on a $\sigma_{XX} = n\sigma^2(X)$ et $\sigma_{YY} = n\sigma^2(Y)$ où $\sigma(X)$ et $\sigma(Y)$ représentent les écarts types respectifs de X et de Y. De même, $\sigma_{XY} = n\mathrm{Cov}(X,Y)$ où $\mathrm{Cov}(X,Y)$ est la covariance de X et de Y. De plus, si on travaille avec des échantillons, les estimateurs de la variance et de la covariance sont sans biais lorsqu'on remplace n par $(n-1)$.

5.2 Le coefficient de corrélation

On est maintenant capable de faire passer "au mieux" une droite parmi un nuage de points selon la méthode des moindres carrés. Cela ne nous dit toujours pas s'il y a une relation (ne serait-ce que linéaire) entre les caractères X et Y. Pour cela, les mathématiciens ont inventé un outil : il s'agit du *coefficient de corrélation* défini par

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_{XX}}\sqrt{\sigma_{YY}}}$$

Propriétés de ce coefficient

Le coefficient de corrélation est toujours compris entre -1 et 1.

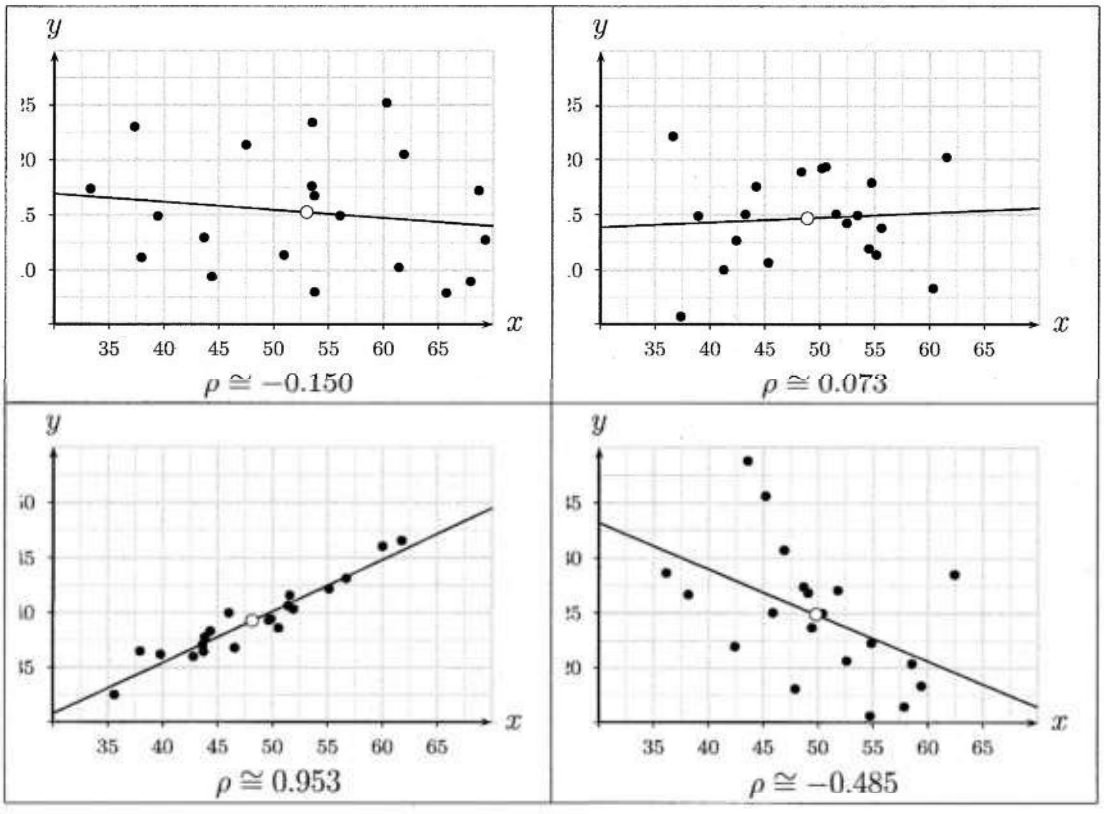
$$-1 \leqslant \rho \leqslant 1$$

C'est un outil qui permet de mesurer si la relation est linéaire, presque linéaire ou très peu linéaire.

- Lorsque ρ est proche de 0, alors la relation n'est pas linéaire. Soit les variables sont indépendantes, soit il y a un autre type de relation (voir page 72).
- Plus ρ est proche de 1, plus les points sont proches de la droite des moindres carrés (qui sera de pente positive).
- Plus ρ est proche de -1, plus les points sont proches de la droite des moindres carrés (qui sera de pente négative).

Moralité Plus $|\rho|$ est proche de 1, meilleure est l'approximation par la droite des moindres carrés. On dit alors que X et Y sont *corrélés* (ou *linéairement dépendants*).

Exemples



5.3 Le coefficient de détermination

Relation évidente

Les valeurs y_i et $\overline{y}$ proviennent directement des données. Les valeurs $\widehat{y}_i$ sont obtenues à partir du modèle. Ces trois valeurs sont liées par la relation suivante.

$$\overbrace{\underbrace{(y_i - \widehat{y}_i)}_{\text{résidu}} + (\widehat{y}_i - \overline{y})}^{\text{dépend aussi du modèle}} = \overbrace{(y_i - \overline{y})}^{\text{ne dépend que des données}}$$

Relation «miraculeuse»

Cette relation est vraie lorsqu'on utilise le modèle de la droite des moindres carrés.

$$\underbrace{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{\substack{\text{variation due} \\ \text{aux résidus}}} + \underbrace{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}_{\substack{\text{variation due} \\ \text{au modèle}}} = \underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{\substack{\text{variation totale} \\ \text{(car somme des} \\ \text{deux autres)}}}$$

Définition du coefficient de détermination

Le coefficient de détermination, noté R^2, est défini par

$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Propriétés de ce coefficient

Lorsque la relation «miraculeuse» est vraie, le coefficient de détermination détermine le rapport entre la variation due au modèle et la variation totale. C'est donc le pourcentage de la variation due au modèle dans la variation totale.

Pour cette raison, le coefficient de détermination est toujours compris entre 0 et 1.

$$0 \leqslant R^2 \leqslant 1$$

- Lorsque R^2 est proche de 0, près du 100% de la variation totale est expliquée par la variation due aux résidus. Cela signifie que le modèle n'est pas adapté aux données.
- Lorsque R^2 est proche de 1, près du 100% de la variation totale est expliquée par la variation due au modèle. Cela signifie que le modèle est bien adapté aux données.

Théorème de retrouvailles

Lorsque le modèle est celui de la droite des moindres carrés, le coefficient de détermination est égal au carré du coefficient de corrélation. Autrement dit

$$R^2 = \rho^2$$

5.4 La droite des moindres carrés forcée à l'origine

Dans ce cas le modèle théorique $y = ax + b$ est construit de manière à ce que

1. la droite passe par l'origine du plan ;
2. la somme des carrés des résidus soit la plus petite possible.

Autrement dit, on veut a et b tels que

1. $b = 0$, ainsi le modèle est $y = ax$.
2. la somme $\sum_{i=1}^{n} e_i^2$ soit minimale.

Théorème des moindres carrés pour la version forcée à l'origine

La valeur du paramètre a pour la droite des moindres carrés $y = ax$ est :

$$\boxed{a = \frac{\sigma_{XY}^{(0)}}{\sigma_{XX}^{(0)}}} \qquad \text{où} \quad \sigma_{XY}^{(0)} = \sum_{i=1}^{n} x_i y_i \quad \text{et} \quad \sigma_{XX}^{(0)} = \sum_{i=1}^{n} x_i^2$$

Cette formule n'est valable que s'il existe au moins deux x_i qui ont des valeurs différentes (sinon, il y a une division par zéro dans la formule pour a).

Ennuis

1. la droite des moindres carrés $y = ax$ ne passe pas forcément par $(\overline{x}; \overline{y})$, qui est le *centre de gravité* des points $(x_i; y_i)$.
2. la droite des moindres carrés $y = ax$ ne vérifie pas forcément la relation $\sum \widehat{y}_i = \sum y_i$, et ainsi la somme des résidus $\sum e_i$ n'est pas forcément nulle.

Ainsi, à l'œil, cette droite peut paraître un peu bizarre.

Exemple

Lors d'une expérience d'osmose en biologie, un arbre chimique grandit dans une solution à 30% de saccharose. On cherche à déterminer la vitesse moyenne de croissance de l'arbre chimique durant les 15 premières minutes qui sera la pente de la droite des moindres carrés. Lors des mesures, des élèves ont obtenus les nombres suivants.

temps (min)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
hauteur (cm)	0.0	1.5	2.9	4.3	5.5	6.6	7.8	9.0	10.3	11.4	12.5	13.7	14.7	15.8	16.8	17.7

On trouve

$$a \cong 1.23 \text{ cm} \cdot \text{min}^{-1}$$

Les relations des pages 68 et 69 ne sont pas toujours vraies dans le cas du modèle $y = ax$. L'exemple précédent infirme chacune de ces formules.

Néanmoins, on peut les retrouver si, dans ces relations, on remplace $\overline{x}$ et $\overline{y}$ par 0, comme on le voit à la page suivante. Ce qui explique la notation avec les exposants $^{(0)}$.

Bien évidemment, si le centre de gravité $(\overline{x}; \overline{y})$ est l'origine du plan, les droites des moindres carrés $y = ax + b$ et le modèle qui force la droite à l'origine sont les mêmes.

5.4.1 Les coefficients de détermination et de corrélation

Relation évidente

$$\overbrace{\underbrace{(y_i - \widehat{y}_i)}_{\text{résidu}} + \widehat{y}_i}^{\text{dépend aussi du modèle}} = \overbrace{y_i}^{\text{ne dépend que des données}}$$

Relation «miraculeuse»

$$\underbrace{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{\substack{\text{variation due} \\ \text{aux résidus}}} + \underbrace{\sum_{i=1}^{n} \widehat{y}_i^2}_{\substack{\text{variation due} \\ \text{au modèle par} \\ \text{rapport à 0}}} = \underbrace{\sum_{i=1}^{n} y_i^2}_{\substack{\text{variation totale} \\ \text{(car somme des} \\ \text{deux autres)}}}$$

Les coefficients de détermination et de corrélations

$$R^2 = \frac{\sum_{i=1}^{n} \widehat{y}_i^2}{\sum_{i=1}^{n} y_i^2} \qquad \text{et} \qquad \rho = \frac{\sigma_{XY}^{(0)}}{\sqrt{\sigma_{XX}^{(0)}}\sqrt{\sigma_{YY}^{(0)}}} \qquad \text{où} \qquad \begin{aligned} \sigma_{XY}^{(0)} &= \sum_{i=1}^{n} x_i y_i \\ \sigma_{XX}^{(0)} &= \sum_{i=1}^{n} x_i^2 \\ \sigma_{YY}^{(0)} &= \sum_{i=1}^{n} y_i^2 \end{aligned}$$

Théorème de retrouvailles

Même dans ce modèle où la droite des moindres carrés est forcée à l'origine, on a

$$R^2 = \rho^2$$

5.4.2 Preuves

Même si on a perdu l'ingrédient $\widehat{y}_i = \sum y_i$. On conserve l'ingrédient $\sum \widehat{y}_i^2 = \sum \widehat{y}_i y_i$. En effet

$$\sum_{i=1}^{n} \widehat{y}_i^2 = \sum_i (a x_i)^2 = a^2 \sum_i x_i^2 = \left(\frac{\sum_i x_i y_i}{\sum_i x_i^2} \right)^2 \sum_i x_i^2 = \frac{\left(\sum_k x_k y_k\right)\left(\sum_i x_i y_i\right)}{\sum_k x_k^2}$$

$$= \sum_i \left(\frac{\sum_k x_k y_k}{\sum_k x_k^2} x_i y_i \right) = \sum_i \left(\underbrace{a x_i}_{\widehat{y}_i} y_i \right) = \sum_{i=1}^{n} \widehat{y}_i y_i$$

Preuve de la relation «miraculeuse»

On peut maintenant prouver la relation «miraculeuse».

$$\sum (y_i - \widehat{y}_i)^2 + \sum \widehat{y}_i^2 = \sum y_i^2 - 2\sum y_i \widehat{y}_i + \sum \widehat{y}_i^2 + \sum \widehat{y}_i^2$$

$$\overset{\text{ingrédient}}{=} \sum y_i^2 - 2\sum \widehat{y}_i^2 + \sum \widehat{y}_i^2 + \sum \widehat{y}_i^2 = \sum y_i^2$$

Preuve du théorème de retrouvailles

$$R^2 = \frac{\sum \widehat{y}_i^2}{\sum y_i^2} = \frac{\sum (a x_i)^2}{\sum y_i^2} = a^2 \cdot \frac{\sum x_i^2}{\sum y_i^2} = \left(\frac{\sigma_{XY}^{(0)}}{\sigma_{XX}^{(0)}} \right)^2 \cdot \frac{\sigma_{XX}^{(0)}}{\sigma_{YY}^{(0)}} = \frac{{\sigma_{XY}^{(0)}}^2}{\sigma_{XX}^{(0)} \sigma_{YY}^{(0)}} = \rho^2$$

5.5 Autres types de régression

5.5.1 Préambule : une autre vision de la régression linéaire

On suppose que les données sont liées par une relation linéaire, non exacte, de la façon suivante.

$$y_i = ax_i + b + e_i$$

En suivant la méthode des moindres carrés, on trouve le minimum de $\sum_{i=1}^{n} e_i^2$ en annulant le gradient [2] : cela revient à résoudre le système suivant d'inconnues a et b.

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

5.5.2 Régression quadratique

On suppose que les données sont liées par une relation quadratique, non exacte, de la façon suivante.

$$y_i = ax_i^2 + bx_i + c + e_i$$

En suivant la méthode des moindres carrés, on trouve le minimum de $\sum_{i=1}^{n} e_i^2$ en annulant le gradient[2] : cela revient à résoudre le système suivant d'inconnues a, b et c.

$$\begin{cases} \sum_{i=1}^{n} y_i x_i^2 &= a \sum_{i=1}^{n} x_i^4 &+ b \sum_{i=1}^{n} x_i^3 &+ c \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} y_i x_i &= a \sum_{i=1}^{n} x_i^3 &+ b \sum_{i=1}^{n} x_i^2 &+ c \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i &= a \sum_{i=1}^{n} x_i^2 &+ b \sum_{i=1}^{n} x_i &+ c\, n \end{cases}$$

5.5.3 Régression hyperbolique

On suppose que les données sont liées par une relation hyperbolique, non exacte, de la façon suivante.

$$y_i = \frac{1}{ax_i + b + e_i} \iff \frac{1}{y_i} = ax_i + b + e_i$$

La méthode des moindres carrés peut s'appliquer [3] en utilisant le changement de variable suivant :

$$z = \frac{1}{y} \iff y = \frac{1}{z} \quad \text{et} \quad z_i = \frac{1}{y_i} \iff y_i = \frac{1}{z_i}$$

On trouve que :

$$y = \frac{1}{ax + b} \quad \left(\iff z = ax + b \right) \qquad \text{où} \quad \boxed{a = \frac{\sigma_{XZ}}{\sigma_{XX}}} \quad \text{et} \quad \boxed{b = \overline{z} - a\,\overline{x}}$$

2. Le gradient est une notion vue à l'université qui ne peut être expliquée qu'en deuxième année de lycée (il faut savoir dériver).

3. La *vraie* méthode des moindres carrés consisterait à minimiser la somme des carrés des ε_i donnés par la relation $y_i = \frac{1}{ax_i+b} + \varepsilon_i$. Néanmoins les calculs sont ici très complexes.

5.5.4 Régression exponentielle

On suppose que les données sont liées par une relation exponentielle, non exacte, de la façon suivante.

$$y_i = b \cdot a^{x_i} \cdot 10^{e_i} \iff \log(y_i) = \log(a)x_i + \log(b) + e_i$$

La méthode des moindres carrés peut s'appliquer[4] en utilisant le changement de variable suivant :

$$w = \log(y) \iff y = 10^w \qquad \text{et} \qquad w_i = \log(y_i) \iff y_i = 10^{w_i}$$

On trouve que :

$$\Big(\begin{array}{c} y = b \cdot a^x \\ \iff w = \log(a)x + \log(b) \end{array} \Big) \text{ où } \begin{array}{l} \log(a) = \dfrac{\sigma_{XW}}{\sigma_{XX}} \\ \Leftrightarrow \boxed{a = \exp_{10}\left(\dfrac{\sigma_{XW}}{\sigma_{XX}}\right)} \end{array} \text{ et } \begin{array}{l} \log(b) = \overline{w} - a\,\overline{x} \\ \Leftrightarrow \boxed{b = 10^{\overline{w} - a\,\overline{x}}} \end{array}$$

5.5.5 Régression d'une puissance

On suppose que les données sont liées par une puissance, non exacte, de la façon suivante.

$$y_i = b \cdot x_i^a \cdot 10^{e_i} \iff \log(y_i) = a\log(x_i) + \log(b) + e_i$$

La méthode des moindres carrés peut s'appliquer[5] en utilisant le changement de variable suivant :

$$\begin{array}{lll} w = \log(y) \iff y = 10^w & \text{et} & w_i = \log(y_i) \iff y_i = 10^{w_i} \\ & \text{et} & \\ v = \log(x) \iff x = 10^v & \text{et} & v_i = \log(x_i) \iff x_i = 10^{v_i} \end{array}$$

On trouve que :

$$\Big(\begin{array}{c} y = b \cdot x^a \\ \iff w = av + \log(b) \end{array} \Big) \text{ où } \boxed{a = \frac{\sigma_{VW}}{\sigma_{VV}}} \text{ et } \begin{array}{l} \log(b) = \overline{w} - a\,\overline{v} \\ \iff \boxed{b = 10^{\overline{w} - a\,\overline{v}}} \end{array}$$

5.5.6 Régression logarithmique

On suppose que les données sont liées par une relation logarithmique, non exacte, de la façon suivante.

$$y_i = a\log(x_i) + b + e_i$$

Cette fois la *vraie* méthode des moindres carrés peut s'appliquer en utilisant le changement de variable suivant :

$$v = \log(x) \iff x = 10^v \qquad \text{et} \qquad v_i = \log(x_i) \iff x_i = 10^{v_i}$$

On trouve que : $\Big(\begin{array}{c} y = a\log(x_i) + b \\ \iff y = av + b \end{array} \Big)$ où $\boxed{a = \frac{\sigma_{VY}}{\sigma_{VV}}}$ et $\boxed{b = \overline{y} - a\,\overline{v}}$

4. La *vraie* méthode des moindres carrés consisterait à minimiser la somme des carrés des ε_i donnés par la relation $y_i = b \cdot a^{x_i} + \varepsilon_i$. Néanmoins les calculs sont ici très complexes.

5. La *vraie* méthode des moindres carrés consisterait à minimiser la somme des carrés des ε_i donnés par la relation $y_i = b \cdot x_i^a + \varepsilon_i$. Néanmoins les calculs sont ici très complexes.

5.6 Preuves des théorèmes

5.6.1 Preuve des théorèmes des moindres carrés

Rappel sur les paraboles

Une parabole d'expression fonctionnelle $p(x) = \alpha x^2 + \beta x + \gamma$ avec $\alpha > 0$ a un minimum pour $x = -\frac{\beta}{2\alpha}$.

Preuve du théorème sur les moindres carrés

Trouvons une valeur de b tel que la somme des résidus au carré soit minimale. En d'autres termes, on veut trouver le minimum de l'expression suivante.

$$\sum_{i=1}^{n} e_i^2$$

En utilisant le fait que $\widehat{y_i} = ax_i + b$, on peut rendre cette somme dépendante du paramètre b. C'est pourquoi, cette somme est momentanément appelée $S(b)$.

$$\begin{aligned}
S(b) &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2 \\
&= \sum_{i=1}^{n} \big((y_i - ax_i) - b\big)^2 = \sum_{i=1}^{n} \big((y_i - ax_i)^2 - 2(y_i - ax_i)b + b^2\big) \\
&= \sum_{i=1}^{n} (y_i - ax_i)^2 - 2b \sum_{i=1}^{n} (y_i - ax_i) + \sum_{i=1}^{n} b^2 \\
&= \sum_{i=1}^{n} (y_i - ax_i)^2 - 2b \left(\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i \right) + \sum_{i=1}^{n} b^2
\end{aligned}$$

Or $\sum_{i=1}^{n} y_i = n\,\overline{y}$ et $\sum_{i=1}^{n} x_i = n\,\overline{x}$, ainsi on a

$$\begin{aligned}
S(b) &= \sum_{i=1}^{n} (y_i - ax_i)^2 - 2b\big(n\,\overline{y} - a\,n\,\overline{x}\big) + nb^2 \\
&= \underbrace{n}_{\alpha>0} b^2 \; \underbrace{-2n(\overline{y} - a\,\overline{x})}_{\beta} b \; + \; \underbrace{\sum_{i=1}^{n} (y_i - ax_i)^2}_{\gamma}
\end{aligned}$$

Par le rappel ci-dessus, la valeur de b qui minimise $S(b)$ est donnée par

$$b = \frac{2n(\overline{y} - a\,\overline{x})}{2n} = \overline{y} - a\,\overline{x}$$

Maintenant qu'on a établi la relation $b = \overline{y} - a\,\overline{x}$, l'expression de la droite des moindres carrés est ainsi devenue

$$y = ax + b = ax + \overline{y} - a\,\overline{x} = a(x - \overline{x}) + \overline{y}$$

Il faut maintenant trouver a tel que la somme des résidus au carré soit minimale. En d'autres termes, on veut trouver le minimum de l'expression suivante.

$$\sum_{i=1}^{n} e_i^2$$

En utilisant le fait que $\widehat{y}_i = ax_i + b = a(x_i - \overline{x}) + \overline{y}$, on peut rendre cette somme uniquement dépendante du paramètre a. C'est pourquoi, on a décidé d'appeler la somme $S(a)$.

$$S(a) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \Big(y_i - \big(a(x_i - \overline{x}) + \overline{y}\big)\Big)^2 = \sum_{i=1}^{n} \Big(y_i - a(x_i - \overline{x}) - \overline{y}\Big)^2$$

On cherche à trouver a tel que $S(a)$ est le plus petit possible. On sait que $S(a) \geqslant 0$ (car une somme de nombres positifs (ou nuls) ne peut être que positive (ou nulle)).
On a donc

$$\begin{aligned}
S(a) &= \sum_{i=1}^{n} \Big(y_i - a(x_i - \overline{x}) - \overline{y}\Big)^2 = \sum_{i=1}^{n} \Big((y_i - \overline{y}) - a(x_i - \overline{x})\Big)^2 \\
&= \sum_{i=1}^{n} \Big((y_i - \overline{y})^2 - 2a(x_i - \overline{x})(y_i - \overline{y}) + a^2(x_i - \overline{x})^2\Big) \\
&= \underbrace{\sum_{i=1}^{n} (y_i - \overline{y})^2}_{\sigma_{YY}} - 2a \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}_{\sigma_{XY}} + a^2 \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})^2}_{\sigma_{XX}} \\
&= \underbrace{\sigma_{XX}}_{>0} a^2 - 2\sigma_{XY} a + \sigma_{YY}
\end{aligned}$$

Par le rappel ci-dessus, la valeur de a qui minimise $S(a)$ est donnée par

$$a = \frac{2\sigma_{XY}}{2\sigma_{XX}} = \frac{\sigma_{XY}}{\sigma_{XX}}$$

Mais, il faut que $\sigma_{XX} > 0$ afin d'avoir un minimum. C'est le cas s'il y a au moins deux x_i qui sont différents. □

Preuve pour la version forcée à l'origine

Trouvons une valeur de a tel que la somme des résidus au carré soit minimale. En d'autres termes, on veut trouver le minimum de l'expression suivante.

$$\begin{aligned}
S(a) &= \sum_{i=1}^{n} \Big(y_i - ax_i\Big)^2 = \sum_{i=1}^{n} \Big(y_i^2 - 2ax_iy_i + a^2x_i^2\Big) = \underbrace{\sum_{i=1}^{n} y_i^2}_{\sigma_{YY}^{(0)}} - 2a \underbrace{\sum_{i=1}^{n} x_iy_i}_{\sigma_{XY}^{(0)}} + a^2 \underbrace{\sum_{i=1}^{n} x_i^2}_{\sigma_{XX}^{(0)}} \\
&= \underbrace{\sigma_{XX}^{(0)}}_{>0} a^2 - 2\sigma_{XY}^{(0)} a + \sigma_{YY}^{(0)}
\end{aligned}$$

Par le rappel ci-dessus, la valeur de a qui minimise $S(a)$ est donnée par

$$a = \frac{2\sigma_{XY}^{(0)}}{2\sigma_{XX}^{(0)}} = \frac{\sigma_{XY}^{(0)}}{\sigma_{XX}^{(0)}}$$

Mais, il faut que $\sigma_{XX}^{(0)} > 0$ afin d'avoir un minimum. C'est le cas s'il y a au moins deux x_i qui sont différents. □

5.6.2 Preuves de la relation «miraculeuse»

On a besoin de deux **ingrédients**.

$$\boxed{\sum \widehat{y}_i = \sum y_i \quad \text{et} \quad \sum \widehat{y}_i^2 = \sum \widehat{y}_i y_i} \ \bigstar$$

Si le modèle vérifie ces deux ingrédients, alors la relation «miraculeuse» est vraie.

En effet, on a

$$\begin{aligned}
& \sum \left(y_i - \widehat{y}_i\right)^2 \ + \ \sum \left(\widehat{y}_i - \overline{y}\right)^2 \\
= \ & \sum \left(y_i^2 - 2\widehat{y}_i y_i + \widehat{y}_i^2\right) \ + \ \sum \left(\widehat{y}_i^2 - 2\widehat{y}_i \overline{y} + \overline{y}^2\right) \\
\overset{\Sigma_1,\Sigma_2,\Sigma_3}{=} \ & \sum y_i^2 - 2\underbrace{\sum \widehat{y}_i y_i} + \sum \widehat{y}_i^2 \ + \ \sum \widehat{y}_i^2 - 2\overline{y}\underbrace{\sum \widehat{y}_i} + n\overline{y}^2 \\
\overset{\bigstar}{=} \ & \sum y_i^2 - 2\sum \widehat{y}_i^2 + \sum \widehat{y}_i^2 \ + \ \sum \widehat{y}_i^2 - 2\overline{y}\sum y_i + n\overline{y}^2 \\
\overset{\sum y_i = n\overline{y}}{=} \ & \sum y_i^2 - 2\sum \widehat{y}_i^2 + \sum \widehat{y}_i^2 \ + \ \sum \widehat{y}_i^2 - 2n\overline{y}^2 + n\overline{y}^2 \\
= \ & \sum y_i^2 - \sum \widehat{y}_i^2 \ + \ \sum \widehat{y}_i^2 - n\overline{y}^2 \\
= \ & \sum y_i^2 \ - \ n\overline{y}^2 \\
\overset{\text{notation } \sigma_{YY}}{\underset{\text{page 67}}{=}} \ & \sum \left(y_i - \overline{y}\right)^2
\end{aligned}$$

5.6.3 Les deux visions pour la droite de régression

À la page 540, on affirme que a et b satisfont le système suivant.

$$\left\{ \begin{array}{rcl} \sum y_i x_i & = & a\sum x_i^2 \ + \ b\sum x_i \\ \sum y_i & = & a\sum x_i \ + \ b\,n \end{array} \right. \iff \left\{ \begin{array}{rcl} \sum y_i x_i & = & a\sum x_i^2 \ + \ b\,n\,\overline{x} \\ n\,\overline{y} & = & a\,n\,\overline{x} \ + \ b\,n \end{array} \right.$$

À la page 536, on a donné les valeurs suivantes de a et b.

$$a = \frac{\sigma_{XY}}{\sigma_{XX}} \quad \text{et} \quad b = \overline{y} - a\,\overline{x}$$

On retrouve ces coefficients en résolvant le système d'équation. En effet, la deuxième ligne est équivalente à

$$\overline{y} = a\overline{x} + b \iff b = \overline{y} - a\overline{x}$$

De plus si, à la première ligne, on soustrait $\overline{x}$ fois la deuxième, cette première ligne devient

$$\begin{aligned}
& \sum y_i x_i - n\,\overline{x}\,\overline{y} = a\sum x_i^2 - a\,n\,\overline{x}^2 \iff \sum y_i x_i - n\,\overline{x}\,\overline{y} = a\Big(\sum x_i^2 - n\,\overline{x}^2\Big) \\
& \iff \sigma_{XY} = a\,\sigma_{XX} \iff a = \frac{\sigma_{XY}}{\sigma_{XX}}
\end{aligned}$$

5.6.4 Preuve des ingrédients pour le modèle linéaire

On se rappelle que a et b sont solutions du système

$$(\bigstar): \left\{ \begin{array}{rcl} \sum y_i x_i & = & a \sum x_i^2 \; + \; b \sum x_i \\ \sum y_i & = & a \sum x_i \; + \; b\, n \end{array} \right.$$

Preuve du premier ingrédient

On utilise le fait que $\widehat{y}_i = ax_i + b$, on développe et on observe le système.

$$\sum \widehat{y}_i = \sum \big(ax_i + b\big) = a \sum x_i + b\, n \overset{(\bigstar)}{=} \sum y_i$$

Preuve du deuxième ingrédient

On utilise le fait que $\widehat{y}_i = ax_i + b$, on développe et on observe le système.

$$\begin{aligned} \sum \widehat{y}_i^2 &= \sum \big(ax_i + b\big)^2 = \sum \big(a^2 x_i^2 + 2abx_i + b^2\big) = a^2 \sum x_i^2 + 2ab \sum x_i + b^2 n \\ &= a\big(a \sum x_i^2 + b \sum x_i\big) + ab \sum x_i + b^2 n \\ &= a\big(a \sum x_i^2 + b \sum x_i\big) + b\big(a \sum x_i + b\, n\big) \\ &\overset{(\bigstar)}{=} a \sum y_i x_i + b \sum y_i \\ &= \sum y_i (ax_i + b) \\ &= \sum y_i \widehat{y}_i \; = \; \sum \widehat{y}_i y_i \end{aligned}$$

5.6.5 Preuve du théorème de retrouvailles

On se rappelle que et $a = \dfrac{\sigma_{XY}}{\sigma_{XX}}$ et $b = \overline{y} - a\,\overline{x}$.

$$\begin{aligned} R^2 &= \frac{\sum \big(\widehat{y}_i - \overline{y}\big)^2}{\sum \big(y_i - \overline{y}\big)^2} = \frac{\sum \big(ax_i + b - \overline{y}\big)^2}{\sum \big(y_i - \overline{y}\big)^2} = \frac{\sum \big(ax_i + \overline{y} - a\,\overline{x} - \overline{y}\big)^2}{\sum \big(y_i - \overline{y}\big)^2} \\ &= \frac{\sum \big(ax_i - a\,\overline{x}\big)^2}{\sum \big(y_i - \overline{y}\big)^2} = a^2 \cdot \frac{\sum \big(x_i - \overline{x}\big)^2}{\sum \big(y_i - \overline{y}\big)^2} = \left(\frac{\sigma_{XY}}{\sigma_{XX}}\right)^2 \cdot \frac{\sigma_{XX}}{\sigma_{YY}} = \frac{\sigma_{XY}^2}{\sigma_{XX}^2} \cdot \frac{\sigma_{XX}}{\sigma_{YY}} \\ &= \frac{\sigma_{XY}^2}{\sigma_{XX}\sigma_{YY}} = \left(\frac{\sigma_{XY}}{\sqrt{\sigma_{XX}}\sqrt{\sigma_{YY}}}\right)^2 = \rho^2 \end{aligned}$$

5.6.6 Preuve des ingrédients pour le modèle quadratique

On se rappelle que a, b et c sont solutions du système

$$(\bigstar): \begin{cases} \sum y_i x_i^2 &= a\sum x_i^4 &+ b\sum x_i^3 &+ c\sum x_i^2 \\ \sum y_i x_i &= a\sum x_i^3 &+ b\sum x_i^2 &+ c\sum x_i \\ \sum y_i &= a\sum x_i^2 &+ b\sum x_i &+ c\,n \end{cases}$$

Preuve du premier ingrédient

On utilise le fait que $\widehat{y}_i = ax_i^2 + bx_i + c$, on développe et on observe le système.

$$\sum \widehat{y}_i = \sum \left(ax_i^2 + bx_i + c\right) = a\sum x_i^2 + b\sum x_i + c\,n \overset{(\bigstar)}{=} \sum y_i$$

Preuve du deuxième ingrédient

On utilise le fait que $\widehat{y}_i = ax_i^2 + bx_i + c$, on développe et on observe le système.

$$\begin{aligned}
\sum \widehat{y}_i^2 &= \sum \left(ax_i^2 + bx_i + c\right)^2 \\
&= \sum \left(a^2x_i^4 + 2abx_i^3 + b^2x_i^2 + 2acx_i^2 + 2bcx_i + c^2\right) \\
&= a^2\sum x_i^4 + 2ab\sum x_i^3 + b^2\sum x_i^2 + 2ac\sum x_i^2 + 2bc\sum x_i + c^2 n \\
&= a\left(a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2\right) \\
&\quad + ab\sum x_i^3 + b^2\sum x_i^2 + ac\sum x_i^2 + 2bc\sum x_i + c^2 n \\
&= a\left(a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2\right) \\
&\quad + b\left(a\sum x_i^3 + b\sum x_i^2 + c\sum x_i\right) \\
&\quad + ac\sum x_i^2 + bc\sum x_i + c^2 n \\
&= a\left(a\sum x_i^4 + b\sum x_i^3 + c\sum x_i^2\right) \\
&\quad + b\left(a\sum x_i^3 + b\sum x_i^2 + c\sum x_i\right) \\
&\quad + c\left(c\sum x_i^2 + b\sum x_i + cn\right) \\
&\overset{(\bigstar)}{=} a\sum y_i x_i^2 + b\sum y_i x_i + c\sum y_i \\
&= \sum y_i(ax_i^2 + bx_i + c) \\
&= \sum y_i \widehat{y}_i \;=\; \sum \widehat{y}_i y_i
\end{aligned}$$