5UAA1 : Statistique 2 variables

1 Rappel vocabulaire statistique à une variable

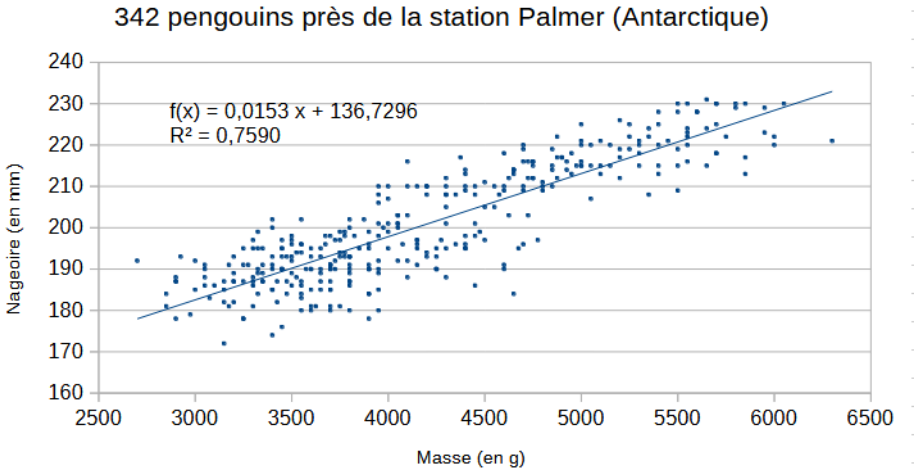
1.1 Population, individus, caractère

Une étude statistique porte sur un ensemble d'**individus** (parfois un échantillon de cet ensemble) appelé **population**.

Elle collecte un (statistique à une variable) ou plusieurs (statistique à plusieurs variables) caractères (propriétés, caractéristiques) que chaque individu possède.

Considérons la taille des nageoires et la masse de 342 pengouins près de la station Palmer, en Antarctique [1].

Traçons le **nuage de points** correspondant.



On note $n = 342$ le nombre d'individus, c'est à dire **l'effectif total** de la population observée.

1.2 Mesures de position, de dispersion

Nous avons défini en quatrième des grandeurs qui permettent de décrire les séries de données collectées.

Certaines donnent des informations sur la position : médiane, moyenne

D'autres donnent des informations sur la dispersion : écart type, ecart interquartile

La moyenne des masses des 342 pengouins : $\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_{342}}{342} = 4201{,}8\,\text{g}$

La variance des masses des 342 pengouins :

$\text{Var}(x) = \dfrac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_{342} - \overline{x})^2}{342} = 641\,250\,\text{g}^2$

L'écart type des masses des 342 pengouins : $\sigma(x) = \sqrt{\text{Var}(x)} = \sqrt{641250} = 800{,}8\,\text{g}$

Les masses se situent essentiellement dans l'intervalle $[\overline{x} - 2 * \sigma; \overline{x} + 2 * \sigma] = [2600\,\text{g}; 5803\,\text{g}]$

La moyenne des longueurs des nageoires des 342 pengouins : $\overline{y} = \dfrac{y_1 + y_2 + \cdots + y_{342}}{342} = 200{,}9\,\text{mm}$

1. Horst AM, Hill AP, Gorman KB (2020). palmerpenguins : Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi : 10.5281/zenodo.3960218. R for datascience https://r4ds.hadley.nz/data-visualize

La variance des longueurs des nageoires des 342 pengouins :
$\mathrm{Var}(y) = \dfrac{(y_1-\overline{y})^2+(y_2-\overline{y})^2+\cdots+(y_{342}-\overline{y})^2}{342} = 197{,}15\,\mathrm{mm}^2$

L'écart type des longueurs des nageoires des 342 pengouins : $\sigma(y) = \sqrt{\mathrm{Var}(y)} = \sqrt{197{,}2} = 14{,}04\,\mathrm{mm}$
La plus grande partie des longueurs se situent dans l'intervalle $[\overline{y}-2*\sigma;\overline{y}+2*\sigma] = [173\,\mathrm{mm};229\,\mathrm{mm}]$

1.3 Définition de la variance et de l'écart type

La variance $\mathrm{Var}(x)$ des x_i est la moyenne des carrés des écarts $x_i-\overline{x}$ à la moyenne $\overline{x}$.

$\mathrm{Var}(x) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$

L'**écart type** (de Pearson) des x_i, noté $\sigma(x)$ en quatrième, est la racine carrée de la variance des données observées : $\boxed{\sigma(x) = \sqrt{\mathrm{Var}(x)}}$.

1.4 Une formule très utile de la variance

La variance est égale à **la moyenne des carrés moins le carré de la moyenne** : $\boxed{\mathrm{Var}(x) = \overline{x^2}-\overline{x}^2}$

Démonstration :

$\mathrm{Var}(x) = \dfrac{1}{n}\sum_i(x_i-\overline{x})^2 = \dfrac{1}{n}\sum_i x_i^2 - 2\overline{x}\cdot\dfrac{1}{n}\sum_i x_i + \dfrac{1}{n}\sum_i\overline{x}^2 = \overline{x^2} - 2\overline{x}^2 + \dfrac{1}{n}\cdot n\cdot\overline{x}^2 = \overline{x^2}-\overline{x}^2$

2 Statistique à deux variables

2.1 Objectif

Observer le nuage constitué de n points de coordonnées $(x_i;y_i)$. Si une relation linéaire semble exister (les points ont l'air de se répartir sur une droite), trouver l'équation $y=ax+b$ de la droite qui approche ou « **ajuste** » au mieux ce nuage **au sens des moindres carrés**.

Remarque : un document annexe, disponible sous moodle, donne des détails pour justifier les formules données ici. Il permet également de calculer une incertitude expérimentale sur la pente a de la droite trouvée sous l'hypothèse qu'il existe vraiment une relation linéaire entre x et y. Le calcul de cette incertitude est par exemple très utile en sciences quand on effectue plusieurs mesures $(x_i;y_i)$ d'un phénomène physique (a est alors une mesure d'une grandeur que l'on cherche à déterminer, et on veut connaître la précision de cette mesure).
Dans l'exemple des 342 pengouins, on cherche juste à déterminer une tendance.

2.2 Méthode des moindres carrés

Soit $\boxed{y=ax+b}$ l'équation de la droite cherchée.
Il faut trouver sa pente a et son ordonnée à l'origine b afin de **minimiser la somme des carrés des écarts verticaux** :

$$\sum_{i=1}^{n}\left(y_i-(ax_i+b)\right)^2=\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2 \qquad \text{en notant : } \boxed{\widehat{y_i}=ax_i+b}$$

$\boxed{e_i=y_i-\widehat{y_i}}$ est **l'écart vertical entre le point $(x_i;y_i)$ du nuage et le point correspondant $(x_i;\widehat{y_i})$ sur la droite de régression (d'abscisse aussi égale à x_i)**.
L'écart vertical e_i est souvent appelé **écart résiduel** ou **résidu**.

$\boxed{e_i \text{ est l'écart entre la valeur observée } y_i \text{ et sa valeur estimée } \widehat{y_i}=ax_i+b \text{ connaissant } x_i}$

$\boxed{\text{La méthode des moindres carrés permet de minimiser } \sum_{i=1}^{n} e_i^2}$

2.3 Calculs de a et b ; Covariance de x et y ; Variables centrées

Un document annexe sur https://ophysis.net donne les justifications pour ceux qui sont intéressés.
$\boxed{a=\dfrac{\overline{xy}-\overline{x}\cdot\overline{y}}{\overline{x^2}-\overline{x}\cdot\overline{x}}}$ et $\boxed{b=\overline{y}-a\overline{x}}$.

On reconnaît au dénominateur de a la variance $\mathrm{Var}(x)=\overline{x^2}-\overline{x}^2=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$ des x_i.

On définit la covariance $\mathrm{Cov}(x;y)$ des x_i et y_i par $\boxed{\mathrm{Cov}(x;y)=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}$

Ainsi, par définition, la covariance $\mathrm{Cov}(x;y)$ des x_i et y_i est la moyenne du produit des **variables centrées** $\boxed{X_i=x_i-\overline{x}}$ et $\boxed{Y_i=y_i-\overline{y}}$.

Démontrer en exercice que $\dfrac{1}{n}\sum_i(x_i-\overline{x})\cdot(y_i-\overline{y})=\overline{xy}-\overline{x}\cdot\overline{y}$
La covariance est égale à **la moyenne des produits moins le produit des moyennes**.

2.4 Autres formules

$$\boxed{a=\frac{\overline{xy}-\overline{x}\cdot\overline{y}}{\overline{x^2}-\overline{x}\cdot\overline{x}}=\frac{\mathrm{Cov}(x;y)}{\mathrm{Var}(x)}=\frac{\sum_{i=1}^{n}X_i\cdot Y_i}{\sum_{i=1}^{n}X_i^2}}$$

La dernière égalité s'obtient en multipliant par n le numérateur et le dénominateur.

L'équation de la droite est $\boxed{y-\overline{y}=\dfrac{\mathrm{Cov}(x;y)}{\mathrm{Var}(x)}\cdot(x-\overline{x})}$ ou encore $\boxed{y=a\cdot x+b}$ avec $\boxed{b=\overline{y}-a\overline{x}}$

En effet, le point moyen $(\overline{x},\overline{y})$ est sur la droite de régression linéaire : $y=a\cdot x+b$.

2.5 Moyenne et variance observées des résidus ; Coefficient de corrélation

Connaissant x_i, on peut estimer y_i en calculant $\widehat{y_i}=ax_i+b=\overline{y}+\frac{\mathrm{Cov}(x;y)}{\mathrm{Var}(x)}\cdot(x_i-\overline{x})$.

$\boxed{\text{La moyenne observée des résidus } e_i=y_i-(a\cdot x_i+b) \text{ est nulle : } \overline{e}=0}$.

$\boxed{\text{La variance observée des } e_i \text{ est } \mathrm{Var}(e)=\overline{e^2}=\mathrm{Var}(y)\cdot(1-\mathrm{R}(x;y)^2)}$ avec : $\boxed{\mathrm{R}(x;y)=\dfrac{\mathrm{Cov}(x;y)}{\sigma(x)\cdot\sigma(y)}}$.

$\mathrm{R}(x;y)$ est appelé **coefficient de corrélation** entre x et y.

Comme $\mathrm{Var}(e)$ et $\mathrm{Var}(y)$ sont positifs, on en déduit que $1-\mathrm{R}(x;y)^2 \geq 0$, soit $\boxed{-1 \leq \mathrm{R}(x;y) \leq 1}$

De plus, si $\mathrm{R}(x;y)^2 = 1$ (donc $\mathrm{R}(x;y) = \pm 1$), alors $e_i = 0$ (puisque $\overline{e^2} = 0$) et donc $y_i = \widehat{y_i}$ pour les n points qui sont donc parfaitement alignés. La connaissance de x_i permet de déterminer la valeur de y_i grâce à la droite de régression.

2.6 Interprétation du coefficient de corrélation $\mathrm{R}(x;y)$

On remarque que :

$$a = \frac{\mathrm{Cov}(x;y)}{\mathrm{Var}(x)} = \frac{\mathrm{Cov}(x;y)}{\sigma(x)\,\sigma(x)} \cdot \frac{\sigma(y)}{\sigma(y)} = \frac{\mathrm{Cov}(x;y)}{\sigma(x)\,\sigma(y)} \cdot \frac{\sigma(y)}{\sigma(x)} = \mathrm{R}(x;y) \cdot \frac{\sigma(y)}{\sigma(x)}$$

Ainsi, si $\mathrm{R}(x;y) = 1$, la pente de la droite de régression est le rapport $\dfrac{s_y}{s_x}$, ce qui est cohérent avec la définition de la pente (rapport des variations en y sur les variations en x).

Si $\mathrm{R}(x;y)$=-1, la pente de la droite de régression est le rapport $-\dfrac{\sigma(y)}{\sigma(x)} < 0$.
De façon générale, si $\mathrm{R}(x;y) < 0$, alors $a < 0$: quand x_i augmente, y_i dimimue.
Si $\mathrm{R}(x;y) > 0$, alors $a > 0$: quand x_i augmente, y_i aussi.

3 Retour en Antarctique

On trouve que $\mathrm{Cov}(x;y) = 9795{,}7\,\text{mm}\cdot\text{g}$ et donc $a = \dfrac{\mathrm{Cov}(x;y)}{\mathrm{Var}(x)} = \dfrac{9795{,}7}{641250} = 0{,}0153\,\text{mm/g}$

Et donc, $b = \overline{y} - a\overline{x} = 200{,}9 - 0{,}0153 \cdot 4201{,}8 = 136{,}6\,\text{mm}$

et $\mathrm{R}(x;y) = \dfrac{\mathrm{Cov}(x;y)}{\sigma(x)\cdot\sigma(y)} = \dfrac{9795{,}7}{800{,}8 \cdot 14{,}04} = 0{,}871$ ce qui donne $\mathrm{R}(x;y)^2 = 0{,}759$

On retrouve ainsi les valeurs calculées par le tableur **Calc** de ***libre office*** lors du calcul de la courbe de tendance linéaire (voir le graphique en première page).
En moyenne, la longueur de la nageoire des pingouins augmente d'environ 1,5 mm par kilogramme supplémentaire.