

TP pluri n°1

Comparaison de 2 méthodes

BTSA Anabiotec

Yannick Le Bastard

LEGTA de l'Hérault

March 13, 2024



Bienvenue sur les
sentiers
mathématiques !

Ce premier TP se propose d'exploiter les résultats que vous avez obtenus en mesurant la concentration en glucose d'une solution via deux méthodes différentes :

- 1 La méthode de Bertrand (méthode de référence)
- 2 La méthode DNS (méthode alternative)

Ce premier TP se propose d'exploiter les résultats que vous avez obtenus en mesurant la concentration en glucose d'une solution via deux méthodes différentes :

- ① La méthode de Bertrand (méthode de référence)
- ② La méthode DNS (méthode alternative)

La mise en œuvre de ces méthodes, donne lieu à des erreurs de mesure de différentes natures.

Ce premier TP se propose d'exploiter les résultats que vous avez obtenus en mesurant la concentration en glucose d'une solution via deux méthodes différentes :

- ① La méthode de Bertrand (méthode de référence)
- ② La méthode DNS (méthode alternative)

La mise en œuvre de ces méthodes, donne lieu à des erreurs de mesure de différentes natures.

Pouvez-vous en citer quelques-unes ?

Ce premier TP se propose d'exploiter les résultats que vous avez obtenus en mesurant la concentration en glucose d'une solution via deux méthodes différentes :

- ① La méthode de Bertrand (méthode de référence)
- ② La méthode DNS (méthode alternative)

La mise en œuvre de ces méthodes, donne lieu à des erreurs de mesure de différentes natures.

Pouvez-vous en citer quelques-unes ?

Nous retiendrons que nous obtenons des couples de mesure portant sur le même individu statistique (votre solution) : nous parlerons d'**échantillons appariés**.

Présentation des résultats

La première étape, après recueil des données, est leur **représentation** (sous forme de tableau numérique ET de graphiques). Voici les concentrations de glucose mesurées en g/L par laboratoire (binômes) via les méthodes de Bertrand et DNS :

Présentation des résultats

La première étape, après recueil des données, est leur **représentation** (sous forme de tableau numérique ET de graphiques). Voici les concentrations de glucose mesurées en g/L par laboratoire (binômes) via les méthodes de Bertrand et DNS :

Labo	Bertrand	DNS
1	5,04	5,03
2	5,92	6,01
3	5,30	5,27
4	5,23	5,18
5	5,09	5,07
6	5,47	5,54
7	6,53	6,31
8	5,14	5,14
9	5,20	4,69
10	6,71	4,56

Présentation des résultats

Le logiciel Geogebra, tout comme le logiciel Python (que vous utiliserez) nous permettent de visualiser ces données séparément sous la forme d'un histogramme (attention, le nombre de classes modifie l'aspect du graphique !):

Présentation des résultats

Le logiciel Geogebra, tout comme le logiciel Python (que vous utiliserez) nous permettent de visualiser ces données séparément sous la forme d'un histogramme (attention, le nombre de classes modifie l'aspect du graphique !):

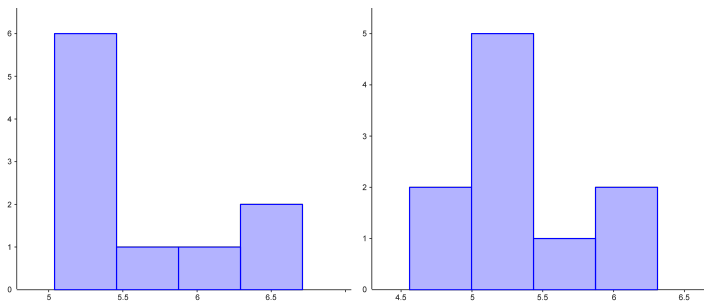


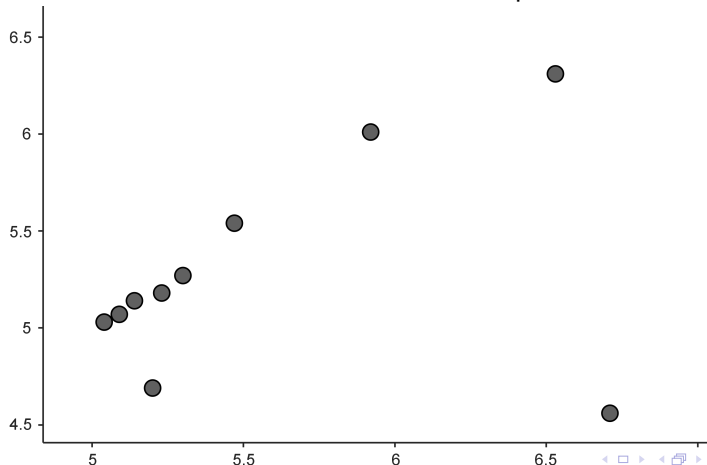
Figure: Concentrations relevées avec la méthode de Bertrand (à gauche) et la méthode DNS (à droite)

Présentation des résultats

Une autre façon de comparer les mesures obtenues par chacune des deux méthodes consiste à représenter le nuage des couples de valeurs observées à l'aide de chaque méthode (en abscisse la mesure obtenue par la méthode de Bertrand et en ordonnée celle obtenue par la méthode DNS):

Présentation des résultats

Une autre façon de comparer les mesures obtenues par chacune des deux méthodes consiste à représenter le nuage des couples de valeurs observées à l'aide de chaque méthode (en abscisse la mesure obtenue par la méthode de Bertrand et en ordonnée celle obtenue par la méthode DNS):



Rappelons que le but de notre étude est de comparer deux méthodes, dont une de référence et l'autre, alternative. On s'attend donc, dans le meilleur des cas, à pouvoir utiliser indifféremment l'une comme l'autre. Les mesures obtenues ne seront pas significativement différentes. On parle de **méthodes concordantes**.

Rappelons que le but de notre étude est de comparer deux méthodes, dont une de référence et l'autre, alternative. On s'attend donc, dans le meilleur des cas, à pouvoir utiliser indifféremment l'une comme l'autre. Les mesures obtenues ne seront pas significativement différentes. On parle de **méthodes concordantes**.

Notre première approche consistera donc à étudier les différences d_i entre la mesure du laboratoire i par la méthode de Bertrand x_i et celle par la méthode DNS y_i : $d_i = x_i - y_i$ et relever des différences trop importantes.

Repérage des valeurs aberrantes

Labo	Bertrand x_i	DNS y_i	Différence $d_i = x_i - y_i$
1	5,04	5,03	0,01
2	5,92	6,01	-0,09
3	5,30	5,27	0,03
4	5,23	5,18	0,05
5	5,09	5,07	0,02
6	5,47	5,54	-0,07
7	6,53	6,31	0,22
8	5,14	5,14	0,00
9	5,20	4,69	0,51
10	6,71	4,56	2,15
\bar{x}	5,56	5,28	0,28
\hat{s}	0,58	0,62	0,64

Classons les différences observées par ordre croissant :

−0,09 ; −0,07 ; 0,00 ; 0,01 ; 0,02 ; 0,03 ; 0,05 ; 0,22 ; 0,51 ; 2,15

Classons les différences observées par ordre croissant :

−0,09 ; −0,07 ; 0,00 ; 0,01 ; 0,02 ; 0,03 ; 0,05 ; 0,22 ; 0,51 ; 2,15

Les deux dernières valeurs semblent s'écarter nettement des précédentes.
Peut-on pour autant les éliminer avec un risque d'erreur de première espèce fixé à l'avance ?

Repérage des valeurs aberrantes

Une méthode graphique : **Bland et Altman**

On représente les couples $M_i((x_i + y_i)/2, x_i - y_i)$ encadré par deux "bandes de sécurité" d'équations respectives $y = \bar{d} - 2\hat{s}_d$ et $y = \bar{d} + 2\hat{s}_d$, où \hat{s}_d désigne l'écart-type corrigé des différences observées.

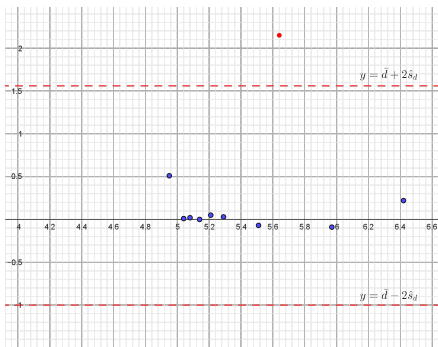
On considère comme aberrants les couples de valeurs (x_i, y_i) qui sont en dehors des bandes de sécurité.

Repérage des valeurs aberrantes

Une méthode graphique : **Bland et Altman**

On représente les couples $M_i((x_i + y_i)/2, x_i - y_i)$ encadré par deux "bandes de sécurité" d'équations respectives $y = \bar{d} - 2\hat{s}_d$ et $y = \bar{d} + 2\hat{s}_d$, où \hat{s}_d désigne l'écart-type corrigé des différences observées.

On considère comme aberrants les couples de valeurs (x_i, y_i) qui sont en dehors des bandes de sécurité.



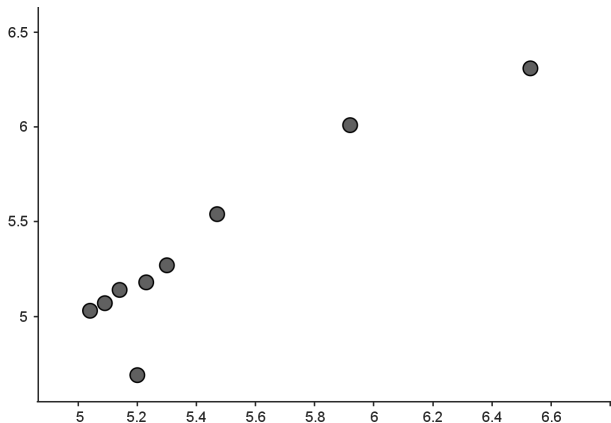
Repérage des valeurs aberrantes

On élimine donc le point figuré en rouge dont les coordonnées initiales sont $(x_i, y_i) = (6, 71; 4, 56)$.

Repérage des valeurs aberrantes

On élimine donc le point figuré en rouge dont les coordonnées initiales sont $(x_i, y_i) = (6, 71; 4, 56)$.

Le nouveau nuage de points *toiletté* est donc :



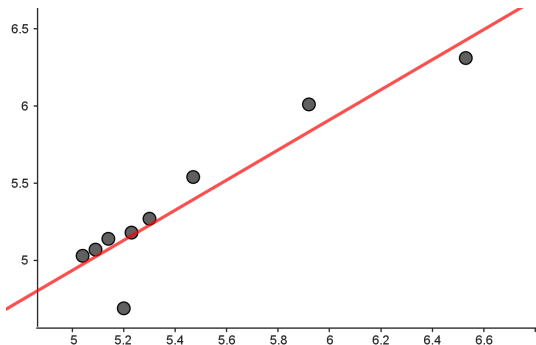
Repérage des valeurs aberrantes

Il semble encore, au vu du nuage de points, que l'un d'entre eux s'écarte manifestement des autres. Traitons son cas à l'aide du test de Dixon (cf polycopié).

Repérage des valeurs aberrantes

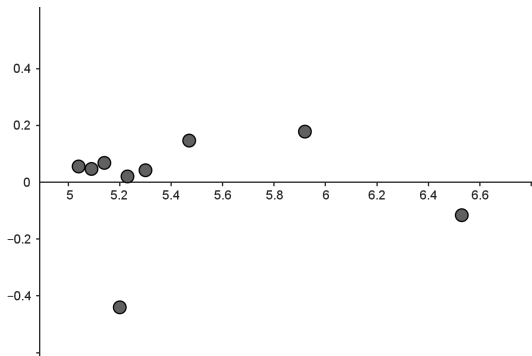
Il semble encore, au vu du nuage de points, que l'un d'entre eux s'écarte manifestement des autres. Traitons son cas à l'aide du test de Dixon (cf polycopié).

L'équation de la droite de régression de Y (DNS) en X (Bertrand) s'écrit : $\hat{y} = 0,9742x + 0,0649$. Et un coefficient de détermination $r^2 = 0,8694$.



Repérage des valeurs aberrantes

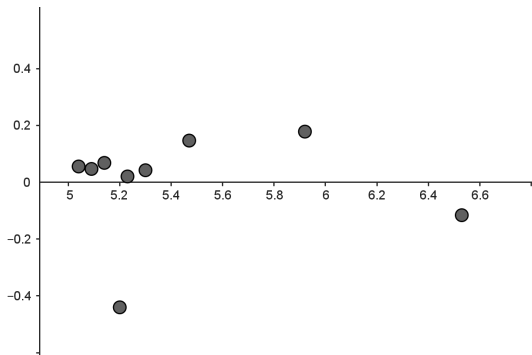
Le nuage des résidus est :



Dont les valeurs $e_i = y_i - \hat{y}_i$ sont données dans le tableau qui suit :

Repérage des valeurs aberrantes

Le nuage des résidus est :



Dont les valeurs $e_i = y_i - \hat{y}_i$ sont données dans le tableau qui suit :

i	1	2	3	4	5	6	7	8	9
e_i	0,06	0,18	0,04	0,02	0,05	0,15	-0,12	0,07	-0,44

Repérage des valeurs aberrantes

Classons les résidus dans l'ordre croissant :

$-0,44$; $-0,12$; $0,02$; $0,04$; $0,05$; $0,06$; $0,07$; $0,15$; $0,18$

Le premier résidu associé au point de coordonnées $(5, 2; 4, 69)$ est douteux.

Repérage des valeurs aberrantes

Classons les résidus dans l'ordre croissant :

$-0,44$; $-0,12$; $0,02$; $0,04$; $0,05$; $0,06$; $0,07$; $0,15$; $0,18$

Le premier résidu associé au point de coordonnées $(5, 2; 4, 69)$ est douteux.

Fixons-nous un risque α de 5%.

Posons H_0 : $-0,44$ n'est pas une valeur critique vs H_1 : $-0,44$ est une valeur critique.

Repérage des valeurs aberrantes

Classons les résidus dans l'ordre croissant :

$-0,44$; $-0,12$; $0,02$; $0,04$; $0,05$; $0,06$; $0,07$; $0,15$; $0,18$

Le premier résidu associé au point de coordonnées $(5, 2; 4, 69)$ est douteux.

Fixons-nous un risque α de 5%.

Posons H_0 : $-0,44$ n'est pas une valeur critique vs H_1 : $-0,44$ est une valeur critique.

Comme $n = 9$, $r_{obs} = \frac{r_2 - r_1}{r_8 - r_1} = \frac{-0,12 + 0,44}{0,15 + 0,44} = 0,542$.

Avec la table, on lit que $r_{critique} = 0,512$.

Au risque α de 5%, on rejette H_0 .

On élimine donc le point associé au résidu $-0,44$, i.e $(5, 20; 4, 69)$

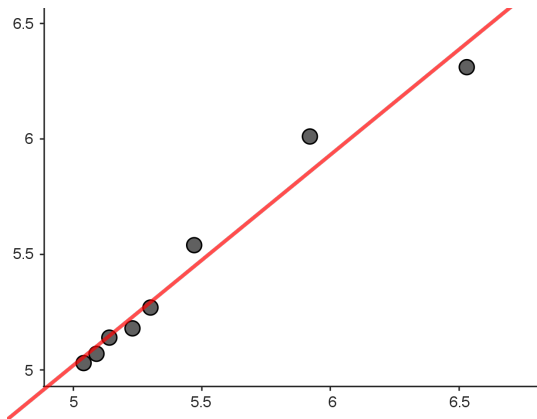
Test de la pente et de l'ordonnée à l'origine

Nous obtenons ainsi un nouveau tableau de données, avec 8 couples d'observation :

Labo	Bertrand	DNS
1	5,04	5,03
2	5,92	6,01
3	5,30	5,27
4	5,23	5,18
5	5,09	5,07
6	5,47	5,54
7	6,53	6,31
8	5,14	5,14

Test de la pente et de l'ordonnée à l'origine

La droite de régression de Y en X du nuage de points toiletté a pour équation : $\hat{y} = 0,91x + 0,46$, avec $r^2 = 0,9703$.



Test de la pente et de l'ordonnée à l'origine

L'étude précédente nous permet de valider le modèle linéaire (bien qu'il y ait une étude complémentaire sur la normalité des résidus à effectuer : faites-la à l'aide de Python et du test de Shapiro-Wilk).

La concordance des deux méthodes se traduit par une équation théorique de la droite de régression de Y en X : $Y = \beta X + \alpha$, avec $\alpha = 0$ (ordonnée à l'origine) et $\beta = 1$ (pente).

Test de la pente et de l'ordonnée à l'origine

Test de la pente

On pose $H_0 : \beta = 1$ et $H_1 : \beta \neq 1$.

On se fixe un risque de première espèce de 5%.

La variable de décision sous H_0 est : $T = \frac{|b - 1|}{s_b}$ qui suit la loi de Student à $n - 2 = 8 - 2 = 6$ ddl.

$$\text{Avec } s_b^2 = \frac{s_y^2/s_x^2 - b^2}{n - 2}.$$

Conclusion ?

Test de la pente et de l'ordonnée à l'origine

Test de l'ordonnée à l'origine

On pose $H_0 : \alpha = 0$ et $H_1 : \alpha \neq 0$.

On se fixe un risque de première espèce de 5%.

La variable de décision sous H_0 est : $T = \frac{|a - 0|}{s_a}$ qui suit la loi de Student à $n - 2 = 8 - 2 = 6$ ddl.

$$\text{Avec } s_a = s_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \text{ et } s_r = \sqrt{\frac{1 - r^2}{n - 2}}.$$

Conclusion ?

Conclusion générale

Que déduisez-vous de l'étude précédente ?