



- 1 Sélection des données dans la feuille de calcul
- 2 La vue Analyse des données
- 3 Calculs statistiques

http://url.univ-irem.fr/ft37





QR Code

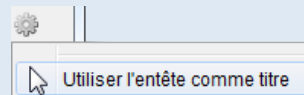
GeoGebra dispose d'un outil permettant de mener une comparaison de plusieurs séries statistiques. Les valeurs prises par les séries doivent être entrées dans la feuille de calcul de la vue **Tableur** et GeoGebra ne permet de comparer que des séries dont les données brutes sont fournies.


1 Sélection des données dans la feuille de calcul

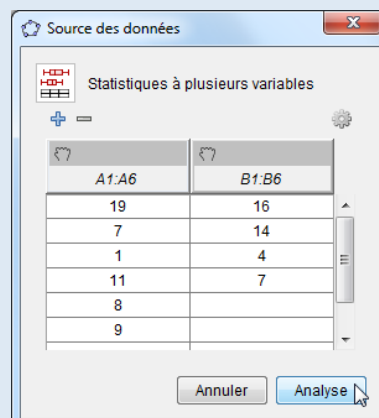
Lorsque les données occupent des colonnes adjacentes, la sélection de celles-ci s'avère très simple.

Méthode

- Sélectionner la plage de cellules contenant les données à comparer.
- Cliquer sur l'icône .
- Dans la boîte de dialogue **Source des données**, cliquer sur  puis choisir Utiliser l'entête comme titre si les colonnes présentent un entête.



- Cliquer sur le bouton .



Tableur		
	A	B
1	19	16
2	7	14
3	1	4
4	11	7
5	8	
6	9	
7		

Remarque :







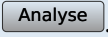
- Comme l'exemple précédent le montre, les effectifs des séries à comparer peuvent être différents (néanmoins, certains tests statistiques nécessitent des séries de même effectif).
- La procédure précédente reste valable pour plus de deux séries à comparer.

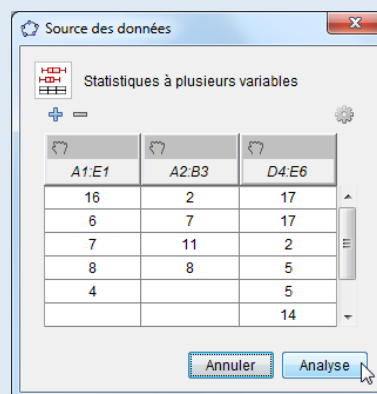
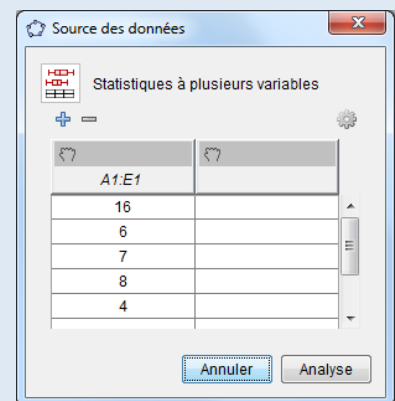
Lorsque les données ne sont pas regroupées dans des colonnes adjacentes (elles peuvent être en ligne ou encore dans des plages de cellules arbitraires), il reste cependant possible de les sélectionner, mais au prix d'un plus grand nombre de manipulations.

Méthode


On désire comparer les trois séries dont les valeurs appartiennent respectivement aux plages de cellules A1 : E1, A2 : B3 et D4 : E6.

Tableur						
	A	B	C	D	E	
1	16	6	7	8	4	
2	2	11				
3	7	8				
4				17	5	
5				17	5	
6				2	14	
7						

- Sélectionner une cellule non vide (par exemple, la cellule A1).
- Cliquer sur l'icône .
- Dans la vue **Tableur**, sélectionner la plage de cellules A1 : E1 puis, dans la boîte de dialogue **Source des données**, cliquer sur l'icône  pour remplir la première colonne avec les valeurs de la plage A1 : E1.
- Cliquer sur le bouton  pour ajouter une série de valeurs : une colonne vide apparaît.
- Dans la vue **Tableur**, sélectionner la plage de cellules A2 : B3 puis, dans la boîte de dialogue **Source des données**, cliquer sur l'icône  située au-dessus de la colonne vide.
- Cliquer sur le bouton .
- Dans la vue **Tableur**, sélectionner la plage de cellules D4 : E6 puis, dans la boîte de dialogue **Source des données**, cliquer sur l'icône  située au-dessus de la colonne vide.
- Cliquer sur le bouton .

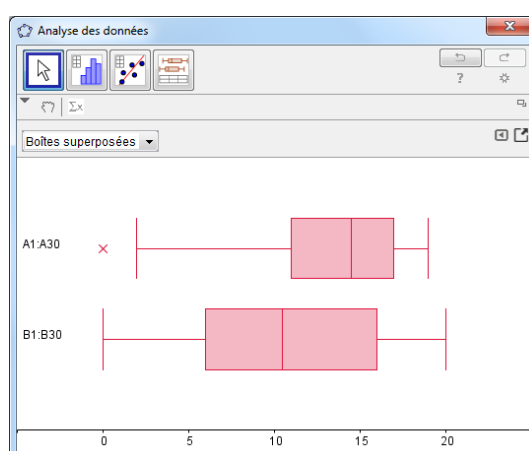





Remarque :


- GeoGebra permet également d'ajouter des séries dont les valeurs sont isolées dans la feuille de calcul : il suffit simplement de sélectionner ces valeurs en maintenant la touche **Ctrl** enfoncée.
- Le bouton  permet de supprimer une série de valeurs dans la boîte de dialogue **Source des données**. Cependant, il n'est pas possible de supprimer n'importe laquelle des colonnes : ce bouton n'agit que sur la dernière colonne.



2 La vue Analyse des données


L'appui sur le bouton **Analyse** de la boîte de dialogue **Source des données** provoque l'affichage de la vue **Analyse des données**.

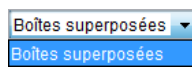


La barre principale permet d'accéder aux outils  (Statistiques à une variable),  (Statistiques à deux variables) et  (Statistiques à plusieurs variables).

La barre d'outils (qui se déplie en cliquant sur le bouton ) propose les outils suivants :


-  : l'outil Afficher la source de données pour afficher la fenêtre **Source des données** et modifier les données à analyser;
-  : l'outil Afficher statistiques pour afficher dans la vue les calculs statistiques concernant les données.

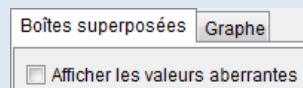
Dans la barre d'outils, le bouton  permet d'ancrer la vue **Analyse des données** dans la fenêtre principale de GeoGebra.

Dans la zone de graphique, les boîtes à moustaches sont les seuls types de graphiques disponibles. GeoGebra permet de modifier quelques paramètres de ce mode de représentation. 

Pour afficher les valeurs aberrantes, c'est-à-dire celles situées hors de l'intervalle $[Q_1 - 1,5 \times EI; Q_3 + 1,5 \times EI]$ (où EI désigne l'intervalle interquartile $Q_3 - Q_1$) :


Méthode

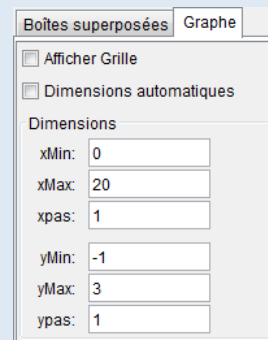
- Cliquer sur le bouton .
- Dans l'onglet **Boîtes superposées**, cocher ou décocher la case **Afficher les valeurs aberrantes**.




Pour modifier le repère :

Méthode

- Cliquer sur le bouton .
- Ouvrir l'onglet **Graphe**.
- Cocher ou décocher la case **Afficher Grille** selon les besoins.
- Décocher la case **Dimensions automatiques** pour modifier les paramètres du repère : les bornes des abscisses et des ordonnées sont modifiables par les champs **xMin**, **xMax**, **yMin** et **yMax**, tandis que les graduations des axes sont modifiables par les champs **xpas** et **ypas**.



Le bouton  permet d'exporter le graphique :

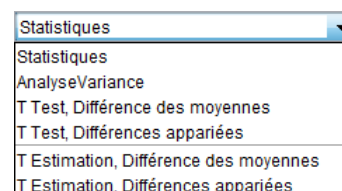
- vers la vue **Graphique** active ;
- vers le presse-papier afin de coller le graphique dans une autre application en tant qu'image ;
- vers un fichier image.



3 Calculs statistiques

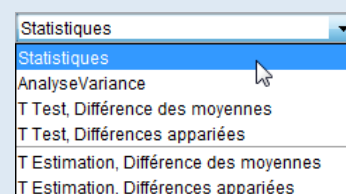
L'appui sur le bouton Σx provoque l'apparition d'une nouvelle zone dans la vue **Analyse des données**. Le choix du type de calculs à afficher s'effectue à l'aide de la liste déroulante située dans la partie supérieure de la zone.

Pour afficher les caractéristiques usuelles des séries statistiques :



Méthode

- Sélectionner **Statistiques** dans la liste déroulante.
- GeoGebra affiche alors, dans un tableau, un certain nombre de caractéristiques des séries sélectionnées :



	n	Moyenne	σ	s	Min	Q1	Médi...	Q3	Max
Série A	30	8.5667	5.4814	5.5751	0	4	7	12	20
Série B	27	11.1852	5.0261	5.1219	1	9	12	16	19

- n : l'effectif total des séries statistiques
- Moyenne : la moyenne des séries statistiques
- σ : l'écart-type standard des séries statistiques (racine carrée de la variance)

$$\sigma = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

- s : l'écart-type sans biais des séries statistiques (racine carrée de la variance sans biais)

$$s = \sigma \times \sqrt{\frac{n}{n-1}}$$

- Min : la valeur minimale des séries statistiques
- Q_1 : le premier quartile des séries statistiques
- Médiane : la médiane des séries statistiques
- Q_3 : le troisième quartile des séries statistiques
- Max : la valeur maximale des séries statistiques

Une analyse de la variance (ou encore ANOVA) à un facteur est un moyen d'apprécier l'effet d'une variable qualitative (appelée facteur) sur une variable numérique et revient à comparer plusieurs moyennes d'échantillons supposés gaussiens et de variance identique. On teste l'hypothèse que les moyennes des échantillons sont identiques contre celle qu'il existe au moins deux échantillons de moyennes différentes. Techniquement, on construit un test de FISHER; en effet si les moyennes des échantillons sont identiques alors on décompose la variance totale comme suit :

$$S^2 = S_A^2 + S_R^2$$

Où S_A^2 représente la variance due au facteur (dite aussi inter-groupes) et S_R^2 la variance résiduelle (dite aussi intra-groupes).

Ainsi, si k représente le nombre de modalités de la variable qualitative et n l'effectif total, on construit la statistique suivante :

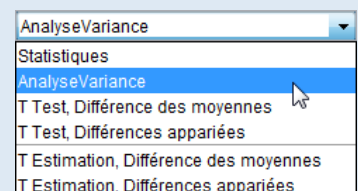
$$\frac{\frac{S_A^2}{k} - 1}{\frac{S_R^2}{n} - k} \sim F(k-1; n-k)$$

Si ce rapport est supérieur à une valeur critique, alors on conclut à une influence significative du facteur et on rejette alors l'hypothèse que toutes les moyennes sont identiques.

Pour effectuer une analyse de la variance à un facteur :

Méthode

- Sélectionner **AnalyseVariance** dans la liste déroulante.
- GeoGebra affiche alors un premier tableau qui permet de connaître toutes les valeurs nécessaires au test de FISHER, à savoir :



	dlib	SS	MoyCarrés	F	P
Entre groupes	1	97.4365	97.4365	3.3844	0.0712
Groupe	55	1583.4407	28.7898		
Total	56	1680.8772			

- dlib : nombre de degrés de liberté associé aux variances inter-groupes et intra-groupes ;
- SS : somme des carrés inter-groupes et intra-groupes ;

- MoyCarrés : quotient des sommes des carrés inter-groupes et intra-groupes par les degrés de liberté respectifs;
- F : quotient de MoyCarrés inter-groupes par MoyCarrés intra-groupes;
- P : probabilité d'avoir une variable aléatoire suivant une loi de FISCHER de paramètres les degrés de liberté inférieure à la valeur donnée par F . En général, si P est supérieur à 0,05, on rejette l'hypothèse d'égalité des moyennes.
- Le second tableau permet de visualiser l'effectif total des séries (n), leur moyenne et leur écart-type sans biais (s).

	n	Moyenne	s
Série A	30	8.5667	5.5751
Série B	27	11.1852	5.1219

GeoGebra permet d'effectuer un t-test de la différence des moyennes de deux populations à partir de la donnée d'échantillons.

On suppose ici que les deux échantillons X_1 et X_2 suivent des lois normales de paramètres respectifs (μ_1, σ_1) et (μ_2, σ_2) .

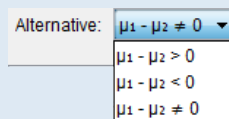
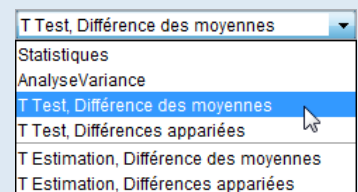
On construit la statistique :

$$T = \frac{\mu_1 - \mu_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad \text{où} \quad S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

S_1^2 et S_2^2 étant les variances observées. Cette statistique suit une loi de STUDENT de paramètre $(n_1 + n_2 - 2)$. GeoGebra permet d'effectuer un test bilatéral ou unilatéral.


Méthode

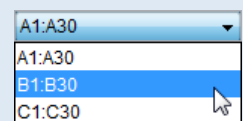
- Sélectionner **T Test, Différences des moyennes** dans la liste déroulante.
- Dans la liste déroulante **Alternative**, sélectionner l'hypothèse alternative souhaitée (l'hypothèse nulle est toujours $\mu_1 - \mu_2 = 0$).



- GeoGebra affiche alors un premier tableau récapitulant la moyenne, l'écart-type sans biais (s) ainsi que l'effectif total (n) des séries sélectionnées.

		Moyenne	s	n
Échantillon 1	B1:B30	9.1091	5.8581	27
Échantillon 2	C1:C30	10.5557	5.7878	29

Dans le cas où plus de deux séries ont été choisies dans la vue **Tableur** à l'occasion de l'utilisation de l'outil , GeoGebra permet de sélectionner les échantillons à comparer à l'aide des listes déroulantes situées sur les lignes **Échantillon 1** et **Échantillon 2**.



- Le second tableau fournit les résultats du test.

Différence	P	t	SE	dlib
1.276	0.4177	0.8167	1.5625	54.4982

- Différence : la différence entre les moyennes des deux échantillons ;
- P : la probabilité associée à la statistique T calculée, à savoir t (si la valeur est inférieure à 0,05 par exemple, on rejette l'hypothèse d'égalité des moyennes) ;
- t : valeur de la statistique de test T pour les échantillons donnés ;
- SE : erreur standard (à priori, il s'agit du quotient de la différence des moyennes par la valeur de t) ;
- dlib : nombre de degrés de liberté associé à t .

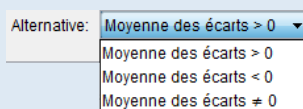
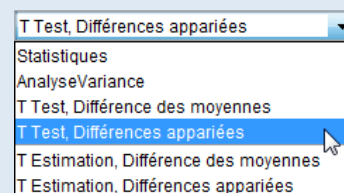
Lorsque les échantillons sont appariés (ils doivent donc avoir le même effectif) et quand les données des deux variables sont distribuées selon une loi normale, on peut effectuer un test t pour comparer leurs moyennes.

On soustrait les observations du premier échantillon et les observations du deuxième entre elles et on étudie la moyenne des écarts. L'hypothèse du test ici est que la moyenne des écarts est nulle contre une des trois hypothèses alternatives en fonction du choix d'un test bilatéral ou unilatéral.

On construit la statistique $T = \frac{m}{\sqrt{\frac{S^2}{n}}}$ qui suit une loi de STUDENT de paramètre $n-1$ (n étant la taille commune des échantillons) où $m = \frac{\sum_{i=1}^n d_i}{n}$, d_i étant l'écart entre les observations i des échantillons et $S^2 = \frac{\sum_{i=1}^n (d_i - m)^2}{n-1}$.

Méthode


- Sélectionner **T Test, Différences appariées** dans la liste déroulante.
- Dans la liste déroulante **Alternative**, sélectionner l'hypothèse alternative souhaitée (l'hypothèse nulle est toujours « la moyenne des différences appariées de la population vaut zéro »).

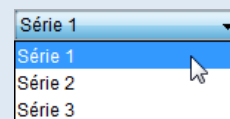


- GeoGebra affiche alors un premier tableau récapitulant la moyenne, l'écart-type sans biais s (s est le quotient de la somme des écarts aux carrés des observations à la moyenne par $n-1$), l'effectif total n des séries sélectionnées ainsi que les différences entre les deux échantillons pour ces valeurs.

		Moyenne	s	n
Échantillon 1	Série 1	12.3816	5.3347	24
Échantillon 2	Série 3	9.8179	5.3982	24
Différences		2.5636	7.0246	24

Dans le cas où plus de deux séries ont été choisies dans la vue **Tableur**

à l'occasion de l'utilisation de l'outil , GeoGebra permet de sélectionner les échantillons à comparer à l'aide des listes déroulantes situées sur les lignes **Échantillon 1** et **Échantillon 2**.



- Le second tableau fournit les résultats du test.

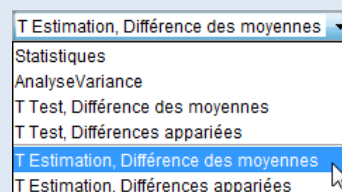
Moyenne des écarts	P	t	SE	dlib
2.5636	0.087	1.7879	1.4339	23

- Différence des écarts : la différence entre les moyennes des deux échantillons;
- P : probabilité associée à la statistique T calculée, à savoir t ;
- t : valeur de la statistique de test T pour les échantillons donnés;
- SE : erreur standard (à priori, il s'agit du quotient de la moyenne des écarts par la valeur de t);
- dlib : nombre de degrés de liberté associé à t .

T Estimation, Différences des moyennes : c'est le même point de départ que le t-test différence des moyennes, sauf qu'à la place d'un test, on fournit un intervalle dans lequel se trouve la différence des moyennes des deux échantillons au niveau de confiance fixé. On considère donc la même statistique T .

Méthode

- Sélectionner **T Estimation, Différences des moyennes** dans la liste déroulante.
- Dans le champ **Niveau de confiance**, entrer une valeur comprise entre 0 et 1 : elle représente la probabilité que la valeur de la différence des moyennes des populations soit dans l'intervalle donné.

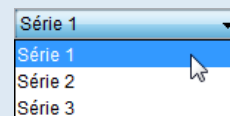


Niveau de confiance:

- GeoGebra affiche alors un premier tableau récapitulant la moyenne, l'écart-type sans biais (s), et l'effectif total (n) des séries sélectionnées.

		Moyenne	s	n
Échantillon 1	Série 1	11.3841	5.6019	31
Échantillon 2	Série 3	9.4841	5.3509	28

Les listes déroulantes situées sur les lignes **Échantillon 1** et **Échantillon 2** permettent de sélectionner les échantillons à comparer.



- Le second tableau fournit les résultats du test.

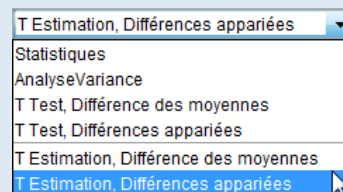
Différence	ME	Limite inférieure	Limite supérieure	SE	dlib
1.9	2.8567	-0.9567	4.7567	1.4265	56.8105

- Différence : la différence entre les moyennes des deux échantillons;
- ME : amplitude de l'intervalle divisée par deux;
- Limite inférieure : valeur minimale de l'intervalle;
- Limite supérieure : valeur maximale de l'intervalle;
- SE : erreur standard (à priori, il s'agit du quotient de la différence des moyennes par la valeur de t);
- dlib : nombre de degrés de liberté associé à t .

T Estimation, Différences appariées : c'est le même point de départ que le t-test, différences appariées, sauf qu'à la place d'un test, on donne un intervalle dans lequel se trouve la moyenne des différences des observations des deux échantillons au niveau de confiance fixé. On considère donc la même statistique T .

Méthode

- Sélectionner **T Estimation, Différences appariées** dans la liste déroulante.
- Dans le champ **Niveau de confiance**, entrer une valeur comprise entre 0 et 1 : elle représente la probabilité que la valeur de la différence des moyennes des populations soit dans l'intervalle donné.

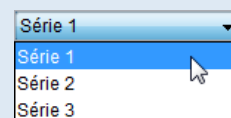


Niveau de confiance:

- GeoGebra affiche alors un premier tableau récapitulant la moyenne, l'écart-type sans biais s (s est le quotient de la somme des écarts aux carrés des observations à la moyenne par $n - 1$), l'effectif total n des séries sélectionnées ainsi que les différences entre les deux échantillons pour ces valeurs.

		Moyenne	s	n
Échantillon 1	Série 1	10.2385	5.073	25
Échantillon 2	Série 2	8.8117	5.8022	25
Différences		1.4267	7.3648	25

Les listes déroulantes situées sur les lignes **Échantillon 1** et **Échantillon 2** permettent de sélectionner les échantillons à comparer.



- Le second tableau fournit les résultats du test.

Moyenne des écarts	ME	Limite inférieure	Limite supérieure	SE	dlib
1.4267	1.9411	-0.5144	3.3679	1.473	24

- Moyenne des écarts : la moyenne des écarts des observations des deux échantillons;
- ME : amplitude divisée par deux de l'intervalle de confiance;
- Limite inférieure : borne inférieure de l'intervalle;
- Limite supérieure : borne supérieure de l'intervalle;
- SE : erreur standard (à priori, il s'agit du quotient de la différence des moyennes par la valeur de t);
- dlib : nombre de degrés de liberté associé à t .