Corrigé - Série 2
Inférence sur les paramètres

Exercice 1 - Les enfants qui dépassent leurs parents

a) Les filles sont-elles plus grandes que leurs mères en moyenne ?

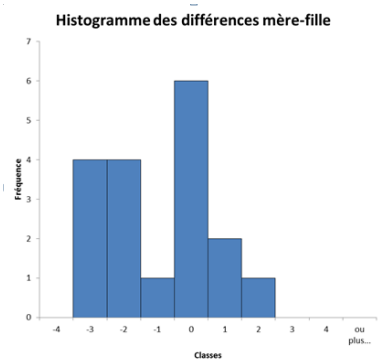
$$H_0 : \mu_{\text{filles}} = \mu_{\text{mères}}$$
$$H_1 : \mu_{\text{filles}} > \mu_{\text{mères}}$$

On répondra par un test de Student sur des données appariées (groupées par paires mère-fille). On voudra donc faire calculer la valeur observée de la statistique du test

$$T_0 = \frac{\overline{D}}{S_D/\sqrt{n}}$$

Dans l'Utilitaire d'analyse, on commande un *Test d'égalité des espérances : observations pairées* et on obtient le résultat ci-dessous :

	Variable 1	*Variable 2*
Moyenne	64,5	63,5555556
Variance	5,2058824	6,37908497
Observations	18	18
Coefficient de corrélation de P	0,7859871	
Différence hypothétique des m	0	
Degré de liberté	17	
Statistique t	2,5210626	
P(T<=t) unilatéral	0,0109892	
Valeur critique de t (unilatéral)	1,7396067	
P(T<=t) bilatéral	0,0219785	
Valeur critique de t (bilatéral)	2,1098156	



Puisque $t_{obs} = 2,521$, le seuil observé du test unilatéral est $P(T > 2,521) = 0,0109892$, où $T \sim t_{17}$. Cette valeur-P étant inférieure à 5% (le seuil du test), on rejette H_0 et on conclut que les filles sont significativement plus grandes que leurs mères en moyenne.

On aurait pu tirer la même conclusion en comparant $t_{obs} = 2,521$ à la valeur critique d'une loi de Student, soit $t_{\alpha;n-1} = t_{0,05;17} = 1,739$. Le test étant unilatéral à droite, on rejette H_0, car $t_{obs} > 1,739$.

Le test de Student suppose que les données sont issues d'une loi normale. Un histogramme des 18 différences nous montre une tendance à la bimodalité, mais le nombre de valeurs étant peu élevé, il est difficile de rejeter catégoriquement la normalité.

b) La différence mère-fille est-elle plus petite que la différence père-fils ?

$$H_0 : \mu_{\text{mère-fille}} = \mu_{\text{père-fils}}$$
$$H_1 : \mu_{\text{mère-fille}} < \mu_{\text{père-fils}}$$

Il faut d'abord calculer les 18 différences concernées. On répondra par un test de Student sur des échantillons indépendants. Pour choisir le bon test, il faut d'abord déterminer si les variances peuvent être considérées égales (à l'aide d'un test de Fisher).

$$H_0 : \sigma^2_{\text{mère-fille}} = \sigma^2_{\text{père-fils}}$$
$$H_1 : \sigma^2_{\text{mère-fille}} \neq \sigma^2_{\text{père-fils}}$$

On voudra faire calculer la valeur observée de la statistique du test

$$F_0 = \frac{S_1^2}{S_2^2}$$

Dans l'Utilitaire d'analyse, on commande un *Test d'égalité des variances (F-test)* et on obtient le résultat ci-dessous :

	Variable 1	*Variable 2*
Moyenne	-0,944444	-0,90909091
Variance	2,5261438	8,09090909
Observations	18	11
Degré de liberté	17	10
F	0,31222	
P(F<=f) unilatéral	0,016875	
Valeur critique pour F (unilatér	0,4081773	

Puisque $f_{obs} = 0,3122$, le seuil observé du test bilatéral est $2 \times P(F < 0,3122) = 2 \times 0,016875 = 0.03375$, où $F \sim F_{17,10}$. Cette valeur-P étant inférieure à 5% (le seuil du test), on rejette H_0 et on conclut que les variances diffèrent significativement.

Le test de Student à utiliser sera donc celui avec variances inégales. On voudra donc faire calculer la valeur observée de la statistique du test

$$T_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Dans l'Utilitaire d'analyse, on commande un *Test d'égalité des espérances : deux observations de variances différentes* et on obtient le résultat ci-dessous :

	Variable 1	*Variable 2*
Moyenne	-0,94444444	-0,90909091
Variance	2,52614379	8,09090909
Observations	18	11
Différence hypothétique	0	
Degré de liberté	14	
Statistique t	-0,03777556	
P(T<=t) unilatéral	0,4852	
Valeur critique de t (unila	1,76131014	
P(T<=t) bilatéral	0,97040001	
Valeur critique de t (bilat	2,14478669	

Puisque $t_{obs} = -0,03777$, le seuil observé du test unilatéral est $P(T < -0,03777) = 0,4852$, où $T \sim t_{14}$. Cette valeur-P étant supérieure à 5% (le seuil du test), on ne rejette pas H_0 et on conclut que la différence mère-fille n'est pas significativement inférieure à la différence père-fils.

c) Estimer la proportion de jeunes qui dépassent le parent du même sexe, avec un niveau de confiance de 95%.

On veut construire un intervalle de confiance sur une proportion. Il faut donc avoir une grande taille d'échantillon, car l'IC est asymptotique. Ici, $n = 29$ est tout juste acceptable.

Il faut définir la variable binaire qui identifie les gens plus grands que leur parent du même sexe à l'aide de la fonction

$$\text{SI(Test\_logique ;Valeur\_si\_vrai ;Valeur\_si\_faux)} = \text{SI(C2>D2 ;1 ;0)}.$$

On calcule ensuite la proportion échantillonnale $\hat{p}$ en faisant la moyenne de cette colonne, puis on complète les calculs en utilisant la formule

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

et on obtient l'intervalle $[0,335;\ 0,699]$.

Exercice 2 - Les donateurs aux partis politiques

a) La valeur moyenne d'un don est-elle la même d'un parti à l'autre ?

L'Utilitaire d'analyse permet de faire un test global de comparaison des moyennes avec la commande *Analyse de variance : un facteur*. Les trois séries de données doivent être placées dans trois colonnes adjacentes. On obtient un tableau des moyennes et des variances échantillonnales, ainsi que la table d'anova :

RAPPORT DÉTAILLÉ

Groupes	*Nombre d'échantillons*	*Somme*	*Moyenne*	*Variance*
CAQ	22	4392,99	199,6813636	34060,608
PLQ	42	12337	293,7380952	65474,539
PQ	70	9925	141,7857143	19763,069

ANALYSE DE VARIANCE

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Probabilité*	*Valeur critique pour F*
Entre Groupes	606115,4043	2	303057,7021	8,3345342	0,00039157	3,065295706
A l'intérieur des groupes	4763380,671	131	36361,68451			
Total	5369496,075	133				

On est tenté de rejeter d'emblée $H_0 : \mu_{CAQ} = \mu_{PLQ} = \mu_{PQ}$ en raison du seuil observé inférieur à 5% :

$$\text{Valeur} - \text{P} = P(F > 8,3345) = 0,00039157 \qquad \text{où} \quad F \sim F_{2,131}$$

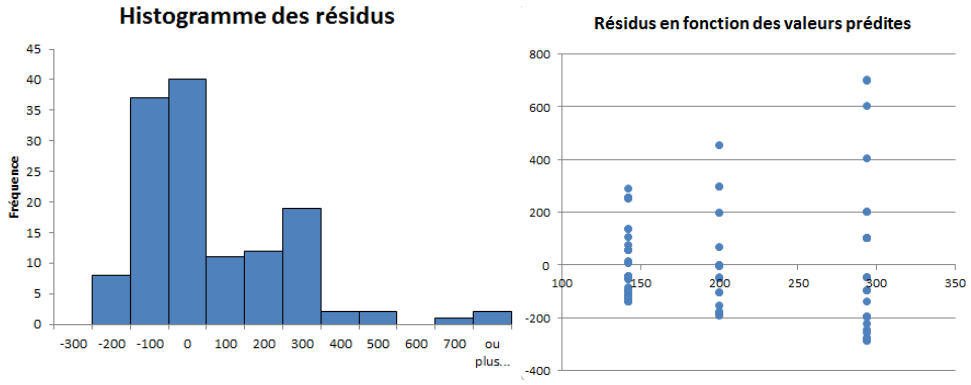
Mais attention...

b) Les postulats du modèle d'analyse de la variance appliqué en a) sont-ils respectés ?

Pour répondre à cette question, il faut faire une analyse de résidus. On doit vérifier que la loi normale est un modèle raisonnable, et que les variances sont similaires d'un échantillon à l'autre.

Pour créer la variable résidus dans une nouvelle colonne, on soustrait à chaque observation sa moyenne échantillonnale locale : $e_{ij} = y_{ij} - \overline{y}_{i\bullet}$.

On construit ensuite l'histogramme des résidus et le graphique des résidus en fonction des valeurs prédites :



On remarque une bonne asymétrie vers la droite dans l'histogramme. De plus, le graphique de droite présente un patron en forme d'entonnoir, donc une hétéroscédasticité assez claire. Ces deux aspects viennent mettre un gros bémol sur la validité du test F effectué en a).

Devant une telle situation, l'option la plus fréquemment envisagée est la transformation de la variable réponse (Y) avec une fonction monotone comme $\sqrt{Y}$, $\ln(Y)$, $1/Y$, Y^2, etc. On refait l'anova avec plusieurs variables transformées jusqu'à ce que les postulats soient respectés.

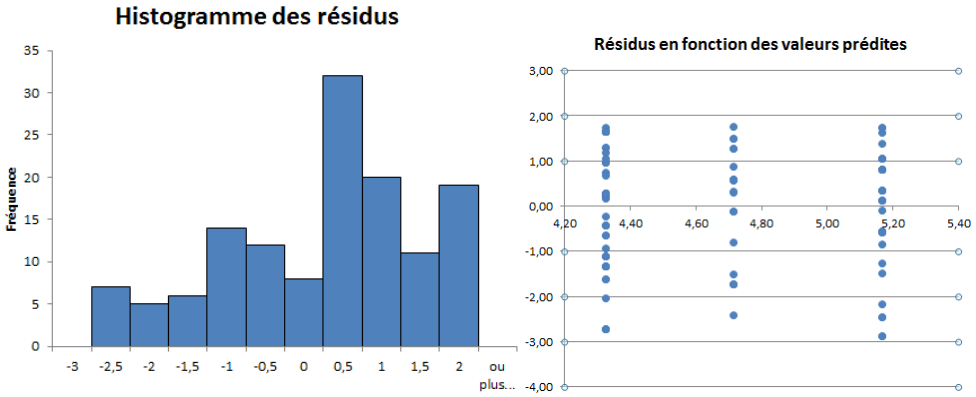
Après quelques essais, on voit que dans notre cas, c'est la transformation logarithmique qui donne les meilleurs résultats. Voici la nouvelle analyse :

RAPPORT DÉTAILLÉ

Groupes	*Nombre*	*Somme*	*Moyenne*	*Variance*
CAQ	22	103,657528	4,71170581	1,62398117
PLQ	42	216,962768	5,16578019	1,50703999
PQ	70	302,759295	4,32513278	1,64675856

ANALYSE DE VARIANCE

Source varia	*S. des carrés*	*Deg. liberté*	*Moy. carrés*	*F*	*Probabilité*	*Val. Crit. F*
Entre Groupe	18,6441203	2	9,32206017	5,8285516	0,00375882	3,06529571
A l'intérieur	209,518585	131	1,59937851			
Total	228,162705	133				



Puisque l'analyse des résidus est plus satisfaisante (malgré une légère asymétrie à gauche), on peut interpréter les résultats du test global de comparaison des moyennes. Le seuil observé étant inférieur à 5% :

$$\text{Valeur} - \text{P} = P(F > 5,82855) = 0,00375882 \qquad \text{où} \quad F \sim F_{2,131},$$

on rejette H_0, et on conclut que la valeur du don moyen à un des trois principaux partis politiques est différente selon le parti.

c) Peut-on voir où se situent les différences significatives ?

Le test global est significatif, on peut donc comparer les moyennes deux à deux. Puisque les tailles d'échantillon sont différentes, on ne peut pas calculer une seule "PPDS". Il faut calculer une différence significative (une marge d'erreur) pour chaque paire de moyennes.

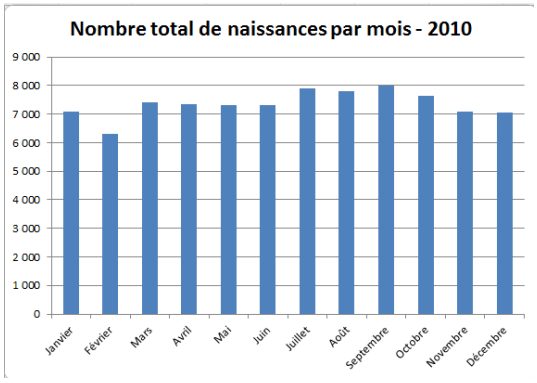
MSE	1,59937851
alpha	0,05
quantile t	1,97823854
CAQ vs PLQ	
err-type(diff. Moy)	0,33283549
PPDS	0,65842799
diff. Moy	-0,45407438
SIGNIF?	
CAQ vs PQ	
err-type(diff. Moy)	0,30910724
PPDS	0,61148786
diff. Moy	0,38657303
SIGNIF?	
PLQ vs PQ	
err-type(diff. Moy)	0,24683741
PPDS	0,48830327
diff. Moy	0,84064741
SIGNIF?	

Ici, on conclut que seuls le PQ et le PLQ reçoivent des dons dont la valeur moyenne diffère significativement. On pourrait représenter schématiquement ces comparaisons deux à deux comme suit :

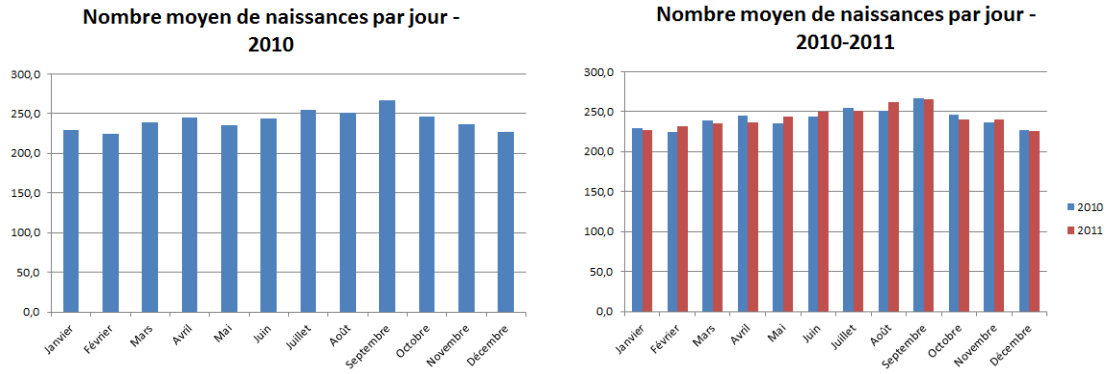
PQ	CAQ	PLQ
141, 79 $	199, 68 $	293, 74 $
4, 32	4, 71	5, 17

Exercice 3 - Distribution des naissances

a)



b) Il y a moins de naissances en février... donc le mois de mai n'est pas propice à la fécondation ? En fait, cela est peut-être dû au fait que février compte moins de jours que les autres mois ? Il serait peut-être plus judicieux de comparer le nombre moyen de naissances par jour d'un mois à l'autre :



Février est toujours le plus bas en 2010 !

c) Observe-t-on le même phénomène en 2011 ?

C'est en décembre et en janvier que les naissances sont les moins nombreuses en 2011. On voit quand même une tendance se dessiner : il semble y avoir plus de naissances en été qu'en hiver. Il serait intéressant d'étudier un plus grand nombre d'années pour voir si ce n'est que ponctuel.

d) Peut-on affirmer que l'accroissement naturel (naissances - décès) est supérieur à 2000 individus en moyenne à chaque mois, au seuil de 5% ?

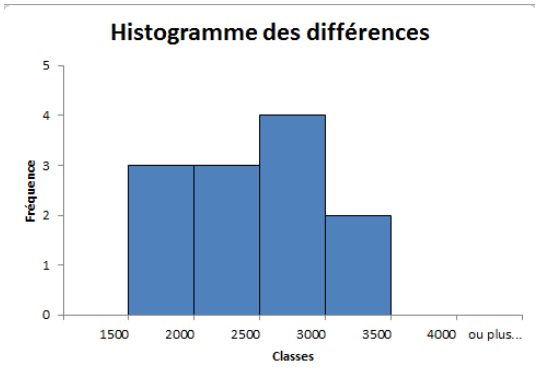
On réalise un test de comparaison de moyennes en considérant les observations appariées. On peut le faire directement dans Excel à partir des naissances et des décès, ou en se ramenant à un seul échantillon de différences : on calcule soi-même les différences entre les naissances et les décès chaque mois, et on fait un test de Student à un échantillon pour vérifier si la moyenne des différences est supérieure à 2000.

$$H_0 : \mu_D = 2000$$

$$H_1 : \mu_D > 2000$$

.

Test d'égalité des espérances: observations pairées		
	Variable 1	*Variable 2*
Moyenne	7354,16667	4862,5
Variance	212935,606	76420,4545
Observations	12	12
Coefficient de corrélation de Pear	-0,04676788	
Différence hypothétique des moy	2000	
Degré de liberté	11	
Statistique t	3,10291986	
P(T<=t) unilatéral	0,00502582	
Valeur critique de t (unilatéral)	1,79588482	
P(T<=t) bilatéral	0,01005164	
Valeur critique de t (bilatéral)	2,20098516	



On rejette H_0 si $t_{obs} = \dfrac{\overline{d} - 2000}{s_D/\sqrt{12}} > t_{11;0,05} = 1,796$. Puisque $t_{obs} = 3,10$, on rejette H_0 au seuil de 5%. La valeur du seuil observé (0,005) nous mène évidemment à la même conclusion, car il est inférieur au seuil du test.

On conclut donc que l'accroissement naturel moyen par mois est significativement supérieur à 2000.

Bien sûr, ce test se base sur le postulat de normalité. L'histogramme des différences n'a pas une forme de cloche parfaite, mais considérant que seulement 12 données le composent, il ne s'en éloigne pas suffisamment pour rejeter l'analyse de Student.

Exercice 4 - Séries éliminatoires

a) Peut-on dire que le nombre de buts comptés par l'équipe locale est supérieur en moyenne au nombre de buts comptés par l'équipe en visite ?

Test de comparaison de deux moyennes sur des données appariées provenant de populations normales à variances inconnues.

$$H_0 : \mu_{\text{local}} = \mu_{\text{visiteur}}$$
$$H_1 : \mu_{\text{local}} > \mu_{\text{visiteur}}$$

Test d'égalité des espérances: observations pairées		
	Variable 1	*Variable 2*
Moyenne	2,83146067	2,7752809
Variance	2,80081716	3,03983657
Observations	89	89
Coefficient de corrélation de Pearson	0,09202357	
Différence hypothétique des moyennes	0	
Degré de liberté	88	
Statistique t	0,23013811	
P(T<=t) unilatéral	0,40925929	
Valeur critique de t (unilatéral)	2,36947227	
P(T<=t) bilatéral	0,81851858	
Valeur critique de t (bilatéral)	2,63285804	

$t_{obs} = 0,230$, à comparer avec le quantile d'une loi de Student $t_{88;0,01} = 2,369$.

Seuil observé : $P(T > 0,230) = 0,409$ où $T \sim t_{88}$.
Au seuil de 1%, on ne rejette pas l'égalité des moyennes. L'équipe locale ne compte pas significativement plus de buts que les visiteurs en moyenne.

b) Le nombre de buts total comptés dans un match de séries est-il plus élevé quand l'équipe locale gagne que quand elle perd ?

Il faut d'abord créer une variable représentant la somme des buts des deux équipes. On crée ensuite une variable binaire pour distinguer si l'équipe locale a gagné ou perdu (aucune nulle en série). On trie les données selon cette variable, et on distingue ainsi deux échantillons de valeurs qu'on considère indépendants puisqu'associés à des matchs différents.

Test de comparaison de deux moyennes provenant de populations normales à variances inconnues, dont les échantillons sont indépendants.

$$H_0 : \mu_{\text{loc\_gagne}} = \mu_{\text{loc\_perd}}$$
$$H_1 : \mu_{\text{loc\_gagne}} > \mu_{\text{loc\_perd}}$$

Test d'égalité des variances (F-Test)

	Variable 1	*Variable 2*
Moyenne	5,90243902	5,35416667
Variance	5,8402439	6,82934397
Observations	41	48
Degré de liberté	40	47
F	0,85516909	
P(F<=f) unilatéral	0,30790011	
Valeur critique pour	0,59933323	

Test d'égalité des espérances: deux observations de variances égales

	Variable 1	*Variable 2*
Moyenne	5,90243902	5,35416667
Variance	5,8402439	6,82934397
Observations	41	48
Variance pondérée	6,37458532	
Différence hypothétique de	0	
Degré de liberté	87	
Statistique t	1,0211462	
P(T<=t) unilatéral	0,15500817	
Valeur critique de t (unilaté	1,66255735	
P(T<=t) bilatéral	0,31001634	
Valeur critique de t (bilatéra	1,98760828	

Le test F de comparaison des variances n'est pas significatif, donc on utilise le test de Student avec variances égales.

$t_{obs} = 1,021$, à comparer avec le quantile d'une loi de Student $t_{87;0,01} = 1,663$.

Seuil observé : $P(T > 1,021) = 0,155$ où $T \sim t_{87}$.

Au seuil de 1%, on ne rejette pas l'égalité des moyennes. Le nombre moyen de buts comptés dans un match n'est pas plus élevé lorsque l'équipe locale gagne.

Exercice 5 - 1, 2, 3... payez !

Nous supposons que le prix de l'essence dans les villes du Canada (X) suit une loi normale. Dans notre échantillon,

$n = 12$

$\overline{x} = 118,7$ cents

$s = 10,3$ cents.

a) Intervalle de confiance à 99% pour le prix moyen réel :

$$\overline{x} \pm t_{11,0,005} s/\sqrt{n} \quad = \quad 118,7 \pm 9,24 \quad = \quad [109.46, 127.94]$$

b) Pour réduire la longueur de cet intervalle de confiance, on peut augmenter la taille d'échantillon ou diminuer le niveau de confiance (i.e. augmenter α).

c) $X \approx N(118.7, 10.3^2)$
(Note : ici, ce sont les paramètres qui sont approximatifs, et non la loi !)
$P(X > 120) \approx P(Z > 0,126) = 0,450.$

d) $\overline{X} \approx N(118,7,\ (10,3)^2/12)$
(Note : ici encore, ce sont les paramètres qui sont approximatifs, et non la loi !)
$P(\overline{X} > 120) \approx P(Z > 0.437) = 0.331$

Exercice 6 - Seuils observés

a) H_0 : Les étudiants trouvent le cours plate.
H_1 : Les étudiants ne trouvent pas le cours plate.

Seuil observé de 0,0246, inférieur au seuil du test (5%) : On rejette H_0.
Ma conclusion ? Les étudiants ne trouvent pas le cours plate (quelle question... !).

b) $n = 25$, population normale de variance inconnue.

$$H_0 : \ \mu = 21$$
$$H_1 : \ \mu < 21$$

1) Valeur-p $= P(T < t_{obs}) = 0,0413$. On déduit que la valeur observée de la statistique du test est négative (car la valeur-p est inférieure à 1/2)., et donc que $\overline{x}$ est inférieure à 21 (significativement).
La valeur-p du test bilatéral aurait été $2 \times P(T < t_{obs}) = 2 \times 0,0413 = 0,0826$.

2) Valeur-p $= P(T < t_{obs}) = 0,3413$, donc $\overline{x}$ inférieure à 21 (mais pas significativement).
La valeur-p du test bilatéral aurait été $2 \times P(T < t_{obs}) = 2 \times 0,3413 = 0,6826$.

3) Valeur-p $= P(T < t_{obs}) = 0,6413$, donc $\overline{x}$ supérieure à 21 (car la valeur-p est supérieure à 1/2).
La valeur-p du test bilatéral aurait été $2 \times P(T > |t_{obs}|) = 2 \times (1 - 0,6413) = 0,7174$.

4) $n = 25$, $\overline{x} = 18$, et $s^2 = 100$.

La valeur observée de la statistique du test sera $t_{obs} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} = \dfrac{18 - 21}{\sqrt{100/25}} = -1,5$

Le seuil observé du test se calcule comme suit :
$P(T < t_{obs}) = P(T < -1,5) = 0,0733, \qquad$ où $T \sim t_{24}$

Remarque : Si vous avez consulté une table de la loi de Student pour évaluer la probabilité, vous avez seulement pu borner la valeur-p entre 0,05 et 0,10 car la valeur observée se situe entre les quantiles -1,318 et -1,711. La valeur ci-dessus a été obtenue par Excel.

5) $n > 25$, $\overline{x} = 18$, et $s^2 = 100$.

La valeur observée de la statistique du test sera plus grande en valeur absolue, car n est plus grand. Puisqu'elle est négative, elle sera plus à gauche, et fera donc diminuer la probabilité d'observer une valeur inférieure sous H_0.
De plus, en faisant augmenter n, on augmente les degrés de liberté de la loi t, qui sera donc moins évasée. L'aire sous la courbe à gauche de la valeur observée sera donc diminuée.
Le seuil observé sera donc plus bas que le précédent, et le test plus significatif. (C'est normal : un écart de 3 unités entre les moyennes théorique et échantillonnale est plus significatif s'il provient de 100 données que de 25 données.)

6) Aucun impact sur le calcul du seuil observé. C'est seulement sur la conclusion que cela peut faire une différence, si le seuil observé se trouve entre 0,05 et 0,10.

c) Faux : l'inverse est vrai.

d) Faux : on rejette H_0 s'il est inférieur au seuil du test.

e) Faux : cette notation indique seulement que Excel a calculé le seuil observé d'un test unilatéral. À vous de déterminer lequel (à droite ou à gauche).

f) Faux : la probabilité que H_0 soit vraie ne se calcule pas.

g) Faux : évidemment.