Exercices de biostatistique

Rappel: pour visualiser la formule associée aux résultats obtenus, il vous suffit d'aller cliquer sur la case concernée(uniquement dans excel et non avec "Adobe Acrobat") !!

Tests d'hypothèse sur les moyennes de population

Exercice 1

On considère que le poids des chiots bergers allemands à la naissance suit une distribution normale, de moyenne 0,150 kg et de variance 0,015. On suspecte cependant que les chiennes diabétiques mettent au monde des chiots qui ont en moyenne un poids inférieur à 0,150 kg. Afin de vérifier cette hypothèse, on a relevé le poids de 25 chiots bergers allemands nés de mères diabétiques et le poids moyen observé a été de 0,2 kg. On demande:

a) quelle est la probabilité d'observer un poids moyen aussi élevé ou plus élevé si les chiots nés de mères diabétiques obéissent à la loi générale?

b) quelle hypothèse acceptera-t-on?

Avant de résoudre l'exercice, je tiens à rappeler l'importance de bien faire la différence entre les 2 formules suivantes: 1) 2)

$$z = \frac{x - \mu}{\sigma} \qquad z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

On utilise la statistique Z car, dans ce cas-ci, la variance est connue!
La première est à utiliser lorsqu'on est en présence d'un individu possédant une valeur X, dans ce cas-ci on l'utiliserait si on devait comparer le poids d'un seul chiot avec la moyenne de population suivant la loi générale. La seconde est par contre celle à utiliser ici car nous sommes face à une comparaison de moyenne d'échantillon d'individus par rapport à la moyenne de population (obéissant à la loi générale)!

a)

Taille de l'échantillon (n):	25
Moyenne de population:	0,15
Moyenne d'échantillon:	0,2
Variance de population:	0,015
Déviation standard de pop.:	0,12247449
Z=	2,041241452

La probabilité d'avoir un poids moyen de 0,2 kg aussi élevé ou plus élevé, si les chiots obéissent à la loi générale est de: 0,02061335 soit 2,0613335%!

b)

Hypothèse: les chiennes de la race berger allemand, diabétiques mettent au monde des chiots qui ont en moyenne un poids de 0,150 kg. La probabilité d'observer un poids moyen de 0,2 kg, aussi élevé ou plus élevé, si les chiots obéissent à la loi générale, est de +/- 2,1%.
Si on travaille au seuil de rejet 2,5%, on en concluera qu'ils n'obéissent pas à la loi générale!

Exercice 2

Dans un élevage de souris, la taille des portées est en moyenne de 10 jeunes, avec déviation standard de 2,5. Combien faut-il produire de nichées pour avoir 97,5 chances sur 100 de produire au moins 50 jeunes? On admet que la taille des portées est distribuée normalement.

La formule à utiliser ici est la même que celle de l'exercice précédent, mais en la modifiant

légèrement comme suit:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\frac{X}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{X - n\mu}{\sigma\sqrt{n}}$$

car en effet, nous ne disposons pas de la moyenne et nous recherchons la taille de l'échantillon!

Une façon différente de voir la chose, mais qui mène au même résultat et vous semblera certainement plus compréhensible pour l'obtention de la formule finale ci-dessus, est de partir de la formule générale de la loi normale: $Z = \frac{X - \mu}{\sigma}$ où le X est ici une somme de variables

(de portées) égale à 50, la moyenne est ici la moyenne d'une somme de variables continues distribuées normalement et égale à la somme des moyennes et donc = $n\mu$ et la variance d'une somme de variables continues distribuées normalement est égale à la somme des variances, ajoutée de n*(n-1)/2 termes de covariance entre ces variables(nuls dans ce cas-ci car les variables sont indépendantes) et donc la déviation standard(=racine de la variance) est: $\sigma\sqrt{n}$
Voir cours théorique page 27!

La valeur de Z associée est celle correspondant à une probabilité de 0,025 puisqu'il s'agit d'un test unilatéral gauche(les 97,5 % supérieurs) et cette valeur s'obtient comme suit: -1,95996108

X:	50
Moyenne:	10
Dév. Stand.:	2,5

En remplaçant dans la formule ci-dessus, on obtient: $-1.96 = \frac{50 - 10n}{2.5\sqrt{n}}$ 1048596,48

Ce qui mène à l'équation à une inconnue, du second degré suivante:

$24.01n = 100n^2 - 1000n + 2500$ ou $100n^2 - 1024.01n + 2500 = 0$

La méthode du delta est à employer ici: $\Delta = b^2 - 4ac$ Soit= 48596,4801

Et la taille de l'échantillon se trouve comme suit: $n = \frac{-b \pm \sqrt{\Delta}}{2a}$ Soit= 6,22228047
Soit= 4,01781953

Evidement, la première solution est la bonne car la deuxième est absurde (pourquoi ?) !

Exercice 3

Un régime particulier, chez des bovins, augmente le poids en moyenne de 20 kgs a un âge donné. Chez les bovins n'ayant pas subi le régime , le poids à cet âge est N(700;50^2).

Un expérimentateur désire vérifier ce gain dans sa population et fait une expérience sur 20 bêtes qu'il soumet au régime. Calculez la puissance de l'expérience de manière théorique et empirique.

Avant tout, il faut générer un échantillon de 20 bêtes ayant subi un régime augmentant le poids de 20 kgs en moyenne, chez des animaux dont le poids suit une distribution normale de moyenne 700 kgs(lorsqu'ils ne reçoivent aucun régime particulier) et de déviation standard 50! Pour y arriver, il faut utiliser la fonction loi normale inverse("fX"dans la barre d'outils, puis "statistiques")où l'on met une probabilité aléatoire(alea()), une moyenne(espérance) de 720, vu

l'augmentation de poids en moyenne et une déviation standard(ecart-type) de 50, il ne reste plus qu'à "copier-coller" la formule pour avoir 20 bêtes prisent au hasard dans la "population" ayant subi le régime! Calculer ensuite la moyenne de cet échantillon de 20 bêtes, on peut ainsi calculer la valeur de Z associée à cet échantillon et savoir si cette valeur est significativement différente ou non, suivant le seuil de signification choisi, de la moyenne de la population n'ayant pas subi le régime (700)!

L'hypothèse nulle de départ est:" le régime n'augmente pas le poids de manière significative, autrement dit pas de différence entre l'échantillon ayant reçu le régime et la population dont le poids moyen est de 700 kgs!"
On répètera l'expérience par exemple 100 fois pour obtenir la puissance empirique, qui sera le nombre de fois où nous avons un Z significatif(quand on rejette l'hypothèse nulle) sur les 100 Z obtenus, un simple "copier-coller" suffit dans ce cas-ci, une méthode un peu plus complexe mais beaucoup plus adaptée aux calculs plus longs et lourds, sera vue plus tard, les "macros"!!

Ceci est un extrait des données générées.
Elles n'apparaîtront pas toutes à l'écran.

Poids des 20 bêtes ayant été soumises au régime:	747,178089	792,136572	724,110575	694,80046	733,945282
	604,66061	646,754526	714,238351	729,362338	715,118515
	740,217783	667,018637	681,674803	649,708906	762,445095
	771,495249	780,911816	708,793093	761,466251	812,087703
	642,363713	721,173703	768,681386	622,633135	722,13231
	776,490762	742,648351	727,977235	647,350927	753,470712
	730,256713	729,10029	671,869654	688,969632	654,585389
	767,57203	782,95295	779,718673	736,409331	601,396434
	774,107431	758,820474	700,218377	769,436949	646,783403
	688,27097	768,195488	662,149543	666,877551	789,79667
	741,985102	753,024094	710,203776	731,644488	657,098892
	723,711364	631,073473	806,775344	841,155608	717,622012
	689,120893	639,370339	700,960183	738,776177	712,930896
	709,538424	764,762032	627,872279	668,892082	734,373597
	779,743456	654,484094	697,775474	692,781893	747,304623
	693,397848	738,154765	777,609782	747,769374	726,863843
	715,495273	741,987432	695,965579	772,538894	779,266085
	677,531145	745,484155	713,621657	741,026381	776,665044
	669,859102	715,252608	686,985797	633,459078	758,530175
	743,144878	688,018699	593,836714	670,173219	813,456947
Somme :	14386,1408	14461,3245	14151,0383	14205,2327	14615,8736
Moyenne de l'échantillon:	719,307042	723,066225	707,551914	710,261634	730,793681
Moyenne de la population (sans le régime):	700				
Déviation standard de la population(sans le régime):	50				
Z associé à la comparaison entre l'échantillon et la population:	1,72687431	2,06310588	0,6754637	0,91782843	2,75427059
Significatif ou non (1 si significatif, 0 si non!), test unilatéral droit, donc sera significatif si >+1,96(au seuil 2,5 % par exemple):	0	1	0	0	1
Nombre de fois où Z est significatif:		43			

Puissance empirique: 0,43
soit: 43 %
La puissance théorique est obtenue de la façon suivante:
1) Trouver le poids de l'animal pour une valeur de Z=1,96 dans la population de poids moyen 700 kgs grâce à la formule:

$$X = \mu + Z(\frac{\sigma}{\sqrt{n}})$$

X= 721,9134662

2) Calculer la probabilité d'avoir un tel individu dans une population de moyenne 720 kgs et de déviation standard 50 grâce à la formule :

$$Z = \frac{(X-\mu)}{\sigma/\sqrt{n}}$$

Z= 0,171145618
Puissance théorique= 0,56794537
soit: 56,79453673 %

Exercice 4

Deux populations bactériennes ont une sensibilité différente à la pénicilline. La première est caractérisée par un diamètre moyen de la zone d'inhibition de la croissance de 11 mm, avec une déviation standard de 1 mm (la distribution est supposée normale). La seconde a un diamètre moyen de 8 mm, avec une déviation standard de 0,8 mm.

On désire déterminer la population d'origine d'une colonie en examinant la taille d'une plage.
On demande:
a) de calculer les valeurs seuils menant à une erreur de type 1 de 5 % et 1 %.
b) pour chacune de ces valeurs seuils, donnez l'erreur de type 2 ainsi que la puissance du test.

La première chose à faire est de choisir arbitrairement(vu que l'énoncé ne nous dit rien à ce sujet!) quelle colonie suivant telle ou telle population est prise comme hypothèse nulle, l'autre devenant l'hypothèse 1.
En effet suivant ce choix, les réponses diffèreront! Pour exemple, je choisirai la colonie suivant la population de moyenne 8 mm comme hypothèse nulle et donc celle de moyenne 11 mm devient l'hypothèse 1.

a) L'erreur de type 1(probabilité de rejeter H0 alors qu'elle est vraie) de 5 % correspond à une valeur de z=1,645 (unilatérale) et donc la valeur seuil que nous appellerons X est : 9,316

L'erreur de type 1(probabilité de rejeter H0 alors qu'elle est vraie) de 1 % correspond à une valeur de z=2,326 (unilatérale) et donc la valeur seuil que nous appellerons X est : 9,8608

b) L'erreur de type 2 (probabilité d'accepter H0 alors qu'elle est fausse) liée à la valeur seuil, pour une erreur de type 1de 5 %, se calcule de la manière suivante:
Trouver la valeur de z associée à cette valeur seuil: -1,684
Ce qui correspond à une erreur de type 2 de : 0,04609081
Et donc la puissance du test est: 0,95390919

L'erreur de type 2 (probabilité d'accepter H0 alors qu'elle est fausse) liée à la valeur seuil, pour une erreur de type 1 de 1 %, se calcule de la manière suivante:
Trouver la valeur de z associée à cette valeur seuil: -1,1392
Ce qui correspond à une erreur de type 2 de : 0,12730992
Et donc la puissance du test est: 0,87269008

Exercice 5

On a comparé le taux sanguin de créatine chez des veaux culards et chez des veaux non culards :

Culards	Non culards
4,2	2,8
3,9	2,1
3,7	3
3,5	2,3
4	2,4
	2,7

La différence entre les deux moyennes est-elle significative?

Hypothèse nulle, pas de ifférence entre culards et non culards!

	Culards	x^2	Non culards	x^2
	4,2	0,1156	2,8	0,0625
	3,9	0,0016	2,1	0,2025
	3,7	0,0256	3	0,2025
	3,5	0,1296	2,3	0,0625
	4	0,0196	2,4	0,0225
			2,7	0,0225
Somme	19,3	0,292	15,3	0,575
Moyenne	3,86		2,55	

La variance étant inconnue, il faut recourir à l'utilisation du test de t de student!

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{(\frac{(\Sigma {x_1}^2 + \Sigma {x_2}^2)}{(n_1 + n_2 - 2)}(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

avec $\Sigma x^2 = \Sigma(X - \overline{x})^2$

Et donc t= 6,970232848

Et donc cette valeur de t correspond à une probabilité de: 6,5364E-05

L'hypothèse nulle selon laquelle il n'y a pas de différence significative entre les moyennes est rejetée car ne se vérifie que dans, si l'hypothèse nulle est vraie: 0,00653635 % des cas!

Exercice 6

On désire vérifier si deux populations données, A et B, ont même moyenne. On prélève un échantillon de taille n=50 dans la population A et un échantillon de taille n=100 dans la population B. On trouve: 35,6 de moyenne pour l'échantillon provenant de A et 34,7 pour celle de B. Quelles sont vos conclusions, sachant que la variance des 2 populations est de 3.

H0: Pas de différence entre les deux moyennes!

Dans ce cas-ci, nous utiliserons un test de Z puique la variance est connue et la formule à utiliser est:

$$Z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sigma\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

x1 (barre)	35,6
x2 (barre)	34,7
variance	3
dév. stand.	1,732050808
n1	50
n2	100
Z	3

La probabilité d'une valeur aussi grande ou plus grande est de: 0,00134997 et donc, nous rejetons l'hypothèse nulle au seuil 1 %!

Exercice 7

Calculez la distribution du statistique S obtenu en prélevant deux échantillons de tailles respectives n1 et n2 dans deux populations normales de déviation standard "sigma" et de moyennes respectives "mu1" et "mu2" et obtenu en calculant:

$$S = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Cette formule, risque de vous inciter à croire que vous allez utiliser un test de Z, il n'en est rien et ce "piège" n'est pas là dans le but de vous embrouiller, mais plutôt afin de vous inciter à bien lire un énoncé, avant de vous lancer dans une résolution et à vous inviter à comprendre ce que vous faîtes, au lieu d'appliquer bêtement des formules apprises par cœur!!

Aucune donnée n'étant fournie, nous pouvons choisir la taille de l'échantillon, inventer des moyennes(mu 1 et mu 2) et une variance (sigma²). En effet, la seule contrainte que l'on retrouve dans l'énoncé est que la variance soit identique pour les 2 échantillons.
Je fixerai donc arbitrairement les valeurs qui suivent:

n1	100
n2	75
mu 1	50
mu 2	40
sigma	10

Générons les échantillons de la manière qui suit:

Echantillon 1	Echantillon 2	x1²	x2²
66,08223101	47,0758006	256,81382	69,7203115
35,36977728	57,3682565	215,709204	347,536278
54,67925929	39,9606644	21,3669513	1,52456493
44,4577894	34,7113565	31,349136	16,1168132
49,75630089	34,4370757	0,09031132	18,3942855
42,8870011	40,8802431	51,4062933	4,64105787
29,31336111	29,5338988	430,291058	84,4934679
53,76792286	30,9779945	13,7722895	60,0305303
57,95337201	28,4071485	62,3555444	106,477285
44,36967073	50,9893108	32,343659	150,390469
66,73206498	41,126034	278,063818	5,76049155
42,09023827	57,9501512	63,4664137	369,57062
52,68555596	38,9550588	6,91025631	0,05249927
47,47449237	23,1522018	6,66841244	242,54106
53,43746933	28,6670594	11,4287945	101,180911
48,85072839	44,0895429	1,45465533	28,7683263
47,81161932	38,35361	5,04092301	0,13862342

37,11844	41,4266107	167,401659	7,29366724
69,80688467	47,2908279	390,065081	73,3574475
45,96300313	29,9491513	16,759331	77,0318725
65,67909749	35,025405	244,055575	13,6938969
40,12786702	29,6125928	98,5840937	83,0529444
66,03793862	26,1553988	255,396174	158,018295
39,47387212	38,0117764	111,998771	0,51001772
53,08800736	49,2823484	9,18810096	111,437935
53,62898618	37,5675905	12,7603761	1,34175409
66,509739	14,2969521	270,698566	596,775039
45,56027888	27,490105	20,2188759	126,2438
55,79066182	20,4087565	32,8769498	335,518903
59,75819603	32,1265498	94,1167101	43,5518398
62,85675353	21,0565896	163,838316	312,205647
41,60062316	53,1014076	71,5072534	206,65431
32,84793037	39,7217856	296,145859	0,99172509
30,32981239	45,7642183	389,154802	49,5374788
55,73318175	25,1489974	32,2210905	184,333142
57,18046067	39,7651685	50,7462661	1,08001323
61,7344598	45,5890041	136,367287	47,1017639
54,11646397	43,439709	16,4807147	22,2196968
38,92797066	34,267182	123,851272	19,8804477
42,86378625	47,7412324	51,7397248	81,2756476
68,06438377	31,5294893	324,272377	51,7887819
44,82827207	40,3306923	27,3377064	2,57525676
42,87874743	33,4442712	51,524716	27,8959364
50,23308075	51,3168426	0,03106809	158,531039
30,05532689	38,6436251	400,059699	0,00677437
52,27624923	53,2550667	4,92586915	211,095767
63,22223397	46,8663667	173,328143	66,2666829
47,34226321	45,6696763	7,36881484	48,21559
50,95922132	42,8609975	0,81432938	17,0987697
67,86747816	41,2156988	317,219569	6,19894077
65,78064257	50,5510026	247,238618	139,832303
55,52490746	37,2222668	29,899988	2,26100802
37,92447968	32,5837869	147,193665	37,7259413
69,226718	47,5558773	367,485015	77,9679416
49,63550863	64,7728167	0,1775027	678,440222
65,04840839	31,6108209	224,747743	50,624801
42,78504674	50,5537765	52,8786764	139,897915
38,04732968	17,6928234	144,227841	442,391639
46,42452622	37,6785375	13,1935532	1,09703442
38,25801413	33,6002496	139,211804	26,2726163
44,14018248	37,9123004	35,0065915	0,66199578
55,81545692	43,9210136	33,161907	26,9888764
33,23291948	39,8915655	283,043606	1,35870233
59,66820153	32,5361408	92,378668	38,3135106
63,03990302	43,0577894	168,560463	18,7649918
42,63382051	36,0506898	55,1009113	7,15691885
55,09530764	44,2783768	25,3863646	30,8296471
65,51643436	32,2011284	238,999697	42,5730569
44,23893087	40,5747324	33,847826	3,41806448
47,44341039	47,3986257	6,8299062	75,2156228
60,15494036	37,7003426	101,972048	1,05183274
52,31393642	41,9997287	5,09457762	10,7177476
44,36611233	18,9938556	32,384146	389,354823

	57,2211833	37,2478122	51,3281112	2,18483705
	41,17003426	50,8513177	78,9749486	147,024988
	53,52948746		12,059424	
	47,12177896		8,61446189	
	47,48106347		6,63451816	
	38,67240265		129,604943	
	56,96470579		47,7188955	
	33,63400664		269,708776	
	54,57323495		20,3980101	
	67,81668288		315,412754	
	62,2356596		148,324151	
	48,04540494		4,04578781	
	56,66871074		43,7171082	
	38,47119969		134,226582	
	56,06862613		36,1418212	
	40,9128519		83,6121401	
	40,83893272		84,9694334	
	35,90619725		200,240105	
	53,02267154		8,79627938	
	41,83903583		67,5319655	
	45,95431746		16,8305215	
	58,7331955		75,2795037	
	22,11416485		780,791939	
	41,89490154		66,6169018	
	35,01984919		226,110471	
	60,44677447		107,951168	
	60,82919425		116,044062	
Somme:	5005,681932	2904,44487	11471,2926	7142,24545
Moyenne:	50,05681932	38,7259316		

Vu que la variance est inconnue, il faut utiliser un test de t de student et le petit s de la formule de l'énoncé est en fait :

$$s = \sqrt{\frac{(\Sigma_1{}^2 + \Sigma x_2{}^2)}{(n_1 + n_2 - 2)}}$$

s:	7,620587542
S=t:	9,733904616

Pour vérifier que cette distribution correspond bien à la distribution théorique de la statistique t, nous allons générer à l'aide d'une macro, 100 valeurs de S (le grand!). Ensuite nous calculerons les effectifs des différentes classes de S, pour pouvoir représenter graphiquement les valeurs de S en fonction des effectifs des différentes classes et ainsi vérifier qu'il s'agit bien d'une statistique t de student.

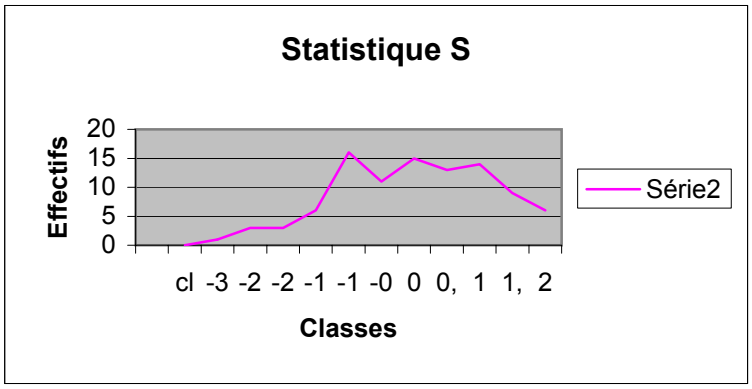
Pour visualiser la macro utilisée ici, aller dans "outils", "macros", sélectionner la macro renseignée(statistique S) et ensuite "modifier". Vous pourrez de cette manière voir la macro correpondant au calcul de plusieurs S(100 dans ce cas-ci)! Il est évident que plus vous répèterez l'opération, plus votre graphique des résultats(t en fonction des effectifs de classes)se rapprochera du graphique théorique(infinité de valeurs!).

Pour réaliser une macro vous-même, référez-vous aux notes suivantes et à la macro "statistique S":
L'accès aux macros se fait par le menu "outils"-"macros"-"nouvelle macro". Vous donnez un nom à la macro et vous l'enregistrez dans votre classeur Excell.

t	classes	effectifs
-3,98667829	-3	1

0,87154011	-2,5	3
-0,73541104	-2	3
-0,29951047	-1,5	6
-1,17130692	-1	16
-0,42356582	-0,5	11
0,07033385	0	15
2,7359448	0,5	13
-2,39229734	1	14
-0,3654141	1,5	9
1,30791233	2	6
1,06983613	2,5	0
-0,51481567	3	2
-0,35087797		
-2,4805694		
-0,95180309		
-0,21179848		
-0,23729712		
-1,13850891		
1,09065875		
0,57807316		
-1,07000678		
0,74435827		
1,78577611		
0,38592057		
-0,17073161		
-1,4639265		
0,58783402		
-1,42854713		
-0,41038491		
-2,4188872		
-0,97063218		
0,92482191		
0,66023764		
0,25345629		
0,0963112		
0,11889013		
-2,92785876		
1,23062753		
-1,53853611		
1,10463316		
-0,01283448		
-1,17837218		
-0,38355848		
-1,30085706		
0,75427697		
-0,66282509		
-2,90725513		
-1,52259813		
0,83366719		
-0,92164951		
-0,53768907		
1,00079362		
0,48228574		
-1,09413223		
0,88955999		
-0,01311208		

Celui-ci enregistre alors dans votre macro tout ce que vous entreprenez, pour vous permettre de refaire la même chose plus tard en exécutant à nouveau la macro. Dans notre cas, on arrête l'enregistrement de la macro sur le carré qui est apparu, et on va modifier manuellement le contenu de la macro("outils", "macros", choisir votre macro dans la liste et "modifier"). Une fois le texte encodé, on revient à la feuille Excel via le menu "fichier"(Alt-Q). La macro est alors prête à être exécutée("outils", "macros", choisir votre macro dans la liste et "exécuter"). Elle vous demande combien de fois vous voulez effectuer l'opération, l'adrese de la case à préserver et celle de la première case de la colonne des résultats. Une fois ces renseignements introduits, la macro effectue le nombre prescrit de simulations et range les résultats dans la colonne indiquée. Vous pouvez alors utiliser ces résultats.



Comme dit plus haut, le manque de ressemblance entre ce graphique et la courbe de Student théorique attendue est du à la taille faible de notre échantillon. Pour obtenir une courbe plus similaire, vous pouvez faire l'expérience avec plus (>1000) de valeurs de t.

1,69164458
0,00916728
1,64971305
-1,87720761
0,41378691
-1,79549681
-0,08380709
0,06219882
0,90545824
-1,26667815
1,82165551
-1,07954616
0,23443817
-2,51331284
-0,94123587
0,06609512
-1,06472262
-0,09579476
-1,15335041
1,18314534
-0,25599101
0,46700447
2,67692266
3,90921641
1,02839211
-1,94446275
1,57356872
-0,36333497
-0,97520787
0,77353761
-0,64251137
-0,72517849
1,11529278
-1,28430582
0,59341062
-1,12088127
0,13546204
-1,03207809
0,51315299
-1,7869744
0,9495699
1,51805508
-1,08759132

Exercice 8

Sur 5 individus, on a effectué une numération globulaire au mois de janvier; on a répété la même numération en avril. On a obtenu les valeurs suivantes:

Individus	**Janvier**	**Avril**
A	46	46 (x 100.000/mm³)
B	38	42
C	42	44
D	43	47

E	45	48

On demande si les 5 individus ont, en moyenne, la même numération globulaire, à ces 2 moments de l'année.

Hypothèse nulle: Pas de différence entre les numérations globulaires des mois de janvier et avril.

Dans ce cas-ci, les données sont dites pairées car il s'agit des mêmes animaux sur lesquels des dosages sont effectués à différentes époques de l'année!

Les formules à utiliser sont: $t = \frac{\overline{d}}{S_{\overline{d}}}$ $S^2_{\overline{d}} = \frac{S^2_d}{n}$ $S^2_d = \frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}$

Individus	**Janvier**	**Avril**	**d**	**d²**
A	46	46	0	0
B	38	42	4	16
C	42	44	2	4
D	43	47	4	16
E	45	48	3	9
Somme:	214	227	13	45
Moyenne:	42,8	45,4	2,6	
S²d	2,8			

t: 3,474396145 et P(t>3,47439614) = 0,02548148

Lorsque l'on regarde dans une table de t 0,025, la valeur correspondant à 4 degrés de liberté(n-1), on trouve la valeur de 2,7764. Cette valeur étant inférieure à celle trouvée, on rejette l'hypothèse nulle au seuil 5 %(bilatéral, donc 2,5 % à gauche et 2,5 % à droite de la courbe de Student!).
Excel permet de résoudre ce problème directement en utilisant la fonction TEST.STUDENT, qui donne la même probabilité que ci-dessus: 0,02548148 .

Exercice 9

H0: pas d'amélioration du caractère du chien μ1=μ

On va tester si obtenir 7.2 dans un échantillon de 10 chiens est vraisemblable sous l'hypothèse nulle d'une distribution normale N(8;3). Si la réponse est < α, H0 sera rejetée.

zs = (7.2-8)/(3/sqrt(10)) = -0,84327404

Test unilatéral gauche: P(z <= zs) = 0,19953755

Le résultat est non significatif.

Exercice 10

Mesures		47	38	39	32	34
		37	41	43	41	43
Moyenne	39,5					
Variance	20,05555556					
s	4,478342948					

ts = (xb-μ)/(s/sqrt(10)) = -1,76531682

Probabilité bilatérale: 0,11132557 =P[(t9<ts) ou (t9>|ts|)]
Probabilité unilatérale: 0,05566278 =P[(t9<ts)]=0.5*=P[(t9<ts) ou (t9>|ts|)]

t(9;0.025;bilat)=	3,11094482	μ=xb-ts*(s/sqrt(10))=	35,0943534
		μ=xb+ts*(s/sqrt(10))=	43,9056466
t(9;0.025;unilat)=	2,68501026	μ=xb-ts*(s/sqrt(10))=	35,6975519
		μ=xb+ts*(s/sqrt(10))=	43,3024481

Exercice 11

Vit A		Témoin
142		175
311		132
337		218
262		151
302		200
195		219
253		234
199		149
236		187
216		123
211		248
176		206
249		179
214		206
235,928571	Moyenne	187,642857
2946,99451	Var	1451,47802

Variance commune 2199,23626

t(18)=(xb1-xb2)/sc/sqrt(1/n1+1/n2)= 2,30233077

Test unilatéral 0,01673493

Exercice 12

Test t des données pairées

M1	M2	d
12	9	3
11	11	0
13	10	3
8	7	1
8	9	-1
		1,2

S²(db)=S²(d)/n= 0,64
S(db)= 0,8

t(4) = 1,5
Test bilatéral 0,208

Pas de différence significative entre les deux méthodes.

Exercice 13

Au sens strict, obtenir un veau femelle après une insémination est un événement binomial, l'événement contraire étant de ne pas obtenir un veau femelle.

L'événement résulte de la conjonction de la réussite de l'insémination (p1=0.6) et de l'obtention d'une femelle (p2=0.5). La probabilité de l'événement est donc p = p1*p2 = 0.3
L'événement contraire a donc une probabilité q = 1-p = 0.7, qui pourrait aussi se calculer en combinant les probabilités des situations menant à cet événement contraire:

(échec de l'insémination) **OU** (succès de l'insémination **ET** veau mâle)

soit, q = 0.4 + 0.6*0.5 = 0.7

La distribution binomiale est donc la loi à utiliser dans cette situation. L'effectif total (n) est à déterminer, et la probabilité d'avoir:
(0 veau femelle **OU** 1 veau femelle **OU** … **OU** 29 veaux femelles)
doit être inférieure à 5%, ce qui se calcule (avec un logiciel) de manière itérative:

n	126
Proba	0,05071906

La formule suivante a été utilisée pour la calcul (cfr TP)
=LOI.BINOMIALE(29;C20;0,3;VRAI)

Si on insémine 126 vaches, on a donc que 5 chances sur 100 d'avoir moins de 30 femelles.

Une méthode plus simple (mais approximative) consiste à utiliser l'approximation de la loi binomiale par la loi normale. La variable aléatoire (discrète, mais approximée par une variable continue) X est le nombre de veaux femelles. Elle sera donc supposée distribuée normalement, avec une moyenne $\mu = n*p = 0.3*n$ et une déviation standard σ = racine(npq) = racine (0.21*n). On aimerait donc calculer la valeur de n qui est telle qu'avoir un nombre de veaux femelles (X) inférieur ou égal à 30 ne se produit que dans 5% des situations. La valeur de z correspondante est z = -1.645, et on peut donc écrire:

$$z = (X - \mu)/\sigma = -1.645 = (30 - 0.3*n) / \text{racine } (0.21*n)$$

soit

$$0,21*n*(1.645)^2 = (30 - 0.3*n)^2$$

soit

$$0.09*n^2 - 18.57*n + 900 = 0$$

ce qui peut être résolu (équation du second degré) et fournit:

n1	77,8328068	soit	78
n2	128,480527	soit	129

La première valeur est rejetée car avec cet effectif, on obtient en moyenne que 23,4 veaux femelles, ce qui est incompatible avec le fait d'avoir plus de 29 femelles dans 95% des cas.
La seconde valeur est donc acceptée, et est en bonne correspondance avec la valeur calculée plus haut.

Exercice 14

On recherche cette fois n, le nombre de portées nécessaires pour obtenir 50 souriceaux (ou +) avec une probabilité d'au moins 97.5%.
Le nombre de souriceaux total après n portées est évidemment:

$$N = X1 + X2 + X3 + \ldots + Xn$$

où Xi est le nombre de souriceaux à la portée i. Xi est donc une variable aléatoire normale, de moyenne 10 et déviation standard 2.5, et N est une variable aléatoire qui est la somme de n autres variables aléatoires. On a vu que:

la moyenne d'une somme de variables aléatoires est la somme de leurs moyennes.
la variance d'une somme de variables aléatoires est la somme de leurs variances si elles sont indépendantes (ce qui est le cas ici, les tailles de portées étant supposées indépendantes l'une de l'autre).
En conséquence, la moyenne de N vaut n*10, et sa variance n*2.5*2.5.
Le problème est alors similaire au précédent:

$$z = -1.96 = (N - \mu) / \sigma = (50 - n*10) / (2.5*\text{racine}(n))$$

soit

$$6{,}25*n*(1.96)^2 = (50 - 10*n)^2$$

soit

$$100*n^2 - 1025*n + 2500 = 0$$

ce qui peut être résolu (équation du second degré) et fournit:

n1	4	soit	4
n2	6,25	soit	7

La première solution est rejetée (on a en moyenne 40 souriceaux, ce qui est trop peu); il faudra donc 7 portées au moins.

Exercice 15

Il s'agit donc dans un premier temps de confronter un résultat obtenu sur un échantillon traité à une population de moyenne et déviation standard connues. On peut utiliser la statistique z:

$$z = (X - \mu) / (\sigma/\text{racine}(n)) = 2/(6/\text{racine}(20)) = 1.49$$

Cette valeur est inférieure à la valeur seuil 1.645 (correspondant à une erreur de type 1 de 5% dans un test unilatéral - on est intéressé que par l'augmentation du poids). On a donc pas de raison de rejeter l'hypothèse nulle dans ce cas, et les espoirs sont déçus à ce stade.
Obtenir une telle valeur (ou une valeur plus grande) par hasard arrive dans 0.0681 des situations (cfr table).

L'erreur de type I est donc de 6,81%, supérieure au seuil toléré de 5%. On accepte H0.

Si en réalité l'effet existe (et vaut 2 kilos), on peut mesurer la puissance du test, qui consiste donc à essayer de détecter cet effet avec 20 individus:

a) l'hypothèse nulle (pas d'effet) sera (correctement) rejetée si la moyenne obtenue sur l'échantillon (M) est telle que:

$$(M - \mu) / (\sigma/\text{racine}(n)) = (M - 40) / (6/\text{racine}(20)) > 1.645$$

soit

M > 40 + (6/racine(20))*1.645 = 42.207

b) la puissance du test est la probabilité d'obtenir cette valeur (ou une valeur plus grande) dans la vraie population (c'est à dire celle correspondant à H1) dont est issu l'échantillon.

z = (M - μ) / (s/racine(n)) = 0.207 / (6/racine(20)) = 0.154

La probabilité d'être plus grand que cette valeur de z est (cfr table) 0,439.
La puissance du test est donc de 43.9 %

c) Pour avoir une puissance de 90%, il faudrait que la valeur de z calculée au point (b) soit (cfr table) de -1,28.

On sait que:

M > 40 + (6/racine(n))*1.645 (*)

et

z = -1.28 = (M - 42) / (6/racine(n)) (**)

tilisant (*) et (**), on peut écrire:

40 + (6/racine(n))*1.645 = 42 - (6/racine(n))*1.28

soit

racine(n) = 8.775 n = 77,000625

soit

n = 77

Exercice 16

Les données sont les suivantes:

	Elevée	Faible
	134	70
	146	118
	104	101
	119	85
	124	107
	161	132
	107	94
	83	
	113	
	124	
	97	
	123	

Estimation			
	Moyennes	119,583333	101
	Variances	451,356061	425,333333
	Effectif	12	7

On veut comparer les moyennes de gains pondéraux des deux groupes. Il s'agit donc de données non pairées (les souris des deux "colonnes" sont différentes), et on désire comparer deux groupes, la variance dans la population générale étant inconnue (elle sera estimée sur les échantillons). On a donc recours à un test de t pour données non-pairées.

La variance dans les groupes est calculée comme une moyenne pondérée des variances à l'intérieur de chaque groupe, les facteurs de pondération étant les effectifs (moins 1) de chacun des groupes. On obtient donc:

s^2 = (11*451.356 + 6*425.333)/(11+6) = 442,171569

L'effectif à considérer est la moyenne harmonique des effectifs de groupes (cfr théorie).

1/n = 1/n1 + 1/n2 = 1/12 + 1/7 = 19/84

Finalement,

t = (μ1 - μ2) / (s/racine(n)) = (μ1 - μ2) / (racine(442,172)/racine(84/19))

soit

t = (119.583 - 101) / (racine(442,172)/racine(84/19)) = 1,858194

avec

(n1-n2-2) = 17 degrés de liberté

La valeur seuil (α = 5%) de t, pour 17 degrés de liberté, est: 1.7396 pour un test unilatéral, et 2.1098 pour un test bilatéral.

Si la question est de savoir si le régime à teneur protéique élevée augmente le gain pondéral entre les 24eme et 84eme jours d'âge, la réponse (au seuil 5%) est oui (augmentation significative car 1.8582 > 1.7396). Si la question est (comme dans l'énoncé): la différence entre les gains pondéraux des deux groupes est elle significative, la réponse (au seuil 5%) est non (différence non significative car 1.8582 < 2.1098).

A noter que l'utilisation du test de t implique que les variances dans les deux groupes soient identiques (remarquez que les estimateurs des deux variances sont proches) et que les données aient une répartition normale **dans** les groupes.

Exercice 17

Dans une analyse de ce type, diverses hypothèses peuvent être testés. Celle qui nous intéresse directement est évidemment l'hypothèse selon laquelle le médicament n'a aucun effet sur la tension artérielle, ce qui pourrait s'écrire:

H0: μ(m) = μ(p), i.e. la tension artérielle des individus ayant reçu le médicament est en moyenne égale à celle de ceux qui ne l'ont pas reçu.

D'autres effets (de nuisance) peuvent néanmoins masquer partiellement l'information que l'on souhaite tirer de l'expérience. Par exemple, les inévitables variations de tension artérielle entre individus pourraient partiellement cacher l'effet s'il existe (variation dans les groupes).
Le remède à ce problème potentiel est d'utiliser des données pairées, ce qui a été fait ici, les variations individuelles étant balancées dans les deux groupes.
Un autre problème potentiel est que le jour auquel on prend le médicament pourrait influencer les résultats. Pour balancer les données, la moitié des individus a reçu le médicament le premier jour et le placebo le 2eme jour, et vice-versa. Egalement, pour éviter l'influence de la prise du médicament le premier jour sur l'observation du 2eme jour, les deux prises doivent être suffisamment espacées pour qu'on puisse exclure toute interférence.
Il pourrait également y avoir une interaction entre le jour de prise et la substance (en d'autres mots, l'effet du médicament n'est pas le même selon le jour auquel il est pris). Il est possible de tester pour ce genre d'effets (voir plus loin dans le cours).

a) Test de l'effet du jour:

H0: μ(j1) = μ(j2)

| Jour 1 | Jour 2 |

	tension artérielle	tension artérielle	différence
Individu 1	17	10	7
Individu 2	17	15	2
Individu 3	15	13	2
Individu 4	13	13	0
Individu 5	12	11	1
Individu 6	14	17	-3
Individu 7	15	13	2
Individu 8	12	19	-7
Individu 9	11	11	0
Individu 10	16	16	0

Moyennes	14,2	13,8	0,4
Variances			13,1555556

La variance des différences moyennes est 13.155555/10 = 1.3155555, ce qui donne une déviation standard des différences moyennes de:

Sdb = 1,1469767

La statistique t vaut donc: db / sdb, soit: 0,34874292 avec 9 degrés de liberté.
Cette valeur est évidemment non significative, ce qui signifie bien qu'il n'y a pas de différence significative de tension artérielle due au jour auquel se fait la prise (pas d'effet 'jour')

On peut maintenant tester l'effet du médicament en réécrivant le tableau de la manière suivante (la notion de jour a été supprimée):

	préparation	tension artérielle	préparation	tension artérielle	différence
individu 1	p	17	m	10	7
individu 2	p	15	m	17	-2
individu 3	p	15	m	13	2
individu 4	p	13	m	13	0
individu 5	p	12	m	11	1
individu 6	p	17	m	14	3
individu 7	p	15	m	13	2
individu 8	p	19	m	12	7
individu 9	p	11	m	11	0
individu 10	p	16	m	16	0

Moyennes	2
Variances	8,88888889

La variance des différences moyennes est cette fois 8.8888/10 = 0.88888, ce qui donne une déviation standard des différences moyennes de:

Sdb = 0,94280904

La statistique t vaut donc: db / sdb, soit: 2,12132034 avec 9 degrés de liberté.
Cette valeur est significative a = 5%, unilatéral), ce qui signifie bien qu'il y a une différence significative de tension artérielle due au traitement. Le médicament a un effet sur la tension.