TP 4 – Statistiques bivariées

Correction

Exercice 1 Données : tabac.xls

Question 1 :

1) partir du caractère « nombre de cigarettes », créer une nouvelle colonne prenant la valeur « fumeur » si le nombre de cigarettes est strictement supérieur à 0 et « non fumeur » sinon.

2) Etablir le tableau de contingence des effectifs entre « problèmes pulmonaires » et « fumeurs ».

3) Quel pourcentage de l'échantillon est fumeur ? Quel pourcentage de l'échantillon présente des problèmes pulmonaires ? Quel pourcentage de l'échantillon est non fumeur et avec des troubles pulmonaires ?

Correction

	Fumeur			
Troubles pulmonaires	fumeur	non fumeur	(vide)	Total général
FALSE	19	46		65
TRUE	30	5		35
(vide)				
Total général	49	51		100

49% de l'échantillon est fumeur.
35% de l'échantillon présente des troubles pulmonaires
5% de l'échantillon est non fumeur avec des troubles pulmonaires.

Question 2 :

1) Etablir le profil ligne.

2) Parmi les personnes présentant un trouble pulmonaire, quel pourcentage est fumeur ?

3) Comparer avec les fréquences marginales. Calculer un indice vous permettant de confirmer vos impressions.

Correction

Troubles pulmonaires	fumeur	non fumeur
FALSE	0,29	0,71
TRUE	0,86	0,14
Freq. Marg.	0,49	0,51

Parmi les personnes présentant un trouble pulmonaire, 86% sont fumeurs.

Dans la population totale, il y a environ 50% de fumeurs et donc 50% de non fumeurs. Dans le cas où fumer n'aurait pas de conséquence sur la santé, on devrait retrouver cette proportion pour chacune des sous-populations avec ou sans trouble pulmonaire. Or cela n'est pas le cas. On peut donc en conclure qu'il y a un lien entre ces caractères.

Les deux caractères ayant le même nombre de modalités, on peut utiliser le coefficient de contingence. On trouve CC=0,47.

Exercice 2 Données : herbicide.xls

Question 1 :

1) Représenter la moyenne du taux de plants survivants en fonction de l'herbicide et de la plante.

Correction

Moyenne de taux de survivants	plante			
herbicide	ble	chiendent	liseron	(vide)	Total général
aucun	0,93	0,95	0,94		0,94
herbicide1	0,85	0,23	0,78		0,62
herbicide2	0,74	0,69	0,25		0,56
herbicide3	0,11	0,16	0,07		0,11
(vide)					
Total général	0,66	0,50	0,51		0,56

2) Quel herbicide vous semble le plus efficace et quel est celui qui est le plus néfaste ?

Correction

L'herbicide 3 a un taux de survivants très faible mais détruit tout y compris le blé. Les herbicides 1 et 2 ont a peu près le même taux. L'herbicide 1 semble plus inoffensif pour le blé. L'efficacité dépend du type de mauvaises herbes, chiendent pour l'herbicide 2 et liseron pour l'herbicide 1.

3) Pouvez-vous vous faire une idée a priori des rapports de corrélation entre le taux et la plante puis entre le taux et l'herbicide. Calculer ces rapports et vérifier.

Correction

Le taux de plants survivants varie assez peu selon le type de plante (sauf pour le blé qu'on ne souhaite pas détruire) alors qu'il varie beaucoup selon l'herbicide. On devrait donc trouver un rapport de corrélation proche de 1 entre taux et herbicide et un rapport plus faible entre taux et plante. Effectivement, on trouve un rapport de corrélation taux-plante de 0,2 et un rapport de corrélation taux-herbicide de 0,81.

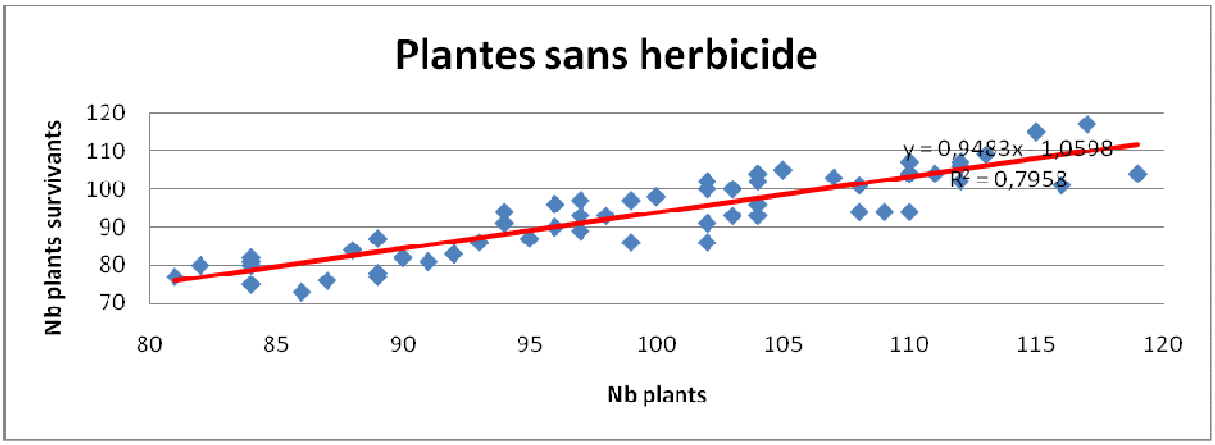
Question 2 :

1) Dans le cas où les plantes n'ont subi aucun traitement, établir le lien entre le nombre de plants (x) et le nombre de plants survivants (y). En déduire le nombre de plants survivants dans le cas ou le nombre de plants est 100.

Correction

Y=0,9483x-1,0598

Y=0,9483*100-1,0598=94



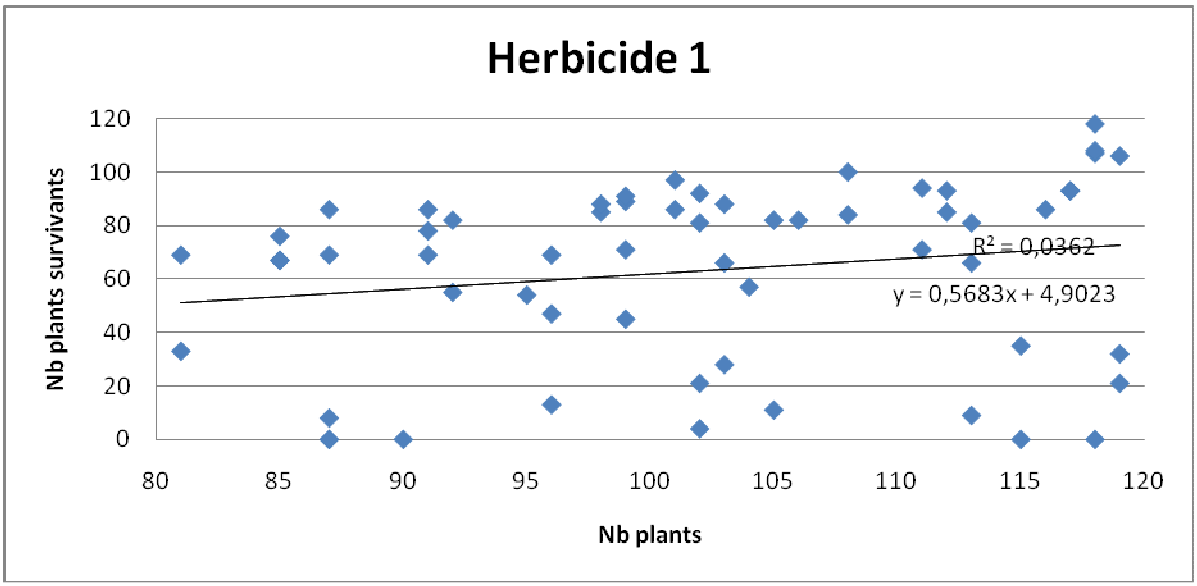
2) Quel pourcentage de la variance de y est expliquée par la droite de régression ? Conclusion.

Correction

R^2=0,7958 donc 79% de la variance du nombre de plants survivants est expliquée par la droite de régression. On peut considérer ce modèle fiable.

Question 3 :

Dans le cas de l'herbicide 1, peut-on établir un lien entre nombre de plants et nombre de plants survivants ? Expliquer pourquoi. Que faudrait-il faire pour déterminer ce lien ?



Correction

R^2=0,0362 donc uniquement 3,6% de la variance du nombre de plants survivants est expliquée par la droite de régression. Ce modèle n'est donc pas utilisable.

Sur le nuage de points, on constate qu'il y a deux ensembles de points, un situé entre 40 et 120 et l'autre avec très peu de survivants. En revenant au tableau de la question 1, on note que l'herbicide 1 est très efficace sur le chiendent uniquement. Il faudrait donc faire un modèle sur le sous-ensemble « chiendent » et un sur le sous-ensemble « blé-liseron ».