Chapitre 2 :Mesure de la liaison entre une variable et un ensemble de variables

Exercice 19

Nous étudions ici un croisant de classe d'âge et de diplôme, pour 90 individus.

Table 1: tableau de contingence de la classe d'âge croisé avec le diplôme

	BEPC	BAC	Licence	Total
Plus de 50 ans	15	12	3	30
Entre 30 et 50 ans	10	18	4	32
Moins de 30 ans	15	5	8	28
Total	40	35	15	90

Le tableau 1 nous donne les effectifs croisée de nos deux variable. On apprend par exemple que dans nos données il y a 40 individus avec un BEPC, et que 15 d'entre eux ont plus de 50 ans.

Table 2: Tableau des fréquences croisées en pourcentage

	BEPC	BAC	Licence	Total
Plus de 50 ans	16.7	13.3	3.3	33.3
Entre 30 et 50 ans	11.1	20.0	4.4	35.6
Moins de 30 ans	16.7	5.6	8.9	31.1
Total	44.4	38.9	16.7	100.0

On peut obtenir les fréquences de notre tableau 1. C'est ce qu'illustre le tableau 2, on voit par exemple que 35.6% de nos individus ont entre 30 et 50 ans , et 20% d'entre eux on un bac.

Table 3: profils lignes en pourcentage

	BEPC	BAC	Licence	Total
Plus de 50 ans	50.00	40.00	10.00	100
Entre 30 et 50 ans	31.25	56.25	12.50	100
Moins de 30 ans	53.57	17.86	28.57	100
Total	44.44	38.89	16.67	100

Le tableau 3 nous donne les fréquence sachant la trache d'age. On voit que 28.57% des moins de 30 ans on une licence.

Table 4: Profils collones en pourcentage

	BEPC	BAC	Licence	Total
Plus de 50 ans	37.5	34.29	20.00	33.33
Entre 30 et 50 ans	25.0	51.43	26.67	35.56
Moins de 30 ans	37.5	14.29	53.33	31.11
Total	100.0	100.00	100.00	100.00

Le tableau 4 nous donne les fréquences sachant le diplome, on apprend que pour ceux ayant un bac, 51.43% ont entre 30 et 50 ans.

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 11.175, df = 9, p-value = 0.2639
```

On retrouve une pvlaur de 0.2639, une valeur conséquente qui nous permet de conclure sur le non rejet de H0, et d'en déduire qu'il y a independance entre les deux variables.

Exercice 20

Pour une population d'effectif de taille 1000 on a mesuré les deux variables qualitatives "Couleur des yeux" et "Etat matrimonial".

Table 5: Tableau de contingence de la Couleur des yeux et Etat matrimonial

	Bleu	Brun
Celib	290	410
Marie	110	190

Le tableau 5 nous apprend que parmis nos 1000 individus 290 sont celibataire avec les yeux bleu. Ou encore que 190 sont mariés avec les yeux brun.

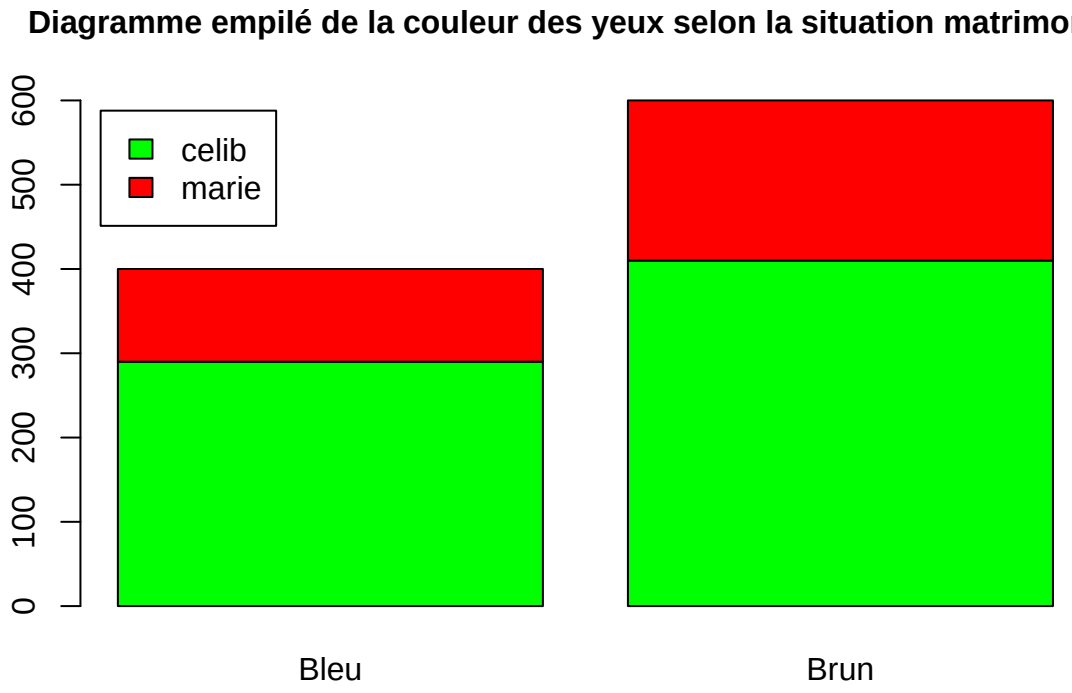

Figure 1: Diagramme empilé de la couleur des yeux selon la situation matrimoniale

On peut rendre graphique le tableau 1. Grace à la figure 1 on remarque que nous avec moins d'individus aux yeux bleu que aux yeux brun.

Quelque commande R utile sur nos données :

n <- margin.table(tableau) => effectif total

m1 <- margin.table(tableau,1) => lois marginale de l'état matrimonial

m2 <- margin.table(tableau,2) => lois marginale de la couleur des yeux

prop.table(tableau) => tableau de contingence en fréquence

Table 6: tableau de contingence en pourcentage

	Bleu	Brun
Celib	29	41
Marie	11	19

Voici par exemple le tableau tableau de contingence en pourcentage cette fois. On apprend que 41% des indivudus sont celibataire au yeux brun.

Table 7: Tableau des effectifs tehorique

	Bleu	Brun
Celib	280	420
Marie	120	180

Le tableau 7 nous montre les effectifs théorique, c'est à dire les effectifs si nos variables étaient parfaitement indépandantes.

```
## Number of cases in table: 1000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.9841, df = 1, p-value = 0.159
```

La stat du khi2 mesure l'écart entre le tableau de contingence et le tableau des effectif théorique.

Les résultat du test du khi2 indique une pvaleur supperieur à 0.05 ce qui nous permet de ne pas rejet H0, l'hypothèse d'indépendance.

```
## Number of cases in table: 1000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.154e-29, df = 1, p-value = 1
```

Si on réalise le test sur le tableau des effectifs théoriques, on trouve une pvaleur de 1. Ce qui prouve que ces données reflete l'indépandence parfaite.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tableau2
## X-squared = 995.84, df = 1, p-value < 2.2e-16
```

quand on réalise le test sur un tableau truqué, ou tous les individus aux yeux bleus sont mariés et tous les autres sont célibataires, on trouve une pvaleur très petit. Donc un rejet de H0, il y a une forte dépendance entre les deux variables.

Exercice 21

Pour cet exercice nous utiliserons le jeu de données car directement implanté dans R.

Table 8: Extrait du jeu de donnée cars

speed	dist
4	2
4	10
7	4
7	22
8	16

speed	dist
9	10

Ces données indiquent la vitesse de 50 voitures et les distances nécessaires pour s'arrêter. Notez qu'elles ont été enregistrées dans les années 1920. Nous retrouvons un extrait de ces données dans le tableau 8.

La Matrice est de taille 50 lignes 2 collones. Il y a donc deux variables qui sont "speed" vitesse en mph, et "dist" distance en ft.

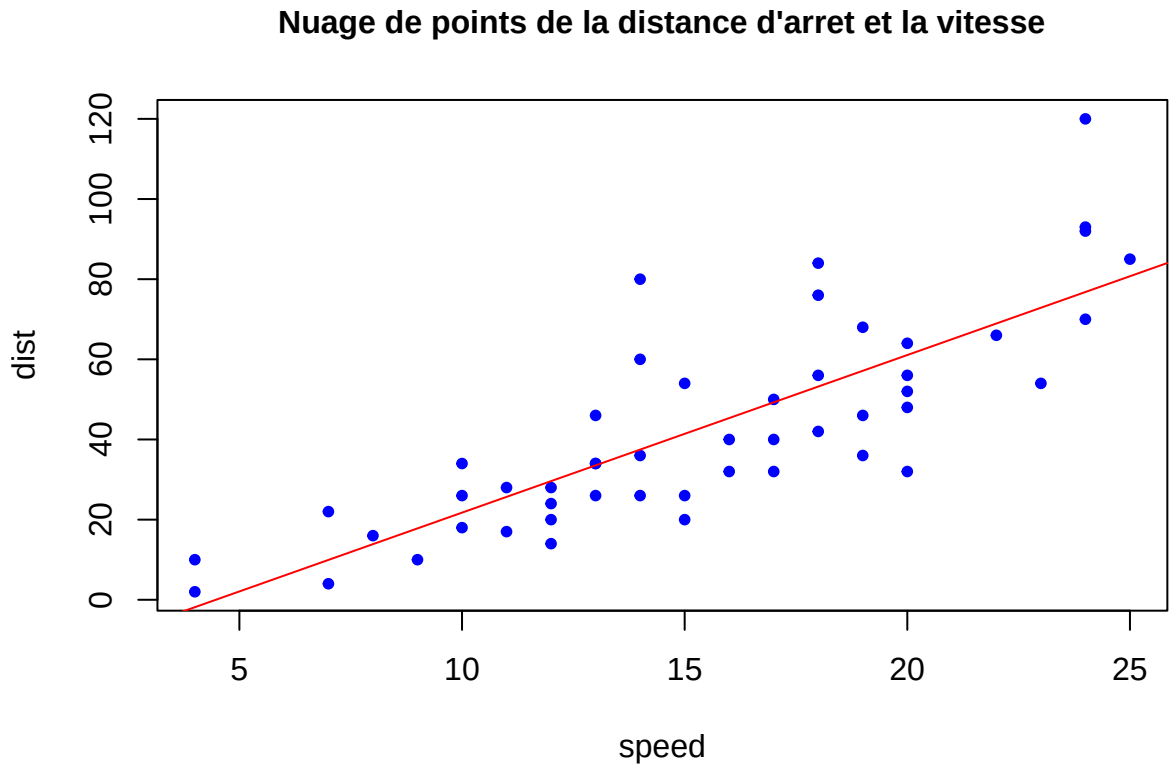

Figure 2: Nuage de points de la distance d'arret et la vitesse

Un nuage de points est une bonne representation entre deux variables quantitative. La figure est donc adapter à nos données.Les points semble liée linéairement de maniere postive et croissante.La droite de regression,en rouge, est celle qui passe le plus près de tout les points.

Le modele predit une distance de frainage de 61.07 ft pour une vitesse de 20 mph.

Intervalle de confiance : [55.25 , 66.89]

Intervalle de prédiction : [29.6 , 92.54]

L'exemple cars est adapté à la sélection de modèles.

update() : Va mettre à jour et (par défaut) réajuster un modèle.

step() : Sélectionnez un modèle basé sur une formule par AIC.