TP 3–Statistiques descriptives – Tableau de contingence

On souhaite se familiariser avec le tableau de contingence d'une série double. On prend pour exemple la série des poids et tailles de bébés :

	Y : **Poids à la naissance**				
X : **Sexe**	**Faible (0,5-2 kg)**	**Moyen (2-3 kg)**	**Élevé (3-4 kg)**	**Très élevé (4 kg et +)**	**Total**
Garçons	830	8 615	30 784	4 839	45 068
Filles	862	11 183	27 566	2 348	41 959
Total	1 692	19 798	58 350	7 187	87 027

SOURCE : Bureau de la Statistique du Québec

3.1 Saisie du tableau de contingence

Le tableau de contingence est une matrice sous R à laquelle nous allons rajouter quelques éléments cosmétiques.

(i) On saisit la matrice ayant pour coefficients les effectifs dans l'ordre du tableau ci dessus. On peut choisir entre deux façon de déclarer :

- concaténer des vecteurs verticaux ("c" pour colonne) :

```
poidstaille=cbind(c(830,862),c(8615,11183),c(30784,2756),c(4839,2348))
```

- concaténer des vecteurs verticaux ("r" pour raw (ligne)) :

```
poidstaille=rbind(c(830,8615,30784,4839),c(862,11183,27566,2348))
```

(ii) On ajoute la colonne modalité de X :

```
rownames(poidstaille)=c("Garçons","filles")
```

(iii) On ajoute la ligne modalité de Y :

```
colnames(poidstaille)=c("Faible","Moyen","Elevé","Tr.élv")
```

3.2 Tableau des fréquences

(i) On peut obtenir l'effectif total de la population par la fonction `sum()` :

```
sum(poidstaille)
```

(ii) On peut obtenir le tableau des fréquences. Il ne faut pas oublier que notre tableau est une matrice et que l'on peut effectuer toutes les opérations usuelles comme diviser la matrice `poidstaille` terme à terme par l'effectif total `sum(poidstaille)`

```
frequences=poidstaille/sum(poidstaille)
```

On obtient :

	Faible	Moyen	Eleve	Tr. elv
Garçons	0.009537270	0.09899227	0.3537293	0.05560343
filles	0.009904972	0.12850035	0.3167523	0.02698013

Remarque 3.1 *On aurait pu obtenir ce tableau par la commande* `prop.table()` *:*

```
prop.table(poidstaille)
```
.

3.3 Distributions conditionnelles

(i) La distribution de Y pour les garçons est donnée par la première ligne de la matrice `poidstaille` soit

```
poidstaille[1,]
```

(on note qu'on fixe la première ligne par le 1, mais on fait dérouler les colonnes). On obtient

Faible	Moyen	Eleve	Tr. elv
830	8615	30784	4839

En calculant les fréquences par rapport à l'effectif de garçons `sum(poidstaille[1,])`, on obtient la distribution conditionnelle de Y étant donné X =garçon :

Faible	Moyen	Eleve	Tr. elv
0.01841661	0.19115559	0.68305671	0.10737108

obtenu par `poidstaille[1,]/sum(poidstaille[1,])`.

(ii) De même, on obtiendra la distribution conditionnelle de Y étant donné X =fille par `poidstaille[2,]/sum(poidstaille[2,])`.

(iii) Tracer sur le même diagramme en bâtons ces deux distributions conditionelles. Que peut-on conclure intuitivement sur la dépendance entre X et Y ?

Réponse : .

On note qu'utiliser des fréquences ou des effectifs ne change que l'échelle. On choisit donc de tracer les effectifs. On utilisera la commande `barplot(poidstaille,beside=TRUE)`.

On observe un léger décallage du pois des bébés filles vers les faibles poids. Ce n'est pas flagrant.

3.4 Distributions marginales

(i) On peut obtenir les colonnes et lignes "Total" (les marges) par la commande `addmargins` :

```
frequences=addmargins(frequences)
```
.

(ii) La série de fréquences de Garçon et de filles dans la population est appelée distribution marginale de X :

Garçons	filles	Sum
0.5178623	0.4821377	1.0000000

.

Il s'agit de la colonne `sum` que l'on extrait par la commande `frequences[,5]` (on note qu'on fixe la dernière colonne par le 5, mais on fait dérouler les lignes).

(iii) De même, on obtient la distribution marginale de Y à partir de la ligne `sum` en extrayant de `frequences` la dernière ligne (on fixe la troisième ligne et on déroule les colonnes) grâce à `frequences[3,]` :

Faible	Moyen	Eleve	Tr. elv	Sum
0.01944224	0.22749262	0.67048157	0.08258357	1.00000000

3.5 V de Cramér

On se propose d'utiliser un indicateur d'association appelé V de Cramér pour vérifier l'observation de la dépendance entre les variables observée dans la section 3.3. Plus V est proche de zéro, plus il y a indépendance entre les deux variables X et Y étudiées. Il vaut 1 en cas de complète dépendance.

Le coefficient V de Cramér nécessite l'utilisation de la statistique du χ^2. La statistique du χ^2 est disponible via le test du même nom :

`chisq.test(poidstaille)` .

On s'aperçoit que R donne plusieurs valeurs et non seul la statistique. Nous verrons la signification de ces valeurs dans les cours ultérieurs. Nous rappelons la formule du V de Cramér :

$$V = \sqrt{\frac{D_\chi^2}{n \times min\{l-1; c-1\}}},$$

où n est l'effectif total de la population, c est le nombre de colonnes (nombre de modalités de Y) et l le nombre de lignes (modalits de X).

On se propose de définir une fonction `Cramer` : (*taper sans fautes !*)

`cramer=function(table){`	La fonction s'appelle Cramer, la variable à qui elle s'applique sera nommée `table` durant la programmation de la fonction
`test=chisq.test(table)`	Le test du χ^2 est stocké dans la variable `test`
`chi2=as.numeric(test$statistic)`	Nous ne prenons que la variable `statistic`[1] dans `test`, on l'affecte à `chi2`
`n=sum(table)`	L'effectif total stocké dans `n`
`c=length(table[1,])`	Le nombre de colonnes est la longueur d'une ligne de `table`
`r=length(table[,1])`	Le nombre de lignes est la longueur d'une colonne de `table`
`m=min(c,r)`	
`V=sqrt(chi2/n*(m-1))`	
`return(V)`	Ne pas oublier de faire afficher `V`
`}`	Fin de la déclaration de la fonction

1 Taper `help(chisq.test)` pour obtenir le mode d'emploi en ligne de `chisq.test`.

Il reste à appliquer notre fonction `Cramer` à notre tableau de contingence `poidstaille` pour lire le V de Cramér :

`cramer(poidstaille)`.

On donne le tableau suivant pour l'interprétation de la valeur du V de Cramér :

Valeur du V de Cramér	Intensité de la relation entre les variables
inférieur à 0,10	relation nulle ou très faible
entre 0,10 et 0,20	relation faible
entre 0,20 et 0,30	relation moyenne
au dessus de 0,30	relation forte

Interpréter le résultat.