Université Claude Bernard Lyon 1

L3 MOS (Semestre 6) statistiques
Session 1 13 Mai 2008

Corrigé de l'examen CCF2 de statistiques

Voir le fichier `FPS0.xls` disponible sous spiral.
Pour plus de lisibilité tous les tableaux et les figures sont renvoyés à la fin du corrigé, à partir de la page 4.

Correction de l'exercice 1.

- Pour la variable numérique année de naissance (appelée `Q28_annee`), on peut tracer un histogramme ou une boîte de dispersion. Voir la figure 1 page 4.
  La ligne de points n'est pas adaptée ici, car le nombre de sujets est trop important (318).
  On peut déterminer la moyenne, l'écart-type et les quartiles (voir chapitre 4) : on obtient

```
    mean       sd   0%  25%  50%  75% 100%   n NA
1970.338 13.70772 1933 1962 1969 1982 1997 317  1
```

- Cette variable est difficile d'interprétation car l'année moyenne de naissance par exemple n'est pas très evocatrice. On introduit une nouvelle variable âge égale à l'année du questionnaire (supposé avoir eu lieu en 2008) moins l'année de naissance. Voir l'histogramme ou la boîte de dispersion en figure 2 page 4.
  Les moyenne, l'écart-type quartiles sont

```
    mean       sd 0% 25% 50% 75% 100%   n NA
37.66246 13.70772 11  26  39  46   75 317  1
```

  On observe sur l'histogramme un pic important à 45 ans et un autre plus faible vers 15 ans. La moyenne est égale à 37.6625 (qui au passage vaut 2008 moins la moyenne des années de naissance, soit $2008 - 1970.3375 = 37.6625$ et un écart type de 13.7077 (qui au passage vaut l'écart-type des années de naissance).

Correction de l'exercice 2.

On étudie maintenant la variable catégorielle `Q3j_prudence`.

- On étudie avec Rcommander, la variable catégorielle `Q3j_prudence`. Si on cherche à dénombrer les catégories, cette variable n'apparaît pas car elle est considérée par Rcommander, comme numérique (prenant les deux valeurs 0 ou 1).
  Il faut la transformer en variable en introduisant une nouvelle variable transformée en facteur en tapant par exemple en ligne de commande dans R

```
as.factor(Q3j_prudence)
```

  ou mieux

```
factor(FPS0$Q3j_prudence, labels = c("Imprudente", "Prudente"))
```

On appellera cette nouvelle variable `fact_Q3j_prudence`

On peut alors dénombrer les catégories ; on trouve

```
Imprudente   Prudente
       181        137
```

Il y a en tout 2 catégories. Les fréquences associées sont :

```
fact_Q3j_prudence
Imprudente   Prudente
  56.91824   43.08176
```

soit 56.92 % pour la valeur Imprudente et 43.08 % pour la valeur Prudente.

- Avec Rcommander, on peut aussi tracer le camembert ou éventuellement un graphe en barre. Voir les figures 3 page 5 et 4 page 6.

Correction de l'exercice 3.

On croise maintenant la variable catégorielle prudence (appelée `fact_Q3j_prudence`) définie dans l'exercice 2 et la variable numérique âge définie dans l'exercice 1. Voir le chapitre 5.

(1) On peut tracer une collection de boîte de dispersion (par groupe) : voir figure 5 page 6.

On peut aussi faire des statistiques par groupe :

```
               mean       sd 0% 25% 50% 75% 100%   n NA
Imprudente 36.48066 14.34596 13  24  37  47   73 181  0
Prudente   39.23529 12.69076 11  31  40  45   75 136  1
```

Comme on peut s'y attendre *a priori*, la moyenne des âges des femmes imprudentes est plus faible que celui des prudentes.

(2) Confirmons cela grâce à R.

On utilise l'ajustement de modèle.

La valeur du rapport de corrélation RC est égale à 0.00992357 ; ainsi, au vu des seuils conventionnels proposés par Cohen (0.01,0.05,0.15), la liaison entre les deux variables peut être considérée comme faible. Par ailleurs, la probabilité critique est égale à 0.0765552, supérieure au seuil de 0.05 ; ainsi, la liaison entre les deux variables peut être considérée comme statistiquement non significative.

On peut donc conclure que la prudence ne dépend pas de l'âge, contrairement à l'observation empirique faite !

Correction de l'exercice 4.

Croisons maintenant les deux variables catégorielles prudence (appelée `fact_Q3j_prudence`) et niveau de pratique.

Voir chapitre 7.

Ici, le tableau croisé était fourni dans l'énoncé et il faut le rentrer à la main dans Rcommander. On pourrait aussi l'obtenir en utilisant la variable `Q6_niveau` avec

– 1 = Débutante ;
– 2 = Débrouillée ;
– 3 = Confirmée ;
– 4 = Experte.

Voir le tableau 1 page 5.

On obtient

```
        Pearson's Chi-squared test

data:  xtable1
X-squared = 10.8704, df = 3, p-value = 0.01245
```

On peut donc faire une statistique du chi-carré et on obtient $X^2 = 10.8704$ On obtient donc une taille d'effet

$$w = \sqrt{\frac{X^2}{n}} = \sqrt{\frac{10.8704}{318}} = 0.184888.$$

Pour la probabilité critique, on obtient 0.0124.
La valeur de la taille d'effet w est égale à 0.1849 ; ainsi, au vu des seuils conventionnels proposés par Cohen (0.1,0.3,0.5), la liaison entre les deux variables peut être considérée comme moyenne. Par ailleurs, la probabilité critique est égale à 0.01245, inférieure au seuil de 0.05 ; ainsi, la liaison entre les deux variables peut être considérée comme statistiquement significative.
On conclut donc que la prudence dépend du niveau de pratique.

Correction de l'exercice 5.
On se réfère aux chapitres 3 et 4 de l'ouvrage [Cha04].

(1) Les trois composantes sont :
   - *Les unités expérimentales* : l'ensemble des 219 sujets ;
   - *Les facteurs* : les deux facteurs étudiés sont l'exercice et le régime alimentaire ;
   - *Les réponses* : 4 mesures différentes faites à partir du sang.

(2) Ces mesures sont au nombre de 4 afin de déterminer avec certitude l'état de santé général de chacun des sujets.

(3) Le dispositif est un *dispositif factoriel* car on forme des groupes avec toutes les combinaisons possibles des facteurs. Pour p facteurs, il a 2^p possibilité. Ici deux facteurs $p = 2$ donc $2^2 = 4$ groupes, qui peuvent définis en numérotation binaire (1 : présence du facteur, 0 : absence, premier chiffre : premier facteur, second chiffre : second facteur) : 00, 01, 10, 11. On compte en base 2 de 0 à $2^p - 1$.

(4) Cette expérience se déroule sur une durée importante puisque les effets des deux facteurs sont *a priori* longs à se manifester. Le danger de cette durée est que les sujets se lassent du traitement subi ; dans ce cas, ils risquent de ne pas mener l'expérience à son terme.

(5) Pour un *dispositif factoriel*, l'ajout d'un facteur multiplie par deux le nombre de groupe ; pour $p = 3$, il a $2^3 = 8$ groupes. De façon pratique, on augmente éventuellement le nombre total de sujets, on divise chaque groupe existant en deux sous groupes ; on donne aux premiers un placebo et aux derniers le médicaments. En numérotation binaire, les groupes sont 000, 001, 010, 011, 100, 101, 110 et 111.

Références

[Cha04] Stéphane Champely. *Statistique vraiment appliquée au sport : cours et exercices*. Sciences et pratiques du sport sciences. De Boeck, 2004. Disponible à la Bibliothèque universitaire de Lyon I sous la cote 519.5 CHA.

Ensemble des tableaux et des figures

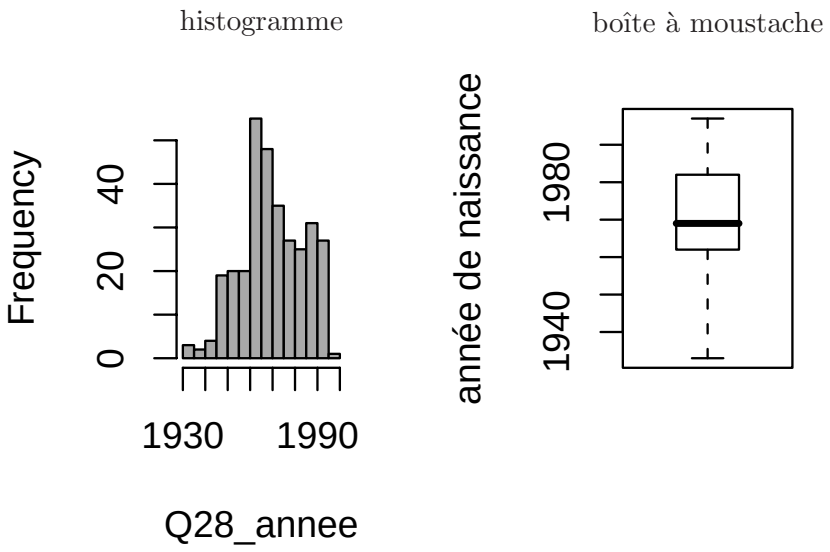

FIG. 1. La boîte de dispersion et l'histogramme de la variable année de naissance.

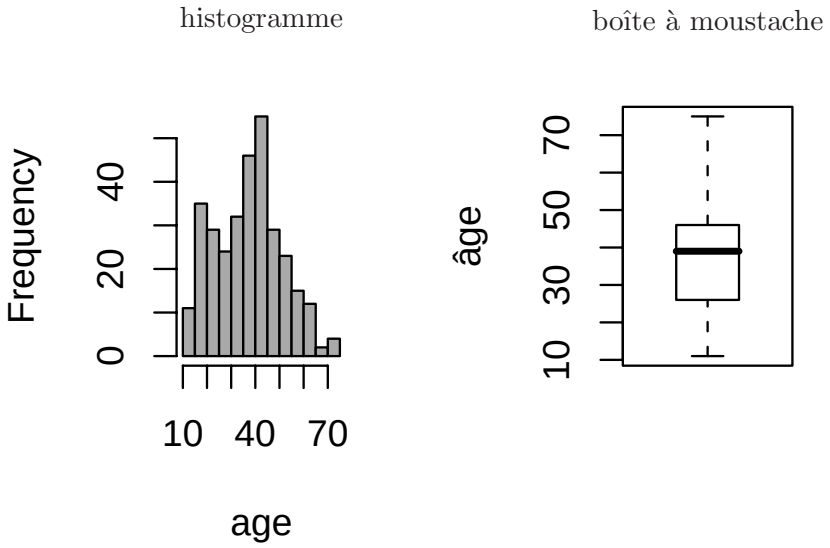

FIG. 2. La boîte de dispersion et l'histogramme de la variable âge.

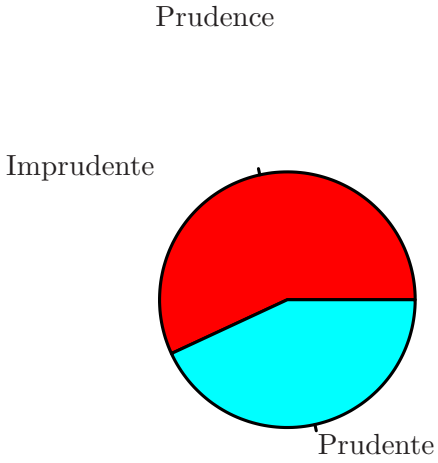

FIG. 3. La camembert de la variable prudence

	Imprudente	Prudente
Débutante	22	20
Débrouillée	65	69
Confirmée	81	45
Experte	13	3

TAB. 1. Le tableau croisé niveau de pratique/prudence

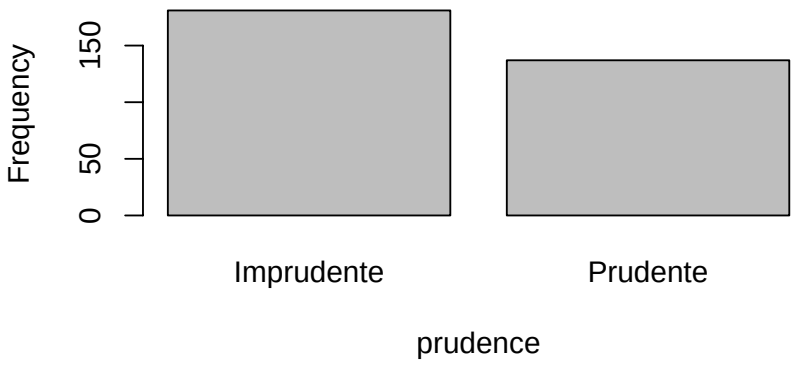

FIG. 4. Le graphe en barre de la variable prudence

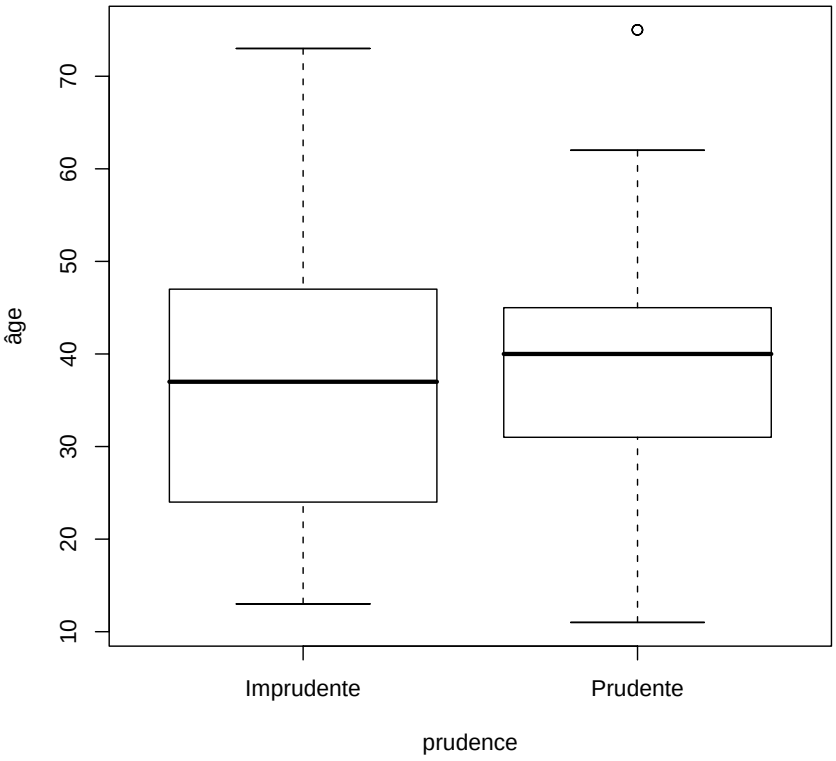

FIG. 5. collection de boîte de dispersion par groupe.