

TD de statistique : estimation

Jean-Baptiste Lamy

6 octobre 2008

1 Estimation avec R

1.1 Moyenne et écart type

Pour cela on utilise les fonctions `mean()` et `sd()` (cf TD 1).

1.2 Calcul d'intervalles de confiance

1.2.1 Variable binaire (loi binomiale ou de Bernoulli)

Les variables binaires, c'est-à-dire ne pouvant prendre que 2 valeurs (par exemple booléenne : malade / pas malade,...), suivent une loi de Bernoulli dont l'espérance est p et la variance est $V = p \times (1 - p)$. On approxime l'espérance par la moyenne.

Le calcul d'un intervalle de confiance à 95% se fait alors ainsi :

```
> n          = length(variable_binaire)
> moyenne    = mean(variable_binaire)
> variance   = moyenne * (1 - moyenne)
> moyenne - 1.96 * sqrt(variance / n); moyenne + 1.96 * sqrt(variance / n)
```

1.2.2 Variable numérique (loi de Student / loi normale)

Pour calculer un intervalle de confiance à 95% de la moyenne d'une variable numérique :

```
> t.test(variable_numerique)$conf.int[1 :2]
```

Il est possible de précéder le taux de confiance (ici 90%) :

```
> t.test(variable_numerique, conf.level=0.90)$conf.int[1 :2]
```

2 Exercice 1

Un échantillon de 100 patients a reçu un nouveau traitement contre la migraine. A la fin du traitement, 57 patients disent avoir perçu une amélioration de leur état migraineux.

1. À quelle loi de probabilité avons-nous à faire ?

Réponse : Loi de Bernoulli.

2. Donner une estimation ponctuelle du pourcentage de patients satisfaits par ce nouveau traitement.

Réponse : 57%.

3. Donner un intervalle de confiance au risque 5% du pourcentage de patients satisfaits par ce nouveau traitement.

Réponse :

```
> xbar = 0.57
> n = 100
> C = 1.96
> sigma = xbar * (1 - xbar)
> xbar - C * sigma / sqrt(n); xbar + C * sigma / sqrt(n)
[1] 0.5219604
[1] 0.6180396

> require(boot)
> norm.ci(t0=xbar, var.t0 = sigma ^ 2 / n)
```

3 Exercice 2

On cherche à étudier la prévalence de la grippe aviaire chez les canards. Pour cela, une première étude a porté sur un échantillon de 10 canards sauvages ; parmi ceux-ci, 1 seul était atteint de grippe.

1. À quelle loi de probabilité avons-nous à faire ?

Réponse : Loi de Bernouilli.

2. Calculer la prévalence de la grippe aviaire sur cet échantillon.

Réponse :

```
> n = 10
> xbar = 1 / n
> xbar
[1] 0.1
```

3. Déterminer un interval de confiance à 95% de la prévalence à partir de cet échantillon.

Réponse :

```
> C = 1.96
> sigma = xbar * (1 - xbar)
> xbar - C * sigma / sqrt(n) ; xbar + C * sigma / sqrt(n)
[1] 0.04421742
[1] 0.1557826
```

4. Une nouvelle étude est réalisé et porte cette fois-ci sur 100 canards ; parmi ceux-ci, 10 étaient atteints de grippe. Calculer la prévalence de la grippe aviaire sur ce nouvel échantillon, ainsi qu'un interval de confiance à 95% de la prévalence.

Réponse :

```
> n = 100
> xbar = 10 / n
> xbar
[1] 0.1
> sigma = xbar * (1 - xbar)
> xbar - C * sigma / sqrt(n) ; xbar + C * sigma / sqrt(n)
[1] 0.08236
[1] 0.11764
```

4 Exercice 3

Pour déterminer la concentration en glucose d'un échantillon sanguin, on effectue des dosages à l'aide d'une technique expérimentale donnée. On considère que le résultat de chaque dosage est une variable aléatoire normale. On effectue 10 dosages indépendants, qui donnent les résultats suivants (en g/l) : 0.96, 1.04, 1.08, 0.92, 1.04, 1.18, 0.99, 0.99, 1.25, 1.08

1. Calculer une estimation de la concentration en glucose de cet échantillon.

Réponse :

```
> g = c(0.96, 1.04, 1.08, 0.92, 1.04, 1.18, 0.99, 0.99, 1.25, 1.08)
> mean(g)
[1] 1.053
```

2. Calculer un interval de confiance de cette concentration à 95%.

Réponse :

```
> t.test(g)$conf.int[1 :2]
[1] 0.981064 1.124936
```

5 Exercice 4

Un laboratoire cosmétique souhaite vérifier l'absence de propriétés irritantes sur trois nouveaux produits, afin de pouvoir les inclure dans des shampoings. Pour cela, des gouttes de produit sont placés dans l'oeil de lapins, et on mesure dix minutes après la rougeur de l'oeil du lapin (unité arbitraire). L'expérience est réalisée sur quarante lapins, et chaque lapin reçoit successivement les trois produits ainsi qu'un témoin (à base de savon de marseille).

1. Le tableau de données est enregistré dans le fichier lapin.csv. Charger ce fichier.

Réponse :

```
> t = read.table("lapin.csv", sep=";", header=TRUE)
> attach(t)
```

2. Tracer un graphique pour représenter ces données.

Réponse :

```
> boxplot(temoin, produit1, produit2, produit3)
```

3. Calculer la rougeur moyenne pour le témoin et pour chaque produit, ainsi que leurs écart-type.

Réponse :

```
> mean(temoin); mean(produit1); mean(produit2); mean(produit3)
[1] 5.005509
[1] 4.985366
[1] 7.900932
[1] 5.241466
> sd(temoin); sd(produit1); sd(produit2); sd(produit3)
[1] 0.3267342
[1] 0.5758515
[1] 0.6088226
[1] 0.5163729
```

4. Calculer un interval de confiance de la rougeur moyenne pour le témoin et pour chaque produit.

Réponse :

```
> t.test(temoin)$conf.int[1 :2]
[1] 4.901015 5.110004
> t.test(produit1)$conf.int[1 :2]
[1] 4.801200 5.169532
> t.test(produit2)$conf.int[1 :2]
[1] 7.706221 8.095643
> t.test(produit3)$conf.int[1 :2]
[1] 5.076322 5.406610
```

6 Exercice 5

Afin de tester la toxicité d'une variété de maïs OGM, 3 groupes de 10 rats a été nourri avec ce maïs. Le maïs OGM représentait 11% de la ration alimentaire dans le premier groupe, 22% dans le second, et 33% dans le troisième. Un quatrième groupe témoin de 60 rats a été nourri avec du maïs non-OGM. Après 90 jours, on mesure le poids du foie de chaque rat.

[d'après Gilles-Eric Séralini *et al.*, New analysis of a rat feeding study with a genetically modified maize reveals signs of hepatorenal toxicity, 2007, Archives of Environmental Contamination and Toxicology, version française :

http://www.criigen.org/full_article.pdf

1. Le tableau de données est enregistré dans le fichier `ogm.csv`. Charger ce fichier. Quelles sont les variables dont nous disposons? Sont-elles numériques ou qualitatives?

Réponse :

```
> t = read.table("ogm.csv", sep=";", header=TRUE)
> attach(t)
```

La variable `ogm` peut être considérée comme numérique ou comme qualitative, car elle n'a que 4 valeurs possibles : 0%, 11%, 22%, 33%. Lorsque l'on veut la traiter comme variable qualitative, il faudra faire : `factor(ogm)`.

2. Représenter graphiquement le nombre de rats dans chacun des 4 groupes (groupe à 0%, 11%, 22% et 33% d'OGM). Peut-on expliquer pourquoi le nombre de rats est plus important dans le groupe à 0%?

Réponse :

```
> hist(ogm)
> pie(summary(factor(ogm)))
```

3. Représenter graphiquement la répartition des rats mâles et femelles au sein des 4 groupes.

Réponse :

```
> plot(factor(ogm), sexe)
```

4. Représenter graphiquement la taille du foie en fonction de la quantité d'OGM présent dans l'alimentation des rats.

Réponse :

```
> plot(foie, ogm)
> boxplot(foie ~ ogm)
```

5. Calculer la moyenne du poids du foie et son interval de confiance à 95%, sur chacun des groupes de rats.

Réponse :

```
> attach(t[t["ogm"] == 0,])
> mean(foie); t.test(foie)$conf.int[1 :2]
[1] 15.97238
[1] 15.71223 16.23254
> attach(t[t["ogm"] == 11,])
> mean(foie); t.test(foie)$conf.int[1 :2]
[1] 17.22328
[1] 16.79696 17.64960
> attach(t[t["ogm"] == 22,])
> mean(foie); t.test(foie)$conf.int[1 :2]
[1] 17.705
[1] 16.80134 18.60866
> attach(t[t["ogm"] == 33,])
> mean(foie); t.test(foie)$conf.int[1 :2]
[1] 18.18769
[1] 17.55258 18.82280
> attach(t)
```

6. Chez cette espèce de rat, on considère que le poids du foie est anormal s'il dépasse 17g. Ajouter une nouvelle colonne au tableau indiquant pour chaque rat si le poids de son foie est anormal ou non.

Réponse :

```
> t["gros_foie"] = foie > 17
> attach(t)
```

7. Représenter graphiquement la répartition foie normal / anormal au sein de chacun des 4 groupes.

Réponse :

```
> plot(factor(ogm), factor(gros_foie))
```

8. Donner une estimation et un interval de confiance à 95% de la probabilité d'avoir un foie anormal pour un rat ayant mangé 33% d'OGM.

Réponse :

```
> n = nrow(t[ogm == 33,]) # 10
> moyenne = mean(t[ogm == 33,]["gros_foie"])
> moyenne = nrow(t[gros_foie & ogm == 33,]) / n
> moyenne
[1] 0.9
> variance = moyenne * (1 - moyenne)
> moyenne - 1.96 * sqrt(variance / n); moyenne + 1.96 * sqrt(variance / n)
0.1900968 0.8099032
```

9. Même question chez les rats mâles, et chez les rats femelles. Qu'en déduisez-vous ?

Réponse :

```
> m = t[t["sexe"] == "male",]
> attach(m)
> n = nrow(m[ogm == 33,]) # 6
> moyenne = mean(m[ogm == 33,]["gros_foie"])
> moyenne = nrow(m[gros_foie & ogm == 33,]) / n
> moyenne
[1] 0.6666667
> variance = moyenne * (1 - moyenne)
> moyenne - 1.96 * sqrt(variance / n); moyenne + 1.96 * sqrt(variance / n)
[1] 0.2894645
[1] 1.043869 1

> m = t[t["sexe"] == "femelle",]...
[1] 0.25
[1] -0.1743524
[1] 0.6743524
```

Lors du prochain TD, nous verrons les test statistiques !