Cours de biostatistique ⋈ Illustrations dans 

Des électeurs, des boules, des cercles, des étudiants satisfaits

A.B. Dufour, J.R. Lobry & D. Chessel

31 mars 2008

Le calcul des probabilités parle de l'échantillon à partir de la population. La statistique inférentielle parle de la population à partir d'un échantillon. Quelques illustrations.

Table des matières

1 Introduction

Une *population* est un ensemble d'objets, d'individus de même nature. Cette population est généralement un ensemble très grand voire infini. Tous les étudiants de France constituent une population de même que tous les résultats possibles du tirage du loto. Une population peut donc être constituée d'individus réels ou d'individus fictifs. On parle alors *individus statistiques*. Cette population se caractérise par un ensemble de propriétés. Reprenons l'exemple des étudiants de France. Nous pouvons ajouter les critères suivants : discipline "mathématiques", âge "18 à 20 ans",...

Généralement, collecter des informations sur l'ensemble de la population est impossible et si cela est possible, le coût en est très élevé. C'est pourquoi, on extrait de la population, de manière judicieuse, un ensemble d'individus. Cet ensemble est appelé *échantillon*. L'opération qui consiste à extrait des individus d'une population s'appelle un *échantillonnage* ou un *sondage*.

Il existe plusieurs types d'échantillonnage.

1. L'*échantillonnage aléatoire simple* consiste à extraire des individus de la population tels que chacun d'entre eux a la même probabilité d'être choisi. Les "extractions" sont indépendantes les unes des autres.
2. Une autre procédure classique est l'*échantillonnage stratifié*. La stratification est un pas en direction du contrôle expérimental. Elle opère par sous-groupes de compositions plus homogènes à l'intérieur de la population. C'est le cas des sondages politiques où l'on classe les électeurs par catégories socio-professionnelles. D'autres divisions sont possibles comme une opposition rural/urbain, niveau d'éducation, sexe, âge ... En d'autres termes, les sous-groupes de la population sont construits à partir d'une variable qui est supposée corrélée de manière significative avec la variable à étudier. Une fois mis en place les variables importantes pour construire l'échantillon, la population totale est étudiée afin d'établir les différents pourcentages nécessaires dans chaque catégorie. Puis un échantillon aléatoire simple est réalisé dans chaque catégorie.
3. Il existe d'autres types d'échantillonnage. L'*échantillon baromètre* est un échantillon choisi arbitrairement parce qu'il est évident qu'il es très représentatif de la population totale. L'expérience a montré que dans les enquêtes d'opinion publique, certains états, certaines régions, reflètent l'opinion nationale. Cette population limitée est alors un bon baromètre de la population totale. Ce n'est cependant pas une très bonne procédure car elle nécessite l'obtention d'un grand nombre d'informations.
4. L'échantillon *boule de neige* consiste à trouver par exemple trois individus possédant une caractéristique spécifique donnée, demander à chacun d'eux de trouver trois individus possédant la caractéristique en question et ainsi de suite.

L'objectif de cette fiche est de montrer les relations entre la population de départ et les échantillons.

2 De la connaissance de la réalité à la simulation

2.1 Des électeurs dans une ville

La *fiabilité* d'un échantillon dépend du nombre de personnes interrogées et non pas du nombre de personnes dans la population. Cela est du au fait que l'on pose la même question à plusieurs personnes et on compte le nombre de réponses. Pour mieux comprendre le phénomène, imaginons des élections politiques dans une ville de 100000 habitants. Un candidat de gauche et un candidat de droite s'affrontent pour conquérir la mairie. 60% des habitants votent à gauche (soit 60000 personnes) et 40% votent à droite (soit 40000 personnes).
Dans un premier temps, un échantillon est constitué d'**un** individu choisi au hasard dans la population. On a 3 chances sur 5 que cette personne soit de gauche (60000 sur 100000) et 2 chances sur 5 qu'elle soit de droite (40000 sur 100000). On dit alors que la probabilité que la personne interrogée soit de gauche est de $\frac{3}{5}$ et que la probabilité qu'elle soit de droite est de $\frac{2}{5}$. Pour notre très mini sondage, nous avons les résultats suivants.

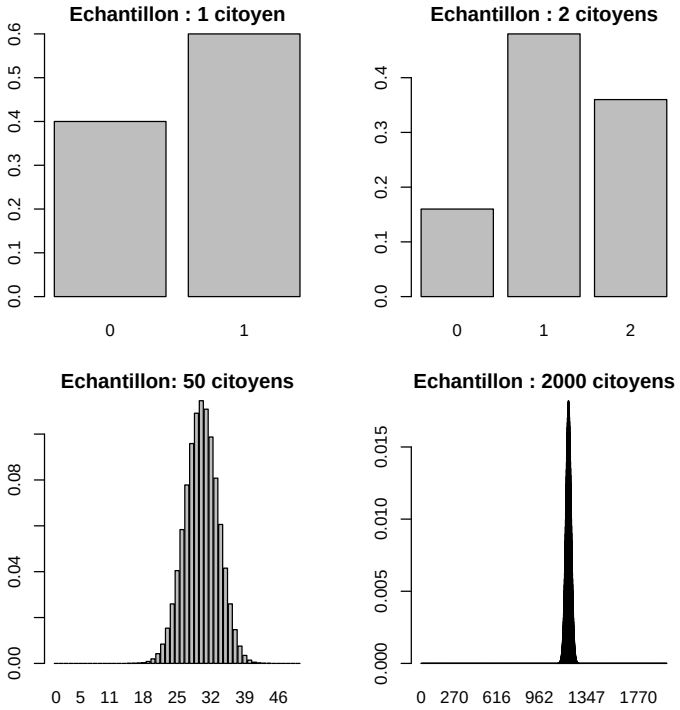
gauche	droite	probabilité
1	0	$\frac{3}{5}$
0	1	$\frac{2}{5}$

Prenons toujours au hasard deux individus de la population. La probabilité pour que le premier soit de gauche et le deuxième de droite s'obtient en multipliant les probabilités entre elles soit $\frac{3}{5} * \frac{2}{5} = \frac{6}{25}$. Mais le premier individu aurait très bien pu voter à droite et le deuxième à gauche. Nous avons donc deux possibilités. Les résultats possibles se retrouvent dans le tableau ci-dessous.

gauche	droite	probabilité
2	0	$\frac{3}{5} * \frac{3}{5} = \frac{9}{25}$
1	1	$(\frac{3}{5} * \frac{2}{5}) * 2 = \frac{12}{25}$
0	2	$\frac{2}{5} * \frac{2}{5} = \frac{4}{25}$

La somme des probabilités vaut 1. On dit alors que le "nombre de votants à gauche" est une *variable aléatoire* dont la distribution est une *loi binomiale*.
Tous ces calculs peuvent être refait en faisant varier le nombre d'individus dans l'échantillon. Avec un sondage de 50 personnes, il y a 51 possibilités pour la variable aléatoire depuis "aucune personne ne vote à gauche" jusqu'à "toutes les personnes votent à gauche". Il en est de même pour un échantillon de 2000 personnes. Les représentations graphiques associées à ces distributions de probabilité sont des représentations en bâtons où l'axe des abscisses donne les valeurs "nombre de personnes votant à gauche" et l'axe des ordonnées les probabilités de la loi binomiale.

```
par(mfrow = c(2, 2))
par(mar = c(3, 3, 2, 2))
barplot(dbinom(x = 0:1, size = 1, prob = 0.6), names.arg = as.character(0:1),
    main = "Echantillon : 1 citoyen")
barplot(dbinom(x = 0:2, size = 2, prob = 0.6), names.arg = as.character(0:2),
    main = "Echantillon : 2 citoyens")
barplot(dbinom(x = 0:50, size = 50, prob = 0.6), names.arg = as.character(0:50),
    main = "Echantillon: 50 citoyens")
barplot(dbinom(x = 0:2000, size = 2000, prob = 0.6), names.arg = as.character(0:2000),
    main = "Echantillon : 2000 citoyens")
```



Raisonnons maintenant avec l'idée que le sondage est le reflet exact de ce que l'on connait dans la population c'est-à-dire 60% de personnes votent à gauche. Calculons les probabilités d'avoir exactement 60% des individus de l'échantillon votant à gauche en fonction de différentes tailles possibles.

```
taillespossibles <- c(1, 50, 100, 150, 200, 500, 1000, 1500, 2000)
exact60 <- as.integer(0.6 * taillespossibles)
exact60
```

```
[1]    0   30   60   90  120  300  600  900 1200
```

```
dbinom(x = exact60, size = taillespossibles, prob = 0.6)
```

```
[1] 0.40000000 0.11455855 0.08121914 0.06637351 0.05750643 0.03639907 0.02574482
[8] 0.02102241 0.01820674
```

La probabilité d'avoir 60 % des électeurs votant à gauche dans un échantillon de 2000 personnes vaut 1.8 %. Il est donc peu probable que le sondage soit parfaitement exact. Nous sommes donc près à tolérer une certaine erreur.
Fixons par exemple une *marge d'erreur* à 1 %. Cela signifie que le pourcentage de personnes votant à gauche doit être compris entre 59 % et 61 % ou encore que le nombre de personnes votant à gauche soit compris entre 1180 et 1220 parmi les 2000 personnes interrogées. Pour calculer cette probabilité, il suffit d'ajouter les probabilités d'avoir exactement 1180 personnes, puis 1181, 1182, ..., 1220.

```
2000 * 0.59
```

```
[1] 1180
```

```
2000 * 0.61
```

```
[1] 1220
```

```
dbinom(x = 1180:1220, size = 2000, prob = 0.6)
```

```
 [1] 0.01196885 0.01246544 0.01295584 0.01343771 0.01390871 0.01436646 0.01480860
 [8] 0.01523278 0.01563667 0.01601806 0.01637476 0.01670473 0.01700603 0.01727688
[15] 0.01751563 0.01772083 0.01789122 0.01802574 0.01812355 0.01818401 0.01820674
[22] 0.01819158 0.01813861 0.01804814 0.01792072 0.01775713 0.01755836 0.01732561
[29] 0.01706027 0.01676394 0.01643836 0.01608543 0.01570718 0.01530576 0.01488340
[36] 0.01444241 0.01398515 0.01351400 0.01303136 0.01253961 0.01204111
```

```
sum(dbinom(x = 1180:1220, size = 2000, prob = 0.6))
```

```
[1] 0.6505674
```

Il ya donc 65 % de chances que le résultat du sondage donne entre 59 % et 61 % de personnes votant à gauche.

Exercice. Calculer le pourcentage pour une marge d'erreur de 2 % dans un échantillon de 2000 individus.

```
[1] 0.935509
```

Exercice. Calculer le pourcentage pour une marge d'erreur de 2 % dans un échantillon de 5000 individus.

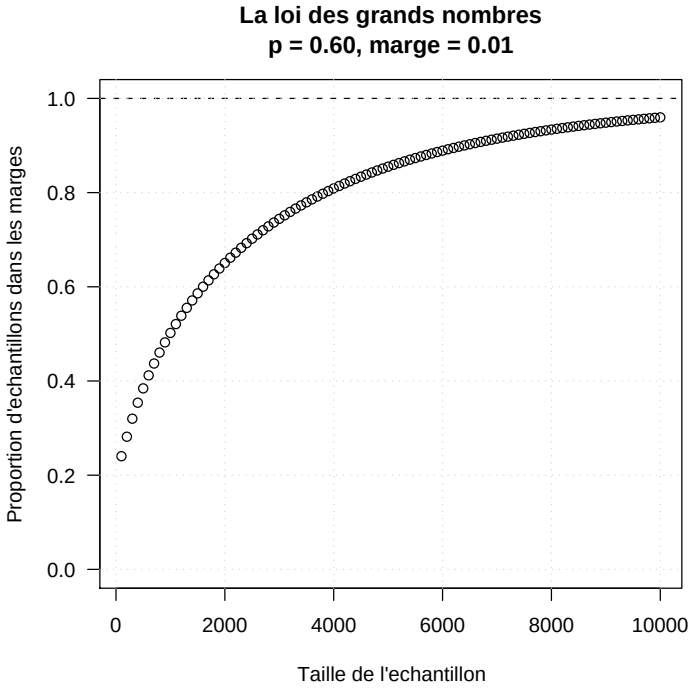
```
[1] 0.9962865
```

Supposons maintenant que nos candidats ne se présentent pas dans une ville de 100000 habitants mais dans un pays de 60 millions d'habitants. Supposons encore que le pourcentage de personnes votant à gauche est de 60 % (36 millions de personnes) et que le pourcentage de personnes votant à droite est de 40 % (24 millions de personnes). Que l'on interroge une personne, deux personnes, les probabilités sont les mêmes que précédemment (cf tableaux). L'information "taille de la population" n'intervient pas dans le calcul des probabilités. Comme l'écrit Gilles Dowek [1] :

> "...Ce qui rend un sondage fiable, ce n'est pas la ressemblance entre le sondage et l'élection, c'est la loi des grands nombres, selon laquelle si, lors d'une épreuve, un évenement a une probabilité p de se produire alors, quand on répète l'épreuve plusieurs fois, la proportion de cas dans lesquels cet événement se produit se rapproche de p quand le nombre de fois que l'on répète l'épreuve augmente. Cette proportion a déjà une chance importante d'être proche de p quand on dépasse quelques milliers de personnes..."

Nous pouvons illustrer ceci avec le graphique suivant :

```
npoints <- 100
pourcentage <- numeric(npoints)
proba = 0.6
marge = 0.01
tailles <- as.integer(seq(from = 100, to = 10000, length = npoints))
for (i in 1:npoints) {
    mini <- as.integer((tailles[i] * (proba - marge)))
    maxi <- as.integer((tailles[i] * (proba + marge)))
    pourcentage[i] <- sum(dbinom(x = mini:maxi, size = tailles[i],
        prob = proba))
}
plot(x = tailles, y = pourcentage, las = 1, ylim = c(0, 1), xlab = "Taille de l'echantillon",
    main = "La loi des grands nombres\np = 0.60, marge = 0.01",
    ylab = "Proportion d'echantillons dans les marges")
abline(h = 1, lty = 2)
grid()
```



2.2 Des boules dans une urne

Une situation expérimentale classique est le tirage de boules dans une urne. Pour cela, on construit une urne contenant 30 boules blanches (notées "B") et 70 boules noires (notées "N").

```
urne <- rep(x = c("B", "N"), times = c(30, 70))
urne
```

```
 [1] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "B"
[21] "B" "B" "B" "B" "B" "B" "B" "B" "B" "B" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N"
[41] "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N"
[61] "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N"
[81] "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N"
```

On extrait maintenant sans remise un échantillon de 20 boules de cette urne ; on compte le nombre de boules blanches et de boules noires ; on compte plus simplement le nombre de boules blanches.

```
echan <- sample(urne, 20)
echan
```

```
 [1] "N" "B" "B" "N" "N" "N" "N" "N" "N" "N" "N" "N" "N" "B" "N" "N" "N" "N" "B" "N" "N"
```

```
table(echan)
```

```
echan
 B  N
 4 16
```

```
sum(echan == "B")
```

```
[1] 4
```

Répéter l'expérience :

```
echan <- sample(urne, 20)
echan
```

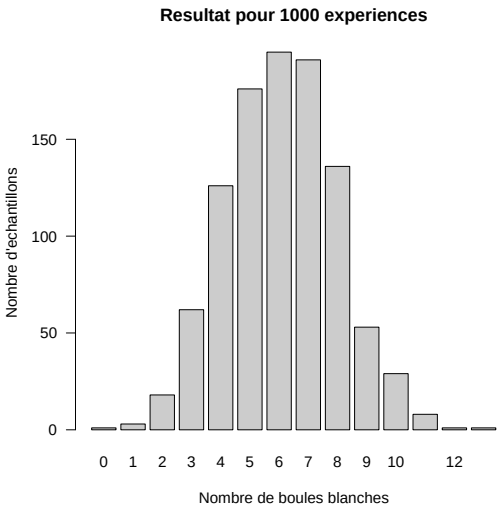
```
 [1] "B" "N" "N" "N" "N" "B" "N" "N" "N" "N" "N" "B" "N" "B" "B" "B" "N" "N" "N" "N"
 table(echan)
echan
 B  N
 6 14
 sum(echan == "B")
[1] 6
```

Répéter encore l'expérience :

```
 echan <- sample(urne, 20)
 echan
 [1] "N" "N" "B" "N" "N" "N" "N" "N" "B" "N" "B" "N" "B" "B" "N" "N" "N" "N" "B" "N"
 table(echan)
echan
 B  N
 6 14
 sum(echan == "B")
[1] 6
```

On n'obtient pas toujours le même résultat, on dit que le nombre de boules blanches est une variable aléatoire. On étudie maintenant la variable aléatoire "nombre de boules blanches". Pour cela, on réalise 1000 fois l'expérience précédente.

```
 resultat <- numeric(1000)
 for (i in 1:1000) resultat[i] <- sum(sample(urne, 20) == "B")
 resultat[1:20]
 [1] 6 7 5 3 4 6 7 4 6 7 5 4 8 5 5 4 6 9 5 6
 table(resultat)
resultat
  1   2   3   4   5   6   7   8   9  10  11
  5  12  56 123 223 210 192 108  48  17   6
 barplot(table(resultat), col = grey(0.8), las = 1, xlab = "Nombre de boules blanches",
     ylab = "Nombre d'echantillons", main = "Resultat pour 1000 experiences")
```

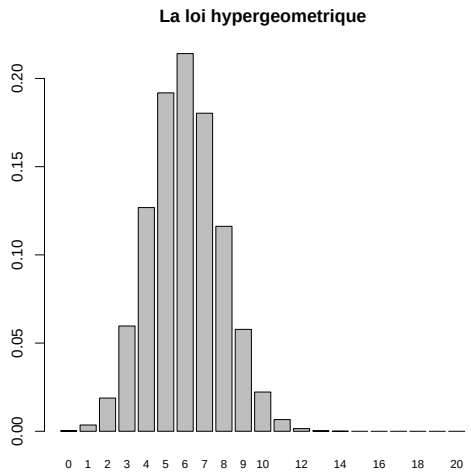


L'aide de la fonction `sample()` montre que les tirages sont *sans remise*. La variable aléatoire "nombre de boules blanches" suit une distribution discrète appelée *loi hypergéométrique*. Elle est parfaitement connue. On donne ci-dessous l'ensemble des probabilités de la variable : de la réponse "aucune boule blanche n'est extraite de l'urne" à "les 20 boules blanches sont extraites de l'urne" ainsi que sa représentation graphique.

```
dhyper(x = 0:20, m = 30, n = 70, k = 20)
```

```
 [1] 3.020329e-04 3.553328e-03 1.882581e-02 5.967425e-02 1.268078e-01 1.918256e-01
 [7] 2.140911e-01 1.802872e-01 1.161765e-01 5.776005e-02 2.223762e-02 6.628202e-03
[13] 1.523417e-03 2.678536e-04 3.557431e-05 3.502701e-06 2.487714e-07 1.223108e-08
[19] 3.897157e-10 7.134384e-12 5.605587e-14
```

```
barplot(dhyper(x = 0:20, m = 30, n = 70, k = 20), names.arg = 0:20,
    main = "La loi hypergeometrique", cex.names = 0.75)
```



On a vu ici deux approches : l'approche informatique par la simulation de 1000 échantillons et l'approche mathématique. Quand on sait faire le calcul, c'est parfait. Quand on a une approximation mathématique (par la simulation), si les conditions d'approximation sont respectées, c'est bon. Quand on ignore la solution, on peut l'approcher et la précision ne dépend que du temps de calcul.

Exercice.

1. Construire une urne contenant 200 boules dont 120 blanches et 80 noires.
2. Extraire 30 boules de l'urne sans remise. Construire les représentations en bâtons de la variable aléatoire "nombre de boules blanches" :
   suite à une simulation de 1000 échantillons,
   suite à une simulation de 10000 échantillons,
   avec la loi de probabilité connue.
3. Extraire 30 boules de l'urne avec *remise* (voir la documentation de la fonction `sample()`). Une fois notée la couleur de la boule tirée, celle-ci est replacée dans l'urne. La variable aléatoire suit alors une loi binomiale (`dbinom()`). Même étude que précédemment.

2.3 Des cercles de rayons différents

Une population est constituée de 100 cercles [2] de rayons différents. On considère la variable "rayon" dont les unités de mesure ne sont pas précisées. On peut calculer la moyenne $\mu = 5.35$ et la variance $\sigma^2 = 7.46$. La représentation graphique associée est l'histogramme.

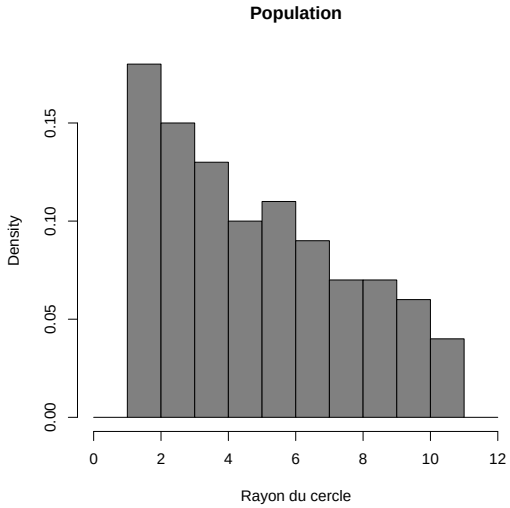
```
cercles <- read.table("http://pbil.univ-lyon1.fr/R/donnees/rayon.txt",
    header = TRUE)
mean(cercles$rayon)
```

```
[1] 5.35
```

```
var(cercles$rayon)
```

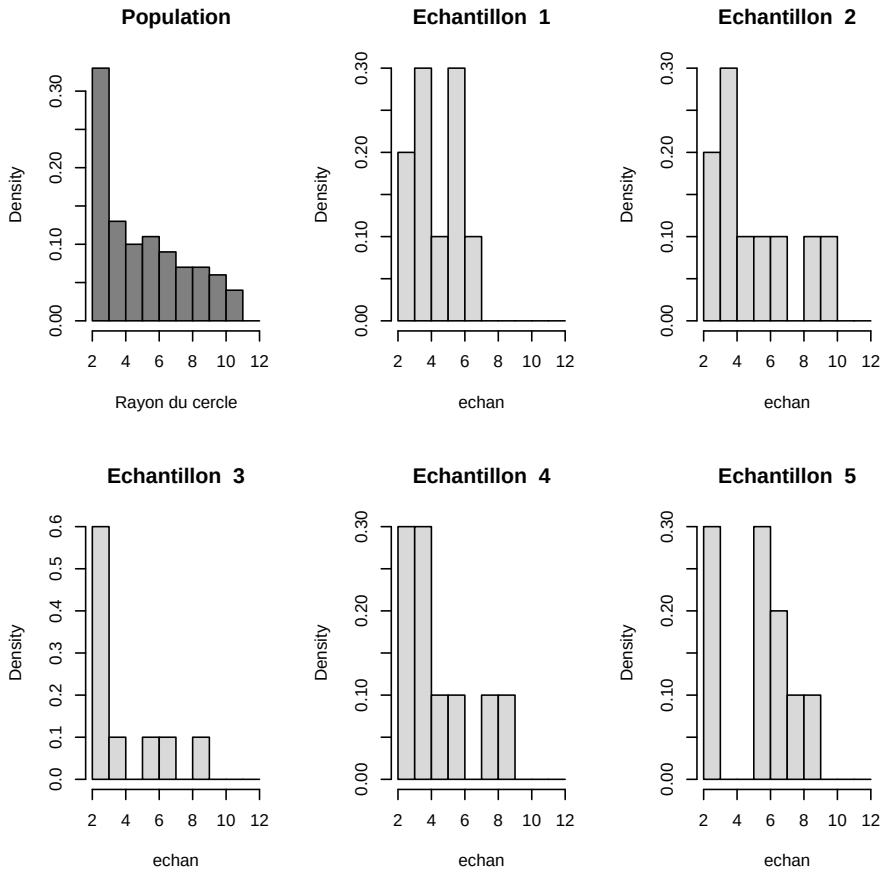
```
[1] 7.462121
```

```
brks <- 0:12
hist(cercles$rayon, main = "Population", breaks = brks, col = grey(0.5),
    proba = TRUE, xlab = "Rayon du cercle")
```



On va construire quelques échantillons extraits de manière aléatoire de cette population et faire varier la taille de ces échantillons. Pour cela, écrire la fonction `simulation()` puis exécuter là avec par exemple des échantillons de taille $n = 10$.

```
simulation <- function(n) {
    brks <- 2:12
    par(mfrow = c(2, 3))
    hist(cercles$rayon, main = "Population", breaks = brks, col = grey(0.5),
        proba = TRUE, xlab = "Rayon du cercle")
    for (i in 1:5) {
        titre <- paste("Echantillon ", i)
        echan <- sample(cercles$rayon, n)
        hist(echan, breaks = brks, proba = TRUE, main = titre, col = grey(0.85))
    }
}
simulation(10)
```



Réaliser d'autres simulations en augmentant la taille `n` des échantillons. Que peut-on conclure ?

3 De l'influence des échantillons sur le calcul des paramètres

3.1 Des électeurs dans une ville

Reprenons l'exemple de notre sondage politique dans une ville de 100000 électeurs dont 60 % votent à gauche :

```
electeurs <- rep(c("G", "D"), c(60000, 40000))
```

Partons du fait que nous ignorons le nombre d'électeurs votant à gauche : faisons comme si la composition de l'urne `electeurs` nous était inconnue. Réalisons un sondage sur 2000 personnes.

```
sondage <- sum(sample(electeurs, 2000) == "G")
sondage
```

```
[1] 1207
```

Nous comptons donc ici 1207 électeurs votant à gauche soit une proportion de 0.604.

Réalisons un deuxième sondage.

```
sondage <- sum(sample(electeurs, 2000) == "G")
sondage
```

```
[1] 1172
```

Nous comptons 1172 électeurs votant à gauche soit une proportion de 0.586.

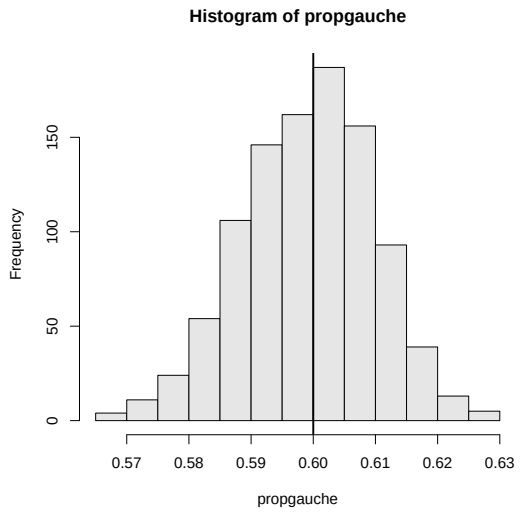
Recommençons encore une fois :

```
sondage <- sum(sample(electeurs, 2000) == "G")
sondage
```

```
[1] 1197
```

Nous comptons 1197 électeurs votant à gauche soit une proportion de 0.598. Nous voyons que la *vraie* valeur de la proportion (0.6) est approchée. Utilisons le logiciel pour faire une simulation sur 1000 échantillons de taille 2000.

```
simsond <- function(n) {
    echantillon <- sample(electeurs, n)
    propgauche <- numeric(1000)
    for (i in 1:1000) {
        propgauche[i] <- sum(sample(electeurs, n) == "G")/n
    }
    hist(propgauche, col = grey(0.9))
    abline(v = 0.6, lwd = 2)
    return(mean(propgauche))
}
moy2000 <- simsond(2000)
```



Nous obtenons 1000 fréquences. Ces valeurs représentent une nouvelle variable aléatoire appelée *distribution d'échantillonnage*. La moyenne de cette distribution d'échantillonnage est 0.599378, ce qui est assez proche de la vraie valeur (0.6)

Exercice. Refaire l'expérience précédente, en utlisant la fonction `simsond()` avec des échantillons de tailles différentes.

3.2 Des cercles de rayons différents

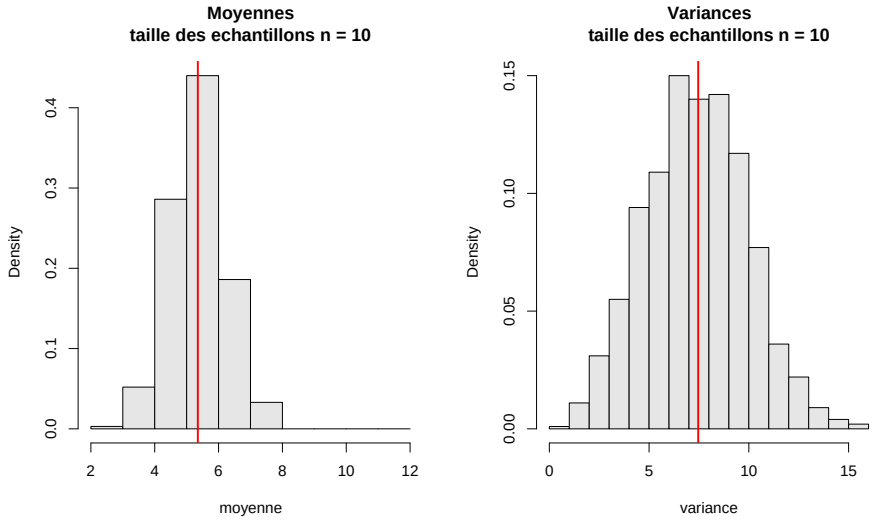
Dans le paragraphe précédent, on a vu que quelle que soit la taille de l'échantillon, avec le hasard, l'échantillon peut prendre une allure plus ou moins proche de la population de départ. Qu'advient-il alors de la moyenne et la variance ? Sur quelques exemples, on a obtenu les résultats suivants :

effectif	moyenne	variance
5	6.40	7.80
5	6.80	6.70
10	4.50	6.28
10	5.60	6.49
20	4.85	7.08
20	5.5	6.89

On va réaliser des simulations de 1000 échantillons de taille n. Chaque échantillon donne une moyenne et une variance. L'ensemble des 1000 moyennes constitue la distribution d'échantillonnage de la moyenne et l'ensemble des 1000 variances constitue la distribution d'échantillonnage de la variance. Que peut-on dire de ces deux distributions ?

Pour cela, utiliser la fonction `simmoyvar()` définie ci-dessous :

```
simmoyvar <- function(n) {
    moyenne <- numeric(1000)
    variance <- numeric(1000)
    for (i in 1:1000) {
        echan <- sample(cercles$rayon, n)
        moyenne[i] <- mean(echan)
        variance[i] <- var(echan)
    }
    brks <- 2:12
    par(mfrow = c(1, 2))
    titre <- paste("Moyennes\ntaille des echantillons n =", n)
    hist(moyenne, col = grey(0.9), main = titre, breaks = brks,
        proba = TRUE)
    abline(v = mean(cercles$rayon), lwd = 2, col = "red")
    titre <- paste("Variances\ntaille des echantillons n =", n)
    hist(variance, col = grey(0.9), main = titre, proba = TRUE)
    abline(v = var(cercles$rayon), lwd = 2, col = "red")
}
simmoyvar(10)
```



Exercice. Refaire l'expérience précédente, en utlisant la fonction `simmoyvar()` avec des échantillons de tailles différentes.

4 Des étudiants satisfaits

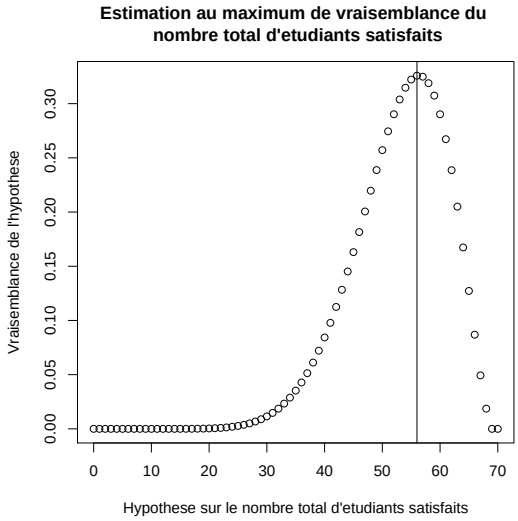
Il y a dans l'amphi 70 étudiants. Un enseignant, Mr Biomod, désire savoir combien d'étudiants sont satisfaits de ce qu'il raconte. Bien sûr le plus efficace serait d'interroger tous les étudiants mais le temps de comptage lui prendrait toute la séance. Il décide d'en interroger 10 au hasard. 8 sont satisfaits et 2 non. Mr Biomdo peut dire alors que 80 % des étudiants interrogés sont satisfaits par ce qu'il dit.
Si l'enseignant s'intéresse au nombre de satisfaits dans l'amphi, il a 71 possibilités : de 0 satisfait à tous satisfaits. L'enseignant peut par exemple émettre l'hypothèse que dans l'amphi, 45 étudiants sont satisfaits de son cours (et donc 25 non satisfaits). Quelle est la probabilité qu'il ait obtenu 8 satisfactions dans son échantillon de 10 ? Nous retrouvons là l'expression de la loi hypergéométrique.

```
dhyper(x = 8, m = 45, n = 25, k = 10)
```

```
[1] 0.1630079
```

Ce calcul peut être réitéré pour les 71 hypothèses possibles. Plaçons en abscisses les 71 hypothèses et en ordonnée la probabilité d'avoir 8 étudiants satisfaits sur 10 compte tenu du nombre supposé d'étudiants satisfaits dans l'amphi.

```
nstot <- 0:70
proba8 <- numeric(71)
for (i in nstot) {
    proba8[i] <- dhyper(x = 8, m = nstot[i], n = 70 - nstot[i],
        k = 10)
}
plot(nstot, proba8, xlab = "Hypothese sur le nombre total d'etudiants satisfaits",
    ylab = "Vraisemblance de l'hypothese", main = paste("Estimation au maximum de vraisemblance du \n",
        "nombre total d'etudiants satisfaits"))
abline(v = nstot[which.max(proba8)])
```



L'hypothèse correspondant à la plus grande probabilité (0.3258470) est l'hypothèse de 56 étudiants satisfaits :

```
 max(proba8)
[1] 0.3258470
 nstot[which.max(proba8)]
[1] 56
```

La probabilité d'observer le résultat sous une hypothèse arbitraire est la *vraisemblance* de cette hypothèse pour l'observation donnée. **Estimer** c'est choisir une des hypothèses. Estimer au *maximum de vraisemblance* c'est choisir l'hypothèse qui donne à l'observation la plus grande vraisemblance. L'estimation au maximum de vraisemblance du nombre total d'étudiants satisfaits dans l'amphi est donc ici de 56. On évite de parler d'estimation au maximum de probabilité parce que :

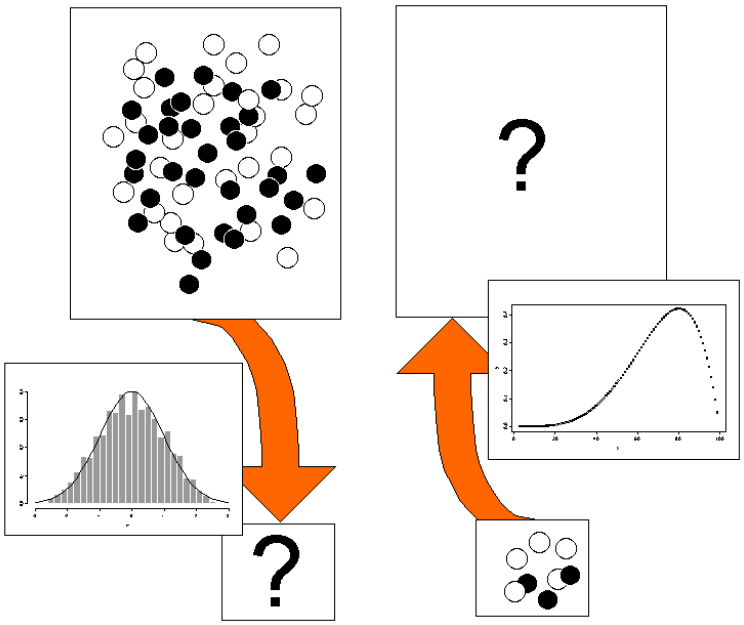
```
 sum(proba8)
[1] 6.454545
```

C'est sans importance pratique parce que nous sommes intéressés par les valeurs relatives des vraisemblances, et non à leur valeurs dans l'absolu, pour déterminer la plus grande.
Exercice. Quelle est l'estimation au maximum de vraisemblance du nombre total d'étudiants satisfaits dans l'amphi si sur 10 étudiants choisis au hasard 7 sont satisfaits et 3 non ? Quelle conjecture pourriez vous avancer ?

5 Conclusion

De la connaissance d'un espace probabilisé, on peut déduire ce qui va se passer sur un échantillon, du genre si on tire une boule au hasard dans une urne contenant 7 rouges et 3 bleues, 7 fois sur 10 on aura une rouge ! Puis nous avons inversé totalement la question : si une urne contient 10 boules et qu'en tirant au hasard on obtient 3 rouges qu'est-ce qu'on peut dire des autres ? C'est

ce qu'on appelle l'inférence statistique. Le premier point de vue est celui des mathématiques, le second est celui des sciences expérimentales. Les deux sont très liés.



Le calcul des probabilités parle de l'échantillon à partir de la population. La statistique inférentielle parle de la population à partir d'un échantillon.

Références

[1] Dowek G. *Peut-on croire les sondages ? Les Petites Pommes du Savoir.* Editions Le Pommier, Dijon, France, 2002.

[2] R. Tomassone, C. Dervin, and J.P. Masson. *Modélisation de phénomènes biologiques.* Masson, Paris, 1993.